

On the Consistency of the Bootstrap Approach for Support Vector Machines and Related Kernel Based Methods

Andreas Christmann and Robert Hable

Abstract It is shown that bootstrap approximations of support vector machines (SVMs) based on a general convex and smooth loss function and on a general kernel are consistent. This result is useful to approximate the unknown finite sample distribution of SVMs by the bootstrap approach.

1 Introduction

Support vector machines and related kernel based methods can be considered as a hot topic in machine learning because they have good statistical and numerical properties under weak assumptions and have demonstrated their often good generalization properties in many applications, see e.g. [14, 15], [10], and [12]. To our best knowledge, the original SVM approach by [1] was derived from the generalized portrait algorithm invented earlier by [16]. Throughout the paper, the term SVM will be used in the broad sense, i.e. for a general convex loss function and a general kernel.

SVMs based on many standard kernels as for example the Gaussian RBF kernel are nonparametric methods. The finite sample distribution of many nonparametric methods is unfortunately unknown because the distribution P from which the data were generated is usually completely unknown and because there are often only asymptotical results describing the consistency or the rate of convergence of such methods known so far. Furthermore, there is in general *no* uniform rate of convergence for such nonparametric methods due to the famous no-free-lunch theo-

Andreas Christmann
University of Bayreuth, Department of Mathematics, Germany, e-mail:
andreas.christmann@uni-bayreuth.de

Robert Hable
University of Bayreuth, Department of Mathematics, Germany, e-mail:
robert.hable@uni-bayreuth.de

rem, see [5] and [6]. Informally speaking, the no-free-lunch theorem states that, for sufficiently malign distributions, the average risk of any statistical (classification) method may tend arbitrarily slowly to zero. These facts are true for SVMs. SVMs are known to be universally consistent and fast rates of convergence are known for broad *subsets* of all probability distributions. The asymptotic normality of SVMs was shown recently by [8] under certain conditions.

Here, we apply a different approach to SVMs, namely Efron’s bootstrap. The goal of this paper is to show that bootstrap approximations of SVMs which are based on a general convex and smooth loss function and a general smooth kernel are consistent under mild assumptions; more precisely, convergence in outer probability is shown. This result is useful to draw statistical decisions based on SVMs, e.g. confidence intervals, tolerance intervals and so on.

We mention that both the sequence of SVMs and the sequence of their corresponding risks are qualitatively robust under mild assumptions, see [2]. Hence, Efron’s bootstrap approach turns out to be quite successful for SVMs from several aspects.

The rest of the paper has the following structure. Section 2 gives a brief introduction into SVMs. Section 3 gives the result. The last section contains the proof and related results.

2 Support Vector Machines

Current statistical applications are characterized by a wealth of large and high-dimensional data sets. In classification and in regression problems there is a variable of main interest, often called “output values” or “response”, and a number of potential explanatory variables, which are often called “input values”. These input values are used to model the observed output values or to predict future output values. The observations consist of n pairs $(x_1, y_1), \dots, (x_n, y_n)$, which will be assumed to be independent realizations of a random pair (X, Y) . We are interested in minimizing the risk or to obtain a function $f : \mathcal{X} \rightarrow \mathcal{Y}$ such that $f(x)$ is a good predictor for the response y , if $X = x$ is observed. The prediction should be made in an automatic way. We refer to this process of determining a prediction method as “statistical machine learning”, see e.g. [14, 15, 10, 3, 11]. Here, by “good predictor” we mean that f minimizes the expected loss, i.e. the risk,

$$\mathcal{R}_{L,P}(f) = \mathbb{E}_P[L(X, Y, f(X))],$$

where P denotes the unknown joint distribution of the random pair (X, Y) and $L : \mathcal{X} \times \mathcal{Y} \times \mathbb{R} \rightarrow [0, +\infty)$ is a fixed loss function. As a simple example, the least squares loss $L(X, Y, f(X)) = (Y - f(X))^2$ yields the optimal predictor $f(x) = \mathbb{E}_P(Y|X = x)$, $x \in \mathcal{X}$. Because P is unknown, we can neither compute nor minimize the risk $\mathcal{R}_{L,P}(f)$ directly.

Support vector machines, see [16], [1], [14, 15], provide a highly versatile framework to perform statistical machine learning in a wide variety of setups. The minimization of regularized empirical risks over reproducing kernel Hilbert spaces was already considered e.g. by [9]. Given a kernel $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ we consider predictors $f \in H$, where H denotes the corresponding reproducing kernel Hilbert space of functions from \mathcal{X} to \mathbb{R} . The space H includes, for example, all functions of the form $f(x) = \sum_{j=1}^m \alpha_j k(x, x_j)$ where x_j are arbitrary elements in \mathcal{X} and $\alpha_j \in \mathbb{R}$, $1 \leq j \leq m$. To avoid overfitting, a support vector machine $f_{L,P,\lambda}$ is defined as the solution of a regularized risk minimization problem. More precisely,

$$f_{L,P,\lambda} = \arg \inf_{f \in H} \mathbb{E}_P L(X, Y, f(X)) + \lambda \|f\|_H^2, \quad (1)$$

where $\lambda \in (0, \infty)$ is the regularization parameter. For a sample $D = ((x_1, y_1), \dots, (x_n, y_n))$ the corresponding estimated function is given by

$$f_{L,D_n,\lambda} = \arg \inf_{f \in H} \frac{1}{n} \sum_{i=1}^n L(x_i, y_i, f(x_i)) + \lambda \|f\|_H^2, \quad (2)$$

where D_n denotes the empirical distribution based on D (see (3) below). Note that the optimization problem (2) corresponds to (1) when using D_n instead of P .

Efficient algorithms to compute $\hat{f}_n := f_{L,D_n,\lambda}$ exist for a number of different loss functions. However, there are often good reasons to consider other convex loss functions, e.g. the hinge loss $L(X, Y, f(X)) = \max\{1 - Y \cdot f(X), 0\}$ for binary classification purposes or the ε -insensitive loss $L(X, Y, f(X)) = \max\{0, |Y - f(X)| - \varepsilon\}$ for regression purposes, where $\varepsilon > 0$. As these loss functions are not differentiable, the logistic loss functions $L(X, Y, f(X)) = \ln(1 + \exp(-Y \cdot f(X)))$ and $L(X, Y, f(X)) = -\ln(4e^{Y-f(X)}/(1 + e^{Y-f(X)})^2)$ and Huber-type loss functions are also used in practice. These loss functions can be considered as smoothed versions of the previous two loss functions.

An important component of statistical analyses concerns quantifying and incorporating uncertainty (e.g. sampling variability) in the reported estimates. For example, one may want to include confidence bounds along the individual predicted values $\hat{f}_n(x_i)$ obtained from (2). Unfortunately, the sampling distribution of the estimated function \hat{f}_n is unknown. Recently, [8] derived the asymptotic distribution of SVMs under some mild conditions. Asymptotic confidence intervals based on those general results are always symmetric.

Here, we are interested in approximating the finite sample distribution of SVMs by Efron's bootstrap approach, because confidence intervals based on the bootstrap approach can be asymmetric. The bootstrap [7] provides an alternative way to estimate the sampling distribution of a wide variety of estimators. To fix ideas, consider a functional $S : \mathcal{M} \rightarrow \mathcal{W}$, where \mathcal{M} is a set of probability measures and \mathcal{W} denotes a metric space. Many estimators can be included in this framework. Simple examples include the sample mean (with functional $S(P) = \int Z dP$) and M-estimators (with functional defined implicitly as the solution to the equation $\mathbb{E}_P \Psi(Z, S(P)) = 0$). Let $\mathcal{B}(\mathcal{Z})$ be the Borel σ -algebra on $\mathcal{Z} = \mathcal{X} \times \mathcal{W}$ and denote the set of all Borel

probability measures on $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$ by $\mathcal{M}_1(\mathcal{X}, \mathcal{B}(\mathcal{X}))$. Then, it follows that (1) defines an operator

$$S : \mathcal{M}_1(\mathcal{X}, \mathcal{B}(\mathcal{X})) \rightarrow H, \quad S(P) = f_{L,P,\lambda},$$

i.e. the support vector machine. Moreover, the estimator in (2) satisfies

$$f_{L,D_n,\lambda} = S(D_n)$$

where

$$D_n = \frac{1}{n} \sum_{i=1}^n \delta_{(x_i, y_i)} \quad (3)$$

is the empirical distribution based on the sample $D = ((x_1, y_1), \dots, (x_n, y_n))$ and $\delta_{(x_i, y_i)}$ denotes the Dirac measure at the point (x_i, y_i) .

More generally, let $Z_i = (X_i, Y_i)$, $i = 1, \dots, n$, be independent and identically distributed (i.i.d.) random variables with distribution P , and let

$$S_n(Z_1, \dots, Z_n) = S(P_n)$$

be the corresponding estimator, where

$$P_n = \frac{1}{n} \sum_{i=1}^n \delta_{Z_i}.$$

Denote the distribution of $S(P_n)$ by $\mathcal{L}_n(S; P) = \mathcal{L}(S(P_n))$. If P was known to us, we could estimate this sampling distribution by drawing a large number of random samples from P and evaluating our estimator on them. The basic idea of Efron's bootstrap approach is to replace the unknown distribution P by an estimate \hat{P} . Here we will consider the natural non-parametric estimator given by the sample empirical distribution P_n . In other words, we estimate the distribution of our estimator of interest by its sampling distribution when the data are generated by P_n . In symbols, the bootstrap proposes to use

$$\widehat{\mathcal{L}_n(S; P)} = \mathcal{L}_n(S; P_n).$$

Since this distribution is generally unknown, in practice one uses Monte Carlo simulation to estimate it by repeatedly evaluating the estimator on samples drawn from D_n . Note that drawing a sample from D_n means that n observations are drawn *with replacement* from the original n observations $(x_1, y_1), \dots, (x_n, y_n)$.

3 Consistency of Bootstrap SVMs

In this section it will be shown under appropriate assumptions that the weak consistency of bootstrap estimators carries over to the Hadamard-differentiable SVM

functional in the sense that the sequence of “conditional random laws” (given $(X_1, Y_1), (X_2, Y_2), \dots$) of $\sqrt{n}(f_{L, \hat{\mathbb{P}}_n, \lambda} - f_{L, \mathbb{P}_n, \lambda})$ is asymptotically consistent in probability for estimating the laws of the random elements $\sqrt{n}(f_{L, \mathbb{P}_n, \lambda} - f_{L, \mathbb{P}, \lambda})$. In other words, if n is large, the “random distribution”

$$\mathcal{L}(\sqrt{n}(f_{L, \hat{\mathbb{P}}_n, \lambda} - f_{L, \mathbb{P}_n, \lambda})) \quad (4)$$

based on bootstrapping an SVM can be considered as a valid approximation of the unknown finite sample distribution

$$\mathcal{L}(\sqrt{n}(f_{L, \mathbb{P}_n, \lambda} - f_{L, \mathbb{P}, \lambda})). \quad (5)$$

Assumption 1 Let $\mathcal{X} \subset \mathbb{R}^d$ be closed and bounded and let $\mathcal{Y} \subset \mathbb{R}$ be closed. Assume that $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ is the restriction of an m -times continuously differentiable kernel $\tilde{k} : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$ such that $m > d/2$ and $k \neq 0$. Let H be the RKHS of k and let \mathbb{P} be a probability distribution on $(\mathcal{X} \times \mathcal{Y}, \mathcal{B}(\mathcal{X} \times \mathcal{Y}))$. Let $L : \mathcal{X} \times \mathcal{Y} \times \mathbb{R} \rightarrow [0, \infty)$ be a convex, \mathbb{P} -square-integrable Nemitski loss function of order $p \in [1, \infty)$ such that the partial derivatives

$$L'(x, y, t) := \frac{\partial L}{\partial t}(x, y, t) \quad \text{and} \quad L''(x, y, t) := \frac{\partial^2 L}{\partial t^2}(x, y, t)$$

exist for every $(x, y, t) \in \mathcal{X} \times \mathcal{Y} \times \mathbb{R}$. Assume that the maps

$$(x, y, t) \mapsto L'(x, y, t) \quad \text{and} \quad (x, y, t) \mapsto L''(x, y, t)$$

are continuous. Furthermore, assume that for every $a \in (0, \infty)$, there is a $b'_a \in L_2(\mathbb{P})$ and a constant $b''_a \in [0, \infty)$ such that, for every $(x, y) \in \mathcal{X} \times \mathcal{Y}$,

$$\sup_{t \in [-a, a]} |L'(x, y, t)| \leq b'_a(x, y) \quad \text{and} \quad \sup_{t \in [-a, a]} |L''(x, y, t)| \leq b''_a. \quad (6)$$

The conditions on the kernel k in Assumption 1 are satisfied for many common kernels, e.g., Gaussian RBF kernel, exponential kernel, polynomial kernel, and linear kernel, but also Wendland kernels $k_{d, \ell}$ based on certain univariate polynomials $p_{d, \ell}$ of degree $\lfloor d/2 \rfloor + 3\ell + 1$ for $\ell \in \mathbb{N}$ such that $\ell > d/4$, see [17].

The conditions on the loss function L in Assumption 1 are satisfied, e.g., for the logistic loss for classification or for regression, however the popular non-smooth loss functions hinge, ε -insensitive, and pinball are not covered. However, [8, Remark 3.5] described an analytical method to approximate such non-smooth loss functions up to an arbitrarily good precision $\varepsilon > 0$ by a convex \mathbb{P} -square integrable Nemitski loss function of order $p \in [1, \infty)$.

We can now state our result on the consistency of the bootstrap approach for SVMs.

Theorem 2. *Let Assumption 1 be satisfied. Let $\lambda \in (0, \infty)$. Then*

$$\sup_{h \in \text{BL}_1(H)} |\mathbb{E}_M h(\sqrt{n}(f_{L, \hat{\mathbb{P}}_n, \lambda} - f_{L, \mathbb{P}_n, \lambda})) - \mathbb{E} h(S'_P(\mathbb{G}))| \rightarrow 0, \quad (7)$$

$$\mathbb{E}_M h(\sqrt{n}(f_{L, \hat{\mathbb{P}}_n, \lambda} - f_{L, \mathbb{P}_n, \lambda}))^* - \mathbb{E}_M h(\sqrt{n}(f_{L, \hat{\mathbb{P}}_n, \lambda} - f_{L, \mathbb{P}_n, \lambda}))_* \rightarrow 0, \quad (8)$$

converge in outer probability, where \mathbb{G} is a tight Borel-measurable Gaussian process, S'_P is a continuous linear operator with

$$S'_P(Q) = -K_P^{-1}(\mathbb{E}_Q(L'(X, Y, f_{L, P, \lambda}(X))\Phi(X))), \quad Q \in \mathcal{M}_1(\mathcal{X} \times \mathcal{Y}) \quad (9)$$

and

$$K_P : H \rightarrow H, \quad f \mapsto 2\lambda f + \mathbb{E}_P(L''(X, Y, f_{L, P, \lambda}(X))f(X)\Phi(X)) \quad (10)$$

is a continuous linear operator which is invertible.

For details on K_P , S'_P , and \mathbb{G} we refer to Lemma 1, Theorem 6, and Lemma 2.

4 Proofs

4.1 Tools for the proof of Theorem 2

We will need two general results on bootstrap methods proven in [13] and adopt their notation, see [13, Chapters 3.6 and 3.9]. Let \mathbb{P}_n be the empirical measure of an i.i.d. sample Z_1, \dots, Z_n from a probability distribution P . The *empirical process* is the signed measure

$$\mathbb{G}_n = \sqrt{n}(\mathbb{P}_n - P).$$

Given the sample values, let $\hat{Z}_1, \dots, \hat{Z}_n$ be an i.i.d. sample from $\hat{\mathbb{P}}_n$. The *bootstrap empirical distribution* is the empirical measure $\hat{\mathbb{P}}_n := n^{-1} \sum_{i=1}^n \delta_{\hat{Z}_i}$, and the *bootstrap empirical process* is

$$\hat{\mathbb{G}}_n = \sqrt{n}(\hat{\mathbb{P}}_n - \mathbb{P}_n) = \frac{1}{\sqrt{n}} \sum_{i=1}^n (M_{ni} - 1) \delta_{Z_i},$$

where M_{ni} is the number of times that Z_i is “redrawn” from the original sample Z_1, \dots, Z_n , $M := (M_{n1}, \dots, M_{nn})$ is stochastically independent of Z_1, \dots, Z_n and multinomially distributed with parameters n and probabilities $\frac{1}{n}, \dots, \frac{1}{n}$. If outer expectations are computed, stochastic independence is understood in terms of a product probability space. Let Z_1, Z_2, \dots be the coordinate projections on the first ∞ coordinates of the product space $(\mathcal{Z}^\infty, \mathcal{B}(\mathcal{Z}), P^\infty) \times (\mathcal{Z}, \mathcal{C}, Q)$ and let the multinomial vectors M depend on the last factor only, see [13, p. 345f].

The following theorem shows (conditional) weak convergence for the empirical bootstrap, where the symbol \rightsquigarrow denotes the weak convergence of finite measures.

We will need only the equivalence between (i) and (iii) from this theorem and list part (ii) only for the sake of completeness.

Theorem 3 ([13, Thm. 3.6.2, p. 347]). *Let \mathcal{F} be a class of measurable functions with finite envelope function. Define $\mathbb{Y}_n := n^{-1/2} \sum_{i=1}^n (M_{N_n,i} - 1)(\delta_{Z_i} - P)$. The following statements are equivalent:*

- (i) \mathcal{F} is Donsker and $P^* \|f - Pf\|_{\mathcal{F}}^2 < \infty$;
- (ii) $\sup_{h \in \text{BL}_1} |\mathbb{E}_{M,N} h(\hat{\mathbb{Y}}_n) - \mathbb{E} h(\mathbb{G})|$ converges outer almost surely to zero and the sequence $\mathbb{E}_{M,N} h(\hat{\mathbb{Y}}_n)^* - \mathbb{E}_{M,N} h(\hat{\mathbb{Y}}_n)_*$ converges almost surely to zero for every $h \in \text{BL}_1$.
- (iii) $\sup_{h \in \text{BL}_1} |\mathbb{E}_M h(\hat{\mathbb{G}}_n) - \mathbb{E} h(\mathbb{G})|$ converges outer almost surely to zero and the sequence $\mathbb{E}_M h(\hat{\mathbb{G}}_n)^* - \mathbb{E}_M h(\hat{\mathbb{G}}_n)_*$ converges almost surely to zero for every $h \in \text{BL}_1$.

Here the asterisks denote the measurable cover functions with respect to M, N , and Z_1, Z_2, \dots jointly.

Consider sequences of random elements $\mathbb{P}_n = \mathbb{P}_n(Z_n)$ and $\hat{\mathbb{P}}_n = \hat{\mathbb{P}}_n(Z_n, M_n)$ in a normed space \mathbb{D} such that the sequence $\sqrt{n}(\mathbb{P}_n - P)$ converges unconditionally and the sequence $\sqrt{n}(\hat{\mathbb{P}}_n - \mathbb{P}_n)$ converges conditionally on Z_n in distribution to a tight random element \mathbb{G} . A precise formulation of the second assumption is

$$\sup_{h \in \text{BL}_1(\mathbb{D})} |\mathbb{E}_M h(\sqrt{n}(\hat{\mathbb{P}}_n - \mathbb{P}_n)) - \mathbb{E} h(\mathbb{G})| \rightarrow 0, \quad (11)$$

$$\mathbb{E}_M h(\sqrt{n}(\hat{\mathbb{P}}_n - \mathbb{P}_n))^* - \mathbb{E}_M h(\sqrt{n}(\hat{\mathbb{P}}_n - \mathbb{P}_n))_* \rightarrow 0, \quad (12)$$

in outer probability, with h ranging over the bounded Lipschitz functions, see [13, p. 378, Formula (3.9.9)]. The next theorem shows that under appropriate assumptions, weak consistency of the bootstrap estimators carries over to any Hadamard-differentiable functional in the sense that the sequence of “conditional random laws” (given Z_1, Z_2, \dots) of $\sqrt{n}(\phi(\hat{\mathbb{P}}_n) - \phi(\mathbb{P}_n))$ is asymptotically consistent in probability for estimating the laws of the random elements $\sqrt{n}(\phi(\mathbb{P}_n) - \phi(P))$, see [13, p. 378].

Theorem 4 ([13, Thm. 3.9.11, p. 378]). *(Delta-method for bootstrap in probability) Let \mathbb{D} and \mathbb{E} be normed spaces. Let $\phi : \mathbb{D}_\phi \subset \mathbb{D} \rightarrow \mathbb{E}$ be Hadamard-differentiable at P tangentially to a subspace \mathbb{D}_0 . Let \mathbb{P}_n and $\hat{\mathbb{P}}_n$ be maps as indicated previously with values in \mathbb{D}_ϕ such that $\mathbb{G}_n := \sqrt{n}(\mathbb{P}_n - P) \rightsquigarrow \mathbb{G}$ and that (11)-(12) holds in outer probability, where \mathbb{G} is separable and takes its values in \mathbb{D}_0 . Then*

$$\sup_{h \in \text{BL}_1(\mathbb{E})} |\mathbb{E}_M h(\sqrt{n}(\phi(\hat{\mathbb{P}}_n) - \phi(\mathbb{P}_n))) - \mathbb{E} h(\phi'_P(\mathbb{G}))| \rightarrow 0, \quad (13)$$

$$\mathbb{E}_M h(\sqrt{n}(\phi(\hat{\mathbb{P}}_n) - \phi(\mathbb{P}_n)))^* - \mathbb{E}_M h(\sqrt{n}(\phi(\hat{\mathbb{P}}_n) - \phi(\mathbb{P}_n)))_* \rightarrow 0, \quad (14)$$

holds in outer probability.

As was pointed out by [13, p. 378], consistency in probability appears to be sufficient for (many) statistical purposes and the theorem above shows this is retained under Hadamard differentiability at the single distribution P .

We now list some results from [8], which will also be essential for the proof of Theorem 2.

Theorem 5 ([8, Theorem 3.1]). *Let Assumption 1 be satisfied. Then, for every regularizing parameter $\lambda_0 \in (0, \infty)$, there is a tight, Borel-measurable Gaussian process $\mathbb{H} : \Omega \rightarrow H$, $\omega \rightarrow \mathbb{H}(\omega)$, such that*

$$\sqrt{n}(f_{L, \mathbf{D}_n, \lambda_{\mathbf{D}_n}} - f_{L, P, \lambda_0}) \rightsquigarrow \mathbb{H} \quad \text{in } H \quad (15)$$

for every Borel-measurable sequence of random regularization parameters $\lambda_{\mathbf{D}_n}$ with $\sqrt{n}(\lambda_{\mathbf{D}_n} - \lambda_0) \rightarrow 0$ in probability. The Gaussian process \mathbb{H} is zero-mean; i.e., $\mathbb{E}\langle f, \mathbb{H} \rangle_H = 0$ for every $f \in H$.

Lemma 1 ([8, Lemma A.5]). *For every $F \in B_S$ defined later in (25),*

$$K_F : H \rightarrow H, \quad f \mapsto 2\lambda_0 f + \int L''(x, y, f_{L, \mathbf{t}(F), \lambda_0}(x)) f(x) \Phi(x) d\mathbf{t}(F)(x, y) \quad (16)$$

is a continuous linear operator which is invertible.

Theorem 6 ([8, Theorem A.8]). *For every $F_0 \in B_S$ which fulfills $F_0(b) < \mathbb{E}_P(b) + \lambda_0$, the map $S : B_S \rightarrow H$, $F \mapsto f_{\mathbf{t}(F)}$, is Hadamard-differentiable in F_0 tangentially to the closed linear span $B_0 = \text{cl}(\text{lin}(B_S))$. The derivative in F_0 is a continuous linear operator $S'_{F_0} : B_0 \rightarrow H$ such that*

$$S'_{F_0}(G) = -K_{F_0}^{-1}(\mathbb{E}_{\mathbf{t}(G)}(L'(X, Y, f_{L, \mathbf{t}(F_0), \lambda_0}(X))\Phi(X))), \quad \forall G \in \text{lin}(B_S). \quad (17)$$

Lemma 2 ([8, Lemma A.9]). *For every data set $D_n = ((x_1, y_1), \dots, (x_n, y_n)) \in (\mathcal{X} \times \mathcal{Y})^n$, let \mathbb{F}_{D_n} denote the element of $\ell_\infty(\mathcal{G})$ which corresponds to the empirical measure $\mathbb{P}_n := \mathbb{P}_{D_n}$. That is, $\mathbb{F}_{D_n}(g) = \int g d\mathbb{P}_n = n^{-1} \sum_{i=1}^n g(x_i, y_i)$ for every $g \in \mathcal{G}$. Then*

$$\sqrt{n}(\mathbb{F}_{D_n} - \mathbf{t}^{-1}(P)) \rightsquigarrow \mathbb{G} \quad \text{in } \ell_\infty(\mathcal{G}), \quad (18)$$

where $\mathbb{G} : \Omega \rightarrow \ell_\infty(\mathcal{G})$ is a tight Borel-measurable Gaussian process such that $\mathbb{G}(\omega) \in B_0$ for every $\omega \in \Omega$.

4.2 Proof of Theorem 2

The proof relies on the application of Theorem 4. Hence, we have to show the following steps:

1. The empirical process $\mathbb{G}_n = \sqrt{n}(\mathbb{P}_n - P)$ weakly converges to a separable Gaussian process \mathbb{G} .

2. SVMs are based on a map ϕ which is Hadamard differentiable at P tangentially to some appropriate subspace.
3. The assumptions (11)-(12) of Theorem 4 are satisfied. For this purpose we will use Theorem 3. Actually, we will show that part (i) of Theorem 3 is satisfied which gives the equivalence to part (iii), from which we conclude that (11)-(12) hold true. For the proof that part (i) of Theorem 3 is satisfied, i.e., that a suitable set \mathcal{F} is a P-Donsker class and that $P^*\|f - Pf\|_{\mathcal{F}}^2 < \infty$, we use several facts recently shown by [8].
4. We put all parts together and apply Theorem 4.

Step 1. To apply Theorem 4, we first have to specify the considered spaces \mathbb{D} , \mathbb{E} , \mathbb{D}_ϕ , \mathbb{D}_0 and the map ϕ . As in [8] we use the following notations. Because L is a P-square-integrable Nemitski loss function of order $p \in [1, \infty)$, there is a function $b \in L_2(P)$ such that

$$|L(x, y, t)| \leq b(x, y) + |t|^p, \quad (x, y, t) \in \mathcal{X} \times \mathcal{Y} \times \mathbb{R}. \quad (19)$$

Let

$$c_0 := \sqrt{\lambda_0^{-1} \mathbb{E}_P(b)} + 1, \quad (20)$$

Define

$$\mathcal{G} := \mathcal{G}_1 \cup \mathcal{G}_2 \cup \mathcal{G}_3, \quad (21)$$

where

$$\mathcal{G}_1 := \{g : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R} : \exists z \in \mathbb{R}^{d+1} \text{ such that } g = I_{(-\infty, z]}\} \quad (22)$$

is the set of all indicator functions $I_{(-\infty, z]}$,

$$\mathcal{G}_2 := \left\{ g : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R} \left| \begin{array}{l} \exists f_0 \in H, \exists f \in H \text{ such that } \|f_0\|_H \leq c_0, \\ \|f\|_H \leq 1, g(x, y) = L'(x, y, f_0(x))f(x) \forall (x, y) \end{array} \right. \right\}, \quad (23)$$

and

$$\mathcal{G}_3 := \{b\}. \quad (24)$$

Now let $\ell_\infty(\mathcal{G})$ be the set of all bounded functions $F : \mathcal{G} \rightarrow \mathbb{R}$ with norm $\|F\|_\infty = \sup_{g \in \mathcal{G}} |F(g)|$. Define

$$B_S := \left\{ F : \mathcal{G} \rightarrow \mathbb{R} \left| \begin{array}{l} \exists \mu \neq 0 \text{ a finite measure on } \mathcal{X} \times \mathcal{Y} \text{ such that} \\ F(g) = \int g d\mu \forall g \in \mathcal{G}, \\ b \in L_2(\mu), b'_a \in L_2(\mu) \forall a \in (0, \infty) \end{array} \right. \right\} \quad (25)$$

and

$$B_0 := \text{cl}(\text{lin}(B_S)) \quad (26)$$

the closed linear span of B_S in $\ell_\infty(\mathcal{G})$. That is, B_S is a subset of $\ell_\infty(\mathcal{G})$ whose elements correspond to finite measures. Hence probability measures are covered as special cases. The elements of B_S can be interpreted as some kind of generalized distributions functions, because $\mathcal{G}_1 \subset \mathcal{G}$. The assumptions on L and P imply that $\mathcal{G} \rightarrow \mathbb{R}$,

$g \mapsto \int g dP$ is a well-defined element of B_S . For every $F \in B_S$, let $\iota(F)$ denote the corresponding finite measure on $(\mathcal{X} \times \mathcal{Y}, \mathcal{B}(\mathcal{X} \times \mathcal{Y}))$ such that $F(g) = \int g d\mu$ for all $g \in \mathcal{G}$. Note that the map ι is well-defined, because by definition of B_S , $\iota(F)$ uniquely exists for every $F \in B_S$.

With these notations, we will apply Theorem 4 for

$$\begin{aligned} \mathbb{D} &:= \ell_\infty(\mathcal{G}), \quad \mathbb{E} := H \text{ (RKHS of the kernel } k), \\ \mathbb{D}_\phi &:= B_S, \quad \mathbb{D}_0 := B_0 := \text{cl}(\text{lin}(B_S)), \\ \lambda_0 &\in (0, \infty), \\ \phi &:= S, \quad S : B_S \rightarrow H, \quad F \mapsto f_{\iota(F)} := f_{L, \iota(F), \lambda_0} := \\ &\quad \arg \inf_{f \in H} \int L(x, y, f(x)) d\iota(F)(x, y) + \lambda_0 \|f\|_H^2. \end{aligned} \tag{27}$$

At first glance this definition of S seems to be somewhat technical. However, this will allow us to use a functional delta method for bootstrap estimators of SVMs with regularization parameter $\lambda = \lambda_0 \in (0, \infty)$.

Lemma 2 guarantees that the empirical process $\mathbb{G}_n := \sqrt{n}(\mathbb{P}_n - P)$ weakly converges to a tight Borel-measurable Gaussian process.

Since a σ -compact set in a metric space is separable, separability of a random variable is slightly weaker than tightness, see [13, p. 17]. Therefore, \mathbb{G} in our Theorem 2 is indeed separable.

Step 2. Theorem 6 showed that the map S indeed satisfies the necessary Hadamard-differentiability in the point $P := \iota^{-1}(F)$.

Step 3. We know that \mathcal{G} is a P-Donsker class, see Lemma 2. Hence, an immediate consequence from [13, Theorem 3.6.1, p. 347] is, that

$$\sup_{h \in \text{BL}_1} |\mathbb{E}_M h(\hat{\mathbb{G}}_n) - \mathbb{E} h(\mathbb{G})| \tag{28}$$

converges in outer probability to zero and $\hat{\mathbb{G}}_n$ is asymptotically measurable.

However, we will prove a somewhat stronger result, namely that \mathcal{G} is a P-Donsker class and $P^* \|g - Pg\|_{\mathcal{G}}^2 < \infty$, which is part (i) of Theorem 3, and then part (iii) of Theorem 3 yields, that the term in (28) converges even outer almost surely to zero and the sequence

$$\mathbb{E}_M h(\hat{\mathbb{G}}_n)^* - \mathbb{E}_M h(\hat{\mathbb{G}}_n)_* \tag{29}$$

converges almost surely to zero for every $h \in \text{BL}_1$.

Because \mathcal{G} is a P-Donsker class, it remains to show that $P^* \|g - Pg\|_{\mathcal{G}}^2 < \infty$. Due to

$$P^* \|g - Pg\|_{\mathcal{G}}^2 := \int \left(\sup_{g \in \mathcal{G}} |g - \mathbb{E}_P(g)| \right)^2 dP^* \tag{30}$$

and $\mathcal{G} = \mathcal{G}_1 \cup \mathcal{G}_2 \cup \mathcal{G}_3$, we obtain the inequality

$$\begin{aligned}
\mathbf{P}^* \|g - \mathbf{P}g\|_{\mathcal{G}}^2 &\leq \mathbf{P}^* \sup_{g \in \mathcal{G}} (g^2 + 2|g| \cdot \mathbf{P}|g| + (\mathbf{P}|g|)^2) \\
&\leq \mathbf{P}^* \sup_{g \in \mathcal{G}} g^2 + 2\mathbf{P}^* \sup_{g \in \mathcal{G}} (|g| \cdot \mathbf{P}|g|) + \sup_{g \in \mathcal{G}} (\mathbf{P}|g|)^2 \\
&\leq \sum_{j=1}^3 \left(\mathbf{P}^* \sup_{g \in \mathcal{G}_j} g^2 + 2\mathbf{P}^* \sup_{g \in \mathcal{G}_j} (|g| \cdot \mathbf{P}|g|) + \sup_{g \in \mathcal{G}_j} (\mathbf{P}|g|)^2 \right). \quad (31)
\end{aligned}$$

We will show that each of the three summands on the right hand side of the last inequality is finite. If $g \in \mathcal{G}_1$, then g equals the indicator function $I_{(-\infty, z]}$ for some $z \in \mathbb{R}^{d+1}$. Hence, $\|g\|_{\infty} = 1$ and the summand for $j = 1$ is finite. If $g \in \mathcal{G}_3$, then $g = b \in L_2(\mathbf{P})$ because L is by assumption a \mathbf{P} -square-integrable Nemitski loss function of order $p \in [1, \infty)$. Hence the summand for $j = 3$ is finite, too. Let us now consider the case that $g \in \mathcal{G}_2$. By definition of \mathcal{G}_2 , for every $g \in \mathcal{G}_2$ there exist $f, f_0 \in H$ such that $\|f_0\|_H \leq c_0$, $\|f\|_H \leq 1$, and $g = L'_{f_0} f$, where we used the notation $(L'_{f_0} f)(x, y) := L'(x, y, f_0(x))f(x)$ for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$. Using $\|f\|_{\infty} \leq \|k\|_{\infty} \|f\|_H$ for every $f \in H$, we obtain

$$\|f_0\|_H \leq c_0 \Rightarrow \|f_0\|_{\infty} \leq c_0 \|k\|_{\infty} \quad \text{and} \quad \|f\|_H \leq 1 \Rightarrow \|f\|_{\infty} \leq \|k\|_{\infty}. \quad (32)$$

Define the constant $a := c_0 \|k\|_{\infty}$ with c_0 given by (20). Hence, for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$,

$$\begin{aligned}
\sup_{f_0 \in H; \|f_0\|_H \leq c_0} |L'(x, y, f_0(x))|^2 &\leq \sup_{f_0 \in H; \|f_0\|_{\infty} \leq a} \sup_{t \in [-a, +a]} |L'(x, y, t)|^2 \\
&\stackrel{(6)}{\leq} \sup_{f_0 \in H; \|f_0\|_{\infty} \leq a} (b'_a(x, y))^2. \quad (33)
\end{aligned}$$

Hence we get

$$\begin{aligned}
&\mathbf{P}^* \sup_{g \in \mathcal{G}_2} g^2 \\
&= \int \sup_{g \in \mathcal{G}_2; \|f_0\|_H \leq c_0, \|f\|_H \leq 1, g = L'_{f_0} f} |L'(x, y, f_0(x))f(x)|^2 d\mathbf{P}^*(x, y) \\
&\leq \int \sup_{f_0 \in H; \|f_0\|_H \leq c_0} |L'(x, y, f_0(x))|^2 \sup_{f \in H; \|f\|_H \leq 1} |f(x)|^2 d\mathbf{P}^*(x, y) \\
&\stackrel{(33), (32)}{\leq} \|k\|_{\infty}^2 \int (b'_a)^2 d\mathbf{P}^* = \|k\|_{\infty}^2 \int (b'_a)^2 d\mathbf{P} < \infty,
\end{aligned}$$

because $b'_a \in L_2(\mathbf{P})$ and $\|k\|_{\infty} < \infty$ by Assumption 1. With the same arguments we obtain, for every $g \in \mathcal{G}_2$,

$$\begin{aligned}
\mathbb{P}|g| &\leq \int \sup_{g \in \mathcal{G}_2} |g| d\mathbb{P}^* \\
&\leq \int \sup_{f_0 \in H; \|f_0\|_H \leq c_0} |L'(x, y, f_0(x))| \sup_{f \in H; \|f\|_H \leq 1} |f(x)| d\mathbb{P}^*(x, y) \\
&\stackrel{(33), (32)}{\leq} \int b'_a(x, y) \|k\|_\infty d\mathbb{P}^*(x, y) \\
&\leq \|k\|_\infty \int b'_a d\mathbb{P} < \infty,
\end{aligned}$$

because $b'_a \in L_2(\mathbb{P})$ and $\|k\|_\infty < \infty$ by Assumption 1. Hence,

$$\mathbb{P}^* \sup_{g \in \mathcal{G}_2} (|g| \mathbb{P}|g|) \leq \|k\|_\infty \int b'_a d\mathbb{P} \int \sup_{g \in \mathcal{G}_2} |g| d\mathbb{P}^* \leq \|k\|_\infty^2 \left(\int b'_a d\mathbb{P} \right)^2 < \infty.$$

Therefore, the sum on the right hand side in (31) is finite and thus the assumption $\mathbb{P}^* \|g - \mathbb{P}g\|_{\mathcal{G}}^2 < \infty$ is satisfied. This yields by part (iii) of Theorem 3 that $\sup_{h \in \text{BL}_1} |\mathbb{E}_M h(\hat{\mathbb{G}}_n) - \mathbb{E} h(\mathbb{G})|$ converges outer almost surely to zero and the sequence

$$\mathbb{E}_M h(\hat{\mathbb{G}}_n)^* - \mathbb{E}_M h(\hat{\mathbb{G}}_n)_* \tag{34}$$

converges almost surely to zero for every $h \in \text{BL}_1$, where the asterisks denote the measurable cover functions with respect to M and Z_1, Z_2, \dots jointly.

Step 4. Due to Step 3, the assumption (11) of Theorem 4 is satisfied. We now show that additionally (12) is satisfied, i.e., that the term in (34) converges to zero in outer probability. In general, one can *not* conclude that almost sure convergence implies convergence in outer probability, see [13, p. 52]. We know that the term in (34) converges almost surely to zero for every $h \in \text{BL}_1$, where the asterisks denote the *measurable* cover functions with respect to M and $(X_1, Y_1), (X_2, Y_2), \dots$ *jointly*. Hence, for every $h \in \text{BL}_1$, the cover functions to be considered in (34) are measurable. Additionally, the multinomially distributed random variable M is stochastically independent of $(X_1, Y_1), \dots, (X_n, Y_n)$ in the bootstrap, where independence is understood in terms of a product probability space, see [13, p. 346] for details. Therefore, an application of the Fubini-Tonelli theorem, see e.g., [4, p. 174, Thm. 2.4.10], yields that the inner integral $\mathbb{E}_M h(\sqrt{n}(\hat{\mathbb{P}}_n - \mathbb{P}_n))^* - \mathbb{E}_M h(\sqrt{n}(\hat{\mathbb{P}}_n - \mathbb{P}_n))_*$ considered by Fubini-Tonelli is *measurable* for every $n \in \mathbb{N}$ and every $h \in \text{BL}_1$. Recall that almost sure convergence of measurable functions implies convergence in probability which is equivalent with convergence in outer probability for measurable functions. Hence we have convergence in outer probability in (34). Therefore, all assumptions of Theorem 4 are satisfied and the assertion of our theorem follows. ■

References

1. B. E. Boser, I. Guyon, and V. Vapnik. A training algorithm for optimal margin classifiers. In *Computational Learning Theory*, pages 144–152, Madison, WI, 1992. ACM.
2. A. Christmann, M. Salibián-Barrera, and S. Van Aelst. *Qualitative Robustness of Bootstrap Approximations for Kernel Based Methods*, chapter 16 in C. Becker, R. Fried, S. Kuhnt (Eds.). “Robustness and Complex Data Structures” (Preprint available on <http://arxiv.org/abs/1111.1876>). Springer, Heidelberg, 2013.
3. F. Cucker and D.X. Zhou. *Learning Theory: An Approximation Theory Viewpoint*. Cambridge University Press, Cambridge, 2007.
4. Z. Denkowski, S. Migórski, and N.S. Papageorgiou. *An introduction to nonlinear analysis: Theory*. Kluwer Academic Publishers, Boston, 2003.
5. L. Devroye. Any discrimination rule can have an arbitrarily bad probability of error for finite sample size. *IEEE Trans. Pattern Anal. Mach. Intell.*, 4:154–157, 1982.
6. L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer, New York, 1996.
7. B. Efron. Bootstrap methods: Another look at the jackknife. *Annals of Statistics*, 7:1–26, 1979.
8. R. Hable. Asymptotic normality of support vector machine variants and other regularized kernel methods. *Journal of Multivariate Analysis*, 106:92–117, 2012.
9. T. Poggio and F. Girosi. Networks for approximation and learning. *Proc. IEEE*, 78:1481–1497, 1990.
10. B. Schölkopf and A. J. Smola. *Learning with Kernels. Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, Cambridge, MA, 2002.
11. S. Smale and D.-X. Zhou. Learning theory estimates via integral operators and their approximations. *Constr. Approx.*, 26:153–172, 2007.
12. I. Steinwart and A. Christmann. *Support Vector Machines*. Springer, New York, 2008.
13. A. W. van der Vaart and J. A. Wellner. *Weak convergence and empirical processes*. Springer, New York, 1996.
14. V. N. Vapnik. *The Nature of Statistical Learning Theory*. Springer, New York, 1995.
15. V. N. Vapnik. *Statistical Learning Theory*. John Wiley & Sons, New York, 1998.
16. V. N. Vapnik and A. Lerner. Pattern recognition using generalized portrait method. *Autom. Remote Control*, 24:774–780, 1963.
17. H. Wendland. *Scattered Data Approximation*. Cambridge University Press, Cambridge, 2005.