Asymptotic Normality of Support Vector Machines for Classification and Regression

Robert Hable

Department of Mathematics

University of Bayreuth

Abstract

In nonparametric classification and regression problems, support vector machines (SVMs) attract much attention in theoretical and in applied statistics. In an abstract sense, SVMs can be seen as regularized M-estimators for a parameter in a (typically infinite dimensional) reproducing kernel Hilbert space. In the article, it is shown that the difference between the empirical SVM $f_{L,\mathbf{D}_n,\lambda_{\mathbf{D}_n}}$ and the theoretical SVM f_{L,P,λ_0} is asymptotically normal with rate $\sqrt{n}$. That is, $\sqrt{n}(f_{L,\mathbf{D}_n,\lambda_{\mathbf{D}_n}} - f_{L,P,\lambda_0})$ converges weakly to a Gaussian process in the reproducing kernel Hilbert space. This is done by an application of the functional delta-method and by showing that the SVM-functional $P \mapsto f_{L,P,\lambda}$ is suitably Hadamard-differentiable.

1 Introduction

One of the most important tasks in statistics is the estimation of the influence of an input variable X on an output variable Y. On the basis of a finite data set $(x_1, y_1), \ldots, (x_n, y_n) \in \mathcal{X} \times \mathcal{Y}$, 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 case of a finite space $\mathcal{Y}$, this is called classification and, in case of an infinite space $\mathcal{Y} \subset \mathbb{R}$, this is called regression. Often, a signal plus noise relationship $y = f_0(x) + \varepsilon$ is assumed and the task is to estimate the unknown regression function f_0. In parametric statistics, it is assumed that f_0 is contained in a known finite-dimensional function space. This assumption is dropped or, at least, considerably weakened in nonparametric

statistics. As a method of nonparametric classification and regression, support vector machines (SVMs) recently attract much attention in theoretical and in applied statistics; see e.g. the comprehensive books Vapnik (1998), Schölkopf and Smola (2002), and Steinwart and Christmann (2008) and the references cited therein. Typically, the weaker assumptions in nonparametric statistics have to be compensated by an increase of observations in order to obtain the same precision of the estimation. Nevertheless, it is well-known that some nonparametric estimators still are asymptotically normal for the same rate $\sqrt{n}$ as many parametric estimators. In this article, it is shown that also support vector machines enjoy an asymptotic normality property for the rate $\sqrt{n}$. For an i.i.d. sample $D_n = \big((x_1, y_1), \ldots, (x_1, y_n)\big)$ from a distribution P, the *empirical SVM* is a function $f_{L, D_n, \lambda_{D_n}}$ which solves the minimization problem

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

where L is a loss function and H is a certain space of functions $f : \mathcal{X} \to \mathbb{R}$, namely a so-called *reproducing kernel Hilbert space*. The first term in (1) is the empirical mean of the losses caused by the predictions $f(x_i)$ and the second term penalizes the complexity of f in order to avoid overfitting; the regularization parameter λ_{D_n} is a positive real number which is typically chosen in a data-driven way, e.g., by cross-validation. In this article, the term "support vector machine" (SVM) is used in a broad sense covering both classification and regression and the use of any (suitable) loss function. Originally, the term "support vector machine" was coined for the special case where $\mathcal{Y} = \{-1, 1\}$ (binary classification) and where the loss function L is the so-called *hinge-loss*.

Depending on the size of the space H, SVMs can be used as a parametric or a non-parametric method. Choosing a finite-dimensional H leads to a parametric setting, choosing an infinite-dimensional H leads to a non-parametric setting. In the parametric setting, asymptotic normality of support vector machines in the original sense (binary classification using the hinge loss) has already been investigated: Jiang et al. (2008) derive asymptotic normality of the estimated prediction error of SVMs with finite-dimensional H. Under some regularity conditions on the distribution of the data, Koo et al. (2008) show asymptotic normality of the coefficients of the linear SVM (i.e., H only contains linear functions). In the following, a general non-parametric setting (covering classification and regression) is considered but, by going over from parametrics to non-parametrics, we have to impose a bound on

the complexity of the predictor. Instead of estimating a solution $f_{L,P}^*$ of the (ill-posed) minimization problem

$$\min_{f \in H} \int L\big((x,y,f(x)\big) \, P\big(d(x,y)\big) \,, \tag{2}$$

we estimate a smoother approximation, namely the solution f_{L,P,λ_0} of the minimization problem

$$\min_{f \in H} \int L\big((x,y,f(x)\big) \, P\big(d(x,y)\big) \,+\, \lambda_0 \|f\|_H^2 \tag{3}$$

for a fixed regularization parameter $\lambda_0 \in (0,\infty)$. The minimizer f_{L,P,λ_0} of (3) is called *theoretical SVM*. This so-called Tikhonov regularization is equivalent to a minimization problem

$$\int L\big((x,y,f(x)\big) \, P\big(d(x,y)\big) \;\; = \;\; \min! \qquad f \in H, \quad \|f\|_H \le r_0$$

where r_0 can be interpreted as an upper bound on the complexity of the function f; a smaller $\lambda_0 > 0$ corresponds to a larger $r_0 > 0$. It will be shown that the sequence of SVM-estimators

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

is asymptotically normal for the rate $\sqrt{n}$ if the empirical SVM $f_{L,D_n,\lambda_{D_n}}$ is shifted by the theoretical SVM f_{L,P,λ_0}. That is,

$$\sqrt{n}\big(f_{L,D_n,\lambda_{D_n}} - f_{L,P,\lambda_0}\big)$$

converges weakly to a (zero-mean) Gaussian process in the function space H. This also implies asymptotic normality of the risk

$$\sqrt{n}\Big(\mathcal{R}_{L,P}\big(f_{L,D_n,\lambda_{D_n}}\big) - \mathcal{R}_{L,P}\big(f_{L,P,\lambda_0}\big)\Big) \;\;\rightsquigarrow\;\; \mathcal{N}(0,\sigma^2)$$

where $\mathcal{R}_{L,P}(f) = \int L(x,y,f(x)) \, P\big(d(x,y)\big)$ denotes the risk of a predictor f and $\sigma \in [0,\infty)$ is a standard deviation. The regularization parameter λ_{D_n} for the empirical SVM may depend on the data. We only need that $\sqrt{n}(\lambda_{D_n} - \lambda_0)$ converges to 0 in probability. These results will be proven by an advanced application of a functional delta-method. Accordingly, it will be shown that the map $P \mapsto f_{L,P,\lambda}$ is suitably Hadamard-differentiable.

Of course, it would be desirable to dispense with the complexity bound and to have asymptotic normality of

$$\sqrt{n}(f_{L,D_n,\lambda_{D_n}} - f_{L,P}^*) \qquad \text{instead of} \qquad \sqrt{n}(f_{L,D_n,\lambda_{D_n}} - f_{L,P,\lambda_0})$$

3

– if $f_{L,P}^*$ exists at all. However, in the non-parametric setting where H is a large infinite-dimensional function space, this is not possible. Such a result would violate the no-free-lunch theorem which, roughly speaking, yields that there is no uniform rate of convergence without such a bound on the complexity. It is only possible to get uniform rates of convergence within special classes of distributions. The investigation of rates of convergence for special cases – e.g. classification under assumptions on the unknown true probability measure such as Tsybakov's noise assumption (Tsybakov, 2004, p. 138) – is one of the most important topics of recent research about support vector machines and related learning methods; see e.g. Steinwart and Scovel (2007), Caponnetto and De Vito (2007), Blanchard et al. (2008), Steinwart et al. (2009), Mendelson and Neeman (2010). It is a matter of further research if similar assumptions on the unknown true probability measures allow asymptotic normality of $\sqrt{n}(f_{L,D_n,\lambda_{D_n}} - f_{L,P}^*)$.

The article is organized as follows: Section 2 briefly recalls the definition of support vector machines in a broad sense and fixes the notation. Section 3.1 contains the main results concerning asymptotic normality of support vector machines and their risks. Since the proof is quite involved, it is deferred to the appendix but Section 3.2 provides a short outline. Finally, Sections 4 contains some concluding remarks.

2   Support Vector Machines

Let $(\Omega, \mathcal{A}, Q)$ be a probability space, let $\mathcal{X}$ be a closed and bounded subset of $\mathbb{R}^d$, and 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 $\mathfrak{B}(\mathcal{X} \times \mathcal{Y})$. Let

$$X_1, \ldots, X_n \; : \; (\Omega, \mathcal{A}, Q) \; \longrightarrow \; \big(\mathcal{X}, \mathfrak{B}(\mathcal{X})\big),$$
$$Y_1, \ldots, Y_n \; : \; (\Omega, \mathcal{A}, Q) \; \longrightarrow \; \big(\mathcal{Y}, \mathfrak{B}(\mathcal{Y})\big)$$

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 P on $\big(\mathcal{X} \times \mathcal{Y}, \mathfrak{B}(\mathcal{X} \times \mathcal{Y})\big)$. Define

$$\mathbf{D}_n \; := \; \big((X_1, Y_1), \ldots, (X_n, Y_n)\big) \qquad \forall \, n \in \mathbb{N} \,.$$

A measurable map $L : \mathcal{X} \times \mathcal{Y} \times \mathbb{R} \to [0, \infty)$ is called *loss function*. A loss function L is called *convex* loss function if it is convex in its third argument, i.e. $t \mapsto L(x, y, t)$ is convex for every $(x, y) \in \mathcal{X} \times \mathcal{Y}$. Furtheremore, a loss

function L is called P-integrable Nemitski loss function of order $p \in [1, \infty)$ if there is a P-integrable function $b : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ such that

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

If b is even P-*square*-integrable, L is called P-*square*-integrable Nemitski loss function of order $p \in [1, \infty)$. The *risk* of a measurable function $f : \mathcal{X} \to \mathbb{R}$ is defined by

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

The goal is to estimate a function $f : \mathcal{X} \to \mathbb{R}$ which minimizes this risk. The estimates obtained from the method of support vector machines are elements of so-called reproducing kernel Hilbert spaces (RKHS) H. A RKHS H is a certain Hilbert space of functions $f : \mathcal{X} \to \mathbb{R}$ which is generated by a *kernel* $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. See e.g. Schölkopf and Smola (2002) or Steinwart and Christmann (2008) for details about these concepts.
Let H be such a RKHS. Then, the *regularized risk* of an element $f \in H$ is defined to be

$$\mathcal{R}_{L,P,\lambda}(f) \;=\; \mathcal{R}_{L,P}(f) + \lambda \|f\|_H^2\,, \qquad \text{where} \quad \lambda \in (0, \infty)\,.$$

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

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

The *SVM-estimator* is defined by

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

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

$$\frac{1}{n} \sum_{i=1}^{n} L\big(x_i, y_i, f(x_i)\big) + \lambda_{D_n} \|f\|_H^2 \tag{4}$$

in H for $D_n = ((x_1, x_2), \ldots, (x_n, y_n)) \in (\mathcal{X} \times \mathcal{Y})^n$. The empirical support vector machine $f_{L,D_n,\lambda_{D_n}}$ uniquely exists for every $\lambda_{D_n} \in (0, \infty)$ and every data-set $D_n \in (\mathcal{X} \times \mathcal{Y})^n$ if $t \mapsto L(x, y, t)$ is convex for every $(x, y) \in \mathcal{X} \times \mathcal{Y}$.
The symbol $\rightsquigarrow$ denotes weak convergence of probability measures or random variables.

3 Asymptotic Normality

3.1 Main Results

The following theorems provide the main results. For random sequences of regularization parameters $(\lambda_{\mathbf{D}_n})_{n \in \mathbb{N}} \subset (0, \infty)$ which converges in probability with rate $\sqrt{n}$ to some $\lambda_0 \in (0, \infty)$, Theorem 3.1 says that the $\sqrt{n}$-standardized difference between the empirical support vector machine $f_{L,\mathbf{D}_n,\lambda_{\mathbf{D}_n}}$ and the theoretical support vector machine f_{L,P,λ_0} is asymptotically normal under some relatively mild conditions. That is, the H-valued random variable

$$\Omega \;\to\; H, \qquad \omega \;\to\; \sqrt{n}(f_{L,\mathbf{D}_n(\omega),\lambda_{\mathbf{D}_n(\omega)}} - f_{L,P,\lambda_0})$$

converges weakly to a random variable

$$\mathbb{H} \;:\; \Omega \;\to\; H, \qquad \omega \;\mapsto\; \mathbb{H}(\omega)$$

which is a Gaussian process in H. Accordingly, for every finite collection of functions $\{f_1, \ldots, f_m\} \subset H$, the random variable

$$\Omega \;\to\; \mathbb{R}^m, \qquad \omega \;\mapsto\; \left(\langle f_1, \mathbb{H}(\omega) \rangle_H, \ldots, \langle f_m, \mathbb{H}(\omega) \rangle_H \right)$$

has a multivariate normal distribution. In particular, the reproducing property of k implies that, for every $x_1, \ldots, x_m \in \mathcal{X}$,

$$\sqrt{n} \begin{pmatrix} f_{L,\mathbf{D}_n,\lambda_{\mathbf{D}_n}}(x_1) - f_{L,P,\lambda_0}(x_1) \\ \vdots \\ f_{L,\mathbf{D}_n,\lambda_{\mathbf{D}_n}}(x_m) - f_{L,P,\lambda_0}(x_m) \end{pmatrix} \;\rightsquigarrow\; \mathcal{N}_m(0, \Sigma)$$

where Σ is a covariance matrix. In addition, Theorem 3.2 provides $\sqrt{n}$-consistency of the risk.

Theorem 3.1 *Let $\mathcal{X} \subset \mathbb{R}^d$ be closed and bounded and let $\mathcal{Y} \subset \mathbb{R}$ be closed. Assume that $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is the restriction of an m - times continuously differentiable kernel $\tilde{k} : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ such that $m > d/2$ and $k \neq 0$. Let H be the RKHS of k and let P be a probability measure on $(\mathcal{X} \times \mathcal{Y}, \mathfrak{B}(\mathcal{X} \times \mathcal{Y}))$. Let*

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

be a convex, P-square-integrable Nemitski loss function of order $p \in [1, \infty)$ such that the partial derivatives

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

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

$$(x, y, t) \mapsto L'(x, y, t) \qquad and \qquad (x, y, t) \mapsto L''(x, y, t)$$

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

$$\sup_{t \in [-a, a]} \left| L'(x, y, t) \right| \le b_a'(x, y) \quad and \quad \sup_{t \in [-a, a]} \left| L''(x, y, t) \right| \le b_a'' . \qquad (5)$$

Then, for every $\lambda_0 \in (0, \infty)$, *there is a tight, Borel-measurable Gaussian process*

$$\mathbb{H} : \Omega \to H, \quad \omega \mapsto \mathbb{H}(\omega)$$

such that,

$$\sqrt{n}\left(f_{L, \mathbf{D}_n, \lambda_{\mathbf{D}_n}} - f_{L, P, \lambda_0}\right) \rightsquigarrow \mathbb{H} \qquad in \ H$$

for every Borel-measurable sequence of random regularization parameters $\lambda_{\mathbf{D}_n}$ *with*

$$\sqrt{n}\left(\lambda_{\mathbf{D}_n} - \lambda_0\right) \xrightarrow[n \to \infty]{} 0 \qquad in \ probability .$$

The Gaussian process $\mathbb{H}$ *is zero-mean; i.e.,* $\mathbb{E}\langle f, \mathbb{H}\rangle_H = 0$ *for every* $f \in H$.

Theorem 3.2 *Under the assumptions of Theorem 3.1, there is, for every* $\lambda_0 \in (0, \infty)$, *a* $\sigma \in [0, \infty)$ *such that*

$$\sqrt{n}\left(\mathcal{R}_{L, P}(f_{L, \mathbf{D}_n, \lambda_{\mathbf{D}_n}}) - \mathcal{R}_{L, P}(f_{L, P, \lambda_0})\right) \rightsquigarrow \mathcal{N}(0, \sigma^2)$$

for every Borel-measurable sequence of random regularization parameters $\lambda_{\mathbf{D}_n}$ *with* $\sqrt{n}\left(\lambda_{\mathbf{D}_n} - \lambda_0\right) \xrightarrow[n \to \infty]{} 0$ *in probability.*

According to the above theorems, the Gaussian process $\mathbb{H}$ and the standard deviation σ do not depend on the sequence $\lambda_{\mathbf{D}_n}$, $n \in \mathbb{N}$, but only on λ_0. Note that is possible that $\mathbb{H}$ degenerates to 0 and the standard deviation σ is equal to 0, i.e., $\mathcal{N}(0, \sigma^2) = \delta_0$. This is due to the generality of the result and happens in degenerated cases, e.g., we have not excluded $L \equiv 0$ or P equal to a Dirac distribution, or $Y \le \varepsilon$ while using a smoothed version of the epsilon-insensitive loss.

As stated above, the results are true under some relatively mild assumptions. In particular, the assumptions on k are fulfilled for all of the most common kernels (e.g. Gaussian RBF kernel, polynomial kernel, exponential kernel, linear kernel). It is assumed that the loss function is two times continuously differentiable in the third argument. On the one hand, this is an

7

obvious restriction because some of the most common loss functions are not differentiable: the epsilon-insensitive loss for regression and the hinge loss for classification. On the other hand, this assumption is not based on any unknown entity such as the model distribution P. In particular, a practitioner can a priori meet this requirement by a suitable choice of the loss function; e.g. the least-squares loss for regression and the logistic loss for classification. This is contrary to the noise assumptions common in order to establish rates of convergence to the Bayes risk because such assumptions depend on the unknown P so that they can hardly be checked in applications. In addition, Remark 3.5 describes how a Lipschitz-continuous loss function (such as the epsilon-insensitive loss and the hinge loss) can always be turned into a differentiable ε-version of the loss function.

In order to ensure mere existence of the theoretical SVM f_{L,P,λ_0}, it is necessary to assume a P-integrabilty condition. For example, it is common to assume that L is a P-integrable Nemitski loss function (Christmann and Steinwart, 2007). In order to obtain asymptotic normality in the above theorems, we assume that L is a P-*square*-integrable Nemitski loss function which seems to be a natural assumption in view of the square-integrability assumptions for usual central limit theorems. In addition, a similar P-integrabilty condition is assumed for the derivative of the loss function. If $\mathcal{Y}$ is bounded (as, e.g., in case of a classification problem) and L, L' and L'' are continuous, all of the integrability assumptions are fulfilled.

In order to fulfill

$$\sqrt{n}\big(\lambda_{\mathbf{D}_n} - \lambda_0\big) \xrightarrow[n\to\infty]{} 0 \qquad \text{in probability,}$$

(which is the only assumption on the random sequence of regularization parameters), it is possible to use any data-driven method for choosing the regularization parameter. The only thing one has to do is to choose a (possibly large) constant $c \in (0, \infty)$ and to make sure that the method (e.g. cross validation) picks a value from $[\lambda_0, c/\sqrt{n\ln(n)}]$. Note that, as the notation suggests, it is possible to use the same data for choosing the regularization parameter as for building the final SVM - just as usually done by practitioneers, e.g., when applying cross validation.

The following examples list some general situations in which Theorems 3.1 and 3.2 are applicable.

Example 3.3 (Classification) *Theorems 3.1 and 3.2 are applicable in the following setting for a classification problem:*

- *$\mathcal{X}$ bounded and closed, $\mathcal{Y} = \{-1; 1\}$*

- *k a Gaussian RBF kernel, a polynomial kernel, an exponential kernel or a linear kernel*

- *L the least-squares loss or the logistic loss*

Example 3.4 (Regression) *Theorems 3.1 and 3.2 are applicable in the following setting for a regression problem:*

- $\mathcal{X}$ *bounded and closed, $\mathcal{Y}$ closed*

- *k a Gaussian RBF kernel, a polynomial kernel, an exponential kernel or a linear kernel*

- *L the least-squares loss*

- *P such that $\int y^4 \, P\big(d(x,y)\big) < \infty$*

The following Remark 3.5 describes how a Lipschitz-continuous loss function can always be turned into a differentiable ε-version of the loss function such that all of the assumptions on the partial derivatives L' and L'' are automatically fulfilled. In particular, the proposed construction works for the epsilon-insensitive loss and the hinge loss.

Remark 3.5 *Let $L : \mathcal{X} \times \mathcal{Y} \times \mathbb{R} \to [0,\infty)$ be a convex P-square-integrable Nemitski loss function of order $p \in [1,\infty)$. Assume that L is also a Lipschitz-continuous loss function. That is, there is a constant $b' \in (0,\infty)$ such that*

$$\sup_{(x,y)\in\mathcal{X}\times\mathcal{Y}} \big|L(x,y,t_1) - L(x,y,t_2)\big| \; \leq \; b'|t_1 - t_2| \qquad \forall\, t_1, t_2 \in \mathbb{R}\,.$$

Then, for every $\varepsilon > 0$, it is possible to construct a loss function L_ε such that

$$\big|L(x,y,t) - L_\varepsilon(x,y,t)\big| \; \leq \; \varepsilon \qquad \forall\,(x,y,t) \in \mathcal{X} \times \mathcal{Y} \times \mathbb{R} \qquad (6)$$

and all of the assumptions of Theorems 3.1 and 3.2 are fulfilled for L_ε.
This can be done in the following way: Take a so-called mollifier function $\varphi : \mathbb{R} \to \mathbb{R}$; e.g.,

$$\varphi \;:\; \mathbb{R} \;\to\; \mathbb{R}\,, \qquad t \;\mapsto\; \gamma^{-1} e^{-\frac{1}{1-t^2}} I_{(-1,1)}(t)$$

where $\gamma \in (0,\infty)$ is chosen so that $\int \varphi \, d\lambda = 1$. (See e.g. Denkowski et al. (2003, p. 341ff) for the concept of mollifiers and their basic properties.) Define $\varphi_\varepsilon(s) = \varphi(sb'/\varepsilon)$ for every $s \in \mathbb{R}$ and

$$L_\varepsilon(x,y,t) \;=\; \frac{b'}{\varepsilon} \int \varphi_\varepsilon(s) L(x,y,t-s) \, \lambda(ds) \qquad \forall\,(x,y,t) \qquad (7)$$

9

*Then, (6) follows from an easy calculation using Lipschitz-continuity of L.
The ε-version L_ε is again a convex P-square-integrable Nemitski loss function of order $p \in [1, \infty)$. For every $(x, y, t) \in \mathcal{X} \times \mathcal{Y} \times \mathbb{R}$, the function $t \mapsto L_\varepsilon(x, y, t)$ is infinitely differentiable and the derivatives are given by*

$$\frac{\partial^m}{\partial^m t} L_\varepsilon(x, y, t) \;=\; \frac{b'}{\varepsilon} \int \frac{\partial^m \varphi_\varepsilon}{\partial^m s}(s) L(x, y, t - s) \, \lambda(ds) \, . \tag{8}$$

Furthermore, for every $(x, y, t) \in \mathcal{X} \times \mathcal{Y} \times \mathbb{R}$,

$$\left| L_\varepsilon'(x, y, t) \right| \;=\; \left| \frac{\partial}{\partial t} L_\varepsilon(x, y, t) \right| \;\leq\; b' \, , \tag{9}$$

$$\left| L_\varepsilon''(x, y, t) \right| \;=\; \left| \frac{\partial^2}{\partial^2 t} L_\varepsilon(x, y, t) \right| \;\leq\; b' \cdot \frac{b'}{\varepsilon} \int \frac{\partial \varphi_\varepsilon}{\partial s}(s) \lambda(ds) \;=:\; b'' \, . \tag{10}$$

Inequality (9) follows from the definition of derivatives by means of difference quotients, (7), and Lipschitz-continuity of L. Inequality (10) follows from the definition of derivatives by means of difference quotients, (8) for $m = 1$, and Lipschitz-continuity of L.

In particular, the construction of such an ε-version of L works for the hinge loss (classification) and, if $\int y^2 \, P(d(x, y)) < \infty$, for the epsilon-insensitive loss (regression). Another approach in order to obtain smooth approximations of loss functions is proposed in Dekel et al. (2005).

3.2 Supplements and Sketch of the Proof

The proof of Theorems 3.1 and 3.2 is an involved application of the functional delta-method. In oder to describe this in some more detail, let us first fix a constant sequence of regularization parameters. That is, $\lambda_{\mathbf{D}_n} \equiv \lambda_0 \in (0, \infty)$ for every $n \in \mathbb{N}$. Then, support vector machines may be represented by a functional S on a set of probability measures on $\left(\mathcal{X} \times \mathcal{Y}, \mathfrak{B}(\mathcal{X} \times \mathcal{Y}) \right)$. This functional

$$S \; : \; P \; \mapsto \; f_{L, P, \lambda_0}$$

is called *SVM-functional* in the following. It represents the SVM-estimator because the empirical support vector machine is equal to $f_{L, D_n, \lambda_0} = S(\mathbb{P}_{D_n})$ for every data set $D_n \in (\mathcal{X} \times \mathcal{Y})^n$ where $\mathbb{P}_{D_n}$ denotes the empirical measure corresponding to D_n. In order to use the functional delta-method, it is crucial that this is true for every sample size n and that S does not depend on n. (In Remark 3.6, it will be explained how it is nevertheless possible to deal with random sequences $\lambda_{\mathbf{D}_n}$.) Theorem 3.1 can be shown in the following way:

1. Show that $\sqrt{n}(\mathbb{P}_{\mathbf{D}_n} - P)$ converges weakly to a Gaussian process.
2. Show that S is Hadamard-differentiable:
(a) Show that S is Gâteaux-differentiable.
(b) Show that the Gâteaux-derivative fulfills a continuity property.
(c) Show that (a) and (b) imply Hadamard-differentiability.
3. Then, it follows from the functional delta-method that

$$\sqrt{n}(f_{L,\mathbf{D}_n,\lambda_0} - f_{L,P,\lambda_0}) \; = \; \sqrt{n}\big(S(\mathbb{P}_{\mathbf{D}_n}) - S(P)\big)$$

converges weakly to a Gaussian process. Theorem 3.2 follows from Theorem 3.1 by another application of the functional delta-method.

Step 1 involves the study of Donsker classes. Among other things, this is based on a bound (57) on the uniform entropy number of balls in the reproducing kernel Hilbert space H. A proof of this bound is given in the proof of Lemma 5.9. In similar settings, such bounds have already been proven, e.g. in Zhou (2003, §V) and Steinwart and Christmann (2008, §6.4). In general, $\sqrt{n}(\mathbb{P}_{\mathbf{D}_n} - P)$ is *not* a measurable random variable so that the proof involves the theory of weak convergence of unmeasurable random variables; see van der Vaart and Wellner (1996). However, this does not affect the statements of Theorems 3.1 and 3.2 because $\omega \to f_{L,\mathbf{D}_n(\omega),\lambda_{\mathbf{D}_n(\omega)}}$ is a measurable random variable as shown in the beginning of the proof of Theorem 3.1 in Subsection 5.4.

Essentially, it has already been known that S is Gâteaux-differentiable because Christmann and Steinwart (2004, 2007) derive the influence function of S which is a (special) Gâteaux-derivative. Therefore, essential steps of the proof of Step 2(a) can be adopted from Christmann and Steinwart (2004, 2007) and Steinwart and Christmann (2008, §10.4) but some care is needed as we also have to deal with signed measures here. In addition, we also have to deal with a sequence of random regularization parameters $\lambda_{\mathbf{D}_n}$ instead of a fixed λ_0; see Remark 3.6. In Step 2(c) it will be shown that S is even Hadamard-differentiable (in a specific sense described in Subsection 5.3). This is done because the application of the delta-method requires Hadamard-differentiability. However, this might also be useful for other purposes since, e.g., the chain rule is valid for Hadamard-differentiability but not for Gâteaux-differentiability. Christmann and Van Messem (2008) show Bouligand-differentiability of the SVM-functional which also allows the chain rule.

Remark 3.6 (Sequences of random regularization parameters $\lambda_{\mathbf{D}_n}$)
For a fixed regularization parameter λ_0, support vector machines can be represented by a functional $S : P \mapsto f_{L,P,\lambda_0}$ and the delta-method can be applied

for S. However, if we have a sequence of (random) regularization parameters $\lambda_{\mathbf{D}_n}$, we get a (random) sequence of functionals

$$S_{\mathbf{D}_n} \; : \; P \; \mapsto \; f_{L,P,\lambda_{\mathbf{D}_n}}$$

for which the delta-method cannot be applied offhand. This problem can be solved in the following way: As described in Subsection 5.1,

$$S_{\mathbf{D}_n}(P) \;=\; f_{L,P,\lambda_{\mathbf{D}_n}} \;=\; f_{L,\frac{\lambda_0}{\lambda_{\mathbf{D}_n}}P,\lambda_0} \;=\; S\big(\tfrac{\lambda_0}{\lambda_{\mathbf{D}_n}}P\big) \qquad \forall\, P \,.$$

so that everything can be traced back to S. In this way, the explicit use of $S_{\mathbf{D}_n}$ can be avoided and the delta-method turns out to be applicable also in this case. The price we have to pay is that we have to deal with general finite measures in the proofs because, in general, $\frac{\lambda_0}{\lambda_{\mathbf{D}_{n(\omega)}}}P$ is not a probability measure any more.

4 Conclusions

In the article, asymptotic properties of support vector machines are investigated. For sequences of random regularization parameters $\lambda_{\mathbf{D}_n}$, $n \in \mathbb{N}$, such that $\sqrt{n}\big(\lambda_{\mathbf{D}_n} - \lambda_0\big) \longrightarrow 0$ in probability, it is shown that the difference between the empirical and the theoretical SVM is asymptotically normal with rate $\sqrt{n}$; that is, $\sqrt{n}(f_{L,\mathbf{D}_n,\lambda_{\mathbf{D}_n}} - f_{L,P,\lambda_0})$ converges to a Gaussian process in the function space H. The value $\lambda_0 > 0$ corresponds to a bound on the complexity of the estimate for the regression function; a smaller λ_0 allows for more complex functions. Therefore, the theoretical SVM f_{L,P,λ_0} serves as a "smoother" approximation of more complex regression functions. The results of this article show that, in nonparametric classification and regression problems, the estimation of this smoother approximation by use of empirical SVMs in an infinite dimensional function space is asymptotically normal with rate $\sqrt{n}$ – just as if it was a parametric problem. The proof is done by showing that the map $P \mapsto f_{L,P,\lambda}$ is suitably Hadamard-differentiable and by an application of a functional delta-method.
Estimating a smoother approximation of the regression function is a comprise between a parametric model and a fully non-parametric model without any assumptions on the regression function or the distribution. Without any of such assumptions, similar results are not possible as follows from the no-free-lunch theorem.

Acknowledgment

5 Appendix: Proof of the Main Results

The assumptions of Theorem 3.1 are valid in the whole appendix.

5.1 Preparations

The map $\Phi : \mathcal{X} \to H$ always denotes the canonical feature map corresponding to the kernel k and the RKHS H. It will frequently be used in the proofs that the reproducing property implies

$$\langle \Phi(x), f \rangle_H \;=\; f(x) \qquad \forall\, x \in \mathcal{X}, \quad \forall\, f \in H \tag{11}$$

or, in shorter notation,

$$\langle \Phi, f \rangle_H \;=\; f \qquad \forall\, f \in H \,. \tag{12}$$

In particular, we have

$$\mathbb{E}_\mu \langle \Phi, f \rangle_H = \int \langle \Phi, f \rangle_H \, d\mu = \int \langle \Phi(x), f \rangle_H \, \mu(dx) = \int f(x)\, \mu(dx). \tag{13}$$

According to Steinwart and Christmann (2008, p. 124), boundedness of k implies:

$$\|k\|_\infty \;:=\; \sup_{x \in \mathcal{X}} \sqrt{k(x,x)} \;=\; \sup_{x \in \mathcal{X}} \big\| \Phi(x) \big\|_H \;<\; \infty \tag{14}$$

$$\|f\|_\infty \;\leq\; \|k\|_\infty \cdot \|f\|_H \qquad \forall\, f \in H \,. \tag{15}$$

In order to shorten notation, define

$$L_f \;:\; \mathcal{X} \times \mathcal{Y} \to \mathbb{R}, \qquad (x,y) \mapsto L_f(x,y) \;=\; L\big(x,y,f(x)\big)$$

for every function $f : \mathcal{X} \to \mathbb{R}$. Accordingly, define

$$L_f'(x,y) = L'\big(x,y,f(x)\big) \qquad \text{and} \qquad L_f''(x,y) = L''\big(x,y,f(x)\big)$$

for every $(x, y) \in \mathcal{X} \times \mathcal{Y}$. As L is a P-square-integrable Nemitski loss function of order $p \in [1, \infty)$, there is a $b \in L_2(P)$ such that

$$\left| L(x, y, t) \right| \leq b(x, y) + |t|^p \qquad \forall (x, y, t) \in \mathcal{X} \times \mathcal{Y} \times \mathbb{R} . \qquad (16)$$

Let

$$\mathcal{G}_1 \; := \; \left\{ g : \mathcal{X} \times \mathcal{Y} \to \mathbb{R} \; \middle| \; \exists z \in \mathbb{R}^{d+1} \text{ such that } g = I_{(-\infty, z]} \right\}$$

be the set of all indicator functions $I_{(-\infty, z]}$. Then, it is well-known that

$$\sqrt{n}\left(\mathbb{F}_n - F \right) \; \rightsquigarrow \; \mathbb{G}_1 \qquad \text{in} \quad \ell_\infty(\mathcal{G}_1)$$

where $\mathbb{F}_n$ denotes the empirical process, F denotes the distribution function of P, $\mathbb{G}_1$ is a Gaussian process, and $\ell_\infty(\mathcal{G}_1)$ denotes the set of all bounded functions $G : \mathcal{G}_1 \to \mathbb{R}$. Provided that the SVM-functional S is Hadamard-differentiable in $\ell_\infty(\mathcal{G}_1)$, an application of the functional delta-method would yield asymptotic normality of $\sqrt{n}\left(S(\mathbb{F}_n) - S(F) \right)$. Unfortunately, the norm-topology of $\ell_\infty(\mathcal{G}_1)$ is too weak in order to ensure Hadamard-differentiability. Therefore, the set of indicator functions $\mathcal{G}_1$ has to be enlarged to a set $\mathcal{G} \supset \mathcal{G}_1$ which leads to the following somewhat technical definition of the domain B_S of the SVM-functional S. Define

$$c_0 \; := \; \sqrt{\frac{1}{\lambda_0} \int b \, dP} \; + \; 1 , \qquad (17)$$

$$\mathcal{G}_2 \; := \; \left\{ g : \mathcal{X} \times \mathcal{Y} \to \mathbb{R} \; \middle| \; \begin{array}{c} \exists f_0 \in H , \; \exists f \in H \text{ such that} \\ \|f_0\|_H \leq c_0 , \; \|f\|_H \leq 1 \text{ and} \\ g = L'_{f_0} f \end{array} \right\} ,$$

and

$$\mathcal{G} \; := \; \mathcal{G}_1 \cup \mathcal{G}_2 \cup \{b\} .$$

Let $\ell_\infty(\mathcal{G})$ be the set of all bounded functions

$$F : \; \mathcal{G} \; \to \; \mathbb{R}$$

with norm $\|F\|_\infty = \sup_{g \in \mathcal{G}} \left| F(g) \right|$. Define

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

14

and $B_0 := \mathrm{cl}\big(\mathrm{lin}(B_S)\big)$ 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. The elements of B_S can be seen as some kind of generalized distribution functions. Note that the assumptions on L and P imply that $\mathcal{G} \to \mathbb{R}$, $g \mapsto \int g \, dP$ is a well-defined element of B_S.

For every $F \in B_S$, let $\iota(F)$ denote the corresponding finite measure μ on $\big(\mathcal{X} \times \mathcal{Y}, \mathfrak{B}(\mathcal{X} \times \mathcal{Y})\big)$ such that

$$F(g) \;=\; \int g \, d\mu \qquad \forall \, g \in \mathcal{G} \; .$$

Note that, by definition of B_S, $\iota(F)$ uniquely exists for every $F \in B_S$ so that

$$\iota \;:\; B_S \;\to\; \mathrm{ca}^+(\mathcal{X} \times \mathcal{Y}, \mathfrak{B}(\mathcal{X} \times \mathcal{Y})), \qquad F \;\mapsto\; \iota(F) \;.$$

is well-defined where $\mathrm{ca}^+(\mathcal{X} \times \mathcal{Y}, \mathfrak{B}(\mathcal{X} \times \mathcal{Y}))$ denotes the set of all finite measures on $(\mathcal{X} \times \mathcal{Y}, \mathfrak{B}(\mathcal{X} \times \mathcal{Y}))$. The set of all finite *signed* measures on $(\mathcal{X} \times \mathcal{Y}, \mathfrak{B}(\mathcal{X} \times \mathcal{Y}))$ is denoted by $\mathrm{ca}(\mathcal{X} \times \mathcal{Y}, \mathfrak{B}(\mathcal{X} \times \mathcal{Y}))$. The set of all continuous functions $f : \mathcal{X} \to \mathbb{R}$ is denoted by $\mathcal{C}(\mathcal{X})$. Since $\mathcal{X}$ is compact by assumption, the elements of $\mathcal{C}(\mathcal{X})$ are bounded and $\mathcal{C}(\mathcal{X})$ is endowed with the sup-norm $\|f\|_\infty = \sup_{x \in \mathcal{X}} |f(x)|$.

By now, support vector machines are only defined for probability measures $\tilde{P}$. However, in order to deal with sequences of random regularization parameters $\lambda_{\mathbf{D}_n}$, we will also have to deal with "support vector machines" for general finite measures μ. For every $F \in B_S$, define

$$f_{L, \iota(F), \lambda} \;:=\; \arg \inf_{f \in H} \int L\big(x, y, f(x)\big) \, \iota(F)\big(d(x,y)\big) \;+\; \lambda \|f\|_H^2 \;.$$

Though $\mu := \iota(F) \in \mathrm{ca}^+(\mathcal{X} \times \mathcal{Y}, \mathfrak{B}(\mathcal{X} \times \mathcal{Y}))$ is not necessarily a probability measure, we have, in effect, not defined any new object. In order to see this, note that dividing the objective function by $M := \mu(\mathcal{X} \times \mathcal{Y})$ does not change the minimizer so that we get

$$f_{L, \mu, \lambda} \;=\; \arg \inf_{f \in H} \int L\big(x, y, f(x)\big) \frac{1}{M} \mu\big(d(x,y)\big) \;+\; \frac{\lambda}{M} \|f\|_H^2 \;=\; f_{L, \frac{1}{M}\mu, \frac{\lambda}{M}}$$

and $f_{L, \frac{1}{M}\mu, \frac{\lambda}{M}}$ is an "ordinary" support vector machine as $\frac{1}{M}\mu$ is a probability measure. This also shows that $f_{L, \mu, \lambda}$ uniquely exists because $f_{L, \frac{1}{M}\mu, \frac{\lambda}{M}}$ uniquely exists for the probability measure $\frac{1}{M}\mu$ according to Steinwart and Christmann (2008, Lemma 5.1 and Theorem 5.2).

The idea is that considering support vector machines for general finite measures μ makes it possible to take λ_0 as a "standard regularization parameter". Define

$$S \;:\; B_S \;\to\; H\,, \qquad F \;\mapsto\; S(F) \;=\; f_{\iota(F)}$$

where

$$f_{\iota(F)} \;:=\; f_{L,\iota(F),\lambda_0} \;=\; \arg\inf_{f \in H} \int L\big(x,y,f(x)\big)\,\iota(F)\big(d(x,y)\big) \;+\; \lambda_0\|f\|_H^2\,.$$

Then, we can deal with other regularization parameters $\lambda > 0$ by use of

$$f_{L,\iota(F),\lambda} \;=\; S\big(\tfrac{\lambda_0}{\lambda}F\big) \qquad \forall\, F \in B_S\,. \tag{18}$$

This is important in order to apply the functional delta-method in case of a sequence of random regularization parameters $\lambda_{\mathbf{D}_n}$; see also Remark 3.6.

It follows from Steinwart and Christmann (2008, Eqn. (5.4) and Lemma 4.23) that

$$\big\|f_{\iota(F)}\big\|_H \;\leq\; \sqrt{\tfrac{1}{\lambda_0}F(b)} \qquad \forall\, F \in B_S\,, \tag{19}$$

$$\big\|f_{\iota(F)}\big\|_\infty \;\leq\; \|k\|_\infty \sqrt{\tfrac{1}{\lambda_0}F(b)} \qquad \forall\, F \in B_S\,. \tag{20}$$

Since $\mathcal{X}$ is separable and k is a continuous kernel, the RKHS H is a *separable* Hilbert space; see Steinwart and Christmann (2008, Lemma 4.33). Separability of H is used several times in the proofs; this is important particularly with regard to the Bochner-integral of H-valued functions $\Psi : \mathcal{Z} \to H$. The Bochner-integral $\int \Psi\,d\mu = \int \Psi\,d\mu^+ - \int \Psi\,d\mu^-$ of such a H-valued function Ψ with respect to a finite signed measure $\mu = \mu^+ - \mu^-$ is again an element of H. If Ψ is suitably measurable, then existence of the Bochner-integral follows from $\int \|\Psi\|_H\,d|\mu| < \infty$ where $|\mu| = \mu^+ + \mu^-$ denotes the total variation of μ. We will also frequently use the fact that, for every Banach space E and every continuous linear operator $A : H \to E$, the existence of the Bochner-integral $\int \Psi\,d\mu$ implies the existence of the Bochner-integral $\int A(\Psi)\,d\mu$ and

$$\int A(\Psi)\,d\mu \;=\; A\left(\int \Psi\,d\mu\right)\,; \tag{21}$$

see, e.g. Denkowski et al. (2003, Theorem 3.10.16 and Remark 3.10.17).

This subsection closes with three lemmas which are used several times. Thereafter, Gâteaux-differentiability of the SVM-functional $S \;:\; B_S \to H$ will be shown in Subsection 5.2. This is strengthened to Hadamard-differentiability in Subsection 5.3. Finally, it will be shown in Subsection 5.4 that

$\sqrt{n}(\mathbb{P}_{\mathbf{D}_n} - P)$ converges weakly to a Gaussian process in $\ell_\infty(\mathcal{G})$ and that this implies asymptotic normality of

$$\sqrt{n}\big(f_{L,\mathbf{D}_n,\lambda_{\mathbf{D}_n}} - f_{L,P,\lambda_0}\big) \quad \text{and} \quad \sqrt{n}\big(\mathcal{R}_{L,P}(f_{L,\mathbf{D}_n,\lambda_{\mathbf{D}_n}}) - \mathcal{R}_{L,P}(f_{L,P,\lambda_0})\big)$$

by applying a functional delta-method.

Lemma 5.1 *Let* $(F_n)_{n\in\mathbb{N}} \subset B_S$ *be a sequence which converges to some* $F_0 \in B_S$ *. Then,* $\lim_{n\to\infty} \iota(F_n)(\mathcal{X}\times\mathcal{Y}) = \iota(F_0)(\mathcal{X}\times\mathcal{Y})$ *and the sequence of finite measures* $\iota(F_n)$, $n \in \mathbb{N}$, *converges weakly to* $\iota(F_0)$ *.*

Proof: Define $M_n := \iota(F_n)(\mathcal{X}\times\mathcal{Y})$ and $a_n = (n,\ldots,n) \in \mathbb{R}^{d+1}$ for every $n \in \mathbb{N} \cup \{0\}$. Then,

$$0 \le \big|M_n - M_0\big| = \lim_{l\to\infty} \big|F_n\big(I_{(-\infty,a_l]}\big) - F_0\big(I_{(-\infty,a_l]}\big)\big| \le \big\|F_n - F_0\big\|_\infty \longrightarrow 0\,.$$

Therefore, the normalized sequence $\tilde{F}_n = M_n^{-1}F_n$, $n \in \mathbb{N}\cup\{0\}$, corresponds to a sequence of probability measures $\iota(\tilde{F}_n)$ such that

$$\begin{aligned}
\lim_{n\to\infty} \iota(\tilde{F}_n)\big((-\infty,a]\cap\mathcal{X}\times\mathcal{Y}\big) &= \lim_{n\to\infty} \frac{1}{M_n}F_n\big(I_{(-\infty,a]}\big) = \frac{1}{M_0}F_0\big(I_{(-\infty,a]}\big) \\
&= \iota(\tilde{F}_0)\big((-\infty,a]\cap\mathcal{X}\times\mathcal{Y}\big)
\end{aligned}$$

for every $a \in \mathbb{R}^{d+1}$. Hence, it follows from the Portmanteau theorem that the sequence of probability measures $(\iota(\tilde{F}_n))_{n\in\mathbb{N}}$ converges weakly to $\iota(\tilde{F}_0)$; see e.g. van der Vaart (1998, Lemma 2.2). Finally, this implies that the sequence of finite measures $(\iota(F_n))_{n\in\mathbb{N}}$ converges weakly to $\iota(F_0)$. $\qquad\square$

Lemma 5.2 *For every* $G \in \mathrm{lin}(B_S)$ *, there is a unique finite signed measure* $\iota(G) = \mu$ *on* $(\mathcal{X}\times\mathcal{Y}, \mathfrak{B}(\mathcal{X}\times\mathcal{Y}))$ *such that*

$$\int g\,d\mu = G(g) \qquad \forall\, g \in \mathcal{G}\,. \tag{22}$$

The map

$$\iota\,:\quad \mathrm{lin}(B_S) \to \mathrm{ca}(\mathcal{X}\times\mathcal{Y}, \mathfrak{B}(\mathcal{X}\times\mathcal{Y})), \qquad G \mapsto \iota(G)\,.$$

defined by (22) is linear. Let $G \in \mathrm{lin}(B_S)$ *and* $\mu = \iota(G)$ *. Then,*

$$b \in L_2\big(|\mu|\big)\,, \qquad b_a' \in L_2\big(|\mu|\big) \quad \forall\, a \in (0,\infty)$$

17

and $L'_f\Phi$ and $L''_f h\Phi$ are Bochner-integrable with respect to μ for every $f, h \in H$. Furthermore,

$$\tilde{A}_f \;:\; C(\mathcal{X}) \;\to\; H\,, \qquad h \;\mapsto\; \int L''_f h\Phi \, d\mu\,,$$

$$A_f \;:\; H \;\to\; H\,, \qquad h \;\mapsto\; \int L''_f h\Phi \, d\mu\,.$$

are continuous linear operators for every $f \in H$.

Proof: For every $G \in \operatorname{lin}(B_S)$, there are $F_1, F_2 \in B_S$ such that $G = F_1 - F_2$. Define $\mu := \iota(F_1) - \iota(F_2)$. Then, μ fulfills (22). From the definition of B_S and

$$|\mu|(C) \;\le\; \iota(F_1)(C) + \iota(F_2)(C) \qquad \forall C \in \mathfrak{B}(\mathcal{X} \times \mathcal{Y})$$

it follows that $b, b'_a \in L_2(|\mu|)$ for every $a \in (0, \infty)$. Next, fix any $f \in H$ and define $a = \|f\|_\infty < \infty$; see (15). Then,

$$\int \big\|L'_f\Phi\big\|_H \, d|\mu| \;\overset{(14)}{\le}\; \|k\|_\infty \cdot \int |L'_f| \, d|\mu| \;\overset{(5)}{\le}\; \|k\|_\infty \cdot \int b'_a \, d|\mu| \;<\; \infty$$

and, therefore, $L'_f\Phi$ is Bochner-integrable; see e.g. Denkowski et al. (2003, Theorem 3.10.3 and Theorem 3.10.9). A similar calculation shows that $L''_f h\Phi$ is Bochner-integrable, too.

In order to prove uniqueness of μ, let μ_1 and μ_2 be finite signed measures such that $\int g \, d\mu_1 = \int g \, d\mu_2$ for every $g \in \mathcal{G}$. From this equation it follows that $\int g \, d(\mu_1^+ + \mu_2^-) = \int g \, d(\mu_2^+ + \mu_1^-)$ for every $g \in \mathcal{G}$. Since $\mu_1^+ + \mu_2^-$ and $\mu_2^+ + \mu_1^-$ are finite (positive) measures and $\mathcal{G}$ contains all indicator functions $I_{(-\infty, z]}$, $z \in \mathbb{R}^{d+1}$, it follows from the uniqueness theorem (e.g. Hoffmann-Jørgensen (1994, §1.7)) that $\mu_1^+ + \mu_2^- = \mu_2^+ + \mu_1^-$. Hence, $\mu_1 = \mu_2$.

Uniqueness and (22) imply linearity of the map ι.

Now let us turn over to $\tilde{A}_f$ for any fixed $f \in H$. Obviously, $\tilde{A}_f$ is linear. In order to prove that $\tilde{A}_f$ is a continuous linear operator, define $a := \|f\|_\infty$, which is a finite number due to (15). Then,

$$\big\|\tilde{A}_f(h)\big\|_H \;\le\; \int \big\|L''_f h\Phi\big\|_H \, d|\mu| \;\overset{(14,5)}{\le}\; \|h\|_\infty \|k\|_\infty \int |b''_a| \; |\mu|\big(d(x,y)\big) \;<\; \infty\,.$$

According to Steinwart and Christmann (2008, Lemma 4.23), the canonical embedding $H \to C(\mathcal{X})$ is a continuous linear operator. Hence, it also follows that A_f is a continuous linear operator. $\qquad\square$

Lemma 5.3 *Let $(\mu_n)_{n\in\mathbb{N}}$ be a tight sequence of finite signed measures on $(\mathcal{X}\times\mathcal{Y},\mathfrak{B}(\mathcal{X}\times\mathcal{Y}))$ such that $\sup_{n\in\mathbb{N}}|\mu_n|(\mathcal{X}\times\mathcal{Y})<\infty$. Let $(f_n)_{n\in\mathbb{N}}\subset H$ be a sequence converging to some $f_0\in H$. Then,*

$$\lim_{n\to\infty}\sup_{\substack{h\in H\\ \|h\|_H\leq 1}}\left\|\int L''_{f_n}h\Phi\,d\mu_n-\int L''_{f_0}h\Phi\,d\mu_n\right\|_H=0.$$

Proof: For every $\varepsilon>0$ there is a compact subset $\mathcal{Z}_\varepsilon\subset\mathcal{X}\times\mathcal{Y}$ such that

$$|\mu_n|(\mathcal{X}\times\mathcal{Y}\setminus\mathcal{Z}_\varepsilon)<\varepsilon\qquad\forall\,n\in\mathbb{N}.\tag{23}$$

Define $a:=\sup_{n\in\mathbb{N}_0}\|f_n\|_\infty\overset{(15)}{\leq}\|k\|_\infty\sup_{n\in\mathbb{N}_0}\|f_n\|_H<\infty$. For every $n\in\mathbb{N}$,

$$\sup_{\substack{h\in H\\ \|h\|_H\leq 1}}\left\|\int L''_{f_n}h\Phi\,d\mu_n-\int L''_{f_0}h\Phi\,d\mu_n\right\|_H=\sup_{\substack{h\in H\\ \|h\|_H\leq 1}}\left\|\int(L''_{f_n}-L''_{f_0})h\Phi\,d\mu_n\right\|_H$$

$$\leq\sup_{\substack{h\in H\\ \|h\|_H\leq 1}}\int\left|L''_{f_n}(x,y)-L''_{f_0}(x,y)\right|\cdot\|h\|_\infty\cdot\left\|\Phi(x)\right\|_H|\mu_n|\big(d(x,y)\big)\leq$$

$$\overset{(14,15)}{\leq}\|k\|_\infty^2\int\left|L''_{f_n}(x,y)-L''_{f_0}(x,y)\right||\mu_n|\big(d(x,y)\big)\leq$$

$$\overset{(5,23)}{\leq}\|k\|_\infty^2\int_{\mathcal{Z}_\varepsilon}\left|L''_{f_n}(x,y)-L''_{f_0}(x,y)\right||\mu_n|\big(d(x,y)\big)+2\|k\|_\infty^2 b''_a\varepsilon\leq$$

$$\leq\|k\|_\infty^2|\mu_n|(\mathcal{X}\times\mathcal{Y})\sup_{(x,y)\in\mathcal{Z}_\varepsilon}\left|L''_{f_n}(x,y)-L''_{f_0}(x,y)\right|+2\|k\|_\infty^2 b''_a\varepsilon$$

Since $\sup_{n\in\mathbb{N}}|\mu_n|(\mathcal{X}\times\mathcal{Y})<\infty$ and $\varepsilon>0$ can be chosen arbitrarily small, it only remains to prove that

$$\lim_{n\to\infty}\sup_{(x,y)\in\mathcal{Z}_\varepsilon}\left|L''_{f_n}(x,y)-L''_{f_0}(x,y)\right|=0\tag{24}$$

Continuity of L'' and compactness of $\mathcal{Z}_\varepsilon\times[-a,a]$ imply that L'' is uniformly continuous on $\mathcal{Z}_\varepsilon\times[-a,a]$. Assertion (24) is an easy consequence of uniform continuity of L'' on $\mathcal{Z}_\varepsilon\times[-a,a]$, inequality $-a\leq f_n\leq a$ for every $n\in\mathbb{N}_0$, and the fact that $\lim_n\|f_n-f_0\|_H=0$ implies $\lim_n\|f_n-f_0\|_\infty=0$. $\qquad\square$

5.2 Gâteaux-Differentiability of the SVM-Functional

In this subsection, it will be shown that the SVM-functional

$$S:\ B_S\ \to\ H,\qquad F\ \mapsto\ f_{\iota(F)}$$

is Gâteaux-differentiable. Essentially, this has already been known because Christmann and Steinwart (2004, 2007) derive the influence function of S which is a (special) Gâteaux-derivative. Therefore, the proofs in this sub-section can essentially be adopted from Christmann and Steinwart (2004, 2007) and Steinwart and Christmann (2008, § 10.4). However, some care is needed as we also have to deal with signed measures and with a (random) sequence of regularization parameters $\lambda_{\mathbf{D}_n}$ instead of a fixed λ_0; see also Remark 3.6.

At first, we have to show Fréchet-differentiability of the "generalized risk" $\mathcal{R}_{L,\mu} : f \mapsto \int L_f \, d\mu$ (and of its derivative) for finite signed measures μ. If μ is a probability measure, then Lemma 5.4(a) is just the well-known Fréchet-differentiability of the ordinary risk $\mathcal{R}_{L,P}$.

Lemma 5.4 *For every finite signed measure μ on $(\mathcal{X} \times \mathcal{Y}, \mathfrak{B}(\mathcal{X} \times \mathcal{Y}))$ such that*

$$\int b \, d|\mu| \; < \; \infty \qquad and \qquad \int b'_a \, d|\mu| \; < \; \infty \quad \forall \, a \in (0, \infty), \qquad (25)$$

the following statements are true:

(a) *The map*

$$H \;\to\; \mathbb{R}, \qquad f \;\mapsto\; \int L_f \, d\mu$$

is Fréchet-differentiable and its Fréchet-derivative in $f \in H$ is given by $H \to \mathbb{R}, \; h \mapsto \big\langle \int L'_f \Phi \, d\mu \,, \, h \big\rangle_H$.

(b) *The map*

$$H \;\to\; H, \qquad f \;\mapsto\; \int L'_f \Phi \, d\mu$$

is Fréchet-differentiable and its Fréchet-derivative in $f \in H$ is given by $H \to H, \; h \mapsto \int L''_f h \Phi \, d\mu$.

Proof: Both statements can be proven essentially by following the lines of Steinwart and Christmann (2008, Lemma 2.21). Since the proofs of (a) and (b) nearly coincide, only the proof of (b) is given in detail.

Define

$$T(f) \;=\; \int L'_f \Phi \, d\mu \qquad and \qquad T'_f(h) \;=\; \int L''_f h \Phi \, d\mu$$

for every $f, h \in H$. Lemma 5.2 guarantees that these Bochner-integrals exist and that $T'_f : H \to H, \; h \mapsto T'_f(h)$ is a continuous linear operator. Now,

20

fix any $f \in H$ and let $(h_n)_{n \in \mathbb{N}} \subset H \setminus \{0\}$ be a sequence which converges to 0 in H. Define

$$\gamma_n(x, y) := \frac{\left| L'\big(x, y, f(x) + h_n(x)\big) - L'\big(x, y, f(x)\big) - h_n(x) L''\big(x, y, f(x)\big) \right|}{|h_n(x)|}$$

for every $(x, y) \in \mathcal{X} \times \mathcal{Y}$ such that $h_n(x) \neq 0$ and $\gamma_n(x, y) = 0$ for every $(x, y) \in \mathcal{X} \times \mathcal{Y}$ such that $h_n(x) = 0$. The maps $\gamma_n : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}, \ (x, y) \mapsto \gamma_n(x, y), \ n \in \mathbb{N}$, are measurable. Since H is a RKHS, $\lim_{n \to \infty} h_n(x) = 0$ for every $x \in \mathcal{X}$. Therefore, the definition of L' as a partial derivative of L implies

$$\lim_{n \to \infty} \gamma_n(x, y) = 0 \qquad \forall (x, y) \in \mathcal{X} \times \mathcal{Y}. \tag{26}$$

Define $a := \|f\|_\infty + \sup_{n \in \mathbb{N}} \|h_n\|_\infty \overset{(15)}{\leq} \|k\|_\infty \big(\|f\|_H + \sup_{n \in \mathbb{N}} \|h_n\|_H\big) < \infty$. Then, by use of the elementary mean value theorem,

$$\big|\gamma_n(x, y)\big| \leq \frac{\left| L'\big(x, y, f(x) + h_n(x)\big) - L'\big(x, y, f(x)\big) \right|}{|h_n(x)|} + \big| L''\big(x, y, f(x)\big) \big| \overset{(5)}{\leq} 2 b_a''$$

for every (x, y) such that $h_n(x) \neq 0$ and every $n \in \mathbb{N}$. Hence, we can use the dominated convergence theorem (e.g. Dudley (2002, Theorem 4.3.5)) in order to finish the proof:

$$\lim_{n \to \infty} \frac{\left\| T(f + h_n) - T(f) - T'_f(h_n) \right\|_H}{\|h_n\|_H} \leq$$

$$\leq \lim_{n \to \infty} \int \frac{|h_n(x)|}{\|h_n\|_H} \cdot \big|\gamma_n(x, y)\big| \cdot \big\|\Phi(x)\big\|_H \, |\mu|\big(d(x, y)\big) \leq$$

$$\overset{(14,15)}{\leq} \lim_{n \to \infty} \|k\|_\infty^2 \int \big|\gamma_n(x, y)\big| \, |\mu|\big(d(x, y)\big) \overset{(26)}{=} 0$$

$\square$

Lemma 5.5 *For every $F \in B_S$,*

$$K_F : \ H \ \to \ H, \qquad f \ \mapsto \ 2\lambda_0 f + \int L''_{f_{\iota(F)}} f \Phi \, d[\iota(F)]$$

is a continuous linear operator which is invertible.

Proof: It follows from Lemma 5.2 that K_F is a continuous linear operator and it only remains to prove that K_F is invertible. This is done by use of the Fredholm alternative (see e.g. Griffel (2002, Theorem 9.29)). The following proof is essentially a variant of the proof of Steinwart and Christmann (2008, Theorem 10.18). We have to show:

(i) K_F is injective.

(ii) $A := A_{f_\iota(F)}$ as defined in Lemma 5.2 is a compact operator.

Define $\mu = \iota(F)$. In order to prove (i), fix any $f \in H \setminus \{0\}$ and note that convexity of L implies $L''_{f_\mu} \geq 0$. Therefore,

$$\begin{aligned}
\|K_F(f)\|^2_H &= \langle 2\lambda_0 f + A(f), 2\lambda_0 f + A(f) \rangle_H = \\
&= 4\lambda_0^2 \|f\|^2_H + 4\lambda_0 \langle f, A(f) \rangle_H + \|A(f)\|^2_H > 4\lambda_0 \langle f, A(f) \rangle_H = \\
&= 4\lambda_0 \langle f, \int L''_{f_\mu} f \Phi \, d\mu \rangle_H \overset{(21)}{=} 4\lambda_0 \int L''_{f_\mu} f \langle f, \Phi \rangle_H \, d\mu = 4\lambda_0 \int L''_{f_\mu} \cdot f^2 \, d\mu \geq 0.
\end{aligned}$$

In the following, (ii) will be shown. To this end, let $M \subset H$ be a (norm-)bounded subset of H. Since $\mathcal{X}$ is compact, it follows from Steinwart and Christmann (2008, Corollary 4.31) that M is a relatively compact subset of $\mathcal{C}(\mathcal{X})$ (with respect to the norm-topology of $\mathcal{C}(\mathcal{X})$). In order to prove compactness of A, we have to show that every sequence $(A(f_j))_{j \in \mathbb{N}} \subset \{A(f) \mid f \in M\}$ contains a convergent subsequence. Relative compactness of M (in $\mathcal{C}(\mathcal{X})$) implies that there is a subsequence $(f_{j_\ell})_{\ell \in \mathbb{N}} \subset (f_j)_{j \in \mathbb{N}}$ which is a Cauchy-sequence in $\mathcal{C}(\mathcal{X})$. Since $\tilde{A}_{f_\iota(F)}$ is a continuous linear operator on $\mathcal{C}(\mathcal{X})$ (Lemma 5.2), this implies that the sequence

$$A(f_{j_\ell}) = \tilde{A}_{f_\iota(F)}(f_{j_\ell}), \qquad \ell \in \mathbb{N},$$

is a Cauchy-sequence in H. Hence, $(A(f_{j_\ell}))_{\ell \in \mathbb{N}}$ converges in H since H is complete. □

By use of these preliminary lemmas, Gâteaux-differentiability of the SVM-functional can be shown now:

Proposition 5.6 *Let $F \in B_S$, $G \in \ell_\infty(\mathcal{G})$ and $\rho > 0$ such that $F + sG \in B_S$ for every $s \in (-\rho, \rho)$. Then, there is a unique finite signed measure μ such that*

$$\int g \, d\mu = G(g) \qquad \forall g \in \mathcal{G}. \tag{27}$$

Furthermore,

$$\lim_{s \to 0} \left\| \frac{S(F + sG) - S(F)}{s} - S'_F(G) \right\|_H = 0$$

where

$$S'_F(G) = -K_F^{-1}\left(\mathbb{E}_\mu\left(L'_{f_\iota(F)} \Phi \right) \right). \tag{28}$$

In particular, S is Gâteaux-differentiable.

Proof: The following proof is similar to the proof of Steinwart and Christmann (2008, Theorem 10.18) but some care is needed because we also have to deal with *signed* measures here.

Part 1: Define $\nu := \iota(F)$. Since $G = s^{-1}\big((F + sG) - F\big) \in \text{lin}(B_S)$ for any $s \in (-\rho, \rho) \setminus \{0\}$, it follows from Lemma 5.2 that there is a unique finite signed measure μ such that

$$\int g \, d\mu \;=\; G(g) \qquad \forall\, g \in \mathcal{G} \,. \tag{29}$$

Define

$$\Gamma \;:\; \mathbb{R} \times H \;\to\; H \,, \qquad (s, f) \;\mapsto\; 2\lambda_0 f + \int L_f' \Phi \, d\nu + s \int L_f' \Phi \, d\mu \,.$$

Lemma 5.4 (b) implies that the maps $H \to H$, $f \mapsto \int L_f' \Phi \, d\nu$ and $H \to H$, $f \mapsto \int L_f' \Phi \, d\mu$ are continuous. Hence, an easy calculation shows that Γ is continuous.

Part 2: In this part, it will be shown that Γ is continuously Fréchet-differentiable. First, it follows from Lemma 5.4 (b) that the map

$$\mathbb{R} \times H \;\to\; H \,, \qquad (s, f) \;\mapsto\; \frac{\partial \Gamma}{\partial s}(s, f) \;=\; \int L_f' \Phi \, d\mu$$

is continuous. Secondly, Lemma 5.4 (b) yields that the partial derivative $\frac{\partial \Gamma}{\partial H}(s, f)$ is given by

$$\frac{\partial \Gamma}{\partial H}(s, f) \;:\; H \;\to\; H \,, \qquad h \;\mapsto\; 2\lambda_0 h + \int L_f'' h \Phi \, d\nu + s \int L_f'' h \Phi \, d\mu$$

for every $(s, f) \in \mathbb{R} \times H$. Let $\mathcal{B}(H, H)$ be the set of all continuous linear operators $T : H \to H$; this is a Banach space with the operator norm. It follows from Lemma 5.3 that

$$\mathbb{R} \times H \;\to\; \mathcal{B}(H, H), \qquad (s, f) \;\mapsto\; \frac{\partial \Gamma}{\partial H}(s, f)$$

is continuous. Since Γ is continuous (as stated above), this implies that Γ is continuously Fréchet-differentiable according to Denkowski et al. (2003, p. 635).

Part 3: Now, we can prove the statement of the lemma by use of an implicit function theorem. It follows from Lemma 5.4 (a) that

$$\Gamma(s, f) \;=\; \frac{\partial \mathcal{R}_{L, \nu + s\mu, \lambda_0}}{\partial H}(f) \qquad \forall\, f \in H \qquad \forall\, s \in (-\rho, \rho) \,. \tag{30}$$

Since $H \to \mathbb{R}$, $f \mapsto \mathcal{R}_{L,\nu+s\mu,\lambda_0}$ is strictly convex and continuously Fréchet-differentiable, the following assertion is valid for every $s \in (-\rho, \rho)$:

$$\Gamma(s, f) = 0 \qquad \Leftrightarrow \qquad f = f_{\nu+s\mu} \ . \tag{31}$$

(Direction "$\Leftarrow$" follows from Luenberger (1969, Theorem 7.4.1) and "$\Rightarrow$" follows from Luenberger (1969, Lemma 8.7.1) and uniqueness of the minimizer.) As shown in Part 2, Γ is continuously Fréchet-differentiable. According to Lemma 5.5,

$$\frac{\partial \Gamma}{\partial H}(0, f_\nu) = K_F$$

is an invertible operator. Therefore, it follows from a classical implicit function theorem (e.g. Akerkar (1999, §4)) that there is a $\delta \in (0, \rho)$ and a Fréchet-differentiable map $\varphi : (-\delta, \delta) \to H$ such that

$$\Gamma\big((s, \varphi(s)\big) = 0 \qquad \forall\, s \in (-\delta, \delta) \tag{32}$$

and the derivative is equal to

$$\varphi'(0) = -\left(\frac{\partial \Gamma}{\partial H}(0, \varphi(0))\right)^{-1} \left(\frac{\partial \Gamma}{\partial s}(0, \varphi(0))\right) = -K_F^{-1}\Big(\mathbb{E}_\mu\big(L'_{f_\nu}\Phi\big)\Big) \ .$$

According to (31) and (32), $\varphi(s) = f_{\nu+s\mu} = S(F+sG)$ for every $s \in (-\delta, \delta)$. Define $S'_F(G) = \varphi'(0)$. Hence,

$$\lim_{s \to 0} \left\|\frac{S(F+sG) - S(F)}{s} - S'_F(G)\right\|_H = \lim_{s \to 0} \left\|\frac{\varphi(s) - \varphi(0)}{s} - \varphi'(0)\right\|_H = 0 \ .$$

$\square$

5.3 Hadamard-Differentiability of the SVM-Functional

In this subsection, the result of the previous Subsection 5.2 is strengthened. In statistics, three different types of differentiability in Banach spaces are particularly important: Gâteaux-differentiability, Hadamard-differentiability and Fréchet-differentiability. Among these, Gâteaux is the weakest and Fréchet is the strongest notion of differentiability. In order to apply the functional delta-method, we need the intermediate Hadamard-differentiability. It is well-known that a Gâteaux-differentiable function is even Fréchet-differentiable (and, therefore, Hadamard-differentiable) if the (Gâteaux-)derivative is continuous. In the following Lemma 5.7, it will be shown that the Gâteaux-derivative of S fulfills a certain continuity property (34). This

property is not strong enough in order to guarantee Fréchet-differentiability. However, it will be shown in the proof of Theorem 5.8 that it is just strong enough in order to guarantee Hadamard-differentiability of S tangentially to the closed linear span of B_S. In order to do this, we only have to slightly change the proof of the well-known interrelationship between Gâteaux- and Fréchet-differentiability (as provided, e.g., by Denkowski et al. (2003, Prop. 5.1.8)).

Lemma 5.7 *Let $B_0 = \mathrm{cl}\big(\mathrm{lin}(B_S)\big)$ be the closed linear span of B_S in $\ell_\infty(\mathcal{G})$. Let $(G_n)_{n\in\mathbb{N}} \subset \mathrm{lin}(B_S)$ be a sequence such that $\lim_{n\to\infty} \|G_n - G_0\|_\infty = 0$ for some $G_0 \in \ell_\infty(\mathcal{G})$ and let $(F_n)_{n\in\mathbb{N}} \subset B_S$ be a sequence such that $\lim_{n\to\infty} \|F_n - F_0\|_\infty = 0$ for some $F_0 \in B_S$ which fulfills*

$$F_0(b) \; < \; \int b\, dP \; + \; \lambda_0 \; . \tag{33}$$

Then, there is a $n_0 \in \mathbb{N}$ such that, for every $F \in \{F_n | n \in \mathbb{N}_{\geq n_0}\} \cup \{F_0\}$, the map $S'_F : G \mapsto S'_F(G)$ defined in Proposition 5.6 can be extended to a continuous linear operator $S'_F : B_0 \to H$. In addition,

$$\lim_{n\to\infty} \big\| S'_{F_n}(G_n) - S'_{F_0}(G_0) \big\|_H \; = \; 0 \; . \tag{34}$$

Proof:
Part 1: Fix any $F \in B_S$ such that $\big\| f_{\iota(F)} \big\|_H \leq c_0$ where c_0 is defined as in (17). That is,

$$L'_{f_{\iota(F)}} f \;\; \in \;\; \mathcal{G} \qquad \forall f \in H \;\; \text{with} \;\; \|f\|_H \leq 1 \; . \tag{35}$$

According to Lemma 5.2, the map $S'_F : G \mapsto S'_F(G)$ defined in Proposition 5.6 can be extended to the map

$$S'_F \; : \;\; \mathrm{lin}(B_S) \; \to \; H \, , \qquad G \; \to \; -K_F^{-1}\Big(\mathbb{E}_{\iota(G)}\big(L'_{f_{\iota(F)}}\Phi\big)\Big)$$

Since ι is linear according to Lemma 5.2, this map is linear. In order to prove that S'_F is a continuous linear operator on $\mathrm{lin}(B_S)$, it is enough to show that

$$W_F \; : \;\; \mathrm{lin}(B_S) \; \to \; H \, , \qquad G \; \to \; \mathbb{E}_{\iota(G)}\big(L'_{f_{\iota(F)}}\Phi\big)$$

is a continuous linear operator because K_F^{-1} is a continuous linear operator according to Lemma 5.5. To this end, note that for every $G \in \mathrm{lin}(B_S)$ and every $f \in H$ such that $\|f\|_H \leq 1$,

$$\Big\langle \mathbb{E}_{\iota(G)}\big(L'_{f_{\iota(F)}}\Phi\big), \, f \Big\rangle_H \;\; \overset{(21,13)}{=} \;\; \mathbb{E}_{\iota(G)}\big(L'_{f_{\iota(F)}} f\big) \;\; \overset{(22,35)}{=} \;\; G\big(L'_{f_{\iota(F)}} f\big) \; .$$

That is, for every $f \in H$ such that $\|f\|_H \leq 1$,

$$\langle W_F(G), f \rangle_H = G\big(L'_{f_{\iota(F)}} f\big) \qquad \forall\, G \in \mathrm{lin}(B_S)\,. \tag{36}$$

Hence,

$$\big\|W_F(G)\big\|_H = \sup_{\substack{f \in H \\ \|f\|_H \leq 1}} \langle W_F(G), f \rangle_H \overset{(36)}{=} \sup_{\substack{f \in H \\ \|f\|_H \leq 1}} G\big(L'_{f_{\iota(F)}} f\big) \leq \|G\|_\infty$$

and, therefore, W_F is a continuous linear operator with operator norm

$$\big\|W_F\big\| \leq 1\,.$$

Since $\mathrm{lin}(B_S)$ is dense in B_0, W_F can be extended to a continuous linear operator $W_F : B_0 \to H$ with operator norm

$$\big\|W_F\big\| \leq 1\,, \tag{37}$$

see e.g. Megginson (1998, Theorem 1.9.1). Hence, S'_F can be extended to the continuous linear map

$$S'_F \;:\; B_0 \;\to\; H\,, \qquad G \;\mapsto\; -K_F^{-1}\big(W_F(G)\big)$$

on $B_0 = \mathrm{cl}\big(\mathrm{lin}(B_S)\big)$. In particular, the latter is eventually true for $F = F_n$ because it follows from $\lim_{n\to\infty} \|F_n - F_0\|_\infty = 0$, $b \in \mathcal{G}$, (17), (19) and (33) that there is some $n_0 \in \mathbb{N}$ such that

$$\big\|f_{\iota(F_n)}\big\|_H \;\leq\; c_0 \qquad \forall\, n \in \mathbb{N}_{\geq n_0} \cup \{0\}\,.$$

and, therefore, $F = F_n$ fulfills (35) for every $n \in \mathbb{N}_{\geq n_0} \cup \{0\}$.

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

$$K_{F_n}^{-1} \;\xrightarrow[n\to\infty]{}\; K_{F_0}^{-1} \qquad \text{in the operator norm}\,. \tag{38}$$

To this end, it suffices to show that

$$K_{F_n} \;\xrightarrow[n\to\infty]{}\; K_{F_0} \qquad \text{in the operator norm}$$

according to Dunford and Schwartz (1958, Lemma VII.6.1). Because of

$$\big\|K_{F_n}(f) - K_{F_0}(f)\big\|_H \leq \left\|\int L''_{f_{\iota(F_n)}} f\Phi \; d[\iota(F_n)] - \int L''_{f_{\iota(F_0)}} f\Phi \; d[\iota(F_n)]\right\|_H +$$

$$+ \left\|\int L''_{f_{\iota(F_0)}} f\Phi \; d[\iota(F_n)] - \int L''_{f_{\iota(F_0)}} f\Phi \; d[\iota(F_0)]\right\|_H,$$

26

this can be done by showing

$$\lim_{n\to\infty} \sup_{\substack{f\in H \\ \|f\|_H \le 1}} \left\| \int L''_{f_{\iota(F_n)}} f\Phi \, d[\iota(F_n)] - \int L''_{f_{\iota(F_0)}} f\Phi \, d[\iota(F_n)] \right\|_H = 0 \qquad (39)$$

and

$$\lim_{n\to\infty} \sup_{\substack{f\in H \\ \|f\|_H \le 1}} \left\| \int L''_{f_{\iota(F_0)}} f\Phi \, d[\iota(F_n)] - \int L''_{f_{\iota(F_0)}} f\Phi \, d[\iota(F_0)] \right\|_H = 0 \,. \qquad (40)$$

In order to prove (39), define

$$\tilde{F}_n := \frac{1}{\iota(F_n)(\mathcal{X}\times\mathcal{Y})} F_n \quad \text{and} \quad \tilde{\lambda}_n := \frac{\lambda_0}{\iota(F_n)(\mathcal{X}\times\mathcal{Y})} \qquad \forall\, n \in \mathbb{N} \cup \{0\}\,,$$

and

$$\tilde{F}_{0,n} := \frac{\lambda_0}{\tilde{\lambda}_n}\tilde{F}_0 \;=\; \frac{\iota(F_n)(\mathcal{X}\times\mathcal{Y})}{\iota(F_0)(\mathcal{X}\times\mathcal{Y})} F_0 \qquad \forall\, n \in \mathbb{N} \cup \{0\}$$

Then, $\iota(\tilde{F}_n)$ is a probability measure and it follows from Lemma 5.1 that $\lim_{n\to\infty} \iota(F_n)(\mathcal{X}\times\mathcal{Y}) = \iota(F_0)(\mathcal{X}\times\mathcal{Y})$ and, therefore, $\lim_{n\to\infty} \|\tilde{F}_n - \tilde{F}_0\|_\infty = 0$. Hence,

$$\lim_{n\to\infty} \left\| f_{\iota(F_n)} - f_{\iota(F_0)} \right\|_H \overset{(18)}{=} \lim_{n\to\infty} \left\| f_{L,\iota(\tilde{F}_n),\tilde{\lambda}_n} - f_{L,\iota(\tilde{F}_0),\tilde{\lambda}_0} \right\|_H \le$$

$$\le \lim_{n\to\infty} \left\| f_{L,\iota(\tilde{F}_n),\tilde{\lambda}_n} - f_{L,\iota(\tilde{F}_0),\tilde{\lambda}_n} \right\|_H + \left\| f_{L,\iota(\tilde{F}_0),\tilde{\lambda}_n} - f_{L,\iota(\tilde{F}_0),\tilde{\lambda}_0} \right\|_H$$

$$\overset{(*)}{\le} \lim_{n\to\infty} \frac{1}{\tilde{\lambda}_n} \left\| \int L'_{f_{L,\iota(\tilde{F}_0),\tilde{\lambda}_n}} \Phi \, d[\iota(\tilde{F}_n] - \int L'_{f_{L,\iota(\tilde{F}_0),\tilde{\lambda}_n}} \Phi \, d[\iota(\tilde{F}_0] \right\|_H$$

$$\overset{(18)}{=} \lim_{n\to\infty} \frac{1}{\tilde{\lambda}_n} \left\| \int L'_{f_{L,\iota(\tilde{F}_{0,n}),\lambda_0}} \Phi \, d[\iota(\tilde{F}_n] - \int L'_{f_{L,\iota(\tilde{F}_{0,n}),\lambda_0}} \Phi \, d[\iota(\tilde{F}_0] \right\|_H$$

$$= \lim_{n\to\infty} \frac{1}{\tilde{\lambda}_n} \left\| W_{\tilde{F}_{0,n}}(\tilde{F}_n) - W_{\tilde{F}_{0,n}}(\tilde{F}_0) \right\|_H \qquad (41)$$

where $(*)$ follows from Steinwart and Christmann (2008, Theorem 5.9 and Corollary 5.19). Since $\lim_{n\to\infty} \iota(F_n)(\mathcal{X}\times\mathcal{Y}) = \iota(F_0)(\mathcal{X}\times\mathcal{Y})$, it follows from (17), (19) and (33) that

$$\left\| f_{\iota(\tilde{F}_{0,n})} \right\|_H \le c_0 \qquad \text{for large enough } n \in \mathbb{N}.$$

Hence,

$$\lim_{n\to\infty} \left\| f_{\iota(F_n)} - f_{\iota(F_0)} \right\|_H \overset{(41)}{\le} \lim_{n\to\infty} \frac{1}{\tilde{\lambda}_n} \left\| W_{\tilde{F}_{0,n}}(\tilde{F}_n) - W_{\tilde{F}_{0,n}}(\tilde{F}_0) \right\|_H \le$$

$$\overset{(37)}{\le} \lim_{n\to\infty} \frac{1}{\tilde{\lambda}_n} \left\| \tilde{F}_n - \tilde{F}_0 \right\|_\infty = 0 \qquad (42)$$

27

Therefore, (39) follows from Lemma 5.3.

In order to prove (40), define $M := \sup_{n \in \mathbb{N} \cup \{0\}} \iota(F_n)(\mathcal{X} \times \mathcal{Y}) < \infty$ (see Lemma 5.1) and note that, according to Steinwart and Christmann (2008, Corollary 4.31),

$$\mathcal{F}_1 = \{f \in H \mid \|f\|_H \leq 1\} \subset \mathcal{C}(\mathcal{X})$$

can be identified with a relatively compact subset of $\mathcal{C}(\mathcal{X})$ (with respect to the norm-topology of $\mathcal{C}(\mathcal{X})$). Hence, for every $\varepsilon > 0$, there is an $m_\varepsilon \in \mathbb{N}$ and functions $f_1, \ldots, f_{m_\varepsilon} \in \mathcal{C}(\mathcal{X})$ such that

$$\|f_j\|_\infty \leq \sup_{f \in \mathcal{F}_1} \|f\|_\infty \overset{(15)}{\leq} \|k\|_\infty \qquad \forall j \in \{1, \ldots, m_\varepsilon\}, \qquad (43)$$

$$\min_{j \in \{1, \ldots, m_\varepsilon\}} \|f - f_j\|_\infty < \varepsilon \qquad \forall f \in \mathcal{F}_1. \qquad (44)$$

Define $a := \|f_{\iota(F_0)}\|_\infty$. Fix any $f \in \mathcal{F}_1$ and take $j_0 \in \{1, \ldots, m_\varepsilon\}$ such that $\|f - f_{j_0}\|_\infty < \varepsilon$. Then,

$$\left\| \int L''_{f_{\iota(F_0)}} f \Phi \, d[\iota(F_n)] - \int L''_{f_{\iota(F_0)}} f \Phi \, d[\iota(F_0)] \right\|_H =$$

$$= \left\| \int L''_{f_{\iota(F_0)}} (f - f_{j_0}) \Phi \, d[\iota(F_n)] - \int L''_{f_{\iota(F_0)}} (f - f_{j_0}) \Phi \, d[\iota(F_0)] - \right.$$

$$\left. - \int L''_{f_{\iota(F_0)}} f_{j_0} \Phi \, d[\iota(F_0)] + \int L''_{f_{\iota(F_0)}} f_{j_0} \Phi \, d[\iota(F_n)] \right\|_H \leq$$

$$\leq \int \left\| L''_{f_{\iota(F_0)}} (f - f_{j_0}) \Phi \right\|_H d[\iota(F_n)] + \int \left\| L''_{f_{\iota(F_0)}} (f - f_{j_0}) \Phi \right\|_H d[\iota(F_0)]$$

$$+ \left\| \int L''_{f_{\iota(F_0)}} f_{j_0} \Phi \, d[\iota(F_0)] - \int L''_{f_{\iota(F_0)}} f_{j_0} \Phi \, d[\iota(F_n)] \right\|_H \leq$$

$$\overset{(5,14)}{\leq} 2 b''_a \|k\|_\infty M \varepsilon + \left\| \int L''_{f_{\iota(F_0)}} f_{j_0} \Phi \, d[\iota(F_0)] - \int L''_{f_{\iota(F_0)}} f_{j_0} \Phi \, d[\iota(F_n)] \right\|_H$$

Hence,

$$\sup_{f \in \mathcal{F}_1} \left\| \int L''_{f_{\iota(F_0)}} f \Phi \, d[\iota(F_n)] - \int L''_{f_{\iota(F_0)}} f \Phi \, d[\iota(F_0)] \right\|_H \leq \qquad (45)$$

$$\leq 2 b''_a \|k\|_\infty M \varepsilon + \max_{j \in \{1, \ldots, m_\varepsilon\}} \left\| \int L''_{f_{\iota(F_0)}} f_j \Phi \, d[\iota(F_0)] - \int L''_{f_{\iota(F_0)}} f_j \Phi \, d[\iota(F_n)] \right\|_H.$$

Convergence of $(F_n)_{n \in \mathbb{N}}$ in $\ell_\infty(\mathcal{G})$ implies weak convergence (Lemma 5.1) and, therefore, tightness of the sequence of finite measures $(\iota(F_n))_{n \in \mathbb{N}}$; see

e.g. Bauer (2001, Theorem 30.8). Hence, there is a compact set $\mathcal{Z}_\varepsilon \subset \mathcal{X} \times \mathcal{Y}$ such that, for its complement $\complement\mathcal{Z}_\varepsilon$, we have $\sup_{n\in\mathbb{N}_0} \iota(F_n)(\complement\mathcal{Z}_\varepsilon) < \varepsilon$. Then,

$$\max_{j\in\{1,\dots,m_\varepsilon\}} \left\| \int L''_{f_{\iota(F_0)}} f_j \Phi \, d[\iota(F_0)] - \int L''_{f_{\iota(F_0)}} f_j \Phi \, d[\iota(F_n)] \right\|_H \leq$$

$$\leq \max_{j\in\{1,\dots,m_\varepsilon\}} \left\| \int_{\mathcal{Z}_\varepsilon} L''_{f_{\iota(F_0)}} f_j \Phi \, d[\iota(F_0)] - \int_{\mathcal{Z}_\varepsilon} L''_{f_{\iota(F_0)}} f_j \Phi \, d[\iota(F_n)] \right\|_H +$$

$$+ \left\| \int_{\complement\mathcal{Z}_\varepsilon} L''_{f_{\iota(F_0)}} f_j \Phi \, d[\iota(F_0)] \right\|_H + \left\| \int_{\complement\mathcal{Z}_\varepsilon} L''_{f_{\iota(F_0)}} f_j \Phi \, d[\iota(F_n)] \right\|_H$$

$$\overset{(5.43)}{\leq} \max_{j\in\{1,\dots,m_\varepsilon\}} \left\| \int_{\mathcal{Z}_\varepsilon} L''_{f_{\iota(F_0)}} f_j \Phi \, d[\iota(F_0)] - \int_{\mathcal{Z}_\varepsilon} L''_{f_{\iota(F_0)}} f_j \Phi \, d[\iota(F_n)] \right\|_H + 2b''_a \|k\|^2_\infty \varepsilon.$$

According to Bourbaki (2004, p. III.40), weak convergence of the sequence of finite (positive) measures $(\iota(F_n))_{n\in\mathbb{N}}$ implies

$$\lim_{n\to\infty} \left\| \int_{\mathcal{Z}_\varepsilon} L''_{f_{\iota(F_0)}} f_j \Phi \, d[\iota(F_0)] - \int_{\mathcal{Z}_\varepsilon} L''_{f_{\iota(F_0)}} f_j \Phi \, d[\iota(F_n)] \right\|_H = 0$$

for every $j \in \{1, \dots, m_\varepsilon\}$. (Since H is a separable Banach space, Pettis integrals and Bochner-integrals coincide; see e.g. Dudley (2002, p. 194f).) As $\varepsilon > 0$ can be arbitrarily small, (40) follows from (45) and the above calculation.

Part 3: In this part of the proof, it will be shown that

$$\lim_{n\to\infty} \left\| W_{F_n}(G_0) - W_{F_0}(G_0) \right\|_H = 0. \tag{46}$$

For every $m \in \mathbb{N}$, we have $G_m \in \mathrm{lin}(B_S)$ and, therefore, $W_{F_n}(G_m) = \int L'_{f_{\iota(F_n)}} \Phi \, d[\iota(G_m)]$ for every $n \in \mathbb{N}_0$. Hence, it follows from (42) and Lemma 5.4 b) that

$$\lim_{n\to\infty} \left\| W_{F_n}(G_m) - W_{F_0}(G_m) \right\|_H = 0 \qquad \forall\, m \in \mathbb{N}. \tag{47}$$

Furthermore, we have

$$\lim_{m\to\infty} \sup_{n\in\mathbb{N}_0} \left\| W_{F_n}(G_m) - W_{F_n}(G_0) \right\|_H \overset{(37)}{\leq} \lim_{m\to\infty} \left\| G_m - G_0 \right\|_\infty = 0 \tag{48}$$

According to Dunford and Schwartz (1958, I.7.6), (47) and (48) imply

$$\lim_{n\to\infty} \left\| W_{F_n}(G_0) - W_{F_0}(G_0) \right\|_H = \lim_{n\to\infty} \lim_{m\to\infty} \left\| W_{F_n}(G_m) - W_{F_0}(G_m) \right\|_H =$$

$$= \lim_{m\to\infty} \lim_{n\to\infty} \left\| W_{F_n}(G_m) - W_{F_0}(G_m) \right\|_H = 0.$$

29

Part 4: By use of the previous parts, we complete the proof by proving (34):

$$\lim_{n\to\infty} \left\|S'_{F_n}(G_n) - S'_{F_0}(G_0)\right\|_H = \lim_{n\to\infty} \left\|K_{F_n}^{-1}\big(W_{F_n}(G_n)\big) - K_{F_0}^{-1}\big(W_{F_0}(G_0)\big)\right\|_H$$

$$\leq \lim_{n\to\infty} \left\|K_{F_n}^{-1}\big(W_{F_n}(G_n)\big) - K_{F_0}^{-1}\big(W_{F_n}(G_n)\big)\right\|_H +$$

$$+ \left\|K_{F_0}^{-1}\big(W_{F_n}(G_n)\big) - K_{F_0}^{-1}\big(W_{F_n}(G_0)\big)\right\|_H +$$

$$+ \left\|K_{F_0}^{-1}\big(W_{F_n}(G_0)\big) - K_{F_0}^{-1}\big(W_{F_0}(G_0)\big)\right\|_H =$$

$$\overset{(46)}{\leq} \lim_{n\to\infty} \left\|K_{F_n}^{-1} - K_{F_0}^{-1}\right\| \cdot \left\|W_{F_n}(G_n)\right\|_H + \left\|K_{F_0}^{-1}\right\| \cdot \left\|W_{F_n}(G_n) - W_{F_n}(G_0)\right\|_H$$

$$\overset{(37)}{\leq} \lim_{n\to\infty} \left\|K_{F_n}^{-1} - K_{F_0}^{-1}\right\| \cdot \left\|G_n\right\|_\infty + \left\|K_{F_0}^{-1}\right\| \cdot \left\|G_n - G_0\right\|_\infty = 0$$

$$\square$$

Theorem 5.8 _For every $F_0 \in B_S$ which fulfills (33), the map_

$$S : \quad B_S \to H, \qquad F \mapsto f_{\iota(F)}$$

is Hadamard-differentiable in F_0 tangentially to the closed linear span $B_0 = \mathrm{cl}\big(\mathrm{lin}(B_S)\big)$. The derivative in F_0 is a continuous linear operator $S'_{F_0} : B_0 \to H$ such that

$$S'_{F_0}(G) = -K_F^{-1}\Big(\mathbb{E}_{\iota(G)}\big(L'_{f_{\iota(F)}}\Phi\big)\Big) \qquad \forall\, G \in \mathrm{lin}(B_S). \tag{49}$$

Proof: Let $(G_n)_{n\in\mathbb{N}} \subset \ell_\infty(\mathcal{G})$ and $(t_n)_{n\in\mathbb{N}} \subset \mathbb{R}\setminus\{0\}$ be sequences such that $\lim_{n\to\infty}\|G_n - G_0\|_\infty = 0$ for some $G_0 \in \ell_\infty(\mathcal{G})$, such that $t_n \searrow 0$, and such that $F_n := F_0 + t_n G_n \in B_S$ for every $n \in \mathbb{N}$. Then, $\lim_{n\to\infty}\|F_n - F_0\|_\infty = 0$ and $G_n \in \mathrm{lin}(B_s)$ for every $n \in \mathbb{N}$. According to Lemma 5.7, there is a $n_0 \in \mathbb{N}$ such that, for every $F \in \{F_n | n \in \mathbb{N}_{\geq n_0}\} \cup \{F_0\}$, there is a continuous linear operator $S'_F : B_0 \to H$ which fulfills (49). We have to show

$$\lim_{n\to\infty} \left\|\frac{S(F_0 + t_n G_n) - S(F_0)}{t_n} - S'_{F_0}(G_0)\right\|_H = 0. \tag{50}$$

Note that the assumptions imply $G_0 \in B_0$. Define

$$h_n := S(F_0 + t_n G_n) - S(F_0) - t_n S'_{F_0}(G_0) \qquad \forall\, n \in \mathbb{N}. \tag{51}$$

That is, for every $f \in H$,

$$\langle f, h_n\rangle_H = \langle f, S(F_0 + t_n G_n) - S(F_0)\rangle_H - \langle f, t_n S'_{F_0}(G_0)\rangle_H. \tag{52}$$

30

In order to prove for every $n \in \mathbb{N}$ that the function

$$[0,1] \;\to\; H, \qquad s \;\mapsto\; S(F_0 + st_nG_n)$$

is well-defined, we have to show that $F_0 + st_nG_n \in B_S$ for every $s \in [0,1]$. It follows from $F_n \in B_S$ that $G_n \in \mathrm{lin}(B_S)$. Therefore, there is a finite signed measure $\mu_{n,s}$ such that $\mu_{n,s} = \iota(F_0 + st_nG_n)$ and $F_0 + st_nG_n \in \mathrm{lin}(B_S)$. Take any $A \in \mathfrak{B}(\mathcal{X} \times \mathcal{Y})$. Then, it follows from $\iota(F_0)(A) \geq 0$, $\iota(F_n)(A) \geq 0$ and $s \in [0,1]$ that $\mu_{n,s}(A) = \iota(F_0 + st_nG_n)(A) \geq 0$. That is, $\mu_{n,s} = \iota(F_0 + st_nG_n)$ is a finite measure. Furthermore, it follows from $F_0 \neq 0$, $F_n \neq 0$ and $s \in [0,1]$ that $\mu_{n,s} \neq 0$. According to the definitions, this shows that $F_0 + st_nG_n \in B_S$.

Fix any $n \in \mathbb{N}$. The function $s \mapsto S(F_0 + st_nG_n)$ is continuous on $[0,1]$ according to (42) and Frechét-differentiable on $(0,1)$ according to Proposition 5.6; the derivative in $s \in (0,1)$ is given by $S'_{F_0+st_nG_n}(t_nG_n)$. Since the map $h \mapsto \langle f, h \rangle_H$ is Frechét-differentiable for every $f \in H$, this implies that

$$(0,1) \;\to\; \mathbb{R}, \qquad s \;\mapsto\; \langle f, S(F_0 + st_nG_n) \rangle_H$$

is differentiable for every $f \in H$; the derivative in $s \in (0,1)$ is given by $\langle f, S'_{F_0+st_nG_n}(t_nG_n) \rangle_H$. Define $\tilde{h}_n = h_n / \|h_n\|_H$. According to the elementary mean value theorem, there is an $\tilde{s}_n \in (0,1)$ such that

$$
\begin{aligned}
\langle \tilde{h}_n, S'_{F_0+\tilde{s}_nt_nG_n}(t_nG_n) \rangle_H &= \langle \tilde{h}_n, S(F_0 + t_nG_n) \rangle_H - \langle \tilde{h}_n, S(F_0) \rangle_H = \\
&= \langle \tilde{h}_n, S(F_0 + t_nG_n) - S(F_0) \rangle_H
\end{aligned}
$$

By use of the definition of h_n, this implies

$$\langle \tilde{h}_n, h_n \rangle_H \;=\; \langle \tilde{h}_n, \, S'_{F_0+\tilde{s}_nt_nG_n}(t_nG_n) - t_nS'_{F_0}(G_0) \rangle_H$$

and, by use of the definition of $\tilde{h}_n$, the latter equality and the Cauchy-Schwarz inequality imply

$$\big\| h_n \big\|_H \;\leq\; \big\| S'_{F_0+\tilde{s}_nt_nG_n}(t_nG_n) - t_nS'_{F_0}(G_0) \big\|_H. \tag{53}$$

Then, (50) follows from

$$\left\| \frac{S(F_0 + t_nG_n) - S(F_0)}{t_n} - S'_{F_0}(G_0) \right\|_H =$$

$$= \frac{\big\| S(F_0 + t_nG_n) - S(F_0) - t_nS'_{F_0}(G_0) \big\|}{t_n} \;\overset{(51)}{=}\; \frac{1}{t_n} \big\| h_n \big\|_H \;\leq$$

$$\overset{(53)}{\leq}\; \frac{1}{t_n} \big\| S'_{F_0+\tilde{s}_nt_nG_n}(t_nG_n) - t_nS'_{F_0}(G_0) \big\|_H = \big\| S'_{F_0+\tilde{s}_nt_nG_n}(G_n) - S'_{F_0}(G_0) \big\|_H$$

because the last expression converges to 0 according to Lemma 5.7. $\qquad \square$

5.4 Donsker-Classes and Application of the Delta-Method

It is well-known that

$$\sqrt{n}\big(\mathbb{F}_n - F\big) \rightsquigarrow \mathbb{G}_1 \qquad \text{in} \quad \ell_\infty(\mathcal{G}_1)$$

where $\mathbb{F}_n$ denotes the empirical process, F denotes the distribution function of P, $\mathbb{G}_1$ is a Gaussian process, and $\mathcal{G}_1$ is the set of all indicator functions. However, as already noted in Subsection 5.1, the set of indicator functions had to be enlarged to a set $\mathcal{G} \supset \mathcal{G}_1$ in order to ensure Hadamard-differentiability of the SVM-functional

$$S \; : \; B_S \; \longrightarrow \; H$$

in a neighborhood of $F \in B_S \subset \ell_\infty(\mathcal{G})$. Therefore, it still has to be proven that weak convergence not only holds in $\ell_\infty(\mathcal{G}_1)$ but also in $\ell_\infty(\mathcal{G})$. This is done in the following Lemma 5.9. After that, the main results can be proven by applications of a functional delta-method.

Lemma 5.9 *For every* $D_n = \big((x_1, y_1), \ldots, (x_n, y_n)\big) \in (\mathcal{X} \times \mathcal{Y})^n$, *let* $\mathbb{F}_{D_n}$ *denote the element of* $\ell_\infty(\mathcal{G})$ *which corresponds to the empirical measure* $\mathbb{P}_{D_n}$. *That is,* $\mathbb{F}_{D_n}(g) = \int g \, d\mathbb{P}_{D_n} = \frac{1}{n} \sum_{i=1}^{n} g(x_i, y_i)$ *for every* $g \in \mathcal{G}$. *Then,*

$$\sqrt{n}\big(\mathbb{F}_{\mathbf{D}_n} - \iota^{-1}(P)\big) \rightsquigarrow \mathbb{G} \qquad \text{in} \quad \ell_\infty(\mathcal{G})$$

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

Proof: In other words, we have to show that $\mathcal{G}$ is a P-Donsker class.
Part 1: Fix any $c \in (0, \infty)$. In *Part 1* of the proof, it will be shown that

$$\mathcal{F}_c \; := \; \big\{ f \in H \,\big|\, \|f\|_H \leq c \big\}$$

has a finite uniform entropy integral. Since $\mathcal{X} \subset \mathbb{R}^d$ is bounded, there is an $r > 0$ such that $\mathcal{X} \subset \big\{ x \in \mathbb{R}^d \,\big|\, \|x\|_{\mathbb{R}^d} < r \big\} =: \tilde{\mathcal{X}}$. Then, $\tilde{\mathcal{X}}$ is a convex, bounded subset of $\mathbb{R}^d$ with non-empty interior. Let $\tilde{H}$ be the RKHS of the restriction of the kernel $\tilde{k}$ on $\tilde{\mathcal{X}} \times \tilde{\mathcal{X}}$ and define

$$\tilde{\mathcal{F}}_c \; := \; \big\{ \tilde{f} \in \tilde{H} \,\big|\, \|\tilde{f}\|_{\tilde{H}} \leq c \big\} \; .$$

It follows from Berlinet and Thomas-Agnan (2004, Theorem 4.2.6) that

$$\mathcal{F}_c \; := \; \big\{ f \in H \,\big|\, f \text{ is the restriction of some } \tilde{f} \in \tilde{H} \big\} \; . \tag{54}$$

According to van der Vaart and Wellner (1996, p. 154), let $\mathcal{C}_1^m(\tilde{\mathcal{X}})$ denote the set of all functions $\tilde{f} : \tilde{\mathcal{X}} \to \mathbb{R}$ which have uniformly bounded partial derivatives up to order $m-1$ and whose partial derivatives of order $m-1$ are Lipschitz-continuous such that

$$\|\tilde{f}\|_1 := \max_{\substack{\alpha \in \mathbb{N}_0 \\ |\alpha| \leq m-1}} \sup_{x \in \tilde{\mathcal{X}}} \left|\partial^\alpha \tilde{f}(x)\right| + \max_{\substack{\alpha \in \mathbb{N}_0 \\ |\alpha| = m-1}} \sup_{\substack{x, x' \in \tilde{\mathcal{X}} \\ x \neq x'}} \frac{\left|\partial^\alpha \tilde{f}(x) - \partial^\alpha \tilde{f}(x')\right|}{\|x - x'\|_{\mathbb{R}^d}} \leq 1 .$$

It follows from convexity of $\tilde{\mathcal{X}}$ and the mean value theorem that

$$\max_{\substack{\alpha \in \mathbb{N}_0 \\ |\alpha| = m-1}} \sup_{\substack{x, x' \in \tilde{\mathcal{X}} \\ x \neq x'}} \frac{\left|\partial^\alpha \tilde{f}(x) - \partial^\alpha \tilde{f}(x')\right|}{\|x - x'\|_{\mathbb{R}^d}} \leq \max_{\substack{\alpha \in \mathbb{N}_0 \\ |\alpha| = m}} \sup_{x \in \tilde{\mathcal{X}}} \left|\partial^\alpha \tilde{f}(x)\right| .$$

Hence, it follows from Steinwart and Christmann (2008, Corollary 4.36) that, for every $\tilde{f} \in \tilde{\mathcal{F}}_c$,

$$\|\tilde{f}\|_1 \leq \max_{\substack{\alpha \in \mathbb{N}_0 \\ |\alpha| \leq m}} \sup_{x \in \tilde{\mathcal{X}}} \left|\partial^\alpha \tilde{f}(x)\right| \leq \|\tilde{f}\|_{\tilde{H}} \max_{\substack{\alpha \in \mathbb{N}_0 \\ |\alpha| \leq m}} \sup_{x \in \tilde{\mathcal{X}}} \left(\partial^{\alpha,\alpha} \tilde{k}(x, x)\right)$$

$$\leq c \cdot \max_{\substack{\alpha \in \mathbb{N}_0 \\ |\alpha| \leq m}} \sup_{x \in \tilde{\mathcal{X}}} \left(\partial^{\alpha,\alpha} \tilde{k}(x, x)\right) =: a_c \in (0, \infty) .$$

That is, $\frac{1}{a_c}\tilde{\mathcal{F}}_c \subset \mathcal{C}_1^m(\tilde{\mathcal{X}})$ and, therefore, it follows from van der Vaart and Wellner (1996, Theorem 2.7.1) that there is a constant $r \in (0, \infty)$ such that, for every $\varepsilon > 0$,

$$\ln N\left(a_c \varepsilon, \tilde{\mathcal{F}}_c, \|\cdot\|_\infty\right) = \ln N\left(\varepsilon, \tfrac{1}{a_c}\tilde{\mathcal{F}}_c, \|\cdot\|_\infty\right) \leq r \cdot \left(\frac{1}{\varepsilon}\right)^{\frac{d}{m}} . \qquad (55)$$

Here and in the following, $N(\cdot, \cdot, \cdot)$ denotes the covering number and $N_{[]}(\cdot, \cdot, \cdot)$ denotes the bracketing number; see e.g. van der Vaart and Wellner (1996, §2.1.1). According to (54), $\mathcal{F}_c$ is the set of restrictions of the elements of $\tilde{\mathcal{F}}_c$ on $\mathcal{X}$. By use of this fact, it is easy to see that

$$\ln N\left(\varepsilon, \mathcal{F}_c, \|\cdot\|_\infty\right) \leq \ln N\left(\varepsilon, \tilde{\mathcal{F}}_c, \|\cdot\|_\infty\right)$$

for every $\varepsilon > 0$. Therefore, it follows from (55) that

$$\ln N\left(\varepsilon, \mathcal{F}_c, \|\cdot\|_\infty\right) \leq r \cdot a_c^{\frac{d}{m}} \left(\frac{1}{\varepsilon}\right)^{\frac{d}{m}} \qquad \forall \varepsilon > 0 . \qquad (56)$$

Now, choose the constant $f_c = \|k\|_\infty c + 1$ as an envelope of $\mathcal{F}_c$. Every element $f \in \mathcal{F}_c$ can be identified with a function $\mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ via $f(x,y) = f(x)$. For every probability measure $\tilde{P}$ on $(\mathcal{X} \times \mathcal{Y}, \mathfrak{B}(\mathcal{X} \times \mathcal{Y}))$, we obtain

$$\|f\|_{L_2(\tilde{P})} \leq \sup_{(x,y)\in\mathcal{X}\times\mathcal{Y}} |f(x,y)| = \sup_{x\in\mathcal{X}} |f(x)| = \|f\|_\infty .$$

Therefore, it follows from (56) that

$$\sup_{\tilde{P}} \ln N\big(\varepsilon\|f_c\|_{L_2(\tilde{P})}, \mathcal{F}_c, \|\cdot\|_{L_2(\tilde{P})}\big) \leq r \left(\frac{a_c}{\|k\|_\infty c + 1}\right)^{\frac{d}{m}} \left(\frac{1}{\varepsilon}\right)^{\frac{d}{m}} \qquad (57)$$

where the supremum is taken over all probability measures $\tilde{P}$ on $(\mathcal{X} \times \mathcal{Y}, \mathfrak{B}(\mathcal{X} \times \mathcal{Y}))$. Since $m > \frac{d}{2}$ by assumption, the function class $\mathcal{F}_c$ has a finite uniform entropy integral. That is,

$$\int_{(0,1)} \sqrt{\sup_{\tilde{P}} \ln N\big(\varepsilon\|f_c\|_{L_2(\tilde{P})}, \mathcal{F}_c, \|\cdot\|_{L_2(\tilde{P})}\big)} \, \lambda(d\varepsilon) < \infty .$$

Part 2: Now, it will be shown that

$$\mathcal{G}' := \left\{ L'_f : (x,y) \mapsto L'(x,y,f(x)) \ \middle| \ f \in \mathcal{F}_{c_0} \right\}$$

also has a finite uniform entropy integral. Since

$$\sup_{x\in\mathcal{X}} |f(x)| \overset{(15)}{\leq} \|k\|_\infty c_0 =: a \qquad \forall f \in \mathcal{F}_{c_0},$$

the assumptions imply that $g' := b''_a + b'_a$ is an envelope function of $\mathcal{G}'$ such that $0 \leq b''_a \leq g'$ and, for every $(x,y) \in \mathcal{X} \times \mathcal{Y}$ and every $f_1, f_2 \in \mathcal{F}_{c_0}$,

$$\left|L'_{f_1}(x,y) - L'_{f_2}(x,y)\right| \overset{(*)}{\leq} b''_a|f_1(x) - f_2(x)| \leq g'(x,y)\|f_1 - f_2\|_\infty \qquad (58)$$

where $(*)$ follows from the assumptions on L'' and the elementary mean value theorem. For every probability measure $\tilde{P}$ on $(\mathcal{X} \times \mathcal{Y}, \mathfrak{B}(\mathcal{X} \times \mathcal{Y}))$ such that $0 < \int (g')^2 \, d\tilde{P} < \infty$, it follows from (58) and van der Vaart and Wellner (1996, p. 84 and Theorem 2.7.11) that, for every $\varepsilon > 0$,

$$\ln N\big(\varepsilon\|g'\|_{L_2(\tilde{P})}, \mathcal{G}', \|\cdot\|_{L_2(\tilde{P})}\big) \leq \ln N_{[]}\big(2\varepsilon\|g'\|_{L_2(\tilde{P})}, \mathcal{G}', \|\cdot\|_{L_2(\tilde{P})}\big) \leq$$

$$\leq \ln N\big(\varepsilon, \mathcal{F}_{c_0}, \|\cdot\|_\infty\big) \overset{(56)}{\leq} r \cdot a_{c_0}^{\frac{d}{m}} \left(\frac{1}{\varepsilon}\right)^{\frac{d}{m}} .$$

34

Hence, the assumption $m > \frac{d}{2}$ implies that $\mathcal{G}'$ has a finite uniform entropy integral.

Part 3: Now, it will be shown that $\mathcal{G}$ is a P-Donsker class. Trivially, $\{b\}$ is a P-Donsker class because $b \in L_2(P)$ by assumption. From van der Vaart and Wellner (1996, Example 2.5.4) it follows that $\mathcal{G}_1$ is P-Donsker. Note that $\mathcal{G}_2 = \mathcal{G}' \cdot \mathcal{F}_c$ for $c = 1$. According to Part 1, the class $\mathcal{F}_c$ has a finite uniform entropy integral relative to the (constant) envelope f_c and, according to Part 2, the class $\mathcal{G}'$ has a finite uniform entropy integral relative to the envelope g'. Therefore, it follows from van der Vaart (1998, Example 19.19) that $\mathcal{G}_2 = \mathcal{G}' \cdot \mathcal{F}_c$ has a finite uniform entropy integral relative to the envelope $f_c g'$. The definitions and assumptions imply $\int (f_c g')^2 \, dP < \infty$.

Hence, it follows from van der Vaart (1998, Theorem 19.4) that $\mathcal{G}_2$ is a P-Donsker class provided that $\mathcal{G}_2$ is "suitably measurable". According to van der Vaart (1998, p. 274), it suffices to show that there is a countable subset $\hat{\mathcal{G}}_2 \subset \mathcal{G}_2$ such that, for every $g \in \mathcal{G}_2$, there is a sequence $(\hat{g}_n)_{n \in \mathbb{N}} \subset \hat{\mathcal{G}}_2$ which converges pointwise to g. According to Steinwart and Christmann (2008, Lemma 4.33), H is a separable Hilbert space and, therefore, the subsets $\mathcal{F}_c \subset H$ are also separable for $c = 1$ and $c = c_0$. That is, there are countable subsets $\hat{\mathcal{F}}_1 \subset \mathcal{F}_1$ and $\hat{\mathcal{F}}_{c_0} \subset \mathcal{F}_{c_0}$ which are dense in $\mathcal{F}_1$ and $\mathcal{F}_{c_0}$ respectively (with respect to the norm topology). Then,

$$\hat{\mathcal{G}}_2 := \left\{ L'_{\hat{f}_0} \hat{f}_1 \mid \hat{f}_0 \in \mathcal{F}_{c_0}, \ \hat{f}_1 \in \mathcal{F}_1 \right\}$$

is again countable. Fix any $g \in \mathcal{G}_2$. That is, there are $f_0 \in \mathcal{F}_{c_0}$ and $f_1 \in \mathcal{F}_1$ such that $g = L'_{f_0} f_1$. Furthermore, there are sequences $\big(\hat{f}_0^{(n)}\big)_{n \in \mathbb{N}} \in \mathcal{F}_{c_0}$ and $\big(\hat{f}_1^{(n)}\big)_{n \in \mathbb{N}} \in \mathcal{F}_1$ such that

$$\lim_{n \to \infty} \big\| \hat{f}_0^{(n)} - f_0 \big\|_H = 0 \qquad \text{and} \qquad \lim_{n \to \infty} \big\| \hat{f}_1^{(n)} - f_1 \big\|_H = 0 \,.$$

Next, define $\hat{g}_n := L'_{\hat{f}_0^{(n)}} \hat{f}_1^{(n)} \in \hat{\mathcal{G}}_2$ for every $n \in \mathbb{N}$. Since H is a reproducing kernel Hilbert space, norm convergence implies pointwise convergence so that, for every $(x, y) \in \mathcal{X} \times \mathcal{Y}$,

$$\lim_{n \to \infty} \hat{g}_n(x, y) = \lim_{n \to \infty} L'\big(x, y, \hat{f}_0^{(n)}(x)\big) \hat{f}_1^{(n)}(x) = L'\big(x, y, f_0(x)\big) f_1(x) = g(x, y)$$

due to continuity of L'.

Part 4: As $\mathcal{G}$ is assured to be a P-Donsker class, we have

$$\sqrt{n}\big(\mathbb{F}_{\mathbf{D}_n} - \iota^{-1}(P)\big) \rightsquigarrow \mathbb{G} \qquad \text{in} \quad \ell_\infty(\mathcal{G})$$

where $\mathbb{G} : \Omega \to \ell_\infty(\mathcal{G})$ is a tight Borel-measurable Gaussian process. Since $\sqrt{n}\big(\mathbb{F}_{\mathbf{D}_n(\omega)} - \iota^{-1}(P)\big) \in B_0$ for every $\omega \in \Omega$ and every $n \in \mathbb{N}$, it follows from closedness of B_0 and the Portmanteau theorem (van der Vaart and Wellner, 1996, Theorem 1.3.4(iii)) that $\mathbb{G}(\omega) \in B_0$ almost surely. Hence, we may assume without loss of generality that $\mathbb{G}(\omega) \in B_0$ for every $\omega \in \Omega$. (Otherwise, replace $\mathbb{G}$ by $\mathbb{G} \cdot (I_{B_0} \circ \mathbb{G})$.) $\hfill\square$

For ease of reference, the following lemma summarizes some facts about Bochner-integrals of tight Gaussian processes in a space $\ell_\infty(T)$. Later on, these facts are needed in order to prove that the Gaussian process $\mathbb{H} : \Omega \to H$ is zero-mean.

Lemma 5.10 *Let T be any set, $\ell_\infty(T)$ the set of all bounded functions $h : T \to \mathbb{R}$ (endowed with the supremum-norm) and $\mathbb{G} : \Omega \to \ell_\infty(T)$ a tight Borel-measurable Gaussian process such that*

$$\int \mathbb{G}(\omega)(t)\, Q(d\omega) \;=\; 0 \qquad \forall\, t \in T\, . \tag{59}$$

Then, the Bochner-integral of $\mathbb{G} : \Omega \to \ell_\infty(T)$ exists and $\int \mathbb{G}(\omega)\, Q(d\omega) = 0$. Furthermore, $\int A(\mathbb{G})\, dQ = 0$ for every Banach space E and every continuous linear operator $A : \ell_\infty(T) \to E$.

Proof: Since $\mathbb{G}$ is tight, it is also separable so that there is a separable subset $\Gamma \subset \ell_\infty(T)$ such that $Q(\mathbb{G} \in \Gamma) = 1$; see van der Vaart and Wellner (1996, 16f). As the closed linear span of a separable subset of a Banach space is again separable (Schechter, 2004, Lemma A.48), we may assume without loss of generality that Γ is a separable Banach space. Define $\hat{\mathbb{G}} = \mathbb{G} \cdot (I_\Gamma \circ \mathbb{G})$. Then, $\hat{\mathbb{G}} : \Omega \to \Gamma$ is a Borel-measurable map. Let $\hat{h}^* : \Gamma \to \mathbb{R}$ be a continuous linear functional. According to the Hahn-Banach-Theorem (Dunford and Schwartz, 1958, Theorem II.3.11), $\hat{h}^*$ can be extended to a continuous linear functional $h^* : \ell_\infty(T) \to \mathbb{R}$. Since $h^*(\mathbb{G})$ is normally distributed according to van der Vaart and Wellner (1996, Lemma 3.9.8) and $\hat{h}^*(\hat{\mathbb{G}}) = h^*(\mathbb{G})\ Q - \text{a.s.}$, the real random variable $\hat{h}^*(\hat{\mathbb{G}})$ is normally distributed. This proves that the Borel-measurable map $\hat{\mathbb{G}} : \Omega \to \Gamma$ is a Gaussian process in the separable Banach space Γ. Hence, it follows from Satô (1971) that $\int \|\hat{\mathbb{G}}\|\, dQ < \infty$ and, therefore,

$$\int \|\mathbb{G}\|\, dQ \;<\; \infty\, . \tag{60}$$

(Fernique (1970) proves a related statement for centered Gaussian processes but we still have to prove that $\mathbb{G}$ is centered and this will be done by

36

use of (60) so that we cannot use Fernique's theorem here.) According to Denkowski et al. (2003, Theorem 3.10.3 and Theorem 3.10.9), (60) is equivalent to the existence of the Bochner-integral $\int \mathbb{G} \, dQ$.

Note that, for every $t \in T$, the map $\tau_t : \ell_\infty(T) \to \mathbb{R}$, $h \mapsto h(t)$ is a continuous linear operator. Then, by use of the fact that the Bochner-integral may be interchanged with continuous linear operators Denkowski et al. (2003, Theorem 3.10.16 and Remark 3.10.17), we get

$$\left(\int \mathbb{G}(\omega) \, Q(d\omega) \right)(t) \;=\; \tau_t \left(\int \mathbb{G}(\omega) \, Q(d\omega) \right) \;=\; \int \tau_t \big(\mathbb{G}(\omega) \big) \, Q(d\omega) \;=\;$$
$$=\; \int \mathbb{G}(\omega)(t) \, Q(d\omega) \;\overset{(59)}{=}\; 0$$

for every $t \in T$. That is, $\int \mathbb{G} \, dQ = 0$. Using again the fact that the Bochner-integral may be interchanged with continuous linear operators, we finally get $\int A(\mathbb{G}) \, dQ = A \left(\int \mathbb{G} \, dQ \right) = A(0) = 0$. $\qquad\qquad \square$

Proof of Theorem 3.1: First, it will be shown that

$$\Omega \;\to\; H \,, \qquad \omega \;\mapsto\; f_{L, \mathbf{D}_n(\omega), \lambda_{\mathbf{D}_n(\omega)}}$$

is Borel-measurable. According to the assumptions, it follows from Steinwart and Christmann (2008, Lemma 5.13 and Corollary 5.19) that $(\mathcal{X} \times \mathcal{Y})^n \to H$, $D_n \mapsto f_{L, D_n, \lambda}$ is continuous for every constant $\lambda \in (0, \infty)$ and that $(0, \infty) \to H$, $\lambda \mapsto f_{L, D_n, \lambda}$ is continuous for every $D_n \in (\mathcal{X} \times \mathcal{Y})^n$. Hence, $(D_n, \lambda) \mapsto f_{L, D_n, \lambda}$ is a Carathéodory function and, therefore, measurable; see, e.g., Denkowski et al. (2003, Theorem 2.5.22). Since $\omega \mapsto \mathbf{D}_n(\omega)$ and $\omega \mapsto \lambda_{\mathbf{D}_n(\omega)}$ are assumed to be measurable, the compound function $\omega \mapsto f_{L, \mathbf{D}_n(\omega), \lambda_{\mathbf{D}_n(\omega)}}$ is again measurable.

In order to apply the functional delta-method van der Vaart and Wellner (1996, Theorem 3.9.4), note that $\ell_\infty(\mathcal{G})$ and H are Banach spaces. Recall from Lemma 5.9 that $\mathbb{F}_{\mathbf{D}_n} : \Omega \to B_S$, $\omega \mapsto \mathbb{F}_{\mathbf{D}_n(\omega)}$ is the random map where $\mathbb{F}_{\mathbf{D}_n(\omega)}$ is that element of B_S which corresponds to the empirical distribution of $\mathbf{D}_n(\omega) = \big((X_1(\omega), Y_1(\omega)), \dots, (X_n(\omega), Y_n(\omega)) \big)$. That is,

$$\mathbb{F}_{\mathbf{D}_n(\omega)} \;:\; \mathcal{G} \;\to\; \mathbb{R} \,, \qquad g \;\mapsto\; \frac{1}{n} \sum_{i=1}^{n} g \big(X_i(\omega), Y_i(\omega) \big) \,.$$

Define

$$F_0 := \iota^{-1}(P) \qquad \text{and} \qquad \xi_n := \frac{\lambda_0}{\lambda_{\mathbf{D}_n}} \mathbb{F}_{\mathbf{D}_n} \,.$$

37

Then, Lemma 5.9 yields

$$\sqrt{n}\bigl(\mathbb{F}_{\mathbf{D}_n} - F_0\bigr) \;\rightsquigarrow\; \mathbb{G} \qquad \text{in} \quad \ell_\infty(\mathcal{G})$$

where $\mathbb{G} : \Omega \to \ell_\infty(\mathcal{G})$ is a tight Borel-measurable Gaussian process which takes it values in B_0. Furthermore,

$$\int \mathbb{G}(\omega)(g)\, Q(d\omega) \;=\; 0 \qquad \forall\, g \in \mathcal{G}\,; \tag{61}$$

see van der Vaart and Wellner (1996, p. 81f). According to van der Vaart and Wellner (1996, p. 16f), $\mathbb{G}$ is also separable (which is important in order to apply Slutsky's lemma for Banach space valued random maps below). Note that $\sqrt{n}\bigl(\lambda_{\mathbf{D}_n} - \lambda_0\bigr) \to 0$ in probability implies $\lambda_0/\lambda_{\mathbf{D}_n} \to 1$ and $\sqrt{n}\bigl(\lambda_{\mathbf{D}_n} - \lambda_0\bigr)/\lambda_{\mathbf{D}_n} \to 0$ in probability; see e.g. van der Vaart (1998, Theorems 2.3 and 2.7vi). Hence, it follows from Slutsky's lemma (van der Vaart and Wellner, 1996, p. 32) that

$$\sqrt{n}\bigl(\xi_n - F_0\bigr) \;=\; \sqrt{n}\bigl(\mathbb{F}_{\mathbf{D}_n} - F_0\bigr) \cdot \frac{\lambda_0}{\lambda_{\mathbf{D}_n}} + \frac{\sqrt{n}\bigl(\lambda_{\mathbf{D}_n} - \lambda_0\bigr)}{\lambda_{\mathbf{D}_n}} \;\rightsquigarrow\; \mathbb{G}$$

in $\ell_\infty(\mathcal{G})$. Then, applying the delta-method (van der Vaart and Wellner, 1996, Theorem 3.9.4) yields

$$\sqrt{n}\bigl(f_{L,\mathbf{D}_n,\lambda_{\mathbf{D}_n}} - f_{L,P,\lambda_0}\bigr) \;\overset{(18)}{=}\; \sqrt{n}\bigl(S(\xi_n) - S(F_0)\bigr) \;\rightsquigarrow\; S'_{F_0}(\mathbb{G})\,.$$

Since S'_{F_0} is a continuous linear operator and $\mathbb{G}$ is a tight Borel-measurable Gaussian process, $S'_{F_0}(\mathbb{G})$ is Gaussian as well; see, e.g., van der Vaart and Wellner (1996, §3.9.2). Since H is a complete and separable metric space, $S'_{F_0}(\mathbb{G})$ is tight; see e.g. Dudley (2002, Theorem 11.5.4).

It follows from (61) and Lemma 5.10 that $S'_{F_0}(\mathbb{G})$ has mean zero. $\qquad\square$

Proof of Theorem 3.2: It follows from Lemma 5.4 that the risk functional $\mathcal{R}_{L,P}$ is Hadamard-differentiable in H tangentially to H; the derivative of $\mathcal{R}_{L,P}$ in $f \in H$ is the continuous linear operator

$$\mathcal{R}'_{L,P;f} \;:\; H \;\to\; \mathbb{R}, \qquad h \mapsto \Bigl\langle \int L'_f \Phi\, dP\,, h \Bigr\rangle_H\,.$$

According to Theorem 3.1, $\sqrt{n}\bigl(f_{L,\mathbf{D}_n,\lambda_{\mathbf{D}_n}} - f_{L,P,\lambda_0}\bigr) \rightsquigarrow \mathbb{H}$ where $\mathbb{H} : \Omega \to H$ is a tight Borel-measurable Gaussian process which has zero-mean and does

not depend on $\lambda_{\mathbf{D}_n}$ but only on λ_0. Then, it follows from the delta-method (van der Vaart and Wellner, 1996, Theorem 3.9.4) that

$$\sqrt{n}\big(\mathcal{R}_{L,P}(f_{L,\mathbf{D}_n,\lambda_{\mathbf{D}_n}}) - \mathcal{R}_{L,P}(f_{L,P,\lambda_0})\big) \ \rightsquigarrow \ \mathcal{R}'_{L,P;f_{L,P,\lambda_0}}(\mathbb{H}) \ .$$

Since $\mathcal{R}'_{L,P;f_{L,P,\lambda_0}}$ is a continuous linear operator, and $\mathbb{H}$ is Gaussian, the (real valued) random variable $\mathcal{R}'_{L,P;f_{L,P,\lambda_0}}(\mathbb{H})$ is normally distributed; see e.g. van der Vaart and Wellner (1996, §3.9.2). Therefore, it only remains to prove that the mean of $\mathcal{R}'_{L,P;f_{L,P,\lambda_0}}(\mathbb{H})$ is equal to 0. This follows from

$$\mathbb{E}\mathcal{R}'_{L,P;f_{L,P,\lambda_0}}(\mathbb{H}) \ = \ \mathbb{E}\Big\langle \int L'_{f_{L,P,\lambda_0}} \Phi \, dP \, , \, \mathbb{H}\Big\rangle_H \ = \ 0$$

as $\mathbb{H} : \Omega \to H$ has zero-mean. $\qquad\qquad\qquad\qquad\qquad\qquad\square$

References

R. Akerkar. *Nonlinear functional analysis.* Narosa Publishing House, New Delhi, 1999.

H. Bauer. *Measure and integration theory.* Walter de Gruyter & Co., Berlin, 2001.

A. Berlinet and C. Thomas-Agnan. *Reproducing kernel Hilbert spaces in probability and statistics.* Kluwer Academic Publishers, Boston, MA, 2004.

G. Blanchard, O. Bousquet, and P. Massart. Statistical performance of support vector machines. *The Annals of Statistics*, 36(2):489–531, 2008.

N. Bourbaki. *Integration. I. Chapters 1–6.* Springer-Verlag, Berlin, 2004. Translated from the 1959, 1965 and 1967 French originals by Sterling K. Berberian.

A. Caponnetto and E. De Vito. Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7(3):331–368, 2007.

A. Christmann and I. Steinwart. On robustness properties of convex risk minimization methods for pattern recognition. *Journal of Machine Learning Research*, 5:1007–1034, 2004.

A. Christmann and I. Steinwart. Consistency and robustness of kernel-based regression in convex risk minimization. *Bernoulli*, 13(3):799–819, 2007.

A. Christmann and A. Van Messem. Bouligand derivatives and robustness of support vector machines for regression. *Journal of Machine Learning Research*, 9:915–936, 2008.

O. Dekel, S. Shalev-Shwartz, and Y. Singer. Smooth ϵ-insensitive regression by loss symmetrization. *Journal of Machine Learning Research*, 6:711–741, 2005.

Z. Denkowski, S. Migórski, and N. Papageorgiou. *An introduction to nonlinear analysis: Theory.* Kluwer Academic Publishers, Boston, 2003.

R. Dudley. *Real analysis and probability.* Cambridge University Press, Cambridge, 2002. Revised reprint of the 1989 original.

N. Dunford and J. Schwartz. *Linear operators. I. General theory.* Wiley-Interscience Publishers, New York, 1958.

X. Fernique. Intégrabilité des vecteurs gaussiens. *Comptes Rendus Hebdomadaires des Séances de l'Académie des Sciences. Séries A et B*, 270: A1698–A1699, 1970.

D. Griffel. *Applied functional analysis.* Dover Publications Inc., Mineola, NY, 2002. Revised reprint of the 1985 edition.

J. Hoffmann-Jørgensen. *Probability with a view toward statistics. Vol. I.* Chapman & Hall, New York, 1994.

B. Jiang, X. Zhang, and T. Cai. Estimating the confidence interval for prediction errors of support vector machine classifiers. *Journal of Machine Learning Research*, 9:521–540, 2008.

J.-Y. Koo, Y. Lee, Y. Kim, and C. Park. A Bahadur representation of the linear support vector machine. *Journal of Machine Learning Research*, 9: 1343–1368, 2008.

D. Luenberger. *Optimization by vector space methods.* John Wiley & Sons Inc., New York, 1969.

R. E. Megginson. *An introduction to Banach space theory.* Springer-Verlag, New York, 1998.

S. Mendelson and J. Neeman. Regularization in kernel learning. *The Annals of Statistics*, 38(1):526–565, 2010.

H. Satô. A remark on Landau-Shepp's theorem. *Sankhyā (Statistics). The Indian Journal of Statistics. Series A*, 33:227–228, 1971.

M. Schechter. *An introduction to nonlinear analysis*. Cambridge University Press, Cambridge, 2004.

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

I. Steinwart and A. Christmann. *Support vector machines*. Springer, New York, 2008.

I. Steinwart and C. Scovel. Fast rates for support vector machines using Gaussian kernels. *The Annals of Statistics*, 35(2):575–607, 2007.

I. Steinwart, D. Hush, and C. Scovel. Optimal rates for regularized least squares regression. *Proceedings of the 22nd Conference on Learning Theory (COLT 2009)*, 2009.

A. B. Tsybakov. Optimal aggregation of classifiers in statistical learning. *The Annals of Statistics*, 32(1):135–166, 2004.

A. van der Vaart. *Asymptotic statistics*. Cambridge University Press, Cambridge, 1998.

A. van der Vaart and J. Wellner. *Weak convergence and empirical processes. With applications to statistics*. Springer, New York, 1996.

V. N. Vapnik. *Statistical learning theory*. John Wiley & Sons, New York, 1998.

D.-X. Zhou. Capacity of reproducing kernel spaces in learning theory. *IEEE Transactions on Information Theory*, 49(7):1743–1752, 2003.