Journal of Multivariate Analysis 102 (2011) 993–1007

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

On qualitative robustness of support vector machines

Robert Hable *, Andreas Christmann

Department of Mathematics, University of Bayreuth, D-95440 Bayreuth, Germany

ARTICLE INFO

Article history:
Received 21 May 2010
Available online 12 March 2011

AMS 2000 subject classifications:
62G08
62G35

Keywords:
Classification
Machine learning
Nonparametric regression
Qualitative robustness
Support vector machines

ABSTRACT

Support vector machines (SVMs) have attracted much attention in theoretical and in applied statistics. The main topics of recent interest are consistency, learning rates and robustness. We address the open problem whether SVMs are qualitatively robust. Our results show that SVMs are qualitatively robust for any fixed regularization parameter λ. However, under extremely mild conditions on the SVM, it turns out that SVMs are not qualitatively robust any more for any null sequence λ_n, which are the classical sequences needed to obtain universal consistency. This lack of qualitative robustness is of a rather theoretical nature because we show that, in any case, SVMs fulfill a finite sample qualitative robustness property.

For a fixed regularization parameter, SVMs can be represented by a functional on the set of all probability measures. Qualitative robustness is proven by showing that this functional is continuous with respect to the topology generated by weak convergence of probability measures. Combined with the existence and uniqueness of SVMs, our results show that SVMs are the solutions of a well-posed mathematical problem in Hadamard's sense.

1. Introduction

Two of the most important topics in statistics are classification and regression. Both settings assume that the outcome $y \in \mathcal{Y}$ of a random variable Y (output variable) is influenced by an observed value $x \in \mathcal{X}$ (input variable). On the basis of a finite data set $((x_1, y_1), \ldots, (x_n, y_n)) \in (\mathcal{X} \times \mathcal{Y})^n$, the goal is to find an "optimal" predictor $f : \mathcal{X} \to \mathcal{Y}$ which makes a prediction $f(x)$ for an unobserved y. In parametric statistics, a signal plus noise relationship

$$y = f_\theta(x) + \varepsilon$$

is often assumed, where f_θ is precisely known except for a finite parameter $\theta \in \mathbb{R}^p$ and ε is an error term (generated from a Normal distribution). In this way, the goal of estimating an "optimal" predictor (which can be any function $f : \mathcal{X} \to \mathcal{Y}$) reduces to the much simpler task of estimating the parameter $\theta \in \mathbb{R}^p$. Since, in many applications, such strong assumptions can hardly be justified, nonparametric regression has been developed which avoids (or at least considerably weakens) such assumptions. In statistical machine learning, the method of support vector machines has been developed as a method of nonparametric regression; see e.g., [34,26,31]. In this framework, the estimation of the predictor (called *empirical SVM*) is a function f which solves the minimization problem

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

where H is a certain function space. The first term in (1) is the empirical mean of the losses caused by the predictions $f(x_i)$, as measured by the loss function L. That is, the first term rates the quality of the predictor f. The second term penalizes the

* Corresponding author.
E-mail addresses: Robert.Hable@uni-bayreuth.de (R. Hable), Andreas.Cristmann@uni-bayreuth.de (A. Christmann).

0047-259X/$ – see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmva.2011.01.009

complexity of f in order to avoid overfitting, λ is a positive real number. The space H is a reproducing kernel Hilbert space (RKHS) which consists of functions $f : \mathcal{X} \to \mathbb{R}$.

Since the rise of robust statistics [32,19], it is well-known that imperceptible small deviations of the real world from model assumptions may lead to arbitrarily wrong conclusions; see e.g. [23,24,22] for some recent books on robust statistics. While many practitioners are aware of the need for robust methods in classical parametric statistics, it is quite often overseen that robustness is also a crucial issue in nonparametric statistics. For example, the sample mean can be seen as a nonparametric procedure which is non-robust since it is extremely sensitive to outliers: Let $X_1, \ldots, X_n$ be i.i.d. random variables with unknown distribution P and the task is to estimate the expectation of P. If the observed data are really generated by the ideal P (and if expectation and variance of P exist), then the sample mean is the optimal estimator. However, it frequently happens in the real world that, due to outliers or small model violations, the observed data are not generated by the ideal P but by another distribution P′. Even if P′ is close to the ideal P, the sample mean may lead to disastrous results. Detailed descriptions and some examples of such effects are given, e.g., in [32,19], and [20, Section 1.1].

In nonparametric regression, similar effects can occur. In this setting, it is often assumed that $(X_1, Y_1), \ldots, (X_n, Y_n)$ are i.i.d. random variables with unknown distribution P. This distribution P determines in which way the output variable Y_i is influenced by the input variable X_i. However, estimating a predictor $f : \mathcal{X} \to \mathcal{Y}$ can be severely distorted if the observed data $(x_1, y_1), \ldots, (x_n, y_n)$ are not generated by P but (due to small errors) by another distribution P′ which may be close to the ideal P. In order to safeguard from severe distortions, an estimator S_n should fulfill some kind of continuity: If the real distribution P′ is close to the ideal distribution P, then the distribution of the estimator S_n should hardly be affected (uniformly in the sample sizes $n \in \mathbb{N}$). This kind of robustness is called *qualitative robustness* and has been formalized in [17,18] for estimators taking values in $\mathbb{R}^p$.

In order to study this notion of robust statistics for support vector machines, we need a generalization of this formalization as given by [11] because, here, the values of the estimator are *functions* $f : \mathcal{X} \to \mathcal{Y}$ which are elements of a (typically infinite dimensional) Hilbert space H. In case of support vector machines, the estimators

$$S_n : (\mathcal{X} \times \mathcal{Y})^n \to H$$

can be represented by a functional

$$S : \mathcal{M}_1(\mathcal{X} \times \mathcal{Y}) \to H$$

on the set $\mathcal{M}_1(\mathcal{X} \times \mathcal{Y})$ of all probability measures on $\mathcal{X} \times \mathcal{Y}$:

$$S_n((x_1, y_1), \ldots, (x_n, y_n)) = S\left(\frac{1}{n}\sum_{i=1}^{n} \delta_{(x_i, y_i)}\right)$$

for every $(x_1, y_1), \ldots, (x_n, y_n) \in \mathcal{X} \times \mathcal{Y}$ where $\frac{1}{n}\sum_{i=1}^{n} \delta_{(x_i,y_i)}$ is the empirical measure and $\delta_{(x_i,y_i)}$ denotes the Dirac measure in (x_i, y_i). It is shown by [11] that, in such cases, the qualitative robustness of a sequence of estimators $(S_n)_{n\in\mathbb{N}}$ follows from the continuity of the functional S (with respect to the topology of weak convergence of probability measures). While quantitative robustness of support vector machines has already been investigated by means of Hampel's influence functions and bounds for the maxbias in [7] and by means of Bouligand influence functions in [8], results about qualitative robustness of support vector machines have not been published so far. The goal of this paper is to fill this gap on research on qualitative robustness of support vector machines.

Under very mild conditions, we obtain the following results: For fixed regularization parameters, support vector machines are qualitatively robust in the sense of [17,11]. If classical null sequences λ_n needed to obtain universal consistency are used, support vector machines are not qualitatively robust any more in this sense. Roughly speaking, qualitative robustness fails for null sequences λ_n because the notion of qualitative robustness originating from [17] not only requires a continuity property but even equicontinuity over all possible sample sizes and this conflicts with universal consistency. However, this lack of robustness is of a rather theoretical nature because, in applications, one is always faced with a finite sample of a fixed size and our results show that support vector machines are "qualitatively robust" for finite samples of any fixed size — a property which we will call *finite sample qualitative robustness*.

The structure of the article is as follows: In Section 2, we recall the basic setup concerning support vector machines, define the functional S which represents the SVM-estimators S_n, $n \in \mathbb{N}$, and give the mathematical definitions of qualitative robustness and finite sample qualitative robustness. In Section 3, we show that the functional S of support vector machines is, in fact, continuous under very mild assumptions (Theorem 3.3). This implies that support vector machines are qualitatively robust for any fixed regularization parameter $\lambda > 0$ and are finite sample qualitatively robust for every sequence of regularization parameters $(\lambda_n)_{n\in\mathbb{N}} \subset (0, \infty)$ (Theorem 3.1). We also demonstrate that, for null sequences $(\lambda_n)_{n\in\mathbb{N}}$, finite sample robustness can neither be strengthened to ordinary qualitative robustness in Hampel's sense [17,11] nor to qualitative robustness in the sense of [4] (qualitative resistance). As a byproduct, it follows from Theorem 3.3 that empirical support vector machines are continuous in the data – i.e., they are hardly affected by slight changes in the data (Corollary 3.5). Under somewhat different assumptions, this has already been shown in [31, Lemma 5.13].

Section 4 contains some concluding remarks. All proofs are given in the Appendix.

2. Support vector machines and qualitative robustness

Let $(\Omega, \mathcal{A}, \mathrm{Q})$ be a probability space and let $\mathcal{X}$ be a Polish space with Borel-σ-algebra $\mathfrak{B}(\mathcal{X})$. That is, $\mathcal{X}$ is a separable completely metrizable topological space (e.g., a closed subset of $\mathbb{R}^d$). Let $\mathcal{Y}$ be a closed subset of $\mathbb{R}$ with Borel-σ-algebra $\mathfrak{B}(\mathcal{Y})$. The Borel-$\sigma$-algebra of $\mathcal{X} \times \mathcal{Y}$ is denoted by $\mathfrak{B}(\mathcal{X} \times \mathcal{Y})$ and the set of all probability measures on $(\mathcal{X} \times \mathcal{Y}, \mathfrak{B}(\mathcal{X} \times \mathcal{Y}))$ is denoted by $\mathcal{M}_1(\mathcal{X} \times \mathcal{Y})$. Let

$$X_1, \ldots, X_n : (\Omega, \mathcal{A}, \mathrm{Q}) \longrightarrow (\mathcal{X}, \mathfrak{B}(\mathcal{X}))$$

and

$$Y_1, \ldots, Y_n : (\Omega, \mathcal{A}, \mathrm{Q}) \longrightarrow (\mathcal{Y}, \mathfrak{B}(\mathcal{Y}))$$

be random variables such that $(X_1, Y_1), \ldots, (X_n, Y_n)$ are independent and identically distributed according to some unknown probability measure $\mathrm{P} \in \mathcal{M}_1(\mathcal{X} \times \mathcal{Y})$.

A measurable map $L : \mathcal{X} \times \mathcal{Y} \times \mathbb{R} \to [0, \infty)$ is called a *loss function*. It is assumed that $L(x, y, y) = 0$ for every $(x, y) \in \mathcal{X} \times \mathcal{Y}$ — that is, the loss is zero if the prediction $f(x)$ equals the observed value y. In addition, we will assume that

$$L(x, y, \cdot) : \mathbb{R} \to [0, \infty), \qquad t \mapsto L(x, y, t)$$

is convex for every $(x, y) \in \mathcal{X} \times \mathcal{Y}$ and that the following uniform Lipschitz property is fulfilled for a positive real number $|L|_1 \in (0, \infty)$:

$$\sup_{(x,y) \in \mathcal{X} \times \mathcal{Y}} |L(x, y, t) - L(x, y, t')| \leq |L|_1 \cdot |t - t'| \quad \forall t, t' \in \mathbb{R}. \tag{2}$$

We restrict our attention to Lipschitz continuous loss functions because the use of loss functions which are not Lipschitz continuous (such as the least squares loss on unbounded domains) usually conflicts with several notions of robustness; see, e.g., [31, Section 10.4].

The *risk* of a measurable function $f : \mathcal{X} \to \mathbb{R}$ is defined by

$$\mathcal{R}_{L,\mathrm{P}}(f) = \int_{\mathcal{X} \times \mathcal{Y}} L(x, y, f(x)) \, \mathrm{P}(d(x, y))$$

where $\mathrm{P}(d(x, y))$ denotes integration by P with respect to x and y.

Let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a bounded and continuous *kernel* with *reproducing kernel Hilbert space* (RKHS) H. See e.g. [26] or [31] for details about these concepts. Note that H is a Polish space since every Hilbert space is complete and, according to [31, Lemma 4.29], H is separable. Furthermore, every $f \in H$ is a bounded and continuous function $f : \mathcal{X} \to \mathbb{R}$; see [31, Lemma 4.28]. In particular, every $f \in H$ is measurable and its *regularized risk* is defined to be

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

for $\lambda \in (0, \infty)$.

An element $f \in H$ is called a *support vector machine* and denoted by $f_{L,\mathrm{P},\lambda}$ if it minimizes the regularized risk in H. That is,

$$\mathcal{R}_{L,\mathrm{P}}(f_{L,\mathrm{P},\lambda}) + \lambda \|f_{L,\mathrm{P},\lambda}\|_H^2 = \inf_{f \in H} (\mathcal{R}_{L,\mathrm{P}}(f) + \lambda \|f\|_H^2).$$

We would like to consider a functional

$$S : \mathrm{P} \mapsto f_{L,\mathrm{P},\lambda}. \tag{3}$$

However, support vector machines $f_{L,\mathrm{P},\lambda}$ need not exist for every probability measure $\mathrm{P} \in \mathcal{M}_1(\mathcal{X} \times \mathcal{Y})$ and, therefore, S cannot be defined on $\mathcal{M}_1(\mathcal{X} \times \mathcal{Y})$ in this way. A sufficient condition for the existence of a support vector machine based on a bounded kernel k is, for example, $\mathcal{R}_{L,\mathrm{P}}(0) < \infty$; see [31, Corollary 5.3]. In order to enlarge the applicability of support vector machines, the following extension has been developed in [9]. Following an idea already used by [21] for M-estimates, a *shifted loss function* $L^* : \mathcal{X} \times \mathcal{Y} \times \mathbb{R} \to \mathbb{R}$ is defined by

$$L^*(x, y, t) = L(x, y, t) - L(x, y, 0) \quad \forall (x, y, t) \in \mathcal{X} \times \mathcal{Y} \times \mathbb{R}.$$

Then, similar to the original loss function L, define the L^*- risk by

$$\mathcal{R}_{L^*,\mathrm{P}}(f) = \int L^*(x, y, f(x)) \, \mathrm{P}(d(x, y))$$

and the regularized L^*- risk by

$$\mathcal{R}_{L^*,\mathrm{P},\lambda}(f) = \mathcal{R}_{L^*,\mathrm{P}}(f) + \lambda \|f\|_H^2$$

for every $f \in H$. In complete analogy to $f_{L,\mathrm{P},\lambda}$, we define the support vector machine based on the shifted loss function L^* by

$$f_{L^*,\mathrm{P},\lambda} = \arg\inf_{f \in H} (\mathcal{R}_{L^*,\mathrm{P}}(f) + \lambda \|f\|_H^2).$$

The following theorem summarizes some basic results derived by [9]:

Theorem 2.1. *For every $\mathrm{P} \in \mathcal{M}_1(\mathcal{X} \times \mathcal{Y})$ and every $\lambda \in (0, \infty)$, there exists a unique $f_{L^*,\mathrm{P},\lambda} \in H$ which minimizes $\mathcal{R}_{L^*,\mathrm{P},\lambda}$, i.e.*

$$\mathcal{R}_{L^*,\mathrm{P}}(f_{L^*,\mathrm{P},\lambda}) + \lambda \|f_{L^*,\mathrm{P},\lambda}\|_H^2 = \inf_{f \in H} (\mathcal{R}_{L^*,\mathrm{P}}(f) + \lambda \|f\|_H^2).$$

If a support vector machine $f_{L,P,\lambda} \in H$ exists (which minimizes $\mathcal{R}_{L,P,\lambda}$ in H), then

$$f_{L^*,P,\lambda} = f_{L,P,\lambda}.$$

According to this theorem, the map

$$S: \mathcal{M}_1(\mathcal{X} \times \mathcal{Y}) \to H, \qquad P \mapsto f_{L^*,P,\lambda}$$

exists, is uniquely defined and extends the functional in (3). Therefore, S may be called an *SVM-functional*.

In order to estimate a measurable map $f: \mathcal{X} \to \mathbb{R}$ which minimizes the risk

$$\mathcal{R}_{L,P}(f) = \int_{\mathcal{X}\times\mathcal{Y}} L(x, y, f(x))\, P(d(x, y)),$$

the *SVM-estimator* is defined by

$$S_n: (\mathcal{X} \times \mathcal{Y})^n \to H, \qquad D_n \mapsto f_{L,D_n,\lambda_n}$$

where f_{L,D_n,λ_n} is that function $f \in H$ which minimizes

$$\frac{1}{n}\sum_{i=1}^{n} L(x_i, y_i, f(x_i)) + \lambda_n \|f\|_H^2$$

in H for $D_n = ((x_1, x_2), \ldots, (x_n, y_n)) \in (\mathcal{X} \times \mathcal{Y})^n$ and $\lambda_n \in (0, \infty)$ is the regularization parameter. Let $\mathbb{P}_{D_n}$ be the empirical measure corresponding to the data D_n for sample size $n \in \mathbb{N}$. Then, for $\lambda = \lambda_n$, the definitions given above yield

$$f_{L,D_n,\lambda_n} = S_n(D_n) = S(\mathbb{P}_{D_n}) = f_{L,\mathbb{P}_{D_n},\lambda_n}. \tag{4}$$

Note that the support vector machine uniquely exists for every empirical measure. In particular, this also implies $f_{L,D_n,\lambda_n} = f_{L^*,\mathbb{P}_{D_n},\lambda_n}$.

The main goal of the article is to investigate qualitative robustness of the sequence of SVM-estimators $(S_n)_{n\in\mathbb{N}}$. According to [17] and [11, Definition 1], the sequence $(S_n)_{n\in\mathbb{N}}$ is called *qualitatively robust* if the functions

$$\mathcal{M}_1(\mathcal{X} \times \mathcal{Y}) \to \mathcal{M}_1(H), \qquad P \mapsto S_n(P^n), \quad n \in \mathbb{N},$$

are equicontinuous with respect to the weak topologies on $\mathcal{M}_1(\mathcal{X} \times \mathcal{Y})$ and $\mathcal{M}_1(H)$. Occasionally, we will replace equicontinuity by continuity and call $(S_n)_{n\in\mathbb{N}}$ *finite sample qualitatively robust* then. Here, $\mathcal{M}_1(H)$ denotes the set of all probability measures on $(H, \mathfrak{B}(H))$, $\mathfrak{B}(H)$ is the Borel-σ-algebra on H, and $S_n(P^n)$ denotes the image measure of P^n with respect to S_n. Hence, $S_n(P^n)$ is the measure on $(H, \mathfrak{B}(H))$ which is defined by

$$(S_n(P^n))(F) = P^n(\{D_n \in (\mathcal{X} \times \mathcal{Y})^n \mid S_n(D_n) \in F\})$$

for every Borel-measurable subset $F \subset H$. Of course, this definition only makes sense if the SVM-estimators are measurable with respect to the Borel-σ-algebras. This measurability is assured by Corollary 3.5 below.

Since the weak topologies on $\mathcal{M}_1(\mathcal{X} \times \mathcal{Y})$ and $\mathcal{M}_1(H)$ are metrizable by the Prokhorov metric d_{Pro} (see Appendix A.1), the sequence of SVM-estimators $(S_n)_{n\in\mathbb{N}}$ is qualitatively robust if and only if for every $P \in \mathcal{M}_1(\mathcal{X} \times \mathcal{Y})$ and every $\rho > 0$ there is an $\varepsilon > 0$ such that

$$d_{\mathrm{Pro}}(Q, P) < \varepsilon \;\Rightarrow\; d_{\mathrm{Pro}}(S_n(Q^n), S_n(P^n)) < \rho \quad \forall n \in \mathbb{N}.$$

The sequence of SVM-estimators $(S_n)_{n\in\mathbb{N}}$ is finite sample qualitatively robust if and only if for every $P \in \mathcal{M}_1(\mathcal{X} \times \mathcal{Y})$ and every $\rho > 0$ and every $n \in \mathbb{N}$, there is an $\varepsilon_n > 0$ such that

$$d_{\mathrm{Pro}}(Q, P) < \varepsilon_n \;\Rightarrow\; d_{\mathrm{Pro}}(S_n(Q^n), S_n(P^n)) < \rho.$$

Roughly speaking, qualitative robustness and finite sample qualitative robustness mean that the SVM-estimator tolerates two kinds of errors in the data: small errors in many observations (x_i, y_i) and large errors in a small fraction of the data set. These two kinds of errors only have slight effects on the distribution and, therefore, on the performance of the SVM-estimator. Fig. 1 gives a graphical illustration of qualitative robustness.

3. Main results

The following theorem is our main result and shows that support vector machines are (finite sample) qualitatively robust under mild conditions.

Theorem 3.1. *Let $\mathcal{X}$ be a Polish space and let $\mathcal{Y}$ be a closed subset of $\mathbb{R}$. Let the loss function be a continuous function $L: \mathcal{X} \times \mathcal{Y} \times \mathbb{R} \to [0, \infty)$ such that $L(x, y, y) = 0$ for every $(x, y) \in \mathcal{X} \times \mathcal{Y}$ and*

$$L(x, y, \cdot): \mathbb{R} \to [0, \infty), \qquad t \mapsto L(x, y, t)$$

is convex for every $(x, y) \in \mathcal{X} \times \mathcal{Y}$. Assume that L fulfills the uniform Lipschitz Property (2) for a real number $|L|_1 \in (0, \infty)$. Furthermore, let $k: \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a bounded and continuous kernel with RKHS H.

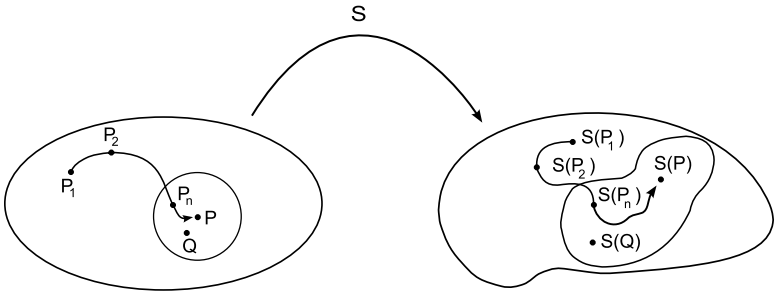

Fig. 1. Sketch: reasoning of robustness of $S(\mathrm{P})$. Left: P, a neighborhood of P, and $\mathcal{M}_1(\mathcal{X} \times \mathcal{Y})$. Right: $S(\mathrm{P})$, a neighborhood of $S(\mathrm{P})$, and the space of all probability measures of $S(\mathrm{P})$ for $\mathrm{P} \in \mathcal{M}_1(\mathcal{X} \times \mathcal{Y})$.

Then, the sequence of SVM-estimators $(S_n)_{n\in\mathbb{N}}$ is finite sample qualitatively robust for every sequence of regularization parameters $(\lambda_n)_{n\in\mathbb{N}} \subset (0, \infty)$. If the regularization parameters $\lambda_n = \lambda \in (0, \infty)$ do not depend on $n \in \mathbb{N}$, then the sequence of SVM-estimators $(S_n)_{n\in\mathbb{N}}$ is qualitatively robust.

Of course, this theorem applies to classification (e.g. $\mathcal{Y} = \{-1, 1\}$) and regression (e.g. $\mathcal{Y} = \mathbb{R}$ or $\mathcal{Y} = [0, \infty)$). In particular, note that every function $g : \mathcal{Y} \to \mathbb{R}$ is continuous if $\mathcal{Y}$ is a discrete set – e.g. $\mathcal{Y} = \{-1, 1\}$. That is, in this case, assuming L to be continuous reduces to the assumption that

$$\mathcal{X} \times \mathbb{R} \to [0, \infty), \qquad (x, t) \mapsto L(x, y, t)$$

is continuous for every $y \in \mathcal{Y}$. Many of the most common loss functions are permitted in the theorem, e.g. the hinge loss and logistic loss for classification, ε-insensitive loss and Huber's loss for regression, and the pinball loss for quantile regression. The least squares loss is ruled out in Theorem 3.1 – which is not surprising as it is the prominent standard example of a loss function which typically conflicts with robustness if $\mathcal{X}$ and $\mathcal{Y}$ are unbounded; see, e.g., [7,8].

Assuming continuity of the kernel k does not seem to be very restrictive as all of the most common kernels are continuous. Assuming k to be bounded is quite natural in order to ensure good robustness properties. While the Gaussian RBF kernel is always bounded, polynomial kernels (except for the constant kernel) and the exponential kernel are bounded if and only if $\mathcal{X}$ is bounded.

Our result shows qualitative robustness in the sense of [17,11] only for fixed regularization parameters λ which do not depend on the sample size. However, it is necessary to choose appropriate null sequences $(\lambda_n)_{n\in\mathbb{N}} \subset (0, \infty)$ in order to prove universal consistency of the risk $\mathcal{R}_{L^*,\mathrm{P}}(f_{L^*,D_n,\lambda_n}) \xrightarrow{\mathrm{P}} \inf_{f\in\mathcal{F}} \mathcal{R}_{L^*,\mathrm{P}}(f)$ and $f_{L^*,D_n,\lambda_n} \xrightarrow{\mathrm{P}} \arg\inf_{f\in\mathcal{F}} \mathcal{R}_{L^*,\mathrm{P}}(f)$ for $n \to \infty$ where $\mathcal{F}$ denotes the set of all *measurable* functions $f : \mathcal{X} \to \mathbb{R}$. Universal consistency of support vector machines was shown by [27, 35,28]. We also refer to [6,1,9,30]. For sequences $(\lambda_n)_{n\in\mathbb{N}} \subset (0, \infty)$, our result only shows *finite sample* qualitative robustness of support vector machines. Though this is weaker than ordinary qualitative robustness, it is comparably meaningful in applications because, in applications, one is always faced with a finite sample of a fixed size.

In case of null sequences $(\lambda_n)_{n\in\mathbb{N}} \subset (0, \infty)$, finite sample robustness *cannot* be strengthened to ordinary qualitative robustness. The following proposition shows that, for null sequences $(\lambda_n)_{n\in\mathbb{N}} \subset (0, \infty)$, support vector machines *cannot* be qualitatively robust in the in the sense of [17,11]. This shows that the asymptotic results on *universal consistency* of support vector machines – which require appropriate null sequences $(\lambda_n)_{n\in\mathbb{N}} \subset (0, \infty)$ – are in conflict with Hampel's standard notion of qualitative robustness. Such a partial conflict between consistency and qualitative robustness also happens in ill-posed estimation problems where the goal is to estimate the value of a discontinuous functional $T : P \mapsto T(P)$. In this case, *every* consistent estimator is *not* qualitatively robust in the ordinary sense as follows from [18, Lemma 3] and [11, Theorem 1]. This happens for example in nonparametric density estimation as has been pointed out in [11, Section 2]. In addition, support vector machines also cannot be qualitatively robust in the sense of [4] (qualitative resistance) for null sequences $(\lambda_n)_{n\in\mathbb{N}}$ because qualitative resistance of support vector machines would imply qualitative robustness in Hampel's sense; see [4, Theorems 3.1, 4.1 and 4.2] and [18, Theorem 3]. For simplicity, the following proposition focuses on regression because it is assumed that $\{0, 1\} \subset \mathcal{Y}$. A similar proposition (with a similar proof) can also be given in case of binary classification with support vector machines where usually $\mathcal{Y} = \{-1, 1\}$.

Proposition 3.2. *Let $\mathcal{X}$ be a Polish space and let $\mathcal{Y}$ be a closed subset of $\mathbb{R}$ such that $\{0, 1\} \subset \mathcal{Y}$. Let k be a bounded kernel with RKHS H. Let L be a convex loss function such that $L(x, y, y) = 0$ for every $(x, y) \in \mathcal{X} \times \mathcal{Y}$. In addition, assume that there are $x_0, x_1 \in \mathcal{X}$ such that*

$$\exists \tilde{f} \in H : \quad \tilde{f}(x_0) = 0, \qquad \tilde{f}(x_1) \neq 0 \tag{5}$$

$$L(x_1, 1, 0) > 0. \tag{6}$$

Let $(\lambda_n)_{n\in\mathbb{N}} \subset (0, \infty)$ be any sequence such that $\lim_{n\to\infty} \lambda_n = 0$. Then, the sequence of estimators

$$S_n : (\mathcal{X} \times \mathcal{Y})^n \to H, \qquad D_n \mapsto f_{L,D_n,\lambda_n}, \quad n \in \mathbb{N},$$

is not qualitatively robust.

Note that, assumptions (5) and (6) in Proposition 3.2 are virtually always fulfilled except for degenerated cases (e.g. k a constant function or $L \equiv 0$).

The proof of our main result, Theorem 3.1, is based on the following theorem which is interesting on its own.

Theorem 3.3. *Let $\lambda \in (0, \infty)$ be fixed. Under the assumptions of Theorem 3.1, the SVM-functional*

$$S : \mathcal{M}_1(\mathcal{X} \times \mathcal{Y}) \to H, \qquad P \mapsto f_{L^*,P,\lambda}$$

is continuous with respect to the weak topology on $\mathcal{M}_1(\mathcal{X} \times \mathcal{Y})$ and the norm topology on H.

As a generalization of earlier results by, e.g., [36,13,29], [9, Theorem 7] derived a representer theorem which showed that for *every* $P_0 \in \mathcal{M}_1(\mathcal{X} \times \mathcal{Y})$, there is a bounded map $h_{P_0} : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ such that $f_{L^*,P_0,\lambda} = -\frac{1}{2\lambda} \int h_{P_0} \Phi \, dP_0$ and

$$\|f_{L^*,P,\lambda} - f_{L^*,P_0,\lambda}\|_H \le \lambda^{-1} \left\| \int h_{P_0} \Phi \, dP - \int h_{P_0} \Phi \, dP_0 \right\| \tag{7}$$

for every $P \in \mathcal{M}_1(\mathcal{X} \times \mathcal{Y})$. The integrals in (7) are Bochner integrals of the H-valued function $h_{P_0}\Phi : \mathcal{X} \times \mathcal{Y} \to H$, $(x, y) \mapsto h_{P_0}(x, y)\Phi(x)$ where Φ is the canonical feature map of k, i.e. $\Phi(x) = k(\cdot, x)$ for all $x \in \mathcal{X}$. This offers an elegant possibility of proving Theorem 3.3 if we would accept some additional smoothness assumptions: The statement of Theorem 3.3 is true if $\int h_{P_0} \Phi \, dP_n$ converges to $\int h_{P_0} \Phi \, dP_0$ for every weakly convergent sequence $P_n \to P_0$. In the following, we show that the integrals indeed converge – under the additional smoothness assumptions that the derivative $\frac{\partial L}{\partial t}(x, y, t)$ exists and is continuous for every $(x, y, t) \in \mathcal{X} \times \mathcal{Y} \times \mathbb{R}$. These assumptions are fulfilled e.g. for the logistic loss function and Huber's loss function. In this case, it follows from [9, Theorem 7] that h_{P_0} is continuous. Since Φ is continuous and bounded (see e.g. [31, p. 124 and Lemma 4.29]), the integrand $h_{P_0}\Phi : \mathcal{X} \times \mathcal{Y} \to H$ is continuous and bounded. Then, it follows from [5, p. III.40] that $\int h_{P_0} \Phi \, dP_n$ converges to $\int h_{P_0} \Phi \, dP_0$ for every weakly convergent sequence $P_n \to P_0$, just as in case of real-valued integrands; see Appendix A.1.

Unfortunately, this short proof only works under the additional assumption of a continuous partial derivative $\frac{\partial L}{\partial t}$ and this assumption rules out many loss functions used in practice, such as hinge, absolute distance and ε-insensitive for regression and pinball for quantile regression. Therefore, our proof of Theorem 3.3 (without this additional assumption) does not use the representer theorem and Bochner integrals; it is mainly based on the theory of Hilbert spaces and weak convergence of measures. In the following, we give some corollaries of Theorem 3.3.

Let $\mathcal{C}_b(\mathcal{X})$ be the Banach space of all bounded, continuous functions $f : \mathcal{X} \to \mathbb{R}$ with norm

$$\|f\|_\infty = \sup_{x \in \mathcal{X}} |f(x)|.$$

Since k is continuous and bounded, we immediately get from Theorem 3.3 and [31, Lemma 4.28]:

Corollary 3.4. *Let $\lambda \in (0, \infty)$ be fixed. Under the assumptions of Theorem 3.1, the SVM-functional*

$$\mathcal{M}_1(\mathcal{X} \times \mathcal{Y}) \to \mathcal{C}_b(\mathcal{X}), \qquad P \mapsto f_{L^*,P,\lambda}$$

is continuous with respect to the weak topology on $\mathcal{M}_1(\mathcal{X} \times \mathcal{Y})$ and the norm topology on $\mathcal{C}_b(\mathcal{X})$.

That is, $\sup_{x \in \mathcal{X}} |f_{L,P',\lambda}(x) - f_{L,P,\lambda}(x)|$ is small if P' is close to P.

The next corollary is similar to [31, Lemma 5.13]. The main difference is that Corollary 3.5 does not assume differentiability of the loss function but assumes Lipschitz continuity of the loss function and boundedness of the kernel instead. Therefore, Corollary 3.5 also covers the popular hinge loss, the ε-insensitive loss and the pinball loss, which are not covered by [31, Lemma 5.13]. In combination with existence and uniqueness of support vector machines (see Theorem 2.1), this result shows that a support vector machine is the solution of a well-posed mathematical problem in the sense of [16].

Corollary 3.5. *Under the assumptions of Theorem 3.1, the SVM-estimator*

$$S_n : (\mathcal{X} \times \mathcal{Y})^n \to H, \qquad D_n \mapsto f_{L,D_n,\lambda_n}$$

is continuous for every $\lambda_n \in (0, \infty)$ and every $n \in \mathbb{N}$.

In particular, it follows from Corollary 3.5 that the SVM-estimator S_n is measurable.

Remark 3.6. Let d_n be a metric which generates the topology on $(\mathcal{X} \times \mathcal{Y})^n$, e.g. the Euclidean metric on $\mathbb{R}^{n(q+1)}$ if $\mathcal{X} \subset \mathbb{R}^q$. Then Corollary 3.5 and [31, Lemma 4.28] imply the following continuity property of the SVM-estimator: For every $\varepsilon > 0$ and every data set $D_n \in (\mathcal{X} \times \mathcal{Y})^n$, there is a $\delta > 0$ such that

$$\sup_{x \in \mathcal{X}} |f_{L,D_n',\lambda_n}(x) - f_{L,D_n,\lambda_n}(x)| < \varepsilon$$

if $D_n' \in (\mathcal{X} \times \mathcal{Y})^n$ is any other data set with n observations and $d_n(D_n', D_n) < \delta$.

We finish this section with a corollary about strong consistency of support vector machines which arises as a by-product of Theorem 3.3. Often, asymptotic results of support vector machines show the convergence in probability of the risk $\mathcal{R}_{L^*,P}(f_{L^*,\mathbb{D}_n,\lambda_n})$ to the so-called *Bayes risk* $\inf_{f \in \mathcal{F}} \mathcal{R}_{L^*,P}(f)$ and of $f_{L^*,\mathbb{D}_n,\lambda_n}$ to $\arg\inf_{f \in \mathcal{F}} \mathcal{R}_{L^*,P}(f)$, where $\mathcal{F}$ is the set of all

measurable functions $f : \mathcal{X} \to \mathbb{R}$ and $(\lambda_n)_{n\in\mathbb{N}}$ is a suitable null sequence. In contrast to that, the following corollary provides for *fixed* $\lambda \in (0, \infty)$ almost sure convergence of $\mathcal{R}_{L^*,P}(f_{L^*,\mathbb{D}_n,\lambda})$ to $\mathcal{R}_{L^*,P}(f_{L^*,P,\lambda})$ and of $f_{L^*,\mathbb{D}_n,\lambda}$ to $f_{L^*,P,\lambda}$. This is an interesting fact, although the limit $\mathcal{R}_{L^*,P}(f_{L^*,P,\lambda})$ will in general differ from the Bayes risk.

Recall from Section 2 that the data points (x_i, y_i) from the data set $D_n = ((x_1, x_2), \ldots (x_n, y_n))$ are realizations of i.i.d. random variables

$$(X_i, Y_i) : (\Omega, \mathcal{A}, Q) \longrightarrow (\mathcal{X} \times \mathcal{Y}, \mathfrak{B}(\mathcal{X} \times \mathcal{Y})), \quad n \in \mathbb{N},$$

such that

$$(X_i, Y_i) \sim P \quad \forall n \in \mathbb{N}.$$

Corollary 3.7. *Define the random vectors*

$$\mathbb{D}_n := ((X_1, Y_1), \ldots, (X_n, Y_n))$$

and the corresponding H-valued random functions

$$f_{L^*,\mathbb{D}_n,\lambda} = \arg\inf_{f\in H} \left(\frac{1}{n} \sum_{i=1}^{n} L^*(X_i, Y_i, f(X_i)) + \lambda \|f\|_H^2 \right), \quad n \in \mathbb{N}.$$

From the assumptions of Theorem 3.1, it follows that

(a) $\lim_{n\to\infty} \|f_{L^*,\mathbb{D}_n,\lambda} - f_{L^*,P,\lambda}\|_H = 0$ *almost sure*
(b) $\lim_{n\to\infty} \sup_{x\in\mathcal{X}} |f_{L^*,\mathbb{D}_n,\lambda}(x) - f_{L^*,P,\lambda}(x)| = 0$ *almost sure*
(c) $\lim_{n\to\infty} \mathcal{R}_{L^*,P,\lambda}(f_{L^*,\mathbb{D}_n,\lambda}) = \mathcal{R}_{L^*,P,\lambda}(f_{L^*,P,\lambda})$ *almost sure*
(d) $\lim_{n\to\infty} \mathcal{R}_{L^*,P}(f_{L^*,\mathbb{D}_n,\lambda}) = \mathcal{R}_{L^*,P}(f_{L^*,P,\lambda})$ *almost sure.*

If the support vector machine $f_{L,P,\lambda}$ exists, then assertions (a)–(d) *are also valid for L instead of L^*.*

4. Conclusions

It is well-known that outliers in data sets or other moderate model violations can pose a serious problem to a statistical analysis. On the one hand, practitioners can hardly guarantee that their data sets do not contain any outliers, while, on the other hand, many statistical methods are very sensitive even to small violations of the assumed statistical model. Since support vector machines play an important role in statistical machine learning, investigating their performance in the presence of moderate model violations is a crucial topic — the more so as support vector machines are frequently applied to large and complex high-dimensional data sets.

In this article, we showed that support vector machines are qualitatively robust for fixed regularization parameters $\lambda \in (0, \infty)$. For sequences of regularization parameters $(\lambda_n)_{n\in\mathbb{N}} \subset (0, \infty)$, we showed that support vector machines still enjoy a finite sample qualitative robustness property. In any case, this means that the performance of support vector machines is hardly affected by the following two kinds of errors: large errors in a small fraction of the data set and small errors in the whole data set. This not only means that these errors do not lead to large errors in the support vector machines but also that even the finite sample distribution of support vector machines is hardly affected.

Acknowledgments

We would like to thank two anonymous reviewers for their very helpful comments.

Appendix

In Appendix A.1, we briefly recall some facts about weak convergence of probability measures. In addition, we show that weak convergence of probability measures on a Polish space implies convergence of the corresponding Bochner integrals of bounded, continuous functions. Appendix A.2 contains all proofs.

A.1. Weak convergence of probability measures and bochner integrals

Let $\mathcal{Z}$ be a Polish space with Borel-σ-algebra $\mathfrak{B}(\mathcal{Z})$, let d be a metric on $\mathcal{Z}$ which generates the topology on $\mathcal{Z}$ and let $\mathcal{M}_1(\mathcal{Z})$ be the set of all probability measures on $(\mathcal{Z}, \mathfrak{B}(\mathcal{Z}))$.

A sequence $(P_n)_{n\in\mathbb{N}}$ of probability measures on $\mathcal{Z}$ converges to a probability measure P_0 in the weak topology on $\mathcal{M}_1(\mathcal{Z})$ if

$$\lim_{n\to\infty} \int g \, dP_n = \int g \, dP_0 \quad \forall g \in \mathcal{C}_b(\mathcal{Z})$$

where $\mathcal{C}_b(\mathcal{Z})$ denotes the set of all bounded, continuous functions $g : \mathcal{Z} \to \mathbb{R}$, see [3, Section 1.1].

The weak topology on $\mathcal{M}_1(\mathcal{Z})$ is metrizable by the Prokhorov metric d_{Pro}; see e.g. [20, Section 2.2]. The Prokhorov metric d_{Pro} on $\mathcal{M}_1(\mathcal{Z})$ is defined by

$$d_{\text{Pro}}(\text{P}_1, \text{P}_2) = \inf\{\varepsilon \in (0, \infty) \mid \text{P}_1(B) < \text{P}_2(B^\varepsilon) + \varepsilon \quad \forall B \in \mathfrak{B}(\mathcal{Z})\}$$

where $B^\varepsilon = \{z \in \mathcal{Z} \mid \inf_{z' \in B} d(z, z') < \varepsilon\}$.

Let $g : \mathcal{Z} \to \mathbb{R}$ be a continuous and bounded function. By definition, we have $\lim_{n\to\infty} \int g \, \text{dP}_n = \int g \, \text{dP}_0$ for every sequence $(\text{P}_n)_{n\in\mathbb{N}} \subset \mathcal{M}_1(\mathcal{Z})$ which converges weakly in $\mathcal{M}_1(\mathcal{Z})$ to some P_0. The following theorem states that this is still valid for Bochner integrals if g is replaced by a vector-valued continuous and bounded function $\Psi : \mathcal{Z} \to H$, where H is a separable Banach space. This follows from a corresponding statement in [5, p. III.40] for locally compact spaces $\mathcal{Z}$. Boundedness of Ψ means that $\sup_{z\in\mathcal{Z}} \|\Psi(z)\|_H < \infty$.

Theorem A.1. *Let $\mathcal{Z}$ be a Polish space with Borel-σ-algebra $\mathfrak{B}(\mathcal{Z})$ and let H be a separable Banach space. If $\Psi : \mathcal{Z} \to H$ is a continuous and bounded function, then*

$$\int \Psi \, \text{dP}_n \longrightarrow \int \Psi \, \text{dP}_0 \quad (n \to \infty)$$

for every sequence $(\text{P}_n)_{n\in\mathbb{N}} \subset \mathcal{M}_1(\mathcal{Z})$ which converges weakly in $\mathcal{M}_1(\mathcal{Z})$ to some P_0.

A.2. Proofs

In order to prove the main theorem, i.e. Theorem 3.1, we have to prove Theorem 3.3 and Corollary 3.5 at first.

Proof of Theorem 3.3. Since the proof is somewhat involved, we start with a short outline. The proof is divided into four parts. Part 1 is concerned with some important preparations. We have to show that $(f_{L^*,\text{P}_n,\lambda})_{n\in\mathbb{N}}$ converges to $f_{L^*,\text{P}_0,\lambda}$ in H if the sequence of probability measures $(\text{P}_n)_{n\in\mathbb{N}}$ weakly converges to the probability measure P_0. According to Part 1, the sequence $(f_{L^*,\text{P}_n,\lambda})_{n\in\mathbb{N}}$ is bounded in the Hilbert space H. Therefore, there is a subsequence $(f_{L^*,\text{P}_{n_\ell},\lambda})_{\ell\in\mathbb{N}}$ of $(f_{L^*,\text{P}_n,\lambda})_{n\in\mathbb{N}}$ which weakly converges in H. Then, it is shown in Part 2 and Part 3 that

$$\lim_{\ell\to\infty} \mathcal{R}_{L^*,\text{P}_{n_\ell}}(f_{L^*,\text{P}_{n_\ell},\lambda}) = \mathcal{R}_{L^*,\text{P}_0}(f_{L^*,\text{P}_0,\lambda}) \tag{8}$$

$$\lim_{\ell\to\infty} \mathcal{R}_{L^*,\text{P}_{n_\ell},\lambda}(f_{L^*,\text{P}_{n_\ell},\lambda}) = \mathcal{R}_{L^*,\text{P}_0,\lambda}(f_{L^*,\text{P}_0,\lambda}). \tag{9}$$

Because of

$$\|f\|_H^2 = \frac{1}{\lambda}(\mathcal{R}_{L^*,\text{P},\lambda}(f) - \mathcal{R}_{L^*,\text{P}}(f)) \quad \forall \text{P} \in \mathcal{M}_1(\mathcal{X} \times \mathcal{Y}) \, \forall f \in H,$$

it follows from (8) and (9) that $\lim_{\ell\to\infty} \|f_{L^*,\text{P}_{n_\ell},\lambda}\|_H = \|f_{L^*,\text{P}_0,\lambda}\|_H$. Since this convergence of the norms together with weak convergence in the Hilbert space H implies (strong) convergence in H, we get that the subsequence $(f_{L^*,\text{P}_{n_\ell},\lambda})_{\ell\in\mathbb{N}}$ converges to $f_{L^*,\text{P}_0,\lambda}$ in H. Part 4 extends this result to the whole sequence $(f_{L^*,\text{P}_n,\lambda})_{n\in\mathbb{N}}$. The main difficulty in the proof is the verification of (8) in Part 3.

In order to shorten notation, define

$$L_f^* : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}, \qquad (x, y) \mapsto L^*(x, y, f(x)) = L(x, y, f(x)) - L(x, y, 0)$$

for every measurable $f : \mathcal{X} \to \mathbb{R}$. Following e.g. [33,25], we use the notation

$$\text{P}g = \int g \, \text{dP}$$

for integrals of real-valued functions g with respect to P. This leads to a very efficient notation which is more intuitive here because, in the following, P rather acts as a linear functional on a function space than as a probability measure on a σ-algebra.

By use of these notations, we may write

$$\text{P}L_f^* = \int L_f^* \, \text{dP} = \mathcal{R}_{L^*,\text{P}}(f)$$

for the (shifted) risk of $f \in H$. Accordingly, the (shifted) regularized risk of $f \in H$ is

$$\mathcal{R}_{L^*,\text{P},\lambda}(f) = \mathcal{R}_{L^*,\text{P}}(f) + \lambda\|f\|_H^2 = \text{P}L_f^* + \lambda\|f\|_H^2.$$

In case of $k \equiv 0$, the statement of Theorem 3.3 is trivial. Therefore, we may assume $k \not\equiv 0$ in the following.

Part 1: Since the loss function L, the shifted loss L^* and the regularization parameter $\lambda \in (0, \infty)$ are fixed, we may drop them in the notation and write

$$f_\text{P} := f_{L^*,\text{P},\lambda} = S(\text{P}) \quad \forall \text{P} \in \mathcal{M}_1(\mathcal{X} \times \mathcal{Y}).$$

Recall from Theorem 2.1 that $f_{L^*,\mathrm{P},\lambda}$ is equal to the support vector machine $f_{L,\mathrm{P},\lambda}$ if $f_{L,\mathrm{P},\lambda}$ exists. That is, we have $f_{\mathrm{P}} = f_{L,\mathrm{P},\lambda}$ in the latter case. According to [9, (17),(16)],

$$\|f_{\mathrm{P}}\|_\infty \leq \frac{1}{\lambda}|L|_1 \cdot \|k\|_\infty^2 \tag{10}$$

$$\|f_{\mathrm{P}}\|_H \leq \sqrt{\frac{1}{\lambda}|L|_1 \int |f_{\mathrm{P}}| \,\mathrm{dP}} \overset{(10)}{\leq} \frac{1}{\lambda}|L|_1 \cdot \|k\|_\infty \tag{11}$$

for every $\mathrm{P} \in \mathcal{M}_1(\mathcal{X} \times \mathcal{Y})$. Since the kernel k is continuous and bounded, [31, Lemma 4.28] yields

$$f \in \mathcal{C}_b(\mathcal{X}) \quad \forall f \in H. \tag{12}$$

Therefore, continuity of L implies continuity of

$$L_f^* : \ \mathcal{X} \times \mathcal{Y} \to \mathbb{R}, \qquad (x, y) \mapsto L(x, y, f(x)) - L(x, y, 0)$$

for every $f \in H$. Furthermore, the uniform Lipschitz property of L implies

$$\begin{aligned} \sup_{x,y} |L_f^*(x, y)| &= \sup_{x,y} |L(x, y, f(x)) - L(x, y, 0)| \\ &\leq \sup_{x'} \sup_{x,y} |L(x, y, f(x')) - L(x, y, 0)| \leq \sup_{x'} |L|_1 \cdot |f(x') - 0| \\ &= |L|_1 \cdot \|f\|_\infty \end{aligned}$$

for every $f \in H$. Hence, we obtain

$$L_f^* \in \mathcal{C}_b(\mathcal{X} \times \mathcal{Y}) \quad \forall f \in H. \tag{13}$$

In particular, the above calculation and (10) imply

$$\|L_{f_{\mathrm{P}}}^*\|_\infty \leq \frac{1}{\lambda}|L|_1^2 \cdot \|k\|_\infty^2 \quad \forall \mathrm{P} \in \mathcal{M}_1(\mathcal{X} \times \mathcal{Y}). \tag{14}$$

For the remaining parts of the proof, let $(\mathrm{P}_n)_{n\in\mathbb{N}_0} \subset \mathcal{M}_1(\mathcal{X} \times \mathcal{Y})$ be any fixed sequence such that

$$\mathrm{P}_n \longrightarrow \mathrm{P}_0 \quad (n \to \infty)$$

in the weak topology on $\mathcal{M}_1(\mathcal{X} \times \mathcal{Y})$ – that is,

$$\lim_{n\to\infty} \mathrm{P}_n g = \mathrm{P}_0 g \quad \forall g \in \mathcal{C}_b(\mathcal{X} \times \mathcal{Y}). \tag{15}$$

In particular, (13) and (15) imply

$$\lim_{n\to\infty} \mathrm{P}_n L_f^* = \mathrm{P}_0 L_f^* \quad \forall f \in H. \tag{16}$$

In order to shorten the notation, define

$$f_n := f_{\mathrm{P}_n} = f_{L^*,\mathrm{P}_n,\lambda} = S(\mathrm{P}_n) \quad \forall n \in \mathbb{N} \cup \{0\}.$$

Hence, we have to show that $(f_n)_{n\in\mathbb{N}}$ converges to f_0 in H – that is,

$$\lim_{n\to\infty} \|f_n - f_0\|_H = 0. \tag{17}$$

Part 2: In this part of the proof, it is shown that

$$\limsup_{n\to\infty} \mathrm{P}_n L_{f_n}^* + \lambda\|f_n\|_H^2 \leq \mathrm{P}_0 L_{f_0}^* + \lambda\|f_0\|_H^2. \tag{18}$$

Due to (13), the mapping

$$\mathcal{M}_1(\mathcal{X} \times \mathcal{Y}) \to \mathbb{R}, \qquad \mathrm{P} \mapsto \mathrm{P}L_f^* + \lambda\|f\|_H^2$$

is defined well and continuous for every $f \in H$. As being the (pointwise) infimum over a family of continuous functions, the function

$$\mathcal{M}_1(\mathcal{X} \times \mathcal{Y}) \to \mathbb{R}, \qquad \mathrm{P} \mapsto \inf_{f\in H} (\mathrm{P}L_f^* + \lambda\|f\|_H^2)$$

is upper semicontinuous; see, e.g., [12, Prop. 1.1.36]. Therefore, the definition of f_n implies

$$\begin{aligned}\limsup_{n\to\infty} (\mathrm{P}_n L^*_{f_n} + \lambda\|f_n\|_H^2) &= \limsup_{n\to\infty} \inf_{f\in H} (\mathrm{P}_n L^*_f + \lambda\|f\|_H^2) \\ &\le \inf_{f\in H} (\mathrm{P}_0 L^*_f + \lambda\|f\|_H^2) = \mathrm{P}_0 L^*_{f_0} + \lambda\|f_0\|_H^2.\end{aligned}$$

Part 3: In this part of the proof, the following statement is shown:

Let $(f_{n_\ell})_{\ell\in\mathbb{N}}$ be a subsequence of $(f_n)_{n\in\mathbb{N}}$ and assume that $(f_{n_\ell})_{\ell\in\mathbb{N}}$ converges weakly in H to some $f_0' \in H$. Then, the following three assertions are true:

$$\lim_{\ell\to\infty} \mathrm{P}_{n_\ell} L^*_{f_{n_\ell}} = \mathrm{P}_0 L^*_{f_0'} \tag{19}$$

$$f_0' = f_0 \tag{20}$$

$$\lim_{\ell\to\infty} \|f_{n_\ell} - f_0\|_H = 0. \tag{21}$$

In order to prove this, we will also have to deal with subsequences of the subsequence $(f_{n_\ell})_{\ell\in\mathbb{N}}$. As this would lead to a somewhat cumbersome notation, we define

$$\mathrm{P}'_\ell := \mathrm{P}_{n_\ell} \quad \text{and} \quad f'_\ell := f_{n_\ell} \quad \ell \in \mathbb{N}.$$

Thus, $f'_\ell = f_{L^*,\mathrm{P}_{n_\ell},\lambda}$ for every $\ell \in \mathbb{N}$. Then, the assumption of weak convergence in the Hilbert space H equals

$$\lim_{\ell\to\infty} \langle f'_\ell, h\rangle_H = \langle f'_0, h\rangle_H \quad \forall h \in H. \tag{22}$$

First of all, we show (19) by proving

$$\limsup_{\ell\to\infty} |\mathrm{P}'_\ell L^*_{f'_\ell} - \mathrm{P}_0 L^*_{f'_0}| \le \varepsilon_0 \tag{23}$$

for every fixed $\varepsilon_0 > 0$. In order to do this, fix any $\varepsilon_0 > 0$ and define

$$\varepsilon := \frac{\varepsilon_0}{|L|_1 \cdot (\frac{1}{\lambda}|L|_1 \cdot \|k\|_\infty^2 + \|f'_0\|_\infty)} > 0. \tag{24}$$

Since $\mathcal{X} \times \mathcal{Y}$ is a Polish space, weak convergence of $(\mathrm{P}'_\ell)_{\ell\in\mathbb{N}}$ implies uniform tightness of $(\mathrm{P}'_\ell)_{\ell\in\mathbb{N}}$ (see e.g. [14, Theorem 11.5.3]). That is, there is a compact subset $K_\varepsilon \subset \mathcal{X} \times \mathcal{Y}$ such that its complement $K_\varepsilon^{\mathrm{c}}$ fulfills

$$\limsup_{\ell\to\infty} \mathrm{P}'_\ell(K_\varepsilon^{\mathrm{c}}) < \varepsilon. \tag{25}$$

Since K_ε is compact and the projection $\tau_{\mathcal{X}} : \mathcal{X} \times \mathcal{Y} \to \mathcal{X}, \ (x, y) \mapsto x$ is continuous, $\tilde{K}_\varepsilon := \tau_X(K_\varepsilon)$ is compact in $\mathcal{X}$. For every $\ell \in \mathbb{N}_0$, the restriction of f'_ℓ to $\tilde{K}_\varepsilon$ is denoted by $\tilde{f}'_\ell$. Let $\tilde{k}$ be the restriction of k to $\tilde{K}_\varepsilon$ and let $\tilde{H}$ be its RKHS. According to [2, Section 4.2], $(\tilde{f}'_\ell)_{\ell\in\mathbb{N}}$ is a bounded sequence in $\tilde{H}$ as $(f'_\ell)_{\ell\in\mathbb{N}}$ is bounded in H according to (11). Hence, $(\tilde{f}'_\ell)_{\ell\in\mathbb{N}}$ is relatively compact in $\mathcal{C}_b(\tilde{K}_\varepsilon)$ according to [31, Corollary 4.31].

The following reasoning shows that $(\tilde{f}'_\ell)_{\ell\in\mathbb{N}}$ converges to $\tilde{f}'_0$ in $\mathcal{C}_b(\tilde{K}_\varepsilon)$ — that is,

$$\lim_{\ell\to\infty} \sup_{x\in\tilde{K}_\varepsilon} |f'_\ell(x) - f'_0(x)| = 0. \tag{26}$$

We will show (26) by contradiction. If (26) is not true, then there is a $\delta > 0$ and a subsequence $(\tilde{f}'_{\ell_j})_{j\in\mathbb{N}}$ such that

$$\sup_{x\in\tilde{K}_\varepsilon} |f'_{\ell_j}(x) - f'_0(x)| > \delta \quad \forall j \in \mathbb{N}. \tag{27}$$

Relative compactness of $(\tilde{f}'_\ell)_{\ell\in\mathbb{N}}$ implies that there is a further subsequence $(\tilde{f}'_{\ell_{j_m}})_{m\in\mathbb{N}}$ which converges in $\mathcal{C}_b(\tilde{K}_\varepsilon)$ to some $\tilde{h}_0 \in \mathcal{C}_b(\tilde{K}_\varepsilon)$. Then,

$$\begin{aligned}\tilde{h}_0(x) &= \lim_{m\to\infty} \tilde{f}'_{\ell_{j_m}}(x) = \lim_{m\to\infty} f'_{\ell_{j_m}}(x) = \lim_{m\to\infty} \langle f'_{\ell_{j_m}}, \Phi(x)\rangle_H \\ &\overset{(22)}{=} \langle f'_0, \Phi(x)\rangle_H = f'_0(x) = \tilde{f}'_0(x)\end{aligned}$$

for every $x \in \tilde{K}_\varepsilon$. That is, $\tilde{f}'_0$ is the limit of $(\tilde{f}'_{\ell_{j_m}})_{m\in\mathbb{N}}$ — which is the desired contradiction to (27). Therefore, (26) is true.

Now, we can prove (23): Firstly, the triangle inequality and the Lipschitz continuity of L yield

$$\begin{aligned}
\limsup_{\ell\to\infty} |\mathrm{P}'_\ell L^*_{f'_\ell} - \mathrm{P}_0 L^*_{f'_0}| &\le \limsup_{\ell\to\infty} |\mathrm{P}'_\ell L^*_{f'_\ell} - \mathrm{P}'_\ell L^*_{f'_0}| + |\mathrm{P}'_\ell L^*_{f'_0} - \mathrm{P}_0 L^*_{f'_0}| \\
&\overset{(16)}{=} \limsup_{\ell\to\infty} |\mathrm{P}'_\ell L^*_{f'_\ell} - \mathrm{P}'_\ell L^*_{f'_0}| \\
&= \limsup_{\ell\to\infty} \left| \int L(x,y,f'_\ell(x)) - L(x,y,f'_0(x)) \mathrm{d}\mathrm{P}'_\ell \right| \\
&\le \limsup_{\ell\to\infty} \int |L|_1 \cdot |f'_\ell(x) - f'_0(x)| \mathrm{P}'_\ell(d(x,y)) \\
&= |L|_1 \cdot \limsup_{\ell\to\infty} \left(\int_{K_\varepsilon} |f'_\ell(x) - f'_0(x)| \mathrm{P}'_\ell(d(x,y)) + \int_{K_\varepsilon^c} |f'_\ell(x) - f'_0(x)| \mathrm{P}'_\ell(d(x,y)) \right).
\end{aligned}$$

Secondly, using $\tilde{K}_\varepsilon = \tau_X(K_\varepsilon)$, we obtain

$$\limsup_{\ell\to\infty} \int_{K_\varepsilon} |f'_\ell(x) - f'_0(x)| \mathrm{P}'_\ell(d(x,y)) \le \limsup_{\ell\to\infty} \sup_{(x,y)\in K_\varepsilon} |f'_\ell(x) - f'_0(x)| = \limsup_{\ell\to\infty} \sup_{x\in\tilde{K}_\varepsilon} |f'_\ell(x) - f'_0(x)| \overset{(26)}{=} 0.$$

Thirdly,

$$\begin{aligned}
\limsup_{\ell\to\infty} \int_{K_\varepsilon^c} |f'_\ell(x) - f'_0(x)| \mathrm{P}'_\ell(d(x,y)) &\le \limsup_{\ell\to\infty} \mathrm{P}'_\ell(K_\varepsilon^c) \cdot (\|f'_\ell\|_\infty + \|f'_0\|_\infty) \\
&\overset{(25)}{\le} \limsup_{\ell\to\infty} \varepsilon \cdot (\|f'_\ell\|_\infty + \|f'_0\|_\infty) \overset{(10),(24)}{=} \frac{\varepsilon_0}{|L|_1}.
\end{aligned}$$

Combining these three calculations proves (23). Since $\varepsilon_0 > 0$ was arbitrarily chosen in (23), this proves (19).

Next, we prove (20): Due to weak convergence of $(f_{n_\ell})_{\ell\in\mathbb{N}}$ in H, it follows from [10, Exercise V.1.9] that

$$\|f'_0\|_H \le \liminf_{\ell\to\infty} \|f_{n_\ell}\|_H. \tag{28}$$

Therefore, the definition of $f_0 = f_{L^*,\mathrm{P}_0,\lambda}$ implies

$$\begin{aligned}
\mathrm{P}_0 L^*_{f_0} + \lambda\|f_0\|_H^2 &= \inf_{f\in H} (\mathrm{P}_0 L^*_f + \lambda\|f\|_H^2) \\
&\le \mathrm{P}_0 L^*_{f'_0} + \lambda\|f'_0\|_H^2 \overset{(19),(28)}{\le} \liminf_{\ell\to\infty} (\mathrm{P}_{n_\ell} L^*_{f_{n_\ell}} + \lambda\|f_{n_\ell}\|_H^2) \\
&\le \limsup_{\ell\to\infty} (\mathrm{P}_{n_\ell} L^*_{f_{n_\ell}} + \lambda\|f_{n_\ell}\|_H^2) \overset{(18)}{\le} \mathrm{P}_0 L^*_{f_0} + \lambda\|f_0\|_H^2.
\end{aligned}$$

Due to this calculation, it follows that

$$\mathrm{P}_0 L^*_{f_0} + \lambda\|f_0\|_H^2 = \inf_{f\in H} (\mathrm{P}_0 L^*_f + \lambda\|f\|_H^2) = \mathrm{P}_0 L^*_{f'_0} + \lambda\|f'_0\|_H^2 \tag{29}$$

and

$$\mathrm{P}_0 L^*_{f_0} + \lambda\|f_0\|_H^2 = \lim_{\ell\to\infty} (\mathrm{P}_{n_\ell} L^*_{f_{n_\ell}} + \lambda\|f_{n_\ell}\|_H^2). \tag{30}$$

According to Theorem 2.1, $f_0 = f_{L^*,\mathrm{P}_0,\lambda}$ is the unique minimizer of the function

$$H \to \mathbb{R}, \quad f \mapsto \mathrm{P}_0 L^*_f + \lambda\|f\|_H^2$$

and, therefore, (29) implies $f_0 = f'_0$ – i.e. (20).

Completing Part 3 of the proof, (21) is shown now:

$$\begin{aligned}
\lim_{\ell\to\infty} \|f_{n_\ell}\|_H^2 &= \lim_{\ell\to\infty} \frac{1}{\lambda} ((\mathrm{P}_{n_\ell} L^*_{f_{n_\ell}} + \lambda\|f_{n_\ell}\|_H^2) - \mathrm{P}_{n_\ell} L^*_{f_{n_\ell}}) \\
&\overset{(19),(30)}{=} \frac{1}{\lambda} ((\mathrm{P}_0 L^*_{f_0} + \lambda\|f_0\|_H^2) - \mathrm{P}_0 L^*_{f_0}) = \|f_0\|_H^2.
\end{aligned}$$

By assumption, the sequence $(f_{n_\ell})_{\ell\in\mathbb{N}}$ converges weakly to some $f'_0 \in H$ and by (20), we know that $f'_0 = f_0$. In addition, we have proven $\lim_{\ell\to\infty} \|f_{n_\ell}\|_H = \|f_0\|_H$ now. This convergence of the norms together with weak convergence implies strong convergence in the Hilbert space H, – see, e.g., [10, Exercise V.1.8]. That is, we have proven (21).

Part 4: In this final part of the proof, (17) is shown. This is done by contradiction: If (17) is not true, there is an $\varepsilon > 0$ and a subsequence $(f_{n_\ell})_{\ell\in\mathbb{N}}$ of $(f_n)_{n\in\mathbb{N}}$ such that

$$\|f_{n_\ell} - f_0\|_H > \varepsilon \quad \forall \ell \in \mathbb{N}. \tag{31}$$

According to (11), $(f_{n_\ell})_{\ell\in\mathbb{N}} = (f_{\mathrm{P}_{n_\ell}})_{\ell\in\mathbb{N}}$ is bounded in H. Hence, the sequence $(f_{n_\ell})_{\ell\in\mathbb{N}}$ contains a further subsequence that weakly converges in H to some f_0'; see e.g. [15, Corollary IV.4.7]. Without loss of generality, we may therefore assume that $(f_{n_\ell})_{\ell\in\mathbb{N}}$ weakly converges in H to some f_0'. (Otherwise, we can choose another subsequence in (31)). Next, it follows from Part 3, that $(f_{n_\ell})_{\ell\in\mathbb{N}}$ strongly converges in H to f_0 – which is a contradiction to (31). □

Proof of Corollary 3.5. Let $(D_{n,m})_{m\in\mathbb{N}}$ be a sequence in $(\mathcal{X}\times\mathcal{Y})^n$ which converges to some $D_{n,0} \in (\mathcal{X}\times\mathcal{Y})^n$. Then, the corresponding sequence of empirical measures $(\mathbb{P}_{D_{n,m}})_{m\in\mathbb{N}}$ weakly converges in $\mathcal{M}_1(\mathcal{X}\times\mathcal{Y})$ to $\mathbb{P}_{D_{n,0}}$. Therefore, the statement follows from Theorem 3.3 and (4). □

Based on [11], the main theorem essentially is a consequence of Theorem 3.3.

Proof of Theorem 3.1. First, assume that $\lambda_n = \lambda$ does not depend on the sample size $n \in \mathbb{N}$. According to Corollary 3.5, the SVM-estimator

$$S_n : (\mathcal{X}\times\mathcal{Y})^n \to H, \qquad D_n \mapsto f_{L,D_n,\lambda}$$

is continuous and, therefore, measurable with respect to the Borel-σ-algebras for every $n \in \mathbb{N}$. The mapping

$$S : \mathcal{M}_1(\mathcal{X}\times\mathcal{Y}) \to H, \qquad \mathrm{P} \mapsto f_{L^*,\mathrm{P},\lambda}$$

is a continuous functional due to Theorem 3.3. Furthermore,

$$S_n(D_n) = S(\mathbb{P}_{D_n}) \quad \forall D_n \in (\mathcal{X}\times\mathcal{Y})^n \ \forall n \in \mathbb{N}.$$

As already mentioned in Section 2, H is a separable Hilbert space and, therefore, a Polish space. Hence, the sequence of SVM-estimators $(S_n)_{n\in\mathbb{N}}$ is qualitatively robust according to [11, Theorem 2].

Next, let $(\lambda_n)_{n\in\mathbb{N}} \subset (0,\infty)$ be any sequence. Fix any $n \in \mathbb{N}$. Then, the previous result implies for the fixed $n \in \mathbb{N}$: for every $\mathrm{P} \in \mathcal{M}_1(\mathcal{X}\times\mathcal{Y})$ and every $\rho > 0$, there is an $\varepsilon_n > 0$ such that

$$d_{\mathrm{Pro}}(\mathrm{Q}, \mathrm{P}) < \varepsilon_n \Rightarrow d_{\mathrm{Pro}}(S_n(\mathrm{Q}^n), S_n(\mathrm{P}^n)) < \rho.$$

That is, $(S_n)_{n\in\mathbb{N}}$ is finite sample qualitatively robust. □

Proof of Proposition 3.2. Without loss of generality, we may assume that

$$\tilde{f}(x_0) = 0 \quad \text{and} \quad \tilde{f}(x_1) = 1. \tag{32}$$

(Otherwise, we can divide $\tilde{f}$ by $\tilde{f}(x_1)$.) Since the function $\mathbb{R} \to [0,\infty), \quad t \mapsto L(x_1, 1, t)$ is convex, it is also continuous. Therefore, (6) implies the existence of an $\gamma \in (0, 1)$ such that

$$L(x_1, 1, \gamma) > 0. \tag{33}$$

Note that convexity of the loss function, $L(x_1, 1, 1) = 0$ and $L(x_1, 1, \gamma) > 0$ imply

$$0 = L(x_1, 1, 1) \le L(x_1, 1, t) < L(x_1, 1, \gamma) \le L(x_1, 1, s) \tag{34}$$

for $0 \le s \le \gamma < t \le 1$. Define $\mathrm{P}_0 := \delta_{(x_0,0)}$. Since $f_{L,\delta_{(x_0,0)},\lambda_n} = 0$, it follows that

$$\mathrm{P}_0^n(\{D_n \in (\mathcal{X}\times\mathcal{Y})^n \mid f_{L,D_n,\lambda_n} = 0\}) = 1. \tag{35}$$

Next, fix any $\varepsilon \in (0, 1)$ and define the mixture distribution

$$\mathrm{P}_\varepsilon := (1-\varepsilon)\mathrm{P}_0 + \varepsilon\delta_{(x_1,1)} = (1-\varepsilon)\delta_{(x_0,0)} + \varepsilon\delta_{(x_1,1)}.$$

For every $n \in \mathbb{N}$, let $\mathcal{Z}_n'$ be the subset of $(\mathcal{X}\times\mathcal{Y})^n$ which consists of all those elements $D_n = (D_n^{(1)}, \ldots, D_n^{(n)}) \in (\mathcal{X}\times\mathcal{Y})^n$ where

$$D_n^{(i)} \in \{(x_0, 0), (x_1, 1)\} \quad \forall i \in \{1, \ldots, n\}.$$

In addition, let $\mathcal{Z}_n''$ be the subset of $(\mathcal{X}\times\mathcal{Y})^n$ which consists of all those elements $D_n = (D_n^{(1)}, \ldots, D_n^{(n)}) \in (\mathcal{X}\times\mathcal{Y})^n$ where

$$\sharp(\{i \in \{1, \ldots, n\} \mid D_n^{(i)} = (x_1, 1)\}) \ge \frac{\varepsilon}{2}. \tag{36}$$

Define $\mathcal{Z}_n := \mathcal{Z}_n' \cap \mathcal{Z}_n''$. Then, we have $\mathrm{P}_\varepsilon^n(\mathcal{Z}_n') = 1$ and, according to the law of large numbers [14, Theorem 8.3.5], $\lim_{n\to\infty} \mathrm{P}_\varepsilon^n(\mathcal{Z}_n'') = 1$. Hence, there is an $n_{\varepsilon,1} \in \mathbb{N}$ such that

$$\mathrm{P}_\varepsilon^n(\mathcal{Z}_n) \geq \frac{1}{2} \quad \forall n \geq n_{\varepsilon,1}. \tag{37}$$

Due to $\lim_{n\to\infty} \lambda_n = 0$ and (33), there is an $n_{\varepsilon,2} \in \mathbb{N}$ such that

$$\lambda_n \|\tilde{f}\|_H^2 < \frac{\varepsilon}{2} L(x_1, 1, \gamma) \quad \forall n \geq n_{\varepsilon,2}. \tag{38}$$

In the following, we show

$$f_{L,D_n,\lambda_n}(x_1) > \gamma \quad \forall D_n \in \mathcal{Z}_n,\ \forall n \geq n_{\varepsilon,2}. \tag{39}$$

To this end, fix any $D_n \in \mathcal{Z}_n$. In order to prove (39), it is enough to show the following assertion for every $n \geq n_{\varepsilon,2}$:

$$f \in H, \quad f(x_1) \leq \gamma \Rightarrow \mathcal{R}_{L,D_n,\lambda_n}(\tilde{f}) \leq \mathcal{R}_{L,D_n,\lambda_n}(f). \tag{40}$$

The definition of $\mathcal{Z}_n$ and (32) imply

$$\mathcal{R}_{L,D_n,\lambda_n}(\tilde{f}) = \mathcal{R}_{L,D_n}(\tilde{f}) + \lambda_n \|\tilde{f}\|_H^2 = \lambda_n \|\tilde{f}\|_H^2.$$

For every $f \in H$ such that $f(x_1) \leq \gamma$, the definition of $\mathcal{Z}_n$ implies

$$\mathcal{R}_{L,D_n,\lambda_n}(f) \geq \mathcal{R}_{L,D_n}(f) \overset{(36)}{\geq} \frac{\varepsilon}{2} L(x_1, 1, f(x_1)) \overset{(34)}{\geq} \frac{\varepsilon}{2} L(x_1, 1, \gamma).$$

Hence, (40) follows from (38) and, therefore, we have proven (39).

Define $n_\varepsilon = \max\{n_{\varepsilon,1}, n_{\varepsilon,2}\}$. By assumption, k is a bounded, non-zero kernel. According to [31, Lemma 4.23], this implies

$$\|f_{L,D_n,\lambda_n}\|_H \geq \frac{\|f_{L,D_n,\lambda_n}\|_\infty}{\|k\|_\infty} \overset{(39)}{\geq} \frac{\gamma}{\|k\|_\infty} \quad \forall D_n \in \mathcal{Z}_n,\ \forall n \geq n_\varepsilon$$

and, therefore,

$$\|f_{L,D_n,\lambda_n}\|_H \geq \min\left\{\frac{\gamma}{\|k\|_\infty},\ 1\right\} =: c \quad \forall D_n \in \mathcal{Z}_n,\ \forall n \geq n_\varepsilon. \tag{41}$$

Define $F := \{f \in H \mid \|f\|_H \geq c\}$ and

$$F^{\frac{c}{2}} := \left\{f \in H \mid \inf_{f' \in F} \|f - f'\|_H \leq \frac{c}{2}\right\} \subset \{f \in H \mid \|f\|_H > 0\}. \tag{42}$$

Hence, for every $n \geq n_\varepsilon$, we obtain

$$\begin{aligned}
\left[S_n(\mathrm{P}_\varepsilon^n)\right](F) &= \mathrm{P}_\varepsilon^n(\{D_n \mid \|f_{L,D_n,\lambda_n}\|_H \geq c\}) \overset{(41)}{\geq} \mathrm{P}_\varepsilon^n(\mathcal{Z}_n) \\
&\overset{(37)}{\geq} \frac{1}{2} \overset{(41)}{\geq} \frac{c}{2} \overset{(35)}{=} \mathrm{P}_0^n(\{D_n \mid \|f_{L,D_n,\lambda_n}\|_H > 0\}) + \frac{c}{2} \\
&= [S_n(\mathrm{P}_0^n)](\{f \in H \mid \|f\|_H > 0\}) + \frac{c}{2} \\
&\overset{(42)}{\geq} [S_n(\mathrm{P}_0^n)]\left(F^{\frac{c}{2}}\right) + \frac{c}{2}.
\end{aligned}$$

According to the definition of the Prokhorov distance (see Appendix A.1), it follows that

$$\sup_{n\in\mathbb{N}} d_{\mathrm{Pro}}(S_n(\mathrm{P}_0^n), S_n(\mathrm{P}_\varepsilon^n)) \geq \frac{c}{2}. \tag{43}$$

In addition, we have $d_{\mathrm{Pro}}(\mathrm{P}_0, \mathrm{P}_\varepsilon) \leq \varepsilon$ because P_ε is an ε-mixture of P_0. Since $c > 0$ does not depend on $\varepsilon \in (0, 1)$ and ε may be arbitrarily small, this proves that $(S_n)_{n\in\mathbb{N}}$ is not qualitatively robust in P_0. □

Proof of Corollary 3.7. Let $\mathbb{P}_{\mathbb{D}_n}$ denote the function which maps $\omega \in \Omega$ to the empirical measure $\frac{1}{n}\sum_{i=1}^n \delta_{(X_i(\omega),Y_i(\omega))}$. According to Varadarajan's Theorem [14, Theorem 11.4.1], there is a set $N \in \mathcal{A}$ such that $\mathrm{Q}(N) = 0$ and $\mathbb{P}_{\mathbb{D}_n(\omega)}$ weakly converges to P for every $\omega \in \Omega \setminus N$. Then, Theorem 3.3 implies

$$\lim_{n\to\infty} \|f_{L^*,\mathbb{D}_n(\omega),\lambda} - f_{L^*,\mathrm{P},\lambda}\|_H \overset{(4)}{=} \lim_{n\to\infty} \|S(\mathbb{P}_{\mathbb{D}_n(\omega)}) - S(\mathrm{P})\|_H = 0$$

for every $\omega \in \Omega \setminus N$. This proves (a) and, due to [31, Lemma 4.28], (b). The Lipschitz continuity of L^* implies

$$\begin{aligned}
|\mathcal{R}_{L^*,\mathrm{P}}(f_{L^*,\mathbb{D}_n(\omega),\lambda}) - \mathcal{R}_{L^*,\mathrm{P}}(f_{L^*,\mathrm{P},\lambda})| &= \left| \int L(x,y,f_{L^*,\mathbb{D}_n(\omega),\lambda}(x)) - L(x,y,f_{L^*,\mathrm{P},\lambda}(x)) \,\mathrm{P}(d(x,y)) \right| \\
&\leq \int \sup_{x',y'} |L(x',y',f_{L^*,\mathbb{D}_n(\omega),\lambda}(x)) - L(x',y',f_{L^*,\mathrm{P},\lambda}(x))| \,\mathrm{P}(d(x,y)) \\
&\leq \int |L|_1 \cdot |f_{L^*,\mathbb{D}_n(\omega),\lambda}(x) - f_{L^*,\mathrm{P},\lambda}(x)| \,\mathrm{P}(d(x,y)) \\
&\leq |L|_1 \cdot \|f_{L^*,\mathbb{D}_n(\omega),\lambda} - f_{L^*,\mathrm{P},\lambda}\|_\infty
\end{aligned}$$

for every $\omega \in \Omega$. According to (b), the last term converges to 0 for Q-almost every $\omega \in \Omega$ and this implies (d). Finally, (c) follows from (a) and (d).

If $f_{L,\mathrm{P},\lambda}$ exists, then $f_{L^*,\mathrm{P},\lambda}$ is equal to $f_{L,\mathrm{P},\lambda}$ (Theorem 2.1). In particular, there is an $f \in H$ such that $(x,y) \mapsto L(x,y,f(x))$ is P- integrable. Since Lipschitz-continuity of L and $H \subset \mathcal{C}_b(\mathcal{X})$ (see [31, Lemma 4.28]) implies P-integrability of $(x,y) \mapsto L^*(x,y,f(x)) = L(x,y,f(x)) - L(x,y,0)$, we get that $(x,y) \mapsto L(x,y,0)$ is also P-integrable. Therefore, $\mathcal{R}_{L^*,\mathrm{P}}(f)$ is equal to $\mathcal{R}_{L,\mathrm{P}}(f) - \mathcal{R}_{L,\mathrm{P}}(0)$ for every $f \in H$, and $\mathcal{R}_{L,\mathrm{P}}(0)$ is a finite constant which does not depend on f. Furthermore, $f_{L^*,D_n,\lambda} = f_{L,D_n,\lambda}$ for every $D_n \in (\mathcal{X} \times \mathcal{Y})^n$; see Section 2. Hence, the original assertions (a)–(d) for L^* turn into the corresponding assertions for L instead of L^*. □

Proof of Theorem A.1. If $\Psi = 0$, the statement is true. Assume $\Psi \neq 0$ now and assume that the statement of the theorem is not true. Then, there is an $\varepsilon > 0$ and a subsequence $(\mathrm{P}_{n_\ell})_{\ell \in \mathbb{N}}$ such that

$$\left\| \int \Psi \,\mathrm{dP}_{n_l} - \int \Psi \,\mathrm{dP}_0 \right\|_H > \varepsilon \quad \forall\, \ell \in \mathbb{N}. \tag{44}$$

Since the sequence $(\mathrm{P}_n)_{n\in\mathbb{N}}$ weakly converges to P_0, it is uniformly tight; see, e.g., [14, Theorem 11.5.3]. That is, there is a compact subset $K \subset \mathcal{Z}$ such that

$$\mathrm{P}_{n_\ell}(\mathcal{Z} \setminus K) < \frac{\varepsilon}{4 \sup_z \|\Psi(z)\|_H} \quad \forall\, \ell \in \mathbb{N}_0. \tag{45}$$

For every $\ell \in \mathbb{N}$, let $\tilde{\mathrm{P}}_{n_\ell}$ denote the restriction of P_{n_ℓ} to the Borel-σ-algebra $\mathfrak{B}(K)$ of K. Let $\tilde{\Psi}$ denote the restriction of Ψ to K. Since K is a compact Polish space, the set $\mathcal{M}(K)$ of all finite signed measures on $\mathfrak{B}(K)$ is the dual space of $\mathcal{C}(K)$ (the set of all continuous functions $f : K \to \mathbb{R}$); see e.g. [14, Theorem 7.1.1 and 7.4.1]. Accordingly, $\mathcal{M}(K)$ is precisely the set of all (real) measures in the sense of [5, Section III.1]; see also [5, Subsection III.1.5 and III.1.8]. Since $(\tilde{\mathrm{P}}_{n_\ell})_{\ell\in\mathbb{N}}$ is relatively compact in the vague topology of $\mathcal{M}(K)$ [5, Subsection III.1.9], we may assume without loss of generality that $(\tilde{\mathrm{P}}_{n_\ell})_{\ell\in\mathbb{N}}$ vaguely converges to some positive finite measure $\tilde{\mathrm{P}}'_0$. (Otherwise, we may replace $(\tilde{\mathrm{P}}_{n_\ell})_{\ell\in\mathbb{N}}$ by a further subsequence.) According to [5, p. III.40], vague convergence implies

$$\int \tilde{\Psi} \,\mathrm{d}\tilde{\mathrm{P}}_{n_\ell} \longrightarrow \int \tilde{\Psi} \,\mathrm{d}\tilde{\mathrm{P}}'_0 \quad (\ell \to \infty) \tag{46}$$

for Pettis and Bochner integrals (since H is assumed to be a separable Banach space, Pettis integrals and Bochner integrals coincide; see e.g. [14, p. 150]).

Let H^* be the dual space of H. Note that $F \circ \Psi$ is continuous and bounded on $\mathcal{Z}$ for every $F \in H^*$. Hence, it follows from weak convergence of $(\mathrm{P}_{n_\ell})_{\ell\in\mathbb{N}}$ to P_0 and a property of the Bochner integral [12, Theorem 3.10.16] that

$$\lim_{\ell\to\infty} F\left(\int \Psi \,\mathrm{dP}_{n_\ell}\right) = \lim_{\ell\to\infty} \int F \circ \Psi \,\mathrm{dP}_{n_\ell} = \int F \circ \Psi \,\mathrm{dP}_0 = F\left(\int \Psi \,\mathrm{dP}_0\right).$$

Accordingly, vague convergence of $(\tilde{\mathrm{P}}_{n_\ell})_{\ell\in\mathbb{N}}$ to $\tilde{\mathrm{P}}'_0$ implies $\lim_{\ell\to\infty} F\left(\int \tilde{\Psi} \,\mathrm{d}\tilde{\mathrm{P}}_{n_\ell}\right) = F\left(\int \tilde{\Psi} \,\mathrm{d}\tilde{\mathrm{P}}'_0\right)$. Hence,

$$\lim_{\ell\to\infty} F\left(\int \Psi \,\mathrm{dP}_{n_\ell} - \int \tilde{\Psi} \,\mathrm{d}\tilde{\mathrm{P}}_{n_\ell}\right) = F\left(\int \Psi \,\mathrm{dP}_0 - \int \tilde{\Psi} \,\mathrm{d}\tilde{\mathrm{P}}'_0\right). \tag{47}$$

For every $\ell \in \mathbb{N}$,

$$\left\| \int \Psi \,\mathrm{dP}_{n_\ell} - \int \tilde{\Psi} \,\mathrm{d}\tilde{\mathrm{P}}_{n_\ell} \right\|_H = \left\| \int_{\mathcal{Z}\setminus K} \Psi \,\mathrm{dP}_{n_\ell} \right\|_H \leq \int_{\mathcal{Z}\setminus K} \|\Psi\|_H \,\mathrm{dP}_{n_\ell} \overset{(45)}{\leq} \frac{\varepsilon}{4}. \tag{48}$$

For every $\ell \in \mathbb{N}$ and every $F \in H^*$ such that $\|F\|_{H^*} \leq 1$, (48) implies $\left|F\left(\int \Psi \, dP_{n_\ell} - \int \tilde{\Psi} \, d\tilde{P}_{n_\ell}\right)\right| \leq \frac{\varepsilon}{4}$ and, because of (47), also $\left|F\left(\int \Psi \, dP_{n_\ell} - \int \tilde{\Psi} \, d\tilde{P}_{n_\ell}\right)\right| \leq \frac{\varepsilon}{4}$. Hence, it follows from [15, Corollary II.3.15] that

$$\left\| \int \Psi \, dP_0 - \int \tilde{\Psi} \, d\tilde{P}'_0 \right\|_H \leq \frac{\varepsilon}{4}. \tag{49}$$

By using the triangle inequality, we obtain

$$\begin{aligned} \left\| \int \Psi \, dP_{n_\ell} - \int \Psi \, dP_0 \right\|_H \leq & \left\| \int \Psi \, dP_{n_\ell} - \int \tilde{\Psi} \, d\tilde{P}_{n_\ell} \right\|_H \\ & + \left\| \int \tilde{\Psi} \, d\tilde{P}_{n_\ell} - \int \tilde{\Psi} \, d\tilde{P}'_0 \right\|_H + \left\| \int \tilde{\Psi} \, d\tilde{P}'_0 - \int \Psi \, dP_0 \right\|_H, \end{aligned}$$

so that (46), (48) and (49) imply $\limsup_{\ell \to \infty} \left\| \int \Psi \, dP_{n_\ell} - \int \Psi \, dP_0 \right\|_H \leq \frac{\varepsilon}{2}$. This is a contradiction to (44). □

References

[1] P.L. Bartlett, M.I. Jordan, J.D. McAuliffe, Convexity, classification, and risk bounds, J. Amer. Statist. Assoc. 101 (2006) 138–156.
[2] A. Berlinet, C. Thomas-Agnan, Reproducing Kernel Hilbert Spaces in Probability and Statistics, Kluwer Academic Publishers, Boston, 2004.
[3] P. Billingsley, Convergence of Probability Measures, John Wiley & Sons, New York, 1968.
[4] G. Boente, R. Fraiman, V.J. Yohai, Qualitative robustness for stochastic processes, Ann. Statist. 15 (3) (1987) 1293–1312.
[5] N. Bourbaki, Integration. I, Springer-Verlag, Berlin, 2004, Translated from the 1959, 1965 and 1967 French originals by Sterling K. Berberian (Chapters 1–6).
[6] O. Bousquet, A. Elisseeff, Stability and generalization, J. Mach. Learn. Res. 2 (2002) 499–526.
[7] A. Christmann, I. Steinwart, Consistency and robustness of kernel-based regression in convex risk minimization, Bernoulli 13 (3) (2007) 799–819.
[8] A. Christmann, A. Van Messem, Bouligand derivatives and robustness of support vector machines for regression, J. Mach. Learn. Res. 9 (2008) 915–936.
[9] A. Christmann, A. Van Messem, I. Steinwart, On consistency and robustness properties of support vector machines for heavy-tailed distributions, Stat. Interface 2 (2009) 311–327.
[10] J.B. Conway, A Course in Functional Analysis, Springer-Verlag, New York, 1985.
[11] A. Cuevas, Qualitative robustness in abstract inference, J. Statist. Plann. Inference 18 (1988) 277–289.
[12] Z. Denkowski, S. Migórski, N. Papageorgiou, An Introduction to Nonlinear Analysis: Theory, Kluwer Academic Publishers, Boston, 2003.
[13] E. De Vito, L. Rosasco, A. Caponnetto, M. Piana, A. Verri, Some properties of regularized kernel methods, J. Mach. Learn. Res. 5 (2004) 1363–1390.
[14] R. Dudley, Real Analysis and Probability, Wadsworth & Brooks/Cole Advanced Books & Software, Pacific Grove, CA, 1989.
[15] N. Dunford, J. Schwartz, Linear Operators. I. General Theory, Wiley-Interscience Publishers, New York, 1958.
[16] J. Hadamard, Sur les problèmes aux dérivées partielles et leur signification physique, Princeton University Bulletin 13 (1902) 49–52.
[17] F.R. Hampel, Contributions to the Theory of Robust Estimation. Ph.D. Thesis, University of California, Berkeley, 1968.
[18] F.R. Hampel, A general qualitative definition of robustness, Ann. Math. Statist. 42 (1971) 1887–1896.
[19] P.J. Huber, Robust estimation of a location parameter, Ann. Math. Statist. 35 (1964) 73–101.
[20] P.J. Huber, Robust Statistics, John Wiley & Sons, New York, 1981.
[21] P.J. Huber, The behavior of maximum likelihood estimates under nonstandard conditions, in: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Vol. I: Statistics, University California Press, Berkeley, 1967, pp. 221–233.
[22] P.J. Huber, E.M. Ronchetti, Robust Statistics, second ed., John Wiley & Sons Inc., New York, 2009.
[23] J. Jurečková, J. Picek, Robust Statistical Methods with *R*, Chapman & Hall/CRC, Boca Raton, FL, 2006.
[24] R. Maronna, R. Martin, V. Yohai, Robust Statistics, John Wiley & Sons Ltd., Chichester, 2006, Theory and methods.
[25] D. Pollard, A User's Guide to Measure Theoretic Probability, Cambridge University Press, Cambridge, 2002.
[26] B. Schölkopf, A.J. Smola, Learning with Kernels, MIT Press, Cambridge, 2002.
[27] I. Steinwart, Support vector machines are universally consistent, J. Complexity 18 (2002) 768–791.
[28] I. Steinwart, Consistency of support vector machines and other regularized kernel classifiers, IEEE Trans. Inform. Theory 51 (2005) 128–142.
[29] I. Steinwart, Sparseness of support vector machines, J. Mach. Learn. Res. 4 (2003) 1071–1105.
[30] I. Steinwart, M. Anghel, Consistency of support vector machines for forecasting the evolution of an unknown ergodic dynamical system from observations with unknown noise, Ann. Statist. 37 (2009) 841–875.
[31] I. Steinwart, A. Christmann, Support Vector Machines, Springer, New York, 2008.
[32] J. Tukey, A survey of sampling from contaminated distributions, in: Contributions to Probability and Statistics, Stanford Univ. Press, Stanford, Calif, 1960, pp. 448–485.
[33] A. van der Vaart, Asymptotic Statistics, Cambridge University Press, Cambridge, 1998.
[34] V.N. Vapnik, Statistical Learning Theory, John Wiley & Sons, New York, 1998.
[35] T. Zhang, Statistical behavior and consistency of classification methods based on convex risk minimization, Ann. Statist. 32 (2004) 56–85.
[36] T. Zhang, Convergence of Large Margin Separable Linear Classification, in: T.K. Leen, T.G. Dietterich, V. Tresp (Eds.), Advances in Neural Information Processing Systems, vol. 13, MIT Press, Cambridge, MA, 2001, pp. 357–363.