Asymptotic Confidence Sets for General Nonparametric Regression and Classification by Regularized Kernel Methods

Robert Hable*
Department of Mathematics
University of Bayreuth
D-95440 Bayreuth, Germany

Abstract

Regularized kernel methods such as, e.g., support vector machines and least-squares support vector regression constitute an important class of standard learning algorithms in machine learning. Theoretical investigations concerning asymptotic properties have mainly focused on rates of convergence during the last years but there are only very few and limited (asymptotic) results on statistical inference so far. As this is a serious limitation for their use in mathematical statistics, the goal of the article is to fill this gap. Based on asymptotic normality of many of these methods, the article derives a strongly consistent estimator for the unknown covariance matrix of the limiting normal distribution. In this way, we obtain asymptotically correct confidence sets for $\psi(f_{P,\lambda_0})$ where $f_{P,\lambda_0}$ denotes the minimizer of the regularized risk in the reproducing kernel Hilbert space $H$ and $\psi : H \rightarrow \mathbb{R}^m$ is any Hadamard-differentiable functional. Applications include (multivariate) pointwise confidence sets for values of $f_{P,\lambda_0}$ and confidence sets for gradients, integrals, and norms.

Keywords: Asymptotic confidence sets, asymptotic normality, least-squares support vector regression, regularized kernel methods, support vector machines.

MSC: 62G08, 62G15

1 Introduction

Regularized kernel methods constitute an important class of standard learning algorithms in machine learning theory. The prominent learning algo-
rithms support vector machine (for classification) and (least-squares) support vector regression (for regression) also belong to this class; see, e.g., [28], [17], and [21]. While these methods are standard in machine learning theory and are widely applied, their propagation in mathematical statistics is still limited. This is partly due to the fact that there is a lack of results on statistical inference for these methods so far. In machine learning theory, the goal in a supervised learning problem is to find a “good” predictor $f : \mathcal{X} \rightarrow \mathcal{Y}$ which maps the observed value $x \in \mathcal{X}$ of an input variable $X$ to a prediction of the unobserved value $y \in \mathcal{Y}$ of an output variable. A learning algorithm $S_n$ is a mapping which maps a set $D_n$ of observed training data $(x_1, y_1), \ldots, (x_n, y_n)$ to a predictor $f_{D_n}$. In mathematical statistics, such a problem would rather be called a (nonparametric) regression (or a classification) problem, $X$ is the covariate, $Y$ is the response variable, $S_n$ is an estimator and $S_n(D_n) = f_{D_n}$ is the estimated function. In both contexts, a “good” predictor/estimate $f$ has a small expected loss (also called risk) $\mathcal{R}_{L,P}(f) = \int L(x, y, f(x)) p(dx, dy)$ where $L$ is a “suitable” loss function and $p$ is the joint distribution of $X$ and $Y$. However, depart form the different terminology, there is also a real difference: In machine learning, the goal is to find any predictor $f$ which has a small risk and, accordingly, a learning algorithm $S_n$ should be risk-consistent, i.e., $\mathcal{R}_{L,P}(S_n(D_n))$ converges to the infimal risk $\inf_f \mathcal{R}_{L,P}(f)$ for $n \rightarrow \infty$. In statistics, it is e.g. common to make a signal plus noise assumption such as $Y = f_0(X) + g(X) \varepsilon$ and the goal is to estimate the unknown regression function $f_0$. Under suitable assumptions, $f_0$ minimizes $\mathcal{R}_{L,P}(f)$ in certain sets of functions $f$. While, in machine learning, one is mainly interested in minimizing the risk, in statistics, one is mainly interested in the minimizer and, accordingly, an estimator $S_n$ should be consistent in the sense that $S_n$ converges to $f_0$. For statistical inference, it is also crucial to have estimates for the error of the estimator, e.g., in order to obtain confidence sets or hypothesis tests. While consistency results for the risk, e.g. [19], [29], [20], and [5], and also for the functions, e.g. [22] §3 and [11] Cor. 3.7, are well-known for regularized kernel methods, there are only very few and limited results concerning statistical inference. In order to fill this gap, asymptotic confidence sets for a wide class of regularized kernel methods are developed in the following. This is possible now because [10] derives asymptotic normality of these methods and, based on this result, estimating the error of the estimate gets tractable. Let $f_{D_n, \Lambda_n}$ be the (nonparametric) estimate obtained by a regularized kernel method ($\Lambda_n$ is a data-driven regularization parameter), fix any $\lambda_0 \in (0, \infty)$, and let $f_{P, \lambda_0}$ be
the minimizer of the regularized problem
\[ f \mapsto \mathcal{R}_{L,P}(f) + \lambda_0 \|f\|_H^2 \] (1)
in the function space \( H \) (a so-called reproducing kernel Hilbert space). According to [10, Theorem 3.1], under some assumptions, the sequence of function-valued random variables \( \sqrt{n}(f_{D_n,\Lambda_n} - f_{P,\lambda_0}) \) weakly converges to a Gaussian process in \( H \). As a consequence, for differentiable functions \( \psi : H \to \mathbb{R}^m \), it follows that
\[ \sqrt{n}\left( \psi(f_{D_n,\Lambda_n}) - \psi(f_{P,\lambda_0}) \right) \xrightarrow{D} \mathcal{N}(0, \Sigma_P). \]
In order to obtain asymptotic confidence sets, e.g., for the vector of values \( f_{P,\lambda_0}(x_j), j \in \{1, \ldots, m\} \), one has to choose \( \psi(f) = (f(x_1), \ldots, f(x_m)) \) and it only remains to estimate the asymptotic covariance matrix \( \Sigma_P \). The derivation of a consistent estimator is not a trivial task and is the main issue of the article. However, pointwise confidence sets for the true values of \( f_{P,\lambda_0} \) are not the only possibility to directly apply the results of the article. We also obtain confidence sets e.g. for integrals of \( f_{P,\lambda_0} \) (choose \( \psi(f) = \int_B f \, d\lambda \)) or for the differential of \( f_{P,\lambda_0} \) in a point \( x_0 \) (choose \( \psi(f) = \partial f(x_0) \)) and many others. Essentially, it is only needed that \( \psi \) takes its values in \( \mathbb{R}^m \) for any \( m \in \mathbb{N} \) and is suitably differentiable.
Note that we are only able to derive asymptotic confidence sets for the unknown solution \( f_{P,\lambda_0} \) of the regularized problem (1). Of course, it would be desirable to obtain asymptotic confidence sets for the minimizer of the unregularized risk \( \mathcal{R}_{L,P} \). However, in our completely nonparametric setting \( (P \text{ is totally unknown}) \), this would require a uniform rate of convergence of the learning algorithm/estimator to the minimizer of \( \mathcal{R}_{L,P} \) (if such a minimizer exists at all) and it is well-known from the no-free-lunch theorem [8] that such a uniform rate of convergence does not exist. That is, similar results for the minimizer of the unregularized \( \mathcal{R}_{L,P} \) can only be obtained under substantial assumptions on the unknown distribution \( P \).
Accordingly, the approach in the present article which focuses on applications in statistical inference considerably differs from the approach common in machine learning theory which focuses on (as fast as possible) rates of convergence of the risk, e.g., [23], [3], [2], [24], [16]. This approach considers large classes \( \mathcal{P} \) of probability measures for which learning rates, e.g., in the form
\[ P^n \left( \mathcal{R}_{L,P}(f_{D_n,\lambda_n}) - \inf_f \mathcal{R}_{L,P}(f) \leq c_{P,\delta} \cdot n^{-\beta} \right) \geq 1 - \delta, \]
exist and where the rate of convergence \( \beta > 0 \) does not depend on \( P \) and the infimum is taken over all measurable functions \( f : \mathcal{X} \to \mathbb{R} \). Such learning
rates are an important tool in order to compare theoretical properties of different learning algorithms. However, these results cannot be applied off-hand for statistical inference in real applications because the constant $c_{P,\delta}$ is usually unknown. Furthermore, the focus lies on maximizing $\beta$ which, typically, results in an increase of the constant $c_{P,\delta}$ so that the bound $c_{P,\delta} \cdot n^{-\beta}$ might be large for ordinary sample sizes $n$. In addition, whether a probability measure belongs to $P$ is often subject to assumptions which are hard to communicate to practitioners and to be satisfactorily checked or made plausible in applications. A common assumption is, e.g., Tsybakov’s noise assumption [25, p. 138].

The present article derives asymptotic confidence sets for $\psi(f_{P,\lambda_0})$ based on the asymptotic normality results of [10]. So far, there are only very few publications which are concerned with statistical inference for regularized kernel methods. In the special case of classification by use of the hinge loss and linear SVMs (i.e. linear kernel), asymptotic normality of the coefficients of the linear SVM is shown in [15] under a number of regularity assumptions (e.g. existence of continuous densities). Though this could yield an alternative way of deriving asymptotic confidence sets in this special case, this has not been done so far. In the special case of classification by use of the hinge loss and SVMs with finite-dimensional kernels (i.e. a parametric setting), [13] shows asymptotic normality of the prediction error estimators and derive confidence intervals for the prediction error of the empirical SVM. In the special case of regression by use of least-squares support vector regression, [6] proposes approximate confidence intervals for the regression function whose derivation is partly based on heuristics; it is not documented whether these intervals approximately hold the intended confidence level in simulated examples.

In the following Section 2 some basics of regularized kernel methods are recalled. The main part of the article, Section 3 consists of two subsections: Subsection 3.1 derives an asymptotically consistent estimator of $\Sigma_P$ and asymptotic confidence intervals; Subsection 3.2 shows how the calculation of the estimator can be done in a computationally tractable way. All proofs are given in the appendix.

## 2 Regularized Kernel Methods

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-$\sigma$-algebra $\mathcal{B}(\mathcal{Y})$. The
Borel-$\sigma$-algebra of $\mathcal{X} \times \mathcal{Y}$ is denoted by $\mathfrak{B}(\mathcal{X} \times \mathcal{Y})$. Let

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

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

$$D_n := ((X_1, Y_1), \ldots, (X_n, Y_n)) \quad \forall n \in \mathbb{N}.$$ 

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 $(x, y) \in \mathcal{X} \times \mathcal{Y}$. Furthermore, a loss function $L$ is called $P$-integrable Nemitski loss function of order $p \in [1, \infty)$ if there is a $P$-integrable function $b : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ and a constant $c \in (0, \infty)$ such that

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

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

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

The goal is to estimate a function $f : \mathcal{X} \to \mathbb{R}$ which minimizes this risk. The estimates obtained from regularized kernel methods are elements of so-called reproducing kernel Hilbert spaces (RKHS) $H$. An 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. [17], [1], [21], or [12] for details about these concepts.

Let $H$ be such an RKHS. Then, the regularized risk of an element $f \in H$ is defined to be

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

where $\lambda \in (0, \infty)$. An element $f \in H$ is denoted by $f_{P, \lambda}$ if it minimizes the regularized risk in $H$. That is,

$$\mathcal{R}_{L, P}(f_{P, \lambda}) + \lambda \|f_{P, \lambda}\|^2_H = \inf_{f \in H} \left( \mathcal{R}_{L, P}(f) + \lambda \|f\|^2_H \right).$$
The estimator is defined by

\[ S_n : (X \times Y)^n \times (0, \infty) \to H, \quad (D_n, \lambda) \mapsto f_{D_n, \lambda} \]

where \( f_{D_n, \lambda} \) is that function \( f \in H \) which minimizes

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

in \( H \) for \( D_n = ((x_1, x_2), \ldots, (x_n, y_n)) \in (X \times Y)^n \). The estimate \( f_{D_n, \lambda} \) uniquely exists for every \( \lambda \in (0, \infty) \) and every data-set \( D_n \in (X \times Y)^n \) if \( t \mapsto L(x, y, t) \) is convex for every \( (x, y) \in X \times Y \).

In the article, the symbol \( \rightsquigarrow \) denotes weak convergence of probability measures or random variables.

3 Asymptotic Confidence Intervals

3.1 Theory

The derivation of asymptotic confidence sets is based on the result in [10] that, under some assumptions,

\[
\sqrt{n}(f_{D_n, \Lambda_n} - f_{P, \lambda_0}) \rightsquigarrow \mathbb{H}_P \quad \text{in } H
\]

where \( \mathbb{H}_P \) is a mean-zero Gaussian process in \( H \) and \( \Lambda_n \) is a random regularization parameter (e.g. data-driven). Therefore, the same assumptions as in [10] are needed; they are collocated in the following:

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

\[ L : X \times Y \times \mathbb{R} \to [0, \infty), \quad (x, y, t) \mapsto L(x, y, t) \]

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

\[
L'(x, y, t) := \frac{\partial L}{\partial t}(x, y, t) \quad \text{and} \quad L''(x, y, t) := \frac{\partial^2 L}{\partial^2 t}(x, y, t)
\]
exist for every \((x, y, t) \in X \times Y \times \mathbb{R}\). Assume that the maps
\[
(x, y, t) \mapsto L'(x, y, t) \quad \text{and} \quad (x, y, t) \mapsto L''(x, y, t)
\]
are continuous. Furthermore, assume that for every \(a \in (0, \infty)\), there is a \(b'_a \in L_2(P)\) and a constant \(b''_a \in [0, \infty)\) such that, for every \((x, y) \in X \times Y\),
\[
\sup_{t \in [-a, a]} |L'(x, y, t)| \leq b'_a(x, y) \quad \text{and} \quad \sup_{t \in [-a, a]} |L''(x, y, t)| \leq b''_a. \tag{3}
\]

These assumptions are relatively mild. 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). Though assuming differentiability of the loss function is an obvious restriction (as it does not cover some of the most popular loss functions as hinge, epsilon-insensitive, and pinball), this assumption is not based on any unknown entity such as the model distribution \(P\). Therefore, a practitioner can a priori meet this requirement by a suitable choice of the loss function (e.g. the least-squares loss for regression, the logistic loss for classification (or smoothed versions of hinge, epsilon-insensitive, and pinball)). This is contrary to the assumptions common in order to establish rates of convergence to the infimal risk. Typically, the assumptions used there depend on the unknown \(P\) so that they can hardly be checked in applications and are mathematically involved so that they can hardly be communicated to practitioners. In Assumption 3.1, the only assumptions on \(P\) are integrability assumptions, which are natural as such assumptions are necessary even for ordinary central limit theorems. Explicit examples where Assumption 3.1 is fulfilled are given in Section 4.

Under these assumptions, we have asymptotic normality:

**Theorem 3.2** [10, Theorem 3.1] Let Assumption 3.1 be fulfilled. Then, for every \(\lambda_0 \in (0, \infty)\), there is a tight, Borel-measurable Gaussian process
\[
\mathbb{H}_P : \Omega \rightarrow H, \quad \omega \mapsto \mathbb{H}_P(\omega)
\]
such that,
\[
\sqrt{n}(f_{D_n,\Lambda_n} - f_{P,\lambda_0}) \overset{\text{law}}{\rightarrow} \mathbb{H}_P \quad \text{in } H \tag{4}
\]
for every Borel-measurable sequence of random regularization parameters \(\Lambda_n\) with
\[
\sqrt{n}(\Lambda_n - \lambda_0) \xrightarrow{n \to \infty} 0 \quad \text{in probability}.
\]
The Gaussian process \(\mathbb{H}_P\) is zero-mean; i.e., \(\mathbb{E}(f, \mathbb{H}_P)_H = 0\) for every \(f \in H\).
Recall that a map \( \psi : H \to \mathbb{R}^m \) is Hadamard differentiable at some \( f_0 \in H \) if and only if there exists a \( \psi'_{f_0} = (\psi'_{f_0,1}, \ldots, \psi'_{f_0,m}) \in H^m \) such that, for every sequence \( t_\ell \downarrow 0 \) in \( \mathbb{R} \), and for every sequence \( h_\ell \to h \) in \( H \),
\[
\lim_{\ell \to \infty} \left| \frac{\psi(f_0 + t_\ell h_\ell) - \psi(f_0) - \langle \psi'_{f_0}, h \rangle_H}{t_\ell} \right| = 0.
\]
The element \( \psi'_{f_0} \in H^m \) is called derivative of \( \psi \) at \( f_0 \). For \( h \in H \) and \( \psi'_{f_0} = (\psi'_{f_0,1}, \ldots, \psi'_{f_0,m}) \in H^m \), the expression \( \langle \psi'_{f_0}, h \rangle_H \) denotes the element of \( \mathbb{R}^m \) whose components are given by \( \langle \psi'_{f_0,j}, h \rangle \), \( j \in \{1, \ldots, m\} \).

By a routine application of the functional delta method [27, Theorem 3.9.4], we get the following corollary:

**Corollary 3.3** Let Assumption 3.1 be fulfilled, let \( \lambda_0 \in (0, \infty) \), and let \( \psi : H \to \mathbb{R}^m \) be Hadamard-differentiable in \( f_{P,\lambda_0} \) with derivative \( \psi'_{f_{P,\lambda_0}} \).

Then, there is a covariance matrix \( \Sigma_P \in \mathbb{R}^{m \times m} \) such that, for every Borel-measurable sequence of random regularization parameters \( \Lambda_n \) with \( \sqrt{n} (\Lambda_n - \lambda_0) \xrightarrow{n \to \infty} 0 \) in probability,
\[
\sqrt{n} \left( \psi(f_{D_n,\Lambda_n}) - \psi(f_{P,\lambda_0}) \right) \xrightarrow{n \to \infty} \mathcal{N}_m(0, \Sigma_P).
\]
The limit \( \mathcal{N}_m(0, \Sigma_P) \) is equal to the distribution of \( \langle \psi'_{f_{P,\lambda_0}}, \Pi_P \rangle_H \) where \( \Pi_P \) is given by (3).

Accordingly, in order to derive asymptotic confidence intervals, the main issue which remains to be solved is to calculate or rather consistently estimate the covariance matrix \( \Sigma_P \). In principle, \( \Sigma_P \) is completely known if \( P \) is known – as can be seen from the proof of Theorem 3.2 given in [10]. This suggests to estimate \( \Sigma_P \) by a plug-in estimator where \( P \) is replaced by the empirical measure \( \mathbb{P}_{D_n} \). However, this is a challenging task because \( \Pi_P \) is given by \( \Pi_P = S'_P(G_P) \) where \( S'_P \) is a (complicated) continuous linear operator and \( G_P \) is a random variable which takes its values in a large function space. Hence, calculating \( \Sigma_P = \text{Cov}(\langle \psi'_{f_{P,\lambda_0}}, \Pi_P \rangle_H) \) means to calculate an integral with respect to a measure on that function space. Fortunately, this can be avoided as follows from Prop. 3.4. There, \( \Sigma_P \) is specified in a way which is more accessible to a plug-in estimator. The consistency of the resulting plug-in estimator is given in Theorem 3.6. Note that \( \Sigma_P \) can be degenerated to 0 in Corollary 3.3. In order to derive asymptotic confidence sets, degeneracy has to be excluded by adding additional assumptions (Assumption 3.8) below.
Proposition 3.4  Let Assumption 3.1 be fulfilled, let $\lambda_0 \in (0, \infty)$, and let $\psi : H \to \mathbb{R}^m$ be Hadamard-differentiable in $f_{P, \lambda_0}$ with derivative $\psi'_{f_{P, \lambda_0}}$. Define

$$g_{P, \lambda_0} : X \times Y \to \mathbb{R}, \quad (x, y) \mapsto -L'(x, y, f_{P, \lambda_0}(x))\langle \psi'_{f_{P, \lambda_0}}, K_P^{-1}(\Phi(x)) \rangle_H$$

where $K_P$ denotes the continuous linear operator defined in (19). Then, the covariance matrix $\Sigma_P$ in Corollary 3.3 is equal to

$$\Sigma_P = \text{Cov}(g_{P, \lambda_0}(X_1, Y_1)) \quad \text{(5)}$$

It follows from Prop. 3.4 that $\Sigma_P$ could be estimated by the standard covariance estimator for the $\mathbb{R}^m$-valued i.i.d. random variables

$$g_{P, \lambda_0}(X_1, Y_1), \ldots, g_{P, \lambda_0}(X_n, Y_n)$$

if $P$ was known. However, as $P$ is unknown, we have to replace $P$ by the empirical measure and $\psi'_{f_{P, \lambda_0}}$ by an estimator $\psi'_{D_n, \Lambda_n}$ of $\psi'_{f_{P, \lambda_0}}$. Then, we may estimate $\Sigma_P$ by the non-i.i.d. random variables

$$g_{D_n, \Lambda_n}(X_1, Y_1), \ldots, g_{D_n, \Lambda_n}(X_n, Y_n)$$

where

$$g_{D_n, \Lambda_n}(x, y) = -L'(x, y, f_{D_n, \Lambda_n}(x))\langle \psi'_{D_n, \Lambda_n}, K_{D_n, \Lambda_n}^{-1}(\Phi(x)) \rangle_H \quad \text{(6)}$$

and $K_{D_n, \Lambda_n}(\omega) : H \to H$ is the continuous linear operator given by

$$K_{D_n, \Lambda_n}(f) = 2\Lambda_n f + \frac{1}{n} \sum_{i=1}^{n} L''(X_i, Y_i, f_{D_n, \Lambda_n}(X_i)) f(X_i) \Phi(X_i) \quad \text{(7)}$$

for every $f \in H$. The following theorem states that the resulting plug-in covariance estimator is strongly consistent. It is also shown that the estimator is measurable. This is not obvious as the proof of Theorem 3.2 is based on the theory of empirical processes and the map $D_n \mapsto \mathbb{P}_{D_n}$ (which maps a set of data to its empirical measure as an element of a certain function space) is typically not Borel-measurable; see e.g. [27, §1.1].

Assumption 3.5  Let $\psi : H \to \mathbb{R}^m$ be Hadamard-differentiable at $f_{P, \lambda_0}$ with derivative $\psi'_{f_{P, \lambda_0}}$ and let $\psi'_{D_n, \Lambda_n}$ be an estimator of $\psi'_{f_{P, \lambda_0}}$ which is strongly consistent, i.e.,

$$\|\psi'_{D_n, \Lambda_n} - \psi'_{f_{P, \lambda_0}}\|_{H^m} \xrightarrow{n \to \infty} 0 \quad \text{almost surely.} \quad \text{(8)}$$
Theorem 3.6 Let Assumption 3.1 and Assumption 3.5 be fulfilled. Fix \(\lambda_0 \in (0, \infty)\) and let \(\Sigma_P \in \mathbb{R}^{m \times m}\) be the covariance matrix in Corollary 3.3. Then, for every Borel-measurable sequence of random regularization parameters \(\Lambda_n\) with

\[
\sqrt{n}(\Lambda_n - \lambda_0) \xrightarrow{n \to \infty} 0 \quad \text{almost surely},
\]
the estimator

\[
\hat{\Sigma}_n(D_n, \Lambda_n) = \frac{1}{n} \sum_{i=1}^{n} \left( \tilde{g}_{D_n, \Lambda_n}(X_i, Y_i) \right) \cdot \left( \tilde{g}_{D_n, \Lambda_n}(X_i, Y_i) \right)^T
\]

with

\[
\tilde{g}_{D_n, \Lambda_n}(X_i, Y_i) := g_{D_n, \Lambda_n}(X_i, Y_i) - \frac{1}{n} \sum_{j=1}^{n} g_{D_n, \Lambda_n}(X_j, Y_j) \quad \forall i \in \{1, \ldots, n\}
\]

is measurable and strongly consistent, i.e.,

\[
\hat{\Sigma}_n(D_n, \Lambda_n) \xrightarrow{n \to \infty} \Sigma_P \quad \text{almost surely.}
\]

The following remark specifies a natural candidate for an estimator of \(\psi'_f \psi_{f, \lambda_0}\); the proof is given in the appendix.

Remark 3.7 If \(\psi\) is Hadamard-differentiable at every \(f \in H\) with derivative \(\psi'_f\) and if \(f \mapsto \psi'_f\) is continuous, then Assumption 3.5 is fulfilled for the estimator

\[
\psi'_{D_n, \Lambda_n} := \psi'_{D_n, \Lambda_n}
\]

– provided that \(\Lambda_n\) converges to \(\lambda_0\) almost surely.

The calculation of the estimator \(\hat{\Sigma}_n(D_n, \Lambda_n)\) for a given data set is an issue of its own because it is burdened by the fact that we have to solve \(n\) equations

\[
K_{D_n, \Lambda_n}(f_i) = \Phi(X_i), \quad i \in \{1, \ldots, n\},
\]
in the typically infinite dimensional function space \(H\). As we will see in Subsection 3.2 below, this problem can satisfactorily be solved (Prop. 3.10). In fact, these equations can be solved jointly, essentially by calculating the Moore-Penrose pseudoinverse of an \(n \times n\)-matrix once only.

In order to derive asymptotic confidence intervals based on Corollary 3.3, it is desirable that the covariance matrix \(\Sigma_P\) has full rank. Lemma 7.2 in the Appendix yields that this can be achieved by the following two weak conditions:
Assumption 3.8 Assume that, for $P_X(dx)$ - a.e. $x \in X$,

$$\exists y_1, y_2 \in \text{supp}(P(dy|x)) \text{ s.t. } L'(x, y_1, f_{P, \lambda_0}(x)) \neq L'(x, y_2, f_{P, \lambda_0}(x)).$$

(9)

For every $j \in \{1, \ldots, m\}$, let $\psi'_{f_{P, \lambda_0}, j} \in H$ denote the $j$-th component of $\psi'_{f_{P, \lambda_0}}$ and assume that

$$\psi'_{f_{P, \lambda_0}, 1}, \ldots, \psi'_{f_{P, \lambda_0}, m} \text{ are linearly independent on } \text{supp}(P_X).$$

(10)

Due to continuity, Assumption (10) can be reformulated to the following condition:

$$a^T \psi'_{f_{P, \lambda_0}} = 0 \quad P_X\text{-a.s. for some } a \in \mathbb{R}^m \quad \Rightarrow \quad a = 0.$$  

(11)

As we will see in the examples in the applications section, Assumption 3.8 indeed provides weak and simple conditions. E.g., in case of the least-squares loss or the logistic loss, Assumption (9) is equivalent to assuming that $P(dy|x)$ is not a Dirac measure.

From the above results and assumptions, it follows that

$$\sqrt{n} \cdot \hat{\Sigma}_n(D_n, \Lambda_n)^{-\frac{1}{2}} \left( \psi(f_{D_n, \Lambda_n}) - \psi(f_{P, \lambda_0}) \right) \overset{\text{d}}{\rightarrow} N_m(0, \text{Id}_{m \times m}).$$

so that we get elliptical confidence sets which are asymptotically correct:

Theorem 3.9 Let $\lambda_0 \in (0, \infty)$ and let Assumption 3.4, Assumption 3.5 and Assumption 3.8 be fulfilled. Let $\Lambda_n$ be a sequence of Borel-measurable random regularization parameters with

$$\sqrt{n}(\Lambda_n - \lambda_0) \xrightarrow{n \to \infty} 0 \quad \text{almost surely.}$$

Fix any $\alpha \in (0, 1)$, let $\chi^2_{m, \alpha}$ be the $(1 - \alpha)$th quantile of the chi-squared distribution with $m$ degrees of freedom and

$$C_{n, \alpha}(D_n, \Lambda_n) := \left\{ w \in \mathbb{R}^m \mid \left\| \hat{\Sigma}_n(D_n, \Lambda_n)^{-\frac{1}{2}} (w - \psi(f_{D_n, \Lambda_n})) \right\|^2_{\mathbb{R}^m} \leq \frac{\chi^2_{m, \alpha}}{n} \right\}.$$ 

Then,

$$Q\left( \psi(f_{P, \lambda_0}) \in C_{n, \alpha}(D_n, \Lambda_n) \right) \xrightarrow{n \to \infty} 1 - \alpha.$$ 

Note that the confidence set $C_{n, \alpha}(D_n, \Lambda_n)$ is an ellipsoid in $\mathbb{R}^m$ which is centered at $\psi(f_{D_n, \Lambda_n})$ and whose principal axes are given by

$$\frac{\sqrt{\chi^2_{m, \alpha} \gamma_1}}{n} \cdot v_1, \ldots, \frac{\sqrt{\chi^2_{m, \alpha} \gamma_m}}{n} \cdot v_m$$

where $\gamma_1, \ldots, \gamma_m$ are the eigenvalues and $v_1, \ldots, v_m$ are corresponding orthonormal eigenvectors of the matrix $\hat{\Sigma}_n(D_n, \Lambda_n)$.
3.2 Computation of Asymptotic Confidence Sets

The calculation of the estimator $\hat{\Sigma}_n(D_n, \Lambda_n)$ for a given data set is burdened by the fact that we have to solve every of the following $n$ equations

$$K_{D_n, \Lambda_n}(f_i) = \Phi(X_i), \quad i \in \{1, \ldots, n\},$$

in the typically infinite dimensional function space $H$. In particular for a large sample size $n$, this seems to be problematic. However, Prop. 3.10 below yields that the problem can essentially be reduced to the calculation of a single Moore-Penrose pseudoinverse of an $n \times n$-matrix after the following preparation: Let $D_n = ((x_1, y_1), \ldots, (x_n, y_n)) \in (X \times Y)^n$. Then, there is always a maximal subset $\{\Phi(x_{i_1}), \ldots, \Phi(x_{i_r})\}$ of $\{\Phi(x_1), \ldots, \Phi(x_n)\}$ such that $\Phi(x_{i_1}), \ldots, \Phi(x_{i_r})$ are linearly independent – i.e. $\{\Phi(x_{i_1}), \ldots, \Phi(x_{i_r})\}$ is a basis of the vector space spanned by $\Phi(x_1), \ldots, \Phi(x_n)$. Accordingly, for every $i \in \{1, \ldots, n\}$, there are $\beta_{1i}, \ldots, \beta_{ri} \in \mathbb{R}$ such that

$$\Phi(x_i) = \sum_{j=1}^{r} \beta_{ji} \Phi(x_{ij}). \quad (12)$$

Define

$$B_{D_n} = \left( \begin{array}{ccc} \beta_{11} & \cdots & \beta_{1n} \\ \vdots & \ddots & \vdots \\ \beta_{r1} & \cdots & \beta_{rn} \end{array} \right) \in \mathbb{R}^{r \times n}. \quad (13)$$

E.g., in case of a Gaussian RBF kernel, vectors $\Phi(x_{i_1}), \ldots, \Phi(x_{i_r})$ are linearly independent if and only if all $x_{i_1}, \ldots, x_{i_r}$ differ; see e.g. [17, Theorem 2.18]. Hence, in this case, finding $B_{D_n}$ only means to identify all ties in the covariates – and, if there are no such ties, $B_{D_n}$ is just the $n \times n$-identity matrix.

**Proposition 3.10** Let Assumption 3.1 be fulfilled. Fix any set of data $D_n = ((x_1, y_1), \ldots, (x_n, y_n)) \in (X \times Y)^n$ and any $\lambda \in (0, \infty)$. Define $B_{D_n}$ according to (12) and (13). Let $L''_{D_n, \lambda} \in \mathbb{R}^{n \times n}$ denote the diagonal matrix with diagonal entries

$$L''(x_1, y_1, f_{D_n, \lambda}(x_1)), \ldots, L''(x_n, y_n, f_{D_n, \lambda}(x_n)),$$

define the $n \times n$-matrix

$$A_{D_n, \lambda} = 2\lambda \cdot \text{Id}_{n \times n} + \frac{1}{n} \cdot L''_{D_n, \lambda} \cdot \begin{pmatrix} k(x_1, x_1) & \cdots & k(x_1, x_n) \\ \vdots & \ddots & \vdots \\ k(x_n, x_1) & \cdots & k(x_n, x_n) \end{pmatrix},$$

12
and let \((B_{D_n}A_{D_n,\lambda})^{-1}\) be the Moore-Penrose pseudoinverse of \(B_{D_n}A_{D_n,\lambda}\). Then, for every \(x \in \mathcal{X}\) and \(y \in \mathcal{Y}\),

\[
K_{D_n,\lambda}^{-1}(\Phi(x)) = \frac{1}{2\lambda} \Phi(x) + \sum_{i=1}^{n} \alpha_i(x) \Phi(x_i)
\]

and

\[
g_{D_n,\lambda}(x, y) = -L'(x, y, f_{D_n,\lambda}(x)) \cdot \left( \frac{1}{2\lambda} \psi'_{f_{D_n,\lambda}}(x) + \sum_{i=1}^{n} \alpha_i(x) \psi'_{f_{D_n,\lambda}}(x_i) \right)
\]

where

\[
\begin{pmatrix}
\alpha_1(x) \\
\vdots \\
\alpha_n(x)
\end{pmatrix} = -\frac{1}{2n\lambda} \cdot (B_{D_n}A_{D_n,\lambda})^{-1} B_{D_n} \cdot 
\begin{pmatrix}
L''(x_1, y_1, f_{D_n,\lambda}(x_1)) k(x_1, x) \\
\vdots \\
L''(x_n, x_n, f_{D_n,\lambda}(x_n)) k(x_n, x)
\end{pmatrix}.
\]

By use of this proposition, the calculation of the estimator \(\hat{\Sigma}_n(D_n, \lambda)\) is unproblematic. According to its definition, it is enough to calculate the values \(g_{D_n,\lambda}(x_i, y_i), i \in \{1, \ldots, n\}\), and, in order to do this, the matrices \(B_{D_n}\) and \(A_{D_n,\lambda}\) have to be defined and \((B_{D_n}A_{D_n,\lambda})^{-1}\) has to be calculated once only. Then, all values \(g_{D_n,\lambda}(x_i, y_i), i \in \{1, \ldots, n\}\), can simultaneously be calculated by matrix calculus. After that, it only remains to calculate the inverse of the matrix \(\hat{\Sigma}_n(D_n, \lambda)\) in order to calculate the elliptical confidence set. In order to obtain the principal axes of the ellipse, one only has to calculate an (orthonormal) eigendecomposition of \(\hat{\Sigma}_n(D_n, \lambda)\) instead.

### 4 Applications

In Subsection 3.1 a general scheme is developed how to derive asymptotic confidence sets for values \(\psi(f_{P,\lambda_0})\) of functionals \(\psi : H \to \mathbb{R}^m\). This general scheme is exemplified in a few possible applications from which it can also be seen that the assumptions made in Subsection 3.1 are moderate and, equally important, not mathematically involved so that they are comprehensible to practitioners.

**The input and the output space.** Let \(\mathcal{X} \subset \mathbb{R}^d\) be closed and bounded and let \(\mathcal{Y} \subset \mathbb{R}\) be closed. That is, the setting covers regression with \(\mathcal{Y} = \mathbb{R}\) and classification with \(\mathcal{Y} = \{-1, +1\}\) as well.

**The kernel \(k\).** Let \(\tilde{k} : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}\) be a kernel which is \(r\) - times continuously differentiable kernel where \(r > d/2\). Let \(k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}\) be the restriction
of $k$ on $\mathcal{X} \times \mathcal{X}$. Let $k \neq 0$. That is, every of the most common kernels can be chosen: a Gaussian RBF kernel, a polynomial kernel, the linear kernel, the exponential kernel, or sums and products of such kernels.

The loss function $L$. We exemplarily consider the following three settings:

(A) Regression with the least-squares loss: Let

$$L(x, y, t) = (y - t)^2 \quad \forall (x, y, t) \in \mathcal{X} \times \mathcal{Y} \times \mathbb{R}$$

and assume that $\mathbb{E}Y^4 < \infty$.

(B) Regression with the logistic loss: Fix a constant $\sigma > 0$ and define

$$L(x, y, t) = -\sigma \cdot \log \frac{4 \exp \left( \frac{y - t}{\sigma} \right)}{(1 + \exp \left( \frac{y - t}{\sigma} \right))^2} \quad \forall (x, y, t) \in \mathcal{X} \times \mathcal{Y} \times \mathbb{R}$$

and assume that $\mathbb{E}Y^2 < \infty$.

(C) Classification: Let $\mathcal{Y} = \{-1, +1\}$ and choose the least-squares loss

$$L(x, y, t) = (1 - yt)^2 \quad \forall (x, y, t) \in \mathcal{X} \times \mathcal{Y} \times \mathbb{R}$$

or the logistic loss

$$L(x, y, t) = \log \left(1 + \exp(y - t)\right) \quad \forall (x, y, t) \in \mathcal{X} \times \mathcal{Y} \times \mathbb{R}.$$ 

In every of these settings, Assumption 3.1 is fulfilled. Furthermore, (9) in Assumption 3.8 can be rewritten as

$$\text{Var}(Y | x) \neq 0 \quad \text{for } P_X(dx) \text{-a.e. } x \in \mathcal{X}. \quad (14)$$

If $\text{Var}(Y | x) = 0$ for some $x \in \mathcal{X}$, then $Y$ is deterministically fixed by $X = x$. Of course, $\text{Var}(Y | x) = 0$ for some $x \in \mathcal{X}$ can happen at most in case of heteroscedastic (or even more complicated) error terms. In addition to (9), the only remaining assumption is Assumption 3.8, which we have to take care of when choosing a functional $\psi$.

The regularization parameter $\Lambda_n$. The regularization parameter can be randomly chosen, e.g. by use of any data-driven method (cross validation etc.). The only requirement is to make sure that $\sqrt{n}(\Lambda_n - \lambda_0) \rightarrow 0$ almost surely for $n \rightarrow \infty$. A simple way to fulfill this condition for any data-driven method is to choose a (possibly large) constant $c \in (0, \infty)$ and to modify the method in such a way that it picks a value from $[\lambda_0, \lambda_0 + c/\sqrt{n \ln(n)}]$. Note that it is
indeed possible to use the same data for choosing the regularization parameter as for building the final estimate - just as usually done by practitioners, e.g., when applying cross validation.

The functional $\psi$. With these choices and assumptions, the asymptotic confidence set (Theorem 3.9) is valid for every functional $\psi : H \rightarrow \mathbb{R}^m$ which is Hadamard-differentiable at $f_{P,\lambda_0}$ and fulfills (10). In the following, some concrete examples for $\psi$ are listed or even worked out in detail. In most cases, $\psi$ is continuous and linear so that the derivative $\psi'_{f_{P,\lambda_0}}$ is exactly known as it does not depend on the unknown $f_{P,\lambda_0}$. If $\psi'_{f_{P,\lambda_0}}$ is exactly known, then Assumption (10) can be checked in real applications by use of the following “test”: Define the $m \times n$-matrix

$$
\Psi := (\psi'_{f_{P,\lambda_0}}(x_1), \ldots, \psi'_{f_{P,\lambda_0}}(x_n))
$$

where $x_1, \ldots, x_n$ are the observed values of the input variables. If Assumption (10) is violated, then the probability that $\Psi$ has rank $m$ (i.e. full rank for $n > m$) is equal to 0. (This follows from continuity of $\psi'_{f_{P,\lambda_0}}$ and the fact that $P_X(\text{supp}(P_X)) = 1$.) That is, if the observed $\Psi$ has full rank, one can assume that (10) is fulfilled.

**Example 1:** Pointwise confidence intervals

Fix some $x_1, \ldots, x_m \in X$ and define

$$
\psi(f) = (f(x_1), \ldots, f(x_m))^T, \quad f \in H.
$$

Since $\psi : H \rightarrow \mathbb{R}^m$ is continuous and linear, $\psi$ is continuously Hadamard-differentiable. The derivative is given by

$$
\psi'_f = (\Phi(x_1), \ldots, \Phi(x_m))^T \in H^m, \quad f \in H.
$$

Condition (10) can be checked as described above. Since

$$
\psi'_f(x) = (k(x, x_1), \ldots, k(x, x_m))^T \quad \forall x \in X, \quad f \in H,
$$

it follows from Prop. 3.10 that

$$
g_{D_n, \Lambda_n}(x, y) = -L'(x, y, f_{D_n, \Lambda_n}(x)) \left( \frac{1}{2\lambda_n} k(x, x_1) + \sum_{i=1}^n \alpha_i(x)k(X_i, x_1) \right.
\left. \vdots \right.
\left. \frac{1}{2\lambda_n} k(x, x_m) + \sum_{i=1}^n \alpha_i(x)k(X_i, x_m) \right)
$$

where the $\alpha_i(x), i \in \{1, \ldots, n\}$, are calculated according to Prop. 3.10. Fix any $\alpha \in (0, 1)$. Then, Theorem 3.9 says that

$$
Q \left( (f_{P,\lambda_0}(x_1), \ldots, f_{P,\lambda_0}(x_m)) \in C_{n,\alpha}(D_n, \Lambda_n) \right) \xrightarrow{n \rightarrow \infty} 1 - \alpha.
$$
where \( C_{n,\alpha}(D_n, \Lambda_n) \) is the elliptical confidence set as defined in Theorem 3.9. \( \square \)

Due to the reproducing property [21, Def. 4.18], Example 1 is a special case of the following example.

**Example 2**: Confidence intervals for inner products

Fix some \( h_1, \ldots, h_m \in H \) which are linearly independent on the support of \( P_X \) and define

\[
\psi(f) = (\langle f, h_1 \rangle_H, \ldots, \langle f, h_m \rangle_H)^T, \quad f \in H.
\]

Since \( \psi : H \to \mathbb{R}^m \) is continuous and linear, \( \psi \) is continuously Fréchet differentiable and the derivative is given by

\[
\psi'(f) = (h_1, \ldots, h_m)^T \in H^m, \quad f \in H,
\]

and condition (10) is fulfilled. It follows from Prop. 3.10 that

\[
g_{D_n, \Lambda_n}(x, y) = -L'(x, y, f_{D_n, \Lambda_n}(x)) \cdot \left( \frac{1}{2\Lambda_n} h_1(x) + \sum_{i=1}^n \alpha_i(x) h_1(x_i) \right) \ldots \left( \frac{1}{2\Lambda_n} h_m(x) + \sum_{i=1}^n \alpha_i(x) h_m(x_i) \right)
\]

where \( \alpha_i(x), i \in \{1, \ldots, n\} \), are calculated according to Prop. 3.10. Fix any \( \alpha \in (0, 1) \). Then, Theorem 3.9 says that

\[
Q \left( (\langle f_{P,\lambda_0}, h_1 \rangle_H, \ldots, \langle f_{P,\lambda_0}, h_m \rangle_H) \in C_{n,\alpha}(D_n, \Lambda_n) \right) \xrightarrow{n \to \infty} 1 - \alpha.
\]

where \( C_{n,\alpha}(D_n, \Lambda_n) \) is the elliptical confidence set as defined in Theorem 3.9. \( \square \)

**Example 3**: Confidence set for the gradient

Fix any \( x_0 \) in the interior of \( X \) and, for every \( f \in H \), let

\[
\psi(f) = \partial f(x_0) \in \mathbb{R}^d
\]

be the gradient vector of \( f \) in \( x_0 \). According to [21, p. 130ff], the partial derivative of \( f \) in \( x_0 \) with respect to the \( j \)-th coordinate of \( x \) is given by \( \partial_j f(x_0) = \langle f, \partial_j \Phi(x_0) \rangle_H \). Hence, this is again a special case of Example 2

and it follows that

\[
\psi'(x) = \frac{\partial}{\partial \tilde{x}} k(x, \tilde{x}) \bigg|_{\tilde{x} = x_0} \quad \forall x \in X, \ f \in H.
\]
Again, Assumption (10) can be checked as described above. \(\square\)

**Example 4:** Confidence set for integrals

Fix any Borel set \(B \subset \mathcal{X}\) and, for every \(f \in H\), define

\[
\psi(f) = \int_B f \, dP_X \in \mathbb{R}^d.
\]

This is again a special case of Example 2, the derivative is given by

\[
\psi_f'(x) = \int_B k(x, \tilde{x}) \, P_X(d\tilde{x}) \quad \forall x \in \mathcal{X}, \ f \in H
\]

and Assumption (10) can be checked as described above. \(\square\)

**Example 5:** Confidence interval for the \(H\)-norm and for the \(L^2\)-norm

The map \(f \mapsto \psi(f) = \|f\|_H^2\) is continuously Hadamard differentiable with derivative \(\psi'_f = 2f\) at \(f\); see e.g. [7, Example 5.1.6(c)]. Condition (10) is fulfilled if \(f_{P,\lambda_0}\) is not \(P_X\)-almost surely equal to 0. Hence, it is possible to construct a confidence interval for \(\|f_{P,\lambda_0}\|_H^2\) according to Theorem 3.9 and, therefore, also for \(\|f_{P,\lambda_0}\|_H\) by taking square roots.

Similarly, for any \(B \subset \mathbb{R}^d\), the map

\[
f \mapsto \psi(f) = \|f\|_{L^2(B,\lambda^d)}^2 = \int_B (f(x))^2 \, dx
\]

is continuously Hadamard-differentiable and the derivative at any \(f \in H\) is equal to \(\psi'_f = \int_B 2f(x) \Phi(x) \, dx\) (this follows from [21, Lemma 2.21] where \(L(x, y, t) = t^2\)). Again, Condition (10) is fulfilled if \(f_{P,\lambda_0}\) is not \(P_X\)-almost surely equal to 0 on \(B\). This can be shown by considering the RKHS which consists of the restrictions of the elements \(f \in H\) on \(\text{supp}(P_X)\). \(\square\)

Similarly, to Example 4, the map \(f \mapsto \|f - f_{P,\lambda_0}\|_H^2\) is continuously differentiable so that one might be tempted to apply \(\psi(f) = \|f - f_{P,\lambda_0}\|_H^2\) in Theorem 3.9 in order to obtain a confidence band for the whole function \(f_{P,\lambda_0}\) and not just for a finite number of points as in Example 1. However, this is not possible because then the derivative is given by \(\psi'_f = 2(f - f_{P,\lambda_0})\) so that \(\psi'_{f_{P,\lambda_0}} = 0\) which violates (10). The mathematical reason behind is that, according to the continuous mapping theorem, \(\sqrt{n}(f_{D_n,\Lambda_n} - f_{P,\lambda_0})\) weakly converges to \(\|H_P\|_H^2\). That is, \(\|f_{D_n,\Lambda_n} - f_{P,\lambda_0}\|^2\) converges with rate \(n\) while the confidence sets obtained from Theorem 3.9 are based on the
rate $\sqrt{n}$. By estimating quantiles of the distribution of $\|H_P\|_H^2$, it would be possible to derive confidence bands for the whole function $f_{P,\lambda_0}$. However, estimating quantiles of the distribution of $\|H_P\|_H^2$ is a matter of its own and cannot be done by use of the results of Subsection 3.1 – among other things because $\|H_P\|_H^2$ is not normally distributed (as $\|H_P\|_H^2 \geq 0$).

5 Simulations

5.1 Confidence sets for function values

The model. The situation

$$Y_i = f_0(X_i) + \varepsilon_i, \quad i \in \{1, \ldots, n\}$$

is considered with the regression function

$$f_0(x) = \log(x + 2) + 0.7 \sin(3x) + 0.7 \cos(2x).$$

(15)

The errors $\varepsilon_i$ are drawn i.i.d. from the standard normal distribution and the covariates $X_i$ are drawn i.i.d. from the uniform distribution on $[0, 5]$. The simulation consists of 5000 data sets with sample sizes $n$ equal to 250, 500, and 1000. The confidence sets apply to $f_{P,\lambda_0}$ with $\lambda_0 = 0.00001$ but the $L_1$-distance between $f_{P,\lambda_0}$ and the actual regression function $f_0$ is approximately equal to 0.026 and the maximal pointwise distance is approximately equal to 0.091 so that the difference between $f_{P,\lambda_0}$ and $f_0$ can be almost ignored for practical purposes here. Three kinds of confidence sets are considered: a univariate one for $f_{P,\lambda_0}(\tilde{x}_0)$ with $\tilde{x}_0 = 3$, a multivariate one for the four values $f_{P,\lambda_0}(\tilde{x})$, $\tilde{x} \in \{1, 2, 3, 4\}$, and a multivariate one for the seven values $f_{P,\lambda_0}(\tilde{x})$, $\tilde{x} \in \{1, 1.5, 2, 2.5, 3, 3.5, 4\}$. The nominal (asymptotic) confidence level is 0.95.

Estimation. The regularized kernel method was applied with the Gaussian RBF kernel $k(x, x') = \exp(\gamma \|x - x'\|_{\mathbb{R}^d}^2)$ and the logistic loss function with parameter $\sigma = 0.5$. Following [11] and [14, p. 9], the hyperparameter $\gamma$ of the kernel was fixed to 0.5 which is about the inverse of the median of the values $\|x_i - x'_j\|_{\mathbb{R}^1}^2$. The regularization parameter was chosen within the values

0.00001, 0.00005, 0.0001, 0.0005, 0.001, 0.005, 0.01

in a data-driven way by a fivefold cross-validation.

Performance results. Table 1 lists the simulated coverage probabilities and, in case of the univariate confidence interval, the average length ($\pm$ standard
deviation) of the intervals obtained by 5000 data sets. Figure 1 shows the boxplots for the estimates of the asymptotic variance $\Sigma_P$ of $\sqrt{n}(f_{D_n,\Lambda_n}(\hat{x}_0) - f_{P,\lambda_0}(\tilde{x}_0))$ for the different sample sizes. In addition, Figure 2 shows the plot of the true function $f_{P,\lambda_0}$ and the pointwise univariate confidence interval for every $\tilde{x} \in [0, 5]$ obtained for four different data sets with $n = 500$. This is only for illustration purposes and must not be mixed with a simultaneous confidence band; the band around the true function is not a simultaneous confidence band.

5.2 Confidence set for the gradient

The model. Two situations are considered, the univariate one

$$Y_i = f_0(X_i) + \varepsilon_i, \quad i \in \{1, \ldots, n\},$$

exactly as in Subsection 5.1 and the multivariate one

$$Y_i = f_0(X_{i,1}) + \sin(1.5X_{i,2}) + \varepsilon_i, \quad i \in \{1, \ldots, n\},$$

where $f_0$ is as in (15). The errors $\varepsilon_i$ are drawn i.i.d. from the standard normal distribution. In the univariate case, the covariates $X_i$ are drawn i.i.d. from the uniform distribution on $[0, 5]$ and, in the multivariate case, the covariates $X_{i,1}$ are also drawn i.i.d. from the uniform distribution on $[0, 5]$ and the covariates $X_{i,2}$ are drawn i.i.d. from the uniform distribution on $[-1, 1]$. In both cases, we consider confidence sets for $\psi(f_{P,\lambda_0}) = \partial f_{P,\lambda_0}(x_0)$ with $\lambda_0 = 0.00001$ where, in the univariate case, $x_0 = 3$ and, in the multivariate case, $x_0 = (3, 0)$. Accordingly, the confidence set is an interval in the univariate case and an ellipse in the multivariate case. The nominal (asymptotic) confidence level is 0.95.

Estimation. The regularized kernel method was applied with the Gaussian RBF kernel and the logistic loss function with parameter $\sigma = 0.5$. Following

\begin{tabular}{l|ccc|ccc|ccc}
 & 1-dim. & & & 4-dim. & & & 7-dim. & \hline
$n$ & Cov. prob. (%) & length & & Cov. prob. (%) & & Cov. prob. (%) & & Cov. prob. (%) & &
250 & 92.7 & 0.61±0.09 & & 91.6 & & 79.4 & &
500 & 94.0 & 0.44±0.04 & & 93.1 & & 91.1 & &
1000 & 94.7 & 0.32±0.02 & & 94.5 & & 93.0 & &
\end{tabular}

Table 1: Simulated coverage probability of the confidence sets obtained by 5000 data sets in Subsection 5.1.
Figure 1: Boxplots for the estimation of the asymptotic variance $\Sigma_P$ of $\sqrt{n}(f_{D_n,\Lambda_n}(\tilde{x}_0) - f_{P,\lambda_0}(\tilde{x}_0))$ for the different sample sizes $n$ in Subsection 5.1.
Figure 2: Estimated pointwise 0.95-confidence intervals (grey area) for four different data sets with sample size \( n = 500 \) and the function \( f_{P,\lambda_0} \) (solid line) in Subsection 5.1.
### Table 2: Simulated coverage probability of the confidence sets obtained by 5000 data sets in Subsection 5.2.

| n   | 1-dim. Cov. prob. (%) | length ± | 2-dim. Cov. prob. (%) |
|-----|-----------------------|----------|-----------------------|
| 250 | 84.0                  | 0.27±0.23| 74.5                  |
| 500 | 90.0                  | 0.17±0.13| 83.4                  |
| 1000| 91.5                  | 0.11±0.09| 91.3                  |

In [4] and [14] p. 9, the hyperparameter $\gamma$ of the kernel was fixed to $1/3$ which is about the inverse of the median of the values $\|x_i - x_j\|_2$. The regularization parameter was chosen as in Subsection 5.1.

**Performance results.** Table 2 lists the simulated coverage probabilities and, in case of the univariate confidence interval, the average length (± standard deviation) of the intervals obtained by 5000 data sets. For $n = 1000$ in the multivariate case, Figure 3 shows the estimates $\psi(f_{D_n, \Lambda_n}) = \partial f_{D_n, \Lambda_n}(x_0)$ obtained in the 5000 runs (gray points), the true value $\psi(f_{P, \lambda_0})$ (as cross ×), and the ellipse (dashed boundary)

$$\left\{ w \in \mathbb{R}^m \left| \left\| \Sigma_P^{-\frac{1}{2}} (w - \psi(f_{P, \lambda_0})) \right\|^2_{\mathbb{R}^m} \leq \frac{\chi^2_{m, \alpha}}{n} \right. \right\},$$

in each plot. Asymptotically, this ellipse contains the estimate $\psi(f_{D_n, \Lambda_n})$ with probability 0.95. In addition, each plot shows the estimate $\psi(f_{D_n, \Lambda_n})$ (as black point) and illustrates the estimated covariance matrix $\hat{\Sigma}_n(D_n, \Lambda_n)$ by showing the ellipse (solid boundary)

$$\left\{ w \in \mathbb{R}^m \left| \left\| \hat{\Sigma}_n(D_n, \Lambda_n)^{-\frac{1}{2}} (w - \psi(f_{P, \lambda_0})) \right\|^2_{\mathbb{R}^m} \leq \frac{\chi^2_{m, \alpha}}{n} \right. \right\}$$

given by the estimate $\hat{\Sigma}_n(D_n, \Lambda_n)$ in one of the first four runs of the simulation.

### 6 Conclusions

Regularized kernel methods constitute an important class of standard learning algorithms in machine learning. As theoretical investigations concerning asymptotic properties have mainly focused on rates of convergence, the lack of (asymptotic) results on statistical inference is a serious limitation for their use in mathematical statistics. Therefore, the article derives asymptotically correct confidence sets for $\psi(f_{P, \lambda_0})$ where $f_{P, \lambda_0}$ denotes the minimizer of the
Figure 3: For \( n = 1000 \) in the multivariate case, each plot shows the estimates \( \psi(f_{D_n,\Lambda_n}) \) obtained in the 5000 runs (gray points), the true value \( \psi(f_{P,\Lambda_0}) \) (as cross \( \times \)), and the ellipse (dashed boundary) which asymptotically contains the estimate \( \psi(f_{D_n,\Lambda_n}) \) with probability 0.95. Each of the four plots shows the estimate \( \psi(f_{D_n,\Lambda_n}) \) (black point) and the ellipse where the true covariance \( \Sigma_P \) is replaced by the estimate \( \tilde{\Sigma}_n(D_n, \Lambda_n) \) (solid boundary) in one of the first four runs of the simulation in Subsection 5.2.
regularized risk in the reproducing kernel Hilbert space $H$ and $\psi : H \to \mathbb{R}^m$ is any Hadamard-differentiable functional. That is, the confidence sets do not apply to the minimizer $f^*$ of the unregularized risk, which would be the quantity of primary interest, but to the minimizer of the regularized risk. On the one hand, this is due to the so-called no-free-lunch theorem and obtaining confidence sets for $f^*$ would require a number of technical assumptions which can hardly be made plausible in practical applications. Without such assumptions, $f^*$ does not need to exist, if it exists, it does not have to be unique, and the rate of convergence depends on unknown properties. Technical assumptions can completely be avoided in this article; all assumptions are simple and can easily be communicated to practitioners. On the other hand, it is exemplified in a simulated example (Subsection 5.1) that the difference between $f^*$ and $f_{P,\lambda_0}$ is negligible for practical purposes even for moderately small $\lambda_0 > 0$.

The derivation of the confidence sets is done by use of asymptotic normality of a large class of regularized kernel methods and by the derivation of a strongly consistent estimator for the unknown covariance matrix of the limiting normal distribution. To this end, the following non-trivial problems had to be solved satisfactorily: (i) the derivation of a manageable formula for the covariance matrix, which is accessible for a plug-in estimator, (ii) strong consistency of the plug-in estimator, (iii) the exclusion of degeneracy of the covariance matrix by simple and week conditions, and (iv) the derivation of an algorithm for the calculation of the estimator which is computationally tractable also for moderately large sample sizes.

Applications include (multivariate) pointwise confidence sets for values of $f_{P,\lambda_0}$ and confidence sets for gradients, integrals, and norms. However, the derivation of simultaneous confidence bands is a matter of further research. It follows from [10, Theorem 3.1] that
$$
\sqrt{n} \left\| f_{D_n,\Lambda_n} - f_{P,\lambda_0} \right\|_{\infty} \sim \left\| H_P \right\|_{\infty}.
$$

Hence, simultaneous confidence bands could be obtained if it is possible to derive a consistent estimator for quantiles of $\left\| H_P \right\|_{\infty}$.

7 Appendix: Proofs

Assumption 3.1 is valid in the whole appendix. Since the results of Section 3 are based on results and proofs in [10], we have to recall the quite technical setting from [10, § A.1] at first.

In order to shorten notation, define
$$
L_f : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}, \quad (x,y) \mapsto L_f(x,y) = L(x,y,f(x))
$$
for every function $f : \mathcal{X} \to \mathbb{R}$. Accordingly, define

$$L'_f(x, y) = L'(x, y, f(x)), \quad L''_f(x, y) = L''(x, y, f(x))$$

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 $b \in L_2(P)$ and a constant $c \in (0, \infty)$ such that

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

Let

$$G_1 := \left\{ g : \mathcal{X} \times \mathcal{Y} \to \mathbb{R} \mid \exists z \in \mathbb{R}^{d+1} \text{ such that } g = I_{(-\infty, z]} \right\}$$

be the set of all indicator functions $I_{(-\infty, z]}$. Define $c_0 := \sqrt{\frac{1}{\lambda_0}} \int b \, dP + 1,$

$$G_2 := \left\{ g : \mathcal{X} \times \mathcal{Y} \to \mathbb{R} \mid \exists f_0 \in H, \exists f \in H \text{ such that} \quad \|f_0\|_H \leq c_0, \quad \|f\|_H \leq 1 \text{ and} \quad g = L'_f f \right\},$$

and

$$G := G_1 \cup G_2 \cup \{b\}.$$ 

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

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

and $B_0 := \text{cl}(\text{lin}(B_S))$ the closed linear span of $B_S$ in $\ell_\infty(G)$. That is, $B_S$ is a subset of $\ell_\infty(G)$ whose elements correspond to finite measures. The assumptions on $L$ and $P$ imply that $G \to \mathbb{R}, \quad g \mapsto \int g \, dP$ is a well-defined element of $B_S$. Most often, we identify an element $G \in B_S$ with its corresponding finite measure $\mu$. That is, we write $\mu(g) = G(g) = \int g \, d\mu$ for every $g \in G$.

Let $\mu \in B_S$. Then,

$$S(\mu) := f_{\mu, \lambda_0} = \arg \inf_{f \in H} \left( \int L(x, y, f(x)) \, d\mu(x, y) + \lambda_0 \|f\|_H^2 \right).$$

This defines a map $S : B_S \to H$. As the multiplication by a strictly positive real number does not change the “arg inf”, we have

$$f_{\lambda \mu, \lambda} = f_{\lambda, \lambda_0} = S \left( \frac{\lambda}{\lambda_0} \mu \right) \quad \forall \mu \in B_S, \quad \lambda \in (0, \infty). \quad (17)$$
Let $\mu \in B_S$ such that $\mu(b) < P(b) + \lambda_0$. Then, it is shown in [10, Theorem A.8] that, $S$ is Hadamard differentiable in $\mu$ tangentially to $B_0$. The derivative in $\mu$ is given by

$$S'_\mu(\nu) = -K^{-1}_\mu \left( \int L'_{f_\mu,\lambda_0}(x,y) \Phi(x) \nu(d(x,y)) \right) \quad \forall \nu \in \text{lin}(B_S) \quad (18)$$

and

$$K_\mu : H \to H, \quad f \mapsto 2\lambda_0 f + \int L''_{f_\mu,\lambda_0}(x,y) f(x) \Phi(x) \mu(d(x,y)). \quad (19)$$

Note that the integrals with respect to the finite signed measure $\nu$ in (18) and the measure $\mu$ in (19) are Bochner integrals as the integrands are $H$-valued functions. According to [10, Lemma A.5], $K_\mu$ is an invertible continuous linear operator and, according to [10, Theorem A.8], the derivative $S'_\mu : B_0 \to H$ is a continuous linear operator. The following relation between $K_\mu$ and the random $K_{D_n,A_n}$ defined in (7) is valid:

$$K_{D_n,A_n} = \frac{A_n}{\lambda_0} K_{\frac{\lambda_0}{A_n}P_{D_n}}. \quad (20)$$

If we identify the empirical measure $P_{D_n}(\omega)$ and $P$ with their corresponding elements in $\ell_\infty(G)$, it is shown in [10, Lemma A.9] that

$$\sqrt{n}(P_{D_n} - P) \Rightarrow G_P \quad \text{in } \ell_\infty(G) \quad (21)$$

where $G_P : \Omega \to \ell_\infty(G)$ is a tight Borel-measurable Gaussian process. Then, it is shown in [10, Proof of Theorem 3.1] that

$$\sqrt{n}(f_{D_n,A_n} - f_{P,\lambda_0}) \Rightarrow H_P = S'_P(G_P) \quad \text{in } H. \quad (22)$$

**Proof of Corollary 3.3** According to the delta-method [27, Theorem 3.9.4], it follows from (22) and Hadamard-differentiability of $\psi$ in $f_{P,\lambda_0}$ that

$$\sqrt{n} \left( \psi(f_{D_n,A_n}) - \psi(f_{P,\lambda_0}) \right) \Rightarrow \langle \psi'_{f_{P,\lambda_0}}, H_P \rangle_H.$$

Since $f \mapsto \langle \psi'_{f_{P,\lambda_0}}, f \rangle_H$ is a continuous linear operator and $H$ is a zero-mean Gaussian process, it follows that the limit distribution is a multivariate normal distribution with mean zero, i.e., the distribution of $\langle \psi'_{f_{P,\lambda_0}}, H_P \rangle_H$ is equal to $N_m(0, \Sigma_P)$ for some covariate matrix $\Sigma_P \in \mathbb{R}^{m \times m}$; see e.g. [27, §3.9.2].
Proof of Prop. 3.4: First, it is a direct consequence of the definition of the continuous linear operator $K_P$ that $K_P$ is self-adjoint and, accordingly, the inverse $K_P^{-1}$ is again self-adjoint; see [9, Lemma VI.2.10]. Define $f_j := K_P^{-1}(\psi_{f_P,\lambda_0, j}) \in H$ and note that [21 (5.4)] implies

$$L'_{f_P,\lambda_0} \|f_j\|^{-1}_H f_j \in \mathcal{G}. \quad (23)$$

Since $K_P^{-1}$ is self-adjoint, it follows for every $G \in \text{lin}(B_S)$ with corresponding signed measure $\mu$ that

$$\langle \psi'_{f_P,\lambda_0, j}, S'_P(G) \rangle_H \overset{(13)}{=} -\left\langle f_j, \int L'_{f_P,\lambda_0} \Phi \, d\mu \right\rangle_H \overset{(\ast)}{=} -\int L'_{f_P,\lambda_0} \langle f_j, \Phi \rangle_H \, d\mu = -\int L'_{f_P,\lambda_0} f_j \, d\mu \overset{(23)}{=} -\|f_j\|_H \cdot G(L'_{f_P,\lambda_0} \|f_j\|^{-1}_H f_j).$$

where $(\ast)$ follows from interchangeability of Bochner integrals with continuous linear operators; see e.g. [7, Theorem 3.10.16 and Remark 3.10.17]. Next, it follows from continuity of $S'_P$ that

$$\langle \psi'_{f_P,\lambda_0, j}, S'_P(G) \rangle_H = -\|f_j\|_H \cdot G(L'_{f_P,\lambda_0} \|f_j\|^{-1}_H f_j)$$

is valid even for every $G \in B_0$ where $B_0$ denotes the closed linear span of $B_S$ in $\ell_\infty(G)$. Since $G_P$ takes its values in $B_0$, it follows from $H_P = S'_P(G_P)$ now that

$$\langle \psi'_{f_P,\lambda_0, j}, H_P \rangle_H = -\|f_j\|_H \cdot G(P(L'_{f_P,\lambda_0} \|f_j\|^{-1}_H f_j) \quad \forall \ j \in \{1, \ldots, m\}. \quad (25)$$

According to (21) and (23),

$$\begin{pmatrix}
G_P(L'_{f_P,\lambda_0} \|f_1\|^{-1}_H f_1) \\
\vdots \\
G_P(L'_{f_P,\lambda_0} \|f_m\|^{-1}_H f_m)
\end{pmatrix} \sim \mathcal{N}_m(0, \tilde{\Sigma}_P)$$

where $\tilde{\Sigma}_P$ is the covariance matrix of

$$\left(L'_{f_P,\lambda_0}(X,Y)\|f_1\|^{-1}_H f_1(X), \ldots, L'_{f_P,\lambda_0}(X,Y)\|f_m\|^{-1}_H f_m(X)\right)^T;$$

see, e.g., [27, p. 81f]. Let $C$ denote the diagonal matrix with diagonal entries $\|f_1\|_H, \ldots, \|f_m\|_H$. Then, it follows from (25) that

$$\Sigma_P = \text{Cov}\left(\langle \psi'_{f_P,\lambda_0}, H_P \rangle_H\right) = C\tilde{\Sigma}_PC. \quad (26)$$
Since, according to the reproducing property and self-adjointness of $K_P^{-1}$,
\[
    f_j(x) = \left( K_P^{-1} \psi_{f_P,\lambda_0}^{j} \right)(x) = \left\langle K_P^{-1} \psi_{f_P,\lambda_0}^{j}, \Phi(x) \right\rangle_H = \left\langle \psi_{f_P,\lambda_0}^{j}, K_P^{-1}(\Phi(x)) \right\rangle_H \quad \forall \, j \in \{1, \ldots, m\},
\]
it follows that $-L'_{f_P,\lambda_0} \|f_j\|_H^{-1} f_j = \|f_j\|_H^{-1} g_{P,\lambda_0,j}$ where $g_{P,\lambda_0,j}$ denotes the $j$-th component of $g_{P,\lambda_0}$. Hence, $	ilde{\Sigma}_P = \text{Cov}(C^{-1} g_{P,\lambda_0}(X,Y))$ so that (26) implies $\Sigma_P = \text{Cov}(g_{P,\lambda_0}(X,Y))$.

**Lemma 7.1** Under the Assumptions of Theorem 3.6, the covariance estimator $\hat{\Sigma}_n(D_n,\Lambda_n)$ is measurable with respect to $A$ and $B^\otimes m^2$.

**Proof of Lemma 7.1.** It has to be shown that $g_{D_n,\Lambda_n}(X_i,Y_i)$ is measurable for every $i \in \{1, \ldots, n\}$. First, note that $\omega \mapsto f_{D_n}(\omega,\Lambda_n(\omega))$ is measurable because: for every fixed $\lambda > 0$, the map $D \mapsto f_{D,\lambda}$ is continuous on $(\mathcal{X} \times \mathcal{Y})^n$ according to [21, Lemma 5.13] and, for every fixed $D \in (\mathcal{X} \times \mathcal{Y})^n$, the map $\lambda \mapsto f_{D,\lambda}$ is continuous on $(0, \infty)$ according to [21, Theorem 5.17]; hence, $(D,\lambda) \mapsto f_{D,\lambda}$ is a Caratheodory function and, therefore, measurable, see e.g. [7, Theorem 2.5.22].

Secondly, we show measurability of $K_{D_n,\Lambda_n}^{-1}(\Phi(X_i))$. To this end, define

\[
    A_{D,\lambda,g} : H \rightarrow H, \quad f \mapsto \frac{1}{n} \sum_{i=1}^{n} L''(x_i, y_i, g(x_i)) f(x_i) \Phi(x_i)
\]

and

\[
    K_{D,\lambda,g} : H \rightarrow H, \quad f \mapsto 2\lambda f + A_{D,\lambda,g}(f)
\]

for every $D = ((x_1, y_1), \ldots, (x_n, y_n)) \in (\mathcal{X} \times \mathcal{Y})^n$, $\lambda \in (0, \infty)$, and $g \in H$. That is, $K_{D_n,\Lambda_n} = K_{D_n,\Lambda_n,f_{D_n,\Lambda_n}}$. The assumptions imply that,

\[
(\mathcal{X} \times \mathcal{Y})^n \times (0, \infty) \times H \rightarrow H, \quad (D,\lambda,g) \mapsto K_{D,\lambda,g}(f)
\]

is continuous (27) for every $f \in H$. Note that

\[
    \left\langle f, A_{D,\lambda,g}(f) \right\rangle_H = \frac{1}{n} \sum_{i=1}^{n} L''(x_i, y_i, g(x_i)) f(x_i) \langle f, \Phi(x_i) \rangle_H = \frac{1}{n} \sum_{i=1}^{n} L''(x_i, y_i, g(x_i)) \langle f(x_i) \rangle^2 \geq 0
\]
because convexity of $t \mapsto L(x, y, t)$ implies $L''(x, y, t) \geq 0$. Hence,

$$
\|K_{D, \lambda, g}(f)\|_H^2 = 4\lambda^2 \|f\|^2 + 2\lambda(f, A_{D, \lambda, g}(f))_H + \|A_{D, \lambda, g}(f)\|_H^2 \geq 4\lambda^2 \|f\|^2
$$

for every $f \in H$, and this implies

$$
\|K_{D, \lambda, g}^{-1}\| \leq \frac{1}{2\lambda} \quad \forall (D, \lambda, g) \in (\mathcal{X} \times \mathcal{Y})^n \times (0, \infty) \times H. \quad (28)
$$

Let the sequence $(D_\ell, \lambda_\ell, g_\ell), \ell \in \mathbb{N}$, converge to some $(D, \lambda, g) \in (\mathcal{X} \times \mathcal{Y})^n \times (0, \infty) \times H$. Fix any $f \in H$ and denote $h := K_{D, \lambda, g}^{-1}(f)$. Then,

$$
\|K_{D_\ell, \lambda_\ell, g_\ell}^{-1}(f) - K_{D, \lambda, g}^{-1}(f)\|_H = \|K_{D_\ell, \lambda_\ell, g_\ell}^{-1}(h) - h\|_H = \|K_{D_\ell, \lambda_\ell, g_\ell}^{-1}(h) - K_{D, \lambda_\ell, g_\ell}(h)\|_H \leq \frac{1}{2\lambda_\ell} \|K_{D_\ell, \lambda_\ell, g_\ell}(h) - K_{D, \lambda_\ell, g_\ell}(h)\|_H \xrightarrow{\ell \to \infty} 0
$$

according to $\mathbf{(27)}$. That is, $(D, \lambda, g) \mapsto K_{D, \lambda, g}^{-1}(f)$ is continuous for every fixed $f \in H$. Since $f \mapsto K_{D, \lambda, g}^{-1}(f)$ is continuous for every fixed $(D, \lambda, g)$, the function $((D, \lambda, g), f) \mapsto K_{D, \lambda, g}^{-1}(f)$ is a Carathéodory function and, therefore, measurable. Since $f_{D_n, \Lambda_n}$ is measurable as shown above and $K_{D_n, \Lambda_n} = K_{D_n, \Lambda_n, f_{D_n, \Lambda_n}}$, it follows that $K_{D_n, \Lambda_n}^{-1}(\Phi(X_i))$ is measurable. Finally, measurability of the estimator $\psi_{D_n, \Lambda_n}$, measurability of $f_{D_n, \Lambda_n}$, and measurability of $K_{D_n, \Lambda_n}^{-1}(\Phi(X_i))$ imply measurability of

$$
g_{D_n, \Lambda_n}(X_i, Y_i) = -L'(X_i, Y_i, f_{D_n, \Lambda_n}(X_i)) \langle \psi_{D_n, \Lambda_n}'(X_i), K_{D_n, \Lambda_n}^{-1}(\Phi(X_i)) \rangle_H.
$$

\[\square\]

**Proof of Theorem 3.6.** For every $j \in \{1, \ldots, m\}$, let $\psi_{fP_0, \lambda_0}^j \in H$ denote the $j$-th component of $\psi_{fP_0}^j$ and, accordingly, let $\psi_{D_n, \Lambda_n, j}^j \in H$, $g_{D_n, \Lambda_n, j}$, and $g_{P_0, \lambda_0, j}$ denote the $j$-th component of $\psi_{D_n, \Lambda_n}^j$, $g_{D_n, \Lambda_n}$, and $g_{P_0, \lambda_0}$ respectively. Define $Z_i = (X_i, Y_i)$ for every $i \in \mathbb{N}$. Measurability of $g_{D_n, \Lambda_n, j}(Z_i)$ is shown in the proof of Lemma 7.1. Define $a := \|f_{P_0, \lambda_0}\|_\infty + 1 \in [1, \infty)$ and $c := \max_j \|\psi_{fP_0, \lambda_0, j}\|_H \cdot \|K_{P_0}^{-1}\| \cdot \|k\|_\infty$ where $\|K_{P_0}^{-1}\|$ denotes the operator norm of the continuous linear operator $K_{P_0}^{-1}$. Then, the definition of $g_{P_0, \lambda_0, j}$ and $\mathbf{(3)}$ imply

$$
|g_{P_0, \lambda_0, j}(z)| \leq c \cdot b'_a(z) \quad \forall z \in \mathcal{X} \times \mathcal{Y}. \quad (29)
$$

29
Hence, \( g_{P,\lambda_0,j} \) is \( P \)-square integrable. Fix any \( j, \ell \in \{1, \ldots, m\} \). We have to show

\[
\frac{1}{n} \sum_{i=1}^{n} g_{D_n, \Lambda_n, j}(Z_i) \xrightarrow{\text{a.s.}} \mathbb{E}[g_{P,\lambda_0,j}(Z_1)] \quad \text{as } n \to \infty \tag{30}
\]

\[
\frac{1}{n} \sum_{i=1}^{n} g_{D_n, \Lambda_n, j}(Z_i) g_{D_n, \Lambda_n, \ell}(Z_i) \xrightarrow{\text{a.s.}} \mathbb{E}[g_{P,\lambda_0,j}(Z_1) g_{P,\lambda_0,\ell}(Z_1)] \quad \text{as } n \to \infty \tag{31}
\]

According to \cite{10} Lemma A.9, \( G \) is a \( P \)-Donsker class and, therefore, a \( P \)-Glivenko-Cantelli class almost sure; see \cite{27} p. 82. Hence, \( \sup_{g \in G} |\mathbb{P}_{D_n}(g) - P(g)| \to 0 \) almost surely and, therefore, there is a measurable set \( \Omega_0 \in \mathcal{A} \) such that \( Q(\Omega_0) = 1 \) and \( \sup_{g \in G} |\mathbb{P}_{D_n(\omega)}(g) - P(g)| \to 0 \) for every \( \omega \in \Omega_0 \); see \cite{27} § 1.9 and Lemma 1.2.3. Due to the law of large numbers, we can choose \( \Omega_0 \in \mathcal{A} \) in such a way that, for every \( \omega \in \Omega_0 \), in addition, \( \frac{1}{n} \sum_{i=1}^{n} g_{P,\lambda_0,j}(Z_i(\omega)) \) and \( \frac{1}{n} \sum_{i=1}^{n} g_{P,\lambda_0,j}(Z_i(\omega)) g_{P,\lambda_0,\ell}(Z_i(\omega)) \) and \( \frac{1}{n} \sum_{i=1}^{n} b'_a(Z_i(\omega)) \) and \( \frac{1}{n} \sum_{i=1}^{n} b'_a(Z_i(\omega))^2 \) converge to their expectations for \( n \to \infty \). Furthermore, due to the assumptions on \( \psi'_{D_n,\Lambda_n} \) and \( \Lambda_n \), the set \( \Omega_0 \) can also be chosen in such a way that, in addition, \( \|\psi'_{D_n(\omega),\Lambda_n(\omega)} - \psi'_{f_P,\lambda_0}\|_H \to 0 \) and \( \Lambda_n(\omega) \to \lambda_0 \) for every \( \omega \in \Omega_0 \). Fix any \( \omega \in \Omega_0 \); define \( D_n := D_n(\omega) \) and \( \lambda_n := \Lambda_n(\omega) \) for every \( n \in \mathbb{N} \) and \( (x_i, y_i) := z_i := Z_i(\omega) \) for every \( i \in \mathbb{N} \). That is, we have

\[
\lim_{n \to \infty} \sup_{g \in G} |\mathbb{P}_{D_n}(g) - P(g)| = 0, \quad \lim_{n \to \infty} \lambda_n = \lambda_0, \tag{32}
\]

\[
\lim_{n \to \infty} \|\psi'_{D_n,\lambda_n} - \psi'_{f_P,\lambda_0}\|_H = 0, \tag{33}
\]

\[
\lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} g_{P,\lambda_0,j}(z_i) = \mathbb{E}_P[g_{P,\lambda_0,j}], \tag{34}
\]

\[
\lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} g_{P,\lambda_0,j}(z_i) g_{P,\lambda_0,\ell}(z_i) = \mathbb{E}_P[g_{P,\lambda_0,j} g_{P,\lambda_0,\ell}], \tag{35}
\]

\[
\lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} b'_a(z_i) = \mathbb{E}_P b'_a, \quad \text{and} \quad \lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} b'_a(z_i)^2 = \mathbb{E}_P b'_a^2. \tag{36}
\]

It is shown in \cite{10} (46) and (47) that \( S : B_\mathcal{S} \to H, \mu \mapsto S(\mu) = f_{\mu,\lambda_0} \) is continuous in \( P \) and, therefore,

\[
\lim_{n \to \infty} f_{D_n,\lambda_n} \xrightarrow{\text{(17)}} \lim_{n \to \infty} S(\frac{\lambda_n}{\lambda_0} P_{D_n}) \overset{\text{(32)}}{=} S(P) = f_{P,\lambda_0}. \tag{37}
\]
In view of (34), it suffices to prove
\[
\frac{1}{n} \sum_{i=1}^{n} g_{D_n, \lambda_n, j}(z_i) - \frac{1}{n} \sum_{i=1}^{n} g_{P, \lambda_0, j}(z_i) \xrightarrow{n \to \infty} 0
\] (38)
in order to prove (30).

First, it is shown in the following that, for the fixed sequence \((z_n)_{n \in \mathbb{N}} \in \mathcal{X} \times \mathcal{Y}\), there is an \(n_j \in \mathbb{N}\) and a sequence \((\varepsilon_{j,n})_{n \in \mathbb{N}} \subset [0, \infty)\) such that
\[
\lim_{n \to \infty} \varepsilon_{j,n} = 0 \quad \text{and, for every } n \geq n_j \quad \text{and every } z = (x, y) \in \mathcal{X} \times \mathcal{Y},
\]
\[
|g_{D_n, \lambda_n, j}(z) - g_{P, \lambda_0, j}(z)| \leq \varepsilon_{j,n} + \varepsilon_{j,n} \cdot b'_a(z) .
\] (39)

To this end, note that it is shown in (10) (43) that \(\mu \mapsto K_{\mu}^{-1}\) is continuous in \(P\) and, therefore, it follows from (32) that
\[
K_{D_n, \lambda_n}^{-1} \frac{\lambda_0}{\lambda_n} K_{x_n P_{D_n}}^{-1} \xrightarrow{n \to \infty} K_{P}^{-1} \quad \text{in operator norm.}
\] (40)

The definitions imply
\[
|g_{D_n, \lambda_n, j}(z) - g_{P, \lambda_0, j}(z)| \leq \left| \langle \psi_{D_n, \lambda_n, j}', K_{D_n, \lambda_n}^{-1}(\Phi(x)) \rangle_H \right| \cdot \left| L_{f_{D_n, \lambda_n}}'(z) - L_{f_{P, \lambda_0}}'(z) \right| + \left| \langle \psi_{D_n, \lambda_n, j}', K_{D_n, \lambda_n}^{-1}(\Phi(x)) \rangle_H - \langle \psi_{P, \lambda_0, j}' \circ K_P^{-1}(\Phi(x)) \rangle_H \right| \cdot \left| L_{f_{P, \lambda_0}}'(z) \right| .
\] (41)

Due to (37), there is an \(n_j \in \mathbb{N}\) such that \(\|f_{D_n, \lambda_n}\|_{\infty} \leq \|f_{P, \lambda_0}\|_{\infty} + 1 = a\) for every \(n \geq n_j\). Hence, the first summand converges to 0 uniformly in \(z \in \mathcal{X} \times \mathcal{Y}\) because of (37), \(\|\Phi(\bar{x})\|_H \leq \|k\|_{\infty} \forall \bar{x} \in \mathcal{X}\), (33), (40), and
\[
\left| \langle \psi_{D_n, \lambda_n, j}', K_{D_n, \lambda_n}^{-1}(\Phi(x)) \rangle_H \right| \cdot \left| L_{f_{D_n, \lambda_n}}'(z) - L_{f_{P, \lambda_0}}'(z) \right| \leq \left\| \psi_{D_n, \lambda_n, j}' \right\|_H \cdot \left\| K_{D_n, \lambda_n}^{-1} \right\| \cdot \sup_{\bar{x} \in \mathcal{X}} \left\| \Phi(\bar{x}) \right\| \cdot b''_d \cdot \left\| f_{D_n, \lambda_n} - f_{P, \lambda_0} \right\|_{\infty} .
\] (3)

For \(\psi_{s,j}' \in H\), let \(\psi_{s,j}' \circ K_{\mu}^{-1}\) denote the continuous linear operator \(h \mapsto \langle \psi_{s,j}' \circ K_{\mu}^{-1}(h) \rangle_H\). Then, the second summand in (41) is bounded via
\[
\left| \langle \psi_{D_n, \lambda_n, j}', K_{D_n, \lambda_n}^{-1}(\Phi(x)) \rangle_H - \langle \psi_{P, \lambda_0, j}' \circ K_P^{-1}(\Phi(x)) \rangle_H \right| \cdot \left| L_{f_{P, \lambda_0}}'(z) \right| \leq \left\| \psi_{D_n, \lambda_n, j}' \circ K_{D_n, \lambda_n}^{-1} - \psi_{P, \lambda_0, j}' \circ K_P^{-1} \right\| \cdot \sup_{\bar{x} \in \mathcal{X}} \left\| \Phi(\bar{x}) \right\|_H \cdot b'_a(z) ,
\] (3)

and \(\left\| \psi_{D_n, \lambda_n, j}' \circ K_{D_n, \lambda_n}^{-1} - \psi_{P, \lambda_0, j}' \circ K_P^{-1} \right\| \cdot \sup_{\bar{x} \in \mathcal{X}} \left\| \Phi(\bar{x}) \right\|_H\) converges to zero because of \(\|\Phi(\bar{x})\|_H \leq \|k\|_{\infty} \forall \bar{x} \in \mathcal{X}\), (33), and (40). This proves that
we can choose a null-sequence \((\varepsilon_{j,n})_{n \in \mathbb{N}} \subset [0, \infty)\) such that \((39)\) is fulfilled for every \(n \geq n_j\) and every \(z \in \mathcal{X} \times \mathcal{Y}\). Accordingly, there is an \(n_\ell \in \mathbb{N}\) and a sequence \((\varepsilon_{\ell,n})_{n \in \mathbb{N}} \subset [0, \infty)\) such that \(\lim_{n \to \infty} \varepsilon_{\ell,n} = 0\) and, for every \(n \geq n_\ell\) and for every \(z \in \mathcal{X} \times \mathcal{Y}\), assertion \((39)\) with \(j\) replaced by \(\ell\) is fulfilled. Then, due to \((29)\),

\[
|g_{D_n,\lambda_n,\ell}(z)| \leq \varepsilon_{\ell,n} + (c + \varepsilon_{\ell,n}) \cdot b'_a(z) \quad \forall z \in \mathcal{X} \times \mathcal{Y}, \ n \geq n_\ell. \quad (42)
\]

Define \(\varepsilon_n := \max\{\varepsilon_{j,n}, \varepsilon_{\ell,n}\}\) for every \(n \in \mathbb{N}\). Then, for every \(n \geq n_j\),

\[
\left| \frac{1}{n} \sum_{i=1}^{n} g_{D_n,\lambda_n,j}(z_i) g_{D_n,\lambda_n,\ell}(z_i) - \frac{1}{n} \sum_{i=1}^{n} g_{P,\lambda,0,j}(z_i) g_{P,\lambda,0,\ell}(z_i) \right| \leq \frac{1}{n} \sum_{i=1}^{n} |g_{D_n,\lambda_n,j}(z_i) - g_{P,\lambda,0,j}(z_i)| \leq \frac{1}{n} \sum_{i=1}^{n} \varepsilon_n + \varepsilon_n \cdot b'_a(z_i) \ \xrightarrow{n \to \infty} \ 0
\]

where convergence to 0 follows from \(\lim_{n \to \infty} \varepsilon_n = 0\) and \((36)\). That is, we have proven \((30)\).

In view of \((35)\), it suffices to prove

\[
\frac{1}{n} \sum_{i=1}^{n} g_{D_n,\lambda_n,j}(z_i) g_{D_n,\lambda_n,\ell}(z_i) - \frac{1}{n} \sum_{i=1}^{n} g_{P,\lambda,0,j}(z_i) g_{P,\lambda,0,\ell}(z_i) \ \xrightarrow{n \to \infty} \ 0 \quad (43)
\]

in order to prove \((31)\). According to \((29)\), \((39)\), and \((42)\),

\[
\left| \frac{1}{n} \sum_{i=1}^{n} g_{D_n,\lambda_n,j}(z_i) g_{D_n,\lambda_n,\ell}(z_i) - \frac{1}{n} \sum_{i=1}^{n} g_{P,\lambda,0,j}(z_i) g_{P,\lambda,0,\ell}(z_i) \right| \leq
\]

\[
\leq \frac{1}{n} \sum_{i=1}^{n} |g_{D_n,\lambda_n,j}(z_i) - g_{P,\lambda,0,j}(z_i)| \cdot |g_{D_n,\lambda_n,\ell}(z_i)| +
\]

\[
+ \frac{1}{n} \sum_{i=1}^{n} |g_{P,\lambda,0,j}(z_i)| \cdot |g_{D_n,\lambda_n,\ell}(z_i) - g_{P,\lambda,0,\ell}(z_i)| \leq
\]

\[
\leq \frac{1}{n} \sum_{i=1}^{n} (\varepsilon_n + \varepsilon_n \cdot b'_a(z_i)) \cdot (\varepsilon_n + (c + \varepsilon_n) \cdot b'_a(z_i)) +
\]

\[
+ \frac{1}{n} \sum_{i=1}^{n} c \cdot b'_a(z_i) \cdot (\varepsilon_n + \varepsilon_n \cdot b'_a(z_i)) =
\]

\[
= \varepsilon_n^2 + 2(\varepsilon_n c + \varepsilon_n^2) \cdot \frac{1}{n} \sum_{i=1}^{n} b'_a(z_i) + (2\varepsilon_n c + \varepsilon_n^2) \cdot \frac{1}{n} \sum_{i=1}^{n} b'_a(z_i)^2
\]

and the last line converges to 0 as \(\lim_{n \to \infty} \varepsilon_n = 0\) and due to \((36)\). \(\square\)
Proof of Remark 3.7: According to \cite{10} Lemma A.9, \( G \) is a \( P \)-Donsker class and, therefore, a \( P \)-Glivenko-Cantelli class almost surely; see \cite{27} p. 82. Hence, \( P_{D_n} \) converges to \( P \) in \( BS \) almost surely. It is shown in \cite{10} (46) and (47) that \( S : BS \to H, \mu \mapsto S(\mu) = f_{\mu, \lambda_0} \) is continuous in \( P \) and, therefore,

\[
f_{D_n, \lambda_n} \xrightarrow{\text{a.s.}} S(P_{D_n}) \xrightarrow{n \to \infty} S(P) = f_{P, \lambda_0}.
\]

\( \square \)

Lemma 7.2 Let the Assumptions 3.1 be fulfilled, let \( \lambda_0 \in (0, \infty) \), and let \( \psi : H \to \mathbb{R}^m \) be Hadamard-differentiable in \( f_{P, \lambda_0} \) with derivative \( \psi_{f_{P, \lambda_0}} \). For every \( j \in \{1, \ldots, m\} \), let \( \psi_{f_{P, \lambda_0}, j} \in H \) denote the \( j \)-th component of \( \psi_{f_{P, \lambda_0}} \). Let \( \Sigma_P \in \mathbb{R}^{m \times m} \) be the covariance matrix in Corollary 3.3. Assume that, for \( P_X(dx) \cdot \text{a.e.} \ x \in X \), there are \( y_1, y_2 \in \text{supp}(P(dy|x)) \) such that

\[
L'(x,y_1,f_{P, \lambda_0}(x)) \neq L'(x,y_2,f_{P, \lambda_0}(x)).
\]

Then,

\[
\Sigma_P \text{ has full rank } \iff \exists \ a \in \mathbb{R}^m \setminus \{0\} \text{ s.th. } a^T \psi_{f_{P, \lambda_0}} = 0 \quad P_X\text{-a.s.} \tag{45}
\]

Proof of Lemma 7.2: According to \cite{5}, we have

\[
\Sigma_P \text{ has full rank } \iff \exists \ a \in \mathbb{R}^m \setminus \{0\}, \ c \in \mathbb{R} : \ a^T g_{P, \lambda_0} = c \quad P\text{-a.s.} \tag{46}
\]

It is a direct consequence of the definition of the continuous linear operator \( K_P \) that \( K_P \) is self-adjoint and, accordingly, \( K_P^{-1} \) is again self-adjoint; see \cite{9} Lemma VI.2.10. Hence, according to the reproducing property, we get

\[
g_{P, \lambda_0}(x,y) = -L'_{f_P, \lambda_0}(x,y)\langle \psi_{f_P, \lambda_0}, K_P^{-1}(\Phi(x)) \rangle_H = L'_{f_P, \lambda_0}(x,y)\langle K_P^{-1}(\psi_{f_P, \lambda_0}), \Phi(x) \rangle_H = -L'_{f_P, \lambda_0}(x,y)[K_P^{-1}(\psi_{f_P, \lambda_0}^j)](x)
\]

and, therefore,

\[
a^T g_{P, \lambda_0} = -L'_{f_P, \lambda_0} K_P^{-1}(a^T \psi_{f_P, \lambda_0}) \quad \forall \ a \in \mathbb{R}^m. \tag{48}
\]

It will be shown below that, for every \( f \in H \),

\[
f = 0 \quad P\text{-a.s.} \iff K_P^{-1}(f) = 0 \quad P\text{-a.s.} \tag{49}
\]

By use of these preparations and \cite{49}, the proof of (45) can be done quickly: First, assume that there is an \( a \in \mathbb{R}^m \setminus \{0\} \) such that \( a^T \psi_{f_P, \lambda_0} = 0 \quad P_X\text{-a.s.} \).
Then, it follows from (49), (48), and (46) that $\Sigma$ does not have full rank. That is we have proven “$\Rightarrow$” in (45). Next, in order to prove “$\Leftarrow$” in (45), assume that $\Sigma$ does not have full rank. Then, according to (46) and (48), there is an $a \in \mathbb{R}^m \setminus \{0\}$ and a $c \in \mathbb{R}$ such that, for $P_X(dx)$-a.e. $x \in \mathcal{X}$

$$-L'_{f_{P,\lambda_0}}(x,\cdot)[K^{-1}_P(a^T\psi'_{f_{P,\lambda_0}})](x) = c \quad P(\cdot|x)-\text{a.s.}$$

Hence, for $P_X(dx)$-a.e. $x \in \mathcal{X}$, it follows from (44) and continuity of $y \mapsto L'_{f_{P,\lambda_0}}(x,y)$ that

$$[K^{-1}_P(a^T\psi'_{f_{P,\lambda_0}})](x) = 0 .$$

According to (49), this implies that $a^T\psi'_{f_{P,\lambda_0}} = 0$ $P_X$-a.s. That is, we have proven “$\Leftarrow$” in (45).

Now, it only remains to prove statement (49). To this end, define $\mathcal{X} := \text{supp}(P_X)$, let $P_{\mathcal{X}}$ be the restriction of $P_X$ on the Borel-$\sigma$-algebra of $\mathcal{X}$, and let $\mathcal{P}$ be the probability measure on $\mathcal{X} \times \mathcal{Y}$ defined by

$$\mathcal{P}(B) = \int \int I_B(x,y) P(dy|x) P_X(dx)$$

for every $B$ in the Borel-$\sigma$-algebra of $\mathcal{X} \times \mathcal{Y}$. In addition, let $\overline{k}$ be the restriction of the kernel $k$ on $\mathcal{X} \times \mathcal{X}$ and $\overline{\Phi}$ the corresponding canonical feature map. Then, the RKHS of $\overline{k}$ is

$$\overline{H} := \{ \overline{f} : \mathcal{X} \rightarrow \mathbb{R} \mid \overline{f} \text{ is the restriction of an } f \in H \text{ on } \mathcal{X} \} ;$$

see e.g. [10 § 4.2]. For every $f \in H$, let $\overline{f}$ denote the restriction of $f$ on $\mathcal{X}$. Define

$$\overline{K}_P : \overline{H} \rightarrow \overline{H} , \quad \overline{f} \mapsto 2\lambda_0 \overline{f} + \int L'_{f_{P,\lambda_0}}(x,y) \overline{f}(x) \overline{\Phi}(x) \mathcal{P}(d(x,y)) .$$

As Assumption 3.1 is also fulfilled for $\mathcal{X}$ and $\mathcal{P}$ instead of $\mathcal{X}$ and $P$, it follows from [10 Lemma A.5] that $\overline{K}_P$ is invertible. The definitions imply $\overline{K}(f) = \overline{K}_P(f)$ for every $f \in H$ and, therefore,

$$\overline{K}_P(\overline{K}_P^{-1}(f)) = \overline{K}_P(\overline{K}_P^{-1}(f)) = \overline{f} \quad \forall f \in H .$$

Hence

$$\overline{K}_P^{-1}(f) = \overline{K}_P^{-1}(f) \quad \forall \overline{f} \in \overline{H} .$$

(51)
Since \( \overline{k} \) is continuous and \( \text{supp}(P_{\overline{X}}) = \overline{X} \), it follows from [21, Exercise 4.6] that, for every \( f \in H \),

\[
f = 0 \quad P_X\text{-a.s.} \iff \overline{f} = 0
\]

Hence, for every \( f \in H \),

\[
K_P^{-1}(f) = 0 \quad P_X\text{-a.s.} \iff K_P^{-1}(\overline{f}) = 0 \iff \overline{f} = 0 \iff f = 0 \quad P_X\text{-a.s.}
\]

\[\square\]

**Proof of Theorem 3.9:** Since taking the square root of a symmetric positive definite matrix is continuous, see e.g. [18, §7.8, Exercise 1], it follows from Theorem 3.6 that \( \hat{\Sigma}_n(D_n, \Lambda_n)^{1/2} \rightarrow \Sigma_P^{1/2} \) almost surely for \( n \rightarrow \infty \). Hence, Corollary 3.3 yields

\[
\sqrt{n} \cdot \hat{\Sigma}_n(D_n, \Lambda_n)^{-\frac{1}{2}} \left( \psi(f_{D_n, \Lambda_n}) - \psi(f_{P, \lambda_0}) \right) \sim \mathcal{N}_m(0, \text{Id}_{m \times m}) ;
\]

see e.g. [26, p. 11]. Finally, weak convergence, the continuous mapping theorem, the portmanteau theorem, and the definition of the chi-squared distribution imply

\[
\lim_{n \rightarrow \infty} Q \left( \psi(f_{P, \lambda_0}) \in C_{n, \alpha}(D_n, \Lambda_n) \right) = \lim_{n \rightarrow \infty} Q \left( \| \sqrt{n} \cdot \hat{\Sigma}_n(D_n, \Lambda_n)^{-\frac{1}{2}} (\psi(f_{P, \lambda_0}) - \psi(f_{D_n, \Lambda_n})) \|^2_{\mathbb{R}^m} \leq \chi^2_{m, \alpha} \right) = 1 - \alpha.
\]

\[\square\]

**Proof of Prop. 3.10:** Let \( \{ \Phi(x_{i1}), \ldots, \Phi(x_{ir}) \} \) be the maximal linearly independent subset of \( \{ \Phi(x_1), \ldots, \Phi(x_n) \} \) which defines \( B_{D_n} \) according to (12) and (13). Fix any \( x \in \mathcal{X} \) and any \( y \in \mathcal{Y} \). We have to find an \( f \in H \) such that \( K_{D_n, \lambda}(f) = \Phi(x) \). (The solution \( f \) depends on \( D_n \) and \( \lambda \) though this is not made explicit in the notation.) Hence, by using \( f(x_i) = \langle f, \Phi(x_i) \rangle_H \),

\[
\Phi(x) = K_{D_n, \lambda}(f) = 2\lambda f + \frac{1}{n} \sum_{i=1}^n L''_{f_{D_n, \lambda}}(x_i, y_i) \langle f, \Phi(x_i) \rangle_H \Phi(x_i).
\]

Rearranging this equality yields

\[
f = \frac{1}{2\lambda} \Phi(x) - \frac{1}{2n\lambda} \sum_{i=1}^n L''_{f_{D_n, \lambda}}(x_i, y_i) \langle f, \Phi(x_i) \rangle_H \Phi(x_i)
\]
and, therefore,
\[ f = \frac{1}{2\lambda} \Phi(x) + h \quad \text{for some } h \in \text{lin}\{\Phi(x_1), \ldots, \Phi(x_n)\}. \]  
(54)

Define
\[ w_i := -\frac{1}{2n\lambda} L''_{f_{D_n,\lambda}}(x_i, y_i) k(x_i, x) \quad \forall \ i \in \{1, \ldots, n\} \]
and \( w := (w_1, \ldots, w_n)\). Putting (54) into (53) again and a simple rearranging of the resulting equation lead to
\[ 2\lambda h + \frac{1}{n} \sum_{i=1}^{n} L''_{f_{D_n,\lambda}}(x_i, y_i) \langle h, \Phi(x_i) \rangle_H \Phi(x_i) = \sum_{i=1}^{n} w_i \Phi(x_i). \]  
(55)

That is, \( f \) solves (53) if and only if \( f \) is of form (54) where \( h \) solves (55). Next, define the linear map
\[ \gamma : \text{lin}\{\Phi(x_1), \ldots, \Phi(x_n)\} \rightarrow \text{lin}\{\Phi(x_1), \ldots, \Phi(x_n)\} \]
by
\[ \gamma(h) = 2\lambda h + \frac{1}{n} \sum_{i=1}^{n} L''_{f_{D_n,\lambda}}(x_i, y_i) \langle h, \Phi(x_i) \rangle_H \Phi(x_i) \]
for every \( h \in \text{lin}\{\Phi(x_1), \ldots, \Phi(x_n)\} \). That is, in order to find \( h \) which fulfills (55) we have to find \( \alpha_1, \ldots, \alpha_n \in \mathbb{R} \) such that
\[ \gamma \left( \sum_{i=1}^{n} \alpha_i \Phi(x_i) \right) = \sum_{i=1}^{n} w_i \Phi(x_i). \]  
(56)

Existence of a solution \( h \) and therefore, of \( \alpha_1, \ldots, \alpha_n \) is guaranteed as \( K_{D_n,\lambda} \) is invertible. Let \( a_{\ell i} \) be the \((\ell, i)\)-entry of the matrix \( A_{D_n,\lambda}, \ell, i \in \{1, \ldots, n\} \). According to the definition of \( A_{D_n,\lambda} \),
\[ \gamma(\Phi(x_i)) = \sum_{\ell=1}^{n} a_{\ell i} \Phi(x_\ell) \quad \forall \ i \in \{1, \ldots, n\}. \]  
(57)

It follows from
\[ \sum_{i=1}^{n} w_i \Phi(x_i) \quad \sum_{i=1}^{n} w_i \left( \sum_{j=1}^{r} \beta_{j i} \Phi(x_j) \right) = \sum_{j=1}^{r} \left( \sum_{i=1}^{n} \beta_{j i} w_i \right) \Phi(x_j) \]
and

\[
\gamma \left( \sum_{i=1}^{n} \alpha_i \Phi(x_i) \right) = \sum_{i=1}^{n} \alpha_i \gamma(\Phi(x_i)) = \sum_{i=1}^{n} \alpha_i \sum_{\ell=1}^{n} a_{\ell i} \Phi(x_\ell) = \gamma \left( \sum_{i=1}^{n} \alpha_i \sum_{\ell=1}^{n} a_{\ell i} \Phi(x_\ell) \right)
\]

that \( \alpha := (\alpha_1, \ldots, \alpha_n)^T \) is a solution of (56) if and only if

\[
\sum_{j=1}^{r} \left( \sum_{i=1}^{n} \sum_{\ell=1}^{n} \beta_{j\ell} a_{\ell i} \alpha_i \right) \Phi(x_{ij}) = \sum_{j=1}^{r} \left( \sum_{i=1}^{n} \beta_{j i} w_i \right) \Phi(x_{ij}). \tag{58}
\]

Linear independence of \( \Phi(x_{i_1}), \ldots, \Phi(x_{i_r}) \) implies that (58) is equivalent to

\[
\sum_{i=1}^{n} \beta_{ji} w_i = \sum_{i=1}^{n} \sum_{\ell=1}^{n} \beta_{j\ell} a_{\ell i} \alpha_i \quad \forall j \in \{1, \ldots, r\}
\]

or, in matrix notation,

\[
B_{Dn} \cdot w = B_{Dn} A_{Dn,\lambda} \cdot \alpha. \tag{59}
\]

Summing up, we have proven that \( \alpha \in \mathbb{R}^n \) is a solution of (56) if and only if \( \alpha \) solves (59). As already stated above, a solution of (56) and, therefore, of (59) exists. Hence,

\[
\alpha := (B_{Dn} A_{Dn,\lambda})^{-1} B_{Dn} w
\]

solves (59) and, therefore (56).

\[ \square \]

Acknowledgment

I would like to thank Andreas Christmann and Thoralf Mildenberger for discussions and valuable suggestions.

References

[1] A. Berlinet and C. Thomas-Agnan. *Reproducing kernel Hilbert spaces in probability and statistics*. Kluwer Academic Publishers, Boston, MA, 2004.
[2] G. Blanchard, O. Bousquet, and P. Massart. Statistical performance of support vector machines. *The Annals of Statistics*, 36:489–531, 2008.

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

[4] B. Caputo, K. Sim, F. Furesjo, and A. Smola. Appearance-based object recognition using SVMs: Which kernel should I use? In *Proceedings of NIPS workshop on Statistical methods for computational experiments in visual processing and computer vision*, Whistler, 2002.

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

[6] K. De Brabanter, J. De Brabanter, J. A. K. Suykens, and B. De Moor. Approximate confidence and prediction intervals for least squares support vector regression. *IEEE Transactions on Neural Networks*, 22:110–120, 2011.

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

[8] L. Devroye. Any discrimination rule can have an arbitrarily bad probability of error for finite sample size. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 4:154–157, 1982.

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

[10] R. Hable. Asymptotic normality of support vector machine variants and other regularized kernel methods. *Journal of Multivariate Analysis*, 106:92–117, 2012.

[11] R. Hable and A. Christmann. On qualitative robustness of support vector machines. *Journal of Multivariate Analysis*, 102:993–1007, 2011.

[12] T. Hofmann, B. Schölkopf, and A. J. Smola. Kernel methods in machine learning. *The Annals of Statistics*, 36:1171–1220, 2008.

[13] 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.
[14] A. Karatzoglou, A. Smola, K. Hornik, and A. Zeileis. kernlab – an S4 package for kernel methods in R. Journal of Statistical Software, 11: 1–20, 2004.

[15] 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.

[16] S. Mendelson and J. Neeman. Regularization in kernel learning. The Annals of Statistics, 38:526–565, 2010.

[17] B. Schölkopf and A. J. Smola. Learning with kernels. MIT Press, Cambridge, 2002.

[18] D. Serre. Matrices. Theory and applications. Springer-Verlag, New York, 2002.

[19] I. Steinwart. Support vector machines are universally consistent. Journal of Complexity, 18:768–791, 2002.

[20] I. Steinwart. Consistency of support vector machines and other regularized kernel classifiers. IEEE Transactions on Information Theory, 51:128–142, 2005.

[21] I. Steinwart and A. Christmann. Support vector machines. Springer, New York, 2008.

[22] I. Steinwart and A. Christmann. Estimating conditional quantiles with the help of the pinball loss. Bernoulli, 17:211–225, 2011.

[23] I. Steinwart and C. Scovel. Fast rates for support vector machines using Gaussian kernels. The Annals of Statistics, 35:575–607, 2007.

[24] 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.

[25] A. B. Tsybakov. Optimal aggregation of classifiers in statistical learning. The Annals of Statistics, 32:135–166, 2004.

[26] A. van der Vaart. Asymptotic statistics. Cambridge University Press, Cambridge, 1998.

[27] A. van der Vaart and J. Wellner. Weak convergence and empirical processes. With applications to statistics. Springer, New York, 1996.
[28] V. N. Vapnik. *Statistical learning theory*. John Wiley & Sons, New York, 1998.

[29] T. Zhang. Statistical behavior and consistency of classification methods based on convex risk minimization. *Annals of Statistics*, 32:56–85, 2004.