Journal of Machine Learning Research 14 (2013) 153-186 Submitted 6/11; Revised 7/12; Published 1/13

Universal Consistency of Localized Versions of Regularized Kernel Methods

Robert Hable ROBERT.HABLE@UNI-BAYREUTH.DE
Department of Mathematics
University of Bayreuth
D-95440 Bayreuth, Germany

Editor: Gábor Lugosi

Abstract

In supervised learning problems, global and local learning algorithms are used. In contrast to global learning algorithms, the prediction of a local learning algorithm in a testing point is only based on training data which are close to the testing point. Every global algorithm such as support vector machines (SVM) can be localized in the following way: in every testing point, the (global) learning algorithm is not applied to the whole training data but only to the k nearest neighbors (kNN) of the testing point. In case of support vector machines, the success of such mixtures of SVM and kNN (called SVM-KNN) has been shown in extensive simulation studies and also for real data sets but only little has been known on theoretical properties so far. In the present article, it is shown how a large class of regularized kernel methods (including SVM) can be localized in order to get a universally consistent learning algorithm.

Keywords: machine learning, regularized kernel methods, localization, SVM, k-nearest neighbors, SVM-KNN

1. Introduction

In a supervised learning problem, the goal is to predict the value y of an unobserved output variable Y after observing the value x of an input variable X. A predictor is a function f which maps the observed input value x (called testing data point) to a prediction $f(x)$ of the unobserved output value y. Choosing a predictor $f = f_{D_n}$ is done on base of previously observed data $D_n = \big((x_1,y_1),\ldots,(x_n,y_n)\big)$ (called trainig data). A learning algorithm is a function $D_n \mapsto f_{D_n}$ which maps training data D_n to a predictor f_{D_n}. Among the learning algorithms commonly used in machine learning, there are local and global algorithms. The most prominent example of a local algorithm is *k-nearest neighbors* (kNN). In case of a local algorithm $D_n \mapsto f_{D_n}$, the prediction $f_{D_n}(x)$ in a testing data point x is not based on the whole training data but only on those training data points (x_i,y_i) which are close to x. In case of a global algorithm, choosing a predictor f_{D_n} is based on a global criterion—such as (penalized) empirical risk minimization—and, accordingly, the prediction $f_{D_n}(x)$ in a point x can also be based on training data points (x_i,y_i) which are not close to x. Typical examples of global algorithms are regularized kernel methods such as *support vector machines* (SVM).

Global algorithms have disadvantages if the complexity of the optimal predictor varies for different areas of the input space. For example, in one part of the the input space, an optimal predictor might be a very simple function and, in another part, it might be a highly complex and volatile func-

tion. This is a problem for global algorithms because the complexity of the selected predictor f_{D_n} is usually regularized by one or several hyperparameters which are fixed for the whole input space. One way to overcome this problem is to separate the input space into several parts in a first step and to separately use a global algorithm for each of the separated parts. For example, the input space is separated by use of decision trees and then SVMs are separately applied on the separated parts of the input space; see, for example, Bennett and Blue (1998), Wu et al. (1999), and Chang et al. (2010). Another possibility is to "localize" a global algorithm. This can be done in the following way: (1) select a few training data points which are close to the testing data point, (2) determine a predictor based on the selected training data points by use of a (global) learning algorithm, and (3) calculate the prediction in the testing data point. A number of algorithms which have been suggested in the literature can be described in this way. These algorithms only differ in the way how data points are selected in (1) and which learning algorithm is used in (2). An early investigation of such methods is Bottou and Vapnik (1992) and Vapnik and Bottou (1993). A number of recent articles apply such an approach to support vector machines (SVM). That is, SVM is used in (2), but there are differences in (1): In Zhang et al. (2006), data points are selected in the same way as for kNN. That is, the prediction in a testing point x is given by that SVM which is calculated based on the k_n training points which are nearest to x; the natural number k_n acts as a hyperparameter. In order to decide which training points are the k_n closest ones to x, a metric on the input space is needed. Zhang et al. (2006) considers different metrics. As this approach is a mixture between kNN and SVM, it is called SVM-KNN. Independently, a similar approach has been developed by E. Blanzieri and others. The main difference to Zhang et al. (2006) is that distances (for selecting the k_n nearest neighbors) are not measured in the input space but in the feature space (i.e., in the RKHS associated with the kernel of the SVM). This approach has been extensively studied in experimental comparisons in Blanzieri and Bryl (2007a), Blanzieri and Bryl (2007b), Segata and Blanzieri (2009) and Blanzieri and Melgani (2008) where the latter publication also derives a local bound on the generalization error. Another slightly different approach is developed in Cheng et al. (2007) and Cheng et al. (2010). There, data points are not selected according to a fixed number k_n of nearest neighbors as in kNN; instead, those training data points are selected which are contained in a fixed neighborhood about the testing point x. That is, not the number of testing points in the neighborhood is fixed (as in kNN), but the area of the neighborhood is fixed. In addition, it is also possible to downweight testing points depending on their distance to the testing point x.

Though all of these approaches have been extensively studied on simulated and real-world data and their success has experimentally been shown, only little is known on theoretical properties so far. In this article, it is shown that some SVM-KNN approaches are universally consistent. Though the above cited approaches only consider SVMs for classification (using the hinge loss) and linear kernels, the following theoretical investigation allows for a large class of loss functions and kernels. That is, not only SVMs but also general regularized kernel methods are considered for classification and regression as well. Here, k_n nearest neighbors are selected by use of the ordinary Euclidean metric on the input space $\mathcal{X} \subset \mathbb{R}^p$ so that this approach is closest to Zhang et al. (2006). All methods based on a kNN approach are faced with the problem of distance ties. This means that, in general, the set of the k_n nearest neighbors to a testing point x is not necessarily unique because different testing points might have the same distance to x. In case of distance ties, a number of tie-breaking strategies have been suggested in the literature; see, for example, Devroye et al. (1994, § 1). E.g. a simple tie-breaking strategy is to generate artificial additional covariates $U_1, \ldots, U_n$ i.i.d. from the uniform distribution on $[0, \varepsilon]$ for some small $\varepsilon > 0$. Then, for the new input variables $X_i' := (X_i, U_i)$,

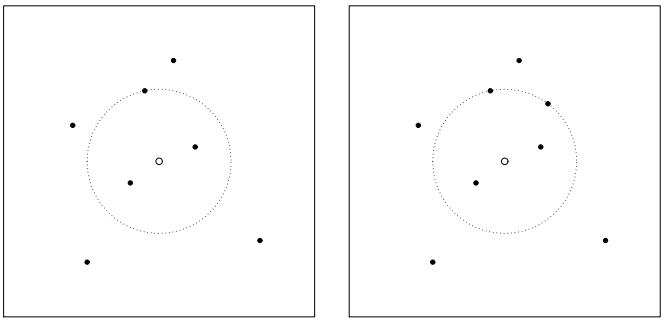

Figure 1: Neighborhood (dotted circle) determined by the k nearest neighbors of a testing point (empty point) for $k = 3$. The left figure shows a situation without distance ties at the border of the neighborhood (dotted circle). The right figure shows a situation with distance ties at the border of the neighborhood (empty point): only one of the two data points (filled points) at the border may belong to the $k = 3$ nearest neighbors; choosing between these two candidates is done by randomization here.

distance ties only occur with zero probability. The drawback of this method is that ε has to be chosen in advance and, in particular if ε is not small enough, this tie-breaking strategy changes the results even if there are no distance ties. Therefore, we use a different strategy where, in case of a distance tie, the k nearest neigbors are chosen by randomization; see Figure 1. Technically, this is done by artificially generated covariates $U_1, \ldots, U_n$ i.i.d. from the uniform distribution on $[0,1]$ where—in contrast to the simple tie-breaking strategy mentioned above—U_i is only taken into account in case of a distance tie in X_i.

It has to be pointed out that the approach of this article differs from the one in Zakai and Ritov (2009); see also Zakai (2008). There, it is shown that every consistent learning algorithm is in a sense localizable. On the one hand, this is of great theoretical importance because, roughly speaking, it says that global methods as SVMs asymptotically act like local methods. On the other hand, this also shows that any consistent method can be localized in a way so that the local version is again consistent. By a superficial inspection of these results, one might suggest that, essentially, this would already show consistency of any localized method such as SVM-KNN. However, this is not the case and these results cannot be used offhand in order to prove consistency of SVM-KNN: Firstly, the way how the methods are localized completely differ. In Zakai and Ritov (2009), localizing is not done by fixed numbers k_n of nearest neighbors (as in kNN and SVM-KNN) but by fixed sizes (radii) R_n of neighborhoods (similar as in Cheng et al. (2010)). Using fixed sizes (radii) of neighborhoods is more convenient for theoretical investigations because whether a data point x_{i_0} lies in such a neighborhood only depends on this data point; that is, variables indicating whether data points belong to such a neighborhood are i.i.d. In contrast, whether a data point x_{i_0} belongs to the k_n nearest neighbors depends on the whole sample; that is, the corresponding indicator variables are not independent and one has to work with random sets of indexes. In particular, the kNN-approach leads to random sizes of neighborhoods which depend on the testing point x while Zakai and Ritov (2009) deal with deterministic sequences of radii R_n which do not depend on the testing point x.

Secondly, due to the generality of the investigation in Zakai and Ritov (2009), it is only shown there that a (deterministic) sequence of radii R_n *exists* such that a suitably,[1] localized method is consistent. This indicates that looking for consistent localized methods may be promising; however, for practical purposes, mere existence is not enough and one also has to know how to choose such entities like R_n in order to get a consistent method. In the special case of SVM-KNN, the main result of the present article precisely specifies possible choices of all involved entities (hyperparameters etc.) which guarantee consistency.

For kNN, consistency requires that the number of selected neighbors k_n goes to infinity but not too fast for $n \to \infty$. Clearly, this will also be crucial for SVM-KNN but, now, an additional difficulty arises: the calculation of the SVM (or any other regularized kernel method) depends on a regularization parameter λ_n which determines to what extend the complexity of a predictor is penalized (in order to avoid overfitting). Consistency of SVMs is only guaranteed if λ_n converges to 0 but not too fast. Accordingly, in case of SVM-KNN, the interplay between the convergence of k_n and the convergence of λ_n is crucial. Theorem 1 below gives precise conditions on k_n and λ_n which guarantee consistency of SVM-KNN. In Therorem 1, it is assumed that k_n, $n \in \mathbb{N}$, is a predefined deterministic sequence. The regularization parameters $\lambda_n = \lambda_{D_n,x}$ are based on the training data and can, to some extend, also be chosen in a data-driven way, for example, by cross-validation. In addition, the choice of the regularization parameter is local, that is, depends on the testing point x. This enables a local regularization of the complexity of the predictor which is an important motivation for localizing a global algorithm as already stated above.

Local approaches such as SVM-KNN are computationally very efficient if the number of testing points is small. However, if the number of testing points is large, then such methods are burdened with high computational costs of the testing phase. Therefore, variants of SVM-KNN have been proposed in Cheng et al. (2007) and Segata and Blanzieri (2010). For example, in Segata and Blanzieri (2010), the computational complexity is reduced by the following modification: the SVM is not calculated on base of the k-nearest neighbors of the testing point but on base of the k-nearest neighbors of a certain training point which is close to the testing point. In this way, only a relatively small number of SVMs has to be calculated. If k is reasonable small (and fixed), then training scales as $O(n\log(n))$ and testing scales as $O(\log(n))$ in the number of training points.

The article is organized as follows: Section 2 recalls the precise mathematical definitions of kNN, regularized kernel methods (in particular, SVM) and SVM-KNN as investigated here. Section 3 contains the main result, that is, consistency of SVM-KNN, Section 4 investigates an illustrative example and Section 5 contains some concluding remarks. All proofs and auxiliary results are given in the Appendix.

2. Setup: kNN, SVM and SVM-KNN

Let $(\Omega, \mathcal{A}, Q)$ be a probability space, let $\mathcal{X}$ be an open subset of $\mathbb{R}^d$, and let $\mathcal{Y}$ be a closed subset of $\mathbb{R}$. For any (topological) space $\mathcal{W}$, its Borel-σ-algebra is denoted by $\mathfrak{B}_{\mathcal{W}}$. Let

$$X_1, \ldots, X_n : (\Omega, \mathcal{A}, Q) \longrightarrow \left(\mathcal{X}, \mathfrak{B}_{\mathcal{X}}\right) \quad \text{and} \quad Y_1, \ldots, Y_n : (\Omega, \mathcal{A}, Q) \longrightarrow \left(\mathcal{Y}, \mathfrak{B}_{\mathcal{Y}}\right)$$

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 $\left(\mathcal{X} \times \mathcal{Y}, \mathfrak{B}_{\mathcal{X}\times\mathcal{Y}}\right)$. In order to find a prediction

1. In Zakai and Ritov (2009), localizing also involves a smoothing operation around the testing point.

$y = f(\xi)$ for a point $\xi \in \mathcal{X}$, a kNN-rule is based on the k_n nearest neighbors of ξ. The k_n nearest neighbors of $\xi \in \mathbb{R}^p$ within $x_1, \ldots, x_n \in \mathbb{R}^p$ are given by an index set $I \subset \{1, \ldots, n\}$ such that

$$\sharp(I) = k_n \qquad \text{and} \qquad \max_{i \in I} |x_i - \xi| \; < \; \min_{j \notin I} |x_j - \xi| \; . \tag{1}$$

However, in case of distance ties, some observations x_i and x_j have the same distance to ξ (i.e., $|x_i - \xi| = |x_j - \xi|$) so that the k_n nearest neighbors are not unique and an index set I as defined above does not exist. In order to break distance ties, we use randomization (see also Figure 1) as done in (Devroye et al., 1994, p. 1373f): We artificially generate data from random variables $U_1, \ldots, U_n$ which are uniformly distributed on $(0,1)$ and such that $(X_1, Y_1), \ldots, (X_n, Y_n), U_1, \ldots, U_n$ are independent. Define $Z_i := (X_i, U_i)$ for every $i \in \{1, \ldots, n\}$. That is, we observe $(Z_1, Y_1), \ldots, (Z_n, Y_n)$ now. Define

$$\mathbf{D}_n \; := \; \big((Z_1, Y_1), \ldots, (Z_n, Y_n)\big) \qquad \forall\, n \in \mathbb{N} \; .$$

We say that $z_i = (x_i, u_i)$ is (strictly) closer to $\zeta = (\xi, u) \in \mathcal{X} \times (0,1)$ than $z_j = (x_j, u_j)$ if $|x_i - \xi| < |x_j - \xi|$; and, in case of a distance tie $|x_i - \xi| = |x_j - \xi|$, we say that $z_i = (x_i, u_i)$ is (strictly) closer to $\zeta = (\xi, u)$ than $z_j = (x_j, u_j)$, if $|u_i - u| < |u_j - u|$. That is, we use some kind of a lexicographic order which guarantees that nothing changes if there are no distance ties. Note that there can also be distance ties for the u_i but these only occur with zero probability. The following is a precise definition of "nearest neighbors" which also takes into account distance ties in the x_i and the u_i. For $n \in \mathbb{N}$, let $k_n \in \{1, \ldots, n\}$. Take any $z_1 = (x_1, u_1), \ldots, z_n = (x_n, u_n), \zeta = (\xi, u) \in \mathbb{R}^p \times (0,1)$ such that there is a $\tau_n(z_1, \ldots, z_n, \zeta) = I \subset \{1, \ldots, n\}$ such that

$$\sharp(I) = k_n \, , \quad \max_{i \in I} |x_i - \xi| \leq \min_{i \notin I} |x_i - \xi| \quad \text{and} \quad \max_{j \in I \cap \mathcal{J}} |u_j - u| \; < \; \min_{j \in \mathcal{J} \setminus I} |u_j - u| \tag{2}$$

where

$$\mathcal{J} \; = \; \Big\{ j \in \{1, \ldots, n\} \Big| \; |x_j - \xi| = \max_{i \in I} |x_i - \xi| \Big\} \; . \tag{3}$$

If such a set $\tau_n(z_1, \ldots, z_n, \zeta) = I$ exists, it is unique. If it does not exist, there are also distance ties in the u_i and we arbitrarily define $\tau_n(z_1, \ldots, z_n, \zeta) := \{1, \ldots, k_n\}$ in this case. Since distance ties in the u_i occur with zero probability, the definition of $\tau_n(z_1, \ldots, z_n, \zeta)$ is meaningless in this case; it is only important to assure measurability of $\tau_n : (z_1, \ldots, z_n, \zeta) \mapsto \tau_n(z_1, \ldots, z_n, \zeta)$; see Appendix B. So, definition (2) and (3) is a modification of (1) in order to deal with distance ties in the x_i. Note that, due to the lexicographic order, the values u_i and u are only relevant in case of distance ties (at the border of the neighborhood given by the k_n nearest neighbors).

Next, define

$$I_{n,\zeta}(\omega) \; := \; \tau_n\big(Z_1(\omega), \ldots, Z_n(\omega), \zeta\big) \qquad \forall\, \omega \in \Omega \, , \quad \forall\, \zeta \in \mathbb{R}^p \times (0,1) \; . \tag{4}$$

That is, $I_{n,\zeta}$ contains the indexes of the k_n-nearest neighbors of ζ. Let $i_1 < i_2 < \ldots < i_{k_n}$ be the (ordered) elements of $I_{n,\zeta}$. Then, the vector of the k_n-nearest neighbors is

$$\mathbf{D}_{n,\zeta} \; := \; \big((Z_{i_1}, Y_{i_1}), \ldots, (Z_{i_{k_n}}, Y_{i_{k_n}})\big) \; . \tag{5}$$

The prediction of the ordinary kNN-rule in ξ is given by the mean

$$\frac{1}{k_n} \sum_{i \in I_{n,\zeta}} Y_i \; .$$

The SVM-KNN method replaces the mean by an SVM. To this end, we recall the definition of SVMs; here, the term "SVM" is used in a wide sense which covers many regularized kernel-based learning algorithms for classification and regression as well; see, for example, Steinwart and Christmann (2008) for these methods.

A measurable map $L : \mathcal{Y} \times \mathbb{R} \to [0,\infty)$ is called *loss function*. A loss function L is called *convex* loss function if it is convex in its second argument, that is, $t \mapsto L(y,t)$ is convex for every $y \in \mathcal{Y}$. The *risk* of a measurable function $f : \mathcal{X} \to \mathbb{R}$ is defined by

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

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. 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, for example, Schölkopf and Smola (2002) or Steinwart and Christmann (2008) 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

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

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

$$\mathcal{R}_P(f_{P,\lambda}) + \lambda\|f_{P,\lambda}\|_H^2 \;=\; \inf_{f\in H}\big(\mathcal{R}_P(f) + \lambda\|f\|_H^2\big)\,. \tag{6}$$

The *empirical SVM* $f_{D_n,\lambda_{D_n}}$ is that function $f \in H$ which minimizes

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

in H for the data $D_n = ((x_1,y_1),\ldots,(x_n,y_n)) \in (\mathcal{X}\times\mathcal{Y})^n$ and a regularization parameter $\lambda_{D_n} \in (0,\infty)$ which is chosen in a data-driven way (e.g., by cross-validation) in applications so that it typically depends on the data. The empirical support vector machine $f_{D_n,\lambda_{D_n}}$ uniquely exists for every $\lambda_{D_n} \in (0,\infty)$ and every data-set $D_n \in (\mathcal{X}\times\mathcal{Y})^n$ if $t \mapsto L(y,t)$ is convex for every $y \in \mathcal{Y}$.

The prediction of the SVM-KNN learning algorithm in $\zeta = (\xi,u) \in \mathcal{X}\times(0,1)$ is given by $f_{\mathbf{D}_{n,\zeta},\Lambda_{n,\zeta}}(\xi)$ with

$$f_{\mathbf{D}_{n,\zeta},\Lambda_{n,\zeta}} \;=\; \arg\min_{f\in H}\left(\frac{1}{k_n}\sum_{i\in I_{n,\zeta}} L\big(Y_i, f(X_i)\big) + \Lambda_{n,\zeta}\|f\|_H^2\right) \tag{7}$$

where $\omega \mapsto \Lambda_{n,\zeta}(\omega)$ is a random regularization parameter depending on n and ζ. That is, the method calculates the empirical SVM $f_{\mathbf{D}_{n,\zeta},\Lambda_{n,\zeta}}$ for the k_n nearest neighbors (given by the index set $I_{n,\zeta}$) and uses the value $f_{\mathbf{D}_{n,\zeta},\Lambda_{n,\zeta}}(\xi)$ for the prediction in ζ. The empirical SVM minimizes the regularized empirical risk where the regularization is done in order to avoid overfitting. Note that—unlike most theoretical investigations on SVMs—the regularization parameter $\Lambda_{n,\zeta}$ is random and, here, also the index set $I_{n,\zeta}$ is random, that is, a set-valued random variable. We will assume that $\mathcal{Y} \subset [-M,M]$ for

some M so that the SVM-KNN can be clipped. The clipped version of the SVM-KNN is denoted by

$$\widehat{f}_{\mathbf{D}_{n,\zeta},\Lambda_{n,\zeta}}(\xi) = \begin{cases} M & f_{\mathbf{D}_{n,\zeta},\Lambda_{n,\zeta}}(\xi) > M \\ f_{\mathbf{D}_{n,\zeta},\Lambda_{n,\zeta}}(\xi) & \text{if} \quad f_{\mathbf{D}_{n,\zeta},\Lambda_{n,\zeta}}(\xi) \in [-M,M] \\ -M & f_{\mathbf{D}_{n,\zeta},\Lambda_{n,\zeta}}(\xi) < -M \end{cases} . \tag{8}$$

This means that we change the prediction to M (or $-M$) if $f_{\mathbf{D}_{n,\zeta},\Lambda_{n,\zeta}}(\xi)$ is larger (or smaller) than M (or $-M$). As we will assume that $\mathcal{Y} \subset [-M,M]$, predictions exceeding $[-M,M]$ are not sensible and, in these cases, clipping obviously improves the accuracy of our predictions.

3. Main Result

This section contains the main result, namely universal consistency of SVM-KNN where the term "SVM" is used in a broad sense. Instead of just SVMs in the original sense (i.e., classification using the hinge loss), a large class of regularized kernel methods for classification and regression as well is covered. However, as already mentioned in the introduction, not any combination of SVM and kNN is possible. In order to get consistency, the choice of the number of neighbors k_n and the data-driven local choice of the regularization parameter $\lambda = \Lambda_{n,\xi}$ needs some care. The following settings guarantee consistency of SVM-KNN. Possible choices for k_n and λ_n are, for example, $k_n = b \cdot n^{0.75}$ for $b \in (0,1]$ and $\lambda_n = a \cdot n^{-0.15}$ for $a \in (0,\infty)$, $n \in \mathbb{N}$.

Settings: Choose a sequence $k_n \in \mathbb{N}$, $n \in \mathbb{N}$, such that

$$k_1 \le k_2 \le k_3 \le \ldots \le \lim_{n\to\infty} k_n = \infty \qquad \text{and} \qquad \frac{k_n}{n} \searrow 0 \quad \text{for } n \to \infty,$$

and a sequence $\lambda_n \in (0,\infty)$, $n \in \mathbb{N}$, such that

$$\lim_{n\to\infty} \lambda_n = 0 \qquad \text{and} \qquad \lim_{n\to\infty} \lambda_n^{\frac{3}{2}} \cdot \frac{k_n}{\sqrt{n}} = \infty \tag{9}$$

and a constant $c \in (0,\infty)$, and a sequence $c_n \in [0,\infty)$ such that $\lim_{n\to\infty} c_n/\sqrt{\lambda_n} = 0$. For every $\zeta = (\xi,u) \in \mathcal{X} \times (0,1)$, define

$$\tilde{\Lambda}_{n,\zeta} = \frac{1}{k_n} \sum_{i \in I_{n,\zeta}} |X_i - \xi|^{\frac{3}{2}}$$

and choose random regularization parameters $\Lambda_{n,\zeta}$ such that

$$\mathcal{X} \times (0,1) \times \Omega \to (0,\infty), \qquad (\xi,u,\omega) = (\zeta,\omega) \mapsto \Lambda_{n,\zeta}(\omega)$$

is measurable and

$$c \cdot \max\left\{\lambda_n, \tilde{\Lambda}_{n,\zeta}\right\} \le \Lambda_{n,\zeta} \le (c + c_n) \cdot \max\left\{\lambda_n, \tilde{\Lambda}_{n,\zeta}\right\} \qquad \forall \zeta \in \mathcal{X} \times (0,1) . \tag{10}$$

Let the kernel $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be continuously differentiable, bounded, and such that its RKHS H is non-degenerated in the following sense:

$$\text{for every } x \in \mathcal{X} \text{ there is an } f \in H \text{ such that } f(x) \neq 0 . \tag{11}$$

Theorem 1 *Let $\mathcal{X} \subset \mathbb{R}^p$ be an open subset and let $\mathcal{Y} \subset [-M,M]$ be closed. Let $L : [-M,M] \times \mathbb{R} \to [0,\infty)$ be a convex loss function with the following local Lipschitz property: there are some $b_0, b_1 \in (0,\infty)$ and $q \in [0,1]$ such that, for every $a \in (0,\infty)$,*

$$\sup_{y \in [-M,M]} \bigl|L(y,t_1) - L(y,t_2)\bigr| \;\le\; |L|_{a,1} \cdot |t_1 - t_2| \qquad \forall t_1, t_2 \in [-a,a] \tag{12}$$

for $|L|_{a,1} = b_0 + b_1 a^q$. In addition, assume that there is an increasing function $\ell : [0,\infty) \to [0,\infty)$ such that $\lim_{s \to 0} \ell(s) = 0$ and

$$\sup_{t \in [-M,M]} \bigl|L(y_1,t) - L(y_2,t)\bigr| \;\le\; \ell\bigl(|y_1 - y_2|\bigr) \qquad \forall y_1, y_2 \in [-M,M]\,. \tag{13}$$

Assume that $(X_1,Y_1),\ldots,(X_n,Y_n)$ are independent and identically distributed according to some unknown probability measure P on $\bigl(\mathcal{X} \times \mathcal{Y}, \mathfrak{B}_{\mathcal{X}\times\mathcal{Y}}\bigr)$ and let $U_1,\ldots,U_n$ be uniformly distributed on $(0,1)$ such that $(X_1,Y_1),\ldots,(X_n,Y_n),U_1,\ldots,U_n$ are independent.

Then, every SVM-KNN defined by (7,8) according to the above settings and clipped at M,

$$f_{\mathbf{D}_n} : \quad \zeta = (\xi, u) \;\mapsto\; \widehat{f}_{\mathbf{D}_{n,\zeta},\Lambda_{n,\zeta}}(\xi)$$

is risk-consistent, that is,

$$\mathcal{R}_P\bigl(f_{\mathbf{D}_n}\bigr) \xrightarrow[n\to\infty]{} \inf_{\substack{f:\mathcal{X}\to\mathbb{R} \\ \text{measurable}}} \mathcal{R}_P(f) \;=:\; \mathcal{R}_P^* \qquad \text{in probability.}$$

Essentially all commonly used loss functions satisfy assumptions (12) and (13): for example, the hinge loss and the logistic loss for classification, the ε-insensitive loss, the least squares loss, the absolute deviation loss, and the Huber loss for regression, and the pinball loss for quantile regression.

The property (11) of a nowhere degenerated RKHS H is a very weak property and replaces strong denseness properties of H which are typically needed in order to assure universal consistency of SVMs.

The settings include a data-driven local choice of the regularization parameter $\lambda = \Lambda_{n,\zeta}$. Here, "local" means that $\Lambda_{n,\zeta}$ depends on the testing point ζ. This is preferable because, in this way, it is possible to allow for different degrees of complexity on different areas of the input space. As already mentioned in the introduction, this is an important motivation for "localizing" a global algorithm. A simple rule of thumb for choosing $\Lambda_{n,\zeta}$ is to predefine a fixed $c \in (0,\infty)$ and use

$$\Lambda_{n,\zeta} \;=\; c \cdot \max\bigl\{\lambda_n, \tilde{\Lambda}_{n,\zeta}\bigr\}\,. \tag{14}$$

The deterministic λ_n prevents the regularization parameters from decreasing to 0 too fast and (9) controls the interplay between k_n and λ_n. (Recall that it is well known that classical SVMs are not consistent if the regularization parameters decrease to 0 too fast.) Note that the calculation of $\tilde{\Lambda}_{n,\zeta}$ is computationally fast as $I_{n,\zeta}$ (the index set of the k_n nearest neighbors) has to be calculated anyway. The behavior of $\tilde{\Lambda}_{n,\zeta}$ is reasonable: if the k_n nearest neighbors are relatively close to the testing point ζ, then $\tilde{\Lambda}_{n,\zeta}$ is relatively small which is favorable because this means that relatively many training points are close to ζ so that the predictor should be allowed to be relatively complex around ζ. Nevertheless, the rule of thumb suggested in (14) will not satisfactorily capture different

degrees of complexity in most cases. Then, it is possible to choose the regularization parameter on base of a (restricted) cross-validation or any other method for selecting the hyperparameter: choose a (very) small $c \in (0,\infty)$ and a (very) large $C \in (0,\infty)$, define $c_n := C\sqrt{\lambda_n}/\ln(n)$ and make sure that your selection method (e.g., cross validation) only picks a value from the interval

$$\left[c \cdot \max\left\{\lambda_n, \tilde{\Lambda}_{n,\zeta}\right\}, (c+c_n) \cdot \max\left\{\lambda_n, \tilde{\Lambda}_{n,\zeta}\right\}\right].$$

As it is assumed in Theorem 1 that $\lim_{n\to\infty} k_n/n = 0$ (i.e., the fraction of data points in the neighborhood diminishes), this SVM-KNN approach is rather a kNN-approach in which the simple (local) constant fitting is replaced by a more advanced (local) SVM fitting. That is, we follow a local modeling paradigm (see Györfi et al., 2002, § 2.1) just as done, for example, when generalizing the Nadaraya-Watson kernel estimator (constant fitting) to the local polynomial kernel estimator (polynomial fitting); for local polynomial fitting and the advantages of generalizing local constant fitting, see, for example, Fan and Gijbels (1996). In case of SVM-KNN, the advantage of generalizing constant fitting (kNN), has been demonstrated in extensive simulation studies in Zhang et al. (2006), Blanzieri and Bryl (2007a), Blanzieri and Bryl (2007b), Segata and Blanzieri (2009), and Blanzieri and Melgani (2008).

Instead, it would also be possible to assume that $\lim_{n\to\infty} k_n/n = 1$ so that the method (asymptotically) acts as an ordinary SVM. If convergence of the fraction k_n/n to 1 is fast enough, then universal consistency of such a method follows from universal consistency of SVM.

4. An Illustrative Example

It is commonly accepted in machine learning that there is no universally consistent learning algorithm which is always better than all other universally consistent learning algorithms and, for two different learning algorithms, there is always a situation in which one learning algorithm is better than the other one and there is also a situation in which it is the other way round; see, for example, (Devroye et al., 1996, § 1). The goal of this section is to illustrate where localizing SVMs provides some gain and where it does not. It has to be pointed out here that it is *not* the goal of this article or this section to empirically show the success of the SVM-KNN approach. This has previously been done; see the references cited in the introduction. The aim of this article is the proof of universal consistency and this section is only for illustrative purposes.

Let us consider the following model

$$Y_i \;=\; f_j(X_i) + \varepsilon_i\,, \qquad i \in \{1,\ldots,n\} \tag{15}$$

where, in the first scenario ($j = 1$), the regression function is given by

$$f_1(x) \;=\; 10(|x|-1)^2 \cdot \operatorname{sign}(x)\,, \qquad x \in [-1,1]$$

and, in the second scenario ($j = 2$), the regression function is given by

$$f_2(x) \;=\; 10x^2 \cdot \operatorname{sign}(x)\,, \qquad x \in [-1,1]\,.$$

As illustrated in Figure 2, the difference between f_1 and f_2 is that the parts of the functions on $(-1,0)$ and $(0,1)$ are interchanged. In both cases, $X_1,\ldots,X_n$ are i.i.d. drawn from the uniform distribution on $[-1,1]$ and $\varepsilon_1,\ldots,\varepsilon_n$ are i.i.d. drawn from $\mathcal{N}(0,\sigma^2)$ for $\sigma = 0.5$.

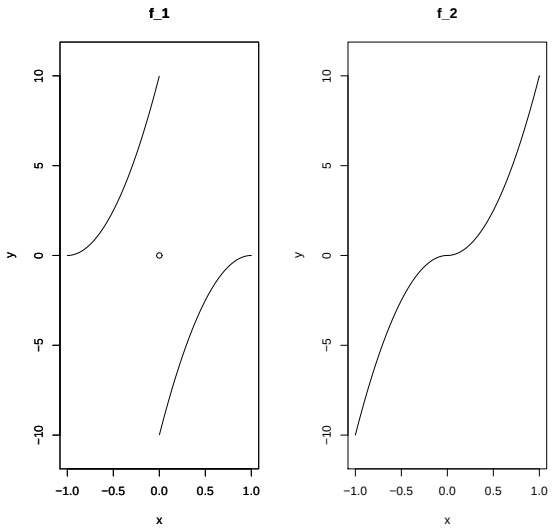

Figure 2: Graph of the regression functions $f_1(x) = 10(|x|-1)^2 \cdot \text{sign}(x)$ and $f_2(x) = 10x^2 \cdot \text{sign}(x)$ in model (15)

Classical SVMs, the localized version SVM-KNN, and classical kNN are applied to simulated data sets of size $n = 200$ for both scenarios each with 500 runs. In case of classical SVMs, the Gaussian RBF kernel $K_\gamma(x,x') = \exp(-\gamma(x-x')^2)$ and the ε-insensitive loss for $\varepsilon = 0.001$ are used. The hyperparameter γ is chosen by a five-fold cross validation among

$$0.001,\ 0.005,\ 0.01,\ 0.05,\ 0.1,\ 0.5,\ 1,\ 2,\ 5,\ 10,\ 15,\ 20,\ 30,\ \ 50,\ \ 75,\ 100,\ 150,\ 200,\ 250,\ 300,\ 350,\ 400,\ 500$$

and the regularization parameter is equal to $\lambda_n = a \cdot n^{-0.45}$ where a is chosen by a five-fold cross validation among

$$0.00001,\ 0.00005,\ 0.0001,\ 0.0005,\ 0.001,\ 0.005,\ 0.01,\ \ \ 0.05,\ 0.1,$$

The choice $\lambda_n = a \cdot n^{-0.45}$ is motivated by the fact that classical SVMs with the ε-insensitive loss are consistent if $\lim_{n\to\infty} \lambda_n = 0$ and $\lim_{n\to\infty} \lambda_n^2 n = \infty$; see (Christmann and Steinwart, 2007, Theorem 12). In case of SVM-KNN, the number of nearest neighbors is equal to $k_n = \lceil b \cdot n^{0.75} \rceil$ where the hyperparameter b is chosen by a five-fold cross validation among

```
0.15, 0.2 ,0.3 ,0.4 ,0.5 ,0.6 ,0.7 ,0.8 ,0.9 ,1 .
```

The exponent 0.75 for the definition of k_n is in accordance with the settings in Section 3. Choosing $k_n = \lceil b \cdot n^{0.75} \rceil$ would also guarantee universal consistency of classical kNN; see, for example, (Györfi et al., 2002, Theorem 6.1). For each testing point ξ, the prediction is calculated by a local SVM on the k_n nearest neighbor. For each local SVM, the polynomial kernel $K(x,x') = (x \cdot x' + 1)^3$ with degree 3 and the ε-insensitive loss for $\varepsilon = 0.001$ are used. In accordance with the settings in Section 3, the regularization parameter is equal to $\Lambda_{n,\xi} = C_{n,\xi} \max\left\{0.01 k_n^{-0.2}, \frac{1}{k_n} \sum_{i \in I_{n,\xi}} |x_i - \xi|^{1.5}\right\}$ where, for every ξ, the hyperparameter $C_{n,\xi}$ is chosen by a five-fold cross validation among

$$0.01,\ 0.1,\ 1,\ 10,\ 100,\ 1000,\ 10000,\ 100000\,.$$

Similarly to the case of SVM-KNN, the number of nearest neighbors in the classical kNN method is equal to $k_n = \lceil c \cdot n^{0.5} \rceil$ where the hyperparameter c is chosen by a five-fold cross validation among

$$0.1,\ 0.2\ ,0.3\ ,0.4\ ,0.5\ ,0.6\ ,0.7\ ,0.8\ ,0.9\ ,1\,.$$

The evaluation of the estimates is done on a test data set which consists of 1001 equidistant grid points ξ_i on $[-1,1]$. For every run $r \in \{1,\ldots,500\}$, the mean absolute error (MAE) is calculated

$$\mathrm{MAE}_{j,r}(f_\star) \;=\; \frac{1}{1001}\sum_{i=1}^{1001}\big|f_\star(x_i) - f_j(x_i)\big| \qquad \text{for } f_\star \in \big\{f_{j,r}^{\mathrm{SVM}}, f_{j,r}^{\mathrm{SVM\text{-}KNN}}, f_{j,r}^{\mathrm{kNN}}\big\}$$

where $f_{j,r}^{\mathrm{SVM}}$ denotes the SVM-estimate, $f_{j,r}^{\mathrm{SVM\text{-}KNN}}$ denotes the SVM-KNN-estimate, and $f_{j,r}^{\mathrm{kNN}}$ denotes the kNN-estimate in the r-th run of scenario j. For every scenario j and every learning algorithm, the values $\mathrm{MAE}_{j,r}(f_\star)$, $r \in \{1,\ldots,500\}$, are shown in a boxplot in Figure 3. In addition, Table 1 shows the average of $\mathrm{MAE}_{j,r}(f_\star)$ over the 500 runs:

$$\mathrm{MAE}_j(f_\star) \;=\; \frac{1}{500}\sum_{r=1}^{500}\mathrm{MAE}_{j,r}(f_\star) \qquad \text{for } f_\star \in \big\{f_{j,r}^{\mathrm{SVM}}, f_{j,r}^{\mathrm{SVM\text{-}KNN}}\big\}\,.$$

	scenario $j=1$	scenario $j=2$
SVM	0.453	0.115
SVM-KNN	0.331	0.216
kNN	0.348	0.189

Table 1: The average MAE_j of the mean absolute error over the 500 runs for classical SVMs and SVM-KNN for scenarios $j=1$ and $j=2$

It turns out that SVM-KNN is clearly better than classical SVM in scenario 1 while classical SVM is clearly better than SVM-KNN in scenario 2. In both examples, the performance of SVM-KNN is similar to that of classical kNN. Function f_2 in scenario 2 is a smooth function and classical SVMs are typically very successful for learning such smooth functions. Function f_1 in scenario 1 nearly coincides with f_2 in scenario 2 in the sense that the parts of the functions on $(-1,0)$ and $(0,1)$ are just interchanged. However, this leads to a considerable jump at $x=0$ which provides some difficulty for classical SVMs. Such jumps can be managed by classical SVMs if the hyperparameter γ and the regularization parameter λ are suitably chosen, namely, if γ is large and/or λ is small. However, such a choice increases the danger of overfitting in those parts of the input space in which the unknown regression function is a simple, smooth function. This problem is avoided by localized learners such as SVM-KNN, which is a main motivation for localizing global learning algorithms. In particular, the difference of the performance between scenario 1 and 2 is much smaller in case of SVM-KNN than in case of classical SVM. Figure 4 shows in a boxplot which values of γ are selected by the cross validation in the 500 runs for each scenario. Obviously, the jump in $x=0$ leads to large values of γ in scenario 1 compared to scenario 2. This in turn facilitates that the SVM-estimate is too volatile in those parts of the input space in which f_1 is relatively simple, for example, in the interval $[-1,-0.5]$. This tendency is exemplarily illustrated in Figure 5 which shows the estimates on the interval $[-1,0]$ of the input space in the first 9 runs of the simulation in case of scenario 1.

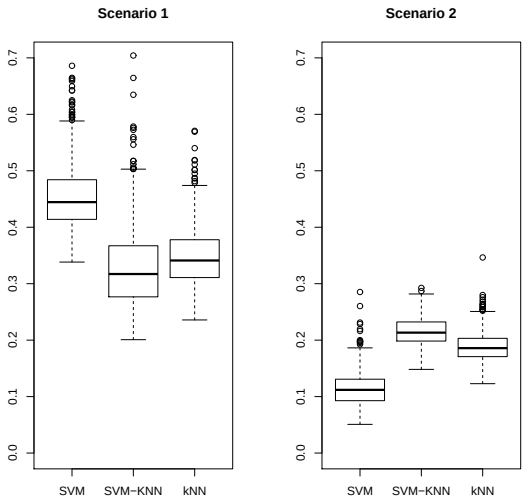

Figure 3: Boxplots of the mean absolute errors $\mathrm{MAE}_{j,r}$ in the runs $r \in \{1,\ldots,500\}$ for classical SVMs and SVM-KNN for scenarios $j = 1$ and $j = 2$

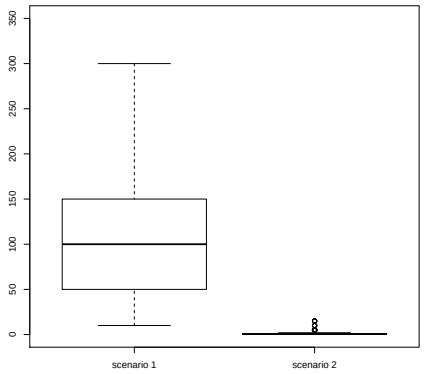

Figure 4: Values of the hyperparameter γ selected by cross validation for the classical SVM in the 500 runs for each scenario

5. Conclusions

Learning algorithms which are defined in a global manner typically can have difficulties if the complexity of the optimal predictor varies for different areas of the input space. One way to overcome this problem is to localize the learning algorithm. That is, the learning algorithm is not applied to the whole training data but only to those training data which are close to the testing point. In a num-

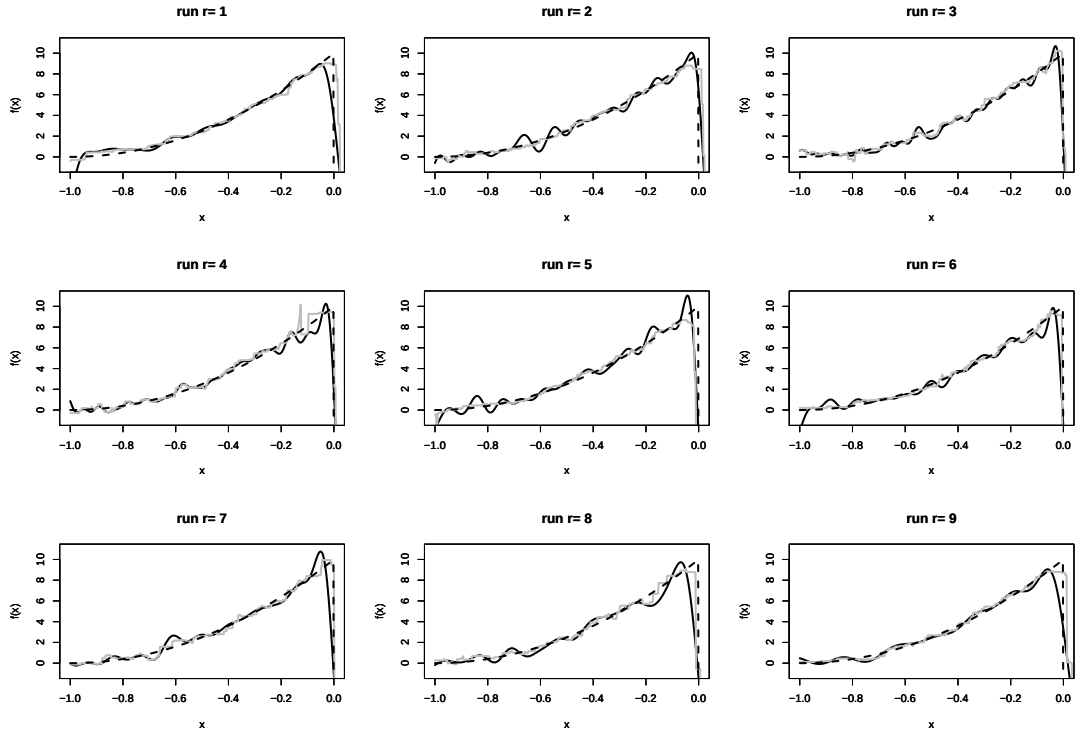

Figure 5: Estimates on the interval $[-1,0]$ in the first nine runs in scenario 1: true function f_1 (dashed black line), SVM (solid black line), SVM-KNN (solid gray line)

ber of recent articles such localizations of support vector machines have been suggested and their success has empirically been shown in extensive simulation studies and on real data sets but only little has been known on theoretical properties. In this article, it has been shown for a large class of regularized kernel methods (including SVM) that suitably localized versions (called SVM-KNN) are universally consistent.

Instead of localizing support vector machines, it would also be possible in principle to localize any other learning algorithm, for example, boosting. If this is done suitably, then localizing a learning algorithm will often lead to an algorithm which is again universally consistent. This article presents one way how this can be done in the special case of regularized kernel methods. However, it is a topic of further research if it is possible to derive a general scheme of localizing learning algorithms which, in combination with properties of the learning algorithm, always guarantees universal consistency.

Acknowledgments

I would like to thank two anonymous reviewers whose valuable comments have led to substantial improvements of the manuscript.

Appendix A. Preparations

Let $P_{\mathcal{X}}$ denote the distribution of the covariates X_i. For every $\zeta = (\xi, u) \in \mathcal{X} \times (0,1)$, there is a smallest $r_{n,\xi} \in [0,\infty]$ such that $Q\big(|X_i - \xi| \le r_{n,\xi}\big) \ge \frac{k_n}{n}$ and there is an $s_{n,\zeta} \in [0,\infty)$ such that

$$Q\Big(|X_i - \xi| < r_{n,\xi} \ \text{ or } \ \big(|X_i - \xi| = r_{n,\xi},\ |U_i - u| < s_{n,\zeta} \big) \Big) \ = \ \frac{k_n}{n} \, .$$

For every $\zeta = (\xi, u) \in \mathcal{X} \times (0,1)$, $r \in [0,\infty)$, and $s \in [0,\infty)$, define the open balls $B_r(\xi) = \{x \in \mathcal{X} \,|\, |x - \xi| < r\}$ and $B_s(u) = \{v \in (0,1) \,|\, |v - u| < s\}$, and define the boundary $\partial B_r(\xi) = \{x \in \mathcal{X} \,|\, |x - \xi| = r\}$. Define

$$B_{n,\zeta} \ = \ \Big(B_{r_{n,\xi}}(\xi) \times (0,1) \Big) \cup \Big(\partial B_{r_{n,\xi}}(\xi) \times B_{s_{n,\zeta}}(u) \Big)$$

Roughly spoken, $B_{n,\zeta}$ is a neighborhood around $\zeta = (\xi, u)$ with probability k_n/n which is in line with our tie-breaking strategy. Then,

$$P_{\mathcal{X}} \otimes \mathrm{Unif}_{(0,1)}\big(B_{n,\zeta}\big) \ = \ Q\big(Z_i \in B_{n,\zeta}\big) \ = \ \frac{k_n}{n}$$

where $\mathrm{Unif}_{(0,1)}$ denotes the uniform distribution on $(0,1)$. Let $P_{n,\zeta}$ be the conditional distribution of Z_i given $Z_i \in B_{n,\zeta}$, that is,

$$P_{n,\zeta}(B) = \frac{Q\big(Z_i \in B \cap B_{n,\zeta}\big)}{Q\big(Z_i \in B_{n,\zeta}\big)} = \frac{n}{k_n} Q\big(Z_i \in B \cap B_{n,\zeta}\big) \quad \forall B \in \mathfrak{B}_{\mathcal{X} \times (0,1)}.$$

Let $x \mapsto P(\cdot|x)$ be any regular version of the factorized conditional distribution of Y_i given $X_i = x$; see, for example, (Dudley, 2002, § 10.2). Due to independence of U_i, this coincides with the conditional distribution of Y_i given $Z_i = z$ (i.e., given $(X_i, U_i) = (x,u)$) and, accordingly, we write $P(\cdot|z) = P(\cdot|x)$. Let $Q_{Z,Y}$ denote the joint distribution of (Z_i, Y_i) and define $\mathcal{Z} := \mathcal{X} \times (0,1)$. Then, for every $\zeta \in \mathcal{Z}$, $n \in \mathbb{N}$, and every integrable $g : \mathcal{Z} \times \mathcal{Y} \to \mathbb{R}$,

$$\frac{n}{k_n} \int_{\mathcal{Z} \times \mathcal{Y}} I_{B_{n,\zeta}}(z) g(z,y) \, Q_{Z,Y}\big(d(z,y)\big) \ = \ \int_{\mathcal{Z}} \int_{\mathcal{Y}} g(z,y) \, P(dy|z) P_{n,\zeta}(dz) \, . \tag{16}$$

When this does not lead to confusion, the conditional distribution of the pair of random variables (Z_i, Y_i) given $Z_i \in B_{n,\zeta}$ is also denoted by $P_{n,\zeta}$. That is, we will also write

$$\frac{n}{k_n} \int_{\mathcal{Z} \times \mathcal{Y}} I_{B_{n,\zeta}}(z) g(z,y) \, Q_{Z,Y}\big(d(z,y)\big) \ = \ \int_{\mathcal{Z} \times \mathcal{Y}} g(z,y) \, P_{n,\zeta}\big(d(z,y)\big) \, . \tag{17}$$

The following lemma is an immediate consequence of the definitions and well known facts about the support of measures, see, for example, Parthasarathy (1967, II. Theorem 2.1). It says that, for almost every $\xi \in \mathcal{X}$, the radii $r_{n,\xi}$ decrease to 0.

Lemma 2 *Define*

$$B_0 \ := \ \big\{ \xi \in \mathcal{X} \ \big| \ \nexists r \in (0,\infty) \text{ such that } P_{\mathcal{X}}(B_r(\xi)) = 0 \big\} \, .$$

Then, $P_{\mathcal{X}}(B_0) = 1$.

Furthermore, for every $\xi \in B_0$,

$$\infty \ge r_{1,\xi} \ge r_{2,\xi} \ge r_{3,\xi} \ge \ldots \ge \lim_n r_{n,\xi} = 0 \, .$$

Similarly to the definition of $I_{n,\zeta}$ and $\mathbf{D}_{n,\zeta}$ in (4) and (5), we define the modifications $I^\star_{n,\zeta}$ and $\mathbf{D}^\star_{n,\zeta}$: For every $n \in \mathbb{N}$, $\zeta = (\xi, u) \in \mathcal{X} \times (0,1)$ and $\omega \in \Omega$, define

$$I^\star_{n,\zeta}(\omega) := \big\{ i \in \{1,\ldots,n\} \,\big|\, Z_i(\omega) \in B_{n,\zeta} \big\} .$$

Fix any $n \in \mathbb{N}$, $\zeta = (\xi, u) \in \mathcal{X} \times (0,1)$ and $\omega \in \Omega$ and let $i_1 < i_2 < \ldots < i_m$ be the (ordered) elements of $I^\star_{n,\zeta}(\omega)$. Then, define

$$\mathbf{D}^\star_{n,\zeta}(\omega) \;=\; \Big(\big(Z_{i_1}(\omega), Y_{i_1}(\omega)\big), \ldots, \big(Z_{i_m}(\omega), Y_{i_m}(\omega)\big) \Big) .$$

That is, $I^\star_{n,\zeta}$ consists of all those indexes $i \in \{1,\ldots,n\}$ and $\mathbf{D}^\star_{n,\zeta}$ consists of all those data points (Z_i, Y_i) such that $Z_i \in B_{n,\zeta}$. This means: while the the sets $I_{n,\zeta}$ and $\mathbf{D}_{n,\zeta}$ consist of a *fixed number* of nearest neighbors, the sets $I^\star_{n,\zeta}$ and $\mathbf{D}^\star_{n,\zeta}$ consist of all those neighbors which lie in a *fixed neighborhood*.

As the probability that $Z_i \in B_{n,\zeta}$ is k_n/n, we expect that, for large n, the index sets $I_{n,\zeta}$ and $I^\star_{n,\zeta}$ and the vectors of data points $\mathbf{D}_{n,\zeta}$ and $\mathbf{D}^\star_{n,\zeta}$ are similar. However, working with $I^\star_{n,\zeta}$ is more comfortable because, whether $i \in I^\star_{n,\zeta}$, only depends on Z_i but, whether $i \in I_{n,\zeta}$, depends on all $Z_1, \ldots, Z_n$.

If a real-valued function f is clipped at M, then the clipped version is denoted by $\widehat{f}$, that is, $\widehat{f}(x) = f(x)$ if $-M \le f(x) \le M$, and $\widehat{f}(x) = -M$ if $f(x) < -M$, and $\widehat{f}(x) = M$ if $M < f(x)$. Note that, for every $f_1, f_2 : \mathcal{X} \to \mathbb{R}$ and $\xi \in \mathcal{X}$, it follows that $\big|\widehat{f_1}(\xi) - \widehat{f_2}(\xi)\big| \le \big|f_1(\xi) - f_2(\xi)\big|$. Furthermore, since K is bounded, every $f \in H$ fulfills $|f(\xi)| \le \|K\|_\infty \cdot \|f\|_H$; see (Steinwart and Christmann, 2008, Lemma 4.23). In combination with (12), this implies that, for every $\xi \in \mathcal{X}$ and for every $f_1, f_2 \in H$,

$$\left| \int L\big(y, \widehat{f_1}(\xi)\big) P(dy|\xi) - \int L\big(y, \widehat{f_2}(\xi)\big) P(dy|\xi) \right| \;\le\; |L|_{M,1} \cdot \|K\|_\infty \cdot \|f_1 - f_2\|_H . \tag{18}$$

Define $\|L(\cdot,0)\|_\infty = \sup_{y \in [-M,M]} \big|L(y,0)\big|$. Then, for every probability measure P_0,

$$\mathcal{R}_{P_0}(0) \;=\; \int L(y,0)\, P_0\big(d(x,y)\big) \;\le\; \|L(\cdot,0)\|_\infty \overset{(13)}{<} \infty . \tag{19}$$

The following lemma is one of the main tools; it is an application of Hoeffding's inequality and will be used several times for $V = H$ and $V = \mathbb{R}$.

Lemma 3 *Let V be a separable Hilbert space and, for every $n \in \mathbb{N}$, let $\Psi_n : \mathcal{Z} \times \mathcal{Y} \to V$ be a Borel-measurable function such that for every bounded subset $B \subset \mathcal{Z}$,*

$$\sup_{n \in \mathbb{N}} \sup_{z \in B, y \in \mathcal{Y}} \big\|\Psi_n(z,y)\big\|_H \;<\; \infty .$$

Then, for every $\zeta \in \mathcal{X}$,

$$\lambda_n^{-\frac{3}{2}} \frac{n}{k_n} \left(\frac{1}{n} \sum_{i=1}^n \Psi_n(Z_i, Y_i) I_{B_{n,\zeta}}(Z_i) - \int \Psi_n(z,y) I_{B_{n,\zeta}}(z)\, Q_{Z,Y}\big(d(z,y)\big) \right) \xrightarrow[n\to\infty]{} 0$$

in probability.

Note that the integral in Lemma 3 is an integral over a Hilbert-space-valued function and, accordingly, is a Bochner integral; see, for example, (Denkowski et al., 2003, § 3.10) for such integrals.

Proof The proof is done by an application of Hoeffding's inequality for functions with values in a separable Hilbert space. According to Lemma 2, there is an $n_0 \in \mathbb{N}$ such that $B_{n_0,\zeta}$ is bounded and $B_{n,\zeta} \subset B_{n_0,\zeta}$ for every $n \geq n_0$. Hence, there is a constant $b \in (0,\infty)$ such that, for every $n \geq n_0$,

$$\sup_{(z,y)\in\mathcal{Z}\times\mathcal{Y}} \left\|\Psi_n(z,y)I_{B_{n,\zeta}}(z)\right\|_V \leq b\,.$$

For every $n \geq n_0$ and $\tau \in (0,\infty)$, define $a_{n,\tau} := 2b\cdot\left(\sqrt{\tau n^{-1}} + \sqrt{n^{-1}} + \tau n^{-1}\right)$ and

$$A_{n,\tau} = \left\{ \left\| \frac{1}{n}\sum_{i=1}^{n} \Psi_n(Z_i,Y_i)I_{B_{n,\zeta}}(Z_i) - \int \Psi_n I_{B_{n,\zeta}}\, dQ_{Z,Y} \right\|_V < a_{n,\tau} \right\}.$$

Then, by Hoeffding's inequality for separable Hilbert spaces (e.g., Steinwart and Christmann, 2008, Corollary 6.15),

$$Q\big(A_{n,\tau}\big) \;\geq\; 1 - e^{-\tau} \qquad \forall n \geq n_0\,, \quad \forall \tau \in (0,\infty)\,. \tag{20}$$

Define $\tau_n := \lambda_n^{\frac{3}{2}} k_n n^{-\frac{1}{2}}$ and $\varepsilon_n := \lambda_n^{-\frac{3}{2}} n k_n^{-1} a_{n,\tau_n}$ for every $n \geq n_0$. Then, for every $\omega \in A_{n,\tau}$,

$$\lambda_n^{-\frac{3}{2}} \frac{n}{k_n} \left\| \frac{1}{n}\sum_{i=1}^{n} \Psi_n(Z_i(\omega),Y_i(\omega))I_{B_{n,\zeta}}(Z_i(\omega)) - \int \Psi_n I_{B_{n,\zeta}}\, dQ_{Z,Y} \right\|_V < \varepsilon_n.$$

According to (9),

$$\varepsilon_n = \frac{n\cdot a_{n,\tau_n}}{\lambda_n^{\frac{3}{2}} k_n} = \frac{2bn}{\lambda_n^{\frac{3}{2}} k_n}\left(\sqrt{\frac{\lambda_n^{\frac{3}{2}} k_n}{\sqrt{n} n}} + \sqrt{\frac{1}{n}} + \frac{\lambda_n^{\frac{3}{2}} k_n}{\sqrt{n} n}\right) = 2b\cdot\left(\sqrt{\frac{\sqrt{n}}{\lambda_n^{\frac{3}{2}} k_n}} + \frac{\sqrt{n}}{\lambda_n^{\frac{3}{2}} k_n} + \frac{1}{\sqrt{n}}\right) \xrightarrow[n\to\infty]{} 0.$$

Hence, for every $\varepsilon > 0$, there is an $n_\varepsilon \in \mathbb{N}$ such that $\varepsilon > \varepsilon_n$ for every $n \geq n_\varepsilon$ and, therefore,

$$Q\left(\lambda_n^{-\frac{3}{2}} \frac{n}{k_n} \left\| \frac{1}{n}\sum_{i=1}^{n} \Psi_n(Z_i,Y_i)I_{B_{n,\zeta}}(Z_i) - \int \Psi_n I_{B_{n,\zeta}}\, dQ_{Z,Y} \right\|_V > \varepsilon\right) \leq Q\Big(\complement A_{n,\tau_n}\Big) \overset{(20)}{\leq} e^{-\tau_n}\,.$$

The last expression converges to 0 because $\lim_{n\to\infty}\tau_n = \infty$ due to (9), ■

Appendix B. Measurability

Measurability is an issue and needs some care because the SVM-KNN is based on a subsample which is randomly chosen. It is not possible to ignore measurability by turning over to outer probabilities here because the final step of the proof of the main theorem is based on an application of Fubini's Theorem and, therefore, heavily relies on (product) measurability.

Lemma 4

(a) The following maps are measurable with respect to the product-σ-algebra $\mathbb{B}^p \otimes \mathfrak{B}_{(0,1)} \otimes \mathcal{A}$ and the respective Borel-σ-Algebra:

(i) $\mathbb{R}^p \times (0,1) \times \Omega \to \mathbb{R}^{(p+1)k_n}, \quad (\xi,u,\omega) = (\zeta,\omega) \mapsto \mathbf{D}_{n,\zeta}(\omega)$

(ii) $\mathbb{R}^p \times (0,1) \times \Omega \to \mathbb{R}, \quad (\xi,u,\omega) = (\zeta,\omega) \mapsto R_{n,\zeta}(\omega) := \max_{i \in I_{n,\zeta}} \left|X_i(\omega) - \xi\right|.$

(iii) $\mathbb{R}^p \times (0,1) \times \Omega \to \mathbb{R}, \quad (\xi,u,\omega) = (\zeta,\omega) \mapsto \tilde{\Lambda}_{n,\zeta}(\omega).$

(b) Let $\Lambda : \mathbb{R}^p \times (0,1) \times \Omega \to (0,\infty)$ be measurable with respect to $\mathbb{B}^p \otimes \mathfrak{B}_{(0,1)} \otimes \mathcal{A}$ and the Borel-σ-Algebra. Then,

$$\mathbb{R}^p \times (0,1) \times \Omega \to \mathbb{R}, \quad (\xi,u,\omega) = (\zeta,\omega) \mapsto f_{\mathbf{D}_{n,\zeta}(\omega),\Lambda(\zeta,\omega)}(\xi)$$

is measurable with respect to $\mathbb{B}^p \otimes \mathfrak{B}_{(0,1)} \otimes \mathcal{A}$ and $\mathbb{B}$.

(c) For every $\zeta = (\xi,u) \in \mathbb{R}^p \times (0,1)$ and every $\Lambda : \Omega \to (0,\infty)$ measurable with respect to $\mathcal{A}$ and the Borel-σ-Algebra, the map

$$\Omega \to \mathbb{R}, \quad \omega \mapsto f_{\mathbf{D}^*_{n,\zeta}(\omega),\Lambda(\omega)}(\xi)$$

is measurable with respect to $\mathcal{A}$ and $\mathbb{B}$.

Proof For every $\zeta = (\xi,u) \in \mathbb{R}^p \times (0,1)$ and $\omega \in \Omega$, define $I_{n,\zeta}(\omega)$ as in Section 2. Let Ind_n denote the set of all subsets of $\{1,\ldots,n\}$ with k_n elements. First, it is shown that

$$\tilde{\tau}_n : \Omega \times \mathbb{R}^p \times (0,1) \to \mathrm{Ind}_n, \quad (\omega,\xi,u) \mapsto I_{n,(\xi,u)}(\omega)$$

is measurable with respect to $\mathcal{A} \otimes \mathbb{B}^p \otimes \mathfrak{B}_{(0,1)}$ and 2^{Ind_n}: Take any $I \in \mathrm{Ind}_n$ such that $I \neq \{1,\ldots,k_n\}$ and, for every $\mathcal{J} \subset \{1,\ldots,n\}$, define

$$\begin{aligned}
B_{\mathcal{J}}^{(1)} &:= \left\{(\omega,\xi,u) \in \Omega \times \mathbb{R}^p \times (0,1) \,\middle|\, \max_{i \in I} |X_i(\omega) - \xi| \leq \min_{\ell \notin I} |X_\ell(\omega) - \xi| \right\} \\
B_{\mathcal{J}}^{(2)} &:= \left\{(\omega,\xi,u) \in \Omega \times \mathbb{R}^p \times (0,1) \,\middle|\, |X_j(\omega) - \xi| = \max_{i \in I} |X_i(\omega) - \xi| \ \ \forall\, j \in \mathcal{J} \right\} \\
B_{\mathcal{J}}^{(3)} &:= \left\{(\omega,\xi,u) \in \Omega \times \mathbb{R}^p \times (0,1) \,\middle|\, |X_\ell(\omega) - \xi| \neq \max_{i \in I} |X_i(\omega) - \xi| \ \ \forall\, \ell \notin \mathcal{J} \right\} \\
B_{\mathcal{J}}^{(4)} &:= \left\{(\omega,\xi,u) \in \Omega \times \mathbb{R}^p \times (0,1) \,\middle|\, \max_{i \in \mathcal{J} \cap I} |U_i(\omega) - u| < \min_{j \in \mathcal{J} \setminus I} |U_j(\omega) - u| \right\}.
\end{aligned}$$

The set $B_{\mathcal{J}}^{(1)}$ says that no X_ℓ is closer to ξ than the k_n nearest neighbors. The sets $B_{\mathcal{J}}^{(2)}$ and $B_{\mathcal{J}}^{(3)}$ states that $\mathcal{J}$ specifies all those X_j which lie at the border of the neighborhood given by the nearest neighbors. The set $B_{\mathcal{J}}^{(4)}$ is concerned with all data points which lie at the border: the nearest neighbors among them have strictly smaller $|U_i - u|$ than those which do not belong to the nearest neighbors. Accordingly, the inverse image $\tilde{\tau}_n^{-1}(\{I\})$ equals

$$\tilde{\tau}_n^{-1}(\{I\}) = \bigcup_{\mathcal{J} \subset \{1,\ldots,n\}} \left(B_{\mathcal{J}}^{(1)} \cap B_{\mathcal{J}}^{(2)} \cap B_{\mathcal{J}}^{(3)} \cap B_{\mathcal{J}}^{(4)} \right).$$

Since $B_{\mathcal{J}}^{(t)}$ is measurable for every $t \in \{1,2,3,4\}$ and $\mathcal{J} \subset \{1,\dots,n\}$, this shows that $\tilde{\tau}_n^{-1}(\{I\})$ is measurable for every $I \neq \{1,\dots,k_n\}$. Hence, $\tilde{\tau}_n$ is measurable. For every $I = \{i_1,\dots,i_{k_n}\}$ such that $i_1 < i_2 < \cdots < i_{k_n}$ and every $D_n = \big((z_1,y_1),\dots,(z_n,y_n)\big) \in ((\mathbb{R}^p \times (0,1)) \times \mathbb{R})^n$ define

$$\varphi_n(I, D_n) \;=\; \big((z_{i_1},y_{i_1}),\dots,(z_{i_{k_n}},y_{i_{k_n}})\big)\,.$$

The map $\varphi_n : \mathrm{Ind}_n \times ((\mathbb{R}^p \times (0,1)) \times \mathbb{R})^n \to ((\mathbb{R}^p \times (0,1)) \times \mathbb{R})^{k_n}$ is continuous (where Ind_n is endowed with the discrete topology). Since

$$\mathbf{D}_{n,\zeta}(\omega) \;=\; \varphi_n\Big(\tilde{\tau}_n(\omega,\xi,u)\,,\mathbf{D}_n(\omega)\Big) \qquad \text{for } \zeta = (\xi,u)\,,$$

statement (i) follows from measurability of $\tilde{\tau}_n$ and φ_n. Next, (ii) follows from measurability of $(x_{i_1},\dots,x_{i_{k_n}},\xi) \mapsto \max_{j\in\{1,\dots,k_n\}} |x_{i_j} - \xi|$ and (iii) follows from

$$\tilde{\Lambda}_{n,\zeta} \;=\; \frac{1}{k_n}\sum_{i=1}^{n} |X_i - \xi|^{\frac{3}{2}} I_{[0,\infty)}(R_{n,\zeta} - X_i)\,.$$

Now, we can prove part (b) and (c): For every $I \subset \{1,2,\dots,n\}$ and every $D = \big((x_1,y_1),\dots,(x_n,y_n)\big) \in ((\mathbb{R}^p \times (0,1)) \times \mathbb{R})^n$, denote $D_I = \big((x_i,y_i)\big)_{i\in I}$. Then, it follows from Lemma 9 (a) and (Steinwart and Christmann, 2008, Lemma 4.23) that the map

$$2^{\{1,2,\dots,n\}} \times ((\mathbb{R}^p \times (0,1)) \times \mathbb{R})^n \times \mathcal{X} \;\to\; H\,, \qquad (I,D,\xi) \;\mapsto\; f_{D_I,\lambda}(\xi)$$

is continuous for every $\lambda > 0$ (where $2^{\{1,2,\dots,n\}}$ is endowed with the discrete topology). Since $\lambda \mapsto f_{D_I,\lambda}(\xi)$ is continuous for every fixed (I,D,ξ) according to (Steinwart and Christmann, 2008, Corollary 5.19 and Lemma 4.23), the map $\big((I,D,\xi),\lambda\big) \mapsto f_{D_I,\lambda}(\xi)$ is a Caratheodory function and, therefore, measurable; see, for example, Denkowski et al. (2003, Definition 2.5.18 and Theorem 2.5.22). Then, (b) follows from (a), and (c) follows from measurability of $\tilde{\tau}^*_{n,\zeta} : \omega \mapsto I^*_{n,\zeta}(\omega)$ for every fixed $\zeta = (\xi,u)$. Measurability of $\tilde{\tau}^*_{n,\zeta}$ follows from

$$\tilde{\tau}^*_{n,\zeta}{}^{-1}(I) \;=\; \bigcap_{i\in I} Z_i^{-1}(B_{n,\zeta}) \cap \bigcap_{i\notin I} Z_i^{-1}(\complement B_{n,\zeta}) \qquad \forall\, I \in 2^{\{1,2,\dots,n\}}\,.$$

■

Appendix C. Proof of Theorem 1

In the main part of the proof, it is shown that for $P_{\mathcal{X}} \otimes \mathrm{Unif}_{(0,1)}$ - almost every $\zeta = (\xi,u) \in \mathcal{X} \times (0,1)$,

$$0 \le \int L\big(y, \widehat{f}_{\mathbf{D}_{n,\zeta},\Lambda_{n,\zeta}}(\xi)\big)\,P(dy|\zeta) - \inf_{t\in\mathbb{R}} \int L(y,t)\,P(dy|\zeta) \;\xrightarrow[n\to\infty]{}\; 0 \tag{21}$$

in probability. Then, statement (21) implies Theorem 1 as follows:

Since, for every fixed $\zeta = (\xi,u)$, the maps

$$\omega \mapsto \int L\big(y, \widehat{f}_{\mathbf{D}_{n,\zeta}(\omega),\Lambda_{n,\zeta}(\omega)}(\xi)\big)\,P(dy|\zeta) - \inf_{t\in\mathbb{R}} \int L(y,t)\,P(dy|\zeta)\,, \quad n \in \mathbb{N},$$

are uniformly bounded, convergence in probability for $P_X \otimes \mathrm{Unif}_{(0,1)}$ - almost every $\zeta = (\xi, u) \in X \times (0,1)$ implies

$$\mathbb{E}_Q \left(\int L\big(y, \widehat{f}_{\mathbf{D}_{n,\zeta},\Lambda_{n,\zeta}}(\xi)\big)\, P(dy|\zeta) - \inf_{t\in\mathbb{R}} \int L(y,t)\, P(dy|\zeta) \right) \xrightarrow[n\to\infty]{} 0$$

for $P_X \otimes \mathrm{Unif}_{(0,1)}$ - almost every $\zeta = (\xi, u) \in X \times (0,1)$. Since the maps

$$\zeta = (\xi, u) \;\mapsto\; \mathbb{E}_Q \left(\int L\big(y, \widehat{f}_{\mathbf{D}_{n,\zeta},\Lambda_{n,\zeta}}(\xi)\big)\, P(dy|\zeta) - \inf_{t\in\mathbb{R}} \int L(y,t)\, P(dy|\zeta) \right), \quad n \in \mathbb{N},$$

are uniformly bounded again, $P_X \otimes \mathrm{Unif}_{(0,1)}$ - almost sure convergence implies

$$\iint \mathbb{E}_Q \left(\int L\big(y, \widehat{f}_{\mathbf{D}_{n,\zeta},\Lambda_{n,\zeta}}(\xi)\big) P(dy|\zeta) - \inf_{t\in\mathbb{R}} \int L(y,t) P(dy|\zeta) \right) P_X(d\xi)\, \mathrm{Unif}_{(0,1)}(du) \longrightarrow 0 \tag{22}$$

for $n \to \infty$. Note that $\zeta \mapsto \inf_{t\in\mathbb{R}} \int L(y,t)\, P(dy|\zeta)$ is measurable, because the assumptions on L imply continuity of $t \mapsto \int L(y,t)\, P(dy|\zeta)$, hence, $\inf_{t\in\mathbb{R}} \int L(y,t)\, P(dy|\zeta) = \inf_{t\in\mathbb{Q}} \int L(y,t)\, P(dy|\zeta)$ for every $\zeta \in X \times (0,1)$. Next, recall that $f_{\mathbf{D}_n}(\zeta) = \widehat{f}_{\mathbf{D}_{n,\zeta},\Lambda_{n,\zeta}}(\xi)$ and $P(\cdot|\xi) = P(\cdot|\zeta)$ for every $\zeta = (\xi, u)$. By a slight abuse of notation, we write

$$\mathcal{R}_P(f_{\mathbf{D}_n}) \;=\; \mathcal{R}_{P\otimes \mathrm{Unif}_{(0,1)}}(f_{\mathbf{D}_n}) \;=\; \iint L\big(y, f_{\mathbf{D}_n}(\xi,u)\big)\, P\big(d(\xi,y)\big)\, \mathrm{Unif}_{(0,1)}(du)\,.$$

Then, applying Fubini's Theorem in (22) yields

$$0 \;\le\; \mathbb{E}_Q \left(\mathcal{R}_P\big(f_{\mathbf{D}_n}\big) - \int \inf_{t\in\mathbb{R}} \int L(y,t)\, P(dy|\xi) P_X(d\xi) \right) \xrightarrow[n\to\infty]{} 0\,. \tag{23}$$

For every measurable $f : X \to \mathbb{R}$,

$$\int L\big(y, f(\xi)\big)\, P(dy|\xi) \;\ge\; \inf_{t\in\mathbb{R}} \int L(y,t)\, P(dy|\xi) \qquad \forall \xi \in X\,.$$

Hence,

$$\mathcal{R}_P^* \;\ge\; \int \inf_{t\in\mathbb{R}} \int L(y,t)\, P(dy|\xi) P_X(d\xi)$$

and, therefore, (23) implies

$$\mathbb{E}_Q \left(\mathcal{R}_P\big(f_{\mathbf{D}_n}\big) - \mathcal{R}_P^* \right) \xrightarrow[n\to\infty]{} 0$$

and, as $\mathcal{R}_P\big(f_{\mathbf{D}_n}\big) \ge \mathcal{R}_P^*$,

$$\mathbb{E}_Q \left| \mathcal{R}_P\big(f_{\mathbf{D}_n}\big) - \mathcal{R}_P^* \right| \xrightarrow[n\to\infty]{} 0\,.$$

In particular, this also implies

$$\mathcal{R}_P\big(f_{\mathbf{D}_n}\big) \xrightarrow[n\to\infty]{} \mathcal{R}_P^* \qquad \text{in probability.}$$

That is, it only remains to prove (21). To this end, note that, for every $\zeta = (\xi, u) \in \mathcal{X} \times (0,1)$, we have $P(\cdot|\zeta) = P(\cdot|\xi)$ and

$$
\begin{aligned}
0 \;\leq\; & \int L\big(y, \widehat{f}_{\mathbf{D}_{n,\zeta},\Lambda_{n,\zeta}}(\xi)\big)\, P(dy|\zeta) \;-\; \inf_{t\in\mathbb{R}} \int L(y,t)\, P(dy|\zeta) \;\leq \\
\leq\; & \left| \int L\big(y, \widehat{f}_{\mathbf{D}_{n,\zeta},\Lambda_{n,\zeta}}(\xi)\big)\, P(dy|\xi) - \int L\big(y, \widehat{f}_{\mathbf{D}^{\star}_{n,\zeta},\Lambda_{n,\zeta}}(\xi)\big)\, P(dy|\xi) \right| && (24) \\
& + \left| \int L\big(y, \widehat{f}_{\mathbf{D}^{\star}_{n,\zeta},\Lambda_{n,\zeta}}(\xi)\big)\, P(dy|\xi) - \int L\big(y, \widehat{f}_{P_{n,\zeta},\Lambda_{n,\zeta}}(\xi)\big)\, P(dy|\xi) \right| && (25) \\
& + \left| \int L\big(y, \widehat{f}_{P_{n,\zeta},\Lambda_{n,\zeta}}(\xi)\big) P(dy|\xi) - \iint L\big(y, \widehat{f}_{P_{n,\zeta},\Lambda_{n,\zeta}}(x)\big) P(dy|x) P_{n,\zeta}(d(x,v)) \right| && (26) \\
& + \left(\iint L\big(y, \widehat{f}_{P_{n,\zeta},\Lambda_{n,\zeta}}(x)\big) P(dy|x) P_{n,\zeta}(d(x,v)) - \inf_{t\in\mathbb{R}} \int L(y,t) P(dy|\xi) \right) \vee 0 && (27)
\end{aligned}
$$

where $a \vee 0 = \max\{a, 0\}$. Therefore, it suffices to prove convergence in probability of each of these four summands. This is done in the following four subsections but, first, we need some more preparations:

Lemma 5 *Fix any $\zeta = (\xi, u) \in B_0 \times (0,1)$ where B_0 is defined as in Lemma 2. Let $\mathbb{P}_{\mathbf{D}_{n,\zeta}}$ and $\mathbb{P}_{\mathbf{D}^{\star}_{n,\zeta}}$ denote the empirical measure corresponding to $\mathbf{D}_{n,\zeta}$ and $\mathbf{D}^{\star}_{n,\zeta}$ respectively. It follows that*

$$
\lambda_n^{-\frac{3}{2}} \frac{\big|\sharp(I^{\star}_{n,\zeta}) - k_n\big|}{k_n} \;\xrightarrow[n\to\infty]{}\; 0 \qquad \textit{in probability}, \tag{28}
$$

$$
\lambda_n^{-\frac{3}{2}} \frac{\big|\sharp(I^{\star}_{n,\zeta}) - k_n\big|}{\sharp(I^{\star}_{n,\zeta})} \;\xrightarrow[n\to\infty]{}\; 0 \qquad \textit{in probability}, \tag{29}
$$

$$
\lambda_n^{-\frac{3}{2}} \Big\| \mathbb{P}_{\mathbf{D}_{n,\zeta}} - \mathbb{P}_{\mathbf{D}^{\star}_{n,\zeta}} \Big\|_{\mathrm{TV}} \;\xrightarrow[n\to\infty]{}\; 0 \qquad \textit{in probability}, \tag{30}
$$

$$
R_{n,\zeta} \;:=\; \max_{i\in I_{n,\zeta}} |X_i - \xi| \;\xrightarrow[n\to\infty]{}\; 0 \qquad \textit{in probability}, \tag{31}
$$

and, for every $\beta \in (0,\infty)$,

$$
\lambda_n^{-\frac{3}{2}} \left| \frac{1}{k_n} \sum_{i\in I_{n,\zeta}} |X_i - \xi|^{\beta} - \int |x-\xi|^{\beta} P_{n,\zeta}(d(x,v)) \right| \;\xrightarrow[n\to\infty]{}\; 0 \qquad \textit{in probability}. \tag{32}
$$

Proof Statement (28) follows from Lemma 3 because the definitions imply

$$
\lambda_n^{-\frac{3}{2}} \frac{\big|\sharp(I^{\star}_{n,\zeta}) - k_n\big|}{k_n} \;=\; \lambda_n^{-\frac{3}{2}} \frac{n}{k_n} \left| \frac{1}{n} \sum_{i=1}^{n} I_{B_{n,\zeta}}(Z_i) - \int I_{B_{n,\zeta}}(z)\, Q_{Z,Y}\big(d(z,y)\big) \right|.
$$

In order to prove (29) note that (28) implies that $\sharp(I^\star_{n,\zeta})/k_n \to 1$ in probability and, therefore, also $k_n/\sharp(I^\star_{n,\zeta}) \to 1$ in probability. Hence, (28) implies (29) because

$$\lambda_n^{-\frac{3}{2}} \frac{\left|\sharp(I^\star_{n,\zeta}) - k_n\right|}{\sharp(I^\star_{n,\zeta})} \;=\; \lambda_n^{-\frac{3}{2}} \frac{\left|\sharp(I^\star_{n,\zeta}) - k_n\right|}{k_n} \cdot \frac{k_n}{\sharp(I^\star_{n,\zeta})} \,.$$

In order to prove (30) note that the definitions imply for almost every $\omega \in \Omega$

$$I_{n,\zeta}(\omega) \;\subset\; I^\star_{n,\zeta}(\omega) \qquad \text{or} \qquad I^\star_{n,\zeta}(\omega) \;\subset\; I_{n,\zeta}(\omega)\,. \tag{33}$$

(Only in case of distance ties in the $U_i(\omega)$, statement (33) is not true.) Therefore,

$$\sharp\big(I_{n,\zeta} \setminus I^\star_{n,\zeta}\big) \;\le\; \left|\sharp(I^\star_{n,\zeta}) - k_n\right| \qquad \text{and} \qquad \sharp\big(I^\star_{n,\zeta} \setminus I_{n,\zeta}\big) \;\le\; \left|\sharp(I^\star_{n,\zeta}) - k_n\right|. \tag{34}$$

almost surely. Then, almost surely,

$$\begin{aligned}
&\sup_{C \in \mathfrak{B}_{\mathcal{Z}\times\mathcal{Y}}} \left|\mathbb{P}_{\mathbf{D}_{n,\zeta}}(C) - \mathbb{P}_{\mathbf{D}^\star_{n,\zeta}}(C)\right| \;= \\
&\quad= \sup_C \left| \frac{1}{k_n}\Big(\sum_{i \in I_{n,\zeta} \cap I^\star_{n,\zeta}} I_C(Z_i,Y_i) + \sum_{i \in I_{n,\zeta} \setminus I^\star_{n,\zeta}} I_C(Z_i,Y_i)\Big) - \right. \\
&\qquad\qquad \left. - \frac{1}{\sharp\big(I^\star_{n,\zeta}\big)}\Big(\sum_{i \in I^\star_{n,\zeta} \cap I_{n,\zeta}} I_C(Z_i,Y_i) + \sum_{i \in I^\star_{n,\zeta} \setminus I_{n,\zeta}} I_C(Z_i,Y_i)\Big) \right| \;\le \\
&\le \sup_C \left| \frac{1}{k_n} - \frac{1}{\sharp\big(I^\star_{n,\zeta}\big)} \right| \sum_{i \in I_{n,\zeta} \cap I^\star_{n,\zeta}} I_C(Z_i,Y_i) \;+ \\
&\qquad + \frac{1}{k_n} \sup_C \sum_{i \in I_{n,\zeta} \setminus I^\star_{n,\zeta}} I_C(Z_i,Y_i) \;+\; \frac{1}{\sharp\big(I^\star_{n,\zeta}\big)} \sup_C \sum_{i \in I^\star_{n,\zeta} \setminus I_{n,\zeta}} I_C(Z_i,Y_i) \;\le \\
&\overset{(34)}{\le} \left| \frac{1}{k_n} - \frac{1}{\sharp\big(I^\star_{n,\zeta}\big)} \right| k_n + \frac{\left|\sharp(I^\star_{n,\zeta}) - k_n\right|}{k_n} + \frac{\left|\sharp(I^\star_{n,\zeta}) - k_n\right|}{\sharp(I^\star_{n,\zeta})}\,.
\end{aligned}$$

Therefore, (30) follows from (28) and (29).

In order to prove (31), fix any $\varepsilon > 0$. As $\xi \in B_0$, we have $P_X\big(B_\varepsilon(\xi)\big) > 0$ and, therefore, $P_X\big(B_\varepsilon(\xi)\big) - k_n/n > \frac{1}{2} P_X\big(B_\varepsilon(\xi)\big) > 0$ for n large enough (see Lemma 2). Then, (31) follows from

$$\begin{aligned}
Q\Big(R_{n,\zeta} > \varepsilon\Big) &= Q\Big(\sharp\big\{i \in \{1,\ldots,n\} \,\big|\, X_i \in B_\varepsilon(\xi)\big\} < k_n\Big) \;=\; Q\left(\frac{1}{n}\sum_{i=1}^n I_{B_\varepsilon(\xi)}(X_i) < \frac{k_n}{n}\right) \\
&= Q\left(P_X\big(B_\varepsilon(\xi)\big) - \frac{1}{n}\sum_{i=1}^n I_{B_\varepsilon(\xi)}(X_i) > P_X\big(B_\varepsilon(\xi)\big) - \frac{k_n}{n}\right) \;\le \\
&\le Q\left(P_X\big(B_\varepsilon(\xi)\big) - \frac{1}{n}\sum_{i=1}^n I_{B_\varepsilon(\xi)}(X_i) > \frac{1}{2} P_X\big(B_\varepsilon(\xi)\big)\right)
\end{aligned}$$

and the law of large numbers.

Now, statement (32) will be proven. An application of Lemma 3 for $\Psi_n\big((x,v),y\big) = |x-\xi|^\beta$ and (16) yield that it suffices to prove

$$\lambda_n^{-\frac{3}{2}} \left| \frac{1}{k_n} \sum_{i \in I_{n,\zeta}} |X_i - \xi|^\beta - \frac{1}{k_n}\sum_{i=1}^n |X_i - \xi|^\beta I_{B_{n,\zeta}}(Z_i) \right| \;\xrightarrow[n\to\infty]{}\; 0 \tag{35}$$

in probability in order to prove statement (32).

According to Lemma 2, there is an $n_0 \in \mathbb{N}$ such that $r_{n,\xi} \leq 1$ for every $n \geq n_0$. Then, for every $\varepsilon > 0$ and every $n \geq n_0$,

$$\begin{aligned} Q\left(\lambda_n^{-\frac{3}{2}}\left|\frac{1}{k_n}\sum_{i\in I_{n,\zeta}}|X_i-\xi|^\beta - \frac{1}{k_n}\sum_{i=1}^n |X_i-\xi|^\beta I_{B_{n,\zeta}}(Z_i)\right| > \varepsilon\right) \leq \\ \leq Q\left(\lambda_n^{-\frac{3}{2}}\left\|\mathbb{P}_{\mathbf{D}_{n,\zeta}} - \tfrac{\sharp(I^\star_{n,\zeta})}{k_n}\mathbb{P}_{\mathbf{D}^\star_{n,\zeta}}\right\|_{\mathrm{TV}} > \varepsilon,\ R_{n,\zeta}\leq 1\right) + Q\Big(R_{n,\zeta}>1\Big) \\ \leq Q\left(\lambda_n^{-\frac{3}{2}}\left\|\mathbb{P}_{\mathbf{D}_{n,\zeta}} - \mathbb{P}_{\mathbf{D}^\star_{n,\zeta}}\right\|_{\mathrm{TV}} > \frac{\varepsilon}{2}\right) + Q\left(\lambda_n^{-\frac{3}{2}}\frac{\left|\sharp(I^\star_{n,\zeta})-k_n\right|}{k_n} > \frac{\varepsilon}{2}\right) + Q\Big(R_{n,\zeta}>1\Big) \end{aligned}$$

so that (35) follows from (30), (28), and (31). ■

Lemma 6 *For every $P_{\mathcal{X}}$-integrable $h : \mathcal{X} \to \mathbb{R}$, there is a set $B_h \in \mathfrak{B}_{\mathcal{X}}$ such that $P_{\mathcal{X}}(B_h) = 1$ and*

$$\lim_{n\to\infty}\int \big|h(x)-h(\xi)\big|\,P_{n,\zeta}\big(d(x,v)\big) = 0 \qquad \forall\,\zeta=(\xi,u)\in B_h\times(0,1). \tag{36}$$

Proof Define

$$\gamma_{n,\xi} := \frac{1}{P_{\mathcal{X}}\big(B_{r_{n,\xi}}(\xi)\big)}\int_{B_{r_{n,\xi}}(\xi)}\big|h(x)-h(\xi)\big|\,P_{\mathcal{X}}(dx)$$

and, analogously, define $\bar{\gamma}_{n,\xi}$ where the open ball $B_{r_{n,\xi}}(\xi)$ is replaced by the closed ball $\overline{B}_{r_{n,\xi}}(\xi)$ around ξ with radius $r_{n,\xi}$. According to Besicovitch's Density Theorem, there is a set $B_h \in \mathfrak{B}_{\mathcal{X}}$ such that $P_{\mathcal{X}}(B_h) = 1$ and, for every $\xi \in B_h$, $\lim_{n\to\infty}\gamma_{n,\xi} = \lim_{n\to\infty}\bar{\gamma}_{n,\xi} = 0$; for $\gamma_{n,\xi}$, see, for example, (Fremlin, 2006, Theorem 472D(b)); for $\bar{\gamma}_{n,\xi}$, this follows from (Krantz and Parks, 2008, Theorem 4.3.5(2)) (exactly in the same way as Fremlin, 2006, Theorem 472D(b) follows from Fremlin, 2006, Theorem 472D(a)). Recall from Appendix A that $B_{n,\zeta} = \big(B_{r_{n,\xi}}(\xi)\times(0,1)\big)\cup\big(\partial B_{r_{n,\xi}}(\xi)\times B_{s_{n,\zeta}}(u)\big)$ and define $\alpha_{n,\zeta} := Q\big(U_i \in B_{s_{n,\zeta}}(u)\big)$, $\beta_{n,\xi} := P_{\mathcal{X}}\big(B_{r_{n,\xi}}(\xi)\big)$ and $\overline{\beta}_{n,\xi} := P_{\mathcal{X}}\big(\overline{B}_{r_{n,\xi}}(\xi)\big)$. Then,

$$\tfrac{k_n}{n} = Q\big(Z_i \in B_{n,\zeta}\big) = \beta_{n,\xi} + \alpha_{n,\zeta}\big(\overline{\beta}_{n,\xi}-\beta_{n,\xi}\big) \tag{37}$$

and

$$\begin{aligned} &\int \big|h(x)-h(\xi)\big|\,P_{n,\zeta}\big(d(x,v)\big) = \\ &= \frac{n}{k_n}\left(\int_{B_{r_{n,\xi}}(\xi)}\big|h(x)-h(\xi)\big|\,P_{\mathcal{X}}(dx) + \alpha_{n,\zeta}\int_{\partial B_{r_{n,\xi}}(\xi)}\big|h(x)-h(\xi)\big|\,P_{\mathcal{X}}(dx)\right) \\ &= \frac{n}{k_n}\Big(\beta_{n,\xi}\gamma_{n,\xi} + \alpha_{n,\zeta}\big(\overline{\beta}_{n,\xi}\bar{\gamma}_{n,\xi} - \beta_{n,\xi}\gamma_{n,\xi}\big)\Big) = \\ &= \frac{n}{k_n}\Big(\beta_{n,\xi}+\alpha_{n,\zeta}\big(\overline{\beta}_{n,\xi}-\beta_{n,\xi}\big)\Big)\bar{\gamma}_{n,\xi} + \frac{n}{k_n}(1-\alpha_{n,\zeta})\beta_{n,\xi}(\gamma_{n,\xi}-\bar{\gamma}_{n,\xi}) \leq \\ &\overset{(37)}{\leq} \bar{\gamma}_{n,\xi} + 1\cdot\big|\bar{\gamma}_{n,\xi}-\gamma_{n,\xi}\big| \xrightarrow[n\to\infty]{} 0 \end{aligned}$$

■

C.1 Convergence of the First Summand (24)

Fix any $\zeta = (\xi, u) \in B_0 \times (0,1)$ where B_0 is defined as in Lemma 2. Again let $\mathbb{P}_{\mathbf{D}_{n,\zeta}}$ and $\mathbb{P}_{\mathbf{D}^\star_{n,\zeta}}$ denote the empirical measure corresponding to $\mathbf{D}_{n,\zeta}$ and $\mathbf{D}^\star_{n,\zeta}$ respectively. It follows from (18), (19), and (51) that

$$\begin{aligned}
&\left| \int L\big(y, \widehat{f}_{\mathbf{D}_{n,\zeta},\Lambda_{n,\zeta}}(\xi)\big)\, P(dy|\xi) - \int L\big(y, \widehat{f}_{\mathbf{D}^\star_{n,\zeta},\Lambda_{n,\zeta}}(\xi)\big)\, P(dy|\xi) \right| \;\leq \\
&\qquad \leq \;\; |L|_{M,1} \|K\|_\infty^2 \Big(b_0 \Lambda_{n,\zeta}^{-1} + b_1 \|K\|_\infty^q \mathcal{R}_{P_1}(0)^{\frac{q}{2}} \Lambda_{n,\zeta}^{-\frac{q}{2}-1} \Big) \cdot \big\| \mathbb{P}_{\mathbf{D}_{n,\zeta}} - \mathbb{P}_{\mathbf{D}^\star_{n,\zeta}} \big\|_{\mathrm{TV}} \;\leq \\
&\qquad \overset{(10)}{\leq} \;\; |L|_{M,1} \|K\|_\infty^2 \Big(b_0 (c\lambda_n)^{-1} + b_1 \|K\|_\infty^q \big\|L(\cdot,0)\big\|_\infty^{\frac{q}{2}} (c\lambda_n)^{-\frac{q}{2}-1} \Big) \cdot \big\| \mathbb{P}_{\mathbf{D}_{n,\zeta}} - \mathbb{P}_{\mathbf{D}^\star_{n,\zeta}} \big\|_{\mathrm{TV}} .
\end{aligned}$$

Therefore, convergence in probability follows from (30) in Lemma 5 and $q \in [0,1]$.

C.2 Convergence of the Second Summand (25)

Fix any $\zeta = (\xi, u) \in B_0 \times (0,1)$.

Lemma 7 *For every $n \in \mathbb{N}$, define*

$$\tilde{\lambda}_{n,\zeta} \;=\; \int |x-\xi|^{\frac{3}{2}}\, P_{n,\zeta}\big(d(x,v)\big) \qquad \textit{and} \qquad \lambda_{n,\zeta} \;:=\; c \cdot \max\left\{\lambda_n, \tilde{\lambda}_{n,\zeta}\right\} .$$

Then,

$$\frac{|\Lambda_{n,\zeta} - \lambda_{n,\zeta}|}{\Lambda_{n,\zeta}\sqrt{\lambda_{n,\zeta}}} \;\xrightarrow[n\to\infty]{}\; 0 \qquad \textit{in probability} .$$

Proof For $a_1, a_2, b \in \mathbb{R}$, denote $a_1 \vee a_2 = \max\{a_1, a_2\}$ and note that $|a_1 \vee b - a_2 \vee b| \leq |a_1 - a_2|$. For every n, the definitions and (10) imply

$$\frac{|\Lambda_{n,\zeta} - \lambda_{n,\zeta}|}{\Lambda_{n,\zeta}\sqrt{\lambda_{n,\zeta}}} \leq \frac{\big|\Lambda_{n,\zeta} - c\cdot\big(\lambda_n \vee \tilde{\Lambda}_{n,\zeta}\big)\big| + c\cdot\big|\lambda_n \vee \tilde{\Lambda}_{n,\zeta} - \lambda_n \vee \tilde{\lambda}_{n,\zeta}\big|}{c^{\frac{3}{2}}\cdot\big(\lambda_n \vee \tilde{\Lambda}_{n,\zeta}\big)\sqrt{\lambda_n}} \leq \frac{c_n}{c^{\frac{3}{2}}\sqrt{\lambda_n}} + \frac{\big|\tilde{\Lambda}_{n,\zeta} - \tilde{\lambda}_{n,\zeta}\big|}{\sqrt{c}\lambda_n\sqrt{\lambda_n}} .$$

Hence, the statement follows from the assumption that $\lim_{n\to\infty} c_n/\sqrt{\lambda_n} = 0$ and from (32) in Lemma 5. ∎

According to (18), it suffices to show

$$\big\| f_{\mathbf{D}^\star_{n,\zeta},\Lambda_{n,\zeta}} - f_{P_{n,\zeta},\Lambda_{n,\zeta}} \big\|_H \;\xrightarrow[n\to\infty]{}\; 0 \qquad \text{in probability}$$

in order to prove convergence to 0 of the the second summand (25). To this end, note that

$$\begin{aligned}
&\big\| f_{\mathbf{D}^\star_{n,\zeta},\Lambda_{n,\zeta}} - f_{P_{n,\zeta},\Lambda_{n,\zeta}} \big\|_H \;\leq \\
&\qquad \leq \; \big\| f_{\mathbf{D}^\star_{n,\zeta},\Lambda_{n,\zeta}} - f_{\mathbf{D}^\star_{n,\zeta},\lambda_{n,\zeta}} \big\|_H + \big\| f_{\mathbf{D}^\star_{n,\zeta},\lambda_{n,\zeta}} - f_{P_{n,\zeta},\lambda_{n,\zeta}} \big\|_H + \big\| f_{P_{n,\zeta},\lambda_{n,\zeta}} - f_{P_{n,\zeta},\Lambda_{n,\zeta}} \big\|_H
\end{aligned}$$

and that $\big\| f_{\mathbf{D}^\star_{n,\zeta},\Lambda_{n,\zeta}} - f_{\mathbf{D}^\star_{n,\zeta},\lambda_{n,\zeta}} \big\|_H$ and $\big\| f_{P_{n,\zeta},\lambda_{n,\zeta}} - f_{P_{n,\zeta},\Lambda_{n,\zeta}} \big\|_H$ converge in probability to 0 according to part (i) of Lemma 9 (b), (19), and Lemma 7. Note that boundedness of the kernel K means that

$\sup_{x\in\mathcal{X}}\|\Phi(x)\|_H = \|K\|_\infty$. By defining $f(x,v)=f(x)$ for every $z=(x,v)\in\mathcal{X}\times(0,1)=\mathcal{Z}$ and $f\in H$, the RKHS H consisting of functions $f:\mathcal{X}\to\mathbb{R}$ can also be identified with an RKHS (again denoted by H) which consists of functions $f:\mathcal{Z}\to\mathbb{R}$; the kernel of this RKHS is given by $K(z,z')=K(x,x')$ for every $z=(x,v),z'=(x',u')\in\mathcal{X}\times(0,1)=\mathcal{Z}$; see, for example, the proof of (Christmann and Hable, 2012, Theorem 2). Fix $a=b_0+b_1\|K\|_\infty^q\big\|L(\cdot,0)\big\|_\infty^{q/2}c^{-q/2}$ and $n_0\in\mathbb{N}$ such that $\lambda_n\leq 1$ for every $n\geq n_0$. According to the definition of $\lambda_{n,\zeta}$, we have $\lambda_{n,\zeta}^{-q/2}\leq c^{-q/2}\lambda_n^{-q/2}\leq c^{-q/2}\lambda_n^{-1/2}$ for every $n\geq n_0$. According to part (ii) of Lemma 9 (b) and (19), for every $n\geq n_0$, there is a measurable function $h_{n,\zeta}:\mathcal{Z}\times\mathcal{Y}\to\mathbb{R}$ such that $\|h_{n,\zeta}\|_\infty\leq a\lambda_n^{-\frac{1}{2}}$ and

$$
\begin{aligned}
&\big\|f_{\mathbf{D}^\star_{n,\zeta},\lambda_{n,\zeta}}-f_{P_{n,\zeta},\lambda_{n,\zeta}}\big\|_H \;\leq\\
&\quad\leq\; \lambda_{n,\zeta}^{-1}\left\|\frac{1}{\sharp(I^\star_{n,\zeta})}\sum_{i\in I^\star_{n,\zeta}}h_{n,\zeta}(Z_i,Y_i)\Phi(X_i)-\int h_{n,\zeta}\Phi\,dP_{n,\zeta}\right\|_H \;\leq\\
&\quad\leq\; c^{-1}\lambda_n^{-1}\left|\frac{1}{\sharp(I^\star_{n,\zeta})}-\frac{1}{k_n}\right|\sum_{i\in I^\star_{n,\zeta}}\Big\|h_{n,\zeta}(Z_i,Y_i)\Phi(X_i)\Big\|_H+\\
&\qquad+c^{-1}\lambda_n^{-1}\left\|\frac{1}{k_n}\sum_{i\in I^\star_{n,\zeta}}h_{n,\zeta}(Z_i,Y_i)\Phi(X_i)-\int h_{n,\zeta}\Phi\,dP_{n,\zeta}\right\|_H \;\leq\\
&\quad\overset{(17)}{\leq}\; \lambda_n^{-\frac{3}{2}}\frac{\big|\sharp(I^\star_{n,\zeta})-k_n\big|}{k_n}\cdot ac^{-1}\|K\|_\infty+\\
&\qquad+\lambda_n^{-1}\frac{n}{k_n}\left\|\frac{1}{n}\sum_{i=1}^n h_{n,\zeta}(Z_i,Y_i)\Phi(X_i)I_{B_{n,\zeta}}(Z_i)-\int h_{n,\zeta}\Phi I_{B_{n,\zeta}}\,dQ_{Z,Y}\right\|_H\cdot c^{-1}.
\end{aligned}
$$

It follows from (28) that

$$
\lambda_n^{-\frac{3}{2}}\frac{\big|\sharp(I^\star_{n,\zeta})-k_n\big|}{k_n}\;\xrightarrow[n\to\infty]{}\;0\qquad\text{in probability}
$$

and it follows from Lemma 3 for $\Psi_n(z,y)=\lambda_n^{\frac{1}{2}}h_{n,\zeta}(z,y)\Phi(x)$, $z=(x,v)$, that

$$
\lambda_n^{-1}\frac{n}{k_n}\left\|\frac{1}{n}\sum_{i=1}^n h_{n,\zeta}(Z_i,Y_i)\Phi(X_i)I_{B_{n,\zeta}}(Z_i)-\int h_{n,\zeta}\Phi I_{B_{n,\zeta}}\,dQ_{Z,Y}\right\|_H\xrightarrow[n\to\infty]{}0\quad\text{in probability.}
$$

C.3 Convergence of the Third Summand (26)

For every $m\in\mathbb{N}$, define

$$
\alpha_m(y)\;=\;\sum_{j=-mM}^{mM}\frac{j}{m}I_{(\frac{j-1}{m},\frac{j}{m}]}(y)\,,\qquad y\in\mathbb{R}\,. \tag{38}
$$

That is, $\alpha_m(\mathcal{Y})\subset\big\{\frac{j}{m}\,\big|\,j\in\{-mM,\ldots,mM\}\big\}$ and

$$
\big|\alpha_m(y)-y\big|<\tfrac{1}{m}\quad\forall\,y\in\mathcal{Y}\,. \tag{39}
$$

According to Lemma 6, there is a set $B_1 \in \mathfrak{B}_X$ such that $P_X(B_1) = 1$ and such that, for all maps

$$h : \mathcal{X} \to \mathbb{R}, \quad x \mapsto P\Big(\big(\tfrac{j-1}{m}, \tfrac{j}{m}\big] \,\Big|\, x\Big), \quad j \in \{-mM, \ldots, mM\}, \ m \in \mathbb{N},$$

(36) is fulfilled with $B_h = B_1$. Fix any $\zeta = (\xi, u) \in \mathcal{X} \times (0,1)$ such that $\xi \in B_0 \cap B_1$. It follows from (13) and (39) that, for every $m \in \mathbb{N}$,

$$\sup_{\substack{t \in [-M,M] \\ x \in \mathcal{X}}} \left| \int L(y,t)\, P(dy|x) - \int L(\alpha_m(y),t)\, P(dy|x) \right| \le \ell(\tfrac{1}{m}).$$

Since $\lim_{m\to\infty} \ell(\frac{1}{m}) = 0$, it is enough to show that, for every $m \in \mathbb{N}$,

$$\left| \int L\big(\alpha_m(y), \widehat{f}_{P_{n,\zeta},\Lambda_{n,\zeta}}(\xi)\big) P(dy|\xi) - \iint L\big(\alpha_m(y), \widehat{f}_{P_{n,\zeta},\Lambda_{n,\zeta}}(x)\big) P(dy|x) P_{n,\zeta}\big(d(x,v)\big) \right|$$

converges to 0 in probability for $n \to \infty$. Next, it follows from

$$\int L\big(\alpha_m(y),t\big)\, P(dy|x) \overset{(38)}{=} \sum_{j=-mM}^{mM} L(\tfrac{j}{m},t) \cdot P\Big(\big(\tfrac{j-1}{m}, \tfrac{j}{m}\big] \,\Big|\, x\Big) \qquad \forall t \in \mathbb{R} \quad \forall x \in \mathcal{X}$$

that it suffices to show that, for every $j \in \{-mM, \ldots, mM\}$ and $m \in \mathbb{N}$,

$$\left| L\big(\tfrac{j}{m}, \widehat{f}_{P_{n,\zeta},\Lambda_{n,\zeta}}(\xi)\big) P\big(\big(\tfrac{j-1}{m}, \tfrac{j}{m}\big] \,\big|\, \xi\big) - \int L\big(\tfrac{j}{m}, \widehat{f}_{P_{n,\zeta},\Lambda_{n,\zeta}}(x)\big) P\big(\big(\tfrac{j-1}{m}, \tfrac{j}{m}\big] \,\big|\, x\big) P_{n,\zeta}\big(d(x,v)\big) \right|$$

converges to 0 in probability for $n \to \infty$. The latter statement is shown in the following:

$$\begin{aligned}
&\left| L\big(\tfrac{j}{m}, \widehat{f}_{P_{n,\zeta},\Lambda_{n,\zeta}}(\xi)\big) P\big(\big(\tfrac{j-1}{m}, \tfrac{j}{m}\big] \,\big|\, \xi\big) - \int L\big(\tfrac{j}{m}, \widehat{f}_{P_{n,\zeta},\Lambda_{n,\zeta}}(x)\big) P\big(\big(\tfrac{j-1}{m}, \tfrac{j}{m}\big] \,\big|\, x\big) P_{n,\zeta}\big(d(x,v)\big) \right| \\
&\le \left| \int L\big(\tfrac{j}{m}, \widehat{f}_{P_{n,\zeta},\Lambda_{n,\zeta}}(\xi)\big) \Big(P\big(\big(\tfrac{j-1}{m}, \tfrac{j}{m}\big] \,\big|\, \xi\big) - P\big(\big(\tfrac{j-1}{m}, \tfrac{j}{m}\big] \,\big|\, x\big) \Big) P_{n,\zeta}\big(d(x,v)\big) \right| \\
&\quad + \left| \int \Big(L\big(\tfrac{j}{m}, \widehat{f}_{P_{n,\zeta},\Lambda_{n,\zeta}}(\xi)\big) - L\big(\tfrac{j}{m}, \widehat{f}_{P_{n,\zeta},\Lambda_{n,\zeta}}(x)\big) \Big) P\big(\big(\tfrac{j-1}{m}, \tfrac{j}{m}\big] \,\big|\, x\big)\, P_{n,\zeta}\big(d(x,v)\big) \right| \\
&\le \sup_{t,y \in [-M,M]} L(y,t) \cdot \int \left| P\big(\big(\tfrac{j-1}{m}, \tfrac{j}{m}\big] \,\big|\, \xi\big) - P\big(\big(\tfrac{j-1}{m}, \tfrac{j}{m}\big] \,\big|\, x\big) \right| P_{n,\zeta}\big(d(x,v)\big) \qquad (40) \\
&\quad + |L|_{M,1} \int \big| \widehat{f}_{P_{n,\zeta},\Lambda_{n,\zeta}}(\xi) - \widehat{f}_{P_{n,\zeta},\Lambda_{n,\zeta}}(x) \big|\, P_{n,\zeta}\big(d(x,v)\big) \qquad (41)
\end{aligned}$$

As $r_{n,\xi} \searrow 0$ (Lemma 2), it follows from the above definition of B_1 and $\xi \in B_1$ that the summand in (40) converges to 0 (in $\mathbb{R}$) for $n \to \infty$. In order to prove convergence (in probability) of the summand in (41), note that, according to the mean value theorem in several variables and Steinwart

and Christmann (2008, Corollary 4.36 and Equation (5.4)),

$$\begin{aligned}
&\int \left|\widehat{f}_{P_{n,\zeta},\Lambda_{n,\zeta}}(\xi) - \widehat{f}_{P_{n,\zeta},\Lambda_{n,\zeta}}(x)\right| P_{n,\zeta}\big(d(x,v)\big) \;\leq \\
&\leq \int \left|f_{P_{n,\zeta},\Lambda_{n,\zeta}}(\xi) - f_{P_{n,\zeta},\Lambda_{n,\zeta}}(x)\right| P_{n,\zeta}\big(d(x,v)\big) \;\leq \\
&\leq \int \sup_{x' \in \overline{B}_{r_{n,\xi}}(\xi)} \left|\frac{\partial}{\partial x} f_{P_{n,\zeta},\Lambda_{n,\zeta}}(x')\right| \cdot |x-\xi| \, P_{n,\zeta}\big(d(x,v)\big) \;\leq \\
&\leq \left\|f_{P_{n,\zeta},\Lambda_{n,\zeta}}\right\|_H \cdot \left(\sup_{x \in \overline{B}_{r_{n,\xi}}(\xi)} \sqrt{\partial^{1,1}K(x,x)}\right) \cdot \int |x-\xi| \, P_{n,\zeta}\big(d(x,v)\big) \;\leq \\
&\leq \left(\sup_{x \in \overline{B}_{r_{n,\xi}}(\xi)} \sqrt{\partial^{1,1}K(x,x)}\right) \cdot \sqrt{\mathcal{R}_{P_{n,\zeta}}(0)} \cdot \frac{\int |x-\xi| \, P_{n,\zeta}\big(d(x,v)\big)}{\sqrt{\Lambda_{n,\zeta}}}
\end{aligned}$$

where $\overline{B}_{r_{n,\xi}}(\xi)$ denotes the closed ball around ξ with radius $r_{n,\xi}$. As

$$\mathcal{R}_{P_{n,\zeta}}(0) \leq \sup_{y \in [-M,M]} L(y,0) < \infty \qquad \text{and} \qquad \lim_{n\to\infty} \sup_{x \in \overline{B}_{r_{n,\xi}}(\xi)} \sqrt{\partial^{1,1}K(x,x)} = \sqrt{\partial^{1,1}K(\xi,\xi)}\,,$$

it remains to show that

$$\frac{\int |x-\xi| \, P_{n,\zeta}\big(d(x,v)\big)}{\sqrt{\Lambda_{n,\zeta}}} \;\xrightarrow[n\to\infty]{}\; 0 \qquad \text{in probability .} \tag{42}$$

In order to prove (42), note that

$$\begin{aligned}
&\frac{\int |x-\xi| \, P_{n,\zeta}\big(d(x,v)\big)}{\sqrt{\Lambda_{n,\zeta}}} \;\leq \\
&\leq \frac{\frac{1}{k_n}\sum_{i \in I_{n,\zeta}} |X_i - \xi|}{\sqrt{\Lambda_{n,\zeta}}} + \frac{\left|\frac{1}{k_n}\sum_{i \in I_{n,\zeta}} |X_i - \xi| - \int |x-\xi| \, P_{n,\zeta}\big(d(x,v)\big)\right|}{\sqrt{\Lambda_{n,\zeta}}} \;\leq \\
&\overset{(10)}{\leq} \frac{\frac{1}{k_n}\sum_{i \in I_{n,\zeta}} |X_i - \xi|}{\sqrt{c\frac{1}{k_n}\sum_{i \in I_{n,\zeta}} |X_i - \xi|^{\frac{3}{2}}}} + \qquad (43) \\
&\quad + c^{-\frac{1}{2}} \lambda_n^{-\frac{1}{2}} \left|\frac{1}{k_n} \sum_{i \in I_{n,\zeta}} |X_i - \xi| - \int |x-\xi| \, P_{n,\zeta}\big(d(x,v)\big)\right| . \qquad (44)
\end{aligned}$$

The summand in (44) converges to 0 in probability according to (32). The summand in (43) converges to 0 in probability because, by convexity of $z \mapsto z^{\frac{3}{2}}$ and $\sharp(I_{n,\zeta}) = k_n$, we get

$$\begin{aligned}
\frac{\frac{1}{k_n}\sum_{i \in I_{n,\zeta}} |X_i - \xi|}{\sqrt{c\frac{1}{k_n}\sum_{i \in I_{n,\zeta}} |X_i - \xi|^{\frac{3}{2}}}} \;&\leq\; \frac{\frac{1}{k_n}\sum_{i \in I_{n,\zeta}} |X_i - \xi|}{\sqrt{c\left(\frac{1}{k_n}\sum_{i \in I_{n,\zeta}} |X_i - \xi|\right)^{\frac{3}{2}}}} \;= \\
&= \; c^{-\frac{1}{2}} \Big(\frac{1}{k_n} \sum_{i \in I_{n,\zeta}} |X_i - \xi|\Big)^{\frac{1}{4}} \;\leq\; c^{-\frac{1}{2}} R_{n,\zeta}^{\frac{1}{4}} \;\xrightarrow[n\to\infty]{}\; 0 \quad \text{in probability}
\end{aligned}$$

according to (31).

C.4 Convergence of the Fourth Summand (27)

Let $\mathcal{M}_1(\mathcal{X}\times\mathcal{Y})$ be the set of all probability measures on $\mathcal{X}\times\mathcal{Y}$. For every $f\in H$, define the map $A_f:\ \mathcal{M}_1(\mathcal{X}\times\mathcal{Y})\times[0,\infty)\to\mathbb{R}$ by

$$A_f(P_0,\lambda)\ =\ \int L\big(y,f(x)\big)\,P_0\big(d(x,y)\big)\ +\ \lambda\|f\|_H^2 \tag{45}$$

for every $P_0\in\mathcal{M}_1(\mathcal{X}\times\mathcal{Y})$ and $\lambda\in[0,\infty)$. For every $f\in H$, the map $(x,y)\ \mapsto\ L\big(y,f(x)\big)$ is continuous and bounded on $\mathcal{X}\times\mathcal{Y}$ and, therefore, A_f is continuous with respect to weak convergence of probability measures (and the ordinary topology on $\mathbb{R}$ and $[0,\infty)$). Hence,

$$(P_0,\lambda)\ \mapsto\ \inf_{f\in H}A_f(P_0,\lambda)\qquad\text{is upper semi-continuous} \tag{46}$$

see, for example, (Denkowski et al., 2003, Prop. 1.1.36).

Let $\mathcal{C}_c(\mathcal{X}\times\mathcal{Y})$ be the set of all continuous functions $g:\mathcal{X}\times\mathcal{Y}\to\mathbb{R}$ with compact support. According to Denkowski et al. (2003, Theorem 2.6.24), there is a countable dense subset $\mathcal{S}\subset\mathcal{C}_c(\mathcal{X}\times\mathcal{Y})$ (with respect to uniform convergence).

According to Lemma 6, there is a set $B_2\in\mathfrak{B}_\mathcal{X}$ such that $P_\mathcal{X}(B_2)=1$ and such that, for all maps

$$h\ :\ \mathcal{X}\ \to\ \mathbb{R}\,,\qquad x\ \mapsto\ \int g(x,y)\,P(dy|x)\,,\qquad g\in\mathcal{S},$$

(36) is fulfilled with $B_h=B_2$. Fix any $\zeta=(\xi,u)\in\mathcal{X}\times(0,1)$ such that $\xi\in B_0\cap B_1\cap B_2$.

Lemma 8 *Let $(\Lambda_{n_j,\zeta})_{j\in\mathbb{N}}$ be a subsequence of $(\Lambda_{n,\zeta})_{n\in\mathbb{N}}$ which converges to zero Q-a.s. for $j\to\infty$. Then, Q - a.s.,*

$$\left(\int\!\!\int L\big(y,\widehat{f}_{P_{n_j,\zeta},\Lambda_{n_j,\zeta}}(x)\big)P(dy|x)P_{n_j,\zeta}\big(d(x,v)\big)-\inf_{t\in\mathbb{R}}\int L(y,t)\,P(dy|\xi)\right)\vee 0\ \ \xrightarrow[j\to\infty]{}\ 0\,.$$

Proof For every $n\in\mathbb{N}$, let $\tilde{P}_{n,\zeta}$ denote the conditional distribution of (X,Y) given $Z\in B_{n,\zeta}$. Then, for every integrable $g:\mathcal{X}\times\mathcal{Y}\to\mathbb{R}$,

$$\int g(x,y)\tilde{P}_{n,\zeta}\big(d(x,y)\big)=\int g(x,y)P_{n,\zeta}\big(d(x,v,y)\big)=\int_\mathcal{Z}\!\int_\mathcal{Y} g(x,y)P(dy|x)P_{n,\zeta}\big(d(x,v)\big)$$

and, according to the definitions (45) and (6),

$$\inf_{f\in H}A_f(\tilde{P}_{n,\zeta},\lambda)\ =\ \int L\big(y,f_{P_{n,\zeta},\lambda}(x)\big)P_{n,\zeta}\big(d(x,v,y)\big)\ +\ \lambda\big\|f_{P_{n,\zeta},\lambda}\big\|_H^2 \tag{47}$$

for every $\lambda\in(0,\infty)$ and $n\in\mathbb{N}$. Analogously to the definition of $\tilde{P}_{n,\zeta}\in\mathcal{M}(\mathcal{X}\times\mathcal{Y})$, define $\tilde{P}_{0,\zeta}\in\mathcal{M}(\mathcal{X}\times\mathcal{Y})$ by

$$\int g(x,y)\,\tilde{P}_{0,\zeta}\big(d(x,y)\big)\ =\ \int_\mathcal{Y} g(\xi,y)\,P(dy|\xi)\qquad\text{for every integrable } g:\mathcal{X}\times\mathcal{Y}\to\mathbb{R}\,.$$

First, it is shown that

$$\tilde{P}_{n,\zeta}\ \xrightarrow[n\to\infty]{}\ \tilde{P}_{0,\zeta}\qquad\text{weakly in }\mathcal{M}(\mathcal{X}\times\mathcal{Y})\,. \tag{48}$$

According to Bauer (2001, Theorem 30.8), we have to show that

$$\int g\,d\tilde{P}_{n,\zeta} \xrightarrow[n\to\infty]{} \int g\,d\tilde{P}_{0,\zeta} \qquad \forall\, g \in \mathcal{C}_c(\mathcal{X}\times\mathcal{Y})\,. \tag{49}$$

Fix any $g \in \mathcal{C}_c(\mathcal{X}\times\mathcal{Y})$. Then, for every $\varepsilon > 0$, there is a $g_\varepsilon \in \mathcal{S}$ such that $\sup_{x,y} |g(x,y) - g_\varepsilon(x,y)| < \varepsilon$ and, therefore,

$$\begin{aligned}\left|\int g\,d\tilde{P}_{n,\zeta} - \int g\,d\tilde{P}_{0,\zeta}\right| \;&\le\; \int |g-g_\varepsilon|\,d\tilde{P}_{n,\zeta} + \left|\int g_\varepsilon\,d\tilde{P}_{n,\zeta} - \int g_\varepsilon\,d\tilde{P}_{0,\zeta}\right| + \int |g-g_\varepsilon|\,d\tilde{P}_{0,\zeta} \;\le \\ &\le\; 2\varepsilon + \int_{\mathcal{Z}} \left| \int_{\mathcal{Y}} g_\varepsilon(x,y)\,P(dy|x) - \int_{\mathcal{Y}} g_\varepsilon(\xi,y)\,P(dy|\xi)\right| P_{n,\zeta}\big(d(x,v)\big)\end{aligned}$$

The second summand converges to 0 for $n\to\infty$ because of $\xi \in B_2$, $g_\varepsilon \in \mathcal{S}$, and the definition of B_2. As $\varepsilon > 0$ can be arbitrarily close to 0, this shows (49) and, therefore, (48).
Next, fix any $\omega \in \Omega$ such that $\gamma_j := \Lambda_{n_j,\xi}(\omega) \longrightarrow 0$ for $j\to\infty$. Then,

$$\begin{aligned}&\limsup_{j\to\infty} \iint L\big(y, \widehat{f}_{P_{n_j,\zeta},\Lambda_{n_j,\zeta}(\omega)}(x)\big) P(dy|x) P_{n_j,\zeta}\big(d(x,v)\big) \;\le \\ &\quad\le\quad \limsup_{j\to\infty} \iint L\big(y, f_{P_{n_j,\zeta},\gamma_j}(x)\big) P(dy|x) P_{n_j,\zeta}\big(d(x,v)\big) + \gamma_j \big\| f_{P_{n_j,\zeta},\gamma_j}\big\|_H^2 \;= \\ &\overset{(47)}{=}\quad \limsup_{j\to\infty} \inf_{f\in H} A_f(\tilde{P}_{n_j,\zeta},\gamma_j) \overset{(46,48)}{\le} \inf_{f\in H} A_f(\tilde{P}_{0,\zeta},0) \;= \\ &\quad=\quad \inf_{f\in H} \int L\big(y, f(\xi)\big)\,P(dy|\xi) \;=\; \inf_{t\in\mathbb{R}} \int L(y,t)\,P(dy|\xi)\,.\end{aligned}$$

■

By use of the above lemma, we can complete the proof of part 4 now. The definition of $\Lambda_{n,\zeta}$ and (31) imply that $\Lambda_{n,\xi} \to 0$ in probability for $n\to\infty$. Then, via the characterization of convergence in probability by use of subsequences and almost sure convergence, it follows from Lemma 8 that

$$\left(\iint L\big(y, \widehat{f}_{P_{n,\zeta},\Lambda_{n,\zeta}}(x)\big) P(dy|x) P_{n,\zeta}\big(d(x,v)\big) - \inf_{t\in\mathbb{R}} \int L(y,t)\,P(dy|\xi)\right) \vee 0 \xrightarrow[n\to\infty]{} 0$$

in probability.

Appendix D. Stability Properties of Support Vector Machines

Part (a) of the following Lemma 9 shows: in order to ensure that empirical SVMs are continuous in the data, continuity of the loss function L is enough. This result strengthens (Steinwart and Christmann, 2008, Lemma 5.13) which assumes differentiability and also (Hable and Christmann, 2011, Corollary 3.5) which assumes Lipschitz-(equi-)continuity. Next, part (i) of Lemma 9 (b) considerably strengthens (Steinwart and Christmann, 2008, Corollary 5.19) in the sense that it quantifies the continuity of the map $\lambda \mapsto f_{P_0,\lambda}$. Finally, parts (ii) and (iii) of Lemma 9 (b) are just simple applications of the stability results in (Steinwart and Christmann, 2008, § 5.3).

Lemma 9 *Let $\mathcal{X}_0$ be a separable metric space and let $\mathcal{Y}_0 \subset \mathbb{R}$ be closed. Let $K : \mathcal{X}_0\times\mathcal{X}_0 \to \mathbb{R}$ be a continuous and bounded kernel with RKHS H and canonical feature map Φ. Let $L : \mathcal{Y}_0\times\mathbb{R} \to [0,\infty)$ be a convex loss function.*

(a) If $L : \mathcal{Y}_0 \times \mathbb{R} \to [0,\infty)$ is continuous, then the map

$$(\mathcal{X}_0 \times \mathcal{Y}_0)^n \;\to\; H\,, \qquad D \;\mapsto\; f_{D,\lambda}$$

is continuous for every $\lambda > 0$ and $n \in \mathbb{N}$.

(b) Assume that L has the local Lipschitz-property that, for every $a \in (0,\infty)$, there is an $|L|_{a,1} \in (0,\infty)$ such that

$$\sup_{y\in\mathcal{Y}_0} \bigl|L(y,t_1) - L(y,t_2)\bigr| \;\le\; |L|_{a,1} \cdot |t_1 - t_2| \qquad \forall\, t_1,t_2 \in [-a,a]\,.$$

(i) Then, for every probability measure P_0 on $(\mathcal{X}_0 \times \mathcal{Y}_0, \mathfrak{B}_{\mathcal{X}_0\times\mathcal{Y}_0})$ such that $\mathcal{R}_{P_0}(0) < \infty$ and for every $\lambda_0,\lambda_1 \in (0,\infty)$, it holds that

$$\bigl\|f_{P_0,\lambda_1} - f_{P_0,\lambda_0}\bigr\|_H \;\le\; \frac{|\lambda_1 - \lambda_0|}{\lambda_1\sqrt{\lambda_0}} 2\sqrt{\mathcal{R}_{P_0}(0)}\,.$$

(ii) If there are some $b_0,b_1 \in (0,\infty)$ and $q \in [0,\infty)$ such that, for every $a \in (0,\infty)$, $|L|_{a,1} = b_0 + b_1 a^q$, then: for every probability measures P_1 on $(\mathcal{X}_0 \times \mathcal{Y}_0, \mathfrak{B}_{\mathcal{X}_0\times\mathcal{Y}_0})$ such that $\mathcal{R}_{P_1}(0) < \infty$ and for every $\lambda \in (0,\infty)$, there is a measurable $h_{P_1,\lambda} : \mathcal{X}_0 \times \mathcal{Y}_0 \to \mathbb{R}$ such that

$$\bigl|h_{P_1,\lambda}(x,y)\bigr| \;\le\; b_0 + b_1\|K\|_\infty^q \left(\frac{\mathcal{R}_{P_1}(0)}{\lambda}\right)^{\frac{q}{2}} \tag{50}$$

and such that, for every P_2 on $(\mathcal{X}_0 \times \mathcal{Y}_0, \mathfrak{B}_{\mathcal{X}_0\times\mathcal{Y}_0})$ with $\mathcal{R}_{P_2}(0) < \infty$,

$$\begin{aligned}\bigl\|f_{P_1,\lambda} - f_{P_2,\lambda}\bigr\|_H \;&\le\; \lambda^{-1}\left\|\int h_{P_1,\lambda}\Phi\,dP_1 - \int h_{P_1,\lambda}\Phi\,dP_1\right\|_H = \\ &=\; \lambda^{-1}\sup_{\substack{f\in H\\ \|f\|_H\le 1}} \bigl|\mathbb{E}_{P_1}h_{P_1,\lambda}f - \mathbb{E}_{P_2}h_{P_1,\lambda}f\bigr|\,.\end{aligned}$$

(iii) If there are some $b_0,b_1 \in (0,\infty)$ and $q \in [0,\infty)$ such that, for every $a \in (0,\infty)$, $|L|_{a,1} = b_0 + b_1 a^q$, then: for every probability measures P_1 and P_2 on $(\mathcal{X}_0 \times \mathcal{Y}_0, \mathfrak{B}_{\mathcal{X}_0\times\mathcal{Y}_0})$ such that $\mathcal{R}_{P_1}(0) < \infty$ and $\mathcal{R}_{P_2}(0) < \infty$ and for every $\lambda \in (0,\infty)$,

$$\bigl\|f_{P_1,\lambda} - f_{P_2,\lambda}\bigr\|_H \;\le\; \|K\|_\infty\Bigl(b_0\lambda^{-1} + b_1\|K\|_\infty^q \mathcal{R}_{P_1}(0)^{\frac{q}{2}}\lambda^{-\frac{q}{2}-1}\Bigr)\bigl\|P_1 - P_2\bigr\|_{\mathrm{TV}}\,. \tag{51}$$

Proof In order to prove (a) by contradiction, assume that $D \mapsto f_{D,\lambda}$ is not continuous. Then, there is an $\varepsilon > 0$ and a sequence, such that

$$D^{(m)} \;\xrightarrow[m\to\infty]{}\; D^{(0)} \qquad \text{and} \qquad \bigl\|f_{D^{(m)},\lambda} - f_{D^{(0)},\lambda}\bigr\|_H \;\ge\; \varepsilon \quad \forall\, m \in \mathbb{N}\,. \tag{52}$$

Define $\mathcal{R}_D(f) = \frac{1}{n}\sum_{i=1}^n L(y_i, f(x_i))$ for every $D = \bigl((x_1,y_1),\ldots,(x_n,y_n)\bigr) \in (\mathcal{X}_0 \times \mathcal{Y}_0)^n$ and $f \in H$. According to (Steinwart and Christmann, 2008, (5.4)) and due to continuity of L,

$$\sup_{m\in\mathbb{N}} \bigl\|f_{D^{(m)},\lambda}\bigr\|_H \;\le\; \sup_{m\in\mathbb{N}} \sqrt{\lambda^{-1}\mathcal{R}_{D^{(m)}}(0)} \;<\; \infty\,. \tag{53}$$

Hence, there is a subsequence such that $f_{D^{(m_\ell)},\lambda}$ weakly converges to some $f_0 \in H$ in the Hilbert space H for $\ell \to \infty$; see, for example, (Dunford and Schwartz, 1958, Corollary IV.4.7). That is, there is also a sequence which fulfills (52) and such that, in addition, $f_{D^{(m)},\lambda}$ weakly converges to f_0 in H for some $f_0 \in H$. This implies

$$\lim_{m\to\infty} f_{D^{(m)},\lambda}(x) \;=\; \lim_{m\to\infty} \left\langle f_{D^{(m)},\lambda}, \Phi(x)\right\rangle_H \;=\; \left\langle f_0, \Phi(x)\right\rangle_H \;=\; f_0(x) \qquad \forall\, x \in \mathcal{X}_0$$

and, for $x^{(m)} \to x^{(0)}$ in $\mathcal{X}_0$,

$$\begin{aligned}
&\lim_{m\to\infty} \left| f_{D^{(m)},\lambda}(x^{(m)}) - f_0(x^{(0)})\right| \;\leq \\
&\quad\leq \lim_{m\to\infty} \left|\left\langle f_{D^{(m)},\lambda}, \Phi(x^{(m)}) - \Phi(x^{(0)})\right\rangle_H\right| + \lim_{m\to\infty} \left| f_{D^{(m)},\lambda}(x^{(0)}) - f_0(x^{(0)})\right| \\
&\quad\leq \lim_{m\to\infty} \left\| f_{D^{(m)},\lambda}\right\|_H \cdot \left\|\Phi(x^{(m)}) - \Phi(x^{(0)})\right\|_H \;=\; 0
\end{aligned}$$

where the last equality follows from (53) and continuity of the kernel K. Hence, it follows that

$$\lim_{m\to\infty} \mathcal{R}_{\boldsymbol{D}^{(m)}}\left(f_{D^{(m)},\lambda}\right) \;=\; \mathcal{R}_{\boldsymbol{D}^{(0)}}\left(f_0\right). \tag{54}$$

Therefore, lower semi-continuity of the H-norm with respect to weak convergence in H (e.g., Conway, 1985, Exercise V.1.9) implies

$$\liminf_{m\to\infty} \mathcal{R}_{\boldsymbol{D}^{(m)}}\left(f_{D^{(m)},\lambda}\right) + \lambda\left\|f_{D^{(m)},\lambda}\right\|_H^2 \;\geq\; \mathcal{R}_{\boldsymbol{D}^{(0)}}\left(f_0\right) + \lambda\left\|f_0\right\|_H^2. \tag{55}$$

Recall that the point-wise infimum of a familiy of continuous functions yields an upper semi-continuous function; see, for example, (Denkowski et al., 2003, Prop. 1.1.36). Then, the definition of $f_{D^{(m)},\lambda}$ and continuity of $D \mapsto \mathcal{R}_D(f) + \lambda\|f\|_H^2$ for every $f \in H$ imply

$$\begin{aligned}
&\mathcal{R}_{\boldsymbol{D}^{(0)}}\left(f_0\right) + \lambda\left\|f_0\right\|_H^2 \;\geq\; \inf_{f\in H}\left(\mathcal{R}_{\boldsymbol{D}^{(0)}}(f) + \lambda\|f\|_H^2\right) \;\geq \\
&\quad\geq\; \limsup_{m\to\infty} \inf_{f\in H}\left(\mathcal{R}_{\boldsymbol{D}^{(m)}}(f) + \lambda\|f\|_H^2\right) \;=\; \limsup_{m\to\infty} \mathcal{R}_{\boldsymbol{D}^{(m)}}\left(f_{D^{(m)},\lambda}\right) + \lambda\left\|f_{D^{(m)},\lambda}\right\|_H^2 \;\geq \\
&\quad\geq\; \liminf_{m\to\infty} \mathcal{R}_{\boldsymbol{D}^{(m)}}\left(f_{D^{(m)},\lambda}\right) + \lambda\left\|f_{D^{(m)},\lambda}\right\|_H^2 \;\overset{(55)}{\geq}\; \mathcal{R}_{\boldsymbol{D}^{(0)}}\left(f_0\right) + \lambda\left\|f_0\right\|_H^2.
\end{aligned}$$

Hence, it follows that $f_0 = f_{D^{(0)},\lambda}$ and

$$\lim_{m\to\infty} \mathcal{R}_{\boldsymbol{D}^{(m)}}\left(f_{D^{(m)},\lambda}\right) + \lambda\left\|f_{D^{(m)},\lambda}\right\|_H^2 \;=\; \mathcal{R}_{\boldsymbol{D}^{(0)}}\left(f_{D^{(0)},\lambda}\right) + \lambda\left\|f_{D^{(0)},\lambda}\right\|_H^2. \tag{56}$$

Then, $f_0 = f_{D^{(0)},\lambda}$, (54), and (56) imply that $\lim_{m\to\infty}\|f_{D^{(m)},\lambda}\|_H = \|f_{D^{(0)},\lambda}\|_H$. Since weak convergence in the Hilbert space H and this convergence of the H-norms imply norm convergence in H (see, e.g., Conway, 1985, Exercise V.1.8), we have shown that $\lim_{m\to\infty}\|f_{D^{(m)},\lambda} - f_{D^{(0)},\lambda}\|_H = 0$, which is a contradiction to (52).

The following proof of part (i) of Lemma 9 (b) is essentially a variant of the proof of (Steinwart and Christmann, 2008, Theorem 5.9) even though the statements are quite different. Let $\partial L(y,t_0)$ denote the subdifferential of the convex map $t \mapsto L(y,t)$ at the point t_0. According to (Steinwart and

Christmann, 2008, Corollary 5.10), there is a bounded measurable map $h :\in \mathcal{X}_0 \times \mathcal{Y}_0 \to \mathbb{R}$ such that $h(x,y) \in \partial L\big(y, f_{P_0,\lambda_0}(x)\big)$ for every $(x,y) \in \mathcal{X}_0 \times \mathcal{Y}_0$ and

$$f_{P_0,\lambda_0} \;=\; -\frac{1}{2\lambda_0} \int h \Phi \, dP_0 \,. \tag{57}$$

The definition of the subdifferential implies

$$h(x,y)\big(f_{P_0,\lambda_1}(x) - f_{P_0,\lambda_0}(x)\big) \;\le\; L\big(y, f_{P_0,\lambda_1}(x)\big) - L\big(y, f_{P_0,\lambda_0}(x)\big)$$

for every $(x,y) \in \mathcal{X}_0 \times \mathcal{Y}_0$ and integrating with respect to P_0 yields

$$\int h(x,y)\big(f_{P_0,\lambda_1}(x) - f_{P_0,\lambda_0}(x)\big) P_0\big(d(x,y)\big) \;\le\; \mathcal{R}(f_{P_0,\lambda_1}) - \mathcal{R}(f_{P_0,\lambda_0}) \,.$$

The reproducing property of the canonical feature map Φ and the property of the Bochner integral (Denkowski et al., 2003, Theorem 3.10.16) imply

$$\begin{aligned}
&\int h(x,y)\big(f_{P_0,\lambda_1}(x) - f_{P_0,\lambda_0}(x)\big) P_0\big(d(x,y)\big) \;= \\
&\quad= \int \big\langle f_{P_0,\lambda_1} - f_{P_0,\lambda_0} \,, h(x,y)\Phi(x)\big\rangle_H \, P_0\big(d(x,y)\big) \;= \\
&\quad= \big\langle f_{P_0,\lambda_1} - f_{P_0,\lambda_0} \,, \textstyle\int h\Phi \, dP_0\big\rangle_H \overset{(57)}{=} \big\langle f_{P_0,\lambda_1} - f_{P_0,\lambda_0} \,, -2\lambda_0 f_{P_0,\lambda_0}\big\rangle_H \,.
\end{aligned}$$

That is,

$$\big\langle f_{P_0,\lambda_1} - f_{P_0,\lambda_0} \,, -2\tfrac{\lambda_0}{\lambda_1} f_{P_0,\lambda_0}\big\rangle_H \;\le\; \tfrac{1}{\lambda_1}\big(\mathcal{R}_{P_0}(f_{P_0,\lambda_1}) - \mathcal{R}_{P_0}(f_{P_0,\lambda_0})\big) \,. \tag{58}$$

An elementary calculation with $\langle \, , \rangle_H$ shows that

$$2\big\langle f_{P_0,\lambda_1} - f_{P_0,\lambda_0}, f_{P_0,\lambda_0}\big\rangle_H + \big\| f_{P_0,\lambda_1} - f_{P_0,\lambda_0}\big\|_H^2 = \big\| f_{P_0,\lambda_1}\big\|_H^2 - \big\| f_{P_0,\lambda_0}\big\|_H^2 . \tag{59}$$

Calculating (58)+(59) yields

$$\begin{aligned}
&\big\langle f_{P_0,\lambda_1} - f_{P_0,\lambda_0} \,, 2(1-\tfrac{\lambda_0}{\lambda_1}) f_{P_0,\lambda_0}\big\rangle_H + \big\| f_{P_0,\lambda_1} - f_{P_0,\lambda_0}\big\|_H^2 \;\le \\
&\qquad\le\; \tfrac{1}{\lambda_1}\big(\mathcal{R}_{P_0}(f_{P_0,\lambda_1}) - \mathcal{R}_{P_0}(f_{P_0,\lambda_0})\big) + \big\| f_{P_0,\lambda_1}\big\|_H^2 - \big\| f_{P_0,\lambda_0}\big\|_H^2 \;= \\
&\qquad=\; \tfrac{1}{\lambda_1}\big(\mathcal{R}_{P_0,\lambda_1}(f_{P_0,\lambda_1}) - \mathcal{R}_{P_0,\lambda_1}(f_{P_0,\lambda_0})\big) \;\le\; 0 \,.
\end{aligned}$$

Hence,

$$\begin{aligned}
\big\| f_{P_0,\lambda_1} - f_{P_0,\lambda_0}\big\|_H^2 \;&\le\; \Big|\big\langle f_{P_0,\lambda_1} - f_{P_0,\lambda_0} \,, 2(1-\tfrac{\lambda_0}{\lambda_1}) f_{P_0,\lambda_0}\big\rangle_H\Big| \;\le \\
&\le\; \big\| f_{P_0,\lambda_1} - f_{P_0,\lambda_0}\big\|_H \cdot 2\big|1-\tfrac{\lambda_0}{\lambda_1}\big| \cdot \big\| f_{P_0,\lambda_0}\big\|_H \,.
\end{aligned}$$

Since $\big\| f_{P_0,\lambda_0}\big\|_H \le \sqrt{\lambda_0^{-1}\mathcal{R}_{P_0}(0)}$ (see, e.g., Steinwart and Christmann, 2008, (5.4)), this implies statement (i) of Lemma 9 (b).

In order to prove (ii) and (iii) of Lemma 9 (b), note that the properties of the Bochner-Integral (see, e.g., Denkowski et al., 2003, Theorem 3.10.16) imply $\langle \int h\Phi \, dP, f\rangle_H = \int hf \, dP$ for every integrable function $h : \mathcal{X}_0 \times \mathcal{Y}_0 \to \mathbb{R}$ because of the reproducing property $\langle \Phi(x), f\rangle_H = f(x)$. Due to the

assumptions on L, it follows from (Steinwart and Christmann, 2008, Corollary 5.10) that there is a measurable function $h_{P_1,\lambda} : \mathcal{X}_0 \times \mathcal{Y}_0 \to \mathbb{R}$ which fulfills (50) and

$$\left\|f_{P_1,\lambda} - f_{P_2,\lambda}\right\|_H \leq \frac{1}{\lambda}\left\|\int h_{P_1,\lambda}\Phi\, dP_1 - \int h_{P_1,\lambda}\Phi\, dP_2\right\|_H =$$
$$= \sup_{\substack{f\in H\\ \|f\|_H\leq 1}} \frac{1}{\lambda}\left\langle \int h_{P_1,\lambda}\Phi\, dP_1 - \int h_{P_1,\lambda}\Phi\, dP_2\ ,\ f\right\rangle_H = \sup_{\substack{f\in H\\ \|f\|_H\leq 1}} \frac{1}{\lambda}\left|\int h_{P_1,\lambda} f\, dP_1 - \int h_{P_1,\lambda} f\, dP_2\right|.$$

That is, we have shown (ii). Then, (iii) follows from (ii) and $\|f\|_\infty \leq \|K\|_\infty \|f\|_H$. ■

References

Heinz Bauer. *Measure and Integration Theory*. Walter de Gruyter & Co., Berlin, 2001.

Kristin P. Bennett and Jennifer A. Blue. A support vector machine approach to decision trees. In *Proceedings IEEE International Joint Conference on Neural Networks*, 1998.

Enrico Blanzieri and Anton Bryl. Instance-based spam filtering using SVM nearest neighbor classifier. In *The 20th International FLAIRS Conference*, pages 441–442, 2007a.

Enrico Blanzieri and Anton Bryl. Evaluation of the highest probability SVM nearest neighbor classifier with variable relative error cost. In *Fourth Conference on Email and Anti-Spam CEAS 2007*, 2007b. URL `http://www.ceas.cc/2007/papers/paper-42_upd.pdf`.

Enrico Blanzieri and Farid Melgani. Nearest neighbor classification of remote sensing images with the maximal margin principle. *IEEE Transactions on Geoscience and Remote Sensing*, 46(6): 1804–1811, 2008.

Léon Bottou and Vladimir Vapnik. Local learning algorithms. *Neural Computation*, 4(6):888–900, 1992.

Fu Chang, Chien-Yang Guo, Xiao-Rong Lin, and Chi-Jen Lu. Tree decomposition for large-scale SVM problems. *Journal of Machine Learning Research*, 11:2935–2972, 2010.

Haibin Cheng, Pang-Ning Tan, and Rong Jin. Localized support vector machine and its efficient algorithm. In *Proceedings of the Seventh SIAM International Conference on Data Mining*, 2007. URL `http://www.siam.org/proceedings/datamining/2007/dm07_045cheng.pdf`.

Haibin Cheng, Pang-Ning Tan, and Rong Jin. Efficient algorithm for localized support vector machine. *IEEE Transactions on Knowledge and Data Engineering*, 22(4):537–549, 2010.

Andreas Christmann and Robert Hable. Consistency of support vector machines using additive kernels for additive models. *Computational Statistics and Data Analysis*, 56:854–873, 2012.

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

John B. Conway. *A Course in Functional Analysis*. Springer-Verlag, New York, 1985.

Zdzislaw Denkowski, Stanislaw Migórski, and Nikolas S. Papageorgiou. *An Introduction to Nonlinear Analysis: Theory*. Kluwer Academic Publishers, Boston, 2003.

Luc Devroye, László Györfi, Adam Krzyżak, and Gábor Lugosi. On the strong universal consistency of nearest neighbor regression function estimates. *The Annals of Statistics*, 22:1371–1385, 1994.

Luc Devroye, László Györfi, and Gábor Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer, New York, 1996.

Richard M. Dudley. *Real Analysis and Probability*. Cambridge University Press, Cambridge, 2002. Revised reprint of the 1989 original.

Nelson Dunford and Jacob T. Schwartz. *Linear Operators. I. General Theory*. Wiley-Interscience Publishers, New York, 1958.

Jinqing Fan and Irène Gijbels. *Local Polynomial Modelling and its Applications*. Chapman & Hall, London, 1996.

David H. Fremlin. *Measure Theory. Vol. 4*. Torres Fremlin, Colchester, 2006.

László Györfi, Michael Kohler, Adam Krzyżak, and Harro Walk. *A Distribution-free Theory of Nonparametric Regression*. Springer, New York, 2002.

Robert Hable and Andreas Christmann. On qualitative robustness of support vector machines. *Journal of Multivariate Analysis*, 102:993–1007, 2011.

Steven G. Krantz and Harold R. Parks. *Geometric Integration Theory*. Birkhäuser, Basel, 2008.

Kalyanapuram R. Parthasarathy. *Probability Measures on Metric Spaces*. Academic Press Inc., New York, 1967.

Bernhard Schölkopf and Alexander J. Smola. *Learning with Kernels*. MIT Press, Cambridge, 2002.

Nicola Segata and Enrico Blanzieri. Empirical assessment of classification accuracy of local SVM. In *The 18th Annual Belgian-Dutch Conference on Machine Learning (Benelearn 2009)*, pages 47–55, Tilburg, Belgium, 2009.

Nicola Segata and Enrico Blanzieri. Fast and scalable local kernel machines. *Journal of Machine Learning Research*, 11:1883–1926, 2010.

Ingo Steinwart and Andreas Christmann. *Support Vector Machines*. Springer, New York, 2008.

Vladimir Vapnik and Léon Bottou. Local algorithms for pattern-recognition and dependencies estimation. *Neural Computation*, 5(6):893–909, 1993.

Donghui Wu, Kristin P. Bennett, Nello Cristianini, John Shawe-Taylor, and Royal Holloway. Large margin trees for induction and transduction. In *Proceedings of International Conference on Machine Learning*, pages 474–483, 1999.

Alon Zakai. *Towards a Theory of Learning in High-Dimensional Spaces*. PhD thesis, The Hebrew University of Jerusalem, 2008. URL `http://icnc.huji.ac.il/phd/theses/files/AlonZakai.pdf`.

Alon Zakai and Ya'acov Ritov. Consistency and localizability. *Journal of Machine Learning Research*, 10:827–856, 2009.

Hao Zhang, Alexander C. Berg, Michael Maire, and Jitendra Malik. SVM-KNN: Discriminative nearest neighbor classification for visual category recognition. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2006)*, pages 2126–2136. IEEE Computer Society, 2006.