On Robustness and Consistency of Support Vector Machines for non-i.i.d. Observations

Von der Universität Bayreuth
zur Erlangung des Grades eines
Doktors der Naturwissenschaften (Dr. rer. nat.)
genehmigte Abhandlung

von

Katharina Strohriegl

aus Pegnitz

1. Gutachter: Prof. Dr. Andreas Christmann
2. Gutachter: Prof. Dr. Ingo Steinwart

Tag der Einreichung: 18.12.2017
Tag des Kolloquiums: 25.06.2018

Abstract

In recent years it becomes more and more important to learn hidden and complex structures from a given data set in an automatic and efficient way. Here statistical machine learning and in particular support vector machines are located. A lot of theoretical work on machine learning has been done under the assumption that the observations are realisations of independent and identically distributed (i.i.d.) random variables. This assumption might be mathematically convenient but it is often violated in practice or at least a doubtful assumption. Recently some work has been done to generalize statistical machine learning theory to non-i.i.d. stochastic processes, which also is the topic of this thesis.
Throughout this work we examine statistical robustness and consistency of estimators, in particular of support vector machines, for data generating stochastic processes with different dependence structures. To get reasonable results, we first introduce stochastic processes which provide convergence of their empirical measures to a limiting distribution. We call such processes weak respectively strong Varadarajan processes. Examples are many α-mixing processes, many Markov chains, and several weakly dependent processes. Concerning qualitative robustness, we prove a generalization of Hampel's famous theorem to Varadarajan processes. Estimators which are continuous and can be represented by a statistical operator on the space of probability measures are qualitatively robust if the data generating stochastic process is a weak Varadarajan process. It is not even necessary to strengthen the assumptions on the estimator, compared to those in Hampel's theorem for the i.i.d. case.
Further, qualitative robustness of bootstrap approximations is a desirable property, as the true distribution of the estimator is unknown in all cases of practical importance and therefore often replaced by a bootstrap approximation. Dropping the assumption of identical distributions, we show that the bootstrap approximation is still qualitatively robust if the empirical bootstrap is used and if the assumptions on the input space are strengthened. Compared to the results of the i.i.d. case, we have the same assumptions on the estimators, but require the process to be a strong Varadarajan process. Assuming uniform continuity instead of continuity of the statistical operator and assuming the input space to be compact, we achieve qualitative robustness for some α-mixing stochastic processes if the blockwise bootstrap is used.

Besides statistical robustness, consistency is of course also an important property of a sequence of estimators. Therefore the second part of this thesis focusses on consistency of support vector machines. We achieve consistency under common assumptions on the loss function and on the kernel. The stochastic process is assumed to be asymptotically mean stationary, which is implied by the Varadarajan property, and it is assumed to fulfil an almost sure convergence condition, similar to a law of large numbers. We show that many asymptotically mean stationary $\mathcal{C}$-mixing, weakly dependent, and α-mixing processes provide this assumption and therefore support vector machines are consistent for such processes. Compared to the i.i.d. case, our assumption on the convergence rate of the sequence of regularization parameters is only slightly stronger.

Zusammenfassung

Heutzutage wird es immer wichtiger, versteckte und komplexe Strukturen in Datensätzen möglichst automatisch und effizient zu finden. Oft werden hierzu Methoden der maschinellen Lerntheorie, zum Beispiel Support Vector Machines, eingesetzt. Die meisten theoretischen Ergebnisse zu Support Vector Machines sind allerdings für den Fall von unabhängig identisch verteilten (u.i.v.) stochastischen Prozessen hergeleitet. Dieser ist zwar mathematisch geeignet, in der Praxis ist die u.i.v.-Annahme aber häufig verletzt oder es ist unklar ob diese gilt. Deswegen versuchen wir zwei wichtige Eigenschaften von Schätzern, statistische Robustheit und Konsistenz, für datenerzeugende stochastische Prozesse zu zeigen, die nicht der u.i.v.-Annahme unterliegen. Dazu führen wir zunächst die sogenannten Varadarajan-Prozesse ein, diese garantieren Konvergenz ihres empirischen Maßes gegen eine Grenzverteilung. Beispiele für solche Prozesse sind einige α-mixing-Prozesse, Markov-Ketten und schwach abhängige Prozesse. Angelehnt an das bekannte Theorem zur qualitativen Robustheit von Hampel betrachten wir Schätzer, die stetig sind und durch einen statistischen Operator auf dem Raum der Wahrscheinlichkeitsmaße repräsentiert werden können. Für solche Schätzer und schwache Varadarajan-Prozesse erhalten wir die qualitative Robustheit des Schätzers. Im Vergleich zu Hampels Theorem für den u.i.v.-Fall ändert sich nur die Voraussetzung an den stochastischen Prozess, die an die Schätzer bleibt gleich.
Zusätzlich ist die Verteilung der datenerzeugenden Prozesse oft unbekannt und wird mit Hilfe eines Bootstrap-Verfahrens angenähert. Auch hierfür ist qualitative Robustheit eine wünschenswerte Eigenschaft. Für den empirischen Bootstrap und stochastische Prozesse, die zwar unabhängig aber nicht identisch verteilt sind, erhalten wir qualitative Robustheit unter den gleichen Voraussetzungen an die Schätzer wie im u.i.v.-Fall, der stochastische Prozess muss die Varadarajan Eigenschaft besitzen und die Voraussetzungen an den zugrundeliegenden Datenraum muss verstärkt werden. Auch für einige α-mixing-Prozesse zeigen wir qualitative Robustheit der Bootstrap-Approximation. Hierzu nehmen wir gleichmäßige Stetigkeit der Schätzer sowie einen kompakten Datenraum an. Die Approximation wird hierbei durch einen „Block-Bootstrap“ erreicht, dieser eignet sich besser für abhängige Daten als der klassische empirische Bootstrap.

Neben der Robustheit ist auch Konsistenz eine zentrale Eigenschaft von Schätzern. Im zweiten Teil der Arbeit zeigen wir Konsistenz für Support Vector Machines. Zusätzlich zu den üblichen Voraussetzungen an den Kern und die Verlustfunktion, benötigen wir einen stochastischen Prozess, der asymptotisch mittelwertstationär ist. Diese Eigenschaft wird zum Beispiel durch die Varadarjan Eigenschaft impliziert. Weiterhin muss der Prozess eine Konvergenzbedingung, ähnlich dem starken Gesetz der großen Zahlen, erfüllen. Für solche Prozesse sind Support Vector Machines konsistent. Wir zeigen, dass einige schwach abhängige, α- und $\mathcal{C}$-mixing Prozesse diese Konvergenzbedingung erfüllen. Verglichen mit u.i.v. stochastischen Prozessen muss die Folge der Regularisierungsparameter nur unmerklich langsamer konvergieren, diese Voraussetzungen sind also fast identisch.

Acknowledgements

My thanks go out to my supervisor Prof. Dr. Andreas Christmann for introducing me to this research area and suggesting open topics and ideas to me, to Dr. habil. Robert Hable for his support during the initial stage of my work and for introducing me to the world of mathematicians, and to all other members of the chair for offering help in many cases.

I want to thank the "Deutsche Forschungsgemeinschaft (DFG)" for supporting my research by financing the project "Support Vector Machines bei stochastischer Abhängigkeit".

Moreover, I would like to thank my parents for always supporting me and my colleagues Florian Dumpert, Manuela Dorn, and Tobias Kreisel for helpful discussions and for making my work days fun.

Bayreuth, 02.07.2018

Contents

1 Introduction **1**

2 Dependence structures **7**

2.1 Weak dependence 8

2.2 Mixing processes 9

2.3 $\mathcal{C}$-mixing processes 11

3 Qualitative robustness **13**

3.1 Qualitative robustness for non-i.i.d. observations 14

3.2 Examples for Varadarajan processes 23

3.2.1 Glivenko-Cantelli theorems, laws of large numbers, and the Varadarajan property 23

3.2.2 Examples 30

3.3 Examples for qualitatively robust estimators 38

3.4 Qualitative robustness for bootstrap estimators 39

3.4.1 Qualitative robustness for independent not necessarily identically distributed stochastic processes 42

3.4.2 Qualitative robustness for the moving block bootstrap of α-mixing processes 58

4 Support vector machines **73**

4.1 A short introduction to support vector machines 73

4.2 Qualitative robustness of support vector machines 79

4.3 Quantitative robustness of support vector machines - maximum bias 82

4.4 Consistency of support vector machines 86

4.4.1 Weakly dependent processes 103

4.4.2 α-mixing processes 117

4.4.3 $\mathcal{C}$-mixing processes 123

5 Conclusion and outlook **131**

A Appendix **135**

Notation

Sets and spaces

$((x_1, y_1), \ldots, (x_n, y_n))$	data set, consisting of $n \in \mathbb{N}$ data points
$(\Omega, \mathcal{A}, \mu)$	probability space
$(\mathcal{Z}^{\mathbb{N}}, \mathcal{B}^{\otimes\mathbb{N}}, \mathcal{M}(\mathcal{Z}^{\mathbb{N}}))$	statistical model
$\mathcal{A}$	σ-algebra
$\mathcal{B}$	Borel σ-algebra
$\mathcal{M}(\mathcal{Z})$	space of all probability measures on $\mathcal{Z}$
$\mathbb{N}$	positive integers, $\mathbb{N} = \{1, 2, 3, ...\}$
$\mathbb{R}$	set of real numbers
$\mathcal{Z}^{\mathbb{N}}$	sample space
$\mathrm{BL}(\mathcal{Z})$	space of Lipschitz continuous functions $f : \mathcal{Z} \to \mathbb{R}$ with bounded Lipschitz norm
$\mathbf{w}_n = (z_1, \ldots, z_n),\ n \in \mathbb{N}$	tuple of points in $\mathcal{Z}$
$\mathcal{X}, \mathcal{Y}, \mathcal{Z}$	sets, often metric spaces
$C^1(\mathcal{Z})$	space of continuously differentiable functions $f : \mathcal{Z} \to \mathbb{R}$
$C_b(\mathcal{Z})$	space of bounded, continuous functions $f : \mathcal{Z} \to \mathbb{R}$

Functions

$\mathbf{W}_n = (Z_1, \ldots Z_n)$	vector of random variables $Z_1, \ldots, Z_n$
$f_{L,P,\lambda}$	support vector machine
$k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$	kernel
$L : \mathcal{X} \times \mathcal{Y} \times \mathbb{R} \to [0, \infty)$	loss function
$L^* : \mathcal{X} \times \mathcal{Y} \times \mathbb{R} \to \mathbb{R}$	shifted loss function
$L_f : \mathcal{X} \times \mathcal{Y} \times \mathbb{R} \to \mathbb{R}$	abbreviated notation for loss function $L(x, y, f(x))$
$R^*_{L,P}$	Bayes risk
$R_{L,P}$	risk function
$S : \mathcal{M}(\mathcal{Z}) \to H$	statistical operator
$Z^*_1, \ldots, Z^*_n$	bootstrap sample
$Z_i : (\Omega, \mathcal{A}, \mu) \to (\mathcal{Z}, \mathcal{B})$	random variable
$(S_n)_{n\in\mathbb{N}}$, $S_n : \mathcal{Z}^n \to H$	sequence of estimators

$(Z_i)_{i\in\mathbb{N}}$	stochastic process

Measures

$K_{\mathbb{N}}$, $\tilde{K}_{\mathbb{N}}$	distributions on $\otimes_{i=1}^{k}\mathcal{Z}^{\mathbb{N}}$, $k\in\mathbb{N}$
μ	general probability measure
$\otimes_{i=1}^{n}P^i,\ n\in\mathbb{N}$	product measure of independent random variables, each with distribution P^i, $i\in\mathbb{N}$
$P_{\mathbb{N}}$, $Q_{\mathbb{N}}$	probability measures in $\mathcal{M}(\mathcal{Z}^{\mathbb{N}})$
$\mathbb{P}_{\mathbf{W}_n}=\frac{1}{n}\sum_{i=1}^{n}\delta_{Z_i}$	empirical measure of $(Z_1,\ldots,Z_n),\ n\in\mathbb{N}$
$\mathbb{P}_{\mathbf{w}_n}=\frac{1}{n}\sum_{i=1}^{n}\delta_{z_i}$	empirical measure of $(z_1,\ldots,z_n),\ n\in\mathbb{N}$
P	probability measure in $\mathcal{M}(\mathcal{Z})$
$P^i,\ i\in\mathbb{N}$	distribution of Z_i, $i\in\mathbb{N}$
$P_n^*,\ n\in\mathbb{N}$	bootstrap approximation of $P_n,\ n\in\mathbb{N}$
$P^{\otimes n},\ n\in\mathbb{N}$	product measure of i.i.d. random variables which have distribution P
$P_n,\ n\in\mathbb{N}$	finite joint distribution of $(Z_1,\ldots,Z_n),\ n\in\mathbb{N}$

Metrics and Norms

π or $\pi_{d_{\mathcal{Z}}}$	Prohorov metric (on $\mathcal{M}(\mathcal{Z},d_{\mathcal{Z}})$)
$\lvert\cdot\rvert_1$	Lipschitz constant
$\lVert\cdot\rVert_{\mathrm{BL}}=\lVert\cdot\rVert_\infty+\lvert\cdot\rvert_1$	bounded Lipschitz norm
$\lVert\cdot\rVert_\infty$	supremum norm
$\lVert\cdot\rVert_{\mathrm{TV}}$	total variation norm
$\lVert\cdot\rVert_p$	L^p-norm
d_{BL}	bounded Lipschitz metric
d_H	metric on the space H
$d_{n,p}$	p-product metric
$e, d_{\mathcal{Z}}$	metrics on $\mathcal{Z}$

Miscellaneous

$\langle\cdot,\cdot\rangle_H$	inner product on H
$\longrightarrow_D$	convergence in distribution (weak convergence)
$\longrightarrow_P$	convergence in probability
$\sharp A$	number of elements of the set A
$\mathcal{O}(\cdot)$	Landau symbol

Chapter 1

Introduction

"If we have data, let's look at data. If all we have are opinions, let's go with mine."

James L. Barksdale

Today, the question is, how to look at data? How to extract information from data? Often the relations and questions are too complex to solve for a human being or the amount of data or variables is too big. Here statistical machine learning is located. Machine learning "gives computers the ability to learn without being explicitly programmed", see Samuel (1959). The goal of supervised statistical learning is to find a function $f\colon \mathcal{X} \to \mathcal{Y}$, $\mathcal{X}, \mathcal{Y}$ sets, by using a given data set $\mathbf{w}_n := ((x_1, y_1), \ldots, (x_n, y_n)) \in (\mathcal{X} \times \mathcal{Y})^n$ to learn the relation between input values $x \in \mathcal{X}$ and output values $y \in \mathcal{Y}$, see for example Vapnik (1995) or Hastie et al. (2001). The learning algorithm is trained by a given data set, in order to be able to predict the outcome of a new input value. Consider, for example, certain characteristics of a vehicle, such as speed, height, or mass, used to assign the vehicle to different groups, for example "car" and "truck". After learning by means of some training data, where height, speed, and mass (input variables) of the vehicle and the kind of vehicle (output variable) is known, the algorithm should be able to classify every new, unknown combination of speed, height, and mass to one of the two groups, with small error probability.

There are various types of machine learning algorithms, the one we focus on are support vector machines (SVMs), see e. g. Boser et al. (1992), Vapnik (1995, 1998), Poggio and Girosi (1998), Schölkopf and Smola (2002), Cucker and Zhou (2007), and Steinwart and Christmann (2008). Support vector machines are considered as a nonparametric learning method

and can, in the case of supervised learning, be used either for classification, regression, or quantile regression. Historically support vector machines have been introduced for classification and linear functions only, see for example Vapnik (1995). Now, they are applied in a much broader sense. In case of support vector machines the function f is implicitly determined by a regularized optimization problem. Therefore we introduce the loss function L, a non-negative measurable function, which measures the distance between the observed output value and the predicted output value, and the risk, which is defined as the expected loss. Given a data set, $\mathbf{w}_n := ((x_1, y_1), \ldots, (x_n, y_n)) \in (\mathcal{X} \times \mathcal{Y})^n$, the statistical estimate is computed by minimizing the empirical risk added to a penalty term over a certain Hilbert space H of functions:

$$f_{L,\mathbb{P}_{\mathbf{w}_n},\lambda} := \arg\min_{f \in H} \frac{1}{n} \sum_{i=1}^{n} L(x_i, y_i, f(x_i)) + \lambda \|f\|_H^2.$$

Note that the penalty term is added in order to prevent overfitting and is weighted by $\lambda > 0$, more details can be found in Section 4.1. For the classification example above, the two groups "car" and "truck" would be labelled either "1" or "-1", and the SVM learned is a function $f_{L,\mathbb{P}_{\mathbf{w}_n},\lambda} \colon \mathcal{X} \to \{-1, 1\}$.

From a theoretical point of view the definition can be generalized to arbitrary probability measures P on $\mathcal{X} \times \mathcal{Y}$ (and the corresponding σ-algebra), that is the risk is computed with respect to the theoretical distribution P, $\lambda \in (0, \infty)$:

$$f_{L,P,\lambda} := \arg\min_{f \in H} \int L(x, y, f(x))\, dP(x, y) + \lambda \|f\|_H^2.$$

So far, the overwhelming part of theoretical works in machine learning has been done under the assumption, that the data can be considered as realisations of independent and identically distributed (i.i.d.) random variables. However, this assumption is not fulfilled in many practical applications so that non-i.i.d. cases increasingly attract attention. In addition to estimators especially designed for certain non-i.i.d. cases, practitioners often also use estimators originally designed for the i.i.d. case even if this assumption is violated. In Mukherjee et al. (1997) and Müller et al. (1997), for example, support vector machines are used for predicting time series with good results. Therefore this thesis focuses especially on non-i.i.d. stochastic processes, for example mixing processes or weakly dependent processes (in the sense of Doukhan and Louhichi (1999)). In particular, we mainly work with stochastic processes $(Z_i)_{i \in \mathbb{N}}$ which provide convergence of the empirical measures $\mathbb{P}_{\mathbf{W}_n}$, $n \in \mathbb{N}$, $\mathbf{W}_n = (Z_1, \ldots, Z_n)$, to a limiting distribution P on the space of probability measures, for

example with respect to the Prohorov metric π. That is

$$\pi(\mathbb{P}_{\mathbf{W}_n}, P) \longrightarrow 0 \text{ almost surely (or in probability)}, \; n \to \infty,$$

to which we refer as Varadarajan property, as it is similar to the result of Varadarajan's theorem for i.i.d. random variables, see Dudley (1989, Theorem 11.4.1). There are many stochastic processes which fulfil this assumption, for example many Markov chains, some martingales, several mixing processes or several weakly dependent process, see Chapter 3.2. Moreover we show: stochastic processes which fulfil a law of large numbers for events, in the sense of Steinwart et al. (2009), are Varadarajan processes under weak assumptions, see Theorem 3.2.1. An even weaker assumption on the stochastic process, also used here, is asymptotically mean stationarity, which is implied by the weak Varadarajan property.

Throughout this thesis some important properties of estimators are shown for those processes. A desirable property for estimators is qualitative robustness, which was first proposed in Hampel (1968). Roughly speaking, statistical robustness in general means that the estimator is only rarely affected by outliers or other small violations. Qualitative robustness in particular means, that the distributions of an estimator differ only slightly, if the underlying distributions of the data generating stochastic process are close together. That is, we assume a data set to be realisations of a stochastic process, with distribution $P_{\mathbb{N}}$, but the real data set may contain some additional errors or the assumption on the distribution is wrong. So the contaminated data set is generated by a stochastic process which may have a slightly different distribution $Q_{\mathbb{N}}$. The goal of qualitative robustness is to guarantee that the distribution of the estimator under the two distributions $P_{\mathbb{N}}$ and $Q_{\mathbb{N}}$ are close, as long as the distributions $P_{\mathbb{N}}$ and $Q_{\mathbb{N}}$ are close. It is well known that many classical estimators are not statistically robust, see for example Huber (1981), Hampel et al. (1986), Jurečková and Picek (2006), and Maronna et al. (2006) for some textbooks on robust statistics. The definition of qualitative robustness can be found in Hampel (1968) for the i.i.d. case, some generalizations can be found in Papantoni-Kazakos and Gray (1979), Cox (1981), and Boente et al. (1987). Throughout this work we use a generalization of Hampel's concept of Π-robustness proposed by Bustos (1980) to define qualitative robustness for non-i.i.d. observations, see Definition 3.1.1. In Theorem 3.1.3, we show that one of the classical results of qualitative robustness in the i.i.d. case, Hampel's theorem, can be generalized to the non-i.i.d. case if the underlying stochastic process fulfils the Varadarajan property. Compared to the i.i.d. case we do not strengthen the assumptions on the estimators and of course the i.i.d. case is included.

Moreover, the finite sample distribution of the data generating stochastic process is com-

monly unknown in practice. One way to get some information about this distribution are bootstrap methods. Here the distribution of the data generating stochastic process is estimated by resampling from the given observations. Historically, the bootstrap was introduced for the i.i.d. case, see Efron (1979). But there are various kinds of bootstrap methods used for different kinds of not necessarily i.i.d. stochastic processes, see for example Efron and Tibshirani (1993) and Shao and Tu (1995) for an introduction and an overview to the bootstrap theory. Regarding the bootstrap approximation for the distribution of the estimator, qualitative robustness is still desirable. The definition of qualitative robustness for bootstrap approximations can be found in Cuevas and Romo (1993). In Christmann et al. (2013) qualitative robustness for SVMs has been shown for the i.i.d. case. Our Theorem 3.4.2 gives a generalization of this result to the case of independent, but not necessarily identically distributed random variables. Additionally the assumptions on the sequence of estimators are slightly weakened. Strengthening the assumptions on the sequence of estimators and the assumptions on the stochastic process, we also achieve qualitative robustness for the bootstrap approximations of some α-mixing sequences, see Theorem 3.4.5 and 3.4.6.

Whereas the first results cover a broader class of estimators than support vector machines, the second part of this thesis focuses on robustness and consistency of support vector machines. For a given data set, the estimator can be computed with respect to this data set, that is we compute the empirical SVM. But for every data generating stochastic process, of course, there is the smallest possible risk, which relies on the distribution of this process. This distribution is commonly unknown, and therefore the empirical estimate is used. Hence, it is crucial to establish some kind of convergence of the empirical solution, that is statistical consistency. Here, we again consider stochastic processes which have the Varadarajan property or are asymptotically mean stationary. We examine convergence in probability of the risk of the empirical SVMs computed with respect to the limiting distribution P to the Bayes-risk $R^*_{L,P}$, which is defined as the smallest possible risk if all measurable functions $f : \mathcal{X} \to \mathcal{Y}$ are considered:

$$\int L(x, y, f_{L,\mathbb{P}_{\mathbf{W}_n},\lambda_n}(x))\, dP(x, y) \longrightarrow R^*_{L,P} \quad \text{in probability, } n \to \infty,$$

where the sequence of regularization parameters $(\lambda_n)_{n\in\mathbb{N}} \subset (0, \infty)$ is a suitable null-sequence. This is called L-risk-consistency. For the i.i.d. case, consistency of support vector machines is already shown, see for example Zhang (2004) and Christmann and Steinwart (2007) and the references in Chapter 4.4. Also learning rates are provided in this case, see e. g. Koltchinskii and Beznosova (2005), De Vito et al. (2005), and Blanchard et al. (2008). In the non-i.i.d. case, there are also some results, which yield that support vector machines are

still consistent and which provide learning rates. Therefore concentration inequalities for different dependence structures have been established, see for example Sun and Wu (2009) and Hang and Steinwart (2015). In Steinwart et al. (2009) consistency of support vector machines and of other regularized kernel methods is shown for a class of stochastic processes which satisfy some mixing conditions, or more generally, fulfil a law of large numbers for events. In Section 4.4, we show that support vector machines are consistent for some α-mixing, several weakly dependent and some $\mathcal{C}$-mixing processes, if they are additionally asymptotically mean stationary.

The next chapters are organised as follows: Chapter 2 gives a short introduction to weakly dependent processes in the sense of Doukhan and Louhichi (1999), α-mixing, and $\mathcal{C}$-mixing processes, as they are often used throughout this work. Chapter 3 focusses on qualitative robustness, including the introduction and definition of qualitative robustness in Section 3.1 and our generalization of Hampel's theorem, see Theorem 3.1.3. Moreover Varadarajan processes are introduced in this section. Examples for Varadarajan process, as well as the relation between laws of large numbers and Varadarajan processes are included in Section 3.2, examples for qualitatively robust estimators can be found in Section 3.3. Section 3.4 contains the definition and the main results about qualitative robustness of the bootstrap approximation, Theorem 3.4.2, Theorem 3.4.5, and Theorem 3.4.6.
The fourth chapter covers the results about support vector machines. A short introduction to support vector machines and reproducing kernel Hilbert spaces is given in Section 4.1. Results on qualitative robustness and the maximum bias of support vector machines are given in Theorem 4.2.1 and Theorem 4.3.2. Consistency of support vector machines is shown in Section 4.4. It contains a general result about consistency of support vector machines requiring a convergence assumption on the stochastic process, Theorem 4.4.4, and examples for stochastic processes which fulfil this assumption, see Theorem 4.4.6, Theorem 4.4.10, and Theorem 4.4.12. We would like to mention, that some results of Chapter 3 as well as Section 4.2 are already published in Strohriegl and Hable (2016), some parts of Section 3.4 are published in Strohriegl (2017) on arXiv. Concluding with Chapter 5 we give a short summary and propose some future research problems.

Chapter 2

Dependence structures

In order to work with general stochastic processes, a lot of different dependence notions have been introduced until now. For example Markov, mixing- and ergodic properties as well as mixingale structures, associated processes or weakly dependent processes. Throughout this thesis we regard qualitative robustness of estimators on general stochastic processes as well as consistency of support vector machines for general stochastic processes and therefore try to show our theorems for different dependence structures. Mainly used are weak dependence, mixing structures and $\mathcal{C}$-mixing processes. These dependence notions are shortly introduced in this chapter. Some results, for example the qualitative robustness, are more general and also work for Markov chains or martingales. The proofs of the results mainly require limit theorems, such as laws of large numbers or convergence conditions on empirical measures. Therefore we regard processes which describe the dependence between "past events" and "future events", which decreases when the gap between past and future increases. Roughly speaking, processes which forget the "past" if only the time gap is big enough. Weak dependence (in the sense of Doukhan and Louhichi (1999)) is based on the covariance between events in the past and events in the future. Whereas the mixing notions used here measure the dependence between the σ-algebras generated by the stochastic process. The $\mathcal{C}$-mixing structure is introduced separately, although it belongs to the mixing structures, but has been introduced in the context of dynamical systems. The $\mathcal{C}$-mixing coefficient is based on the covariance between the stochastic process and an arbitrary, bounded measurable function with respect to the σ-algebra generated by the stochastic process.

2.1 Weak dependence

This dependence notion has been introduced by Doukhan and Louhichi (1999) and Bickel and Bühlmann (1999). Roughly speaking, the dependence structure of a weakly dependent process is described through the covariance of a function f of "elementary events in the past" and another function g of "elementary events in the future". A process is considered to be weakly dependent if the covariance tends to zero as the distance between events in "past" and "future" increases. There are different types of weak dependence, named with different dependence coefficients. For the following results, we only consider non causal cases of weak dependence: η-, λ-, κ-, ζ- and θ-dependence. Therefore, we reduce the definition of weak dependence from Dedecker et al. (2007, Definition 2.2) to these cases. Let $(\Omega, \mathcal{A}, P)$ be a probability space, $\mathcal{Z}$ a Polish space, and $(Z_i)_{i\in\mathbb{N}}$, $Z_i : \Omega \to \mathcal{Z}$, $i \in \mathbb{N}$, a stochastic process. For every $u, v \in \mathbb{N}$, let $\mathcal{F}_u$ and $\mathcal{G}_v$ be classes of measurable functions $f : \mathcal{Z}^u \to \mathbb{R}$ respectively $g : \mathcal{Z}^v \to \mathbb{R}$; define $\mathcal{F} := \bigcup_{u\in\mathbb{N}} \mathcal{F}_u$, $\mathcal{G} := \bigcup_{v\in\mathbb{N}} \mathcal{G}_v$ and fix a function $\Psi : \mathcal{F} \times \mathcal{G} \to (0, \infty]$. For every $u, v \in \mathbb{N}$, let $\Gamma(u, v, \ell)$ be the set of $(i, j) \in \mathbb{Z}^u \times \mathbb{Z}^v$ such that $i_1 < \ldots < i_u \leq i_u + \ell \leq j_1 < \ldots < j_v$, $\ell \in \mathbb{N}$.
Then, the $(\mathcal{F}, \mathcal{G}, \Psi)$-dependence coefficient $\varepsilon(\ell)$ for the stochastic process $(Z_i)_{i\in\mathbb{N}}$ is defined by

$$\varepsilon(\ell) \;=\; \sup_{u,v\in\mathbb{N}} \sup_{(i,j)\in\Gamma(u,v,\ell)} \sup_{f\in\mathcal{F}_u, g\in\mathcal{G}_v} \frac{\big|\mathrm{Cov}\big(f(Z_{i_1}, ..., Z_{i_u}), g(Z_{j_1}, ..., Z_{j_v})\big)\big|}{\Psi(f, g)}. \tag{2.1}$$

The stochastic process $(Z_i)_{i\in\mathbb{N}}$ is called $(\mathcal{F}, \mathcal{G}, \Psi)$-*dependent* if

$$\lim_{\ell\to\infty} \varepsilon(\ell) \;=\; 0\,.$$

For our cases the functions $f : \mathcal{Z}^u \to \mathbb{R}$ are Lipschitz continuous with respect to the distance $d_{u,1}$ on $\mathcal{Z}^u$ defined by $d_{u,1}(z, z') := \sum_{i=1}^{u} d_{\mathcal{Z}}(z_i, z'_i)$, where $d_{\mathcal{Z}}$ is a metric on $\mathcal{Z}$, and the class $\mathcal{G}$ equals $\mathcal{F}$ for the non causal cases. Depending on the choice of the function Ψ and additional regularity assumptions on the functions in $\mathcal{F}$, different dependence coefficients are defined, see Dedecker et al. (2007): Here $|f|_1 := \sup_{z\neq z'} \frac{|f(z)-f(z')|}{d_{n,1}(z,z')}$ denotes the Lipschitz constant of f, $\|\cdot\|_\infty$ the supremum norm, and for $f \in \mathcal{F}_u$, $d_f := u$.

- The coefficient η corresponds to the choice $\Psi(f, g) = d_f\|g\|_\infty|f|_1 + d_g\|f\|_\infty|g|_1$, and $\mathcal{F}_u = \mathcal{G}_u$ is the set of all bounded Lipschitz functions $f : \mathcal{Z}^u \to \mathbb{R}$.
- The coefficient λ corresponds to the choice $\Psi(f, g) = d_f\|g\|_\infty|f|_1 + d_g\|f\|_\infty|g|_1$ $+ d_g d_f|g|_1|f|_1$, and $\mathcal{F}_u = \mathcal{G}_u$ is again the set of all bounded Lipschitz continuous functions.

- The coefficient κ corresponds to the function $\Psi(f,g) = d_f d_g |f|_1 |g|_1$ and $\mathcal{F}_u = \mathcal{G}_u$ is the set of all integrable Lipschitz continuous functions.
- The coefficient ζ corresponds to the choice $\Psi(f,g) = \min\{d_f, d_g\}|f|_1|g|_1$ and $\mathcal{F}_u = \mathcal{G}_u$ is again the set of all integrable Lipschitz continuous functions.
- Finally, the coefficient θ corresponds to the choice $\Psi(f,g) = d_g \|f\|_\infty |g|_1$, $\mathcal{F}_u$ is the set of all bounded functions $f : \mathcal{Z}^u \to \mathbb{R}$ and $\mathcal{G}_u$ is the class of Lipschitz continuous functions $g : \mathcal{Z}^u \to \mathbb{R}$. Moreover the random variables Z_i, $i \in \mathbb{N}$, are assumed to be L^1 integrable.

A good overview of result and definitions as well as examples for weakly dependent processes can be found in Dedecker et al. (2007).

2.2 Mixing processes

Another dependence structure which is used throughout this thesis are mixing processes. Mixing conditions of a stochastic process $(Z_i)_{i\in\mathbb{N}}$ are defined via various mixing coefficients which quantify the degree of dependence of the process. There exist several types of mixing coefficients, but all of them are based on differences between probabilities $\mu(A_1 \cap A_2) - \mu(A_1)\mu(A_2)$. There is a large literature on this dependence structure. For a detailed overview on mixing, see Bradley (2005), Bradley (2007a,b,c), and Doukhan (1994) and the references therein. We mainly use the α-mixing structure, which has been introduced in Rosenblatt (1956). Also examples of relations between dependence structures and mixing coefficients can be found in the references above.

Let Ω be a set equipped with two σ-algebras $\mathcal{A}_1$ and $\mathcal{A}_2$ and a probability measure μ. Let $\mathcal{L}^p(\mathcal{A}, \mu, H)$ be the space of all H-valued, $\mathcal{A}$-measurable, p-integrable functions. Analogously to e. g. Bradley (2005), using the convention $\frac{0}{0} = 0$, we can define the following measures of dependence:

$$\alpha(\mathcal{A}_1, \mathcal{A}_2, \mu) := \sup\{|\mu(A_1 \cap A_2) - \mu(A_2)\mu(A_2)| \mid A_1 \in \mathcal{A}_1,\ A_2 \in \mathcal{A}_2\}, \tag{2.2}$$

$$R^{\mathbb{R}}_\infty(\mathcal{A}_1, \mathcal{A}_2, \mu) := \sup\left\{ \left| \frac{\mathbb{E}_\mu fg - \mathbb{E}_\mu f \mathbb{E}_\mu g}{\|f\|_\infty \|g\|_\infty} \right| \;\middle|\; f \in \mathcal{L}^\infty(\mathcal{A}_1, \mu, \mathbb{R}),\ g \in \mathcal{L}^\infty(\mathcal{A}_2, \mu, \mathbb{R}) \right\}, \tag{2.3}$$

$$\phi(\mathcal{A}_1, \mathcal{A}_2, \mu) := \sup\{|\mu(A_2|A_1) - \mu(A_2)| \mid A_1 \in \mathcal{A}_1,\ A_2 \in \mathcal{A}_2,\ \mu(A_1) > 0\}, \tag{2.4}$$

$$\psi(\mathcal{A}_1, \mathcal{A}_2, \mu) := \sup\left\{ \left| \frac{\mu(A_1 \cap A_2)}{\mu(A_1)\mu(A_2)} - 1 \right| \;\middle|\; A_i \in \mathcal{A}_i,\ \mu(A_i) > 0,\ i \in \{1,2\} \right\}, \tag{2.5}$$

$$\rho(\mathcal{A}_1, \mathcal{A}_2, \mu) := \sup\{|\mathrm{Corr}(f, g)| \mid f \in \mathcal{L}^2(\mathcal{A}_1, \mu, \mathbb{R}),\ g \in \mathcal{L}^2(\mathcal{A}_2, \mu, \mathbb{R})\}, \tag{2.6}$$

$$\begin{aligned}\beta(\mathcal{A}_1, \mathcal{A}_2, \mu) &:= \sup \frac{1}{2} \sum_{i=1}^{I} \sum_{j=1}^{J} |\mu(A_{1,i} \cap A_{2,j}) - \mu(A_{1,i})\mu(A_{2,j})|, \\ &\text{where the supremum is taken over all (finite) partitions } \{A_{1,1}, \ldots, A_{1,I}\} \text{ and} \\ &\{A_{2,1}, \ldots, A_{2,J}\} \text{ of } \Omega, \text{ such that } A_{1,i} \in \mathcal{A}_1, \text{ for all } i \text{ and } A_{2,j} \in \mathcal{A}_2 \text{ for all } j.\end{aligned} \tag{2.7}$$

By definition the coefficients equal zero, if the σ-algebras are independent. Moreover the coefficients, besides ϕ, are symmetric in $\mathcal{A}_1$ and $\mathcal{A}_2$. Among those mixing properties α-mixing is the weakest condition:

$$\begin{aligned}2\alpha(\mathcal{A}_1, \mathcal{A}_2) &\leq \beta(\mathcal{A}_1, \mathcal{A}_2) \leq \phi(\mathcal{A}_1, \mathcal{A}_2) \\ 4\alpha(\mathcal{A}_1, \mathcal{A}_2) &\leq \rho(\mathcal{A}_1, \mathcal{A}_2) \leq \psi(\mathcal{A}_1, \mathcal{A}_2),\end{aligned} \tag{2.8}$$

see Bradley (2005, page 109). Again there are many other inequalities, which can be found therein. An important relation for the proofs of qualitative robustness and for the consistency of α-mixing sequences is the equivalence between the α-mixing coefficient and the $R_\infty^{\mathbb{R}}$-coefficient, see Bradley (1985), as it directly links the covariance to the α-mixing coefficient. According to this we have:

$$R_\infty^{\mathbb{R}}(\mathcal{A}_1, \mathcal{A}_2, \mu) \leq 2\pi\alpha(\mathcal{A}_1, \mathcal{A}_2, \mu). \tag{2.9}$$

Moreover mixing can be defined for stochastic processes. We follow Steinwart et al. (2009, Definition 3.1):

Definition 2.2.1 *Let $(Z_i)_{i\in\mathbb{N}}$ be a stochastic process, $Z_i : \Omega \to \mathcal{Z}$, $i \in \mathbb{N}$, and let $\sigma(Z_i)$ be the σ-algebra generated by Z_i, $i \in \mathbb{N}$. Then the α-bi-, the α- and $\overline{\alpha}$-mixing coefficients are defined by*

$$\begin{aligned}\alpha((Z)_{i\in\mathbb{N}}, \mu, i, j) &= \alpha(\sigma(Z_i), \sigma(Z_j), \mu) \\ \alpha((Z)_{i\in\mathbb{N}}, \mu, n) &= \sup_{i\geq 1} \alpha(\sigma(Z_i), \sigma(Z_{i+n}), \mu) \\ \overline{\alpha}((Z)_{i\in\mathbb{N}}, \mu, n) &= \sup_{i\geq 1} \alpha(\sigma(Z_1, \ldots, Z_i), \sigma(Z_{i+n}, Z_{i+n+1}, \ldots), \mu).\end{aligned}$$

A stochastic process $(Z_i)_{i\in\mathbb{N}}$ is called α- respectively $\overline{\alpha}$-mixing with respect to μ if

$$\lim_{n\to\infty} \alpha((Z)_{i\in\mathbb{N}}, \mu, n) = 0,$$
$$\textit{respectively} \quad \lim_{n\to\infty} \overline{\alpha}((Z)_{i\in\mathbb{N}}, \mu, n) = 0.$$

It is called weakly α- respectively weakly α-bi-mixing with respect to μ if

$$\lim_{n\to\infty} \frac{1}{n} \sum_{\ell=1}^{n} \alpha((Z)_{i\in\mathbb{N}}, \mu, \ell) = 0,$$
$$\textit{respectively} \quad \lim_{n\to\infty} \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{i-1} \alpha((Z)_{i\in\mathbb{N}}, \mu, i, j) = 0.$$

Of course these definitions can be used similarly for other mixing coefficients. Obviously $\alpha((Z)_{i\in\mathbb{N}}, \mu, n) \leq \overline{\alpha}((Z)_{i\in\mathbb{N}}, \mu, n)$. In most of the literature α-mixing for stochastic processes is defined similar to the $\overline{\alpha}$-mixing coefficient above. Also the inequalities can be expressed in terms of random variables, important for our proofs is:

$$R_{\infty}^{\mathbb{R}}(\sigma(Z_i), \sigma(Z_j), \mu,) \leq 2\pi\alpha(Z, \mu, i, j). \tag{2.10}$$

Similar to Steinwart et al. (2009), the following results only assume the process to be weakly α-*bi*-mixing, which is a slightly weaker assumption than the usual α-mixing condition, and is therefore introduced here.

2.3 $\mathcal{C}$-mixing processes

$\mathcal{C}$-mixing processes also belong to the group of mixing processes. They have been introduced especially to cover dynamical systems, as there are several examples of dynamical systems which are not α-mixing, see e. g. Doukhan and Louhichi (1999, page 41) and Dedecker and Prieur (2005) for other examples of stochastic processes which are not α-mixing. In Maume-Deschamps (2006), Hang and Steinwart (2015), and the references therein, examples of $\mathcal{C}$-mixing dynamical systems can be found. The $\mathcal{C}$-mixing coefficient as well as the α-mixing coefficient generalizes Φ-mixing. But in general neither $\mathcal{C}$-mixing implies α-mixing nor the other implication is right. According to Maume-Deschamps (2006, Definition 1) and Hang and Steinwart (2015, Definition 2.5) we define $\mathcal{C}$-mixing for stochastic processes $(Z_i)_{i\in\mathbb{N}}$, $Z_i : \Omega \to \mathcal{Z}$ for a measurable space $\mathcal{Z}$.

Let $\mathcal{C}$ be the Banach space of bounded functions $f\colon \mathcal{Z} \to \mathbb{R}$ with respect to the $\mathcal{C}$-norm $\|\cdot\|_{\mathcal{C}}$:

$$\|f\|_{\mathcal{C}} := \|f\|_\infty + \|f\| \tag{2.11}$$

where $\|\cdot\|_\infty$ denotes the supremum norm and $\|\cdot\|$ is a semi-norm on a vector space of bounded measurable functions $f\colon \mathcal{Z} \to \mathbb{R}$. For example consider the space of Lipschitz continuous functions with semi-norm $\|f\| = |f|_1 = \sup_{x\neq y} \frac{|f(x)-f(y)|}{d(x,y)}$, where $|f|_1$ is the Lipschitz constant of f, the space of $C^1 := \{f : \mathcal{Z} \to \mathbb{R} \mid \text{f bounded and continuously differentiable}\}$ functions on $\mathcal{Z} \subset \mathbb{R}$ open, equipped with semi-norm $\|f\| = \sup_{z\in\mathcal{Z}} |f'(z)|$, or the space of functions with bounded total variation with $\|f\| = \|f\|_{\mathrm{BV}}$. Moreover let $\mathcal{C}_1$ be the closed unit ball of functions f with respect to $\|\cdot\|_{\mathcal{C}}$.
Let $\|\cdot\|_1$ be the usual L^1-Norm on $\mathcal{Z}$, then $\mathcal{C}$-mixing processes are defined as follows:

Definition 2.3.1 ($\mathcal{C}$-mixing processes) *Let $(\Omega, \mathcal{A}, \mu)$ be a probability space and $(\mathcal{Z}, \mathcal{B})$ be a measurable space. Let $(Z_i)_{i\in\mathbb{N}}$, $Z_i : \Omega \to \mathcal{Z}$ be a stochastic process and let $\mathcal{A}_i^\ell$ be the σ-algebra on Ω generated by $(Z_i, \ldots, Z_\ell)$, $i \leq \ell \in \mathbb{N}$. Now define*

- *the $\mathcal{C}$-mixing coefficient by:*

$$\begin{aligned}\Phi_{\mathcal{C}}(\mathcal{Z}, n) := \sup \{ &|\mathbb{E}(f \circ Z_{i+n})\varphi - \mathbb{E}\varphi\mathbb{E}f \circ Z_{i+n}| \mid \\ &i \in \mathbb{N},\ f \in \mathcal{C}_1,\ \varphi\ (\mathcal{A}_1^i, \mathcal{B}) \text{ measurable with } \|\varphi\|_1 \leq 1 \},\end{aligned} \tag{2.12}$$

- *the time reversed $\mathcal{C}$-mixing coefficient by:*

$$\begin{aligned}\Phi_{\mathcal{C},rev}(\mathcal{Z}, n) := \sup \{ &|\mathbb{E}(f \circ Z_i)\varphi - \mathbb{E}f \circ Z_i\mathbb{E}\varphi| \mid \\ &i \in \mathbb{N},\ f \in \mathcal{C}_1,\ \varphi\ (\mathcal{A}_{i+n}^\infty, \mathcal{B}) \text{ measurable with } \|\varphi\|_1 \leq 1 \}.\end{aligned} \tag{2.13}$$

A stochastic process is called $\mathcal{C}$-mixing or time reversed $\mathcal{C}$-mixing if the coefficients $\Phi_{\mathcal{C}}$ respectively $\Phi_{\mathcal{C},rev}$ are summable.

Throughout the thesis, we are concerned with $\mathcal{C}$-mixing with respect to the class of bounded Lipschitz functions $\mathrm{BL}(\mathcal{Z}) := \{f : \mathcal{Z} \to \mathbb{R} \mid ||f||_{\mathrm{BL}} < \infty\}$ and therefore have:

$$\|f\|_{\mathcal{C}} := \|f\|_\infty + |f|_1 = \|f\|_{\mathrm{BL}},$$

where $\|\cdot\|_{\mathrm{BL}}$ is called the bounded Lipschitz norm.

Chapter 3

Qualitative robustness

Qualitative robustness is a continuity property of the estimator and means roughly speaking: small changes in the distribution of the data only lead to small changes in the distribution (i.e. the performance) of the estimator. In this way the following kinds of "small errors" are covered: small errors in all data points and large errors in only a small fraction of the data points (gross errors, outliers). Qualitative robustness of estimators has been defined originally in Hampel (1968) and Hampel (1971) in the i.i.d. case and has been generalized to estimators for stochastic processes in various ways, for example, in Papantoni-Kazakos and Gray (1979), Bustos (1980), which will be the one used here, Cox (1981), Boente et al. (1987), Zähle (2015), and Zähle (2016), for a more local consideration of qualitative robustness, see for example Krätschmer et al. (2017).
In the i.i.d. case, qualitative robustness is often proved by use of Hampel's theorem, see Hampel (1971) and also Cuevas (1988), as it is usually hard to be shown directly. By Hampel's theorem, qualitative robustness of an estimator is ensured if the estimator can be represented by a continuous statistical operator on the space of all probability measures. Here we generalize this theorem to those non-i.i.d. processes which provide convergence of their corresponding empirical measure. We also show that the empirical measure converges if the process satisfies a law of large numbers; this leads to various generalizations of Varadarajan's theorem to non-i.i.d. cases. Alternative generalizations of Hampel's theorem can be found in Zähle (2015) and Zähle (2016). Here only independence is weakened, while the data still have to be identically distributed. For a slightly different generalization of qualitative robustness, Hampel's theorem has been formulated for strongly stationary and ergodic processes in Cox (1981) and Boente et al. (1982); these processes are covered as a special case of our result.

3.1 Qualitative robustness for non-i.i.d. observations

Let $(\mathcal{Z}, d_{\mathcal{Z}})$ be a complete separable metric space with Borel σ-algebra $\mathcal{B}$. Denote by $\mathcal{M}(\mathcal{Z}^{\mathbb{N}})$ the set of all probability measures on $(\mathcal{Z}^{\mathbb{N}}, \mathcal{B}^{\otimes\mathbb{N}})$. Let $(\mathcal{Z}^{\mathbb{N}}, \mathcal{B}^{\otimes\mathbb{N}}, \mathcal{M}(\mathcal{Z}^{\mathbb{N}}))$ be the underlying statistical model. If nothing else is stated, we always use Borel σ-algebras for all topological spaces. Let $(Z_i)_{i\in\mathbb{N}}$ be the coordinate process on $\mathcal{Z}^{\mathbb{N}}$, that is $Z_i : \mathcal{Z}^{\mathbb{N}} \to \mathcal{Z}$, $(z_j)_{j\in\mathbb{N}} \mapsto z_i$, $i \in \mathbb{N}$. Then the process has law $P_{\mathbb{N}}$ under $P_{\mathbb{N}} \in \mathcal{M}(\mathcal{Z}^{\mathbb{N}})$. Moreover let $P_n := (Z_1, \ldots, Z_n)(P_{\mathbb{N}})$ be the n-th order marginal distribution of $P_{\mathbb{N}}$ for every $n \in \mathbb{N}$ and $P_{\mathbb{N}} \in \mathcal{M}(\mathcal{Z}^{\mathbb{N}})$. We are concerned with a sequence of estimators $(S_n)_{n\in\mathbb{N}}$ on the stochastic process $(Z_i)_{i\in\mathbb{N}}$. The estimator may take its values in any complete separable metric space H; that is, $S_n : \mathcal{Z}^n \to H$ for every $n \in \mathbb{N}$.

Following Boente et al. (1987), we use a definition originating from Bustos (1980) which generalizes Hampel's concept of Π-robustness:

Definition 3.1.1 (Qualitative robustness (Bustos (1980))) *Let π_n be the Prohorov metric on $\mathcal{M}(\mathcal{Z}^n)$ for every $n \in \mathbb{N}$. Then, the sequence of estimators $(S_n)_{n\in\mathbb{N}}$ is called qualitatively $(\pi_n)_{n\in\mathbb{N}}$-robust at $P_{\mathbb{N}}$ if, for every $\varepsilon > 0$, there is a $\delta > 0$ such that, for all $n \in \mathbb{N}$ and $Q_{\mathbb{N}} \in \mathcal{M}(\mathcal{Z}^{\mathbb{N}})$,*

$$\pi_n(P_n, Q_n) < \delta \;\Rightarrow\; \pi_{d_H}(\mathcal{L}_{P_n}(S_n), \mathcal{L}_{Q_n}(S_n)) < \varepsilon$$

where $\mathcal{L}_{P_n}(S_n)$ (and $\mathcal{L}_{Q_n}(S_n)$) denotes the distribution of the estimator S_n under P_n (and Q_n respectively) and π_{d_H} denotes the Prohorov metric on $\mathcal{M}(H)$.

Note that qualitative $(\pi_n)_{n\in\mathbb{N}}$-robustness at $P_{\mathbb{N}}$ is a local property.
Recall that the Prohorov metric π_e of two probability measures P and Q on any metric space $(\mathcal{X}, e)$ is given by

$$\pi_e(P, Q) \;=\; \inf\left\{\varepsilon > 0 :\; P(A) \le Q(A^\varepsilon) + \varepsilon \text{ for all measurable } A \subset \mathcal{X}\right\}$$

where $A^\varepsilon = \{x \in \mathcal{X} : e(x, A) < \varepsilon\}$.

Even in the i.i.d. case, it is usually hard to directly show qualitative robustness of estimators. Instead, qualitative robustness in the i.i.d. case is typically shown by use of Hampel's theorem (Hampel (1971, page 1892)); see also Cuevas (1988, Theorem 2) for estimators taking values in an arbitrary complete separable metric spaces. This theorem applies to

estimators which can be represented by a statistical operator S. This means, that there is a map $S : \mathcal{M}(\mathcal{Z}) \to H$ such that:

$$S\big(\mathbb{P}_{\mathbf{w}_n}\big) \;=\; S_n(\mathbf{w}_n) \;=\; S_n(z_1, \ldots, z_n) \;\; \forall\, \mathbf{w}_n = (z_1, \ldots, z_n) \in \mathcal{Z}^n \;\; \forall\, n \in \mathbb{N} \tag{3.1}$$

where $\mathbb{P}_{\mathbf{w}_n}$ denotes the empirical measure defined by $\mathbb{P}_{\mathbf{w}_n}(B) := \frac{1}{n}\sum_{i=1}^{n} I_B(z_i)$, $B \in \mathcal{B}$, for the observations $\mathbf{w}_n = (z_1, ..., z_n) \in \mathcal{Z}^n$. Then, according to Hampel's theorem, a sequence of estimators which can be represented by a operator via (3.1) is qualitatively robust with respect to the Prohorov metric π on $\mathcal{M}(\mathcal{Z})$ in the i.i.d. case if S is continuous (with respect to the Prohorov metric on $\mathcal{M}(\mathcal{Z})$).

The goal of this section is to obtain a similar result also in the non-i.i.d. case: accordingly, we restrict our attention to estimators which can be represented by a statistical operator. These estimators can be seen as plug-in estimators using the empirical measure. In case of non-i.i.d. data, applying an estimator based on the empirical measure is not always sensible because the empirical measure does not need to be meaningful then. However, using the empirical measure is possible if it converges for increasing sample size n. As will be seen, such a convergence of the empirical measure is the only assumption we need for $(Z_i)_{i\in\mathbb{N}}$, respectively $P_{\mathbb{N}}$. When working through the original proof of Hampel's theorem in Hampel (1971), it turns out that the i.i.d. assumption is only needed in one step of the proof in which Varadarajan's theorem is used: if $Z_i \sim P$ i.i.d., then, for almost every $(z_j)_{j\in\mathbb{N}} \in \mathcal{Z}^{\mathbb{N}}$, the empirical measure $\mathbb{P}_{\mathbf{W}_n(z_1,\ldots)}$ converges weakly to P for $n \to \infty$ and $\mathbf{W}_n = (Z_1, \ldots, Z_n)$. That is, in order to generalize Hampel's theorem, it is crucial to generalize Varadarajan's theorem to the non-i.i.d. case. This is the goal of the following section in which it is shown that Varadarajan's theorem can be generalized to many other processes such as certain mixing processes, strongly stationary ergodic processes, and certain weakly dependent processes. In particular, the independence assumption in Varadarajan's classical theorem can be relaxed to pairwise independence. Recall that weak convergence of probability measures on Polish spaces can be expressed by use of the Prohorov metric so that a reformulated version of Varadarajan's theorem says that, for $Z_i \sim P$ i.i.d.,

$$\pi_{d_{\mathcal{Z}}}\big(\mathbb{P}_{\mathbf{W}_n}, P\big) \;\xrightarrow[n\to\infty]{}\; 0 \quad \text{almost surely} \qquad \text{for} \quad \mathbf{W}_n = (Z_1, \ldots, Z_n)\,. \tag{3.2}$$

As shown in Section 3.2, also many non-i.i.d. processes fulfil (3.2) and we call any such process a *(strong) Varadarajan process* – and, if a.s.-convergence is replaced by convergence in probability, we use the term *weak Varadarajan process.* Recall that the convergence above depends on the probability measure, i. e. the Varadarajan property is a local property.

Definition 3.1.2 *Let $(\Omega, \mathcal{A}, \mu)$ be a probability space and $(\mathcal{Z}, d_{\mathcal{Z}})$ a separable metric space. Define $\mathbf{W}_n = (Z_1, \ldots, Z_n)$ for every $n \in \mathbb{N}$. Then the stochastic process $(Z_i)_{i\in\mathbb{N}}$, $Z_i : \Omega \to \mathcal{Z}$, $i \in \mathbb{N}$, is called* (strong) Varadarajan process *if there exists a probability measure $P \in \mathcal{M}(\mathcal{Z})$ such that*

$$\pi(\mathbb{P}_{\mathbf{W}_n}, P) \xrightarrow[n\to\infty]{} 0 \quad \textit{almost surely.}$$

It is called weak Varadarajan process *if*

$$\pi(\mathbb{P}_{\mathbf{W}_n}, P) \xrightarrow[n\to\infty]{} 0 \quad \textit{in probability,}$$

where π is the Prohorov metric on $\mathcal{M}(\mathcal{Z})$.

Now, we can state our generalization of Hampel's theorem, which is one of our main results. It says that, by use of our definition of Varadarajan processes, Hampel's theorem can be generalized to Bustos' notion of qualitative robustness for dependent data. A second result, stated later on (Theorem 3.2.1), then yields many examples for Varadarajan processes: whenever a process fulfils a law of large numbers, then it is a Varadarajan process. There are different kinds of generalizations of Hampel's theorem to the non-i.i.d. case. For example Cox (1981, Corollary 1) and Boente et al. (1982, Theorem 4.3) derive qualitative robustness at a probability measure $P_{\mathbb{N}}$ for strongly stationary ergodic processes. The assumptions on the statistical operator S and the estimator S_n, namely the continuity in $P_{\mathbb{N}}$ and the continuity on $\mathcal{Z}^{\mathbb{N}}$, are the same as in Theorem 3.1.3 below. As shown in Section 3.2, strongly stationary ergodic processes also have the Varadarajan property so that we cover these processes as a special case for qualitative robustness in the sense of Definition 3.1.

Theorem 3.1.3 *Let $\mathcal{Z}, H$ be complete separable metric spaces. Let the sequence of estimators $(S_n)_{n\in\mathbb{N}}$ be represented by an operator $S : \mathcal{M}(\mathcal{Z}) \to H$ via (3.1). Let $P_{\mathbb{N}} \in \mathcal{M}(\mathcal{Z}^{\mathbb{N}})$. If $(Z_i)_{i\in\mathbb{N}}$, $Z_i : \mathcal{Z}^{\mathbb{N}} \to \mathcal{Z}$, $(z_j)_{j\in\mathbb{N}} \mapsto z_i$, $i \in \mathbb{N}$ is a weak Varadarajan process under $P_{\mathbb{N}}$ with limiting distribution P, $S : \mathcal{M}(\mathcal{Z}) \to H$ is continuous (with respect to the Prohorov metric on $\mathcal{M}(\mathcal{Z})$) in P and the estimators $S_n : \mathcal{Z}^n \to H$, $n \in \mathbb{N}$, are continuous, then the sequence of estimators $(S_n)_{n\in\mathbb{N}}$ is qualitatively $(\pi_{d_n})_{n\in\mathbb{N}}$-robust at $P_{\mathbb{N}}$ where the metric d_n on $\mathcal{Z}^n$ is defined as*

$$d_n\big((z_1, \ldots, z_n), (z'_1, \ldots, z'_n)\big) \;=\; \inf\left\{\varepsilon > 0 : \sharp\{i : d_{\mathcal{Z}}(z_i, z'_i) \geq \varepsilon\}/n \leq \varepsilon\right\}. \tag{3.3}$$

Before we prove the result, it is advisable to have a closer look on the metrics, which should be used here. For the metric π_n on $\mathcal{M}(\mathcal{Z}^n)$ it is tempting to use a p-product metric $d_{n,p}$ on

$\mathcal{Z}^n$, that is,

$$d_{n,p}\big((z_1,\ldots,z_n),(z'_1,\ldots,z'_n)\big) \;=\; \big\|\big(d_{\mathcal{Z}}(z_1,z'_1),\ldots,d_{\mathcal{Z}}(z_n,z'_n)\big)\big\|_p \tag{3.4}$$

where $\|\cdot\|_p$ is the p_n-norm on $\mathbb{R}^n$ for $1 \le p \le \infty$. For example, $d_{n,2}$ is the Euclidean metric and $d_{n,\infty}\big((z_1,\ldots,z_n),(z'_1,\ldots,z'_n)\big) = \max_i d_{\mathcal{Z}}(z_i,z'_i)$; all these metrics are strongly equivalent (see Definition A1). However, some more care is needed here because, with these common metrics, the sample mean would turn out to be qualitatively $(\pi_{d_{n,p}})_{n\in\mathbb{N}}$-robust at every $P_{\mathbb{N}} \in \mathcal{M}(\mathcal{Z}^{\mathbb{N}})$; see Proposition 3.1.4 below. Following Boente et al. (1987) again, we use the metric d_n on $\mathcal{Z}^n$ defined in (3.3). This metric covers the intuitive meaning of robustness: two points in $\mathcal{Z}^n$ (i. e., two data sets) are close if only a small fraction of the coordinates are far-off (gross errors) and all other coordinates are close (small rounding errors). The ordinary p-product metrics $d_{n,p}$ would only cover rounding errors but exclude gross errors so that the sample mean becomes "robust", see Proposition 3.1.4. Though d_n is not strongly equivalent to $d_{n,p}$ in general, it is always topologically equivalent; see Lemma 3.1.5 in the Appendix. This is important as we consider $\mathcal{Z}^n$ as the n-fold product space of the Polish space $(\mathcal{Z}, d_{\mathcal{Z}})$. The product space $\mathcal{Z}^n$ is again a Polish space (in the product topology) and, according to Lemma 3.1.5, it is metrizable also with metric d_n. By use of $\pi_n = \pi_{d_n}$ in Definition 3.1.1, this notion of qualitative robustness indeed generalizes Hampel's Π-robustness: if $(Z_n)_{n\in\mathbb{N}}$, $Z_i \sim P$ i.i.d., then any sequence of estimators $(S_n)_{n\in\mathbb{N}}$ is qualitatively $(\pi_{d_n})_{n\in\mathbb{N}}$-robust at $P_{\mathbb{N}}$ if and only if it is Π-robust in P_1; see Boente et al. (1987, Theorem 3.1).

The following Proposition shows that the robustness of the sample mean depends on the metric; in a somewhat different setting, a similar result is given by Cox (1981, Proposition 3).

Proposition 3.1.4 *Let $\mathcal{Z} = \mathbb{R}$, $d_{\mathcal{Z}}(z,z') = |z - z'|$ for all $z, z' \in \mathbb{R}$.*

(a) The sample mean is $(\pi_{d_{n,p}})_{n\in\mathbb{N}}$-robust at every $P_{\mathbb{N}} \in \mathcal{M}(\mathcal{Z}^{\mathbb{N}})$.

(b) Let $(\mathcal{Z}^{\mathbb{N}}, \mathcal{B}^{\otimes\mathbb{N}}, P_{\mathbb{N}})$, $P_{\mathbb{N}} \in \mathcal{M}(\mathcal{Z}^{\mathbb{N}})$ be an arbitrary probability space and let $(Z_i)_{i\in\mathbb{N}}$, $Z_i : \mathcal{Z}^{\mathbb{N}} \to \mathcal{Z}$, $(z_j)_{j\in\mathbb{N}} \mapsto z_i$, $i \in \mathbb{N}$, be a stochastic process. If $(Z_i)_{i\in\mathbb{N}}$ satisfies

$$\frac{1}{n}\sum_{i=1}^{n} Z_i \;\to\; c \quad \textit{in probability}$$

for a constant $c > 0$ then the sample mean is not $(\pi_{d_n})_{n\in\mathbb{N}}$-robust at $P_{\mathbb{N}}$.

Note that if the assumption in part (b) of Proposition 3.1.4 is violated for $(Z_i)_{i\in\mathbb{N}}$, then using the sample mean is pointless anyway.

To prove Proposition 3.1.4 we need the following lemma on the topological equivalence of the metrics d_n and $d_{n,p}$, mentioned above.

Lemma 3.1.5 *Let $(\mathcal{Z}, d_{\mathcal{Z}})$ be a metric space. Then, for every $n \in \mathbb{N}$ and $p \in [1,\infty]$, the metrics $d_{n,p}$ and d_n defined in (3.4) and (3.3) are topologically equivalent on $\mathcal{Z}^n$.*

Proof: Let $\mathbf{w}_n^{(k)} = (z_1^{(k)}, \dots, z_n^{(k)}) \in \mathcal{Z}^n$ for all $k \in \mathbb{N}$ and $\mathbf{w}_n = (z_1, \dots, z_n) \in \mathcal{Z}^n$.
First, let $d_{n,p}(\mathbf{w}_n^{(k)}, \mathbf{w}_n) \to 0$ for $k \to \infty$. Then, according to (3.4) and (3.3) we have:

$$d_n(\mathbf{w}_n^{(k)}, \mathbf{w}_n) \;\le\; \max_{i\in\{1,\dots,n\}} d_{\mathcal{Z}}(z_i^{(k)}, z_i) \;\le\; d_{n,p}(\mathbf{w}_n^{(k)}, \mathbf{w}_n) \to 0, \quad \text{for } k \to \infty.$$

Conversely let $d_n(\mathbf{w}_n^{(k)}, \mathbf{w}_n) \to 0$ for $k \to \infty$. For every $\varepsilon_0 \in (0, \frac{1}{n})$ there is a $k_0 \in \mathbb{N}$ such that $d_n(\mathbf{w}_n^{(k)}, \mathbf{w}_n) \le \varepsilon_0$ for all $k \ge k_0$. Therefore the definition of d_n yields:

$$\sharp\{i \in \{1,\dots,n\} \mid d_{\mathcal{Z}}(z_i^{(k)}, z_i) \ge \varepsilon_0\} \le \varepsilon_0 n < 1, \text{ for all } k \ge k_0.$$

So, $\sharp\{i \in \{1,\dots,n\} \mid d_{\mathcal{Z}}(z_i^{(k)}, z_i) \ge \varepsilon_0\} = 0$ and therefore $d(z_i^{(k)}, z_i) < \varepsilon_0$ for all $i \in \{1,\dots,n\}$ and $k \ge k_0$. Hence,

$$d_{n,p}(\mathbf{w}_n^{(k)}, \mathbf{w}_n) < n^{1/p}\varepsilon_0, \text{ for all } k \ge k_0. \qquad \square$$

Now, we prove Proposition 3.1.4 concerning the qualitative robustness of the sample mean.

Proof of Proposition 3.1.4: For $\varepsilon > 0$, chose $\delta = \frac{1}{2}\varepsilon$. Let $P_{\mathbb{N}} \in \mathcal{M}(\mathcal{Z}^{\mathbb{N}})$ be an arbitrary probability measure, $(Z_i)_{i\in\mathbb{N}}$, $Z_i : \mathcal{Z}^{\mathbb{N}} \to \mathcal{Z}$, $(z_j)_{j\in\mathbb{N}} \mapsto z_i$, $i \in \mathbb{N}$, the i-th coordinate projection and define $P_n := (Z_1, \dots, Z_n)(P_{\mathbb{N}})$. Now choose $Q_{\mathbb{N}} \in \mathcal{M}(\mathcal{Z}^{\mathbb{N}})$, $Q_n = (Z_1, \dots, Z_n)(Q_{\mathbb{N}})$, such that $\pi_{d_{n,p}}(P_n, Q_n) < \delta$, for all $n \in \mathbb{N}$ and let the estimate $S_n(\mathbf{w}_n)$ be the sample mean $\frac{1}{n}\sum_{i=1}^n z_i$. According to the definition of the Prohorov distance:

$$P_n(A) \le Q_n(A^\delta) + \delta \quad \forall A \in \mathcal{B}^{\otimes n},\ n \in \mathbb{N}.$$

Hence with $A := S_n^{-1}(B), B \in \mathcal{B}$:

$$\mathcal{L}_{P_n}(S_n)(B) = P_n(A) \le Q_n(A^\delta) + \delta,\ n \in \mathbb{N}.$$

As $d_{n,p}(\mathbf{w}_n, \mathbf{w}'_n) < \delta$ implies $|S_n(\mathbf{w}_n) - S_n(\mathbf{w}'_n)| = |\frac{1}{n}\sum_{i=1}^n (z_i - z'_i)| \leq d_{n,p}(\mathbf{w}_n, \mathbf{w}'_n) < \delta$, we see $A^\delta \subset S_n^{-1}(B^\delta)$, $n \in \mathbb{N}$. Therefore $\mathcal{L}_{P_n}(S_n)(B) \leq Q_n(S_n^{-1}(B^\delta)) + \delta$, respectively

$$\pi_d(\mathcal{L}_{P_n}(S_n)(B), \mathcal{L}_{Q_n}(S_n)(B)) \leq \delta < \varepsilon \text{ for all } n \in \mathbb{N}$$

which implies the qualitative robustness at $P_{\mathbb{N}}$ and proves part (a) of Proposition 3.1.4.

For the second part choose $\varepsilon = \frac{1}{4}$ and $B = [c-1, c+1]$.

We show that for every $\delta > 0$, there is an $n \in \mathbb{N}$ and a $Q_n \in \mathcal{M}(\mathcal{Z}^n)$ such that $\pi_{d_n}(P_n, Q_n) < \delta$ but $\mathcal{L}_{P_n}(S_n)(B) > \mathcal{L}_{Q_n}(S_n)(B^\varepsilon) + \varepsilon$; this proves part (b). There is $n_1 \in \mathbb{N}$ such that for every $n \geq n_1$: $\mathcal{L}_{P_n}(S_n)(B) > \frac{1}{2}$, as $\frac{1}{n}\sum_{i=1}^n Z_i$ converges in probability to c. Furthermore define $Q_n = \mathcal{L}_{Q_{\mathbb{N}}}(Z_1, \ldots, Z_n)$ with $Q_n((z_1 + 2n, z_2, \ldots, z_n)) = P_n(z_1, z_2, \ldots, z_n)$. Hence

$$\frac{1}{n}\sum_{i=1}^n Z_i \xrightarrow[n\to\infty]{} c + 2 \quad \text{in probability}$$

and therefore there is $n_2 \in \mathbb{N}$ such that for all $n > n_2$: $\mathcal{L}_{Q_n}(S_n)(B^\varepsilon) < \frac{1}{4}$.
Now choose an arbitrary $\delta > 0$, and $n_3 \in \mathbb{N}$ such that $\frac{1}{n_3} < \delta$.
Since $d_n((z_1, \ldots, z_n), (z_1 + 2n, z_2, \ldots, z_n) \leq \frac{1}{n} < \delta$ for every $n \geq n_3$ it follows,

$$P_n(B) \leq Q_n(B^\delta) + \delta, \ \forall B \in \mathcal{B}^{\otimes n},$$

respectively $\pi_{d_n}(P_n, Q_n) < \delta$, for all $n \geq n_3$. But for any $n \geq \max\{n_1, n_2, n_3\}$ we have:

$$\mathcal{L}_{P_n}(S_n)(B) > \frac{1}{2} > \mathcal{L}_{Q_n}(S_n)(B^\varepsilon) + \varepsilon$$

and therefore the sample mean is not qualitatively $(\pi_{d_n})_{n\in\mathbb{N}}$-robust. □

The proof of Theorem 3.1.3 follows the lines of the proof of Hampel (1971, Theorem 1). However, some care is needed as independence is dropped and we have to work with probability measures on the product space $\mathcal{Z}^n$ and with the special metric d_n. First, we need the following Lemma which gives us a condition that implies qualitative robustness. It is a generalization of Hampel (1971, Lemma 1) but the proof is only a variant of the original proof. Let $\mathcal{Z}, H$ be complete separable metric spaces.

Lemma 3.1.6 *Let $(S_n)_{n\in\mathbb{N}}$, $S_n : \mathcal{Z}^n \to H$, be a sequence of estimators. Let $P_{\mathbb{N}} \in \mathcal{M}(\mathcal{Z}^{\mathbb{N}})$ be probability measures with n-th order marginal distribution $P_n = (Z_1, \ldots, Z_n)(P_{\mathbb{N}})$, such*

that for all $\varepsilon > 0$ and for all $\eta > 0$, there exists a $\delta > 0$ such that, for all $n \in \mathbb{N}$, there is a $B_n \in \mathcal{B}^{\otimes n}$ with the following properties

$$(i)\ P_n(B_n) > 1 - \eta \tag{3.5}$$

$$(ii)\ \text{If } d_n(\mathbf{w}_n, \mathbf{w}'_n) < \delta,\ \mathbf{w}_n \in B_n,\ \mathbf{w}'_n \in \mathcal{Z}^n \text{ then } d_H(S_n(\mathbf{w}_n), S_n(\mathbf{w}'_n)) < \varepsilon. \tag{3.6}$$

Then the estimator S_n, $n \in \mathbb{N}$, is qualitatively $(\pi_{d_n})_{n\in\mathbb{N}}$-robust at $P_{\mathbb{N}}$.

Proof: Let $\varepsilon > 0$, $n \in \mathbb{N}$ and $\eta := \frac{1}{2}\varepsilon$. By assumption, there is a $\delta > 0$ such that (3.5) and (3.6) applies. Define $\tilde{\delta} := \min\{\frac{1}{2}\delta, \frac{\varepsilon}{2}\}$ and choose $Q_{\mathbb{N}} \in \mathcal{M}(\mathcal{Z}^{\mathbb{N}})$ such that $\pi_{d_n}(P_n, Q_n) \leq \tilde{\delta}$, $n \in \mathbb{N}$. Then, according to Dudley (1989, Theorem 11.6.2), there exists $K_n \in \mathcal{M}(\mathcal{Z}^n \times \mathcal{Z}^n)$ with:

$$K_n(B_1 \times \mathcal{Z}^n) = P_n(B_1) \quad \forall B_1 \in \mathcal{B}^{\otimes n} \tag{3.7}$$

$$K_n(\mathcal{Z}^n \times B_2) = Q_n(B_2) \quad \forall B_2 \in \mathcal{B}^{\otimes n} \tag{3.8}$$

$$K_n\left(\left\{(\mathbf{w}_n, \mathbf{w}'_n) \in \mathcal{Z}^n \times \mathcal{Z}^n \mid d_n(\mathbf{w}_n, \mathbf{w}'_n) > \tilde{\delta}\right\}\right) < \tilde{\delta}. \tag{3.9}$$

With $\eta = \frac{1}{2}\varepsilon$, it follows that

$$\begin{aligned}
&K_n\left(\left\{(\mathbf{w}_n, \mathbf{w}'_n) \in \mathcal{Z}^n \times \mathcal{Z}^n \mid d_H(S_n(\mathbf{w}_n), S_n(\mathbf{w}'_n)) \leq \varepsilon\right\}\right) \\
&\quad \geq \quad K_n\left(\left\{(\mathbf{w}_n, \mathbf{w}'_n) \in \mathcal{Z}^n \times \mathcal{Z}^n \mid d_n(\mathbf{w}_n, \mathbf{w}'_n) < \delta,\ \mathbf{w}_n \in B_n\right\}\right) \\
&\quad \geq \quad K_n\left(\left\{(\mathbf{w}_n, \mathbf{w}'_n) \in \mathcal{Z}^n \times \mathcal{Z}^n \mid d_n(\mathbf{w}_n, \mathbf{w}'_n) \leq \tilde{\delta},\ \mathbf{w}_n \in B_n\right\}\right) \\
&\quad = \quad 1 - K_n\left(\left\{(\mathbf{w}_n, \mathbf{w}'_n) \in \mathcal{Z}^n \times \mathcal{Z}^n \mid d_n(\mathbf{w}_n, \mathbf{w}'_n) > \tilde{\delta} \text{ or } \mathbf{w}_n \notin B_n\right\}\right) \\
&\quad \overset{(3.7)}{\geq} \quad 1 - K_n\left(\left\{(\mathbf{w}_n, \mathbf{w}'_n) \in \mathcal{Z}^n \times \mathcal{Z}^n \mid d_n(\mathbf{w}_n, \mathbf{w}'_n) > \tilde{\delta}\right\}\right) - P_n(B_n^C) \\
&\quad \overset{(3.9),(3.5)}{>} \quad 1 - \tilde{\delta} - \eta \ \geq\ 1 - \varepsilon
\end{aligned}$$

and $K_n\left(\{(\mathbf{w}_n, \mathbf{w}'_n) \in \mathcal{Z}^n \times \mathcal{Z}^n \mid d_H(S_n(\mathbf{w}_n), S_n(\mathbf{w}'_n)) > \varepsilon\}\right) < \varepsilon$.

Now define $K_{\mathbb{N}} \in \mathcal{M}(\mathcal{Z}^{\mathbb{N}} \times \mathcal{Z}^{\mathbb{N}})$ such that $(\mathbf{W}_n, \mathbf{W}_n)(K_{\mathbb{N}}) = K_n$, $n \in \mathbb{N}$, where $\mathbf{W}_n = (Z_1, \ldots, Z_n) : \mathcal{Z}^{\mathbb{N}} \to \mathcal{Z}^n$ the projection on the first n coordinates.

Then we have:

$$\begin{aligned}
(\mathbf{W}_n, \mathbf{W}_n)(K_{\mathbb{N}})(B_1 \times \mathcal{Z}^n) &= K_n((B_1 \times \mathcal{Z}^n) = P_n(B_1) \quad \forall B_1 \in \mathcal{B}^{\otimes n} \\
(\mathbf{W}_n, \mathbf{W}_n)(K_{\mathbb{N}})(\mathcal{Z}^n \times B_2) &= K_n(\mathcal{Z}^n \times B_2) = Q_n(B_2) \quad \forall B_2 \in \mathcal{B}^{\otimes n}.
\end{aligned}$$

The boundedness of the Prohorov metric by the Ky Fan metric, see Dudley (1989, Theorem 11.3.5), yields for the Prohorov distance:

$$\begin{aligned}
\pi_{d_H}(S_n(P_n), S_n(Q_n)) &= \pi_{d_H}(S_n \circ \mathbf{W}_n(P_{\mathbb{N}}), S_n \circ \mathbf{W}_n(Q_{\mathbb{N}})) \\
&\leq \inf\{\tilde{\varepsilon} > 0 \mid K_{\mathbb{N}}(d_H(S_n \circ \mathbf{W}_n, S_n \circ \mathbf{W}_n) > \tilde{\varepsilon}) \leq \tilde{\varepsilon}\} \\
&= \inf\left\{\tilde{\varepsilon} > 0 \mid (\mathbf{W}_n, \mathbf{W}_n)(K_{\mathbb{N}})(\{(\mathbf{w}_n, \mathbf{w}'_n) \mid d_H(S_n(\mathbf{w}_n), S_n(\mathbf{w}'_n)) > \tilde{\varepsilon}\}) \leq \tilde{\varepsilon}\right\} \\
&= \inf\left\{\tilde{\varepsilon} > 0 \mid K_n(\{(\mathbf{w}_n, \mathbf{w}'_n) \mid d_H(S_n(\mathbf{w}_n), S_n(\mathbf{w}'_n)) > \tilde{\varepsilon}\}) \leq \tilde{\varepsilon}\right\} \leq \varepsilon
\end{aligned}$$

and therefore, the assertion. □

Proof of Theorem 3.1.3: As in the original proof of Hampel (1971, Theorem 1), we show at first that the conditions of Lemma 3.1.6 are satisfied for sufficiently large n.
Let $\varepsilon > 0$ and $\eta > 0$. With S being continuous at P, there exists a $\delta_0 > 0$ such that, for every $\mathbf{w}_n \in \mathcal{Z}^n$:

$$\pi(P, \mathbb{P}_{\mathbf{w}_n}) < 2\delta_0 \quad \Rightarrow \quad d_H(S(P), S(\mathbb{P}_{\mathbf{w}_n})) < \frac{\varepsilon}{2}. \tag{3.10}$$

Now, let $d_{\mathcal{Z}}$ denote the metric on $\mathcal{Z}$ and d_n is defined as in (3.3). For $\mathbf{w}_n = (z_1, \ldots, z_n)$ and $\mathbf{w}'_n = (z'_1, \ldots, z'_n)$, define $\mathcal{I} = \{i \in \{1, \ldots, n\} \mid d_{\mathcal{Z}}(z_i, z'_i) \geq \delta_0\}$. Then $d_n(\mathbf{w}_n, \mathbf{w}'_n) \leq \delta_0$ implies $\sharp\mathcal{I} \leq n\delta_0$ and therefore:

$$\mathbb{P}_{\mathbf{w}_n}(B) = \frac{1}{n}\sum_{i=1}^{n} I_B(z_i) = \frac{1}{n}\sum_{i \notin \mathcal{I}} I_B(z_i) + \frac{1}{n}\sum_{i \in \mathcal{I}} I_B(z_i) \leq \frac{1}{n}\sum_{i=1}^{n} I_{B^{\delta_0}}(z'_i) + \delta_0 = \mathbb{P}_{\mathbf{w}'_n}(B^{\delta_0}) + \delta_0.$$

With the definition of the Prohorov distance π it follows that:

$$d_n(\mathbf{w}_n, \mathbf{w}'_n) < \delta_0 \;\Rightarrow\; \pi_{d_{\mathcal{Z}}}(\mathbb{P}_{\mathbf{w}_n}, \mathbb{P}_{\mathbf{w}'_n}) \leq \delta_0. \tag{3.11}$$

Knowing that $(Z_i)_{i \in \mathbb{N}}$ is a weak Varadarajan process, we can find an $n_0 \in \mathbb{N}$ with

$$P_n\left(\{\mathbf{w}_n \in \mathcal{Z}^n \mid \pi_{d_{\mathcal{Z}}}(P, \mathbb{P}_{\mathbf{w}_n}) \geq \delta_0\}\right) < \eta \quad \forall n \geq n_0$$

Define the set $B_n := \{\mathbf{w}_n \in \mathcal{Z}^n \mid \pi_{d_{\mathcal{Z}}}(P, \mathbb{P}_{\mathbf{w}_n}) < \delta_0\}$, then: $P_n(B_n) > 1 - \eta$.

Therefore, for $\mathbf{w}_n \in B_n$ and $\mathbf{w}'_n \in \mathcal{Z}^n$ with $d_n(\mathbf{w}_n, \mathbf{w}'_n) < \delta_0$:

$$\pi_{d_{\mathcal{Z}}}(\mathbb{P}_{\mathbf{w}'_n}, P) \leq \pi_{d_{\mathcal{Z}}}(\mathbb{P}_{\mathbf{w}_n}, P) + \pi_{d_{\mathcal{Z}}}(\mathbb{P}_{\mathbf{w}_n}, \mathbb{P}_{\mathbf{w}'_n}) \overset{(3.11)}{<} 2\delta_0 \quad \text{as } \mathbf{w}_n \in B_n.$$

So, for every $\mathbf{w}_n \in B_n, \mathbf{w}'_n \in \mathcal{Z}^n$, (3.10) leads to:

$$d_H(S_n(\mathbf{w}_n), S_n(\mathbf{w}'_n)) \leq d_H(S_n(\mathbf{w}_n), S(P)) + d_H(S_n(\mathbf{w}'_n), S(P)) < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon. \qquad (3.12)$$

Due to Lemma 3.1.6 we can conclude: for all $\varepsilon > 0$, there is a δ_0 such that for all $n \geq n_0$,

$$\pi_{d_n}(P_n, Q_n) < \delta_0,\ Q_n \in \mathcal{M}(\mathcal{Z}^n) \quad \Rightarrow \quad \pi_{d_H}(\mathcal{L}_{P_n}(S_n), \mathcal{L}_{Q_n}(S_n)) < \varepsilon.$$

For $n < n_0$ we proceed as follows: as $\mathbf{w}_n \longmapsto S_n(\mathbf{w}_n) = S(\mathbb{P}_{\mathbf{w}_n})$ is continuous, so is $Q_n \longmapsto \mathcal{L}_{Q_n}(S_n)$ with respect to the weak topology. To show this, consider a sequence $(Q_{n,k})_{k\in\mathbb{N}} \subset \mathcal{M}(\mathcal{Z}^n)$ with $Q_{n,k} \to Q_{n,0}$ as $k \to \infty$ in the weak topology on $\mathcal{M}(\mathcal{Z}^n)$. Then, for every continuous and bounded f, the composition $f \circ S_n$ is again continuous and bounded so that:

$$\int f d(S_n(Q_{n,k})) = \int f \circ S_n \, dQ_{n,k} \xrightarrow{k\to\infty} \int f \circ S_n \, dQ_{n,0} = \int f d(S_n(Q_{n,0})).$$

So, for every $n < n_0$, for every $\varepsilon > 0$ there exists a δ_n such that:

$$\pi_{d_n}(P_n, Q_n) < \delta_n \quad \Rightarrow \quad \pi_{d_H}(\mathcal{L}_{P_n}(S_n), \mathcal{L}_{Q_n}(S_n)) < \varepsilon.$$

By choosing $\delta = \min\{\delta_0, \delta_1, \ldots, d_{n_0-1}\}$ the assertion of Theorem 3.1.3 follows. □

Another short remark should be made about the required continuity of S_n:

Remark 3.1.7 *The continuity of S_n on $\mathcal{Z}^n$ is with respect to the product topology on $\mathcal{Z}^n$ which is also generated by the p-metrics $d_{n,p}$. As already mentioned above, these metrics are topologically equivalent to d_n. Continuity of S_n is automatically fulfilled if S is not only continuous in P but on the whole domain $\mathcal{M}(\mathcal{Z})$. This follows from* (3.11) *and* (3.12) *in the proof of the above theorem, as we can use the continuity of S to show the continuity of $\mathbf{w}_n \to S_n(\mathbf{w}_n)$ there.*

In many cases, estimators originally developed for i.i.d. data are also used by practitioners in their data analysis for non-i.i.d. data. In this situation, a pleasant consequence of Theorem 3.1.3 is: any estimator which has been shown to be qualitatively robust by use of Hampel's theorem in the i.i.d. case is also qualitatively robust for the non-i.i.d. case without further ado – as long as $(Z_i)_{i\in\mathbb{N}}$ is a Varadarajan process on $(\mathcal{Z}^{\mathbb{N}}, \mathcal{B}^{\otimes\mathbb{N}}, P_{\mathbb{N}})$. Note that $P_{\mathbb{N}}$ plays the role of the ideal, uncontaminated distribution and that we only assume

the Varadarajan property for $(Z_i)_{i\in\mathbb{N}}$ for this ideal distribution. The observations may be contaminated and, accordingly, come from a different distribution $Q_{\mathbb{N}}$. The observed, contaminated process $(Z_i)_{i\in\mathbb{N}}$ on $(\mathcal{Z}^{\mathbb{N}}, \mathcal{B}^{\otimes\mathbb{N}}, Q_{\mathbb{N}})$ does not need to be Varadarajan. In view of the examples presented in the following section, this means that our results also cover violations of properties such as stationarity, ergodicity, mixing etc. This is contrary to Zähle (2015) and Zähle (2016) in which an alternative generalization of Hampel's theorem for non-i.i.d. cases is shown. There, the empirical measure has to converge not only for the ideal, uncontaminated process $(Z_i)_{i\in\mathbb{N}}$ for $P_{\mathbb{N}}$ but also for the observed, contaminated process $(Z_i)_{i\in\mathbb{N}}$ on $(\mathcal{Z}^{\mathbb{N}}, \mathcal{B}^{\otimes\mathbb{N}}, Q_{\mathbb{N}})$. Furthermore, only independence is dropped in Zähle (2015) and Zähle (2016) but the Z_i, $i \in \mathbb{N}$, are still assumed to be identically distributed for $P_{\mathbb{N}}$ as well as for $Q_{\mathbb{N}}$. However, the continuity assumption on S is less restrictive in Zähle (2015) and Zähle (2016) than in our Theorem 3.1.3; it is only assumed that S is continuous in P.
Different kinds of generalizations of Hampel's definition of qualitative robustness and their relationship can be found in Cox (1981) and Boente et al. (1982).

3.2 Examples for Varadarajan processes

3.2.1 Glivenko-Cantelli theorems, laws of large numbers, and the Varadarajan property

In order to find examples for Varadarajan processes, we connect the Varadarajan property to two classical concepts concerning convergence of the empirical distribution. Glivenko-Cantelli theorems are the first concept. The classical Glivenko-Cantelli theorem assumes i.i.d. stochastic processes with values in $\mathbb{R}$, and states the uniform convergence of the empirical distribution function $\mathbb{F}_n$ to the distribution function F:

$$\sup_{t\in\mathbb{R}} |\mathbb{F}_n(t) - F(t)| \longrightarrow 0 \text{ almost surely} \tag{3.13}$$

As we are especially interested in results for dependent observations, we now consider an arbitrary stochastic process with values in $\mathbb{R}$, that is a process which is not necessarily i.i.d. If this process also fulfils (3.13) as in the classical Glivenko-Cantelli theorem for i.i.d. processes, then it easily follows (from the Portmanteau theorem) that the process is also a a strong Varadarajan process. If convergence almost surely is replaced by convergence in probability, then it is a weak Varadarajan process. It is even possible to reformulate the definition of the Varadarajan property in terms of Glivenko-Cantelli theorems. A class $\mathcal{F}$ of

measurable functions $f : \mathcal{Z} \to \mathbb{R}$ is called Glivenko-Cantelli class if there exists a probability measure $P \in \mathcal{M}(\mathcal{Z})$ such that: $\sup_{f\in\mathcal{F}} |\frac{1}{n}\sum_{i=1}^{n} f(Z_i) - \int f\, dP| \to 0$ almost surely, see e. g. (van der Vaart, 1998, p. 269).
Now, let $\mathcal{F} := \mathrm{BL}_1(\mathcal{Z}, d_{\mathcal{Z}}) = \{f : \mathcal{Z} \to \mathbb{R} \mid \|f\|_{\mathrm{BL}} \leq 1\}$ be the set of bounded Lipschitz functions with $\|f\|_{\mathrm{BL}} \leq 1$, where $\|\cdot\|_{\mathrm{BL}} := |\cdot|_1 + \|\cdot\|_\infty$ denotes the bounded Lipschitz norm with $|f|_1 = \sup_{x\neq y} \frac{|f(x)-f(y)|}{d_{\mathcal{Z}}(x,y)}$ and $\|\cdot\|_\infty$ the supremum norm $\|f\|_\infty := \sup_x |f(x)|$ and $d_{\mathcal{Z}}$ is a metric on $\mathcal{Z}$. Then, it follows from Dudley (1989) Theorem 11.1.2 that $(Z_i)_{i\in\mathbb{N}}$ is a Varadarajan process if and only if $\mathcal{F}$ is a Glivenko-Cantelli class for $(Z_i)_{i\in\mathbb{N}}$.
To verify that a stochastic process fulfils a Glivenko-Cantelli theorem it is always necessary to show uniform convergence of the empirical distribution function. As it is often hard to show uniform convergence in applications we relate the Varadarajan property to a second classical concept, namely laws of large numbers. Theorem 3.2.1 below shows that any process which fulfils a (weak) law of large numbers is a (weak) Varadarajan process. This is of great practical value because, usually, it is much easier to show a non-uniform law of large numbers than Glivenko-Cantelli theorems or convergence in the Prohorov distance.

According to Definition 2.1 in Steinwart et al. (2009), a $\mathcal{Z}$-valued stochastic process on a measurable space $(\mathcal{Z}, \mathcal{B})$ satisfies the weak law of large numbers for events (WLLNE) if, for all $B \in \mathcal{B}$, there exists a constant $c_B \in \mathbb{R}$ such that: $\frac{1}{n}\sum_{i=1}^{n} I_B \circ Z_i \to c_B$ in probability as n tends to infinity. The process $(Z_i)_{i\in\mathbb{N}}$ is said to satisfy a strong law of large numbers for events (SLLNE) if the above convergence applies almost surely.

Theorem 3.2.1 *Let $(\Omega, \mathcal{A}, \mu)$ be a probability space, $(\mathcal{Z}, d_{\mathcal{Z}})$ a separable metric space, and $(Z_i)_{i\in\mathbb{N}}$ a stochastic process with $Z_i : \Omega \to \mathcal{Z}$.*

(a) If $(Z_i)_{i\in\mathbb{N}}$ satisfies the SLLNE then $(Z_i)_{i\in\mathbb{N}}$ is a strong Varadarajan process.

(b) If $(Z_i)_{i\in\mathbb{N}}$ satisfies the WLLNE then $(Z_i)_{i\in\mathbb{N}}$ is a weak Varadarajan process.

This theorem does not only provide us with many examples of (weak) Varadarajan processes in the next subsection, but is also interesting on its own as it can be seen as a generalization of Varadarajan's theorem for non-i.i.d. cases. In particular, from Etemadi's law of large numbers (see, e. g., Hoffmann-Jørgensen (1994, Chapter 4.12)) it follows then that the assumption of independence in Varadarajan's theorem can be relaxed to pairwise independence. Furthermore, from Birkhoff's ergodic theorem (see, e. g., Breiman (1968, Chapter 6)), it follows that Varadarajan's theorem is also valid for strongly stationary ergodic processes.

We need the following lemma which provides the fact that the set of bounded Lipschitz functions $\mathrm{BL}(\mathcal{Z}, e) := \{f : \mathcal{Z} \to \mathbb{R} \mid \|f\|_{\mathrm{BL}} < \infty\}$ is separable with respect to $\|\cdot\|_\infty$ if $(\mathcal{Z}, e)$ is totally bounded, in order to prove Theorem 3.2.1(a) and 3.2.1(b) The proof can be found in Dudley (1989, included in the proof of Theorem 11.4.1).

Lemma 3.2.2 *If $(\mathcal{Z}, e)$ is a totally bounded metric space, then $\mathrm{BL}(\mathcal{Z}, e)$ is separable with respect to $\|\cdot\|_\infty$.*

With this result we can give the proof of Theorem 3.2.1(a) and 3.2.1(b) for processes $(Z_i)_{i\in\mathbb{N}}$ with values in arbitrary separable metric spaces $(\mathcal{Z}, d_{\mathcal{Z}})$.

Proof of Theorem 3.2.1(a): According to Dudley (1989, Theorem 2.8.2) we can find a metric e on $\mathcal{Z}$ defining the same topology as $d_{\mathcal{Z}}$ such that $(\mathcal{Z}, e)$ is totally bounded. Then Lemma 3.2.2 yields existence of a countable and dense subset G of $\mathrm{BL}(\mathcal{Z}, e)$ with respect to $\|\cdot\|_\infty$. As $(Z_i)_{i\in\mathbb{N}}$ satisfies the SLLNE, there exists a probability measure P such that, for all $f \in \mathcal{L}^\infty(\mathcal{Z})$:

$$\mathbb{E}_P f = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} f \circ Z_i \quad \mu\text{-almost surely,}$$

see Steinwart et al. (2009, Lemma 2.5). Then, for all $g \in G$, we have a subset $N_g \in \mathcal{A}$ with $\mu(N_g) = 0$ such that

$$\mathbb{E}_P g = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} g \circ Z_i(w) \quad \forall \omega \in \Omega \backslash N_g \tag{3.14}$$

Due to the countability of G, we find $N = \bigcup_{g\in G} N_g$ with $\mu(N) = 0$ and for all $\omega \in \Omega\backslash N$, $g \in G$, (3.14) applies.
Let f be in $\mathrm{BL}(\mathcal{Z}, e)$, then for every $\varepsilon > 0$ there is a $g_\varepsilon \in G$ such that $\|f - g_\varepsilon\|_\infty < \varepsilon$ and

$$\begin{aligned}
&\left|\mathbb{E}_P f - \frac{1}{n}\sum_{i=1}^{n} f \circ Z_i\right| \\
&\leq \left|\mathbb{E}_P f - \mathbb{E}_P g_\varepsilon\right| + \left|\frac{1}{n}\sum_{i=1}^{n}(f \circ Z_i - g_\varepsilon \circ Z_i)\right| + \left|\mathbb{E}_P g_\varepsilon - \frac{1}{n}\sum_{i=1}^{n} g_\varepsilon \circ Z_i\right| \\
&\leq 2\|f - g_\varepsilon\|_\infty + \left|\mathbb{E}_P g_\varepsilon - \frac{1}{n}\sum_{i=1}^{n} g_\varepsilon \circ Z_i\right|.
\end{aligned}$$

Hence, it follows from the definition of N and (3.14) that

$$\limsup_{n\to\infty} \left| \mathbb{E}_P f - \frac{1}{n}\sum_{i=1}^{n} f \circ Z_i(\omega) \right| \leq 2\varepsilon \quad \forall \varepsilon > 0, \forall \omega \in \Omega \backslash N$$

and, therefore,

$$\mu\left(\left\{\omega \in \Omega \;\middle|\; \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} f \circ Z_i(\omega) = \int_{\mathcal{Z}} f\, dP, \quad f \in \mathrm{BL}(\mathcal{Z}, e)\right\}\right) = 1.$$

Due to the Portmanteau theorem, e. g. see Dudley (1989, Theorem 11.3.3), this implies $\mathbb{P}_{\mathbf{W}_n(\omega)} \to P$ weakly for almost every $\omega \in \Omega$, i. e., if $C_b(\mathcal{Z})$ is the set of all continuous and bounded functions $f : \mathcal{Z} \to \mathbb{R}$:

$$\mu\left(\left\{\omega \in \Omega \mid \lim_{n\to\infty} \int f\, d\mathbb{P}_{\mathbf{W}_n(\omega)} = \int f\, dP, f \in C_b(\mathcal{Z})\right\}\right) = 1.$$

Now the continuity of a function is a topological property and does not depend on the metric $d_{\mathcal{Z}}$ or e, if they define the same topology. Then we follow, again with the Portmanteau theorem, $\mu\left(\left\{\omega \in \Omega \mid \lim_{n\to\infty} \pi_{d_{\mathcal{Z}}}(P, \mathbb{P}_{\mathbf{W}_n(\omega)}) = 0\right\}\right) = 1$ and therefore, the assertion. □

To prove the second part of Theorem 3.2.1 we need the following lemma:

Lemma 3.2.3 *Let* $\mathrm{BL}_1 := \{f \in \mathrm{BL}(\mathcal{Z}, e) \mid \|f\|_{\mathrm{BL}} \leq 1\}$. *If* $(\mathcal{Z}, e)$ *is a totally bounded metric space, then* $\mathrm{BL}_1(\mathcal{Z}, e)$ *is totally bounded.*

Proof: Let (C, e_0) be the completion of $(\mathcal{Z}, e)$, according to Dudley (1989, Theorem 2.5.1). That is, there is a bijective isometry I: $(\mathcal{Z}, e) \to (A, e_0)$ such that $A \subset C$ is dense. With $(\mathcal{Z}, e)$ being totally bounded, (A, e_0) is also totally bounded. This applies, because for every $\varepsilon > 0$ there are $x_1^\varepsilon, ..., x_k^\varepsilon \in (\mathcal{Z}, e)$ such that for every $y \in (\mathcal{Z}, e)$ there is a $j \in \{1, ..k\}$ such that $e(y, x_j^\varepsilon) < \varepsilon$.
Now, choose an arbitrary $\varepsilon > 0$ and define $s_1^\varepsilon := I(x_1^\varepsilon), ..., s_k^\varepsilon := I(x_k^\varepsilon)$. For every $s \in A$ there is a $x \in \mathcal{Z}$ with $I(x) = s$ and there is a x_j^ε with $e_1(x, x_j^\varepsilon) < \varepsilon$. Then, applying that I is an isometry, $e_2(s, s_j^\varepsilon) = e_2(I(x), I(x_j^\varepsilon)) = e_1(x, x_j^\varepsilon) < \varepsilon$. So, for every $\varepsilon > 0$ one can find $s_1^\varepsilon, ..., s_k^\varepsilon$ such that $A \subset \bigcup_{i=1}^k B_\varepsilon(s_i^\varepsilon)$ where $B_\varepsilon(s)$ denotes the ball around x with radius ε.

So, the completion (C, e_0) is compact, as A is dense in C. Define the set $G := \{g \in \mathrm{BL}(A, e_0) \mid \|g\|_{\mathrm{BL}(A,e_0)} \leq 1\}$. Then we see from Dudley (1989, Proposition 11.2.3) that every $g \in G$ has an extension $h \in \mathrm{BL}(C, e_0)$ with $h_{|A} = g$ and $\|h\|_{\mathrm{BL}(C,e_0)} = \|g\|_{\mathrm{BL}(A,e_0)}$.
Moreover, the set $H := \{h \in \mathrm{BL}(C, e_0) \mid \text{h is an extension of } g \in G\}$ is uniformly bounded

in $C(C)$, where $C(C)$ denotes the set of all continuous functions $f : C \to \mathbb{R}$ because $\|h\|_\infty \leq \|h\|_{\mathrm{BL}(C,e_0)} = \|g\|_{\mathrm{BL}(A,e_0)} \leq 1$ for every $h \in H$, and H is equicontinuous as every $h \in H$ is Lipschitz with $|h|_1 \leq \|h\|_{\mathrm{BL}} \leq 1$. Applying the Arzelà-Ascoli theorem, see e. g. Conway (1985, Theorem VI 3.8), the set H, considered as a subset of $C(C)$, is totally bounded with respect to $\|\cdot\|_\infty$. i. e., for every $\varepsilon > 0$ there is a $k = k_\varepsilon \in \mathbb{N}$ such that there are $h_1^\varepsilon, ..., h_k^\varepsilon$ such that $H \subset \bigcup_{i=1}^k B_\varepsilon(h_i^\varepsilon)$.
Define $g_1^\varepsilon := h_{1_{|A}}^\varepsilon, ..., g_k^\varepsilon := h_{k_{|A}}^\varepsilon$ for every $\varepsilon > 0$. Using that H is totally bounded, we can find, for every $g \in G$, a $j \in \{1, \ldots, k\}$ such that $\|g - g_j^\varepsilon\|_\infty = \sup_{s\in A} |g(s) - g_j^\varepsilon(s)| = \sup_{s\in A} |h(s) - h_{j_{|A}}^\varepsilon(s)| \leq \sup_{s\in C} |h(s) - h_j^\varepsilon(s)| < \varepsilon$. So G is totally bounded with respect to $\|\cdot\|_\infty$.
A simple computation using the properties of I shows that, for all $g \in G$, the composition $g \circ I$ is an element of $\mathrm{BL}(\mathcal{Z}, e)$ with $\|g \circ I\|_{\mathrm{BL}(Z,e)} = \|g\|_{\mathrm{BL}(A,e_0)} \leq 1$. And therefore $\{g \circ I \,|g \in G\} \subset \mathrm{BL}_1$. An analogous computation shows that, for every $f \in \mathrm{BL}_1(\mathcal{Z}, e)$, the composition $f \circ I^{-1}$ is an element of $\mathrm{BL}(A, e_0)$ with $\|f\|_{\mathrm{BL}(\mathcal{Z},e)} = \|f \circ I^{-1}\|_{\mathrm{BL}(A,e_0)}$ and therefore, $f \circ I^{-1} \in G$. Hence we find, for every $f \in \mathrm{BL}_1(\mathcal{Z}, e)$, a $g \in G$ such that $g = f \circ I^{-1}$ and therefore, $f = g \circ I$, respectively $\mathrm{BL}_1(\mathcal{Z}, e) = \{f \in \mathrm{BL}(\mathcal{Z}, e) \mid \|f\|_{\mathrm{BL}} \leq 1\} \subset \{g \circ I \,|g \in G\}$. So both sets are equal. Now define, for every $\varepsilon > 0$, $f_1^\varepsilon := g_1^\varepsilon \circ I, ..., f_k^\varepsilon := g_k^\varepsilon \circ I$. As G is totally bounded we find, for every $\varepsilon > 0$ and every $g \in G$, a $j \in \{1, \ldots, k\}$ such that $\|g - g_j^\varepsilon\|_\infty < \varepsilon$. As there is, for every $f \in \mathrm{BL}_1(\mathcal{Z}, e)$, a g such that $f = g \circ I$ we can conclude for all $f \in \mathrm{BL}_1(\mathcal{Z}, e)$: $\|f - g_j^\varepsilon \circ I\|_\infty = \|g \circ I - g_j^\varepsilon \circ I\|_\infty < \varepsilon$, i. e. $\mathrm{BL}_1(\mathcal{Z}, e)$ is totally bounded. □

Proof of Theorem 3.2.1(b): Using that $(\mathcal{Z}, d_\mathcal{Z})$ is a separable metric space, Dudley (1989, Theorem 2.8.2) states that there is a metric e defining the same topology as $d_\mathcal{Z}$ such that $(\mathcal{Z}, e)$ is totally bounded.
As required the process $(Z_i)_{i\in\mathbb{N}}$ satisfies the WLLNE, and therefore, see Steinwart et al. (2009, Lemma 2.5), for all $\varepsilon > 0$ and for all $f \in \mathcal{L}^\infty(\mathcal{Z})$:

$$\lim_{n\to\infty} \mu\left(\left\{\omega \in \Omega \;\middle|\; \left|\mathbb{E}_P f - \frac{1}{n}\sum_{i=1}^n f \circ Z_i(\omega)\right| > \varepsilon\right\}\right) = 0.$$

Particularly this is true for every $f \in \mathrm{BL}_1(\mathcal{Z}, e)$. Because the space $\mathrm{BL}_1(\mathcal{Z}, e)$ is totally bounded, see Lemma 3.2.3, for every $\varepsilon > 0$, there are $k = k_\varepsilon \in \mathbb{N}$ and $f_j^\varepsilon, \ldots, f_k^\varepsilon \in \mathrm{BL}_1(\mathcal{Z}, e)$

such that, for every $f \in \mathrm{BL}_1(\mathcal{Z}, e)$ we can find a $j \in \{1, \ldots, k\}$ with $\|f - f_j^\varepsilon\|_\infty \leq \varepsilon$.

$$\begin{aligned}\left|\int_{\mathcal{Z}} f\, dP - \int_{\mathcal{Z}} f\, d\mathbb{P}_{\mathbf{w}_n}\right| &\leq \left|\int_{\mathcal{Z}} (f - f_j^\varepsilon)\, dP\right| + \left|\int_{\mathcal{Z}} (f - f_j^\varepsilon)\, d\mathbb{P}_{\mathbf{w}_n}\right| \\ &\quad + \left|\int_{\mathcal{Z}} f_j^\varepsilon\, dP - \int_{\mathcal{Z}} f_j^\varepsilon d\mathbb{P}_{\mathbf{w}_n}\right| \\ &\leq 2\|f - f_j^\varepsilon\|_\infty + \left|\int_{\mathcal{Z}} f_j^\varepsilon\, dP - \int_{\mathcal{Z}} f_j^\varepsilon d\mathbb{P}_{\mathbf{w}_n}\right|.\end{aligned}$$

Hence, for all $\varepsilon > 0$:

$$\sup_{f \in BL_1} \left|\int_{\mathcal{Z}} f\, dP - \int_{\mathcal{Z}} f\, d\mathbb{P}_{\mathbf{w}_n}\right| \leq 2\varepsilon + \max_{j \in \{1,\ldots,k\}} \left|\int_{\mathcal{Z}} f_j^\varepsilon\, dP - \int_{\mathcal{Z}} f_j^\varepsilon\, d\mathbb{P}_{\mathbf{w}_n}\right|$$

and, for all $\varepsilon' > 0$

$$\begin{aligned}&\mu\left(\left\{w \in \Omega \;\middle|\; \max_{j \in \{1,\ldots,k\}} \left|\int_{\mathcal{Z}} f_j^\varepsilon\, dP - \int_{\mathcal{Z}} f_j^\varepsilon\, d\mathbb{P}_{\mathbf{W}_n(\omega)}\right| > \varepsilon'\right\}\right) \\ &\leq \sum_{j=1}^{k} \mu\left(\left\{\omega \in \Omega \;\middle|\; \left|\int_{\mathcal{Z}} f_j^\varepsilon\, dP - \int_{\mathcal{Z}} f_j^\varepsilon\, d\mathbb{P}_{\mathbf{W}_n(\omega)}\right| > \varepsilon'\right\}\right) \longrightarrow 0, \;\; n \to \infty\end{aligned}$$

because of the required properties of $(Z_i)_{i \in \mathbb{N}}$.
For all $\varepsilon' > 0$ we obtain with $\varepsilon = \frac{\varepsilon'}{3}$:

$$\begin{aligned}&\mu\left(\left\{\omega \in \Omega \;\middle|\; \sup_{f \in \mathrm{BL}_1} \left|\int_{\mathcal{Z}} f\, dP - \int_{\mathcal{Z}} f\, d\mathbb{P}_{\mathbf{W}_n(\omega)}\right| > \varepsilon'\right\}\right) \\ &\leq \mu\left(\left\{\omega \in \Omega \;\middle|\; 2\varepsilon + \max_{j \in \{1,\ldots,k\}} \left|\int_{\mathcal{Z}} f_j^\varepsilon\, dP - \int_{\mathcal{Z}} f_j^\varepsilon\, d\mathbb{P}_{\mathbf{W}_n(\omega)}\right| > \varepsilon'\right\}\right) \\ &= \mu\left(\left\{\omega \in \Omega \;\middle|\; \max_{j \in \{1,\ldots,k\}} \left|\int_{\mathcal{Z}} f_j^\varepsilon\, dP - \int_{\mathcal{Z}} f_j^\varepsilon\, d\mathbb{P}_{\mathbf{W}_n(\omega)}\right| > \frac{\varepsilon'}{3}\right\}\right) \longrightarrow 0, \; n \to \infty.\end{aligned}$$

We choose the metric β_e on the set of all distributions on $(\mathcal{Z}, e)$:

$$\beta_e(P, Q) := \sup\left\{\left|\int_{\mathcal{Z}} f\, dP - \int_{\mathcal{Z}} f\, dQ\right| : \; f \in \mathrm{BL}(\mathcal{Z}, e), \; \|f\|_{\mathrm{BL}} \leq 1\right\}.$$

Then the above convergence yields: for all $\varepsilon > 0$, $\mu\left(\{\omega \in \Omega \mid \beta_e(P, \mathbb{P}_{\mathbf{w}_n}) > \varepsilon\}\right) \to 0$ for $n \to \infty$, see e. g. Dudley (1989, Proposition 11.3.2 and Theorem 11.3.3). It is also shown there

that for two distributions P and Q:

$$\pi_e(P,Q) \leq C(\beta_e(P,Q))^{\frac{1}{2}}$$

for some constant $C \in (0,\infty)$. Therefore, for all $\varepsilon > 0$ and $n \to \infty$:

$$\mu\left(\left\{\omega \in \Omega \mid \pi_e(P, \mathbb{P}_{\mathbf{W}_n(\omega)}) > \varepsilon\right\}\right) \longrightarrow 0. \tag{3.15}$$

The last step is to show that this applies not only for the metric e, but also for $d_{\mathcal{Z}}$. Therefore, we choose an arbitrary subsequence of $\mathbb{P}_{\mathbf{w}_n}$. As this subsequence satisfies the convergence in (3.15), there has to be a sub-subsequence $(\mathbb{P}_{\mathbf{w}_{n_k}})_{n_k \in \mathbb{N}}$ which converges almost surely, that is $\mu\left(\left\{\omega \in \Omega \mid \lim_{n_k \to \infty} \pi_e(P, \mathbb{P}_{\mathbf{W}_{n_k}(\omega)}) = 0\right\}\right) = 1$. According to Dudley (1989, Theorem 11.3.3) this is equivalent to

$$\mu\left(\left\{\omega \in \Omega \mid \lim_{n_k \to \infty} \int_{\mathcal{Z}} f \, d\mathbb{P}_{\mathbf{W}_{n_k}(\omega)} = \int_{\mathcal{Z}} f \, dP, \ f \in C_b(\mathcal{Z})\right\}\right) = 1$$

where $C_b(\mathcal{Z})$ denotes the set of all continuous and bounded functions on $\mathcal{Z}$. As the metrics $d_{\mathcal{Z}}$ and e define the same topology on $\mathcal{Z}$, it follows again from Dudley (1989, Theorem 11.3.3) that $\mu\left(\left\{(\omega \in \Omega \mid \lim_{n_k \to \infty} \pi_{d_{\mathcal{Z}}}(P, \mathbb{P}_{\mathbf{W}_{n_k}(\omega)}) = 0\right\}\right) = 1$. So, for every subsequence of $\mathbb{P}_{\mathbf{w}_n}$ there always exists a sub-subsequence $\mathbb{P}_{\mathbf{w}_{n_k}}$ with $\lim_{n_k \to \infty} \pi_{d_{\mathcal{Z}}}(P, \mathbb{P}_{\mathbf{w}_{n_k}}) = 0$ μ-almost surely. Hence, we can conclude that the whole sequence satisfies for all $\varepsilon > 0$:

$$\mu\left(\left\{\omega \in \Omega \mid \pi_{d_{\mathcal{Z}}}(P, \mathbb{P}_{\mathbf{W}_n(\omega)}) > \varepsilon\right\}\right) = 0,$$

which means, $(Z_i)_{i \in \mathbb{N}}$ is a weak Varadarajan process. □

The proofs above show, that the process has to satisfy, for all $\varepsilon > 0$ and for all $f \in \mathrm{BL}(\mathcal{Z}, e)$:

$$\lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} f \circ Z_i = \mathbb{E}_P f \quad \mu\text{-almost surely}$$

respectively, for all $\varepsilon > 0$ and for all $f \in \mathrm{BL}_1(\mathcal{Z}, e)$:

$$\lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} f \circ Z_i = \mathbb{E}_P f \quad \text{in probability,}$$

in order to be a strong, respectively a weak Varadarajan process. This is a slightly weaker condition which follows from the SLLNE, respectively the WLLNE property, see Steinwart et al. (2009, Lemma 2.5). Therefore we can weaken the assumptions in Theorem 3.2.1:

Theorem 3.2.4 *Let $(\Omega, \mathcal{A}, \mu)$ be a probability space, $(\mathcal{Z}, e)$ be totally bounded, and $(Z_i)_{i\in\mathbb{N}}$ a stochastic process with $Z_i : \Omega \to \mathcal{Z}$.*

(a) If there is a probability measure P on $(\mathcal{Z}, \mathcal{B})$ such that $(Z_i)_{i\in\mathbb{N}}$ satisfies

$$\mathbb{E}_P f = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} f \circ Z_i \quad \mu\text{-almost surely}$$

for all $f \in \mathrm{BL}(\mathcal{Z}, e)$, then $(Z_i)_{i\in\mathbb{N}}$ is a strong Varadarajan process.

(b) If there is a probability measure P on $(\mathcal{Z}, \mathcal{B})$ such that $(Z_i)_{i\in\mathbb{N}}$ satisfies

$$\mathbb{E}_P f = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} f \circ Z_i \quad \text{in probability}$$

for all $f \in \mathrm{BL}_1(\mathcal{Z}, e)$, then $(Z_i)_{i\in\mathbb{N}}$ is a weak Varadarajan process.

3.2.2 Examples

In the following, we briefly list examples for processes which satisfy a law of large numbers. The examples listed in Subsection 3.2.2 are all taken from Steinwart et al. (2009, Section 2.2 and 3.1). Then, we show in Subsection 3.2.2 that weakly dependent processes in the sense of Doukhan and Louhichi (1999) also satisfy a law of large numbers and, therefore, have the Varadarajan property. Here $(Z_i)_{i\in\mathbb{N}}$ is always a stochastic process with values in a Polish metric space $\mathcal{Z}$ equipped with some metric $d_{\mathcal{Z}}$.

Stationary ergodic processes, Markov chains and mixing processes

Let $(Z_i)_{i\in\mathbb{N}}$ be a strongly stationary ergodic process. Then, for every measurable $f : \mathcal{Z} \to \mathbb{R}$, the process $(f \circ Z_i)_{i\in\mathbb{N}}$ is again strongly stationary and ergodic. (Stationarity is an easy consequence of the definition; for ergodicity, see, e. g., Krengel (1985, Proposition 4.3)). Accordingly, it follows from Birkhoff's ergodic theorem (see, e. g., Breiman (1968, Chapter 6)) that

$$\frac{1}{n} \sum_{i=1}^{n} f \circ Z_i \xrightarrow[n\to\infty]{} \mathbb{E} f \circ Z_1 \qquad \text{almost surely}$$

provided that $\mathbb{E}|f \circ Z_1| < \infty$. Hence, by choosing indicator functions $f = I_B$, it follows that $(Z_i)_{i\in\mathbb{N}}$ satisfies the SLLNE and, therefore, is a strong Varadarajan process.

Markov chains are another example; these are often used when a future event depends only on the current state, and not on the past. We assume that $(Z_i)_{i\in\mathbb{N}}$ is a strongly stationary Markov chain so that, in particular, $\mu\big(Z_{n+1} \in B \,\big|\, Z_n\big) = \mu\big(Z_2 \in B \,\big|\, Z_1\big)$ for every $n \in \mathbb{N}$ and assume that the so-called "Doeblin condition" is fulfilled: there is a finite measure Q on $\mathcal{B}$, an $n \in \mathbb{N}$, and an $\varepsilon > 0$ such that, for all $B \in \mathcal{B}$ with $Q(B) \leq \varepsilon$, we have $\mu(Z_{n+1} \in B \,|\, Z_1 = \cdot) \leq 1 - \varepsilon$. Then, $(Z_i)_{i\in\mathbb{N}}$ satisfies the SLLNE and, therefore, is a strong Varadarajan process; see Steinwart et al. (2009, Theorem 2.12) and the references therein. As the Doeblin condition does not imply ergodicity, these processes are not covered by the example above.

Finally, many mixing processes also have the Varadarajan property. Mixing conditions of a process $(Z_i)_{i\in\mathbb{N}}$ are defined via various mixing coefficients which quantify the degree of dependence of the process. There exist several types of mixing coefficients but all of them are based on differences between probabilities $\mu(A\cap B)$ and $\mu(A)\mu(B)$. According to Steinwart et al. (2009, Proposition 3.2), a weakly α-*bi*-mixing processes $(Z_i)_{i\in\mathbb{N}}$ satisfies the WLLNE if it is also asymptotically mean stationary, i. e., $\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \mathbb{E} I_B \circ Z_i$ exists for every $B \in \mathcal{B}$. For example, let $(Z_i = T^{i-1})_{i\in\mathbb{N}}$ be an asymptotically mean stationary dynamical system, with strong mixing property $\lim_{n\to\infty} \sup_{A,B\in\mathcal{A}} |\mu(T^{-n}A\cap B) - \mu(T^{-n}A)(\mu(B))| = 0$. Then the process is α-mixing and therefore satisfies the WLLNE and hence is a weak Varadarajan process. Although strong mixing for asymptotically mean stationary dynamical systems implies ergodicity, see Gray (1988, p. 212) these processes are, due to the non-stationarity, not covered by the results of Cox (1981) and Boente et al. (1982).

Additionally, Bradley (2005, Theorem 3.3) shows, that for Markov chains boundedness of some mixing coefficients, such as ψ, ϕ or ρ- mixing, implies exponentially fast decay of these mixing coefficients, which implies α-mixing, see Bradley (2005, p. 112). If the Markov chains are additionally asymptotically mean stationary, they also satisfy the WLLNE. Obviously, any strongly stationary process is asymptotically mean stationary, so these processes are covered, too. If α-*bi*-mixing is replaced by $\overline{\alpha}$-mixing, then $(Z_i)_{i\in\mathbb{N}}$ even satisfies the SLLNE; see Steinwart et al. (2009, §3.1) and the references cited therein.

Weakly dependent processes

Another dependence structure which often leads to the Varadarajan property is the concept of weak dependence, introduced by Doukhan and Louhichi (1999) and Bickel and Bühlmann (1999). As introduced in Section 2.1 we examine the non-causal case of weak dependence, in particular η-, λ-, ζ-, κ-mixing, and θ-mixing processes.

The following theorem shows the Varadarajan property for strongly stationary processes, which are weakly dependent.

Theorem 3.2.5 *Let $(\Omega, \mathcal{A}, \mu)$ be a probability space, $(\mathcal{Z}, d_{\mathcal{Z}})$ be totally bounded and let $(Z_i)_{i\in\mathbb{N}}$, $Z_i : \Omega \to \mathcal{Z}$, $i \in \mathbb{N}$, be a stochastic process. If the process $(Z_i)_{i\in\mathbb{N}}$ is strongly stationary and weakly dependent for one of the cases mentioned above, then it is a weak Varadarajan process.*

Proof of Theorem 3.2.5: The proof shows that a stochastic process whose dependence coefficients behave as required fulfils the conditions of Theorem 3.2.4, and therefore is a weak Varadarajan process. As the proofs for the different dependence coefficients follow the same lines we will treat the coefficients separately only where necessary.

Due to the stationarity of $(Z_i)_{i\in\mathbb{N}}$,

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_\mu 1_B \circ Z_i = \mathbb{E}_\mu 1_B \circ Z_1;$$

In particular, the limit exists for every $B \in \mathcal{B}$.

Let $f : \mathcal{Z} \to \mathbb{R}$ be a function in $\mathrm{BL}_1(\mathcal{Z}, d_{\mathcal{Z}}) = \{f : \mathcal{Z} \to \mathbb{R} \mid \|f\|_{\mathrm{BL}} \leq 1\}$, such that f is not constant, i. e. $f \neq c$, $c \in \mathbb{R}$. For all $f \in \mathrm{BL}_1(\mathcal{Z}, d_{\mathcal{Z}})$, which are constant, the condition of Theorem 3.2.4 is clearly right. As convergence in $\mathcal{L}^p(\mu)$ implies convergence in probability we compute:

$$\begin{aligned}
&\mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n} f\circ Z_i - \mathbb{E}_\mu f \circ Z_1\right)^2 = \\
&= \frac{1}{n^2}\mathbb{E}\left[\sum_{i=1}^{n}\left(f\circ Z_i - \mathbb{E}_\mu f\circ Z_1\right)^2 + 2\sum_{i=1}^{n}\sum_{j=1}^{i-1}\left(f\circ Z_i - \mathbb{E}_\mu f\circ Z_1\right)\left(f\circ Z_j - \mathbb{E}_\mu f\circ Z_1\right)\right] \\
&= \frac{1}{n^2}\left(\sum_{i=1}^{n}\mathrm{Var}(f\circ Z_i) + 2\sum_{i=1}^{n}\sum_{j=1}^{i-1}\mathrm{Cov}(f\circ Z_i, f\circ Z_j)\right) \\
&= \frac{1}{n^2}\left(\sum_{i=1}^{n}\mathrm{Var}(f\circ Z_i) + 2\sum_{i=1}^{n}\sum_{j=1}^{i-1}\Psi(f,f)\frac{\mathrm{Cov}(f\circ Z_i, f\circ Z_j)}{\Psi(f,f)}\right) \\
&\leq \frac{1}{n^2}\left(n + 2\Psi(f,f)\sum_{i=1}^{n}\sum_{j=1}^{i-1}\frac{\mathrm{Cov}(f\circ Z_i, f\circ Z_j)}{\Psi(f,f)}\right).
\end{aligned}$$

Note that the assumption, that f is not constant, yields $\|f\|_\infty > 0$ and $|f|_1 > 0$. This implies $\Psi(f,f) > 0$. Also $f \in \mathrm{BL}_1(\mathcal{Z}, d_\mathcal{Z})$ implies $\mathrm{Var}(f \circ Z_i) \leq 1$.
Moreover $f \in \mathrm{BL}_1(\mathcal{Z}, d_\mathcal{Z})$ yields $\Psi(f,f) \leq 3$, for every considered dependence coefficient, i. e. the function $\Psi(f,f)$ is uniformly bounded for every $f \in \mathrm{BL}_1(\mathcal{Z}, d_\mathcal{Z})$.
Therefore we have:

$$\begin{aligned}
\mathbb{E}\left(\sum_{i=1}^{n} f \circ Z_i - \mathbb{E}_\mu f \circ Z_1\right)^2 &\leq \frac{1}{n}\left(1 + \frac{6}{n}\sum_{i=1}^{n}\sum_{j=1}^{i-1}\frac{|\mathrm{Cov}(f \circ Z_i, f \circ Z_j)|}{\Psi(f,f)}\right) \\
&\leq \frac{1}{n}\left(1 + \frac{6}{n}\sum_{\ell=1}^{n}(n-\ell)\varepsilon(\ell)\right) \\
&\leq \frac{1}{n}\left(1 + 6\sum_{\ell=1}^{n}\varepsilon(\ell)\right) \\
&\longrightarrow 0, \quad n \to \infty,
\end{aligned}$$

where the convergence of the second term follows from the fact, that the sequence $\varepsilon(\ell)$ converges to 0 for $\ell \to \infty$ and, accordingly, the arithmetic mean $\frac{1}{n}\sum_{\ell=1}^{n}\varepsilon(\ell)$ converges to 0 for $n \to \infty$, by Kronecker's Lemma, see Hoffmann-Jørgensen (1994, Theorem 4.9, Equation 4.9.1). Applying Theorem 3.2.4 yields, the weak Varadarajan property of the stochastic process $(Z_i)_{i\in\mathbb{N}}$. □

$\mathcal{C}$-mixing processes

Another example for weak Varadarajan processes are $\mathcal{C}$-mixing processes, which are introduced in Section 2.3. We use Cc-mixing with respect to the space of bounded, Lipschitz continuous functions $f\colon \mathcal{Z} \to \mathbb{R}$. That is the class $\mathcal{C}$ of functions equals the set of bounded Lipschitz functions $\mathrm{BL} := \{f : \mathcal{Z} \to \mathbb{R} \mid ||f||_{\mathrm{BL}} < \infty\}$ equipped with semi-norm $\|f\|_\mathcal{C} := \|f\|_{\mathrm{BL}} = \|f\|_\infty + |f|_1$.

Theorem 3.2.6 *Let $(\Omega, \mathcal{A}, \mu)$ be a probability space, let $(\mathcal{Z}, d_\mathcal{Z})$ be a totally bounded measurable space, and let $(Z_i)_{i\in\mathbb{N}}$ be an asymptotically mean stationary and $\mathcal{C}$-mixing stochastic process with $Z_i : \Omega \to \mathcal{Z}$, $i \in \mathbb{N}$. Then $(Z_i)_{i\in\mathbb{N}}$ is a weak Varadarajan process.*

Before we proof the result above, we need the following technical lemma which generalizes the AMS property to bounded and continuous functions.

Lemma 3.2.7 *Let $\mathcal{Z}$ be a metric space and let $(Z_i)_{i\in\mathbb{N}}$ be an asymptotically mean stationary stochastic process with limiting distribution P. Then, for every bounded and continuous function $f\colon \mathcal{Z} \to \mathbb{R}$:*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_\mu f \circ Z_i = \mathbb{E}_P f. \tag{3.16}$$

Proof: Since $(Z_i)_{i\in\mathbb{N}}$ is asymptotically mean stationary we have

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_\mu I_B \circ Z_i = P(B), \text{ for all } B \in \mathcal{B}.$$

Let $f\colon \mathcal{Z} \to \mathbb{R}$ be a continuous bounded function. As every measurable function can be approximated by simple functions, see for example Denkowski et al. (2003, Theorem 2.1. 68) we have: for every $\varepsilon > 0$, there is a simple function $g = \sum_{j=1}^{\ell} a_j I_{A_j}$, $\ell \in \mathbb{N}$, $A_j \subset \mathcal{Z}$, $a_j \in \mathbb{R}$, $j \in \{1, \ldots, \ell\}$, such that $\|f - g\|_\infty \leq \varepsilon$.

Hence,

$$\begin{aligned}
&\left| \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_\mu (f \circ Z_i) - \mathbb{E}_P f \right| \\
&\quad \leq \quad \left| \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_\mu (f \circ Z_i - g \circ Z_i) \right| + \left| \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_\mu (g \circ Z_i) - \mathbb{E}_P g \right| + |\mathbb{E}_P (g - f)| \\
&\quad \overset{\|f-g\|_\infty \leq \varepsilon}{\leq} \quad 2\varepsilon + \left| \frac{1}{n} \sum_{i=1}^{n} \left(\mathbb{E}_\mu \left[\sum_{j=1}^{\ell} a_j I_{A_j} \circ Z_i \right] \right) - \mathbb{E}_P \left(\sum_{j=1}^{\ell} a_j I_{A_j} \right) \right| \\
&\quad \leq \quad 2\varepsilon + \left| \sum_{j=1}^{\ell} a_j \left(\frac{1}{n} \sum_{i=1}^{n} \left(\mathbb{E}_\mu I_{A_j} \circ Z_i - \mathbb{E}_P I_{A_j} \right) \right) \right|.
\end{aligned}$$

As $\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_\mu I_B \circ Z_i = P(B)$, for all $B \in \mathcal{B}$, we have

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_\mu f \circ Z_i = \mathbb{E}_P f. \qquad \square$$

Proof of Theorem 3.2.6: Let $\mathcal{A}_i^k$ be the σ-algebra on Ω generated by $(Z_i, \ldots, Z_k)$, $i \leq k \in \mathbb{N}$. As $(Z_i)_{i\in\mathbb{N}}$ is asymptotically mean stationary, there exists a probability measure

$P \in \mathcal{M}(\mathcal{Z})$ such that

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_\mu I_B \circ Z_i = P(B) \text{ for all } B \in \mathcal{A}.$$

With Lemma 3.2.7, we have for every $f \in C_b(\mathcal{Z})$:

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_\mu f \circ Z_i = \mathbb{E}_P f.$$

Respectively, there is $n_0^f \in \mathbb{N}$ such that for all $n \geq n_0^f$:

$$\left| \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_\mu f \circ Z_i - \mathbb{E}_P f \right| \leq \frac{\varepsilon}{4}. \tag{3.17}$$

As $(\mathcal{Z}, d_{\mathcal{Z}})$ is a totally bounded metric space Lemma 3.2.3 yields that $\mathrm{BL}_1(\mathcal{Z}, d_{\mathcal{Z}})$ is totally bounded with respect to $\|\cdot\|_\infty$. That is, there is a finite subset $G \subset \mathrm{BL}_1(\mathcal{Z}, d_{\mathcal{Z}})$ such that for every $\varepsilon > 0$ and for every $f \in \mathrm{BL}_1(\mathcal{Z}, d_{\mathcal{Z}})$ there is $g_\varepsilon \in G$ such that

$$\|f - g_\varepsilon\|_\infty \leq \frac{\varepsilon}{4}.$$

Hence,

$$\begin{aligned}
&\left| \frac{1}{n} \sum_{i=1}^{n} f \circ Z_i - \int f \, dP \right| \\
&\qquad \leq \frac{1}{n} \sum_{i=1}^{n} \|f \circ Z_i - g_\varepsilon \circ Z_i\|_\infty + \left| \frac{1}{n} \sum_{i=1}^{n} g_\varepsilon \circ Z_i - \int g_\varepsilon \, dP \right| + \int \|f - g_\varepsilon\|_\infty \, dP \\
&\qquad \leq \frac{\varepsilon}{2} + \left| \frac{1}{n} \sum_{i=1}^{n} g_\varepsilon \circ Z_i - \int g_\varepsilon \, dP \right|.
\end{aligned}$$

And therefore

$$\sup_{f \in \mathrm{BL}_1(\mathcal{Z})} \left| \frac{1}{n} \sum_{i=1}^{n} f \circ Z_i - \int f \, dP \right| \leq \frac{\varepsilon}{2} + \max_{g \in \mathcal{G}} \left| \frac{1}{n} \sum_{i=1}^{n} g \circ Z_i - \int g \, dP \right|. \tag{3.18}$$

Now, (3.17) , (3.18) and Markov's inequality, see for example Hoffmann-Jørgensen (1994, Theorem 3.9) yield for all $n \geq \max_{g\in G}\{n_0^g\}$:

$$\begin{aligned}
&\mu\left(\left\{\omega\in\Omega \mid \sup_{f\in \mathrm{BL}_1}\left|\frac{1}{n}\sum_{i=1}^n f\circ Z_i(\omega) - \int f\,dP\right| > \varepsilon\right\}\right)\\
&\overset{(3.18)}{\leq} \quad \mu\left(\left\{\omega\in\Omega \mid \max_{g\in\mathcal{G}}\left|\frac{1}{n}\sum_{i=1}^n g\circ Z_i(\omega) - \int g\,dP\right| > \frac{\varepsilon}{2}\right\}\right)\\
&\leq \quad \sum_{g\in\mathcal{G}}\mu\left(\left\{\omega\in\Omega \mid \left|\frac{1}{n}\sum_{i=1}^n g\circ Z_i(\omega) - \int g\,dP\right| > \frac{\varepsilon}{2}\right\}\right)\\
&\overset{(3.17)}{\leq} \quad \sum_{g\in\mathcal{G}}\mu\left(\left\{\omega\in\Omega \mid \left|\frac{1}{n}\sum_{i=1}^n g\circ Z_i(\omega) - \frac{1}{n}\sum_{i=1}^n \mathbb{E}_\mu g\circ Z_i\right| > \frac{\varepsilon}{4}\right\}\right)\\
&\leq \quad \sum_{g\in\mathcal{G}}\frac{16}{\varepsilon^2}\mathbb{E}_\mu\left(\frac{1}{n}\sum_{i=1}^n g\circ Z_i - \frac{1}{n}\sum_{i=1}^n \mathbb{E}_\mu g\circ Z_i\right)^2\\
&\leq \quad \sum_{g\in\mathcal{G}}\frac{16}{\varepsilon^2 n^2}\left[\sum_{i=1}^n \mathbb{E}_\mu\left(g\circ Z_i - \mathbb{E}_\mu g\circ Z_i\right)^2\right.\\
&\qquad\qquad \left. +2\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\mathbb{E}_\mu(g\circ Z_i - E_\mu g\circ Z_i)(g\circ Z_j - \mathbb{E}_\mu g\circ Z_j)\right].
\end{aligned}$$

As $g\in \mathrm{BL}_1(\mathcal{Z}, d_{\mathcal{Z}})$, we have $\|g\|_\infty \leq 1$ and therefore for every $g\in G$:

$$\sum_{i=1}^n \mathbb{E}_\mu\left(g\circ Z_i - \mathbb{E}_\mu g\circ Z_i\right)^2 \leq 4n.$$

Moreover $(Z_i)_{i\in\mathbb{N}}$ is $\mathcal{C}$-mixing by assumption, that is

$$\begin{aligned}
&\sum_{\ell=1}^\infty \sup\left\{\left|\mathbb{E}\varphi(f\circ Z_{i+\ell}) - \mathbb{E}\varphi\mathbb{E}f\circ Z_{i+\ell}\right|;\right.\\
&\qquad \left. i\in\mathbb{N},\ f\in\mathcal{C}_1,\ \varphi\ (\mathcal{A}_1^i,\mathcal{B})\text{ measurable with }\|\varphi\|_1\leq 1\right\} < \infty,
\end{aligned}$$

see Definition 2.12 in Section 2.3.

Now, $\|g\|_\infty \leq 1$ implies $\|g\|_1 \leq 1$. Hence, we have for the sum of covariances above:

$$\begin{aligned}
&\sum_{i=1}^{n}\sum_{j=i+1}^{n} \mathbb{E}_\mu(g\circ Z_i - E_\mu g\circ Z_i)(g\circ Z_j - \mathbb{E}_\mu g\circ Z_j) \\
&\qquad = \sum_{i=1}^{n}\sum_{k=1}^{n-i} \mathbb{E}_\mu(g\circ Z_i - E_\mu g\circ Z_i)(g\circ Z_{i+k} - \mathbb{E}_\mu g\circ Z_{i+k}) \\
&\qquad = \sum_{i=1}^{n}\sum_{k=1}^{n-i} \mathbb{E}_\mu(g\circ Z_i)(g\circ Z_{i+k}) - E_\mu(g\circ Z_i)\mathbb{E}_\mu(g\circ Z_{i+k}) \\
&\qquad \leq n\sum_{k=1}^{n} \sup_{i\in\{1,\dots,n\}} |\mathbb{E}_\mu(g\circ Z_i)(g\circ Z_{i+k}) - E_\mu(g\circ Z_i)\mathbb{E}_\mu(g\circ Z_{i+k})| \\
&\qquad \leq n\sum_{k=1}^{n} \sup_{i\in\{1,\dots,n\},\,\varphi} |\mathbb{E}_\mu(\varphi(g\circ Z_{i+k})) - E_\mu\varphi\mathbb{E}_\mu(g\circ Z_{i+k})|
\end{aligned}$$

for $\mathcal{A}_1^i$-measurable functions φ with $\|\varphi\|_1 \leq 1$.

Moreover as the process is $\mathcal{C}$-mixing the last sum is finite.

Therefore:

$$\begin{aligned}
&\mu\left(\left\{\omega\in\Omega \mid \sup_{f\in \mathrm{BL}_1}\left|\frac{1}{n}\sum_{i=1}^{n} f\circ Z_i(\omega) - \int f\,dP\right| > \varepsilon\right\}\right) \\
&\qquad \leq \sum_{g\in\mathcal{G}} \frac{16}{\varepsilon^2 n^2}\left[\sum_{i=1}^{n} \mathbb{E}_\mu\left(g\circ Z_i - \mathbb{E}_\mu g\circ Z_i\right)^2\right. \\
&\qquad\quad \left. +2\sum_{i=1}^{n}\sum_{j=i+1}^{n} \mathbb{E}_\mu(g\circ Z_i - E_\mu g\circ Z_i)(g\circ Z_j - \mathbb{E}_\mu g\circ Z_j)\right] \\
&\qquad \leq \sum_{g\in\mathcal{G}} \frac{16}{\varepsilon^2 n^2}[4n + 2n\sum_{\ell=1}^{n} \Phi_{\mathcal{C}}(Z,\ell)] \\
&\qquad < \frac{C}{n} \longrightarrow 0, \quad n\to\infty,
\end{aligned}$$

for a constant $C > 0$. Hence, $(Z_i)_{i\in\mathbb{N}}$ is a weak Varadarajan process. □

3.3 Examples for qualitatively robust estimators

In this chapter estimators which are qualitatively robust even for non-i.i.d. observations are given. Another example are support vector machines or, more generally, regularized kernel methods, which are discussed in Section 4.2.

A first example for qualitatively robust estimators are maximum likelihood type estimators (M-estimators). These are defined as solutions of

$$\sum_{i=1}^{n} \rho(z_i, S_n) = \min!$$

or implicitly by

$$\sum_{i=1}^{n} \psi(z_i, S_n) = 0,$$

see Huber (1981). Especially we consider estimators for location, that is $\psi(z, S_n) = \psi(z - S_n)$. In Hampel (1971) these estimators are already taken as examples for qualitatively robust estimators in the i.i.d. case. As we are not requiring additional properties on the estimators then those needed in the i.i.d. case, M-estimators are also qualitatively robust for the non-i.i.d. case. We take a result from Huber (1981): If $S : \mathcal{M}(X) \to \mathcal{X}$ is the operator representing the estimators S_n i. e. S is the solution of $\int \psi(z - S(P))\, dP = 0$, it is shown in Huber (1981, Chapter 3, Theorem 2.6 and Example 2.2) that this operator is continuous for every P as long as ψ is bounded and strictly monotone and if the solution of the "true" distribution P_0 is unique. Examples for suitable functions ψ are the Huber estimators see Hampel (1971) or the Φ-estimator, see Hampel (1968). Therefore, according to Theorem 3.1.3, these estimators are also qualitatively robust in the non-i.i.d. case.

A second example are R-estimators. R estimators are based on a rank test for two independent samples of size m and n and of shifted distributions $F(x)$ and $G(x) = F(x - \Delta)$. The test statistic for a rank test for $\Delta = 0$ versus $\Delta > 0$ is

$$S_{m,n} = \frac{1}{m} \sum_{i=1}^{m} a_i(R_i) \tag{3.19}$$

and is based on the ranks R_i of one sample in the combined sample and on the scores a_i, $i \in \{1, \ldots, m\}$. The scores a_i are determined by a function J, which is $(m+n) \int_{i-1/(m+n)}^{i/(m+n)} J(s)ds$, moreover $\int J(s)ds = 0$.
According to Hampel et al. (1986, Definition 3), an estimator S_n of location can be defined

such that (3.19) is almost zero for the samples $X_1, \ldots, X_n$ and $2S_n - X_1, \ldots, 2S_n - X_n$. Hence the estimator derives from a statistical operator $S(F)$ which is defined implicitly by:

$$\int J\left(\frac{1}{2}[s + 1 - F(2S(F) - F^{-1}(s))]\right) ds = 0. \tag{3.20}$$

According to Huber (1981, Chapter 3, Theorem 4.1) the operator S is continuous at F as long as the function J is monotone increasing, integrable, and symmetric $J(1-t) = J(t)$, and as long as it is uniquely defined by (3.20). Hence the estimate is qualitatively robust due to Hampel's theorem for i.i.d. observations, see Hampel (1968, Example 7(iii)) and due to Theorem 3.1.3 it is also qualitatively robust for Varadarajan processes.

More examples can be found in Hampel (1968, Section 7). Moreover, qualitative robustness for support vector machines, is shown in Chapter 4.2.1, Theorem 4.2.1.

3.4 Qualitative robustness for bootstrap estimators

Often the finite sample distribution of the estimator or of the stochastic process of interest is unknown, hence an approximation of the distribution is needed. Commonly, the bootstrap is used to receive an approximation of the unknown finite sample distribution by resampling from the given sample.

The classical bootstrap, also called the empirical bootstrap, has been introduced by Efron (1979) for i.i.d. random variables. This concept is based on drawing a bootstrap sample $(Z_1^*, \ldots, Z_m^*)$ of size $m \in \mathbb{N}$ with replacement out of the original sample $(Z_1, \ldots, Z_n)$, $n \in \mathbb{N}$, and approximate the theoretical distribution P_n of $(Z_1, \ldots, Z_n)$ using the bootstrap sample. For the empirical bootstrap the approximation of the distribution via the bootstrap is given by the empirical distribution of the bootstrap sample $(Z_1^*, \ldots, Z_m^*)$, hence $P_n^* = \otimes_{i=1}^n \left(\frac{1}{m}\sum_{i=1}^m \delta_{Z_i^*}\right)$. The bootstrap sample itself has distribution $\otimes_{i=1}^m \left(\frac{1}{n}\sum_{i=1}^n \delta_{Z_i}\right)$.

For an introduction to the bootstrap see for example Efron and Tibshirani (1993) and van der Vaart (1998, Chapter 3.6). Besides the empirical bootstrap many other bootstrap methods have been developed in order to find good approximations also for non-i.i.d. observations, see for example Singh (1981), Lahiri (2003), and the references therein. In Section 3.4.2 the moving block bootstrap introduced by Künsch (1989) and Liu and Singh (1992) is used to approximate the distribution of an α-mixing stochastic process.

It is, also in the non-i.i.d. case, still desirable that the estimator is qualitatively robust even for the bootstrap approximation. That is, the distribution of the estimator under the bootstrap approximation $\mathcal{L}_{P_n^*}(S_n)$, $n \in \mathbb{N}$, of the assumed, ideal distribution P_n should still be close to the distribution of the estimator under the bootstrap approximation $\mathcal{L}_{Q_n^*}(S_n)$, $n \in \mathbb{N}$, of the real contaminated distribution Q_n. Remember that this is a random object as P_n^* respectively Q_n^* are random. For notational convenience all bootstrap values are noted as usual with an asterisk. To show qualitative robustness often generalizations of Hampel's theorem are used. Accordingly we try to find results similar to Hampel's theorem for the case of bootstrap approximations. Cuevas and Romo (1993) describes a concept of qualitative robustness of bootstrap approximations for the i.i.d. case and for real valued estimators. Also a generalization of Hampel's theorem to this case is given. In Christmann et al. (2013, 2011) qualitative robustness of Efron's bootstrap approximation is shown for the i.i.d. case for a class of regularized kernel based learning methods, i. e. not necessarily real valued estimators. Moreover Beutner and Zähle (2016) describes consistency of the bootstrap for plug in estimators. In this chapter estimators with values in a complete separable metric space, which can be represented by a continuous statistical operator on the space of all probability measures are considered.

Based on the generalization of Hampel's concept of Π-robustness from Bustos (1980), we define qualitative robustness for bootstrap approximations for non-i.i.d sequences of random variables. The stronger concept of Π-robustness is needed here, similar to Definition 3.1.1 in Chapter 3, as we do not assume to have i.i.d. random variables, which are used in Cuevas and Romo (1993).

Therefore the definition of qualitative robustness stated below is stronger than the definition in Cuevas and Romo (1993), i. e. if we use this definition for the i.i.d. case the assumption $d_{\mathrm{BL}}(P_n, Q_n) = d_{\mathrm{BL}}(\otimes_{i=1}^n P, \otimes_{i=1}^n Q) < \delta$ implies $d_{\mathrm{BL}}(P, Q) < \delta$. This can be seen similar to the proof of Lemma 3.4.4 in Section 3.4.1.
Remember the statistical model from Chapter 3: $(\mathcal{Z}^{\mathbb{N}}, \mathcal{B}^{\otimes\mathbb{N}}, \mathcal{M}(\mathcal{Z}^{\mathbb{N}}))$, where $(\mathcal{Z}, d_{\mathcal{Z}})$ is a complete separable metric space and $(Z_i)_{i\in\mathbb{N}}$ is the coordinate process on $\mathcal{Z}^{\mathbb{N}}$. $(S_n)_{n\in\mathbb{N}}$ is a sequence of estimators on the stochastic process $(Z_i)_{i\in\mathbb{N}}$. The estimator may take its values in any complete separable metric space H; that is, $S_n : \mathcal{Z}^n \to H$ for every $n \in \mathbb{N}$. Moreover let $P_{\mathbb{N}}^*$ be the approximation of $P_{\mathbb{N}}$ with respect to the bootstrap. Define the bootstrap sample $(Z_1^*, \ldots, Z_n^*)$ as the first n coordinate projections $Z_i^* : \mathcal{Z}^{\mathbb{N}} \to \mathcal{Z}$, where the law of the stochastic process $(Z_i^*)_{i\in\mathbb{N}}$ has to be chosen according to the bootstrap procedure. For the empirical bootstrap, for example, the bootstrap sample is chosen via drawing with replacement from the given observations $z_1, \ldots, z_\ell$, $\ell \in \mathbb{N}$. Hence the

distribution of the bootstrap sample is $\otimes_{n\in\mathbb{N}} \frac{1}{\ell}\sum_{i=1}^{\ell}\delta_{z_i}$, with finite sample distributions $\otimes_{j=1}^{n}\frac{1}{\ell}\sum_{i=1}^{\ell}\delta_{z_i} = (Z_1^*,\dots,Z_n^*)\left(\otimes_{n\in\mathbb{N}}\frac{1}{\ell}\sum_{i=1}^{\ell}\delta_{z_i}\right)$.
Contrarily to the classical case of qualitative robustness the distribution of the estimator under P_n^*, $\mathcal{L}_{P_n^*}(S_n)$ is a random probability measure, as the distribution $P_n^* = \otimes_{i=1}^{n}\frac{1}{\ell}\sum_{i=1}^{\ell}\delta_{Z_i^*}$, $Z_i^* : \mathcal{Z}^{\mathbb{N}} \to \mathcal{Z}$, is random. Hence the mapping $z_{\mathbb{N}} \mapsto \mathcal{L}_{P_n^*}(S_n)$, $z_{\mathbb{N}} \in \mathcal{Z}^{\mathbb{N}}$, is itself a random variable with values in $\mathcal{M}(H)$, i. e. on the space of probability measures on H, equipped with the weak topology on $\mathcal{M}(H)$. The measurability of this mapping is ensured by Beutner and Zähle (2016, Lemma D1).

Contrarily to the original definitions of qualitative robustness in Bustos (1980) the bounded Lipschitz metric d_{BL} is used instead of the Prohorov metric π for the definition of qualitative robustness of the bootstrap approximation below. This is equivalent to Cuevas and Romo (1993). Let $\mathcal{X}$ be a separable metric space, then the bounded Lipschitz metric on the space of probability measures $\mathcal{M}(\mathcal{X})$ on $\mathcal{X}$ is defined by:

$$d_{\mathrm{BL}}(P,Q) := \sup\left\{\left|\int f\,dP - \int f\,dQ\right|;\ f\in \mathrm{BL}(\mathcal{X}), \|f\|_{\mathrm{BL}} \le 1\right\}$$

where $\|\cdot\|_{\mathrm{BL}} := |\cdot|_1 + \|\cdot\|_\infty$ denotes the bounded Lipschitz norm with $|f|_1 = \sup_{x\neq y}\frac{|f(x)-f(y)|}{d(x,y)}$ and $\|\cdot\|_\infty$ the supremum norm $\|f\|_\infty := \sup_x |f(x)|$. This is due to technical reasons only. Both metrics metricize the weak topology on the space of all probability measures $\mathcal{M}(\mathcal{X})$, for Polish spaces $\mathcal{X}$, see, for example, Huber (1981, Chapter 2, Corollary 4.3) or Dudley (1989, Theorem 11.3.3), and therefore can be replaced while adapting δ on the left hand-side of implication (3.21). If $\mathcal{X}$ is a Polish space, so is $\mathcal{M}(\mathcal{X})$ with respect to the weak topology, see Huber (1981, Chapter 2, Theorem 3.9). Hence the bounded Lipschitz metric on the right-hand side of implication (3.21) operates on a space of probability measures on the Polish space $\mathcal{M}(\mathcal{X})$. Therefore the Prohorov metric and the bounded Lipschitz metric are again strongly equivalent and can be replaced while adapting ε in (3.21). Similar to Cuevas and Romo (1993) the proof of the theorems below rely on the fact that the set of bounded Lipschitz functions BL is a uniform Glivenko-Cantelli class (see Definition A3), which implies uniform convergence of the bounded Lipschitz metric of the empirical measure to a limiting distribution, see Dudley et al. (1991). Therefore the definition is given with respect to the bounded Lipschitz metric.

Definition 3.4.1 (Qualitative robustness for bootstrap approximations)
Let $\mathcal{Z}$, H be complete separable metric spaces. Let $P_{\mathbb{N}} \in \mathcal{M}(\mathcal{Z}^{\mathbb{N}})$ and let $P_{\mathbb{N}}^* \in \mathcal{M}(\mathcal{Z}^{\mathbb{N}})$ be the bootstrap approximation of $P_{\mathbb{N}}$. Let $\mathcal{P} \subset \mathcal{M}(\mathcal{Z}^{\mathbb{N}})$ with $P_{\mathbb{N}} \in \mathcal{P}$. Let $S_n : \mathcal{Z}^n \to H$, $n \in \mathbb{N}$,

be a sequence of estimators. Then the sequence of bootstrap approximations $(\mathcal{L}_{P_n^*}(S_n))_{n\in\mathbb{N}}$ *is called qualitatively robust at* $P_{\mathbb{N}}$ *with respect to* $\mathcal{P}$ *if, for every* $\varepsilon > 0$*, there is* $\delta > 0$ *such that there is* $n_0 \in \mathbb{N}$ *such that for every* $n \geq n_0$ *and for every* $Q_{\mathbb{N}} \in \mathcal{P}$*,*

$$d_{\mathrm{BL}}(P_n, Q_n) < \delta \;\Rightarrow\; d_{\mathrm{BL}}(\mathcal{L}(\mathcal{L}_{P_n^*}(S_n)), \mathcal{L}(\mathcal{L}_{Q_n^*}(S_n))) < \varepsilon. \tag{3.21}$$

Here $\mathcal{L}(\mathcal{L}_{P_n^*}(S_n))$ *(respectively* $\mathcal{L}(\mathcal{L}_{Q_n^*}(S_n))$*) denotes the distribution of the bootstrap approximation of the estimator* S_n *under* P_n^* *(respectively* Q_n^**).*

This definition of qualitative robustness with respect to the subset $\mathcal{P}$ indicates that we do not show (3.21) for arbitrary probability measures $Q_{\mathbb{N}} \in \mathcal{M}(\mathcal{Z}^{\mathbb{N}})$. All of our results require the contaminated process to at least have the same structure as the ideal process. This is due to the use of the bootstrap procedure. The empirical bootstrap, which is used below, only works well for a few processes, see for example Lahiri (2003), hence the assumptions on the contaminated process are necessary. To our best knowledge there are no results concerning qualitative robustness of the bootstrap approximation for general stochastic processes without any assumptions on the second process and it is probably very hard to show this for every $Q_{\mathbb{N}} \in \mathcal{M}(\mathcal{Z}^{\mathbb{N}})$, respectively $\mathcal{P} = \mathcal{M}(\mathcal{Z}^{\mathbb{N}})$. Another difference to Definition 3.1.1 is the restriction to $n \geq n_0$. As the results for the bootstrap are asymptotic results, we can not achieve the equicontinuity for every $n \in \mathbb{N}$, but only asymptotically.

The next two sections establish results about qualitative robustness of the bootstrap approximation. First we examine stochastic processes with independent but not necessarily identically distributed random variables, the second kind of stochastic processes are α-mixing processes.

3.4.1 Qualitative robustness for independent not necessarily identically distributed stochastic processes

In this section we relax the i.i.d. assumption in view of the identical distribution. We assume the random variables Z_i, $i \in \mathbb{N}$, to be independent, but not necessarily identically distributed.

The result below generalizes Christmann et al. (2013, Theorem 3) and Christmann et al. (2011), as the assumptions on the stochastic process are weaker as well as those on the statistical operator. Compared to Theorem 3 in Cuevas and Romo (1993), which shows qualitative robustness of the sequence of bootstrap estimators with values in $\mathbb{R}$, we have

to strengthen the assumptions on the sample space, but do not need the estimator to be uniformly continuous. But keep in mind, that the assumption $d_{\mathrm{BL}}(P_n, Q_n) < \delta$ implies $d_{\mathrm{BL}}(P, Q) < \delta$, which is used for the i.i.d. case, in Christmann et al. (2013) and Cuevas and Romo (1993).

Theorem 3.4.2 *Let the sequence of estimators $(S_n)_{n\in\mathbb{N}}$ be represented by a statistical operator $S : \mathcal{M}(\mathcal{Z}) \to H$ via (3.1) for a complete separable metric space (H, d_H) and let $(\mathcal{Z}, d_{\mathcal{Z}})$ be a totally bounded metric space.*
Let $P_{\mathbb{N}} = \otimes_{i\in\mathbb{N}} P^i$, $P^i \in \mathcal{M}(\mathcal{Z})$ be an infinite product measure such that the coordinate process $(Z_i)_{i\in\mathbb{N}}$, $Z_i \colon \mathcal{Z}^{\mathbb{N}} \to z_i$, $i \in \mathbb{N}$, is a strong Varadarajan process with limiting distribution P. Moreover define $\mathcal{P} := \left\{Q_{\mathbb{N}} \in \mathcal{M}(\mathcal{Z}^{\mathbb{N}});\ Q_{\mathbb{N}} = \otimes_{i\in\mathbb{N}} Q^i,\ Q^i \in \mathcal{M}(\mathcal{Z})\right\}$. Let $S : \mathcal{M}(\mathcal{Z}) \to H$ be continuous at P with respect to d_{BL} and let the estimators $S_n : \mathcal{Z}^n \to H$, $n \in \mathbb{N}$, be continuous.
Then the sequence of bootstrap approximations $(\mathcal{L}_{P_n^*}(S_n))_{n\in\mathbb{N}}$, is qualitatively robust at $P_{\mathbb{N}}$ with respect to $\mathcal{P}$.

Remark 3.4.3 *The required properties on the statistical operator S and on the sequence of estimators $(S_n)_{n\in\mathbb{N}}$ in Theorem 3.4.2 ensure the qualitative robustness of $(S_n)_{n\in\mathbb{N}}$, as long as the assumptions on the underlying stochastic processes are fulfilled.*
The proof shows that the bootstrap approximation of every sequence of estimators $(S_n)_{n\in\mathbb{N}}$ which is qualitatively robust in the sense of the definitions in Bustos (1980) and Definition 3.1.1 is qualitatively robust in the sense of Theorem 3.4.2.

All estimators $(S_n)_{n\in\mathbb{N}}$ which are mentioned in Section 3.3 and support vector machines, see Theorem 4.2.1, are included. Hence Hampel's theorem for the i.i.d. case can be generalized to bootstrap approximations and to the case of not necessarily identically distributed random variables if qualitative robustness is based on the definition of Π-robustness.

Unfortunately, the assumption on the space $(\mathcal{Z}, d_{\mathcal{Z}})$ to be totally bounded seems to be necessary. In the proof of Theorem 3.4.2 we use a result of Dudley et al. (1991) to show uniformity on the space of probability measures $\mathcal{M}(\mathcal{Z})$. This result needs the bounded Lipschitz functions to be a uniform Glivenko-Cantelli class, which is equivalent to $(\mathcal{Z}, d_{\mathcal{Z}})$ being totally bounded, see Dudley et al. (1991, Proposition 12). In order to weaken the assumption on $(\mathcal{Z}, d_{\mathcal{Z}})$, probably another way to show uniformity on the space of probability measures $\mathcal{M}(\mathcal{Z})$ has to be found.

Before proving Theorem 3.4.1, we state a rather technical lemma, connecting the product measure $\otimes_{i=1}^n P^i \in \mathcal{M}(\mathcal{Z}^n)$ of independent random variables to their mixture measure $\frac{1}{n}\sum_{i=1}^n P^i \in \mathcal{M}(\mathcal{Z})$. Let $(\mathcal{Z}, d_{\mathcal{Z}})$ be a Polish space.

Lemma 3.4.4 *Let $P_n, Q_n \in \mathcal{M}(\mathcal{Z}^n)$ such that $P_n = \otimes_{i=1}^n P^i$ and $Q_n = \otimes_{i=1}^n Q^i$, $P^i, Q^i \in \mathcal{M}(\mathcal{Z})$, $i \in \mathbb{N}$. Then for all $\delta > 0$:*

$$d_{\mathrm{BL}}(P_n, Q_n) \leq \delta \quad \Rightarrow \quad d_{\mathrm{BL}}\left(\frac{1}{n}\sum_{i=1}^n P^i, \frac{1}{n}\sum_{i=1}^n Q^i\right) \leq \delta.$$

Proof: By assumption we have $d_{\mathrm{BL}}(P_n, Q_n) \leq \delta$. Moreover for a function $f : \mathcal{Z} \to \mathbb{R}$:

$$\int_{\mathcal{Z}} f(z_i)\, dP^i(z_i) = \int_{\mathcal{Z}^{n-1}} \int_{\mathcal{Z}} f(z_i)\, dP^i(z_i)\, d\left(\otimes_{j\neq i} P^j(z_j)\right). \tag{3.22}$$

Then,

$$\begin{aligned}
&\sup_{f\in \mathrm{BL}_1(\mathcal{Z})} \left| \int_{\mathcal{Z}} f(z_i)\, d\left[\frac{1}{n}\sum_{i=1}^n P^i(z_i)\right] - \int_{\mathcal{Z}} f(z_i)\, d\left[\frac{1}{n}\sum_{i=1}^n Q^i(z_i)\right] \right| \\
&\quad = \sup_{f\in \mathrm{BL}_1(\mathcal{Z})} \left| \frac{1}{n}\sum_{i=1}^n \left[\int_{\mathcal{Z}} f(z_i)\, dP^i(z_i) - \int_{\mathcal{Z}} f(z_i)\, dQ^i(z_i) \right] \right| \\
&\quad \overset{(3.22)}{=} \sup_{f\in \mathrm{BL}_1(\mathcal{Z})} \left| \frac{1}{n}\sum_{i=1}^n \left[\int_{\mathcal{Z}^{n-1}} \int_{\mathcal{Z}} f(z_i)\, dP^i(z_i)\, d\left(\otimes_{j\neq i} P^j(z_j)\right) \right.\right. \\
&\qquad \left.\left. - \int_{\mathcal{Z}^{n-1}} \int_{\mathcal{Z}} f(z_i)\, dQ^i(z_i)\, d\left(\otimes_{j\neq i} Q^j(z_j)\right) \right] \right| \\
&\quad = \sup_{f\in \mathrm{BL}_1(\mathcal{Z})} \left| \frac{1}{n}\sum_{i=1}^n \left[\int_{\mathcal{Z}^n} f(z_i)\, d\left(\otimes_{j=1}^n P^j(z_j)\right) - \int_{\mathcal{Z}^n} f(z_i)\, d\left(\otimes_{j=1}^n Q^j(z_j)\right) \right] \right| \\
&\quad \leq \frac{1}{n}\sum_{i=1}^n \sup_{f\in \mathrm{BL}_1(\mathcal{Z})} \left| \int_{\mathcal{Z}^n} f(z_i)\, d\left(\otimes_{j=1}^n P^j(z_j)\right) - \int_{\mathcal{Z}^n} f(z_i)\, d\left(\otimes_{j=1}^n Q^j(z_j)\right) \right|.
\end{aligned}$$

Now every function $f \in \mathrm{BL}_1(\mathcal{Z})$ can be identified as a function $\tilde{f} : \mathcal{Z}^n \to \mathcal{Z}$, $(z_1, \ldots, z_n) \mapsto \tilde{f}(z_1, \ldots, z_n) := f(z_i)$. This function is also Lipschitz continuous on $\mathcal{Z}^n$:

$$\begin{aligned}
|\tilde{f}(z_1, \ldots, z_n) - \tilde{f}(z_1', \ldots, z_n')| &= |f(z_i) - f(z_i')| \\
&\leq |f|_1 d(z_i, z_i') \leq |f|_1 (d_{\mathcal{Z}}(z_1, z_1') + \ldots + d_{\mathcal{Z}}(z_i, z_i') + \ldots + d_{\mathcal{Z}}(z_n, z_n')),
\end{aligned}$$

where $d_{\mathcal{Z}}(z_1, z_1') + \ldots + d_{\mathcal{Z}}(z_i, z_i') + \ldots + d_{\mathcal{Z}}(z_n, z_n')$ induces the product topology on $\mathcal{Z}^n$. That is $\tilde{f} \in \mathrm{BL}_1(\mathcal{Z}^n)$. Note that this is also true for every p-product metric $d_{n,p}$ in $\mathcal{Z}^n$, $1 \leq p \leq \infty$, as they are strongly equivalent. Hence,

$$\begin{aligned} d_{\mathrm{BL}}\left(\frac{1}{n}\sum_{i=1}^n P^i, \frac{1}{n}\sum_{i=1}^n Q^i\right) &\leq \frac{1}{n}\sum_{i=1}^n \sup_{g\in \mathrm{BL}_1(\mathcal{Z}^n)} \left|\int_{\mathcal{Z}^n} g\, dP_n - \int_{\mathcal{Z}^n} g\, dQ_n\right| \\ &\leq \frac{1}{n}\sum_{i=1}^n d_{\mathrm{BL}}\left(P_n, Q_n\right) \leq \delta, \end{aligned}$$

which yields the assertion. □

Proof of Theorem 3.4.2: To prove Theorem 3.4.2 we first use the triangle inequality to split the bounded Lipschitz distance between the distribution of the estimator S_n, $n \in \mathbb{N}$, into two parts regarding the distribution of the estimator under the joint distribution P_n of $(Z_1, \ldots, Z_n)$:

$$d_{\mathrm{BL}}(\mathcal{L}_{P_n^*}(S_n), \mathcal{L}_{Q_n^*}(S_n)) \leq \underbrace{d_{\mathrm{BL}}(\mathcal{L}_{P_n^*}(S_n), \mathcal{L}_{P_n}(S_n))}_{I} + \underbrace{d_{\mathrm{BL}}(\mathcal{L}_{P_n}(S_n), \mathcal{L}_{Q_n^*}(S_n))}_{II}.$$

Then the representation of the estimator S_n by the statistical operator S and the continuity of this operator in P together with the Varadarajan property and the independence assumption on the stochastic process yield the assertion.

First we regard part I: Define the distribution $P_{\mathbb{N}} \in \mathcal{M}(\mathcal{Z}^{\mathbb{N}})$ and let $P_{\mathbb{N}}^*$ be the bootstrap approximation of $P_{\mathbb{N}}$. Define, for $n \in \mathbb{N}$, the random variables
$\mathbf{W}_n : \mathcal{Z}^{\mathbb{N}} \to \mathcal{Z}^n$, $\mathbf{W}_n = (Z_1, \ldots, Z_n)$, $z_{\mathbb{N}} \mapsto \mathbf{W}_n(z_{\mathbb{N}}) = \mathbf{w}_n = (z_1, \ldots, z_n)$, and
$\mathbf{W}_n' : \mathcal{Z}^{\mathbb{N}} \to \mathcal{Z}^n$, $\mathbf{W}_n' = (Z_1', \ldots, Z_n')$, $z_{\mathbb{N}} \mapsto \mathbf{w}_n'$,
such that $\mathbf{W}_n(P_{\mathbb{N}}) = P_n$ and $\mathbf{W}_n'(P_{\mathbb{N}}^*) = P_n^*$.
Denote the bootstrap sample by $\mathbf{W}_n^* := (Z_1^*, \ldots, Z_n^*)$, $\mathbf{W}_n^* : \mathcal{Z}^{\mathbb{N}} \to \mathcal{Z}^n$, $z_{\mathbb{N}} \mapsto \mathbf{w}_n^*$.
As Efron's empirical bootstrap is used, the bootstrap sample, which is chosen via resampling with replacement out of $Z_1, \ldots, Z_\ell$, $\ell \in \mathbb{N}$, has distribution $Z_i^* \sim \mathbb{P}_{\mathbf{W}_\ell} = \frac{1}{\ell}\sum_{j=1}^\ell \delta_{Z_j}$, $i \in \mathbb{N}$, respectively $\mathbf{W}_n^* := (Z_1^*, \ldots, Z_n^*) \sim \otimes_{i=1}^n \mathbb{P}_{\mathbf{W}_\ell}$. The bootstrap approximation of P_ℓ, $\ell \in \mathbb{N}$, is the empirical measure of the bootstrap sample $P_\ell^* = \otimes_{i=1}^\ell \frac{1}{n}\sum_{j=1}^n \delta_{Z_j^*}$.

Further denote the joint distribution of $\mathbf{W}_{\mathbb{N}}$, $\mathbf{W}_{\mathbb{N}}^*$, and $\mathbf{W}_{\mathbb{N}}'$ by $K_{\mathbb{N}} \in \mathcal{M}(\mathcal{Z}^{\mathbb{N}} \times \mathcal{Z}^{\mathbb{N}} \times \mathcal{Z}^{\mathbb{N}})$. Then, $K_{\mathbb{N}}$ has marginal distributions $K_{\mathbb{N}}(B_1 \times \mathcal{Z}^{\mathbb{N}} \times \mathcal{Z}^{\mathbb{N}}) = P_{\mathbb{N}}(B_1)$ for all $B_1 \in \mathcal{B}^{\otimes\mathbb{N}}$,

$K_{\mathbb{N}}(\mathcal{Z}^{\mathbb{N}} \times B_2 \times \mathcal{Z}^{\mathbb{N}}) = \otimes_{i\in\mathbb{N}} \mathbb{P}_{\mathbf{W}_n}(B_2)$ for all $B_2 \in \mathcal{B}^{\otimes\mathbb{N}}$, and $K_{\mathbb{N}}(\mathcal{Z}^{\mathbb{N}} \times \mathcal{Z}^{\mathbb{N}} \times B_3) = P^*_{\mathbb{N}}(B_3)$ for all $B_3 \in \mathcal{B}^{\otimes\mathbb{N}}$.

Then,

$$\mathcal{L}_{P_n}(S_n) = S_n(P_n) = S_n \circ \mathbf{W}_n(P_{\mathbb{N}}) \quad \text{and} \quad \mathcal{L}_{P^*_n}(S_n) = S_n(P^*_n) = S_n \circ \mathbf{W}'_n(P^*_{\mathbb{N}})$$

and therefore

$$d_{\mathrm{BL}}(\mathcal{L}_{P^*_n}(S_n), \mathcal{L}_{P_n}(S_n)) = d_{\mathrm{BL}}(\mathcal{L}(S_n \circ W'_n), \mathcal{L}(S_n \circ W_n)).$$

By assumption the coordinate process $(Z_i)_{i\in\mathbb{N}}$ consists of independent random variables, hence we have $P_n = \otimes_{i=1}^n P^i$, for $P^i = Z_i(P_{\mathbb{N}})$, $i \in \mathbb{N}$.

Moreover $(\mathcal{Z}, d_{\mathcal{Z}})$ is assumed to be a totally bounded metric space. Then, due to Dudley et al. (1991, Proposition 12), the set $\mathrm{BL}_1(\mathcal{Z}, d_{\mathcal{Z}})$ is a uniform Glivenko-Cantelli class (see Definition A3). That is, if $Z_i \sim P$ i.i.d. $i \in \mathbb{N}$, we have for all $\eta > 0$:

$$\lim_{n\to\infty} \sup_{P\in\mathcal{M}(\mathcal{Z})} P_{\mathbb{N}}\left(\left\{z_{\mathbb{N}} \in \mathcal{Z}^{\mathbb{N}} \mid \sup_{m\geq n} d_{\mathrm{BL}}(\mathbb{P}_{\mathbf{W}_m(z_{\mathbb{N}})}, P) > \eta\right\}\right) = 0.$$

Applying this to the bootstrap sample $(Z^*_1, \ldots, Z^*_m)$, $m \in \mathbb{N}$, which is found by resampling with replacement out of the original sample $(Z_1, \ldots, Z_n)$, we have, for all $\mathbf{w}_n \in \mathcal{Z}^n$,

$$\lim_{n\to\infty} \sup_{\mathbb{P}_{\mathbf{w}_n}\in\mathcal{M}(\mathcal{Z})} \otimes_{i\in\mathbb{N}} \mathbb{P}_{\mathbf{w}_n}\left(\left\{z_{\mathbb{N}} \in \mathcal{Z}^{\mathbb{N}} \mid \sup_{m\geq n} d_{\mathrm{BL}}(\mathbb{P}_{\mathbf{W}^*_m(z_{\mathbb{N}})}, \mathbb{P}_{\mathbf{w}_n}) > \eta\right\}\right) = 0.$$

Let $\varepsilon > 0$ be arbitrary but fixed. Then, for every $\delta_0 > 0$ there is $n_1 \in \mathbb{N}$ such that for all $n \geq n_1$ and all $\mathbb{P}_{\mathbf{w}_n} \in \mathcal{M}(\mathcal{Z})$:

$$\otimes_{i=1}^n \mathbb{P}_{\mathbf{w}_n}\left(\left\{\mathbf{w}^*_n \in \mathcal{Z}^n \mid d_{\mathrm{BL}}(\mathbb{P}_{\mathbf{w}^*_n}, \mathbb{P}_{\mathbf{w}_n}) \leq \frac{\delta_0}{4}\right\}\right) \geq 1 - \frac{\varepsilon}{8}. \tag{3.23}$$

And, using the same argumentation for the sequence of random variables Z'_i, $i \in \mathbb{N}$, which are i.i.d. and have distribution $\frac{1}{n}\sum_{i=1}^n \delta_{Z^*_i} = \mathbb{P}_{\mathbf{W}^*_n}$:

$$\lim_{n\to\infty} \sup_{\mathbb{P}_{\mathbf{w}^*_n}\in\mathcal{M}(\mathcal{Z})} P^*_{\mathbb{N}}\left(\left\{z_{\mathbb{N}} \in \mathcal{Z}^{\mathbb{N}} \mid \sup_{m\geq n} d_{\mathrm{BL}}(\mathbb{P}_{\mathbf{W}'_m(z_{\mathbb{N}})}, \mathbb{P}_{\mathbf{w}^*_n}) > \eta\right\}\right) = 0.$$

Respectively, for every $\delta_0 > 0$ there is $n_2 \in \mathbb{N}$ such that for all $n \geq n_2$ and all $\mathbb{P}_{\mathbf{w}_n^*} \in \mathcal{M}(\mathcal{Z})$:

$$P_n^* \left(\left\{ \mathbf{w}_n' \in \mathcal{Z}^n \mid d_{\mathrm{BL}}(\mathbb{P}_{\mathbf{w}_n'}, \mathbb{P}_{\mathbf{w}_n^*}) \leq \frac{\delta_0}{2} \right\} \right) \geq 1 - \frac{\varepsilon}{8}. \tag{3.24}$$

As the process $(Z_i)_{i \in \mathbb{N}}$ is a strong Varadarajan process by assumption, there exists a probability measure $P \in \mathcal{M}(\mathcal{Z})$ such that

$$d_{\mathrm{BL}}(\mathbb{P}_{\mathbf{W}_n}, P) \longrightarrow 0 \text{ almost surely with respect to } P_{\mathbb{N}},\ n \to \infty.$$

That is, for every $\delta_0 > 0$ there is $n_3 \in \mathbb{N}$ such that for all $n \geq n_3$:

$$P_n \left(\left\{ \mathbf{w}_n \in \mathcal{Z}^n \mid d_{\mathrm{BL}}(\mathbb{P}_{\mathbf{w}_n}, P) \leq \frac{\delta_0}{2} \right\} \right) \geq 1 - \frac{\varepsilon}{4}. \tag{3.25}$$

The continuity of the statistical operator $S : \mathcal{M}(\mathcal{Z}) \to H$ in $P \in \mathcal{M}(\mathcal{Z})$ yields: for every $\varepsilon > 0$ there exists $\delta_0 > 0$ such that for all $Q \in \mathcal{M}(\mathcal{Z})$:

$$d_{\mathrm{BL}}(P, Q) \leq \delta_0 \quad \Rightarrow \quad d_H(S(P), S(Q)) \leq \frac{\varepsilon}{4}. \tag{3.26}$$

As the Prohorov metric π_{d_H} is bounded by the Ky Fan metric, see Dudley (1989, Theorem 11.3.5) we conclude:

$$\begin{aligned}
&\pi_{d_H}(\mathcal{L}_{P_n^*}(S_n), \mathcal{L}_{P_n}(S_n)) = \pi_{d_H}(S_n \circ \mathbf{W}_n', S_n \circ \mathbf{W}_n) \\
&\leq \inf \left\{ \tilde{\varepsilon} > 0 \mid K_{\mathbb{N}} \left(\left\{ d_H(S_n \circ \mathbf{W}_n', S_n \circ \mathbf{W}_n) > \tilde{\varepsilon} \right\} \right) \leq \tilde{\varepsilon} \right\} \\
&= \inf \left\{ \tilde{\varepsilon} > 0 \mid (\mathbf{W}_n, \mathbf{W}_n^*, \mathbf{W}_n')(K_{\mathbb{N}}) \left(\left\{ (\mathbf{w}_n, \mathbf{w}_n^*, \mathbf{w}_n') \in \mathcal{Z}^n \times \mathcal{Z}^n \times \mathcal{Z}^n \mid \right. \right. \right. \\
&\qquad \left. \left. \left. d_H(S_n(\mathbf{w}_n'), S_n(\mathbf{w}_n)) > \tilde{\varepsilon}, \mathbf{w}_n^* \in \mathcal{Z}^n \right\} \right) \leq \tilde{\varepsilon} \right\}.
\end{aligned} \tag{3.27}$$

Due to the definition of the statistical operator S, this is equivalent to

$$\begin{aligned}
&\inf \left\{ \tilde{\varepsilon} > 0 \mid (\mathbf{W}_n, \mathbf{W}_n^*, \mathbf{W}_n')(K_{\mathbb{N}}) \left(\left\{ (\mathbf{w}_n, \mathbf{w}_n^*, \mathbf{w}_n') \in \mathcal{Z}^n \times \mathcal{Z}^n \times \mathcal{Z}^n \mid \right. \right. \right. \\
&\qquad \left. \left. \left. d_H(S(\mathbb{P}_{\mathbf{w}_n'}), S(\mathbb{P}_{\mathbf{w}_n})) > \tilde{\varepsilon}, \mathbf{w}_n^* \in \mathcal{Z}^n \right\} \right) \leq \tilde{\varepsilon} \right\}.
\end{aligned}$$

The triangle inequality

$$d_H(S(\mathbb{P}_{\mathbf{w}_n'}), S(\mathbb{P}_{\mathbf{w}_n})) \leq d_H(S(\mathbb{P}_{\mathbf{w}_n'}), S(P)) + d_H(S(P), S(\mathbb{P}_{\mathbf{w}_n})),$$

and the continuity of the statistical operator S, see (3.26), then yield, for all $\varepsilon > 0$,

$$\begin{aligned}
&(\mathbf{W}_n, \mathbf{W}_n^*, \mathbf{W}_n')(K_{\mathbb{N}}) \left(\left\{ (\mathbf{w}_n, \mathbf{w}_n^*, \mathbf{w}_n') \in \mathcal{Z}^n \times \mathcal{Z}^n \times \mathcal{Z}^n \mid d_H(S(\mathbb{P}_{\mathbf{w}_n'}), S(\mathbb{P}_{\mathbf{w}_n})) > \frac{\varepsilon}{2}, \mathbf{w}_n^* \in \mathcal{Z}^n \right\} \right) \\
&\leq \quad (\mathbf{W}_n, \mathbf{W}_n^*, \mathbf{W}_n')(K_{\mathbb{N}}) \left(\left\{ (\mathbf{w}_n, \mathbf{w}_n^*, \mathbf{w}_n') \in \mathcal{Z}^n \times \mathcal{Z}^n \times \mathcal{Z}^n \mid d_H(S(\mathbb{P}_{\mathbf{w}_n'}), S(P)) > \frac{\varepsilon}{4} \right. \right. \\
&\qquad\qquad \left. \left. \text{or } d_H(S(P), S(\mathbb{P}_{\mathbf{w}_n})) > \frac{\varepsilon}{4}, \mathbf{w}_n^* \in \mathcal{Z}^n \right\} \right) \\
&\overset{(3.26)}{\leq} \quad (\mathbf{W}_n, \mathbf{W}_n^*, \mathbf{W}_n')(K_{\mathbb{N}}) \left(\left\{ (\mathbf{w}_n, \mathbf{w}_n^*, \mathbf{w}_n') \in \mathcal{Z}^n \times \mathcal{Z}^n \times \mathcal{Z}^n \mid d_{\mathrm{BL}}(\mathbb{P}_{\mathbf{w}_n'}, P) > \delta_0 \right. \right. \\
&\qquad\qquad \left. \left. \text{or } d_{\mathrm{BL}}(P, \mathbb{P}_{\mathbf{w}_n}) > \delta_0, \mathbf{w}_n^* \in \mathcal{Z}^n \right\} \right).
\end{aligned}$$

Using the triangle inequality,

$$d_{\mathrm{BL}}(\mathbb{P}_{\mathbf{w}_n'}, P) \leq d_{\mathrm{BL}}(\mathbb{P}_{\mathbf{w}_n'}, \mathbb{P}_{\mathbf{w}_n^*}) + d_{\mathrm{BL}}(\mathbb{P}_{\mathbf{w}_n^*}, P) \tag{3.28}$$

$$\text{and} \qquad d_{\mathrm{BL}}(\mathbb{P}_{\mathbf{w}_n^*}, P) \leq d_{\mathrm{BL}}(\mathbb{P}_{\mathbf{w}_n^*}, \mathbb{P}_{\mathbf{w}_n}) + d_{\mathrm{BL}}(\mathbb{P}_{\mathbf{w}_n}, P), \tag{3.29}$$

gives for all $n \geq \max\{n_1, n_2, n_3\}$:

$$\begin{aligned}
&(\mathbf{W}_n, \mathbf{W}_n^*, \mathbf{W}_n')(K_{\mathbb{N}}) \left(\left\{ (\mathbf{w}_n, \mathbf{w}_n^*, \mathbf{w}_n') \in \mathcal{Z}^n \times \mathcal{Z}^n \times \mathcal{Z}^n \mid d_{\mathrm{BL}}(\mathbb{P}_{\mathbf{w}_n'}, P) > \delta_0 \right. \right. \\
&\qquad\qquad \left. \left. \text{or } d_{\mathrm{BL}}(P, \mathbb{P}_{\mathbf{w}_n}) > \delta_0, \mathbf{w}_n^* \in \mathcal{Z}^n \right\} \right) \\
&\overset{(3.28)}{\leq} (\mathbf{W}_n, \mathbf{W}_n^*, \mathbf{W}_n')(K_{\mathbb{N}}) \left(\left\{ (\mathbf{w}_n, \mathbf{w}_n^*, \mathbf{w}_n') \in \mathcal{Z}^n \times \mathcal{Z}^n \times \mathcal{Z}^n \mid d_{\mathrm{BL}}(\mathbb{P}_{\mathbf{w}_n'}, \mathbb{P}_{\mathbf{w}_n^*}) > \frac{\delta_0}{2} \right. \right. \\
&\qquad\qquad \left. \left. \text{or } d_{\mathrm{BL}}(\mathbb{P}_{\mathbf{w}_n^*}, P) > \frac{\delta_0}{2} \text{ or } d_{\mathrm{BL}}(P, \mathbb{P}_{\mathbf{w}_n}) > \delta_0 \right\} \right) \\
&\overset{(3.29)}{\leq} (\mathbf{W}_n, \mathbf{W}_n^*, \mathbf{W}_n')(K_{\mathbb{N}}) \left(\left\{ (\mathbf{w}_n, \mathbf{w}_n^*, \mathbf{w}_n') \in \mathcal{Z}^n \times \mathcal{Z}^n \times \mathcal{Z}^n \mid d_{\mathrm{BL}}(\mathbb{P}_{\mathbf{w}_n'}, \mathbb{P}_{\mathbf{w}_n^*}) > \frac{\delta_0}{2} \right. \right. \\
&\qquad\qquad \left. \left. \text{or } d_{\mathrm{BL}}(\mathbb{P}_{\mathbf{w}_n^*}, \mathbb{P}_{\mathbf{w}_n}) > \frac{\delta_0}{4} \text{ or } d_{\mathrm{BL}}(P, \mathbb{P}_{\mathbf{w}_n}) > \frac{\delta_0}{4} \right\} \right) \\
&\leq P_n^* \left(\left\{ \mathbf{w}_n' \in \mathcal{Z}^n \mid d_{\mathrm{BL}}(\mathbb{P}_{\mathbf{w}_n'}, \mathbb{P}_{\mathbf{w}_n^*}) > \frac{\delta_0}{2} \right\} \right) + P_n \left(\left\{ \mathbf{w}_n \in \mathcal{Z}^n \mid d_{\mathrm{BL}}(\mathbb{P}_{\mathbf{w}_n}, P) > \frac{\delta_0}{4} \right\} \right) \\
&\qquad\qquad + \otimes_{i=1}^n \mathbb{P}_{\mathbf{w}_n} \left(\left\{ \mathbf{w}_n^* \in \mathcal{Z}^n \mid d_{\mathrm{BL}}(\mathbb{P}_{\mathbf{w}_n^*}, \mathbb{P}_{\mathbf{w}_n}) > \frac{\delta_0}{4} \right\} \right) \\
&\overset{(3.23),(3.24),(3.25)}{<} \frac{\varepsilon}{8} + \frac{\varepsilon}{4} + \frac{\varepsilon}{8} \quad = \quad \frac{\varepsilon}{2}.
\end{aligned}$$

Hence, for all $\varepsilon > 0$ there are $n_1, n_2, n_3 \in \mathbb{N}$ such that vor all $n \geq \max\{n_1, n_2, n_3\}$, the infimum in (3.27) is bounded by $\frac{\varepsilon}{2}$. Therefore

$$\pi_{d_H}(\mathcal{L}_{P_n^*}(S_n), \mathcal{L}_{P_n}(S_n)) < \frac{\varepsilon}{2}.$$

The equivalence between the Prohorov metric and the bounded Lipschitz metric for Polish spaces, see Huber (1981, Chapter 2, Corollary 4.3), yields the existence of $n_{0,1} \in \mathbb{N}$ such that for all $n \geq n_{0,1}$:

$$d_{\mathrm{BL}}(\mathcal{L}_{P_n^*}(S_n), \mathcal{L}_{P_n}(S_n)) < \frac{\varepsilon}{2}. \tag{3.30}$$

To prove the convergence of the term in part II, consider the distribution $Q_{\mathbb{N}} \in \mathcal{M}(\mathcal{Z}^{\mathbb{N}})$ and let $Q_{\mathbb{N}}^*$ be the bootstrap approximation of $Q_{\mathbb{N}}$. Define, for $n \in \mathbb{N}$, the random variables
$\tilde{\mathbf{W}}_n : \mathcal{Z}^{\mathbb{N}} \to \mathcal{Z}^n$, $\tilde{\mathbf{W}}_n = (\tilde{Z}_1, \ldots, \tilde{Z}_n)$, $z_{\mathbb{N}} \mapsto \tilde{\mathbf{w}}_n$ with distribution $\tilde{\mathbf{W}}_n(Q_{\mathbb{N}}) = Q_n$,
$\tilde{\mathbf{W}}'_n : \mathcal{Z}^{\mathbb{N}} \to \mathcal{Z}^n$, $\tilde{\mathbf{W}}'_n = (\tilde{Z}'_1, \ldots, \tilde{Z}'_n)$, $z_{\mathbb{N}} \mapsto \tilde{\mathbf{w}}'_n$, with distribution $\tilde{\mathbf{W}}'_n(Q_{\mathbb{N}}^*) = Q_n^*$, and
the bootstrap sample $\tilde{\mathbf{W}}_n^* : \mathcal{Z}^{\mathbb{N}} \to \mathcal{Z}^n$, $\tilde{\mathbf{W}}_n^* = (\tilde{Z}_1^*, \ldots, \tilde{Z}_n^*)$, $z_{\mathbb{N}} \mapsto \tilde{\mathbf{w}}_n^*$, with distribution $\otimes_{i=1}^n \mathbb{Q}_{\tilde{\mathbf{W}}_\ell} = \otimes_{i=1}^n \frac{1}{\ell} \sum_{i=1}^{\ell} \delta_{\tilde{Z}_i}$.

Moreover let $\tilde{K}_{\mathbb{N}} \in \mathcal{M}(\mathcal{Z}^{\mathbb{N}} \times \mathcal{Z}^{\mathbb{N}} \times \mathcal{Z}^{\mathbb{N}} \times \mathcal{Z}^{\mathbb{N}})$ denote the joint distribution of $\mathbf{W}_{\mathbb{N}}$, $\tilde{\mathbf{W}}_{\mathbb{N}}$, $\tilde{\mathbf{W}}_{\mathbb{N}}^*$, and $\tilde{\mathbf{W}}'_{\mathbb{N}}$. Then, $\tilde{K}_{\mathbb{N}} \in \mathcal{M}(\mathcal{Z}^{\mathbb{N}} \times \mathcal{Z}^{\mathbb{N}} \times \mathcal{Z}^{\mathbb{N}} \times \mathcal{Z}^{\mathbb{N}})$ has marginal distributions $P_{\mathbb{N}}$, $Q_{\mathbb{N}}$, $\otimes_{i \in \mathbb{N}} \mathbb{Q}_{\tilde{\mathbf{W}}_n}$, and $Q_{\mathbb{N}}^*$.

First, similar to the argumentation for part I, Efron's bootstrap and Dudley et al. (1991, Proposition 12) give for $\tilde{\mathbf{w}}_n \in \mathcal{Z}^n$:

$$\lim_{n\to\infty} \sup_{\mathbb{Q}_{\tilde{\mathbf{w}}_n} \in \mathcal{M}(\mathcal{Z})} \otimes_{n \in \mathbb{N}} \mathbb{Q}_{\tilde{\mathbf{w}}_n} \left(\left\{ z_{\mathbb{N}} \in \mathcal{Z}^{\mathbb{N}} \mid \sup_{m \geq n} d_{\mathrm{BL}}(\mathbb{Q}_{\tilde{\mathbf{W}}_m^*(z_{\mathbb{N}})}, \mathbb{Q}_{\tilde{\mathbf{w}}_n}) > \eta \right\} \right) = 0.$$

Hence, for arbitrary, but fixed $\varepsilon > 0$, for every $\delta_0 > 0$ there is $n_4 \in \mathbb{N}$ such that for all $n \geq n_4$ and all $\mathbb{Q}_{\tilde{\mathbf{w}}_n} \in \mathcal{M}(\mathcal{Z})$:

$$\otimes_{i=1}^n \mathbb{Q}_{\tilde{\mathbf{w}}_n} \left(\left\{ \tilde{\mathbf{w}}_n^* \in \mathcal{Z}^n \mid d_{\mathrm{BL}}(\mathbb{Q}_{\tilde{\mathbf{w}}_n^*}, \mathbb{Q}_{\tilde{\mathbf{w}}_n}) \leq \frac{\delta_0}{6} \right\} \right) \geq 1 - \frac{\varepsilon}{10}. \tag{3.31}$$

Further,

$$\lim_{n\to\infty} \sup_{\mathbb{Q}_{\tilde{\mathbf{w}}_n^*} \in \mathcal{M}(\mathcal{Z})} Q_{\mathbb{N}}^* \left(\left\{ z_{\mathbb{N}} \in \mathcal{Z}^{\mathbb{N}} \mid \sup_{m \geq n} d_{\mathrm{BL}}(\mathbb{Q}_{\tilde{\mathbf{W}}'_m(z_{\mathbb{N}})}, \mathbb{Q}_{\tilde{\mathbf{w}}_n^*}) > \eta \right\} \right) = 0.$$

Respectively, for every $\delta_0 > 0$ there is $n_5 \in \mathbb{N}$ such that for all $n \geq n_5$ and all $\mathbb{Q}_{\tilde{\mathbf{w}}_n^*} = \frac{1}{n}\sum_{i=1}^n \delta_{\tilde{z}_i^*} \in \mathcal{M}(\mathcal{Z})$:

$$Q_n^*\left(\left\{\tilde{\mathbf{w}}_n' \in \mathcal{Z}^n \mid d_{\mathrm{BL}}(\mathbb{Q}_{\tilde{\mathbf{w}}_n'}, \mathbb{Q}_{\tilde{\mathbf{w}}_n^*}) \leq \frac{\delta_0}{6}\right\}\right) \geq 1 - \frac{\varepsilon}{10}. \tag{3.32}$$

Moreover, as the random variables Z_i, $Z_i \sim P^i$, $i \in \mathbb{N}$, are independent, the bounded Lipschitz distance between the empirical measure and $\frac{1}{n}\sum_{i=1}^n P^i$ can be bounded, due to Dudley et al. (1991, Theorem 7). As totally bounded spaces are particularly separable, see Denkowski et al. (2003, below Corollary 1.4.28), Dudley et al. (1991, Proposition 12) provides that $\mathrm{BL}_1(\mathcal{Z}, d_{\mathcal{Z}})$ is a uniform Glivenko-Cantelli class. The proof of this proposition does not depend on the distributions of the random variables Z_i, $i \in \mathbb{N}$, and is therefore also valid for independent and not necessarily identically distributed random variables. Hence Dudley et al. (1991, Theorem 7) yields for all $\eta > 0$:

$$\lim_{n\to\infty} \sup_{(P^i)_{i\in\mathbb{N}} \in (\mathcal{M}(\mathcal{Z}))^{\mathbb{N}}} P_{\mathbb{N}}\left(\left\{z_{\mathbb{N}} \in \mathcal{Z}^{\mathbb{N}} \mid \sup_{m\geq n} d_{\mathrm{BL}}\left(\mathbb{P}_{\mathbf{W}_m(z_{\mathbb{N}})}, \frac{1}{n}\sum_{i=1}^n P^i\right) > \eta\right\}\right) = 0,$$

as long as the assumptions of Proposition 12 in Dudley et al. (1991) apply. As $\mathrm{BL}_1(\mathcal{Z}, d_{\mathcal{Z}})$ is bounded, we have $\mathcal{F}_0 = \mathrm{BL}_1(\mathcal{Z}, d_{\mathcal{Z}})$, see Dudley et al. (1991, page 499, before Proposition 10), hence it is sufficient to show that $\mathrm{BL}_1(\mathcal{Z}, d_{\mathcal{Z}})$ is image admissible Suslin (see Definition A5). By assumption $(\mathcal{Z}, d_{\mathcal{Z}})$ is totally bounded, hence $\mathrm{BL}_1(\mathcal{Z}, d_{\mathcal{Z}})$ is separable with respect to $\|\cdot\|_\infty$, see Lemma 3.2.3. As $f \in \mathrm{BL}_1(\mathcal{Z}, d_{\mathcal{Z}})$ implies $\|f\|_\infty \leq 1$, the space $\mathrm{BL}_1(\mathcal{Z}, d_{\mathcal{Z}})$ is a bounded subset of $(C_b(\mathcal{Z}, d_{\mathcal{Z}}), \|\cdot\|_\infty)$, which is due to Dudley (1989, Theorem 2.4.9) a complete space. Now, $\mathrm{BL}_1(\mathcal{Z}, d_{\mathcal{Z}})$ is a closed subset of $(C_b(\mathcal{Z}, d_{\mathcal{Z}}), \|\cdot\|_\infty)$ with respect to $\|\cdot\|_\infty$. Hence $\mathrm{BL}_1(\mathcal{Z}, d_{\mathcal{Z}})$ is complete, due to Denkowski et al. (2003, Proposition 1.4.17). Therefore $\mathrm{BL}_1(\mathcal{Z}, d_{\mathcal{Z}})$ is separable and complete with respect to $\|\cdot\|_\infty$ and particularly a Suslin space (see Definition A4), see Dudley (2014, p.229). As Lipschitz continuous functions are also equicontinuous, Dudley (2014, Theorem 5.28 (c)) gives that $\mathrm{BL}_1(\mathcal{Z}, d_{\mathcal{Z}})$ is image admissible Suslin.

Hence, Dudley et al. (1991, Theorem 7) yields

$$\sup_{(P^i)_{i\in\mathbb{N}} \in (\mathcal{M}(\mathcal{Z}))^{\mathbb{N}}} d_{\mathrm{BL}}\left(\mathbb{P}_{\mathbf{W}_n}, \frac{1}{n}\sum_{i=1}^n P^i\right) \longrightarrow 0 \text{ almost surely with respect to } P_{\mathbb{N}},\ n \to \infty,$$

and

$$\sup_{(Q^i)_{i\in\mathbb{N}}\in(\mathcal{M}(\mathcal{Z}))^{\mathbb{N}}} d_{\mathrm{BL}}\left(\mathbb{Q}_{\tilde{\mathbf{W}}_n}, \frac{1}{n}\sum_{i=1}^{n} Q^i\right) \longrightarrow 0 \text{ almost surely with respect to } Q_{\mathbb{N}},\ n\to\infty.$$

That is, there is $n_6 \in \mathbb{N}$ such that for all $n \geq n_6$

$$P_n\left(\left\{\mathbf{w}_n \in \mathcal{Z}^n \mid d_{\mathrm{BL}}\left(\mathbb{P}_{\mathbf{w}_n}, \frac{1}{n}\sum_{i=1}^{n} P^i\right) \leq \frac{\delta_0}{6}\right\}\right) \geq 1-\frac{\varepsilon}{10}, \tag{3.33}$$

$$\text{and}\quad Q_n\left(\left\{\tilde{\mathbf{w}}_n \in \mathcal{Z}^n \mid d_{\mathrm{BL}}\left(\mathbb{Q}_{\tilde{\mathbf{w}}_n}, \frac{1}{n}\sum_{i=1}^{n} Q^i\right) \leq \frac{\delta_0}{6}\right\}\right) \geq 1-\frac{\varepsilon}{10}. \tag{3.34}$$

Moreover, due to Lemma 3.4.4, we have

$$d_{\mathrm{BL}}(P_n, Q_n) \leq \frac{\delta_0}{6} \quad \Rightarrow \quad d_{\mathrm{BL}}\left(\frac{1}{n}\sum_{i=1}^{n} P^i, \frac{1}{n}\sum_{i=1}^{n} Q^i\right) \leq \frac{\delta_0}{6}. \tag{3.35}$$

Then the strong Varadarajan property of $(Z_i)_{i\in\mathbb{N}}$ yields that there is $n_7 \in \mathbb{N}$ such that for all $n \geq n_7$:

$$P_n\left(\left\{\mathbf{w}_n \in \mathcal{Z}^n \mid d_{\mathrm{BL}}(\mathbb{P}_{\mathbf{w}_n}, P) \leq \frac{\delta_0}{6}\right\}\right) \geq 1-\frac{\varepsilon}{10}. \tag{3.36}$$

Similar to the argumentation for part I we conclude, using again the boundedness of the Prohorov metric π_{d_H} by the Ky Fan metric, see Dudley (1989, Theorem 11.3.5):

$$\begin{aligned}
&\pi_{d_H}(\mathcal{L}_{P_n}(S_n), \mathcal{L}_{Q_n^*}(S_n)) \ = \pi_{d_H}(S_n\circ\mathbf{W}_n, S_n\circ\tilde{\mathbf{W}}'_n)\\
&= \inf\{\tilde{\varepsilon}>0 \mid (\mathbf{W}_n, \tilde{\mathbf{W}}_n, \tilde{\mathbf{W}}^*_n, \tilde{\mathbf{W}}'_n)(\tilde{K}_{\mathbb{N}})\left(\left\{(\mathbf{w}_n, \tilde{\mathbf{w}}_n, \tilde{\mathbf{w}}^*_n, \tilde{\mathbf{w}}'_n) \in \mathcal{Z}^n\times\mathcal{Z}^n\times\mathcal{Z}^n\times\mathcal{Z}^n \mid \right.\right.\\
&\qquad\qquad \left.\left. d_H(S_n(\mathbf{w}_n), S_n(\tilde{\mathbf{w}}'_n)) > \tilde{\varepsilon},\ \tilde{\mathbf{w}}_n, \tilde{\mathbf{w}}^*_n \in \mathcal{Z}^n\right\}\right) \leq \tilde{\varepsilon}\}.
\end{aligned}$$

Due to the definition of the statistical operator S, this is equivalent to

$$\begin{aligned}
&\inf\{\tilde{\varepsilon}>0 \mid (\mathbf{W}_n, \tilde{\mathbf{W}}_n, \tilde{\mathbf{W}}^*_n, \tilde{\mathbf{W}}'_n)(\tilde{K}_{\mathbb{N}})\left(\left\{(\mathbf{w}_n, \tilde{\mathbf{w}}_n, \tilde{\mathbf{w}}^*_n, \tilde{\mathbf{w}}'_n) \in \mathcal{Z}^n\times\mathcal{Z}^n\times\mathcal{Z}^n\times\mathcal{Z}^n \mid \right.\right.\\
&\qquad\qquad \left.\left. d_H(S(\mathbb{P}_{\mathbf{w}_n}), S(\mathbb{Q}_{\tilde{\mathbf{w}}'_n})) > \tilde{\varepsilon},\ \tilde{\mathbf{w}}_n, \tilde{\mathbf{w}}^*_n \in \mathcal{Z}^n\right\}\right) \leq \tilde{\varepsilon}\}.
\end{aligned}$$

Moreover the triangle inequality yields

$$d_H(S(\mathbb{P}_{\mathbf{w}_n}), S(\mathbb{Q}_{\tilde{\mathbf{w}}'_n})) \leq d_H(S(\mathbb{P}_{\mathbf{w}_n}), S(P)) + d_H(S(P), S(\mathbb{Q}_{\tilde{\mathbf{w}}'_n})).$$

Hence, for all $n \geq \max\{n_4, n_5, n_6, n_7\}$, we obtain

$$\begin{aligned}
&(\mathbf{W}_n, \tilde{\mathbf{W}}_n, \tilde{\mathbf{W}}_n^*, \tilde{\mathbf{W}}_n')(\tilde{K}_{\mathbb{N}})\Big(\Big\{(\mathbf{w}_n, \tilde{\mathbf{w}}_n, \tilde{\mathbf{w}}_n^*, \tilde{\mathbf{w}}_n') \in \mathcal{Z}^n \times \mathcal{Z}^n \times \mathcal{Z}^n \times \mathcal{Z}^n \mid \\
&\qquad\qquad d_H(S(\mathbb{P}_{\mathbf{w}_n}), S(\mathbb{Q}_{\tilde{\mathbf{w}}_n'})) > \frac{\varepsilon}{2},\ \tilde{\mathbf{w}}_n, \tilde{\mathbf{w}}_n^* \in \mathcal{Z}^n\Big\}\Big) \\
&\leq (\mathbf{W}_n, \tilde{\mathbf{W}}_n, \tilde{\mathbf{W}}_n^*, \tilde{\mathbf{W}}_n')(\tilde{K}_{\mathbb{N}})\Big(\Big\{(\mathbf{w}_n, \tilde{\mathbf{w}}_n, \tilde{\mathbf{w}}_n^*, \tilde{\mathbf{w}}_n') \in \mathcal{Z}^n \times \mathcal{Z}^n \times \mathcal{Z}^n \times \mathcal{Z}^n \mid \\
&\qquad\qquad d_H(S(\mathbb{P}_{\mathbf{w}_n}), S(P)) > \frac{\varepsilon}{4} \text{ or } d_H(S(P), S(\mathbb{Q}_{\tilde{\mathbf{w}}_n'})) > \frac{\varepsilon}{4},\ \tilde{\mathbf{w}}_n, \tilde{\mathbf{w}}_n^* \in \mathcal{Z}^n\Big\}\Big).
\end{aligned}$$

The continuity of the statistical operator S in P, see (3.26), gives

$$\begin{aligned}
d_{\mathrm{BL}}(P, \mathbb{Q}_{\tilde{\mathbf{W}}_n'}) \leq \delta_0 \quad &\Rightarrow \quad d_H(S(P), S(\mathbb{Q}_{\tilde{\mathbf{W}}_n'})) \leq \frac{\varepsilon}{4}, \\
\text{and} \quad d_{\mathrm{BL}}(P, \mathbb{P}_{\mathbf{W}_n}) \leq \delta_0 \quad &\Rightarrow \quad d_H(S(P), S(\mathbb{P}_{\mathbf{W}_n})) \leq \frac{\varepsilon}{4}.
\end{aligned}$$

Further, the triangle inequality yields

$$\begin{aligned}
d_{\mathrm{BL}}(P, \mathbb{Q}_{\tilde{\mathbf{w}}_n'}) \leq\ & d_{\mathrm{BL}}(P, \mathbb{P}_{\mathbf{w}_n}) + d_{\mathrm{BL}}\left(\mathbb{P}_{\mathbf{w}_n}, \frac{1}{n}\sum_{i=1}^n P^i\right) + d_{\mathrm{BL}}\left(\frac{1}{n}\sum_{i=1}^n P^i, \frac{1}{n}\sum_{i=1}^n Q^i\right) \\
&+ d_{\mathrm{BL}}\left(\frac{1}{n}\sum_{i=1}^n Q^i, \mathbb{Q}_{\tilde{\mathbf{w}}_n}\right) + d_{\mathrm{BL}}(\mathbb{Q}_{\tilde{\mathbf{w}}_n}, \mathbb{Q}_{\tilde{\mathbf{w}}_n^*}) + d_{\mathrm{BL}}(\mathbb{Q}_{\tilde{\mathbf{w}}_n^*}, \mathbb{Q}_{\tilde{\mathbf{w}}_n'}). \qquad (3.37)
\end{aligned}$$

Therefore we conclude, for all $n \geq \max\{n_4, n_5, n_6, n_7\}$,

$$\begin{aligned}
&(\mathbf{W}_n, \tilde{\mathbf{W}}_n, \tilde{\mathbf{W}}_n^*, \tilde{\mathbf{W}}_n')(\tilde{K}_{\mathbb{N}})\left(\left\{(\mathbf{w}_n, \tilde{\mathbf{w}}_n, \tilde{\mathbf{w}}_n^*, \tilde{\mathbf{w}}_n') \in \mathcal{Z}^n \times \mathcal{Z}^n \times \mathcal{Z}^n \times \mathcal{Z}^n \mid \right.\right. \\
&\qquad\qquad \left.\left. d_H(S(\mathbb{P}_{\mathbf{w}_n}), S(P)) > \frac{\varepsilon}{4} \text{ or } d_H(S(P), S(\mathbb{Q}_{\tilde{\mathbf{w}}_n'})) > \frac{\varepsilon}{4},\ \tilde{\mathbf{w}}_n, \tilde{\mathbf{w}}_n^* \in \mathcal{Z}^n\right\}\right) \\
&\overset{(3.26)}{\leq} (\mathbf{W}_n, \tilde{\mathbf{W}}_n, \tilde{\mathbf{W}}_n^*, \tilde{\mathbf{W}}_n')(\tilde{K}_{\mathbb{N}})\left(\left\{(\mathbf{w}_n, \tilde{\mathbf{w}}_n, \tilde{\mathbf{w}}_n^*, \tilde{\mathbf{w}}_n') \in \mathcal{Z}^n \times \mathcal{Z}^n \times \mathcal{Z}^n \times \mathcal{Z}^n \mid \right.\right. \\
&\qquad\qquad \left.\left. d_{\mathrm{BL}}(\mathbb{P}_{\mathbf{w}_n}, P) > \delta_0 \text{ or } d_{\mathrm{BL}}(P, \mathbb{Q}_{\tilde{\mathbf{w}}_n'}) > \delta_0,\ \tilde{\mathbf{w}}_n, \tilde{\mathbf{w}}_n^* \in \mathcal{Z}^n\right\}\right) \\
&\overset{(3.37)}{\leq} (\mathbf{W}_n, \tilde{\mathbf{W}}_n, \tilde{\mathbf{W}}_n^*, \tilde{\mathbf{W}}_n')(\tilde{K}_{\mathbb{N}})\left(\left\{(\mathbf{w}_n, \tilde{\mathbf{w}}_n, \tilde{\mathbf{w}}_n^*, \tilde{\mathbf{w}}_n') \in \mathcal{Z}^n \times \mathcal{Z}^n \times \mathcal{Z}^n \times \mathcal{Z}^n \mid \right.\right. \\
&\qquad\qquad d_{\mathrm{BL}}(\mathbb{P}_{\mathbf{w}_n}, P) > \frac{\delta_0}{6} \text{ or } d_{\mathrm{BL}}\left(\mathbb{P}_{\mathbf{w}_n}, \frac{1}{n}\sum_{i=1}^n P^i\right) > \frac{\delta_0}{6} \\
&\qquad\qquad \text{or } d_{\mathrm{BL}}\left(\frac{1}{n}\sum_{i=1}^n P^i, \frac{1}{n}\sum_{i=1}^n Q^i\right) > \frac{\delta_0}{6} \text{ or } d_{\mathrm{BL}}\left(\frac{1}{n}\sum_{i=1}^n Q^i, \mathbb{Q}_{\tilde{\mathbf{w}}_n}\right) > \frac{\delta_0}{6} \\
&\qquad\qquad \left.\left. \text{or } d_{\mathrm{BL}}(\mathbb{Q}_{\tilde{\mathbf{w}}_n}, \mathbb{Q}_{\tilde{\mathbf{w}}_n^*}) > \frac{\delta_0}{6} \text{ or } d_{\mathrm{BL}}(\mathbb{Q}_{\tilde{\mathbf{w}}_n^*}, \mathbb{Q}_{\tilde{\mathbf{w}}_n'}) > \frac{\delta_0}{6}\right\}\right).
\end{aligned}$$

Now, assume $d_{\mathrm{BL}}(P_n, Q_n) \leq \frac{\delta_0}{6}$, then (3.35) yields $d_{\mathrm{BL}}\left(\frac{1}{n}\sum_{i=1}^n P^i, \frac{1}{n}\sum_{i=1}^n Q^i\right) \leq \frac{\delta_0}{6}$, therefore this term can be omitted. Note that this is only proven for the p-product metrics on $\mathcal{Z}^n$ and not for the metric d_n from (3.3). For this metric we need a different argumentation, which is stated below the next calculation.

Hence, for all $n \geq \max\{n_4, n_5, n_6, n_7\}$,

$$
\begin{aligned}
&(\mathbf{W}_n, \tilde{\mathbf{W}}_n, \tilde{\mathbf{W}}_n^*, \tilde{\mathbf{W}}_n')(\tilde{K}_{\mathbb{N}})\left(\left\{(\mathbf{w}_n, \tilde{\mathbf{w}}_n, \tilde{\mathbf{w}}_n^*, \tilde{\mathbf{w}}_n') \in \mathcal{Z}^n \times \mathcal{Z}^n \times \mathcal{Z}^n \times \mathcal{Z}^n \mid \right.\right.\\
&\qquad\qquad \left.\left. d_H(S(\mathbb{P}_{\mathbf{w}_n}), S(\mathbb{Q}_{\tilde{\mathbf{w}}_n'})) > \varepsilon,\ \tilde{\mathbf{w}}_n, \tilde{\mathbf{w}}_n^* \in \mathcal{Z}^n\right\}\right)\\
&\overset{(3.35)}{\leq} (\mathbf{W}_n, \tilde{\mathbf{W}}_n, \tilde{\mathbf{W}}_n^*, \tilde{\mathbf{W}}_n')(\tilde{K}_{\mathbb{N}})\left(\left\{(\mathbf{w}_n, \tilde{\mathbf{w}}_n, \tilde{\mathbf{w}}_n^*, \tilde{\mathbf{w}}_n') \in \mathcal{Z}^n \times \mathcal{Z}^n \times \mathcal{Z}^n \times \mathcal{Z}^n \mid \right.\right.\\
&\qquad d_{\mathrm{BL}}(\mathbb{P}_{\mathbf{w}_n}, P) > \frac{\delta_0}{6} \text{ or } d_{\mathrm{BL}}\left(\mathbb{P}_{\mathbf{w}_n}, \frac{1}{n}\sum_{i=1}^n P^i\right) > \frac{\delta_0}{6} \text{ or } d_{\mathrm{BL}}\left(\frac{1}{n}\sum_{i=1}^n Q^i, \mathbb{Q}_{\tilde{\mathbf{w}}_n}\right) > \frac{\delta_0}{6}\\
&\qquad \left.\left. \text{or } d_{\mathrm{BL}}(\mathbb{Q}_{\tilde{\mathbf{w}}_n}, \mathbb{Q}_{\tilde{\mathbf{w}}_n^*}) > \frac{\delta_0}{6} \text{ or } d_{\mathrm{BL}}(\mathbb{Q}_{\tilde{\mathbf{w}}_n^*}, \mathbb{Q}_{\tilde{\mathbf{w}}_n'}) > \frac{\delta_0}{6}\right\}\right)\\
&\leq P_n\left(\left\{\mathbf{w}_n \in \mathcal{Z}^n \mid d_{\mathrm{BL}}(\mathbb{P}_{\mathbf{w}_n}, P) > \frac{\delta_0}{6}\right\}\right)\\
&\quad + P_n\left(\left\{\mathbf{w}_n \in \mathcal{Z}^n \mid d_{\mathrm{BL}}\left(\mathbb{P}_{\mathbf{w}_n}, \frac{1}{n}\sum_{i=1}^n P^i\right) > \frac{\delta_0}{6}\right\}\right)\\
&\quad + Q_n\left(\left\{\tilde{\mathbf{w}}_n \in \mathcal{Z}^n \mid d_{\mathrm{BL}}\left(\frac{1}{n}\sum_{i=1}^n Q^i, \mathbb{Q}_{\tilde{\mathbf{w}}_n}\right) > \frac{\delta_0}{6}\right\}\right)\\
&\quad + \otimes_{i=1}^n \mathbb{Q}_{\tilde{\mathbf{w}}_n}\left(\left\{\tilde{\mathbf{w}}_n^* \in \mathcal{Z}^n \mid d_{\mathrm{BL}}\left(\mathbb{Q}_{\tilde{\mathbf{w}}_n}, \mathbb{Q}_{\tilde{\mathbf{w}}_n^*}\right) > \frac{\delta_0}{6}\right\}\right)\\
&\quad + Q_n^*\left(\left\{\tilde{\mathbf{w}}_n' \in \mathcal{Z}^n \mid d_{\mathrm{BL}}\left(\mathbb{Q}_{\tilde{\mathbf{w}}_n^*}, \mathbb{Q}_{\tilde{\mathbf{w}}_n'}\right) > \frac{\delta_0}{6}\right\}\right)\\
&\overset{(3.31),(3.32)(3.33),(3.34),(3.36)}{<} \frac{\varepsilon}{10} + \frac{\varepsilon}{10} + \frac{\varepsilon}{10} + \frac{\varepsilon}{10} + \frac{\varepsilon}{10} = \frac{\varepsilon}{2}.
\end{aligned}
$$

In order to show the above bound for the metric d_n, see (3.3), on $\mathcal{Z}^n$, we use another variant of the triangle inequality in (3.37):

$$
d_{\mathrm{BL}}(P, \mathbb{Q}_{\tilde{\mathbf{w}}_n'}) \leq d_{\mathrm{BL}}(P, \mathbb{P}_{\mathbf{w}_n}) + d_{\mathrm{BL}}\left(\mathbb{P}_{\mathbf{w}_n}, \mathbb{Q}_{\tilde{\mathbf{w}}_n}\right) + d_{\mathrm{BL}}(\mathbb{Q}_{\tilde{\mathbf{w}}_n}, \mathbb{Q}_{\tilde{\mathbf{w}}_n^*}) + d_{\mathrm{BL}}(\mathbb{Q}_{\tilde{\mathbf{w}}_n^*}, \mathbb{Q}_{\tilde{\mathbf{w}}_n'}). \tag{3.38}
$$

Assume $d_{\mathrm{BL}}(P_n, Q_n) \leq \frac{\delta_0^2}{64}$. Then, the strong equivalence between the Prohorov metric and the bounded Lipschitz metric on Polish spaces, see Huber (1981, Chapter 2, Corollary 4.3), yields $\pi_{d_n}(P_n, Q_n) \leq \sqrt{d_{\mathrm{BL}}(P_n, Q_n)} \leq \frac{\delta_0}{8}$. Due to Dudley (1989, Theorem 11.6.2),

$\pi_{d_n}(P_n, Q_n) \leq \frac{\delta_0}{8}$ implies the existence of a probability measure $\mu \in \mathcal{M}(\mathcal{Z}^n \times \mathcal{Z}^n)$ with marginal distributions P_n and Q_n, such that $\mu\left(\left\{(\mathbf{w}_n, \tilde{\mathbf{w}}_n) \in \mathcal{Z}^n \times \mathcal{Z}^n \mid d_n(\mathbf{w}_n, \tilde{\mathbf{w}}_n) > \frac{\delta_0}{8}\right\}\right) \leq \frac{\delta_0}{8}$. As $d_n(\mathbf{w}_n, \tilde{\mathbf{w}}_n) \leq \frac{\delta_0}{8}$ implies $\pi_{d_n}\left(\frac{1}{n}\sum_{i=1}^n \delta_{z_i}, \frac{1}{n}\sum_{i=1}^n \delta_{\tilde{z}_i}\right) \leq \frac{\delta_0}{8}$, see (3.11), we have:

$$\mu\left(\left\{(\mathbf{w}_n, \tilde{\mathbf{w}}_n) \in \mathcal{Z}^n \times \mathcal{Z}^n \mid \pi_{d_n}(\mathbb{P}_{\mathbf{w}_n}, \mathbb{Q}_{\tilde{\mathbf{w}}_n}) > \frac{\delta_0}{8}\right\}\right) \leq \frac{\delta_0}{8}.$$

Again the equivalence between the metrics π and d_{BL} yields:

$$\mu\left(\left\{(\mathbf{w}_n, \tilde{\mathbf{w}}_n) \in \mathcal{Z}^n \times \mathcal{Z}^n \mid d_{\mathrm{BL}}(\mathbb{P}_{\mathbf{w}_n}, \mathbb{Q}_{\tilde{\mathbf{w}}_n}) > \frac{\delta_0}{4}\right\}\right) \leq \frac{\delta_0}{8}.$$

Now we choose the joint distribution $\tilde{K}_{\mathbb{N}}$ of $\mathbf{W}_{\mathbb{N}}$, $\tilde{\mathbf{W}}_{\mathbb{N}}$, $\tilde{\mathbf{W}}^*_{\mathbb{N}}$, and $\tilde{\mathbf{W}}'_{\mathbb{N}}$ such that the distribution of $(\mathbf{W}_n, \tilde{\mathbf{W}}_n) : \mathcal{Z}^{\mathbb{N}} \times \mathcal{Z}^{\mathbb{N}} \to \mathcal{Z}^n \times \mathcal{Z}^n$ is $\mu \in \mathcal{M}(\mathcal{Z}^n \times \mathcal{Z}^n)$. Then we conclude:

$$\begin{aligned}
&(\mathbf{W}_n, \tilde{\mathbf{W}}_n, \tilde{\mathbf{W}}^*_n, \tilde{\mathbf{W}}'_n)(\tilde{K}_{\mathbb{N}})\left(\left\{(\mathbf{w}_n, \tilde{\mathbf{w}}_n, \tilde{\mathbf{w}}^*_n, \tilde{\mathbf{w}}'_n) \in \mathcal{Z}^n \times \mathcal{Z}^n \times \mathcal{Z}^n \times \mathcal{Z}^n \mid \right.\right.\\
&\qquad \left.\left. d_H(S(\mathbb{P}_{\mathbf{w}_n}), S(P)) > \frac{\varepsilon}{4} \text{ or } d_H(S(P), S(\mathbb{Q}_{\tilde{\mathbf{w}}'_n})) > \frac{\varepsilon}{4},\ \tilde{\mathbf{w}}_n, \tilde{\mathbf{w}}^*_n \in \mathcal{Z}^n\right\}\right)\\
\overset{(3.26),(3.38)}{\leq}\ &(\mathbf{W}_n, \tilde{\mathbf{W}}_n, \tilde{\mathbf{W}}^*_n, \tilde{\mathbf{W}}'_n)(\tilde{K}_{\mathbb{N}})\left(\left\{(\mathbf{w}_n, \tilde{\mathbf{w}}_n, \tilde{\mathbf{w}}^*_n, \tilde{\mathbf{w}}'_n) \in \mathcal{Z}^n \times \mathcal{Z}^n \times \mathcal{Z}^n \times \mathcal{Z}^n \mid \right.\right.\\
&\qquad d_{\mathrm{BL}}(\mathbb{P}_{\mathbf{w}_n}, P) > \frac{\delta_0}{4} \text{ or } d_{\mathrm{BL}}\left(\mathbb{P}_{\mathbf{w}_n}, \mathbb{Q}_{\tilde{\mathbf{w}}_n}\right) > \frac{\delta_0}{4}\\
&\qquad \left.\left. \text{or } d_{\mathrm{BL}}(\mathbb{Q}_{\tilde{\mathbf{w}}_n}, \mathbb{Q}_{\tilde{\mathbf{w}}^*_n}) > \frac{\delta_0}{4} \text{ or } d_{\mathrm{BL}}(\mathbb{Q}_{\tilde{\mathbf{w}}^*_n}, \mathbb{Q}_{\tilde{\mathbf{w}}'_n}) > \frac{\delta_0}{4}\right\}\right).\\
\leq\ &P_n\left(\left\{\mathbf{w}_n \in \mathcal{Z}^n \mid d_{\mathrm{BL}}(\mathbb{P}_{\mathbf{w}_n}, P) > \frac{\delta_0}{4}\right\}\right)\\
&+ \mu\left(\left\{(\mathbf{w}_n, \tilde{\mathbf{w}}_n) \in \mathcal{Z}^n \times \mathcal{Z}^n \mid d_{\mathrm{BL}}\left(\mathbb{P}_{\mathbf{w}_n}, \mathbb{Q}_{\tilde{\mathbf{w}}_n}\right) > \frac{\delta_0}{4}\right\}\right)\\
&+ \otimes_{i=1}^n \mathbb{Q}_{\tilde{\mathbf{w}}_n}\left(\left\{\tilde{\mathbf{w}}^*_n \in \mathcal{Z}^n \mid d_{\mathrm{BL}}\left(\mathbb{Q}_{\tilde{\mathbf{w}}_n}, \mathbb{Q}_{\tilde{\mathbf{w}}^*_n}\right) > \frac{\delta_0}{4}\right\}\right)\\
&+ Q^*_n\left(\left\{\tilde{\mathbf{w}}'_n \in \mathcal{Z}^n \mid d_{\mathrm{BL}}\left(\mathbb{Q}_{\tilde{\mathbf{w}}^*_n}, \mathbb{Q}_{\tilde{\mathbf{w}}'_n}\right) > \frac{\delta_0}{4}\right\}\right).
\end{aligned}$$

Now, adapting the inequalities in (3.31), (3.32), and (3.36) in ε respectively n yields the boundedness of the above term by $\frac{\varepsilon}{2}$ for $d_{\mathrm{BL}}(P_n, Q_n) \leq \frac{\delta_0^2}{64}$ and for all $n \geq \{n_4, n_5, n_7\}$.

Now we can go on with the proof similar for both kinds of metrics on $\mathcal{Z}^n$.

The equivalence between the Prohorov metric and the bounded Lipschitz metric on Polish spaces, see Huber (1981, Chapter 2, Corollary 4.3), yields the existence of $n_{0,2} \in \mathbb{N}$ such

that for all $n \geq n_{0,2}$, $d_{\mathrm{BL}}(P_n, Q_n) \leq \frac{\delta_0}{6}$ (respectively $d_{\mathrm{BL}}(P_n, Q_n) \leq \frac{\delta_0^2}{64}$) implies

$$d_{\mathrm{BL}}(\mathcal{L}_{P_n}(S_n), \mathcal{L}_{Q_n^*}(S_n)) < \frac{\varepsilon}{2}. \tag{3.39}$$

Now, (3.30) and (3.39) yield for all $n \geq \max\{n_{0,1}, n_{0,2}\}$:

$$d_{\mathrm{BL}}(\mathcal{L}_{P_n^*}(S_n), \mathcal{L}_{Q_n^*}(S_n)) < \varepsilon. \tag{3.40}$$

Recall that $\mathcal{L}_{P_n^*}(S_n) =: \zeta_n$ and $\mathcal{L}_{Q_n^*}(S_n) =: \xi_n$ are random quantities with values in $\mathcal{M}(H)$. Hence (3.40) is equivalent to

$$\mathbb{E}\left[d_{\mathrm{BL}}(\mathcal{L}_{P_n^*}(S_n), \mathcal{L}_{Q_n^*}(S_n))\right] < \varepsilon, \text{ for all } n \geq \max\{n_{0,1}, n_{0,2}\},$$

respectively

$$\mathbb{E}\left[d_{\mathrm{BL}}(\zeta_n, \xi_n)\right] < \varepsilon, \text{ for all } n \geq \max\{n_{0,1}, n_{0,2}\}.$$

Therefore, for all $f \in \mathrm{BL}_1(\mathcal{M}(\mathcal{Z}))$ and for all $n \geq \max\{n_{0,1}, n_{0,2}\}$:

$$\begin{aligned}\left|\int f d(\mathcal{L}(\zeta_n)) - \int f d(\mathcal{L}(\xi_n))\right| = |\mathbb{E}f(\zeta_n) - \mathbb{E}f(\xi_n)| &\leq \mathbb{E}\left|f(\zeta_n) - f(\xi_n)\right| \\ &\leq \mathbb{E}\left(|f|_1 \, d_{BL}(\zeta_n, \xi_n)\right) \quad < \varepsilon,\end{aligned}$$

by a variant of Strassen's Theorem, see Huber (1981, Chapter 2, Theorem 4.2, (2)⇒(1)). That is,

$$d_{\mathrm{BL}}(\mathcal{L}(\mathcal{L}_{P_n^*}(S_n)), \mathcal{L}(\mathcal{L}_{Q_n^*}(S_n))) < \varepsilon \text{ for all } n \geq \max\{n_{0,1}, n_{0,2}\}.$$

Hence for every $\varepsilon > 0$ we find $\delta = \frac{\delta_0}{6}$ and $n_0 = \max\{n_{0,1}, n_{0,2}\}$ such that for all $n \geq n_0$:

$$d_{\mathrm{BL}}(P_n, Q_n) < \delta \quad \Rightarrow \quad d_{\mathrm{BL}}(\mathcal{L}(\mathcal{L}_{P_n^*}(S_n)), \mathcal{L}(\mathcal{L}_{Q_n^*}(S_n))) < \varepsilon,$$

which yields the assertion. □

The next part gives two examples of stochastic processes of independent, but not necessarily identically distributed random variables, which are Varadarajan processes. In particular these stochastic processes even satisfy a strong law of large numbers for events (SLLNE) in the sense of Steinwart et al. (2009) and therefore are, due to Theorem 3.2.1, strong Varadarajan processes. The first example is rather simple and describes a sequence of univariate normal distributions.

Example 1 *Let $(a_i)_{i\in\mathbb{N}} \subset \mathbb{R}$ be a sequence with $\lim_{i\to\infty} a_i = a \in \mathbb{R}$ and let $|a_i| \leq c$, for some constant $c > 0$ for all $i \in \mathbb{N}$. Let $(Z_i)_{i\in\mathbb{N}}$, $Z_i : \Omega \to \mathbb{R}$, be a stochastic process where Z_i, $i \in \mathbb{N}$, are independent and $Z_i \sim N(a_i, 1)$, $i \in \mathbb{N}$. Then the process $(Z_i)_{i\in\mathbb{N}}$ is a strong Varadarajan process.*

Proof: Without any restriction we assume $a = 0$. Otherwise regard the process $Z_i - a$, $i \in \mathbb{N}$. By assumption, the random variables Z_i, $i \in \mathbb{N}$, are independent. Hence $I_B \circ Z_i$, $i \in \mathbb{N}$, are independent, see for example Hoffmann-Jørgensen (1994, Theorem 2.10.6) for all measurable $B \in \mathcal{B}$, as I_B is a measurable function. According to Steinwart et al. (2009, Proposition 2.8), $(Z_i)_{i\in\mathbb{N}}$ satisfies the SLLNE if there is a probability measure P in $\mathcal{M}(\mathcal{Z})$ such that $\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_\mu I_B \circ Z_i = P(B)$ for all measurable $B \in \mathcal{B}$. Hence:

$$\frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_\mu I_B \circ Z_i \;=\; \frac{1}{n} \sum_{i=1}^{n} \int I_B \, dZ_i(\mu) \;=\; \frac{1}{n} \sum_{i=1}^{n} \int I_B f_i \, d\lambda^1,$$

where $f_i(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-a_i)^2}$ denotes the density of the normal distribution $N(0,1)$ with respect to the Lebesgue measure λ^1. Moreover define $g : \mathbb{R} \to \mathbb{R}$ by

$$g(x) = \begin{cases} e^{-\frac{1}{2}(x+c)^2}, & x < -c \\ \frac{1}{\sqrt{2\pi}}, & -c \leq x \leq c \\ e^{-\frac{1}{2}(x-c)^2}, & c < x \end{cases} \qquad x \in \mathbb{R}.$$

Therefore $|f_i| \leq |g|$, for all $i \in \mathbb{N}$, g is integrable and due to Lebesgue's Theorem, see for example Hoffmann-Jørgensen (1994, Theorem 3.6):

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \int I_B f_i \, d\lambda^1 \;=\; \lim_{n\to\infty} \int \frac{1}{n} \sum_{i=1}^{n} I_B f_i \, d\lambda^1 \;=\; \int \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} I_B f_i \, d\lambda^1. \tag{3.41}$$

We have $f_i \to f_0$, where $f_0 = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$ for all $x \in \mathbb{R}$, as $a_i \to 0$ and therefore the Lemma of Kronecker, see for example Hoffmann-Jørgensen (1994, Theorem 4.9, Equation 4.9.1) yields: $\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} f_i(x) = f_0(x)$ for all $x \in \mathcal{X}$.

Now (3.41) yields the SLLNE:

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \int I_B f_i \, d\lambda^1 = \int I_B f_0 \, d\lambda^1 = P(B), \text{ for al } B \in \mathcal{B}.$$

With Theorem 3.2.1 the Varadarajan property is given. □

The second example are stochastic processes where the distributions of the random variables Z_i, $i \in \mathbb{N}$, are lying in a so-called shrinking ε-neighbourhood of a probability measure P.

Example 2 *Let $(\mathcal{Z}, \mathcal{B})$ be a measurable space and let $(Z_i)_{i\in\mathbb{N}}$ be a stochastic process with independent random variables $Z_i : \Omega \to \mathcal{Z}$, $Z_i \sim P^i$, where*

$$P^i = (1 - \varepsilon_i)P + \varepsilon \tilde{P}^i$$

for a sequence $\varepsilon_i \to 0$, $i \to \infty$, $\varepsilon_i > 0$ and $\tilde{P}^i$, $P \in \mathcal{M}(\mathcal{Z})$, $i \in \mathbb{N}$. Then the process $(Z_i)_{i\in\mathbb{N}}$ is a strong Varadarajan process.

Proof: Similar to the proof of Example 1, we first show the SLLNE, that is there exists a probability measure $P \in \mathcal{M}(\mathcal{Z})$ such that

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \int I_B \circ Z_i \, d\mu = P(B), \text{ for all measurable } B \subset \Omega.$$

Now let $B \subset \Omega$ be an arbitrary measurable set. Then:

$$\begin{aligned} \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \int I_B \circ Z_i \, d\mu &= \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \int_{\mathcal{Z}} I_B \, dP^i = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \int_{\mathcal{Z}} I_B \, d[(1-\varepsilon_i)P + \varepsilon_i \tilde{P}^i] \\ &= \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \int_{\mathcal{Z}} I_B \, dP - \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i \int_{\mathcal{Z}} I_B \, dP + \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i \int_{\mathcal{Z}} I_B \, d\tilde{P}^i. \end{aligned} \tag{3.42}$$

As, $0 \le \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i \int I_B \, dP \le \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i$ and $\varepsilon_i \to 0$, we have

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i \int I_B \, dP \le \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i \longrightarrow 0, \; n \to \infty$$

and similarly

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i \int I_B \, d\tilde{P}^i \le \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i \longrightarrow 0, \quad n \to \infty.$$

Hence (3.42) yields

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} I_B \circ Z_i = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \int I_B \, dP = P(B)$$

and therefore, due to Theorem 3.2.1, the assertion. □

In the next chapter a result for more general stochastic process, namely α-mixing processes, is established. Due to the possible dependence, stronger assumptions on the statistical operator are required.

3.4.2 Qualitative robustness for the moving block bootstrap of α-mixing processes

Dropping the independence assumption we now focus on real valued mixing processes, in particular on strongly stationary α-mixing or strong mixing stochastic processes. The mixing notion is an often used and well-accepted dependence notion which quantifies the degree of dependence of a stochastic process.

Instead of Efron's empirical bootstrap another bootstrap approach is used in order to represent the dependence structure of an α-mixing process. Künsch (1989) and Liu and Singh (1992) introduced the moving block bootstrap (MBB). Often resampling of single observations can not preserve the dependence structure of the process, therefore they decided to take blocks of length b of observations instead. The dependence structure of the process is preserved, within these blocks. The block length b increases with the number of observations n for asymptotic considerations. A slight modification of the original moving block bootstrap, see for example Politis and Romano (1990) and Shao and Yu (1993), is used in the next two theorems in order to avoid edge effects.

The following proofs are based on central limit theorems for empirical processes. There are several results concerning the moving block bootstrap of the empirical process in case of mixing processes, see for example Bühlmann (1994), Naik-Nimbalkar and Rajarshi (1994), and Peligrad (1998, Theorem 2.2) for α-mixing sequences and Radulović (1996) and Bühlmann (1995) for β-mixing sequences. To our best knowledge there are so far no results concerning qualitative robustness for bootstrap approximations of estimators for α-mixing stochastic processes. Therefore, Theorem 3.4.5 shows qualitative robustness for a stochastic process with values in $\mathbb{R}$. The proof is based on Peligrad (1998, Theorem 2.2), which provides a central limit theorem under assumptions on the process, which are weaker than those in Bühlmann (1994) and Naik-Nimbalkar and Rajarshi (1994). In the case of $\mathbb{R}^d$-valued, $d > 1$, stochastic processes, stronger assumptions on the stochastic process are needed, as the central limit theorem in Bühlmann (1994) requires stronger assumptions, see Theorem 3.4.6.

Let $Z_1, \ldots, Z_n$, $n \in \mathbb{N}$, be the first n projections of a real valued stochastic process $(Z_i)_{i\in\mathbb{N}}$ and let $b \in \mathbb{N}, b < n$, be the block length. Then, for fixed $n \in \mathbb{N}$, the sample can be divided into blocks $B_{i,b} := (Z_i, \ldots, Z_{i+b-1})$. If $i > n - b + 1$, we define $Z_{n+j} = Z_j$, for the missing elements of the blocks. To get the MBB bootstrap sample $\mathbf{W}_n^* = (Z_1^*, \ldots, Z_n^*)$, ℓ numbers $I_1, \ldots, I_\ell$ from the set $\{1, \ldots, n\}$ are randomly chosen with replacement. Without loss of generality it is assumed that $n = \ell b$, if n is not a multiple of b we simply cut the last block, which is usually done in literature. Then the sample consists of the blocks $B_{I_1,b}, B_{I_2,b}, \ldots, B_{I_\ell,b}$, that is $Z_1^* = Z_{I_1}, Z_2^* = Z_{I_1+1}, \ldots, Z_b^* =_{I_1+b-1}, Z_{b+1}^* = Z_{I_2}, \ldots, Z_{\ell b}^* = Z_{I_\ell+b-1}$.

As we are interested in estimators S_n, $n \in \mathbb{N}$, which can be represented by a statistical operator $S\colon \mathcal{M}(\mathcal{Z}) \to H$ via $S(\mathbb{P}_{\mathbf{w}_n}) = S_n(z_1, \ldots, z_n)$, for a complete separable metric space H, see (3.1), the empirical measure of the bootstrap sample $\mathbb{P}_{\mathbf{W}_n^*} = \frac{1}{n}\sum_{i=1}^n \delta_{Z_i^*}$ should approximate the empirical measure of the original sample $\mathbb{P}_{\mathbf{W}_n} = \frac{1}{n}\sum_{i=1}^n \delta_{Z_i}$. Contrarily to qualitative robustness in the case of independent and not necessarily identically distributed random variables (Theorem 3.4.2), the assumptions on the statistical operator S are strengthened for the case of α-mixing sequences. In particular the statistical operator S is assumed to be uniformly continuous for all $P \in (\mathcal{M}(\mathcal{Z}), d_{\mathrm{BL}})$. For the first theorem we assume the random variables Z_i, $i \in \mathbb{N}$, to be real valued and bounded. Without loss of generality we assume $0 \le Z_i \le 1$, otherwise a transformation leads to this assumption. For the bootstrap for the true as well as for the contaminated process, we assume the block length $b(n)$ and the number of blocks $\ell(n)$ to be sequences of integers satisfying

$$n^h \in \mathcal{O}(b(n)),\ b(n) \in \mathcal{O}(n^{1/3-a}), \text{ for some } 0 < h < \frac{1}{3} - a,\ 0 < a < \frac{1}{3},$$

$b(n) = b(2^q)$ for $2^q \le n < 2^{q+1}$, $q \in \mathbb{N}$, $b(n) \to \infty$, $n \to \infty$ and $b(n) \cdot \ell(n) = n$, $n \in \mathbb{N}$.

Theorem 3.4.5 *Let $P_{\mathbb{N}} \in \mathcal{M}(\mathbb{R}^{\mathbb{N}})$ be a probability measure on $(\mathbb{R}^{\mathbb{N}}, \mathcal{B}^{\otimes\mathbb{N}})$ such that the coordinate process $(Z_i)_{i\in\mathbb{N}}$, $Z_i : \mathbb{R}^{\mathbb{N}} \to \mathbb{R}$ is bounded, strongly stationary, and α-mixing with*

$$\sum_{m>n} \alpha(\sigma(Z_1, \ldots, Z_i), \sigma(Z_{i+m}, \ldots), P_{\mathbb{N}}) = \mathcal{O}(n^{-\gamma}),\ i \in \mathbb{N}, \text{ for some } \gamma > 0. \tag{3.43}$$

Let $\mathcal{P} \subset \mathcal{M}(\mathbb{R}^{\mathbb{N}})$ be the set of probability measures such that the coordinate process fulfils the properties above for the same $\gamma > 0$. Let (H, d_H) be a complete separable metric space, let $(S_n)_{n\in\mathbb{N}}$ be a sequence of estimators which can be represented by a statistical operator $S : \mathcal{M}(\mathbb{R}) \to H$ via (3.1). Moreover let S_n be continuous and let S be additionally uniformly

continuous with respect to d_{BL}. Then the sequence of estimators $(S_n)_{n\in\mathbb{N}}$ is qualitatively robust at $P_{\mathbb{N}}$ with respect to $\mathcal{P}$.

The assumptions on the stochastic process are on the one hand, together with the assumptions on the block length, used to ensure the validity of the bootstrap approximation and on the other hand, together with the assumptions on the statistical operator, respectively the sequence of estimators, to ensure the qualitative robustness.

Proof of Theorem 3.4.5: Let $P^*_{\mathbb{N}}, Q^*_{\mathbb{N}} \in \mathcal{M}(\mathcal{Z}^{\mathbb{N}})$ be the bootstrap approximations of the true distribution $P_{\mathbb{N}}$ and the contaminated distribution $Q_{\mathbb{N}}$. First, the triangle inequality yields:

$$\begin{aligned}
&d_{\mathrm{BL}}(\mathcal{L}_{P^*_n}(S_n), \mathcal{L}_{Q^*_n}(S_n)) \\
&\leq \underbrace{d_{\mathrm{BL}}(\mathcal{L}_{P^*_n}(S_n), \mathcal{L}_{P_n}(S_n))}_{I} + \underbrace{d_{\mathrm{BL}}(\mathcal{L}_{P_n}(S_n), \mathcal{L}_{Q_n}(S_n))}_{II} + \underbrace{d_{\mathrm{BL}}(\mathcal{L}_{Q_n}(S_n), \mathcal{L}_{Q^*_n}(S_n))}_{III}.
\end{aligned}$$

First, we regard the term in part II. Let $\sigma(Z_i),\ i \in \mathbb{N}$, be the σ-algebra generated by Z_i. Due to the assumptions on the mixing process $\sum_{m>n} \alpha(\sigma(Z_1,\ldots,Z_i), \sigma(Z_{i+m},\ldots), P_{\mathbb{N}}) = \mathcal{O}(n^{-\gamma}),\ i \in \mathbb{N}, \gamma > 0$, the sequence $(\alpha(\sigma(Z_1,\ldots,Z_i), \sigma(Z_{i+m},\ldots), \mu))_{m\in\mathbb{N}}$ is a null sequence. Moreover it is bounded by the definition of the α-mixing coefficient which, due to the strong stationarity, does not depend on i. Therefore

$$\begin{aligned}
\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha((Z_i)_{i\in\mathbb{N}}, P_{\mathbb{N}}, i, j) &= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha(\sigma(Z_i), \sigma(Z_j), P_{\mathbb{N}}) \\
&\leq \frac{2}{n^2}\sum_{i=1}^{n}\sum_{j\geq i}^{n} \alpha(\sigma(Z_i), \sigma(Z_j), P_{\mathbb{N}}) \\
&\leq \frac{2}{n^2}\sum_{i=1}^{n}\sum_{j\geq i}^{n} \alpha(\sigma(Z_1,\ldots,Z_i), \sigma(Z_j,\ldots), P_{\mathbb{N}}) \\
&= \frac{2}{n^2}\sum_{i=1}^{n}\sum_{\ell=0}^{n-i} \alpha(\sigma(Z_1,\ldots,Z_i), \sigma(Z_{i+\ell},\ldots), P_{\mathbb{N}}) \\
&\overset{\text{stationarity}}{\leq} \frac{2}{n}\sum_{\ell=0}^{n} \alpha(\sigma(Z_1,\ldots,Z_i), \sigma(Z_{i+\ell},\ldots), P_{\mathbb{N}}),\ i \in \mathbb{N} \\
&\longrightarrow 0,\ n \to \infty.
\end{aligned}$$

Hence, the process is weakly α-*bi*-mixing with respect to $P_{\mathbb{N}}$, see Definition 2.2.1. Due to the stationarity assumption, the process $(Z_i)_{i\in\mathbb{N}}$ is additionally asymptotically mean stationary,

that is $\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \mathbb{E} I_B \circ Z_i = P(B)$ for all $B \in \mathcal{A}$ for a probability measure P. Therefore the process satisfies the WLLNE, see Steinwart et al. (2009, Proposition 3.2), and therefore is a weak Varadarajan process, see Theorem 3.2.1.

As the process is assumed to be a Varadarajan process and due to the assumptions on the sequence of estimators $(S_n)_{n\in\mathbb{N}}$, qualitative robustness of $(S_n)_{n\in\mathbb{N}}$ is ensured by Theorem 3.1.3. Together with the equivalence between the Prohorov metric and the bounded Lipschitz metric for Polish spaces, see Huber (1981, Chapter 2, Corollary 4.3), it follows:

For every $\varepsilon > 0$ there is $\delta > 0$ such that for all $n \in \mathbb{N}$ and for all $Q_n \in \mathcal{M}(\mathcal{Z}^n)$ we have:

$$d_{\mathrm{BL}}(P_n, Q_n) < \delta \quad \Rightarrow \quad d_{\mathrm{BL}}(\mathcal{L}_{P_n}(S_n), \mathcal{L}_{Q_n}(S_n)) < \frac{\varepsilon}{3}.$$

This implies

$$\mathbb{E}\left[d_{\mathrm{BL}}(\mathcal{L}_{P_n}(S_n), \mathcal{L}_{Q_n}(S_n))\right] < \frac{\varepsilon}{3}. \tag{3.44}$$

Hence the convergence of the term in part II is shown.

To prove the convergence of the term in part I, consider the distribution $P_{\mathbb{N}} \in \mathcal{M}(\mathcal{Z}^{\mathbb{N}})$ and let $P_{\mathbb{N}}^*$ be the bootstrap approximation of $P_{\mathbb{N}}$, via the blockwise bootstrap. Define, for $n \in \mathbb{N}$, the random variables
$\mathbf{W}_n : \mathcal{Z}^{\mathbb{N}} \to \mathcal{Z}^n$, $\mathbf{W}_n = (Z_1, \ldots, Z_n)$, $z_{\mathbb{N}} \mapsto \mathbf{w}_n = (z_1, \ldots, z_n)$, and
$\mathbf{W}'_n : \mathcal{Z}^{\mathbb{N}} \to \mathcal{Z}^n$, $\mathbf{W}'_n = (Z'_1, \ldots, Z'_n)$, $z_{\mathbb{N}} \mapsto \mathbf{w}'_n$,
such that $\mathbf{W}_n(P_{\mathbb{N}}) = P_n$ and $\mathbf{W}'_n(P_{\mathbb{N}}^*) = P_n^*$.
Moreover denote the bootstrap sample by $\mathbf{W}_n^* : \mathcal{Z}^{\mathbb{N}} \to \mathcal{Z}^n$, $\mathbf{W}_n^* := (Z_1^*, \ldots, Z_n^*)$, $z_{\mathbb{N}} \mapsto \mathbf{w}_n^*$, and the distribution of $\mathbf{W}_n^*$ by $\overline{P}_n$. The blockwise bootstrap approximation of P_m, $m \in \mathbb{N}$, is $P_m^* = \otimes_{j=1}^{m} \frac{1}{n} \sum_{i=1}^{n} \delta_{Z_i^*}$, $m \in \mathbb{N}$. Note that the sample $Z_1^*, \ldots, Z_n^*$ depends and on the blocklength $b(n)$ and on the number of blocks $\ell(n)$.

Further denote the joint distribution of $\mathbf{W}_{\mathbb{N}}$, $\mathbf{W}_{\mathbb{N}}^*$, and $\mathbf{W}'_{\mathbb{N}}$ by $K_{\mathbb{N}} \in \mathcal{M}(\mathcal{Z}^{\mathbb{N}} \times \mathcal{Z}^{\mathbb{N}} \times \mathcal{Z}^{\mathbb{N}})$. Then, $K_{\mathbb{N}}$ has marginal distributions $K_{\mathbb{N}}(B_1 \times \mathcal{Z}^{\mathbb{N}} \times \mathcal{Z}^{\mathbb{N}}) = P_{\mathbb{N}}(B_1)$ for all $B_1 \in \mathcal{B}^{\otimes\mathbb{N}}$, $K_{\mathbb{N}}(\mathcal{Z}^{\mathbb{N}} \times B_2 \times \mathcal{Z}^{\mathbb{N}}) = \overline{P}_{\mathbb{N}}(B_2)$ for all $B_2 \in \mathcal{B}^{\otimes\mathbb{N}}$, and $K_{\mathbb{N}}(\mathcal{Z}^{\mathbb{N}} \times \mathcal{Z}^{\mathbb{N}} \times B_3) = P_{\mathbb{N}}^*(B_3)$ for all $B_3 \in \mathcal{B}^{\otimes\mathbb{N}}$.
Then,

$$\mathcal{L}_{P_n}(S_n) = S_n(P_n) = S_n \circ \mathbf{W}_n(P_{\mathbb{N}}) \quad \text{and} \quad \mathcal{L}_{P_n^*}(S_n) = S_n(P_n^*) = S_n \circ \mathbf{W}'_n(P_{\mathbb{N}}^*)$$

and therefore

$$d_{\mathrm{BL}}(\mathcal{L}_{P_n^*}(S_n), \mathcal{L}_{P_n}(S_n)) = d_{\mathrm{BL}}(\mathcal{L}(S_n \circ W_n'), \mathcal{L}(S_n \circ W_n)).$$

By assumption we have $0 \leq z_i \leq 1$, $i \in \mathbb{N}$. Hence $Z_i(z_{\mathbb{N}}) = z_i \in [0,1]$, i. e. $\mathcal{Z} = [0,1]$, which is a totally bounded metric space. Therefore the set $\mathrm{BL}_1([0,1])$ is a uniform Glivenko-Cantelli class, due to Dudley et al. (1991, Proposition 12). Similar to part I of the proof of Theorem 3.4.2, the blockwise bootstrap structure and the Glivenko-Cantelli property yield:

$$\lim_{n\to\infty} \sup_{\mathbb{P}_{\mathbf{w}_n^*} \in \mathcal{M}(\mathcal{Z})} P_{\mathbb{N}}^* \left(\left\{ z_{\mathbb{N}} \in \mathcal{Z}^{\mathbb{N}} \mid \sup_{m \geq n} d_{\mathrm{BL}}(\mathbb{P}_{\mathbf{W}'_m(z_{\mathbb{N}})}, \mathbb{P}_{\mathbf{w}_n^*}) > \eta \right\} \right) = 0.$$

Respectively, for fixed $\varepsilon > 0$, for every $\delta_0 > 0$ there is $n_1 \in \mathbb{N}$ such that for all $n \geq n_1$ and all $\mathbb{P}_{\mathbf{w}_n^*} \in \mathcal{M}(\mathcal{Z})$:

$$P_n^* \left(\left\{ \mathbf{w}_n' \in \mathcal{Z}^n \mid d_{\mathrm{BL}}(\mathbb{P}_{\mathbf{w}_n'}, \mathbb{P}_{\mathbf{w}_n^*}) \leq \frac{\delta_0}{2} \right\} \right) \geq 1 - \frac{\varepsilon}{6}. \tag{3.45}$$

Regard the process $G_n(t) = \frac{1}{\sqrt{n}} \sum_{i=1}^n I_{\{Z_i^* \leq t\}} - \frac{1}{\sqrt{n}} \sum_{i=1}^n I_{\{Z_i \leq t\}}$, $t \in \mathbb{R}$. Due to the assumptions on the process and on the moving block bootstrap, Theorem 2.3 in Peligrad (1998) yields the almost sure convergence in distribution to a Brownian bridge G:

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n I_{\{Z_i^* \leq t\}} - \frac{1}{\sqrt{n}} \sum_{i=1}^n I_{\{Z_i \leq t\}} \longrightarrow_{\mathcal{D}} G(t), \quad t \in \mathbb{R} \tag{3.46}$$

almost surely with respect to $P_{\mathbb{N}}$, $n \to \infty$, in the Skorohod topology on $D[0,1]$. Here $\longrightarrow_{\mathcal{D}}$ indicates convergence in distribution and $D[0,1]$ denotes the space of cadlag functions on $[0,1]$, for details see for example Billingsley (1999, p. 121).

This is equivalent to

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n I_{\{Z_i^* \leq t\}} - \frac{1}{\sqrt{n}} \sum_{i=1}^n I_{\{Z_i \leq t\}} \longrightarrow_{\mathcal{D}} G(t), \text{ almost surely with respect to } P_{\mathbb{N}},\ n \to \infty,$$

for all continuity points t of G, see Billingsley (1999, (12.14), p. 124).

Multiplying by $\frac{1}{\sqrt{n}}$ yields for any fixed continuity point $t \in \mathbb{R}$:

$$\frac{1}{n} \sum_{i=1}^n I_{\{Z_i^* \leq t\}} - \frac{1}{n} \sum_{i=1}^n I_{\{Z_i \leq t\}} - \frac{1}{\sqrt{n}} G(t) \longrightarrow_{\mathcal{D}} 0 \text{ almost surely with respect to } P_{\mathbb{N}},\ n \to \infty.$$

As convergence in distribution to a finite constant implies convergence in probability, see for example van der Vaart (1998, Theorem 2.7(iii)), and as $\frac{1}{\sqrt{n}}G(t) \to 0$ in probability, for all $t \in \mathbb{R}$:

$$\frac{1}{n}\sum_{i=1}^{n} I_{\{Z_i^* \leq t\}} - \frac{1}{n}\sum_{i=1}^{n} I_{\{Z_i \leq t\}} \longrightarrow_P 0 \text{ almost surely with respect to } P_{\mathbb{N}},\ n \to \infty,$$

for all continuity points t of G, where $\longrightarrow_P$ denotes the convergence in probability.

Hence, Dudley (1989, Theorem 11.12) yields the convergence of the corresponding probability measures:

$$d_{\mathrm{BL}}\left(\frac{1}{n}\sum_{i=1}^{n}\delta_{Z_i^*}, \frac{1}{n}\sum_{i=1}^{n}\delta_{Z_i}\right) \longrightarrow_P 0 \text{ almost surely with respect to } P_{\mathbb{N}},\ n \to \infty.$$

Respectively

$$d_{\mathrm{BL}}(\mathbb{P}_{\mathbf{W}_n^*}, \mathbb{P}_{\mathbf{W}_n}) \longrightarrow_P 0 \text{ almost surely with respect to } P_{\mathbb{N}},\ n \to \infty.$$

Define the set $B_n = \left\{\mathbf{w}_n \in \mathcal{Z}^n \mid d_{\mathrm{BL}}(\mathbb{P}_{\mathbf{W}_n^*}, \mathbb{P}_{\mathbf{w}_n}) \longrightarrow_P 0,\ n \to \infty\right\}$. Hence,

$$P_n(B_n) = P_{\mathbb{N}}\left(\left\{z_{\mathbb{N}} \in \mathcal{Z}^{\mathbb{N}} \mid \mathbf{W}_n(z_{\mathbb{N}}) \in B_n\right\}\right) = 1 \tag{3.47}$$

and, for all $\mathbf{w}_n \in B_n$, there is $n_{2,\mathbf{w}_n} \in \mathbb{N}$ such that for all $n \geq n_{2,\mathbf{w}_n} \in \mathbb{N}$:

$$\overline{P}_n\left(\left\{\mathbf{w}_n^* \in \mathcal{Z}^n \mid d_{\mathrm{BL}}\left(\mathbb{P}_{\mathbf{w}_n^*}, \mathbb{P}_{\mathbf{w}_n}\right) > \frac{\delta_0}{4}\right\}\right) < \frac{\varepsilon}{6}. \tag{3.48}$$

By assumption we have $0 \leq z_i \leq 1$, $i \in \mathbb{N}$. Hence the space of probability measures $\{\mathbb{P}_{\mathbf{w}_n} \mid \mathbf{w}_n \in [0,1]^n\}$ is a subset of $\mathcal{M}([0,1])$ and therefore tight (see Definition A6), as [0,1] is a compact space, see e. g. (Klenke, 2013, Example 13.28). Then Prohorov's Theorem, see for example Billingsley (1999, Theorem 5.1) yields relative compactness of $\mathcal{M}([0,1], d_{\mathrm{BL}})$ and in particular the relative compactness of the set $\{\mathbb{P}_{\mathbf{w}_n} \mid \mathbf{w}_n \in [0,1]^n\}$. As $\mathcal{M}([0,1], d_{\mathrm{BL}})$ is a complete space, see Dudley (1989, Theorem 11.5.5), relative compactness equals total boundedness. That is, there exists a finite dense subset $\tilde{\mathcal{P}}$ of $\{\mathbb{P}_{\mathbf{w}_n} \mid \mathbf{w}_n \in [0,1]^n\}$ such that for all $\rho > 0$ and $\mathbb{P}_{\mathbf{w}_n} \in \{\mathbb{P}_{\mathbf{w}_n} \mid \mathbf{w}_n \in [0,1]^n\}$ there is $\tilde{P}_\rho \in \tilde{\mathcal{P}}$ such that

$$d_{\mathrm{BL}}(\tilde{P}_\rho, \mathbb{P}_{\mathbf{w}_n}) \leq \rho. \tag{3.49}$$

The triangle inequality yields:

$$d_{\mathrm{BL}}\left(\mathbb{P}_{\mathbf{w}_n^*}, \mathbb{P}_{\mathbf{w}_n}\right) \leq d_{\mathrm{BL}}\left(\mathbb{P}_{\mathbf{w}_n^*}, \tilde{P}_\rho\right) + d_{\mathrm{BL}}\left(\tilde{P}^\rho, \mathbb{P}_{\mathbf{w}_n}\right).$$

Define $\rho = \frac{\delta_0}{4}$. Then (3.48) yields for every $\tilde{P}_\rho \in \tilde{\mathcal{P}}$ the existence of an integer $n \geq n_{2,\tilde{P}} \in \mathbb{N}$ such that, for all $n \geq n_{2,\tilde{P}}$ and all $\mathbf{w}_n \in B_n$:

$$\begin{aligned}
&\overline{P}_n\left(\left\{\mathbf{w}_n^* \in \mathcal{Z}^n \mid d_{\mathrm{BL}}\left(\mathbb{P}_{\mathbf{w}_n^*}, \mathbb{P}_{\mathbf{w}_n}\right) > \frac{\delta_0}{2}\right\}\right) \\
&\quad \leq \ \overline{P}_n\left(\left\{\mathbf{w}_n^* \in \mathcal{Z}^n \mid d_{\mathrm{BL}}\left(\mathbb{P}_{\mathbf{w}_n^*}, \tilde{P}_\rho\right) > \frac{\delta_0}{4} \text{ or } d_{\mathrm{BL}}\left(\tilde{P}_\rho, \mathbb{P}_{\mathbf{w}_n}\right) > \frac{\delta_0}{4}\right\}\right) \\
&\quad \overset{(3.49)}{\leq} \overline{P}_n\left(\left\{\mathbf{w}_n^* \in \mathcal{Z}^n \mid d_{\mathrm{BL}}\left(\mathbb{P}_{\mathbf{w}_n^*}, \tilde{P}_\rho\right) > \frac{\delta_0}{4}\right\}\right) \overset{(3.48)}{<} \frac{\varepsilon}{6}.
\end{aligned}$$

Hence, for all $n \geq n_2 := \max_{\tilde{P}\in\tilde{\mathcal{P}}}\{n_{2,\tilde{P}}\}$ and for all $\mathbf{w}_n \in B_n$, we have:

$$\sup_{\mathbb{P}_{\mathbf{w}_n} \in \mathcal{M}(\mathcal{Z})} \overline{P}_n\left(\left\{\mathbf{w}_n^* \in \mathcal{Z}^n \mid d_{\mathrm{BL}}\left(\mathbb{P}_{\mathbf{w}_n^*}, \mathbb{P}_{\mathbf{w}_n}\right) > \frac{\delta_0}{2}\right\}\right) < \frac{\varepsilon}{6}. \tag{3.50}$$

Due to the uniform continuity of the operator S, for every $\varepsilon > 0$ there is $\delta_0 > 0$ such that for all $P, Q \in \mathcal{M}(\mathcal{Z})$:

$$d_{\mathrm{BL}}(P, Q) \leq \delta_0 \quad \Rightarrow \quad d_H(S(P), S(Q)) \leq \frac{\varepsilon}{3}. \tag{3.51}$$

Moreover, the triangle inequality yields:

$$d_{\mathrm{BL}}(\mathbb{P}_{\mathbf{w}'_n}, \mathbb{P}_{\mathbf{w}_n}) \leq d_{\mathrm{BL}}(\mathbb{P}_{\mathbf{w}'_n}, \mathbb{P}_{\mathbf{w}_n^*}) + d_{\mathrm{BL}}(\mathbb{P}_{\mathbf{w}_n^*}, \mathbb{P}_{\mathbf{w}_n}). \tag{3.52}$$

Again we use the relation between the Prohorov metric π_{d_H} and the Ky Fan metric, Dudley (1989, Theorem 11.3.5):

$$\begin{aligned}
\pi_{d_H}\left(\mathcal{L}_{P_n^*}(S_n),\ \mathcal{L}_{P_n}(S_n)\right) &= \pi_{d_H}(S_n \circ \mathbf{W}'_n, S_n \circ \mathbf{W}_n) \\
&\leq \inf\left\{\tilde{\varepsilon} > 0 \mid K_{\mathbb{N}}\Big(\left\{\ d_H(S_n \circ \mathbf{W}'_n, S_n \circ \mathbf{W}_n) > \tilde{\varepsilon}, \mathbf{w}_{\mathbb{N}}^* \in \mathcal{Z}^{\mathbb{N}}\right\}\Big) \leq \tilde{\varepsilon}\right\} \\
&= \inf\left\{\tilde{\varepsilon} > 0 \mid (\mathbf{W}_n, \mathbf{W}_n^*, \mathbf{W}'_n)(K_{\mathbb{N}})\left(\left\{(\mathbf{w}_n, \mathbf{w}_n^*, \mathbf{w}'_n) \in \mathcal{Z}^n \times \mathcal{Z}^n \times \mathcal{Z}^n \mid\right.\right.\right. \\
&\qquad\qquad \left.\left.\left. d_H(S_n(\mathbf{w}'_n), S_n(\mathbf{w}_n)) > \tilde{\varepsilon}, \mathbf{w}_n^* \in \mathcal{Z}^n\right\}\right) \leq \tilde{\varepsilon}\right\}.
\end{aligned}$$

Due to the definition of the statistical operator S, this is equivalent to

$$\inf\{\tilde{\varepsilon} > 0 \mid (\mathbf{W}_n, \mathbf{W}_n^*, \mathbf{W}_n')(K_{\mathbb{N}}) \left(\left\{(\mathbf{w}_n, \mathbf{w}_n^*, \mathbf{w}_n') \in \mathcal{Z}^n \times \mathcal{Z}^n \times \mathcal{Z}^n \mid \right.\right.$$
$$\left.\left. d_H(S(\mathbb{P}_{\mathbf{w}_n'}), S(\mathbb{P}_{\mathbf{w}_n}) > \tilde{\varepsilon}, \mathbf{w}_n^* \in \mathcal{Z}^n\right\}\right) \leq \tilde{\varepsilon}\}.$$

Due to the uniform continuity of S, see (3.51), we obtain, for all $n \geq \max\{n_1, n_2\}$:

$$\begin{aligned}
&(\mathbf{W}_n, \mathbf{W}_n^*, \mathbf{W}_n')(K_{\mathbb{N}}) \left(\left\{(\mathbf{w}_n, \mathbf{w}_n^*, \mathbf{w}_n') \in \mathcal{Z}^n \times \mathcal{Z}^n \times \mathcal{Z}^n \mid d_H(S(\mathbb{P}_{\mathbf{w}_n'}), S(\mathbb{P}_{\mathbf{w}_n})) > \frac{\varepsilon}{3}, \mathbf{w}_n^* \in \mathcal{Z}^n\right\}\right) \\
&\overset{(3.51)}{\leq} (\mathbf{W}_n, \mathbf{W}_n^*, \mathbf{W}_n')(K_{\mathbb{N}}) \left(\left\{(\mathbf{w}_n, \mathbf{w}_n^*, \mathbf{w}_n') \in \mathcal{Z}^n \times \mathcal{Z}^n \times \mathcal{Z}^n \mid d_{\mathrm{BL}}(\mathbb{P}_{\mathbf{w}_n'}, \mathbb{P}_{\mathbf{w}_n}) > \delta_0, \mathbf{w}_n^* \in \mathcal{Z}^n\right\}\right) \\
&= (\mathbf{W}_n, \mathbf{W}_n^*, \mathbf{W}_n')(K_{\mathbb{N}}) \left(\left\{(\mathbf{w}_n, \mathbf{w}_n^*, \mathbf{w}_n') \in \mathcal{Z}^n \times \mathcal{Z}^n \times \mathcal{Z}^n \mid \right.\right. \\
&\qquad \left.\left. \{\mathbf{w}_n \notin B_n,\ d_{\mathrm{BL}}(\mathbb{P}_{\mathbf{w}_n'}, \mathbb{P}_{\mathbf{w}_n}) > \delta_0\} \text{ or } \{\mathbf{w}_n \in B_n,\ d_{\mathrm{BL}}(\mathbb{P}_{\mathbf{w}_n'}, \mathbb{P}_{\mathbf{w}_n}) > \delta_0\}, \mathbf{w}_n^* \in \mathcal{Z}^n\right\}\right) \\
&\leq (\mathbf{W}_n, \mathbf{W}_n^*, \mathbf{W}_n')(K_{\mathbb{N}}) \left(\left\{(\mathbf{w}_n, \mathbf{w}_n^*, \mathbf{w}_n') \in \mathcal{Z}^n \times \mathcal{Z}^n \times \mathcal{Z}^n \mid \right.\right. \\
&\qquad \left.\left. \mathbf{w}_n \notin B_n,\ d_{\mathrm{BL}}(\mathbb{P}_{\mathbf{w}_n'}, \mathbb{P}_{\mathbf{w}_n}) > \delta_0, \mathbf{w}_n^* \in \mathcal{Z}^n\right\}\right) \\
&\quad + (\mathbf{W}_n, \mathbf{W}_n^*, \mathbf{W}_n')(K_{\mathbb{N}}) \left(\left\{(\mathbf{w}_n, \mathbf{w}_n^*, \mathbf{w}_n') \in \mathcal{Z}^n \times \mathcal{Z}^n \times \mathcal{Z}^n \mid \right.\right. \\
&\qquad \left.\left. \mathbf{w}_n \in B_n,\ d_{\mathrm{BL}}(\mathbb{P}_{\mathbf{w}_n'}, \mathbb{P}_{\mathbf{w}_n}) > \delta_0, \mathbf{w}_n^* \in \mathcal{Z}^n\right\}\right) \\
&\overset{(3.47)}{=} (\mathbf{W}_n, \mathbf{W}_n^*, \mathbf{W}_n')(K_{\mathbb{N}}) \left(\left\{(\mathbf{w}_n, \mathbf{w}_n^*, \mathbf{w}_n') \in \mathcal{Z}^n \times \mathcal{Z}^n \times \mathcal{Z}^n \mid \right.\right. \\
&\qquad \left.\left. \mathbf{w}_n \in B_n,\ d_{\mathrm{BL}}(\mathbb{P}_{\mathbf{w}_n'}, \mathbb{P}_{\mathbf{w}_n}) > \delta_0, \mathbf{w}_n^* \in \mathcal{Z}^n\right\}\right).
\end{aligned}$$

The triangle inequality, (3.52), then yields for all $n \geq \max\{n_1, n_2\}$:

$$\begin{aligned}
&(\mathbf{W}_n, \mathbf{W}_n^*, \mathbf{W}_n')(K_{\mathbb{N}}) \left(\left\{(\mathbf{w}_n, \mathbf{w}_n^*, \mathbf{w}_n') \in \mathcal{Z}^n \times \mathcal{Z}^n \times \mathcal{Z}^n | \ \mathbf{w}_n \in B_n, d_{\mathrm{BL}}(\mathbb{P}_{\mathbf{w}_n'}, \mathbb{P}_{\mathbf{w}_n}) > \delta_0, \mathbf{w}_n^* \in \mathcal{Z}^n\right\}\right) \\
&\overset{(3.52)}{\leq} (\mathbf{W}_n, \mathbf{W}_n^*, \mathbf{W}_n')(K_{\mathbb{N}}) \left(\left\{(\mathbf{w}_n, \mathbf{w}_n^*, \mathbf{w}_n') \in \mathcal{Z}^n \times \mathcal{Z}^n \times \mathcal{Z}^n \mid \{\mathbf{w}_n \in B_n \right.\right. \\
&\qquad \left.\left. \text{and } d_{\mathrm{BL}}(\mathbb{P}_{\mathbf{w}_n'}, \mathbb{P}_{\mathbf{w}_n^*}) > \frac{\delta_0}{2}\} \text{ or } \{\mathbf{w}_n \in B_n \text{ and } d_{\mathrm{BL}}(\mathbb{P}_{\mathbf{w}_n^*}, \mathbb{P}_{\mathbf{w}_n}) > \frac{\delta_0}{2}\}\right\}\right) \\
&\leq P_n^* \left(\left\{\mathbf{w}_n' \in \mathcal{Z}^n \mid \mathbf{w}_n \in \mathcal{B}_n,\ d_{\mathrm{BL}}(\mathbb{P}_{\mathbf{w}_n'}, \mathbb{P}_{\mathbf{w}_n^*}) > \frac{\delta_0}{2}\right\}\right) \\
&\qquad + \overline{P}_n \left(\left\{\mathbf{w}_n^* \in \mathcal{Z}^n \mid \mathbf{w}_n \in \mathcal{B}_n,\ d_{\mathrm{BL}}(\mathbb{P}_{\mathbf{w}_n^*}, \mathbb{P}_{\mathbf{w}_n}) > \frac{\delta_0}{2}\right\}\right) \\
&\overset{(3.45),(3.48)}{<} \frac{\varepsilon}{6} + \frac{\varepsilon}{6} \quad = \quad \frac{\varepsilon}{3}.
\end{aligned}$$

The equivalence between the Prohorov metric and the bounded Lipschitz metric on Polish spaces, see Huber (1981, Chapter 2, Corollary 4.3), yields the existence of $\tilde{n}_1$ such that for

every $n \geq \tilde{n}_1$:

$$d_{\mathrm{BL}}(\mathcal{L}_{P_n^*}(S_n), \mathcal{L}_{P_n}(S_n)) < \frac{\varepsilon}{3}.$$

And therefore

$$\mathbb{E}\left[d_{\mathrm{BL}}\left(\mathcal{L}_{P_n^*}(S_n), \mathcal{L}_{P_n}(S_n)\right)\right] < \frac{\varepsilon}{3}. \tag{3.53}$$

For the convergence of the term in part III the same argumentation as for part I can be applied, as the assumptions on $Q_{\mathbb{N}}$ and $Q_{\mathbb{N}}^*$ are the same as for $P_{\mathbb{N}}$ and $\mathbb{P}_{\mathbb{N}}^*$. In particular for every $\varepsilon > 0$ there is $\tilde{n}_2 \in \mathbb{N}$ such that for all $n \geq \tilde{n}_2$:

$$d_{\mathrm{BL}}\left(\mathcal{L}_{Q_n^*}(S_n), \mathcal{L}_{Q_n}(S_n)\right) < \frac{\varepsilon}{3},$$

respectively

$$\mathbb{E}\left[d_{\mathrm{BL}}\left(\mathcal{L}_{Q_n^*}(S_n), \mathcal{L}_{Q_n}(S_n)\right)\right] < \frac{\varepsilon}{3}. \tag{3.54}$$

Hence, (3.44), (3.53), and (3.54) yield, for all $n \geq \max\{\tilde{n}_1, \tilde{n}_2\}$:

$$\mathbb{E}\left[d_{\mathrm{BL}}\left(\mathcal{L}_{P_n^*}(S_n), \mathcal{L}_{Q_n^*}(S_n)\right)\right] < \frac{\varepsilon}{3} + \frac{\varepsilon}{3} + \frac{\varepsilon}{3} = \varepsilon.$$

As $\mathcal{L}_{P_n^*}(S_n)$ and $\mathcal{L}_{Q_n^*}(S_n)$ are random variables itself we have, due to Huber (1981, Chapter 2 Theorem 4.2, (2)⇒(1)), for all $n \geq \max\{\tilde{n}_1, \tilde{n}_2\}$:

$$d_{\mathrm{BL}}\left(\mathcal{L}(\mathcal{L}_{P_n^*}(S_n)), \mathcal{L}(\mathcal{L}_{Q_n^*}(S_n))\right) < \varepsilon.$$

Hence, for all $\varepsilon > 0$ there is $\delta > 0$ such that there is $n_0 = \max\{\tilde{n}_1, \tilde{n}_2\} \in \mathbb{N}$ such that, for all $n \geq n_0$:

$$d_{\mathrm{BL}}(P_n, Q_n) < \delta \;\Rightarrow\; d_{\mathrm{BL}}(\mathcal{L}(\mathcal{L}_{P_n^*}(S_n)), \mathcal{L}(\mathcal{L}_{Q_n^*}(S_n))) < \varepsilon$$

and therefore the assertion. □

The next theorem generalizes this result to stochastic processes with values in $[0,1]^d$, $d > 1$, instead of $[0,1] \subset \mathbb{R}$. Therefore, for example, the bootstrap version of the SVM estimator is qualitatively robust under weak conditions. The proof of the next theorem follows the same lines as the proof of the theorem above, but another central limit theorem, which is shown in Bühlmann (1994), is used. Therefore the assumptions on the mixing property of the stochastic process are stronger and the random variables Z_i, $i \in \mathbb{N}$, are assumed to

have continuous marginal distributions. Again the bootstrap sample results of a moving block bootstrap where $\ell(n)$ blocks of length $b(n)$ are chosen, again assuming $\ell(n) \cdot b(n) = n$. Moreover, let $b(n)$ be a sequences of integers satisfying

$$b(n) = \mathcal{O}(n^{\frac{1}{2}-a}) \text{ for some } a > 0.$$

Theorem 3.4.6 *Assume $\mathcal{Z} = [0,1]^d$, $d > 1$. Let $P_{\mathbb{N}}$ be a probability measure such that the coordinate process $(Z_i)_{i\in\mathbb{N}}$, $Z_i : \mathcal{Z}^{\mathbb{N}} \to \mathcal{Z}$, is strongly stationary and α-mixing with*

$$\sum_{m=0}^{\infty} (m+1)^{8d+7} (\alpha(\sigma(Z_1,\ldots,Z_i), \sigma(Z_{i+m},\ldots), P_{\mathbb{N}}))^{\frac{1}{2}} < \infty, \; i \in \mathbb{N}. \tag{3.55}$$

Assume that Z_i has continuous marginal distributions for all $i \in \mathbb{N}$. Define the set of probability measures $\mathcal{P} \subset \mathcal{M}(\mathcal{Z})$ such that the coordinate process is strongly stationary and α-mixing as in (3.55).
Let (H, d_H) be a complete separable metric space, $(S_n)_{n\in\mathbb{N}}$ be a sequence of estimators such that $S_n : \mathcal{Z}^n \to H$ is continuous and assume that S_n can be represented by a statistical operator $S : \mathcal{M}(\mathcal{Z}) \to H$ via (3.1) *which is additionally uniformly continuous with respect to d_{BL}.*

Then the sequence of estimators $(S_n)_{n\in\mathbb{N}}$ is qualitatively robust at $P_{\mathbb{N}}$ with respect to $\mathcal{P}$.

Proof of Theorem 3.4.6: The proof follows the same lines as the proof of Theorem 3.4.5 and therefore we only state the different steps. Again we start with the triangle inequality:

$$\begin{aligned} &d_{\mathrm{BL}}(\mathcal{L}_{P_n^*}(S_n), \mathcal{L}_{Q_n^*}(S_n)) \\ &\leq \underbrace{d_{\mathrm{BL}}(\mathcal{L}_{P_n^*}(S_n), \mathcal{L}_{P_n}(S_n))}_{I} + \underbrace{d_{\mathrm{BL}}(\mathcal{L}_{P_n}(S_n), \mathcal{L}_{Q_n}(S_n))}_{II} + \underbrace{d_{\mathrm{BL}}(\mathcal{L}_{Q_n}(S_n), \mathcal{L}_{Q_n^*}(S_n))}_{III}. \end{aligned}$$

To proof the convergence of the term in part II, we need the weak Varadarajan property of the stochastic process. Due to the definition $\alpha(\sigma(Z_1,\ldots,Z_i), \sigma(Z_{i+\ell},\ldots), \mu) \leq 2$ for all $\ell \in \mathbb{N}$, $i \in \mathbb{N}$, and obviously:

$$\alpha(\sigma(Z_1,\ldots,Z_i), \sigma(Z_{i+\ell},\ldots), P_{\mathbb{N}}) \leq \ell + 1, \; \ell > 0. \tag{3.56}$$

Hence, due to the strong stationarity of the stochastic process, we have:

$$\begin{aligned}
\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha((Z_i)_{i\in\mathbb{N}},P_{\mathbb{N}},i,j) &= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha(\sigma(Z_i),\sigma(Z_j),P_{\mathbb{N}})\\
&\leq \frac{2}{n^2}\sum_{i=1}^{n}\sum_{j\geq i}^{n}\alpha(\sigma(Z_i),\sigma(Z_j),P_{\mathbb{N}})\\
&\leq \frac{2}{n^2}\sum_{i=1}^{n}\sum_{j\geq i}^{n}\alpha(\sigma(Z_1,\ldots,Z_i),\sigma(Z_j,\ldots),P_{\mathbb{N}})\\
&= \frac{2}{n^2}\sum_{i=1}^{n}\sum_{\ell=0}^{n-i}\alpha(\sigma(Z_1,\ldots,Z_i),\sigma(Z_{i+\ell},\ldots),P_{\mathbb{N}})\\
&\overset{\text{stationarity}}{\leq} \frac{2}{n}\sum_{\ell=0}^{n}\alpha(\sigma(Z_1,\ldots,Z_i),\sigma(Z_{i+\ell},\ldots),P_{\mathbb{N}}),\ i\in\mathbb{N}\\
&= \frac{2}{n}\sum_{\ell=0}^{n}(\alpha(\sigma(Z_1,\ldots,Z_i),\sigma(Z_{i+\ell},\ldots),P_{\mathbb{N}}))^{\frac{1}{2}}\,(\alpha(\sigma(Z_1,\ldots,Z_i),\sigma(Z_{i+\ell},\ldots),P_{\mathbb{N}}))^{\frac{1}{2}},\ i\in\mathbb{N}\\
&\overset{(3.56)}{\leq} \frac{2}{n}\sum_{\ell=0}^{n}(\ell+1)\,(\alpha(\sigma(Z_1,\ldots,Z_i),\sigma(Z_{i+\ell},\ldots),P_{\mathbb{N}}))^{\frac{1}{2}},\ i\in\mathbb{N}\\
&\overset{(3.55)}{\longrightarrow} 0,\ n\to\infty.
\end{aligned}$$

Now, the same argumentation as in the proof of Theorem 3.4.5 yields the weak Varadarajan property and therefore, for all $\varepsilon > 0$,

$$\mathbb{E}\left[d_{\mathrm{BL}}(\mathcal{L}_{P_n}(S_n),\mathcal{L}_{Q_n}(S_n))\right] < \frac{\varepsilon}{3}. \tag{3.57}$$

Regarding the term in part I, we use a central limit theorem for the blockwise bootstrapped empirical process by Bühlmann (1994, Corollary 1 and remark) to show its convergence. Again, regard the distribution $P_{\mathbb{N}} \in \mathcal{M}(\mathcal{Z}^{\mathbb{N}})$ and let $P_{\mathbb{N}}^*$ be the bootstrap approximation of $P_{\mathbb{N}}$, via the blockwise bootstrap. Define, for all $n \in \mathbb{N}$, the random variables
$\mathbf{W}_n : \mathcal{Z}^{\mathbb{N}} \to \mathcal{Z}^n,\ \mathbf{W}_n = (Z_1,\ldots,Z_n),\ z_{\mathbb{N}} \mapsto \mathbf{w}_n$, and
$\mathbf{W}'_n : \mathcal{Z}^{\mathbb{N}} \to \mathcal{Z}^n,\ \mathbf{W}'_n = (Z'_1,\ldots,Z'_n),\ z_{\mathbb{N}} \mapsto \mathbf{w}'_n$,
such that $\mathbf{W}_n(P_{\mathbb{N}}) = P_n$ and $\mathbf{W}'_n(P_{\mathbb{N}}^*) = P_n^*$.
Moreover denote the bootstrap sample by $\mathbf{W}_n^* : \mathcal{Z}^{\mathbb{N}} \to \mathcal{Z}^n,\ \mathbf{W}_n^* := (Z_1^*,\ldots,Z_n^*),\ z_{\mathbb{N}} \mapsto \mathbf{w}_n^*$, and the distribution of $\mathbf{W}_n^*$ by $\overline{P}_n$. The bootstrap approximation of P_m is $P_m^* = \otimes_{j=1}^m \frac{1}{n}\sum_{i=1}^n \delta_{Z_i^*} = \otimes_{j=1}^m \mathbb{P}_{\mathbf{W}_n^*}$, $m \in \mathbb{N}$, by definition of the bootstrap procedure. Note that

the sample $Z_1^*, \ldots, Z_n^*$ depends and on the blocklength $b(n)$ and on the number of blocks $\ell(n)$.

Further denote the joint distribution of $\mathbf{W}_{\mathbb{N}}$, $\mathbf{W}_{\mathbb{N}}^*$, and $\mathbf{W}'_{\mathbb{N}}$ by $K_{\mathbb{N}} \in \mathcal{M}(\mathcal{Z}^{\mathbb{N}} \times \mathcal{Z}^{\mathbb{N}} \times \mathcal{Z}^{\mathbb{N}})$. Then, $K_{\mathbb{N}}$ has marginal distributions $K_{\mathbb{N}}(B_1 \times \mathcal{Z}^{\mathbb{N}} \times \mathcal{Z}^{\mathbb{N}}) = P_{\mathbb{N}}(B_1)$ for all $B_1 \in \mathcal{B}^{\otimes\mathbb{N}}$, $K_{\mathbb{N}}(\mathcal{Z}^{\mathbb{N}} \times B_2 \times \mathcal{Z}^{\mathbb{N}}) = \overline{P}_{\mathbb{N}}(B_2)$ for all $B_2 \in \mathcal{B}^{\otimes\mathbb{N}}$, and $K_{\mathbb{N}}(\mathcal{Z}^{\mathbb{N}} \times \mathcal{Z}^{\mathbb{N}} \times B_3) = P_{\mathbb{N}}^*(B_3)$ for all $B_3 \in \mathcal{B}^{\otimes\mathbb{N}}$.

Then,

$$\mathcal{L}_{P_n}(S_n) = S_n(P_n) = S_n \circ \mathbf{W}_n(P_{\mathbb{N}}) \quad \text{and} \quad \mathcal{L}_{P_n^*}(S_n) = S_n(P_n^*) = S_n \circ \mathbf{W}'_n(P_{\mathbb{N}}^*)$$

and therefore

$$d_{\mathrm{BL}}(\mathcal{L}_{P_n^*}(S_n), \mathcal{L}_{P_n}(S_n)) = d_{\mathrm{BL}}(\mathcal{L}(S_n \circ W'_n), \mathcal{L}(S_n \circ W_n)).$$

As $\mathcal{Z} = [0,1]^d$ is compact, it is in particular totally bounded. Hence the set $\mathrm{BL}_1(\mathcal{Z}, d_{\mathcal{Z}})$ is a uniform Glivenko-Cantelli class, due to Dudley et al. (1991, Proposition 12). Similar to part I of the proof of Theorem 3.4.5, the bootstrap structure and the Glivenko-Cantelli property given above yield for arbitrary, but fixed $\varepsilon > 0$:
for every $\delta_0 > 0$ there is $n_0 \in \mathbb{N}$ such that, for all $n \geq n_0$ and all $\mathbb{P}_{\mathbf{w}_n^*} \in \mathcal{M}(\mathcal{Z})$,

$$P_n^*\left(\left\{\mathbf{w}'_n \in \mathcal{Z}^n \mid d_{\mathrm{BL}}(\mathbb{P}_{\mathbf{w}'_n}, \mathbb{P}_{\mathbf{w}_n^*}) \leq \frac{\delta_0}{2}\right\}\right) \geq 1 - \frac{\varepsilon}{6}.$$

Now, regard the empirical process of $(Z_1, \ldots, Z_n)$. Set $\mathbf{t} = (t_1, \ldots, t_d) \in \mathbb{R}^d$. Moreover $\mathbf{t} < \mathbf{b}$ means $t_i < b_i$ for all $i \in \{1, \ldots, d\}$. Hence we can define the empirical process and the blockwise bootstrapped empirical process by

$$\frac{1}{n}\sum_{i=1}^n I_{\{Z_i \leq \mathbf{t}\}} \quad \text{and} \quad \frac{1}{n}\sum_{i=1}^n I_{\{Z_i^* \leq \mathbf{t}\}}.$$

Regard the process $G_n(\mathbf{t}) = \frac{1}{\sqrt{n}}\sum_{i=1}^n I_{\{Z_i^* \leq \mathbf{t}\}} - \frac{1}{\sqrt{n}}\sum_{i=1}^n I_{\{Z_i \leq \mathbf{t}\}}$, $\mathbf{t} \in [0,1]^d$. Now, due to the assumptions on the stochastic process and on the moving block bootstrap, Bühlmann (1994, Corollary 1 and remark) yields the almost sure convergence in distribution to a Gaussian process G:

$$\frac{1}{\sqrt{n}}\sum_{i=1}^n I_{\{Z_i^* \leq \mathbf{t}\}} - \frac{1}{\sqrt{n}}\sum_{i=1}^n I_{\{Z_i \leq \mathbf{t}\}} \longrightarrow_{\mathcal{D}} G(\mathbf{t}), \quad \mathbf{t} \in [0,1]^d,$$

almost surely with respect to $P_{\mathbb{N}}$, $n \to \infty$, in the (extended) Skorohod topology on $D^d([0,1])$. The space $D^d([0,1])$ is a generalization of the space of cadlag functions on $[0,1]$, see Billingsley (1999, Chapter 12), and consists of functions $f : [0,1]^d \to \mathbb{R}$. A detailed description of this space and the extended Skorohod topology can be found in Straf (1972, 1969a) and Bickel and Wichura (1971). The definition of the space $D^d([0,1])$ can, for example, be found in Bickel and Wichura (1971, Chapter 3).

Straf (1972, Lemma 5.4) yields, that the above convergence in the Skorohod topology is equivalent to the convergence for all continuity points $\mathbf{t}$ of G. Hence,

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} I_{\{Z_i^* \leq \mathbf{t}\}} - \frac{1}{\sqrt{n}} \sum_{i=1}^{n} I_{\{Z_i \leq \mathbf{t}\}} \longrightarrow_{\mathcal{D}} G(\mathbf{t}) \text{ almost surely with respect to } P_{\mathbb{N}},\ n \to \infty,$$

for all continuity points $\mathbf{t}$ of G.

Multiplying by $\frac{1}{\sqrt{n}}$ yields, for every continuity point $\mathbf{t}$ of G,

$$\frac{1}{n} \sum_{i=1}^{n} I_{\{Z_i^* \leq \mathbf{t}\}} - \frac{1}{n} \sum_{i=1}^{n} I_{\{Z_i \leq \mathbf{t}\}} - \frac{1}{\sqrt{n}} G(\mathbf{t}) \longrightarrow_{\mathcal{D}} 0 \text{ almost surely with respect to } P_{\mathbb{N}},\ n \to \infty.$$

As convergence in distribution to a constant implies convergence in probability, see e. g. van der Vaart (1998, Theorem 2.7(iii)) and as $\frac{1}{\sqrt{n}} G(\mathbf{t})$ converges in probability to 0, for all fixed continuity points $\mathbf{t} \in [0,1]^d$ of G:

$$\frac{1}{n} \sum_{i=1}^{n} I_{\{Z_i^* \leq \mathbf{t}\}} - \frac{1}{n} \sum_{i=1}^{n} I_{\{Z_i \leq \mathbf{t}\}} \longrightarrow_{P} 0 \text{ almost surely with respect to } P_{\mathbb{N}},\ n \to \infty.$$

This yields the convergence of the corresponding probability measures, see for example Billingsley (1995, Chapter 29) for a theory on $\mathbb{R}^d$:

$$d_{\mathrm{BL}}\left(\frac{1}{n} \sum_{i=1}^{n} \delta_{Z_i^*}, \frac{1}{n} \sum_{i=1}^{n} \delta_{Z_i}\right) \longrightarrow_{P} 0 \text{ almost surely with respect to } P_{\mathbb{N}},\ n \to \infty,$$

respectively

$$d_{\mathrm{BL}}(\mathbb{P}_{\mathbf{W}_n^*}, \mathbb{P}_{\mathbf{W}_n}) \longrightarrow_{P} 0 \text{ almost surely with respect to } P_{\mathbb{N}},\ n \to \infty.$$

As the space $[0,1]^d$ is compact, we can use an argumentation similar to the proof of Theorem

3.4.5. Then, for every $\varepsilon > 0$, there is $n_1 \in \mathbb{N}$ such that for all $n \geq n_1$

$$d_{\mathrm{BL}}\left(\mathcal{L}_{P_n^*}(S_n), \mathcal{L}_{P_n}(S_n)\right) < \frac{\varepsilon}{3},$$

respectively,

$$\mathbb{E}\left[d_{\mathrm{BL}}\left(\mathcal{L}_{P_n^*}(S_n), \mathcal{L}_{P_n}(S_n)\right)\right] < \frac{\varepsilon}{3}. \tag{3.58}$$

The convergence of the term in part III follows simultaneously to part I for the distributions $Q_{\mathbb{N}}$ and $Q_{\mathbb{N}}^*$. Hence, for every $\varepsilon > 0$, there is $n_2 \in \mathbb{N}$ such that for all $n \geq n_2$

$$\mathbb{E}\left[d_{\mathrm{BL}}\left(\mathcal{L}_{Q_n^*}(S_n), \mathcal{L}_{Q_n}(S_n)\right)\right] < \frac{\varepsilon}{3}. \tag{3.59}$$

The combination of (3.57), (3.58), and (3.59) yields for all $n \geq \max\{n_1, n_2\}$:

$$\mathbb{E}\left[d_{\mathrm{BL}}\left(\mathcal{L}_{P_n^*}(S_n), \mathcal{L}_{Q^*}(S_n)\right)\right] < \frac{\varepsilon}{3} + \frac{\varepsilon}{3} + \frac{\varepsilon}{3} = \varepsilon.$$

As $\mathcal{L}_{P_n^*}(S_n)$ and $\mathcal{L}_{Q_n^*}(S_n)$ are random variables itself we have, due to Huber (1981, Chapter 2, Theorem 4.2, (2)⇒(1)), for all $n \geq \max\{n_1, n_2\}$:

$$d_{\mathrm{BL}}\left(\mathcal{L}(\mathcal{L}_{P_n^*}(S_n)), \mathcal{L}(\mathcal{L}_{Q_n^*}(S_n))\right) < \varepsilon.$$

Hence, for all $\varepsilon > 0$ there is $\delta > 0$ such that there is $n_0 = \max\{n_1, n_2\} \in \mathbb{N}$ such that for all $n \geq n_0$:

$$d_{\mathrm{BL}}(P_n, Q_n) < \delta \;\Rightarrow\; d_{\mathrm{BL}}(\mathcal{L}(\mathcal{L}_{P_n^*}(S_n)), \mathcal{L}(\mathcal{L}_{Q_n^*}(S_n))) < \varepsilon.$$

This yields the assertion. □

Although the assumptions on the statistical operator S, compared to Theorem 3.4.2, were strengthened in order to generalize the qualitative robustness to α-mixing sequences in Theorem 3.4.6 and 3.4.5, the M-estimators introduced in Chapter 3.3 are still an example for qualitative robust estimators if the sample space $(\mathcal{Z}, d_{\mathcal{Z}})$, $\mathcal{Z} \subset \mathbb{R}$ is compact. The compactness of $(\mathcal{Z}, d_{\mathcal{Z}})$ implies the compactness of the space $(\mathcal{M}(\mathcal{Z}), d_{\mathrm{BL}})$, see Parthasarathy (1967, Theorem 6.4). As the statistical operator S is continuous, the compactness of $\mathcal{M}(\mathcal{Z})$ implies the uniform continuity of S. Another example of M-estimators which are uniformly continuous even if the input space is not compact is given in Cuevas and Romo (1993, Theorem 4). Chapter 4.2.1 shows, that the SVM estimator is still qualitatively robust for the empirical bootstrap. If the space $\mathcal{Z} = \mathcal{X} \times \mathcal{Y} \subset \mathbb{R}^d$, $d \geq 1$, is compact, the same holds for the blockwise bootstrap for the given α-mixing sequences.

Chapter 4

Support vector machines

The following chapter contains robustness and consistency results concerning support vector machines. First we give a short introduction to SVMs, the ensuing section contains robustness and the last sections gives the consistency result. Again, if nothing else is stated we consider Borel σ-algebras throughout this chapter.

4.1 A short introduction to support vector machines

In recent years statistical machine learning and hence support vector machines became more and more important. A lot of introductory literature on support vector machines is available, for example Vapnik (1995, 1998) and Schölkopf and Smola (2002), Cristianini and Shawe-Taylor (2000), and Cucker and Zhou (2007). Most of the definitions below can be found in Steinwart and Christmann (2008). The goal of SVMs is to learn a relation between input variables $x \in \mathcal{X}$ and output variables $y \in \mathcal{Y}$, that is a function $f\colon \mathcal{X} \to \mathcal{Y}$, $\mathcal{X}, \mathcal{Y}$ sets. This function should give a prediction of the output value y for a given input value x. Therefore the algorithm is given a set of training data, consisting of pairs of input values and output values (x_i, y_i), $i \in \{1, \ldots, n\}$, $n \in \mathbb{N}$. Then, based on the knowledge of the training data, the predictor f is learned. The quality of the prediction is given in terms of the loss function L and the risk R. The loss function L measures the distance between the true value y and the predicted value $f(x)$ and is defined as follows:

Definition 4.1.1 (Loss function) *Let $\mathcal{X}$ be a measurable space and $\mathcal{Y} \subset \mathbb{R}$ a closed subset, then a function $L : \mathcal{X} \times \mathcal{Y} \times \mathbb{R} \to [0, \infty)$ is called loss function if it is measurable.*

Obviously a perfect prediction, i. e. the prediction equals the true value, should not be punished. Therefore it is assumed, that $L(x, y, y) = 0$ for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$, that is the loss is zero if the prediction equals the true value. A few useful properties and examples of common loss functions are stated later. By means of the expected loss, the risk, a predictor f is considered to be "good" or "bad". The risk is defined as follows:

Definition 4.1.2 (Risk) *Let $L : \mathcal{X} \times \mathcal{Y} \times \mathbb{R} \to [0, \infty)$ be a loss function and P be a probability distribution on $\mathcal{X} \times \mathcal{Y}$, where $\mathcal{Y}$ is a Polish space. For a measurable function $f : \mathcal{X} \to \mathcal{Y}$, the L-risk is defined by*

$$R_{L,P}(f) := \int_{\mathcal{X}\times\mathcal{Y}} L(x, y, f(x))\, dP(x, y) = \int_{\mathcal{X}} \int_{\mathcal{Y}} L(x, y, f(x))\, dP(y|x)\, dP_{\mathcal{X}}(x), \tag{4.1}$$

where $P_{\mathcal{X}}$ denotes the marginal distribution on $\mathcal{X}$ and $P(\cdot|x)$ denotes the regular conditional probability for a given $X = x \in \mathcal{X}$ on $\mathcal{Y}$.

Moreover we define the smallest possible risk, the so-called Bayes risk $R^*_{L,P}$, by $R^*_{L,P}(f) := \inf\{R_{L,P}(f) \mid f : \mathcal{X} \to \mathbb{R} \text{ measurable}\}$. A measurable function $f^* : \mathcal{X} \to \mathbb{R}$ such that $R_{L,P}(f^*) = R^*_{L,P}$ is called a Bayes decision function.

Instead of minimizing over all measurable functions, the support vector machine minimizes over a special Hilbert space consisting of functions, a so-called reproducing kernel Hilbert space (RKHS) H with corresponding kernel k. Some properties of RKHS are listed below, for a detailed description see Berlinet and Thomas-Agnan (2004) and Steinwart and Christmann (2008, Chapter 4) for an overview.

This leads to the definition of the support vector machine:

Definition 4.1.3 (Support vector machine) *Let $L : \mathcal{X} \times \mathcal{Y} \times \mathbb{R} \to [0, \infty)$ be a loss function and let H be a reproducing kernel Hilbert space. Let P be a probability distribution on $\mathcal{X} \times \mathcal{Y}$ and let $\lambda \in \mathbb{R}$, $\lambda > 0$ be an integer. Then the support vector machine $f_{L,P,\lambda}$ is defined via:*

$$f_{L,P,\lambda} := \arg\inf_{f \in H} R_{L,P}(f) + \lambda \|f\|^2_H. \tag{4.2}$$

Additionally to the risk $R_{L,P}$ the definition of the SVM includes a regularization term $\|f\|^2_H$ to prevent overfitting.

Moreover, for a given data set $\mathbf{w}_n = ((x_1, y_1), \ldots, (x_n, y_n)) \in (\mathcal{X} \times \mathcal{Y})^n$ the SVM computed with respect to the empirical measure $\mathbb{P}_{\mathbf{w}_n} = \frac{1}{n}\sum_{i=1}^n \delta_{(x_i,y_i)}$ is called the empirical SVM:

$$f_{L,\mathbb{P}_{\mathbf{w}_n},\lambda} = \arg\inf_{f \in H} \frac{1}{n}\sum_{i=1}^n L(x_i, y_i, f(x_i)) + \lambda \|f\|_H^2. \tag{4.3}$$

To justify this definition for the non-i.i.d. case we again regard stochastic processes which are Varadarajan processes or fulfil a law of large numbers. Then, the existence of a limiting distribution P of the empirical measure is assured. The even weaker assumption of an asymptotically mean stationary process is used in Chapter 4.4 to show consistency of the SVM, that is stochastic convergence of the risk of the empirical estimate to the Bayes risk.

Steinwart and Christmann (2008, Lemma 5.1) provides the uniqueness of a SVM under some mild conditions on the loss function L and the risk $R_{L,P}$. Existence of a SVM is, again under mild conditions on the loss function L, also shown in Steinwart and Christmann (2008, Theorem 5.2).

Moreover representer theorems for the empirical SVM and for general SVMs are shown in Steinwart and Christmann (2008, Theorem 5.8 and Theorem 5.6) and in De Vito et al. (2003/04). In Steinwart et al. (2009) the representer theorems are also used for the non-i.i.d. case.

While working with SVMs, often assumptions on the existence of moments with respect to P are needed. As these assumptions restrict the applicability of SVMs, often the trick of shifting the loss function L is used. Then the shifted loss $L^* : \mathcal{X} \times \mathcal{Y} \times \mathbb{R} \to \mathbb{R} = L(x, y, t) - L(x, y, 0)$ is used instead of the loss function L. For details on this concept see Huber (1981) and Christmann et al. (2009).
For the shifted loss L^* existence and uniqueness as well as a representer Theorem can be found in Christmann et al. (2009, Theorem 5,6,7) if the loss function L is Lipschitz continuous. This concept is applied in Section 4.2 and 4.3 to show that the SVM estimator is qualitatively robust under some assumptions on the statistical operator S for weak Varadarajan processes and to give bounds on the maxbias.

As those results do not depend on the distribution of the data, they are also valid for general stochastic process $(X_i, Y_i)_{i \in \mathbb{N}}$. In particular the proofs of these results do not rely on an i.i.d. assumption of the stochastic process.

Some properties of loss functions and reproducing kernel Hilbert spaces

Due to computational feasibility convex losses are often used, moreover continuity and especially Lipschitz continuity are useful properties of loss functions:

Definition 4.1.4 *Let $L : \mathcal{X} \times \mathcal{Y} \times \mathbb{R} \to [0, \infty)$ be a loss function, then*

*L is said to be **convex** if $L(x, y, \cdot) : \mathbb{R} \to [0, \infty)$ is convex for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$.*

*L is said to be **continuous** if $L(x, y, \cdot) : \mathbb{R} \to [0, \infty)$ is continuous for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$.*

*L is said to be **locally Lipschitz continuous** if for all $a > 0$:*

$$|L|_{a,1} = \sup_{t,t' \in [-a,a],\, t \neq t'} \sup_{(x,y) \in \mathcal{X} \times \mathcal{Y}} \frac{|L(x, y, t) - L(x, y, t')|}{|t - t'|} < \infty. \tag{4.4}$$

*L is said to be **Lipschitz continuous** if $|L|_1 := \sup_{a>0} |L|_{a,1} < \infty$.*

There are several examples of loss functions. For example the classification loss $L(y, t) = 1_{(-\infty,0]}(y\text{sign}(t))$, which is not convex. Therefore, often the hinge loss $L(y, t) = \max\{0, 1 - yt\}$, $y = \pm 1$, $t \in \mathbb{R}$, is used as a surrogate loss. Moreover it is distinguished between supervised losses $L : \mathcal{Y} \times \mathbb{R} \to [0, \infty)$, which are independent of the input value x, and unsupervised losses $L : \mathcal{X} \times \mathbb{R} \to [0, \infty)$, which are independent of the output value y. In the following only supervised losses are considered. Examples of supervised losses are the least squares loss $L(y, t) = (y-t)^2$, which is strictly convex but not Lipschitz continuous, the logistic loss for regression $L(y, t) = -\ln \frac{4e^{y-t}}{(1+e^{y-t})^2}$ and classification $L(y, t) = \ln(1 + e^{-yt})$, which are Lipschitz continuous and strictly convex. For more examples see Steinwart and Christmann (2008, Chapter 2). According to their applicability the supervised losses can be margin-based, used for classification problems, or distance-based, used for regression problems:

Definition 4.1.5 (Margin- and distance-based losses) *A supervised loss $L : \mathcal{Y} \times \mathbb{R} \to [0, \infty)$ is called*

margin-based, if there exists a representing function $\psi : \mathbb{R} \to [0, \infty)$ such that

$$L(y, t) = \psi(yt),\ y \in \mathcal{Y},\ t \in \mathbb{R}.$$

distance-based, if there exists a representing function $\psi : \mathbb{R} \to [0, \infty)$ *such that* $\psi(0) = 0$ *and*

$$L(y,t) = \psi(y-t),\ y \in \mathcal{Y},\ t \in \mathbb{R}.$$

Examples of margin-based losses are the hinge loss, the logistic loss for classification, or the least squares loss $\psi(y-t) = (1-yt)^2,\ y = \pm 1,\ t \in \mathbb{R}$. For distance-based losses and $y, t \in \mathbb{R}$ there is the least squares loss $\psi(y-t) = (y-t)^2$, the logistic loss for regression $\psi(y-t) = -\ln \frac{e^{y-t}}{(1+e^{y-t})^2}$, the ε-insensitive loss $\psi(y-t) = \max\{0, |y-t| - \varepsilon\}$, and the pinball loss

$$\psi(y-t) = \begin{cases} -(1-\tau)(y-t), & if (y-t) < 0 \\ \tau(y-t), & if (y-t) \geq 0 \end{cases}.$$

The representing function ψ inherits some properties from the loss function L, in the margin-based, as well as in the distance-based case. For example ψ is continuous, Lipschitz continuous, and convex if and only if L is continuous, Lipschitz continuous, and convex, see e. g. Steinwart and Christmann (2008, Lemma 2.25 and 2.33).

Another important tool for the analysis of SVMs is the reproducing kernel Hilbert space H, respectively the corresponding reproducing kernel k. For detailed information on kernels and RKHS see Berlinet and Thomas-Agnan (2004), Aronszajn (1950), and Steinwart and Christmann (2008, Chapter 4). A kernel is defined as follows:

Definition 4.1.6 (Kernel) *Let* $\mathcal{X}$ *be a non-empty set. Then a function* $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ *is called a kernel on* $\mathcal{X}$ *if there exists a* $\mathbb{R}$*-Hilbert space* H *and a map* $\Phi : \mathcal{X} \to H$ *such that for all* $x, x' \in \mathcal{X}$ *we have*

$$k(x,x') = \left\langle \Phi(x), \Phi(x') \right\rangle.$$

We call Φ *a feature map and* H *a feature space of* k*.*

There are several examples of kernels, a kernel which is often used in practice is the Gaussian RBF kernel $k_{\gamma,\mathbb{R}^d}(x,x') = \exp(-\frac{\|x-x'\|_2^2}{\gamma^2}),\ \gamma > 0$.

In general the feature space and the feature map are not uniquely determined. Therefore the reproducing kernel Hilbert space (RKHS) is defined, which is in some sense a canonical choice of feature space and uniquely determined.

Definition 4.1.7 (Reproducing kernel) *Let* $\mathcal{X} \neq \emptyset$ *and* H *be a* $\mathbb{R}$*-Hilbert function space over* $\mathcal{X}$*, i. e., a* $\mathbb{R}$*-Hilbert space that consists of functions mapping from* $\mathcal{X}$ *into* $\mathbb{R}$*.*

A function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called a reproducing kernel of H if $k(\cdot, x) \in H$ for all $x \in \mathcal{X}$ and the reproducing property

$$f(x) = \langle f, k(\cdot, x) \rangle_H$$

applies for all $f \in H$ and all $x \in \mathcal{X}$.

The space H is called a reproducing kernel Hilbert space (RKHS) over $\mathcal{X}$ if for all $x \in \mathcal{X}$ the Dirac functional $\delta_x : H \to \mathbb{R}$, defined by

$$\delta_x(f) = f(x), \ f \in H,$$

is continuous.

The canonical feature map $\Phi : \mathcal{X} \to H$ is given by

$$\Phi(x) = k(\cdot, x), \ x \in \mathcal{X}.$$

In Steinwart and Christmann (2008, Theorem 4.20 and Theorem 4.21) the correspondence between kernels and RKHS is given. Every RKHS corresponds to exactly one reproducing kernel, which is a kernel, and every kernel has exactly one RKHS, for which it is a reproducing kernel.

The next theorem states a few inequalities, which are frequently used in the next sections.

Theorem 4.1.8 *Let $\mathcal{X}$ be topological space and k a kernel on $\mathcal{X}$ with RKHS H.*

Then k is bounded and $k(\cdot, x) : \mathcal{X} \to \mathbb{R}$ is continuous for all $x \in \mathcal{X}$ if and only if every $f \in H$ is a bounded and continuous function.

Then we have

$$\|f\|_\infty \leq \|f\|_H \|k\|_\infty, \tag{4.5}$$

$$\|\Phi\|_\infty = \sup_{x' \in \mathcal{X}} |\Phi(x)(x')| \leq \|k\|_\infty^2 \quad \textit{and} \tag{4.6}$$

$$\|\Phi(x)\|_H^2 = \langle \Phi(x), \Phi(x) \rangle = k(x, x) \leq \|k\|_\infty^2. \tag{4.7}$$

For the proofs of this results, see Steinwart and Christmann (2008, Lemma 4.23 and 4.24). Continuity of the kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ as well as measurability and further useful properties of RKHS and kernels can also be found in Steinwart and Christmann (2008, Chapter 4).

4.2 Qualitative robustness of support vector machines

In this chapter, we use Theorem 3.1.3 to show qualitative robustness of support vector machines for non-i.i.d. observations, that is, we show that the estimator S_n can be represented by a functional S, which is continuous in P. For SVMs the estimator S_n maps the training data $((x_1, y_1), \ldots, (x_n, y_n))$ to a function $f_{L,\mathbb{P}_{\mathbf{w}_n},\lambda} \in H$ and is given by the function which minimizes $\lambda \|f\|_H^2 + \frac{1}{n} \sum_{i=1}^n L(x_i, y_i, f(x_i))$.

To use Theorem 3.1.3 we would like to consider a statistical operator $S : \mathcal{M}(\mathcal{X} \times \mathcal{Y}) \to H,\ P \mapsto f_{L,P,\lambda}$, but the SVM need not exist for every $P \in \mathcal{M}(\mathcal{X} \times \mathcal{Y})$. By using the L^*-trick, see Section 4.1, we gain the existence of a SVM $f_{L^*,P,\lambda}$ for every $P \in \mathcal{M}(\mathcal{X} \times \mathcal{Y})$, where the SVM $f_{L^*,P,\lambda}$ is analogously defined as $f_{L,P,\lambda}$ with the use of L^* instead of L. Therefore, it is easy to see: if $f_{L,P,\lambda}$ exists it equals $f_{L^*,P,\lambda}$, see Christmann et al. (2009).

Now we can define a statistical operator by

$$\begin{aligned} S: \quad & \mathcal{M}(\mathcal{X} \times \mathcal{Y}) \to H \\ & P \mapsto f_{L^*,P,\lambda} \end{aligned} \tag{4.8}$$

in the sense that $S(\mathbb{P}_{\mathbf{w}_n}) = S_n(\mathbf{w}_n) = f_{L^*,\mathbb{P}_{\mathbf{w}_n},\lambda}$.

Using the shifted loss function L^*, qualitative robustness of the SVM estimator $(S_n)_{n\in\mathbb{N}}$ is ensured for any fixed regularization parameter $\lambda > 0$ and under mild conditions on the loss L and the kernel k, see Hable and Christmann (2011, Theorem 3.1).

However, the estimators $(S_n)_{n\in\mathbb{N}}$ are not consistent for fixed regularization parameter λ. To obtain consistency the fixed λ has to be replaced by a sequence λ_n converging to zero, as n tends to ∞, see e. g. Steinwart and Christmann (2008). But then Hable and Christmann (2011, Proposition 5.2) yields that this sequence of estimators is not qualitatively robust any more. This is not a special property of SVMs but an unavoidable consequence of the fact that risk minimization is an ill-posed problem. For such a problem it follows from Hampel's second theorem that no estimator can simultaneously be consistent and robust, see Hable and Christmann (2013). In order to find a good compromise between consistency and robustness, we fix a possibly small $\lambda_0 > 0$ and allow for a sequence of regularization parameters with $\lambda_n \to \lambda_0$. Then, the following theorem shows qualitative robustness for the sequence of estimators $S_{\lambda_n} : \mathbf{w}_n \mapsto f_{L,\mathbb{P}_{\mathbf{w}_n},\lambda_n}$, even in the non-i.i.d. case.

Theorem 4.2.1 *Let $\mathcal{Z}, H$ be complete separable metric spaces, let $(Z_i)_{i\in\mathbb{N}}$, $Z_i : \Omega \to \mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, be a stochastic process satisfying the weak Varadarajan property, $\mathcal{Y} \subset \mathbb{R}$ closed, $(\lambda_n)_{n\in\mathbb{N}}$ a sequence of positive real valued numbers with $\lambda_n \to \lambda_0$, $n \to \infty$, for a $\lambda_0 > 0$. Let $S_{\lambda_n} : (\mathcal{X} \times \mathcal{Y})^n \to H$ be the SVM estimator, which maps $\mathbf{w}_n$ to $f_{L^*,\mathbb{P}_{\mathbf{w}_n},\lambda_n}$ for a continuous and convex loss function $L : \mathcal{X} \times \mathcal{Y} \times \mathbb{R} \to [0,\infty)$. Assume that $L(x,y,y) = 0$ for every $(x,y) \in \mathcal{X} \times \mathcal{Y}$, that L is additionally Lipschitz continuous in the last argument, and that the kernel k is continuous and bounded.*

Then, the sequence of estimators $(S_{\lambda_n})_{n\in\mathbb{N}}$ is qualitatively $(\pi_{d_n})_{n\in\mathbb{N}}$-robust at $P_{\mathbb{N}}$.

Remember, that the metric π_{d_n} is defined by:

$$d_n\big((z_1,\dots,z_n),(z_1',\dots,z_n')\big) \;=\; \inf\big\{\varepsilon > 0 : \sharp\{i : d_{\mathcal{Z}}(z_i,z_i') \geq \varepsilon\}/n \leq \varepsilon\big\}.$$

Proof of Theorem 4.2.1: To prove qualitative robustness of the SVMs we choose an arbitrary $\varepsilon > 0$. Similarly to Hable (2013, Lemma 9(b)(i)) we have:

$$\|f_{L^*,\mathbb{P}_{\mathbf{w}_n},\lambda_n} - f_{L^*,\mathbb{P}_{\mathbf{w}_n},\lambda_0}\|_H \leq \frac{\lambda_n - \lambda_0}{\lambda_n\lambda_0} 2|L|_1\|k\|_\infty, \tag{4.9}$$

where L^* is the shifted loss function and $|L|_1$ denotes the Lipschitz constant of L respectively L^*. In Hable (2013, Lemma 9(b)(i)) the above result is given for the regular loss L, but the proof is the same for the shifted loss L^* except for the last step. Here $\|f_{L^*,\mathbb{P}_{\mathbf{w}_n},\lambda_0}\|_H \leq \frac{1}{\lambda_0}\|L\|_L\|k\|_\infty$, see Christmann et al. (2009, Proposition 3(iv)), can be used instead of the corresponding bound by use of the risk.
According to (4.9), with $\lambda_n \to \lambda_0$, there exists $n_0 \in \mathbb{N}$ such that, for every $n \geq n_0$, $\mathbf{w}_n \in \mathcal{Z}^n$: $\|f_{L^*,\mathbb{P}_{\mathbf{w}_n},\lambda_n} - f_{L^*,\mathbb{P}_{\mathbf{w}_n},\lambda_0}\|_H \leq \frac{\varepsilon}{3}$. Now let $n < n_0$. Due to the regularity assumptions on the loss function L and the kernel k, the qualitative $(\pi_{d_n})_{n\in\mathbb{N}}$-robustness for the estimator $\mathbf{w}_n \mapsto \arg\inf_{f\in H} \lambda\|f\|_H^2 + \frac{1}{n}\sum_{i=1}^n L^*(x,y,f(x))$ follows from Hable and Christmann (2011, Theorem 3.1). Hence we have for the estimator $S_n : \mathbf{w}_n \mapsto f_{L^*,\mathbb{P}_{\mathbf{w}_n},\lambda_n}$: for every $\varepsilon > 0$ and for every $n < n_0$ there is $\delta_n > 0$ such that:

$$\pi_{d_n}(P_n,Q_n) \leq \delta_n \;\Rightarrow\; \pi_{d_H}(S_n(P_n),S_n(Q_n)) \leq \varepsilon.$$

For $n \geq n_0$, choose δ_{n_0} such that $\pi_{d_n}(P_n,Q_n) \leq \delta_{n_0}$ implies $\pi_{d_H}(S_n^{\lambda_0}(P_n),S_n^{\lambda_0}(Q_n)) \leq \frac{\varepsilon}{3}$, where $S_n^{\lambda_0} : \mathbf{w}_n \mapsto f_{L^*,\mathbb{P}_{\mathbf{w}_n},\lambda_0}$ for the fixed λ_0, which is again possible due to Hable and Christmann (2011, Theorem 3.1). Now let a measurable $A \subset H$ be arbitrarily chosen and

define $\mathcal{D}_{\lambda_n} := S_n^{-1}(A)$, then

$$S_n(P_n)(A) \;=\; P_n(S_n^{-1}(A)) \;=\; P_n(\mathbf{W}_{\lambda_n}) \;\leq\; P_n((S_n^{\lambda_0})^{-1}(A^{\varepsilon/3}))$$

because $\|S_n(\mathbf{w}_n) - S_n^{\lambda_0}(\mathbf{w}_n)\|_H = \|f_{\mathbb{P}_{\mathbf{w}_n},\lambda_n} - f_{\mathbb{P}_{\mathbf{w}_n},\lambda_0}\|_H \leq \frac{\varepsilon}{3}$ by assumption and therefore $\mathbf{W}_{\lambda_n} \subset (S_n^{\lambda_0})^{-1}(A^{\varepsilon/3})$. Remember, that $A^\varepsilon = \{x \in \mathcal{H} : d_H(x, A) < \varepsilon\}$. By use of the qualitative robustness of S_{λ_0} and the choice of δ_{n_0} it follows, that:

$$S_n(P_n)(A) \;\leq\; P_n((S_n^{\lambda_0})^{-1}(A^{\varepsilon/3})) \;\leq\; Q_n((S_n^{\lambda_0})^{-1}(A^{\varepsilon/3+\varepsilon/3})) + \frac{\varepsilon}{3}$$

and with the same argument as before: $Q_n((S_n^{\lambda_0})^{-1}(A^{2\varepsilon/3})) \leq Q_n(S_n^{-1}(A^{2\varepsilon/3+\varepsilon/3}))$. So,

$$S_n(P_n)(A) \;\leq\; Q_n(S_n^{-1}(A^\varepsilon)) + \varepsilon/3 \;\leq\; Q_n(S_n^{-1}(A^\varepsilon)) + \varepsilon$$

and therefore for every $n \geq n_0$: if $\pi_{d_n}(P_n, Q_n) \leq \delta_{n_0}$, then $\pi_{d_H}(S_n(P_n), S_n(Q_n)) \leq \varepsilon$. Now choose $\delta = \min\{\delta_1, \ldots, \delta_{n_0}\}$. □

The proof above shows, if we have qualitative robustness for the sequence of estimators $S_n : \mathcal{Z}^n \to H,\ \mathbf{w}_n \mapsto f_{L^*,\mathbb{P}_{\mathbf{w}_n},\lambda}$ for fixed $\lambda > 0$, the sequence of estimators $S_n : \mathcal{Z}^n \to H,\ \mathbf{w}_n \mapsto f_{L^*,\mathbb{P}_{\mathbf{w}_n},\lambda_n}$ is also qualitatively robust for $\lambda_n \to \lambda_0,\ n \to \infty,\ \lambda_0 > 0$. A similar argument as above can be used to show qualitative robustness of the bootstrap approximation for the SVM estimator for independent, not necessarily identically distributed stochastic processes, see Theorem 3.4.2, Chapter 3.4.

Corollary 4.2.2 *Let $P_{\mathbb{N}} = \otimes_{i\in\mathbb{N}} P^i$, $P^i \in \mathcal{M}(\mathcal{Z})$ be an infinite product measure such that the coordinate process $(Z_i)_{i\in\mathbb{N}}$ is a strong Varadarajan process. Define the set of product measures on $\mathcal{Z}^{\mathbb{N}}$, $\mathcal{P} := \left\{Q_{\mathbb{N}} \in \mathcal{M}(\mathcal{Z}^{\mathbb{N}});\ Q_{\mathbb{N}} = \otimes_{i\in\mathbb{N}} Q^i,\ Q^i \in \mathcal{M}(\mathcal{Z})\right\}$. Let $(\lambda_n)_{n\in\mathbb{N}}$ be a sequence of positive real valued numbers with $\lambda_n \to \lambda_0,\ n \to \infty$, for a $\lambda_0 > 0$ and let $S_{\lambda_n} : (\mathcal{X} \times \mathcal{Y})^n \to H$ be the SVM estimator, which maps $\mathbf{w}_n$ to $f_{L^*,\mathbb{P}_{\mathbf{w}_n},\lambda_n}$ for a continuous and convex loss function $L : \mathcal{X} \times \mathcal{Y} \times \mathbb{R} \to [0, \infty)$. Let $L(x, y, y) = 0$ for every $(x, y) \in \mathcal{X} \times \mathcal{Y}$, let L be Lipschitz continuous in the last argument, and let the kernel k be continuous and bounded.*

Then the sequence of bootstrap approximations $(\mathcal{L}_{P_n^*}(S_n))_{n\in\mathbb{N}}$ is qualitatively robust at $P_{\mathbb{N}}$ with respect to $\mathcal{P}$.

Proof: The regularity assumptions on the loss function L and the kernel k imply the continuity of the statistical operator $S : \mathcal{M}(\mathcal{X} \times \mathcal{Y}) \to H$, see Hable and Christmann (2011,

Theorem 3.2), as well as the continuity of the estimators $S_n : (\mathcal{X} \times \mathcal{Y})^n \to H$, $\mathbf{w}_n \mapsto f_{L^*,\mathbb{P}_{\mathbf{w}_n},\lambda_n}$ for every $\lambda_n \in (0,\infty)$, $n \in \mathbb{N}$. Hence, for fixed λ the bootstrap approximation of the SVM estimator $S_n : \mathbf{w}_n \mapsto f_{L^*,\mathbb{P}_{\mathbf{w}_n},\lambda}$ is qualitatively robust, that is, for every $\varepsilon > 0$ there is $\delta > 0$ such that there is $n_0 \in \mathbb{N}$ such that for all $n \geq n_0$ and for all $Q_{\mathbb{N}} \in \mathcal{P}$:

$$d_{\mathrm{BL}}(P_n, Q_n) < \delta \;\Rightarrow\; d_{\mathrm{BL}}(\mathcal{L}(\mathcal{L}_{P_n^*}(S_n)), \mathcal{L}(\mathcal{L}_{Q_n^*}(S_n))) < \varepsilon.$$

Moreover the proof of Theorem 3.4.2, (3.40), and the strong equivalence between the bounded Lipschitz metric and the Prohorov distance on Polish spaces, see e. g. Huber (1981, Chapter 2, Corollary 4.3), yield: for every $\varepsilon > 0$ there is $\delta > 0$ such that there is $n_0 \in \mathbb{N}$ such that for all $n \geq n_0$ and if $d_{\mathrm{BL}}(P_n, Q_n) \leq \delta$:

$$\pi(\mathcal{L}_{P_n^*}(S_n), \mathcal{L}_{Q_n^*}(S_n)) < \varepsilon \text{ almost surely.}$$

Similarly to the proof above, for every $\varepsilon > 0$ there is n_ε such that for all $n \geq n_\varepsilon$:

$$\|f_{L^*,\mathbb{P}_{\mathbf{w}_n},\lambda_n} - f_{L^*,\mathbb{P}_{\mathbf{w}_n},\lambda_0}\|_H \leq \frac{\varepsilon}{3}.$$

Now, the same argumentation as in the proof above for the cases $n_0 \leq n \leq n_\varepsilon$ and $n > n_\varepsilon$ for the sequence of estimators $S_{\lambda_n} : \mathbf{w}_n \mapsto f_{L,\mathbb{P}_{\mathbf{w}_n},\lambda_n}$ yields the assertion. □

As already described for M-estimators at the end of Chapter 3.4.2, the statistical operator $S : \mathcal{M}(\mathcal{Z}) \to H$ is uniformly continuous if the space $\mathcal{Z}$ is compact. Therefore qualitative robustness of the bootstrap approximation of the SVM estimator for α-mixing sequences with values in $[0,1]^d$, $d \geq 1$, follows in the same way as above. By assuming the space $\mathcal{X} \times \mathcal{Y}$ to be compact, Theorem 3.4.5 and 3.4.6 yield the qualitative robustness of the bootstrap approximation for the SVM estimator for fixed regularization parameter λ under the assumptions on the kernel and the loss function given above. Then the same argumentation as above yields the qualitative robustness of the bootstrap approximation of $(S_{\lambda_n})_{n\in\mathbb{N}}$.

4.3 Quantitative robustness of support vector machines - maximum bias

Besides qualitative robustness we shortly regard the maximum bias of SVMs, which is a quantitative approach to robustness. Quantitative robustness describes the influence of a small change in the underlying distribution to the test statistic $S_n : \mathcal{Z}^n \to H$ or to the

distribution of the estimator $\mathcal{L}(S_n)$. This can be useful, for example, to select a statistical procedure, whereas the qualitative robustness does not give a quantitative measure to compare two stochastic procedures. There are many more different kinds of quantitative robustness, for example the influence function, the sensitivity curve, and the breakdown point, see Huber (1981, Chapter 1.4 and 1.5) for a detailed description.

Again we are concerned with the SVM estimator $S_n : \mathcal{Z}^n \to H,\ \mathbf{w}_n \mapsto f_{L^*,\mathbb{P}_{\mathbf{w}_n},\lambda}$ for fixed $\lambda > 0$ and a dataset $\mathbf{w}_n = ((x_1, y_1), \ldots, (x_n, y_n))$, respectively the operator $S : \mathcal{M}(\mathcal{Z}) \to H,\ P \mapsto f_{L^*,P,\lambda}$, see (4.8).

To describe the term "small change" of the underlying distribution, neighbourhoods of the true distribution P are investigated. Commonly used neighbourhoods, see for example Huber (1981), are the *contamination neighbourhood* N_{con}

$$N_{con,\varepsilon}(P) := \{P_\varepsilon \mid Q = (1-\varepsilon)P + \varepsilon Q,\ Q \in \mathcal{M}(\mathcal{Z})\}, \tag{4.10}$$

which is not a neighbourhood in the topological sense. And the *total variation neighbourhood* N_{TV}

$$N_{\mathrm{TV},\varepsilon}(P) := \{Q \in \mathcal{M}(\mathcal{Z}) \mid d_{TV}(P,Q) \leq \varepsilon\}, \tag{4.11}$$

where $\varepsilon > 0$ and

$$d_{\mathrm{TV}}(P,Q) := \sup_{A \in \mathcal{B}(\mathcal{Z})} \|P(A) - Q(A)\| = \sup_{\|f\|_\infty \leq 1} \frac{1}{2} \left| \int f\, dP - \int f\, dQ \right|$$

is the total variation metric. Note that $d_{\mathrm{TV}}(P,Q) \leq 1$, for every $P, Q \in \mathcal{M}(\mathcal{Z})$.

A characteristic which is often used to describe quantitative robustness is the *maximum bias*. An estimator is said to be quantitative robust if the maximum bias is bounded for sufficiently large ε. To compare two statistical methods the estimator with the smaller maximum bias is considered to be better. The following definition is a straight forward modification of the definition in Huber (1981, p.11) to our set-up.

Definition 4.3.1 (Maximum bias) *Let $\mathcal{Z}, H$ be complete separable metric spaces, let $S : \mathcal{M}(\mathcal{Z}) \to H$ be a statistical operator, then the maximum bias of S is defined by:*

$$b(\varepsilon, P) := \sup_{Q \in N_\varepsilon} \|S(P) - S(Q)\|_H. \tag{4.12}$$

The following theorem shows, that there exists a linear bound on the maximum bias of the SVM estimator $S : P \mapsto f_{L^*,P,\lambda}$. This result is similar to the results in the i.i.d. case which also provide a linear bound on the maxbias, see e. g. Christmann and Steinwart (2004, Remark 14) and Christmann et al. (2009, Theorem 12).

Theorem 4.3.2 *Let $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ be a complete separable metric space, $\mathcal{Y} \subset \mathbb{R}$, let $S : \mathcal{M}(\mathcal{Z}) \to H$, $P \mapsto f_{L^*,P,\lambda}$ be the SVM operator in (4.8), let $L : \mathcal{X} \times \mathcal{Y} \times \mathbb{R} \to [0, \infty[$ be a convex, Lipschitz continuous loss function and $L^* : \mathcal{X} \times \mathcal{Y} \times \mathbb{R} \to \mathbb{R}$ the shifted loss function, let H be the RKHS to a continuous, bounded kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ and $\lambda > 0$, then the maximum bias $b(\varepsilon, P)$ is bounded:*

i) for the contamination neighbourhood

$$b_{con}(\varepsilon, P) \leq \frac{1}{\lambda} C\varepsilon, \tag{4.13}$$

ii) for the total variation neighbourhood

$$b_{\mathrm{TV}}(\varepsilon, P) \leq \frac{1}{\lambda} C\varepsilon, \tag{4.14}$$

where $C > 0$ depends on the loss function L and the kernel k.

Proof: The proof of Theorem 4.3.2 is based on the representer theorem which can be found in Christmann et al. (2009, Theorem 6).

i): Let $P \in \mathcal{M}(\mathcal{X} \times \mathcal{Y})$ be a fixed probability measure. Then for every $\varepsilon > 0$, for every $P_\varepsilon \in N_{con,\varepsilon}(P)$, i. e. for every $P_\varepsilon = (1 - \varepsilon)P + \varepsilon Q$, $Q \in \mathcal{M}(\mathcal{X} \times \mathcal{Y})$, the maximum bias is given by:

$$\begin{aligned} b_{con}(\varepsilon, P) &= \sup_{P_\varepsilon \in N_{con,\varepsilon}} \|S(P) - S(P_\varepsilon)\|_H \\ &= \sup_{P_\varepsilon \in N_{con,\varepsilon}} \|f_{L^*,P,\lambda} - f_{L^*,P_\varepsilon,\lambda}\|_H. \end{aligned}$$

The representer theorem, see Christmann et al. (2009, Theorem 6), ensures the existence of a bounded function $h_P : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, element of the subdifferential (see Definition A8)

$\partial L(x, y, f_{L^*,P,\lambda}(x))$ of the loss function L, such that for all $\lambda > 0$:

$$\begin{aligned}
\|f_{L^*,P,\lambda} - f_{L^*,P_\varepsilon,\lambda}\|_H &\leq \frac{1}{\lambda} \|E_P h_P \Phi - E_{P_\varepsilon} h_P \Phi\|_H \\
&= \frac{1}{\lambda} \left\| \int_{\mathcal{X}\times\mathcal{Y}} h_P \Phi \, dP - \int_{\mathcal{X}\times\mathcal{Y}} h_P \Phi d \, ((1-\varepsilon)P + \varepsilon Q) \right\|_H \\
&= \frac{1}{\lambda} \left\| \int_{\mathcal{X}\times\mathcal{Y}} h_P \Phi \, d(\varepsilon(P-Q)) \right\|_H \\
&= \left\| \varepsilon \int_{\mathcal{X}\times\mathcal{Y}} h_P \Phi \, d(P-Q) \right\|_H \\
&\leq \frac{1}{\lambda} \varepsilon \|h_P\|_\infty \sup_{x\in\mathcal{X}} \|\Phi(x)\|_H d_{TV}(P,Q).
\end{aligned}$$

As L is Lipschitz continuous, the function h is bounded by the Lipschitz constant $|L|_1$ of L, respectively L^*, see Christmann et al. (2009, Theorem 6). Moreover $\|\Phi(x)\|_H \leq \|k\|_\infty$, see (4.7), and $d_{TV}(P,Q) \leq 1$, for all $P, Q \in \mathcal{M}(\mathcal{X}\times\mathcal{Y})$. Hence:

$$b_{con}(\varepsilon, P) \leq \frac{1}{\lambda} \varepsilon |L|_1 \|k\|_\infty.$$

The proof of part ii) is similar to the first part: Fix any $P \in \mathcal{M}(\mathcal{X}\times\mathcal{Y})$ and $\lambda > 0$. Then, for every $\varepsilon > 0$, for every $Q \in N_{\varepsilon,\mathrm{TV}}$, i. e. for every $Q \in \mathcal{M}(\mathcal{X}\times\mathcal{Y})$, with $d_{\mathrm{TV}}(P,Q) \leq \varepsilon$:

$$\begin{aligned}
b_{\mathrm{TV}}(\varepsilon, P) := & \sup_{Q\in N_{\mathrm{TV},\varepsilon}} \|S(P) - S(Q)\|_H \\
= & \sup_{Q\in N_{\mathrm{TV},\varepsilon}} \|f_{L^*,P,\lambda} - f_{L^*,Q,\lambda}\|_H \\
\leq & \sup_{Q\in N_{\mathrm{TV},\varepsilon}} \frac{1}{\lambda} \|E_P h_P \Phi - E_Q h_P \Phi\|_H \\
\leq & \frac{1}{\lambda} \|h_P\|_\infty \sup_{x\in\mathcal{X}} \|\Phi(x)\|_H d_{TV}(P,Q) \\
\overset{(4.7)}{\leq} & \frac{1}{\lambda} \varepsilon |L|_1 \|k\|_\infty.
\end{aligned}$$

□

Clearly, the maximum bias b is bounded if the kernel k is bounded, and therefore the SVM $f_{L^*,P,\lambda}$ is bounded.

As the computation of empirical SVMs is based on a data set, we shortly discuss the maximum bias for two empirical measures: Let $P_{\mathbf{w}_n}$ be the empirical measure for the data set $(z_1, \ldots, z_n)$ and $P_{\mathbf{w}'_{n'}}$ the empirical measure for $(z'_1, \ldots, z'_{n'})$.

If $(z'_1, \dots, z'_{n'})$ equals $(z_1, \dots, z_n)$ except from a fraction α, then the maximum bias, analogously to the theorem above, is bounded by $\frac{1}{\lambda}C\alpha$. That means, if an experiment is done twice for the same input variables x and the output values are the same except from a few, then the bias of the estimates is smaller then $\frac{1}{\lambda}C\alpha$. This can be seen by following the proof of Theorem 4.3.2 and by computing $d_{\mathrm{TV}}(P_{\mathbf{w}_n}, P_{\mathbf{w}'_n}) = \sup_{||f||_\infty \leq 1} \frac{1}{2} \left| \frac{1}{n} \sum_{i=1}^n f(z_i) - f(z'_i) \right| \leq \frac{1}{2n} \sum_{i, z_i \neq z'_i} |f(z_i) - f(z'_i)| \leq \alpha$.
Regarding rounding errors, that is the difference between the observations $|z_i - z'_i|$ of two data sets is smaller than δ for all $i \in 1, \dots, n$, the maximum bias b is smaller than $\frac{1}{\lambda}C$. Again computing d_{TV}: $d_{\mathrm{TV}}(P_{\mathbf{w}_n}, P_{\mathbf{w}'_n}) = \sup_{||f||_\infty \leq 1} \frac{1}{2} \left| \frac{1}{n} \sum_{i=1}^n f(z_i) - f(z'_i) \right| \leq 1$ leads the assertion.
Now assume, that $(z'_1, \dots, z'_{n'})$ results from $(z_1, \dots, z_n)$ by adding $n' - n, n' > n$ data points, then $b < \frac{1}{\lambda}C \cdot \frac{n'-n}{n'}$.

Besides robustness another important property of an estimator is consistency. The next chapter introduces L-risk-consistency of support vector machines, which justifies, that the estimate f can be learned on a given data set and converges to the theoretical solution.

4.4 Consistency of support vector machines

As the theoretical distribution of the data generating random variables is commonly unknown, the predictor f is learned from a given data set. That is, an empirical estimate is used instead of the theoretical solution. Therefore it is crucial to claim some kind of convergence of the empirical result to the true theoretical solution, that is consistency in a probabilistic sense. Here, we examine L-risk-consistency of support vector machines, i. e. convergence in probability of the expected loss of the empirical estimate to the theoretically expected loss.

In the i.i.d. case the risk $R_{L,P}(f)$ is computed with respect to the distribution $P = \mathcal{L}(Z_i), i \in \mathbb{N}$. For general stochastic processes we do not require the random variables to be identically distributed and independent, so no intuitive choice of distribution exists. Moreover, as working with the empirical SVM, we need to ensure that this definition is reasonable for the non-i.i.d. case. Hence, we assume convergence of the empirical measure to a limiting distribution. Therefore we work with processes which are asymptotically mean stationary (AMS). Remember from Chapter 3.2.2, a process is called *asymptotically mean stationary*,

if there exists a probability measure $P \in \mathcal{M}(\mathcal{Z})$ such that

$$P(B) = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_\mu I_B \circ Z_i, \text{ for all } B \in \mathcal{B}. \tag{4.15}$$

In particular, every strongly stationary process is AMS. The AMS property indicates that there exists a limiting distribution P such that the distribution of the random variables asymptotically equal each other, hence this choice intuitively implies the computation of the risk with respect to the limiting distribution P.

AMS processes are introduced for dynamical systems in Gray (1988) and are used for general stochastic processes in Steinwart et al. (2009). Many examples for AMS processes are provided via Varadarajan processes, see Chapter 3.2, and processes which satisfy a law of large numbers for events, see Steinwart et al. (2009). Both notions imply convergence of the empirical measure to a limiting distribution P and the AMS property, see (Steinwart et al., 2009, Theorem 2.4) and Lemma 4.4.1 below.

Additionally, a process which satisfies a (weak) law of large numbers for events is a (weak) Varadarajan process, see Theorem 3.2.1. Examples are α-mixing processes, certain Markov chains, weakly dependent processes or strongly stationary ergodic processes, see Steinwart et al. (2009) and Chapter 3.2.1 for more examples.

Lemma 4.4.1 *Let $(\Omega, \mathcal{A}, \mu)$ be a probability space and let $\mathcal{Z}$ be a Polish space equipped with the Borel σ-algebra $\mathcal{B}$. Then, for a stochastic process $(Z_i)_{i\in\mathbb{N}}$, $Z_i : \Omega \to \mathcal{Z}$, $i \in \mathbb{N}$, the weak Varadarajan property implies the AMS property. That is:*

$$\pi(\mathbb{P}_{\mathbf{W}_n}, P) \longrightarrow 0 \textit{ in probability} \quad \Rightarrow \quad \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_\mu I_B \circ Z_i = P(B), \textit{ for all } B \in \mathcal{B},$$

where P is the limiting distribution of the Varadarajan process.

Proof of Lemma 4.4.1: Let $B \in \mathcal{B}$ be a Borel set. Therefore, countable many open subsets $B_i \subset \mathcal{Z}$, $i \in \mathbb{N}$, exist such that $B = \bigcup_{i\in\mathbb{N}} B_i$. Without loss of generality we assume B_i to be pairwise disjoint. Hence $I_B(z) = \sum_{i\in\mathbb{N}} I_{B_i}(z)$.

By assumption $(Z_i)_{i\in\mathbb{N}}$ is a weak Varadarajan process, that is, due to Dudley (1989, Theorem

11.3.3), equivalent to

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} f\circ Z_i = \int f\,dP \text{ in probability, for all } f\in C_b(\mathcal{Z}).$$

Hence, for all $f \in C_b(\mathcal{Z})$, there is $m \in \mathbb{N}$ such that for all $n > m$:

$$\mu\left(\left\{\omega\in\Omega \;\middle|\; \left|\frac{1}{n}\sum_{i=1}^{n} f\circ Z_i(\omega) - \int f\,dP\right| > \varepsilon\right\}\right) \leq \varepsilon. \tag{4.16}$$

Now for every B_i, $i \in \mathbb{N}$, define the function $f_{i,n} : \mathcal{Z} \to \mathbb{R}$, $f_{i,n}(z) = \min\{1, nd(z, B_i^c)\}$, $n \in \mathbb{N}$, where $B_i^c = \mathcal{Z}\backslash B_i$ and $d(z, B_i^c) := \inf_{\tilde{z}\in B_i^c} d_{\mathcal{Z}}(z, \tilde{z})$ measures the distance between the point $z \in \mathcal{Z}$ and the set $B_i^c \subset \mathcal{Z}$ and $d_{\mathcal{Z}}$ denotes a metric on $\mathcal{Z}$. It can easily be seen, that this function is continuous and for every $\delta > 0$ and $i \in \mathbb{N}$ there exists $n_i \in \mathbb{N}$ such that $\|I_{B_i} - f_{i,n_i}\|_\infty \leq \delta$. Then, for every $\varepsilon > 0$ there are $n_i \in \mathbb{N}$ such that $\|\sum_{i\in\mathbb{N}} f_{i,n_i} - I_B\|_\infty \leq \frac{\varepsilon}{2}$. By choosing a suitable partial sum we approximate $\sum_{i\in\mathbb{N}} f_{i,n_i}$ as follows: for every $\varepsilon > 0$ there exists $n_0 \in \mathbb{N}$ such that $\|\sum_{i\in\mathbb{N}} f_{i,n_i} - \sum_{i=1}^{n_0} f_{i,n_i}\|_\infty \leq \frac{\varepsilon}{2}$.

Therefore, for every $\varepsilon > 0$ there is $n_0 \in \mathbb{N}$ and there are positive integers $n_1, n_2, \ldots, n_{n_0}$ such that

$$\|\sum_{i=1}^{n_0} f_{i,n_i} - I_B\|_\infty \leq \varepsilon.$$

The function $\sum_{i=1}^{n_0} f_{i,n_i} := F$ is continuous as it is a finite sum of continuous functions. Additionally $|F| \leq 1$ by definition of the functions f_{i,n_i}. Hence, for all $n \geq \max\{n_0, n_1, \ldots, n_{n_0}, m\}$, $n \in \mathbb{N}_0$, the triangle inequality yields:

$$\begin{aligned}
\left|\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_\mu I_B\circ Z_i - P(B)\right| &= \left|\int \frac{1}{n}\sum_{i=1}^{n} I_B\circ Z_i\,d\mu - \int I_B\,dP\right| \\
&= \left|\int \frac{1}{n}\sum_{i=1}^{n}\left[I_B\circ Z_i - \int I_B\,dP\right]\,d\mu\right| \\
&\leq \int \frac{1}{n}\sum_{i=1}^{n} |I_B\circ Z_i - F\circ Z_i|\,d\mu + \left|\int\left[\frac{1}{n}\sum_{i=1}^{n} F\circ Z_i - \int F\,dP\right]\,d\mu\right| \\
&\quad + \int\left[\int |F - I_B|\,dP\right]\,d\mu \\
&\leq 2\varepsilon + \left|\int\left[\frac{1}{n}\sum_{i=1}^{n} F\circ Z_i - \int F\,dP\right]\,d\mu\right|.
\end{aligned}$$

Due to the Varadarajan property and the continuity and boundedness of F, (4.16), yields for $A_\varepsilon := \left\{\omega \in \Omega \;\middle|\; \left|\frac{1}{n}\sum_{i=1}^n F \circ Z_i(\omega) - \int F\,dP\right| > \varepsilon\right\}$ and $A_\varepsilon^c = \Omega \backslash A_\varepsilon$:

$$\begin{aligned}
&\left|\int \left[\frac{1}{n}\sum_{i=1}^n F \circ Z_i - \int F\,dP\right] d\mu\right| \\
&\quad\leq \left|\int_{A_\varepsilon} \left[\frac{1}{n}\sum_{i=1}^n F \circ Z_i - \int F\,dP\right] d\mu\right| + \left|\int_{A_\varepsilon^c} \frac{1}{n}\sum_{i=1}^n \left(F \circ Z_i - \int F\,dP\right) d\mu\right| \\
&\quad\overset{(4.16)}{\leq} \varepsilon + \int_{A_\varepsilon^c} \left|\frac{1}{n}\sum_{i=1}^n F \circ Z_i - \int F\,dP\right| d\mu \\
&\quad\leq 2\varepsilon.
\end{aligned}$$

Therefore, for all $B \subset \mathcal{B}$,

$$\lim_{n\to\infty} \left|\frac{1}{n}\sum_{i=1}^n \int I_B \circ Z_i - \int I_B\,dP\right| d\mu = 0. \qquad \square$$

Now, we define L-risk-consistency of SVMs for AMS stochastic processes with different dependence structures. Using the empirical SVM $f_{L,\mathbb{P}_{\mathbf{W}_n},\lambda_n}$ as an estimate for the true solution f_{L,P,λ_n}, it is important to show consistency. In our case we require the L-risk-consistency of the SVM. That is the stochastic convergence of the risk computed for the empirical SVM to the Bayes risk.

Definition 4.4.2 (L-risk-consistency of support vector machines) *Let P be a probability distribution on a Polish space $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ and let $L\colon \mathcal{X} \times \mathcal{Y} \times \mathbb{R} \to [0,\infty)$ be a loss function. Then a learning method is said to be L-risk-consistent for P if, for every $\varepsilon > 0$,*

$$R_{L,P}(f_{\mathbb{P}_{\mathbf{W}_n},\lambda_n}) \to R_{L,P}^* \quad \textit{in probability, } n \to \infty,$$

where $R_{L,P}^*$ is the Bayes risk. Moreover, the learning method is called universally L-risk-consistent if it is L-risk-consistent for all probability distributions P on $\mathcal{Z}$.

For general stochastic processes this definition is only reasonable if the distribution P is related to the process $(Z_i)_{i\in\mathbb{N}}$. Hence processes which are asymptotically mean stationary and therefore provide the existence of a limiting distribution P are regarded. Then the risk is computed with respect to this distribution.

As explained in Section 4.1, instead of searching for the minimizer f among all measurable functions, SVMs are computed for a RKHS of functions. Informally spoken, we can still achieve convergence against the Bayes risk if the RKHS is large enough. The term "large enough" can for example be defined via universal kernels:

Definition 4.4.3 (Universal kernel) *A continuous kernel k on a compact metric space $(\mathcal{X}, d_{\mathcal{X}})$ is called universal if the RKHS H of k is dense in $C(\mathcal{X})$, i. e. for every function $g \in C(\mathcal{X})$ and all $\varepsilon > 0$ there exists $f \in H$ such that*

$$\|f - g\|_\infty \leq \varepsilon.$$

For universal kernels, Steinwart and Christmann (2008, Corollary 5.28) shows that the Bayes risk can be approximated by the minimal risk computed over all functions in the RKHS, for continuous integrable Nemitsky losses, in particular for Lipschitz continuous losses. The Gaussian RBF kernel and the exponential kernel, for example are universal, see Steinwart and Christmann (2008, Corollary 4.58).

For the i.i.d. case universal L-risk-consistency, also for non-compact input spaces $\mathcal{X}$, is for example established in Steinwart (2002), Zhang (2004), Steinwart (2005), and Christmann and Steinwart (2007). Moreover learning rates for SVMs corresponding to different loss functions can be found in Koltchinskii and Beznosova (2005), Steinwart and Scovel (2007), Blanchard et al. (2008), for classification, and in De Vito et al. (2005), Steinwart and Christmann (2011), and Eberts and Steinwart (2011) for regression. Unfortunately, universal consistency, that is consistency for general stochastic processes, can not be achieved without any assumptions for the non-i.i.d. case. In Steinwart et al. (2009, Theorem 2.2), for example, it is shown that it is impossible to show universal consistency for processes which satisfy a law of large numbers for events. Therefore special classes of dependencies, namely α-mixing, $\mathcal{C}$-mixing and weakly dependent processes (in the sense of Doukhan and Louhichi (1999)) are investigated throughout the next chapters.

Often consistency is proven via concentration inequalities, for example using Hoeffding's inequality or Bernstein-type inequalities, in order to additionally achieve learning rates, see Boucheron et al. (2013) for an overview of different concentration inequalities. For the non-i.i.d. case there has been some effort in showing consistency of SVMs via concentration inequalities. A dependence notion which is widely used, is the mixing notion. In Xu and Chen (2008), Sun and Wu (2009), Pan and Xiao (2009) consistency and learning rates are achieved for SVMs using the least squares loss function under α-mixing conditions.

The article Hang and Steinwart (2015) shows a Bernstein-type inequality for α- and $\mathcal{C}$-mixing, which implies the consistency of empirical risk minimization (ERM) algorithms and support vector machines, while Kulkarni et al. (2005) establishes consistency of regularized boosting algorithms for β-mixing sequences. Zou et al. (2009a) gives generalization bounds of ERM for α-mixing sequences and Zou et al. (2009b) provides consistency of the ERM algorithm for uniformly ergodic Markov chains. Based on Markov's inequality, Steinwart et al. (2009) presents consistency of support vector machines for α-mixing processes, which provide an uniform decay of the mixing coefficients and a stability assumption. In Smale and Zhou (2009) consistency for regularized online learning for Markov chains is given. Fender (2003) examines ERM for martingale and mixingale structures. As the properties of these dependence structures are hard to transfer to the loss function, the dependence structures therein are not defined for the observations but for the losses. Moreover a Bernstein-type inequality for weakly dependent random variables is shown in Doukhan and Louhichi (1999, Theorem 4.5). As we have a slightly different consistency result we do not work with this Bernstein-type inequality but show consistency for weakly dependent processes using Markov's inequality.

The consistency for SVM estimators is established, in Theorem 4.4.4, under common assumptions on the reproducing kernel k and on the loss function L. Moreover we assume almost sure convergence of $\frac{1}{n^{1-r}} \sum_{i=1}^{n} L_{f_n} \circ Z_i - \int L_{f_n} \circ Z_i \, d\mu \longrightarrow 0$, $n \to \infty$, for some $0 < r < \frac{1}{2}$ and for uniformly bounded functions f_n, $n \in \mathbb{N}$. The proof is based on Markov's inequality and the convergence above. Contrarily to Steinwart et al. (2009), where consistency of the SVM estimator for α-mixing processes is shown in a similar way, we do not need strict assumptions on the stochastic process or on the decay of the mixing coefficients, but require the stochastic process to be asymptotically mean stationary, as long as the convergence assumption is fulfilled. On the other hand our restriction on the input space $\mathcal{X}$ to be compact is stronger than in Steinwart et al. (2009). Theorem 4.4.10 shows that an assumption on the α-mixing process, as used in Steinwart et al. (2009, Theorem 3.4), already leads to the required convergence (4.17) and therefore guarantees consistency without additional assumptions on the process. In particular we need the AMS property as well as either the convergence assumption or certain dependence conditions on the stochastic process. Theorem 4.4.6 and Theorem 4.4.12 show that several weakly dependent (in the sense of Doukhan and Louichi) and $\mathcal{C}$-mixing processes satisfy (4.17). That is again, convergence is given by conditions on the weak dependence coefficients respectively on the mixing coefficients. Moreover Theorem 4.4.12 covers Lipschitz continuous loss functions, whereas the L-risk-consistency which is shown in Hang and Steinwart (2015) via the Bernstein-type

inequality, applies for the least squares loss, which is not Lipschitz continuous. Also (4.19) covers more processes, as the assumptions on the process are weaker, see also Theorem 4.4.12. But, unlike Hang and Steinwart (2015), we do not achieve learning rates.
For notational convenience we write: $L_f \circ Z_i := L(X_i, Y_i, f(X_i))$ and $\mathbb{E}_\mu f \circ Z_i = \int f \, dP^i$.

Theorem 4.4.4 (L-risk-consistency of support vector machines) *Let $(\Omega, \mathcal{A}, \mu)$ be a probability space, let $(\mathcal{Z}, d_\mathcal{Z}) = (\mathcal{X} \times \mathcal{Y}, d_{\mathcal{X}\times\mathcal{Y}})$ be a separable metric space, let $(\mathcal{X}, d_\mathcal{X})$ be compact, and let $\mathcal{Y} \subset \mathbb{R}$ be closed. Let $L\colon \mathcal{X} \times \mathcal{Y} \times \mathbb{R} \to [0, \infty)$ be a convex and Lipschitz continuous loss function which is also continuous in (x, y) for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$, and $\sup_{(x,y)\in\mathcal{X}\times\mathcal{Y}} L(x, y, 0) \leq S$, for some constant $S \in (0, \infty)$. Moreover let H be the reproducing kernel Hilbert space of a bounded continuous kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ and let $(Z_i)_{i\in\mathbb{N}}$, $Z_i : \Omega \to \mathcal{Z}$ be an asymptotically mean stationary stochastic process. Let $0 < r < \frac{1}{2}$ be a real number such that:*

$$\frac{1}{n^{1-r}} \sum_{i=1}^{n} \left(L_{f_n} \circ Z_i - \int L_{f_n} \circ Z_i \, d\mu \right) \longrightarrow 0 \quad \text{almost surely}, \quad n \to \infty, \ f_n \in \mathcal{G}, \tag{4.17}$$

where $\mathcal{G}$ is any uniformly bounded subset of functions $f \in H$, i. e. there is a constant $M > 0$ such that $\|f\|_H \leq M$ for all $f \in \mathcal{G}$.
Let $(\lambda_n)_{n\in\mathbb{N}} \subset (0, \infty)$ such that $\lambda_n \to 0$ and $\lambda_n n^r \to \infty$, and let the sequences $(f_{\frac{1}{n}\sum P^i, \lambda_n})_{n\in\mathbb{N}}$ and $(f_{\mathbb{P}_{\mathbf{W}_n(\omega)}, \lambda_n})_{n\in\mathbb{N}}$ be bounded for all $\omega \in \Omega$, i. e. there are constants $M, \tilde{M} > 0$ such that $\|f_{\frac{1}{n}\sum P^i, \lambda_n}\|_H \leq M$ and $\|f_{\mathbb{P}_{\mathbf{W}_n(\omega)}, \lambda_n}\|_H \leq \tilde{M}$, $n \in \mathbb{N}$.

Then:

$$R_{L,P}(f_{\mathbb{P}_{\mathbf{W}_n}, \lambda_n}) \to R^*_{L,P,H} \quad \text{in probability}, n \to \infty, \tag{4.18}$$

where $R^*_{L,P,H} := \inf_{f\in H} \int L(x, y, f(x)) \, dP$ is the Bayes risk over H.

Remark 4.4.5 *For practical purposes, convexity and Lipschitz continuity are common assumptions on the loss function L.*
Moreover the continuity assumption in (x, y) on the loss function L is not restrictive. For example, every supervised, distance-based continuous loss is also continuous in (y, t). As $(y, t) \mapsto y - t$ is continuous and $\psi(r)$ is continuous the composition is also continuous. The same applies for continuous margin-based loss functions, as again $(y, t) \mapsto yt$ is continuous. As we also assume the loss function L to be continuous in the last argument we implicitly ensure the continuity of the representing function ψ.
The assumption on the uniform boundedness of the sequences of SVMs $(f_{\frac{1}{n}\sum P^i, \lambda_n})_{n\in\mathbb{N}}$ and $(f_{\mathbb{P}_{\mathbf{W}_n(\omega)}, \lambda_n})_{n\in\mathbb{N}}$ for all $\omega \in \Omega$ with respect to $\|\cdot\|_H$, however is not easy to check in practice.

Proof of Theorem 4.4.4: With help of the triangle inequality we split the proof in two parts:

$$\begin{aligned}&\left|R_{L,P}(f_{\mathbb{P}_{\mathbf{W}_n},\lambda_n}) - R^*_{L,P,H}\right| \\ &\leq \underbrace{\left|R_{L,P}(f_{\mathbb{P}_{\mathbf{W}_n},\lambda_n}) - R_{L,P}(f_{\frac{1}{n}\sum_{i=1}^n P^i,\lambda_n})\right|}_{I} + \underbrace{\left|R_{L,P}(f_{\frac{1}{n}\sum_{i=1}^n P^i,\lambda_n}) - R^*_{L,P,H}\right|}_{II} \end{aligned} \tag{4.19}$$

where $\frac{1}{n}\sum_{i=1}^n P^i = \frac{1}{n}\sum_{i=1}^n Z_i(\mu)$.

Part I: The first part of the proof shows the convergence in probability of term I in (4.19).

By assumption the kernel k is bounded. Therefore $f \in H$ is bounded, see Steinwart and Christmann (2008, Lemma 4.23). Hence, the function $L(\cdot,\cdot,f(\cdot))$ satisfies for all $f \in H$:

$$\begin{aligned}\sup_{(x,y)\in\mathcal{X}\times\mathcal{Y}} L(x,y,f(x)) &\leq \sup_{(x,y)\in\mathcal{X}\times\mathcal{Y}} |L(x,y,f(x)) - L(x,y,0)| + \sup_{(x,y)\in\mathcal{X}\times\mathcal{Y}} |L(x,y,0)| \\ &\leq |L|_1\|f\|_\infty + S < \infty,\end{aligned}$$

because $\sup_{(x,y)\in\mathcal{X}\times\mathcal{Y}} L(x,y,0) \leq S$ by assumption. Using (4.5), $\|f\|_\infty \leq \|k\|_\infty\|f\|_H$, we have:

$$\sup_{(x,y)\in\mathcal{X}\times\mathcal{Y}} L(x,y,f(x)) \leq S + |L|_1\|f\|_H\|k\|_\infty < \infty. \tag{4.20}$$

Moreover, this yields $R_{L,Q}(f) < \infty$ for all probability measures $Q \in \mathcal{M}(\mathcal{Z})$, $f \in H$, and in particular the existence of the risk $R_{L,\frac{1}{n}\sum P^i}(0)$ for every $\frac{1}{n}\sum_{i=1}^n P^i$.
Additionally for a uniformly bounded (with respect to $\|\cdot\|_H$) class of functions $\mathcal{G} \subset H$, this yields the existence of a constant $C_L > 0$ such that

$$\sup_{(x,y)\in\mathcal{X}\times\mathcal{Y}} L(x,y,f(x)) \leq C_L.$$

That is $L(\cdot,\cdot,f(\cdot))$ is uniformly bounded for all $f \in G$.
According to the Lipschitz continuity of L we have:

$$\begin{aligned}&\mu\left(\left\{\omega\in\Omega \mid \left|R_{L,P}(f_{\mathbb{P}_{\mathbf{W}_n}(\omega),\lambda_n}) - R_{L,P}(f_{\frac{1}{n}\sum P^i,\lambda_n})\right| \geq \varepsilon\right\}\right) \\ &\qquad\leq \mu\Big(\Big\{\omega\in\Omega \mid |L|_1 \left\|f_{\mathbb{P}_{\mathbf{W}_n}(\omega),\lambda_n} - f_{\frac{1}{n}\sum P^i,\lambda_n}\right\|_\infty \geq \varepsilon\Big\}\Big) \\ &\qquad\overset{(4.5)}{\leq} \mu\left(\left\{\omega\in\Omega \mid |L|_1\|k\|_\infty \left\|f_{\mathbb{P}_{\mathbf{W}_n}(\omega),\lambda_n} - f_{\frac{1}{n}\sum P^i,\lambda_n}\right\|_H \geq \varepsilon\right\}\right).\end{aligned}$$

Markov's inequality, see for example Hoffmann-Jørgensen (1994, Theorem 3.9), and the boundedness of the functions $f \in H$, due to the boundedness of the kernel, see (Steinwart and Christmann, 2008, Lemma 4.23), yields:

$$\mu\left(\left\{\omega \in \Omega \;\middle|\; |L|_1 \left\| f_{\mathbb{P}_{\mathbf{W}_n}(\omega),\lambda_n} - f_{\frac{1}{n}\sum P^i,\lambda_n} \right\|_H \geq \varepsilon \right\}\right) \leq \frac{|L|_1^2 \|k\|_\infty^2}{\varepsilon^2} \mathbb{E}_\mu \left\| f_{\mathbb{P}_{\mathbf{W}_n},\lambda_n} - f_{\frac{1}{n}\sum P^i,\lambda_n} \right\|_H^2 . \tag{4.21}$$

Now, as $\mathcal{X}$ is compact and therefore separable, the RKHS H is separable, see Steinwart and Christmann (2008, Lemma 4.33). According to the generalized representer theorem in Steinwart and Christmann (2008, Theorem 5.10) and due to the Lipschitz continuity of the loss function L, there is a function $h_Q : \mathcal{Z} \to \mathbb{R}$, $Q \in \mathcal{M}(\mathcal{Z})$, which is element of the subdifferential (see Definition A8) $\partial L(x, y, \cdot)$ of $L(x, y, f_{Q,\lambda_n}(x))$, such that:

$$f_{Q,\lambda_n} = -\frac{1}{2\lambda_n} \mathbb{E}_Q \left(h_Q \Phi\right), \quad \text{for all } Q \in \mathcal{M}(\mathcal{Z}).$$

Here $\Phi : \mathcal{X} \to H$ denotes again the canonical feature map of the kernel k and the integral with respect to Q is a Bochner integral. In particular we have

$$f_{\mathbb{P}_{\mathbf{W}_n},\lambda_n} = -\frac{1}{2\lambda_n} \mathbb{E}_{\mathbb{P}_{\mathbf{W}_n}} \left(h_{\mathbb{P}_{\mathbf{W}_n}} \Phi\right) \text{ and } f_{\frac{1}{n}\sum P^i,\lambda_n} = -\frac{1}{2\lambda_n} \mathbb{E}_{\frac{1}{n}\sum P^i} \left(h_{\frac{1}{n}\sum P^i} \Phi\right). \tag{4.22}$$

Hence,

$$\begin{aligned}
\left\| f_{\mathbb{P}_{\mathbf{W}_n},\lambda_n} - f_{\frac{1}{n}\sum P^i,\lambda_n} \right\|_H^2 &= \langle f_{\mathbb{P}_{\mathbf{W}_n},\lambda_n} - f_{\frac{1}{n}\sum P^i,\lambda_n}, f_{\mathbb{P}_{\mathbf{W}_n},\lambda_n} - f_{\frac{1}{n}\sum P^i,\lambda_n} \rangle_H \\
&\overset{(4.22)}{=} \langle f_{\mathbb{P}_{\mathbf{W}_n},\lambda_n} - f_{\frac{1}{n}\sum P^i,\lambda_n}, -\frac{1}{2\lambda_n} \mathbb{E}_{\mathbb{P}_{\mathbf{W}_n}} \left(h_{\mathbb{P}_{\mathbf{W}_n}} \Phi\right) + \frac{1}{2\lambda_n} \mathbb{E}_{\frac{1}{n}\sum P^i} \left(h_{\frac{1}{n}\sum P^i} \Phi\right) \rangle_H \\
&= \frac{1}{2\lambda_n} \langle f_{\mathbb{P}_{\mathbf{W}_n},\lambda_n} - f_{\frac{1}{n}\sum P^i,\lambda_n}, \mathbb{E}_{\frac{1}{n}\sum P^i} \left(h_{\frac{1}{n}\sum P^i} \Phi\right) \rangle_H \\
&\quad - \frac{1}{2\lambda_n} \langle f_{\mathbb{P}_{\mathbf{W}_n},\lambda_n} - f_{\frac{1}{n}\sum P^i,\lambda_n} \mathbb{E}_{\mathbb{P}_{\mathbf{W}_n}} \left(h_{\mathbb{P}_{\mathbf{W}_n}} \Phi\right) \rangle_H .
\end{aligned} \tag{4.23}$$

Now the reproducing property

$$f(x) = \langle f, \Phi(x) \rangle_H = \langle f, k(\cdot, x) \rangle_H , \; x \in \mathcal{X}, \; f \in H, \tag{4.24}$$

of the kernel and (4.23) yields

$$\begin{aligned}&\left\| f_{\mathbb{P}_{\mathbf{W}_n},\lambda_n} - f_{\frac{1}{n}\sum P^i,\lambda_n} \right\|_H^2 \\ &= \frac{1}{2\lambda_n}\left(\mathbb{E}_{\frac{1}{n}\sum P^i}\left(h_{\frac{1}{n}\sum P^i}\left(f_{\mathbb{P}_{\mathbf{W}_n},\lambda_n} - f_{\frac{1}{n}\sum P^i,\lambda_n}\right)\right) + \mathbb{E}_{\mathbb{P}_{\mathbf{W}_n}}\left(h_{\mathbb{P}_{\mathbf{W}_n}}\left(f_{\frac{1}{n}\sum P^i,\lambda_n} - f_{\mathbb{P}_{\mathbf{W}_n},\lambda_n}\right)\right)\right).\end{aligned} \tag{4.25}$$

As the functions $h_{\mathbb{P}_{\mathbf{W}_n}}, h_{\frac{1}{n}\sum P^i} : \mathcal{X}\times\mathcal{Y}\to\mathbb{R}$ are elements of the corresponding subdifferentials $\partial L(x,y,\cdot)$, the following inequalities apply, see Denkowski et al. (2003, Definition 5.3.20),

$$\begin{aligned} h_{\mathbb{P}_{\mathbf{W}_n}}(x,y)\left(f_{\frac{1}{n}\sum P^i,\lambda_n}(x) - f_{\mathbb{P}_{\mathbf{W}_n},\lambda_n}(x)\right) &\le L(x,y,f_{\frac{1}{n}\sum P^i,\lambda_n}(x)) - L(x,y,f_{\mathbb{P}_{\mathbf{W}_n},\lambda_n}(x)),\\ h_{\frac{1}{n}\sum P^i}(x,y)\left(f_{\mathbb{P}_{\mathbf{W}_n},\lambda_n}(x) - f_{\frac{1}{n}\sum P^i,\lambda_n}(x)\right) &\le L(x,y,f_{\mathbb{P}_{\mathbf{W}_n},\lambda_n}(x)) - L(x,y,f_{\frac{1}{n}\sum P^i,\lambda_n}(x)),\end{aligned}$$

for all $(x,y)\in\mathcal{X}\times\mathcal{Y}$. Using this, (4.25) gives:

$$\begin{aligned}&\mathbb{E}_{\frac{1}{n}\sum P^i}\left(h_{\frac{1}{n}\sum P^i}\left(f_{\mathbb{P}_{\mathbf{W}_n},\lambda_n} - f_{\frac{1}{n}\sum P^i,\lambda_n}\right)\right) + \mathbb{E}_{\mathbb{P}_{\mathbf{W}_n}}\left(h_{\mathbb{P}_{\mathbf{W}_n}}\left(f_{\frac{1}{n}\sum P^i,\lambda_n} - f_{\mathbb{P}_{\mathbf{W}_n},\lambda_n}\right)\right)\\ &\le \mathbb{E}_{\frac{1}{n}\sum P^i}\left(L_{f_{\mathbb{P}_{\mathbf{W}_n},\lambda_n}} - L_{f_{\frac{1}{n}\sum P^i,\lambda_n}}\right) + \mathbb{E}_{\mathbb{P}_{\mathbf{W}_n}}\left(L_{f_{\frac{1}{n}\sum P^i,\lambda_n}} - L_{f_{\mathbb{P}_{\mathbf{W}_n},\lambda_n}}\right)\\ &= \frac{1}{n}\sum_{i=1}^n\left(\int L_{f_{\mathbb{P}_{\mathbf{W}_n},\lambda_n}}\,dP^i - L_{f_{\mathbb{P}_{\mathbf{W}_n},\lambda_n}}\circ Z_i + L_{f_{\frac{1}{n}\sum P^i,\lambda_n}}\circ Z_i - \int L_{f_{\frac{1}{n}\sum P^i,\lambda_n}}\,dP^i\right).\end{aligned} \tag{4.26}$$

Applying (4.23), (4.25), and (4.26) to (4.21) we have:

$$\begin{aligned}&\mu\left(\left\{\omega\in\Omega \mid |L|_1\|k\|_\infty \left\| f_{\mathbb{P}_{\mathbf{W}_n(\omega)},\lambda_n} - f_{\frac{1}{n}\sum P^i,\lambda_n}\right\|_H \ge \varepsilon\right\}\right)\\ &\overset{(4.21)}{\le} \frac{|L|_1^2\|k\|_\infty^2}{\varepsilon^2}\mathbb{E}_\mu\left\| f_{\mathbb{P}_{\mathbf{W}_n},\lambda_n} - f_{\frac{1}{n}\sum P^i,\lambda_n}\right\|_H^2\\ &\overset{(4.23)}{\le} \frac{|L|_1^2\|k\|_\infty^2}{2\varepsilon^2\lambda_n}\mathbb{E}_\mu\left[\langle f_{\mathbb{P}_{\mathbf{W}_n},\lambda_n} - f_{\frac{1}{n}\sum P^i,\lambda_n}, \mathbb{E}_{\frac{1}{n}\sum P^i} h_{\frac{1}{n}\sum P^i}\Phi\rangle_H\right.\\ &\qquad\qquad \left. -\langle f_{\mathbb{P}_{\mathbf{W}_n},\lambda_n} - f_{\frac{1}{n}\sum P^i,\lambda_n}, \mathbb{E}_{\mathbb{P}_{\mathbf{W}_n}} h_{\mathbb{P}_{\mathbf{W}_n}}\Phi\rangle_H\right]\\ &\overset{(4.25),(4.26)}{\le} \frac{|L|_1^2\|k\|_\infty^2}{2\varepsilon^2\lambda_n}\mathbb{E}_\mu\left[\frac{1}{n}\sum_{i=1}^n\left(\int L_{f_{\mathbb{P}_{\mathbf{W}_n},\lambda_n}}\,dP^i - L_{f_{\mathbb{P}_{\mathbf{W}_n},\lambda_n}}\circ Z_i\right.\right.\\ &\qquad\qquad \left.\left. +L_{f_{\frac{1}{n}\sum P^i,\lambda_n}}\circ Z_i - \int L_{f_{\frac{1}{n}\sum P^i,\lambda_n}(x)}\,dP^i\right)\right]\\ &= \frac{|L|_1^2\|k\|_\infty^2}{2\varepsilon^2\lambda_n}\mathbb{E}_\mu\left[\frac{1}{n}\sum_{i=1}^n\int L_{f_{\mathbb{P}_{\mathbf{W}_n},\lambda_n}}\,dP^i - \frac{1}{n}\sum_{i=1}^n L_{f_{\mathbb{P}_{\mathbf{W}_n},\lambda_n}}\circ Z_i\right]\end{aligned}$$

$$= \frac{|L|_1^2 \|k\|_\infty^2}{2\varepsilon^2 \lambda_n n^r} \mathbb{E}_\mu \left[\frac{1}{n^{1-r}} \sum_{i=1}^n \left(\int L_{f_{\mathbb{P}_{\mathbf{W}_n},\lambda_n}} \, dP^i - L_{f_{\mathbb{P}_{\mathbf{W}_n},\lambda_n}} \circ Z_i \right) \right].$$

Contrarily to $f_{\frac{1}{n}\sum_{i=1}^n P^i,\lambda_n}$, the function $f_{\mathbb{P}_{\mathbf{W}_n},\lambda_n}$ is a random element with respect to μ.

By assumption the kernel k is continuous. Therefore every $f \in H$ is continuous, see Berlinet and Thomas-Agnan (2004, Theorem 17), in particular every SVM $f_{\mathbb{P}_{\mathbf{W}_n},\lambda_n}$.
Let $\mathcal{K} := \left\{ f_{\mathbb{P}_{\mathbf{W}_{n(\omega)}},\lambda_n}, \omega \in \Omega \mid n \in \mathbb{N} \right\}$ be the set of support vector machines for the probability measures $\mathbb{P}_{\mathbf{W}_{n(\omega)}}$, $\omega \in \Omega$, $n \in \mathbb{N}$. By assumption the sequence $(f_{\mathbb{P}_{\mathbf{W}_{n(\omega)}},\lambda_n})_{n\in\mathbb{N}}$ is bounded by $\tilde{M}$, for all $\omega \in \Omega$, and therefore $\mathcal{K}$ is a uniformly bounded subset of H.

The reproducing property of the kernel yields the equicontinuity of the functions $f \in \mathcal{K}$: Let $d_{\mathcal{X}}$ be the metric on $\mathcal{X}$. By assumption the kernel k is continuous, that is, for every $\varepsilon > 0$, there is $\delta > 0$ such that for all $x' \in \mathcal{X}$:

$$d_{\mathcal{X}}(x, x') \leq \delta \quad \Rightarrow \quad \|k(\cdot, x) - k(\cdot, x')\|_H \leq \varepsilon.$$

Due to the reproducing property of the kernel, see (4.24), we have for $x' \in \mathcal{X}$ with $d_{\mathcal{X}}(x, x') \leq \delta$:

$$\begin{aligned} |f(x) - f(x')| &\overset{(4.24)}{=} |\langle f, k(\cdot, x)\rangle_H - \langle f, k(\cdot, x')\rangle_H| \; = |\langle f, k(\cdot, x) - k(\cdot, x')\rangle_H| \\ &\leq \|f\|_H \|k(\cdot, x) - k(\cdot, x')\|_H \; \leq \|f\|_H \varepsilon. \end{aligned}$$

And by assumption $\|f\|_H$ is bounded by $\tilde{M}$, hence

$$|f(x) - f(x')| \; \leq \|f\|_H \varepsilon \leq \tilde{M}\varepsilon.$$

Hence $\mathcal{K} \subset C(\mathcal{X})$ is equicontinuous. As $\mathcal{X}$ is compact and $\mathcal{K}$ uniformly bounded by assumption, the Theorem of Arzelà-Ascoli, see for example Dudley (1989, Theorem 2.4.7), states: $\mathcal{K}$ is totally bounded with respect to $\|\cdot\|_\infty$ on $\mathcal{X}$. That is for every $\varepsilon > 0$ there is a finite subset $K \subset \mathcal{K}$ such that for every $f \in \mathcal{K}$ there is $g_\varepsilon \in K$ such that $\|f - g_\varepsilon\|_\infty \leq \varepsilon$.

In particular, for every $n \in \mathbb{N}$ there is a finite subset K_n of $\mathcal{K}$ such that for all $n \in \mathbb{N}$ and for all functions $f_{\mathbb{P}_{\mathbf{W}_{n(\omega)}},\lambda_n}$, there is a function $g_{n,\omega} \in K$ such that

$$\|f_{\mathbb{P}_{\mathbf{W}_{n(\omega)}},\lambda_n} - g_{n,\omega}\|_\infty \leq \frac{1}{n^r}. \tag{4.27}$$

Note, that $g_{n,\omega}$ depends on n and ω as it is the corresponding function to $f_{\mathbb{P}_{\mathbf{W}_n(\omega)},\lambda_n}$, but is an element of a finite subset $K_n \subset \mathcal{K}$. And remember, that the loss function L is Lipschitz continuous.

Then,

$$
\begin{aligned}
&\frac{|L|_1^2\|k\|_\infty^2}{2\varepsilon^2\lambda_n n^r}\mathbb{E}_\mu\left[\frac{1}{n^{1-r}}\sum_{i=1}^n\left(\int L_{f_{\mathbb{P}_{\mathbf{W}_n},\lambda_n}}\,dP^i - L_{f_{\mathbb{P}_{\mathbf{W}_n},\lambda_n}}\circ Z_i\right)\right]\\
&\quad= \frac{|L|_1^2\|k\|_\infty^2}{2\varepsilon^2\lambda_n n^r}\mathbb{E}_\mu\left[\frac{1}{n^{1-r}}\sum_{i=1}^n\left[\left(\int L_{f_{\mathbb{P}_{\mathbf{W}_n},\lambda_n}}\,dP^i - \int L_{g_{n,\omega}}\,dP^i\right)\right.\right.\\
&\qquad\qquad\left.\left.+\left(\int L_{g_{n,\omega}}\,dP^i - L_{g_{n,\omega}}\circ Z_i\right) + \left(L_{g_{n,\omega}}\circ Z_i - L_{f_{\mathbb{P}_{\mathbf{W}_n},\lambda_n}}\circ Z_i\right)\right]\right]\\
&\quad\le \frac{|L|_1^2\|k\|_\infty^2}{2\varepsilon^2\lambda_n n^r}\mathbb{E}_\mu\left[\frac{1}{n^{1-r}}\sum_{i=1}^n\left(\left|\int L_{f_{\mathbb{P}_{\mathbf{W}_n},\lambda_n}}\,dP^i - \int L_{g_{n,\omega}}\,dP^i\right|\right.\right.\\
&\qquad\qquad\left.\left.+\left(\int L_{g_{n,\omega}}\,dP^i - L_{g_{n,\omega}}\circ Z_i\right) + \left|L_{g_{n,\omega}}\circ Z_i - L_{f_{\mathbb{P}_{\mathbf{W}_n},\lambda_n}}\circ Z_i\right|\right)\right]\\
&\quad\le \frac{|L|_1^2\|k\|_\infty^2}{2\varepsilon^2\lambda_n n^r}\mathbb{E}_\mu\left[\frac{1}{n^{1-r}}\sum_{i=1}^n\left(\int |L|_1\|f_{\mathbb{P}_{\mathbf{W}_n},\lambda_n} - L_{g_{n,\omega}}\|_\infty\,dP^i\right.\right.\\
&\qquad\qquad\left.\left.+\int L_{g_{n,\omega}}\,dP^i - L_{g_{n,\omega}}\circ Z_i + |L|_1\|g_n\circ Z_i - f_{\mathbb{P}_{\mathbf{W}_n},\lambda_n}\circ Z_i\|_\infty\right)\right]\\
&\quad\overset{(4.27)}{\le} \frac{|L|_1^2\|k\|_\infty^2}{2\varepsilon^2\lambda_n n^r}\mathbb{E}_\mu\left[\frac{1}{n^{1-r}}\sum_{i=1}^n\left(\frac{2|L|_1}{n^r} + \int L_{g_{n,\omega}}\,dP^i - L_{g_{n,\omega}}\circ Z_i\right)\right]\\
&\quad\le \frac{|L|_1^3\|k\|_\infty^2}{\varepsilon^2\lambda_n n^r} + \frac{|L|_1^2\|k\|_\infty^2}{2\varepsilon^2\lambda_n n^r}\mathbb{E}_\mu\max_{g\in K_n}\left[\frac{1}{n^{1-r}}\sum_{i=1}^n\left(\int L_g\,dP^i - L_g\circ Z_i\right)\right].
\end{aligned}
$$

Now, Assumption (4.17) yields the existence of a set $N \subset \Omega$ with $\mu(N) = 0$ for every arbitrary sequence $(f_n)_{n\in\mathbb{N}}$, $f_n \in H$, which is uniformly bounded, such that:

$$
\frac{1}{n^{1-r}}\sum_{i=1}^n\left(\int L_{f_n}\,dP^i - L_{f_n}\circ Z_i(\omega)\right) \longrightarrow 0 \text{ for all } \omega\in\Omega\backslash N,\ n\to\infty.
$$

Choose $f_n := \arg\max_{g\in K^n}\frac{1}{n^{1-r}}\sum_{i=1}^n\left(\int L_g\,dP^i - L_g\circ Z_i\right)$, $i\in\mathbb{N}$. By construction K_n, $n\in\mathbb{N}$, are subsets of $\mathcal{K}$, and therefore for every $n\in\mathbb{N}$ uniformly bounded by the same constant. Hence the sequence $(f_n)_{n\in\mathbb{N}}$ is uniformly bounded and a subset of H.

Then,

$$\begin{aligned}
&\mathbb{E}_\mu\left[\frac{1}{n^{1-r}}\sum_{i=1}^{n}\left(\int L_{f_n}\,dP^i - L_{f_n}\circ Z_i\right)\right] \\
&\quad=\int_{\Omega\setminus N}\left[\frac{1}{n^{1-r}}\sum_{i=1}^{n}\left(\int L_{f_n}\,dP^i - L_{f_{n,}}\circ Z_i\right)\right]d\mu \overset{(4.17)}{\longrightarrow} 0,\ n\to\infty
\end{aligned} \tag{4.28}$$

and, due to the implications given above and the assumption $\lambda_n n^r \to \infty,\ n\to\infty$:

$$\begin{aligned}
&\mu\left(\left\{\omega\in\Omega \;\middle|\; |L|_1\|k\|_\infty\left\|f_{\mathbb{P}_{\mathbf{W}_n(\omega)},\lambda_n} - f_{\frac{1}{n}\sum P^i,\lambda_n}\right\|_H \geq \varepsilon\right\}\right) \\
&\leq \quad \frac{|L|_1^3\|k\|_\infty^2}{\varepsilon^2\lambda_n n^r} + \frac{|L|_1^2\|k\|_\infty^2}{2\varepsilon^2\lambda_n n^r}\mathbb{E}_\mu\left[\frac{1}{n^{1-r}}\sum_{i=1}^{n}\left(\int L_{f_{\mathbb{P}_{\mathbf{W}_n},\lambda_n}}\,dP^i - L_{f_{\mathbb{P}_{\mathbf{W}_n},\lambda_n}}\circ Z_i\right)\right] \\
&\leq \quad \frac{|L|_1^3\|k\|_\infty^2}{\varepsilon^2\lambda_n n^r} + \frac{|L|_1^2\|k\|_\infty^2}{2\varepsilon^2\lambda_n n^r}\mathbb{E}_\mu\left[\frac{1}{n^{1-r}}\sum_{i=1}^{n}\left(\int L_{f_n}\,dP^i - L_{f_n}\circ Z_i\right)\right] \\
&\overset{(4.28)}{\longrightarrow} \quad 0,\ n\to\infty.
\end{aligned}$$

This proves part I.

The next part proves the convergence of the term in part II of (4.19):

$$\left|R_{L,P}(f_{\frac{1}{n}\sum_{i=1}^n P^i,\lambda_n}) - R^*_{L,P,H}\right| \longrightarrow 0,\ n\to\infty.$$

First we show that there is a weakly convergent subsequence of $\left(f_{\frac{1}{n}\sum P^i,\lambda_n}\right)_{n\in\mathbb{N}}$ converging to the Bayes decision function f^* in H, then we conclude the strong convergence of $f_{\frac{1}{n}\sum P^i,\lambda_n}$ to f^* and therefore the convergence of the risks.

By assumption the sequence $(f_{\frac{1}{n}\sum P^i,\lambda_n})_{n\in\mathbb{N}}$ is uniformly bounded, i.e. $\|f_{\frac{1}{n}\sum P^i,\lambda_n}\|_H \leq M$. Since H is a Hilbert space and therefore reflexive, see Dunford and Schwartz (1958, Theorem II.4.6), there exists, according to Dunford and Schwartz (1958, Theorem II.3.28), a subsequence $(f_{\frac{1}{n_k}\sum P^i,\lambda_{n_k}})_{n_k\in\mathbb{N}}$ which converges weakly in H. i. e. there exists $\tilde{f}\in H$ such that

$$\langle f_{\frac{1}{n_k}\sum P^i,\lambda_{n_k}}, f\rangle_H \longrightarrow \langle\tilde{f}, f\rangle_H,\ n_k\to\infty, \tag{4.29}$$

for all $f\in H$, see Dunford and Schwartz (1958, Definition 3.25). Moreover Dunford and Schwartz (1958, Lemma II.3.27) yields

$$\|\tilde{f}\|_H \leq \liminf_{n_k\to\infty}\|f_{\frac{1}{n_k}\sum P^i,\lambda_{n_k}}\|_H. \tag{4.30}$$

The sequence $\left(\|f_{\frac{1}{n_k}\sum P^i,\lambda_{n_k}}\|_H\right)_{k\in\mathbb{N}}$ is bounded by assumption. As it is a sequence in $\mathbb{R}$, the Bolzano-Weierstrass theorem yields the existence of a convergent subsequence of $\left(\|f_{\frac{1}{n_k}\sum P^i,\lambda_{n_k}}\|_H\right)_{k\in\mathbb{N}}$. Hence there exists a weakly convergent subsequence, which additionally to (4.30) possesses the following property:

$$\|f_{\frac{1}{n_{k_l}}\sum P^i,\lambda_{n_{k_l}}}\|_H \longrightarrow c, \text{ for a constant } c > 0. \tag{4.31}$$

Now, (4.30) yields for this sub-subsequence:

$$\|\tilde{f}\|_H \leq c. \tag{4.32}$$

Following the Riesz' Representation theorem, see for example Conway (1985, Theorem 3.4), the weak convergence in (4.29) is equivalent to:

$$\lim_{n_{k_l}\to\infty} h^*(f_{\frac{1}{n_{k_l}}\sum P^i,\lambda_{n_{k_l}}}) \longrightarrow h^*(\tilde{f}), \quad \text{for all } h^* \in H^*,$$

where H^* denotes the dual space of H. As the Dirac functional $\delta_x(f) = f(x)$ is continuous on H, see Berlinet and Thomas-Agnan (2004, Lemma 8), it is an element of H^*, see Dudley (1989, Theorem 6.1.2). Then the above convergence implies for all $x \in \mathcal{X}$,

$$f_{\frac{1}{n_{k_l}}\sum P^i,\lambda_{n_{k_l}}}(x) = \delta_{f_{\frac{1}{n_{k_l}}\sum P^i,\lambda_{n_{k_l}}}}(x) \longrightarrow \delta_{\tilde{f}}(x) = \tilde{f}(x),\ n_{k_l} \to \infty$$

i. e. the pointwise convergence of $f_{\frac{1}{n_{k_l}}\sum P^i,\lambda_{n_{k_l}}}(x)$ to $\tilde{f}(x)$, $x \in \mathcal{X}$.

As the kernel k is continuous, f is continuous, see Berlinet and Thomas-Agnan (2004, Theorem 17).

Due to the assumptions on the continuity of the loss function L for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$, the function $L \circ f$ is continuous. Then the dominated convergence theorem, see for example Hoffmann-Jørgensen (1994, Theorem 3.6), yields:

$$\begin{aligned}
\lim_{n_{k_l}\to\infty} R_{L,P}(f_{\frac{1}{n_{k_l}}\sum P^i,\lambda_{n_{k_l}}}) &= \lim_{n_{k_l}\to\infty} \int L_{f_{\frac{1}{n_{k_l}}\sum P^i,\lambda_{n_{k_l}}}}\, dP \qquad (4.33)\\
&= \int \lim_{n_{k_l}\to\infty} L(x, y, f_{\frac{1}{n_{k_l}}\sum P^i,\lambda_{n_{k_l}}}(x))\, dP(x, y)\\
&= \int_{\mathcal{X}\times\mathcal{Y}} L(x, y, \tilde{f}(x))\, dP(x, y) \;=\; R_{L,P}(\tilde{f}). \qquad (4.34)
\end{aligned}$$

Now we show the convergence of the risks

$$R_{L,\frac{1}{n_{k_l}}\sum P^i}(f_{\frac{1}{n_{k_l}}\sum P^i,\lambda_{n_{k_l}}}) - R_{L,P}(f_{\frac{1}{n_{k_l}}\sum P^i,\lambda_{n_{k_l}}}) \longrightarrow 0,\ n_{k_l} \to \infty.$$

Regard the set $\mathcal{U} := \left\{ f_{\frac{1}{n}\sum P^i,\lambda_n} : \mathcal{X} \to \mathbb{R},\ n \in \mathbb{N} \right\}$ of support vector machines for the probability measures $\frac{1}{n}\sum P^i$, $n \in \mathbb{N}$. The same argumentation as in part I shows the equicontinuity of $\mathcal{U}$. As $\mathcal{U}$ is uniformly bounded and $\mathcal{X}$ compact by assumption, the Theorem of Arzelà-Ascoli, see e. g. Dudley (1989, Theorem 2.4.7), states the uniform boundedness of $\mathcal{U}$ with respect to $\|\cdot\|_\infty$. That is for every $\varepsilon > 0$ there is a finite dense subset $U \subset \mathcal{U}$ such that for every $f \in \mathcal{U}$ there is $g_\varepsilon \in U$ such that $\|f - g_\varepsilon\|_\infty \leq \varepsilon$.

Hence the triangle inequality yields:

$$\begin{aligned}
&\left| R_{L,\frac{1}{n_{k_l}}\sum P^i}(f_{\frac{1}{n_{k_l}}\sum P^i,\lambda_{n_{k_l}}}) - R_{L,P}(f_{\frac{1}{n_{k_l}}\sum P^i,\lambda_{n_{k_l}}}) \right| \\
&= \left| \frac{1}{n_{k_l}} \sum_{i=1}^{n_{k_l}} \int_{\mathcal{Z}} L_{f_{\frac{1}{n_{k_l}}\sum P^i,\lambda_{n_{k_l}}}}\, dP^i - \int_{\mathcal{Z}} L_{f_{\frac{1}{n_{k_l}}\sum P^i,\lambda_{n_{k_l}}}}\, dP \right| \\
&= \left| \frac{1}{n_{k_l}} \sum_{i=1}^{n_{k_l}} \int_{\mathcal{Z}} \left(L_{f_{\frac{1}{n_{k_l}}\sum P^i,\lambda_{n_{k_l}}}} - L_{g_\varepsilon} \right) dP^i \right| + \left| \frac{1}{n_{k_l}} \sum_{i=1}^{n_{k_l}} \int_{\mathcal{Z}} L_{g_\varepsilon}\, dP^i - \int_{\mathcal{Z}} L_{g_\varepsilon}\, dP \right| \\
&\quad + \left| \int_{\mathcal{Z}} \left(L_{g_\varepsilon} - L_{f_{\frac{1}{n_{k_l}}\sum P^i,\lambda_{n_{k_l}}}} \right) dP \right|.
\end{aligned}$$

Due to the Lipschitz continuity of L in the last argument and the approximation of the SVM by g_ε we obtain

$$\begin{aligned}
&\left| \frac{1}{n_{k_l}} \sum_{i=1}^{n_{k_l}} \int_{\mathcal{Z}} \left(L_{f_{\frac{1}{n_{k_l}}\sum P^i,\lambda_{n_{k_l}}}} - L_{g_\varepsilon} \right) dP^i \right| + \left| \frac{1}{n_{k_l}} \sum_{i=1}^{n_{k_l}} \int_{\mathcal{Z}} L_{g_\varepsilon}\, dP^i - \int_{\mathcal{Z}} L_{g_\varepsilon}\, dP \right| \\
&\qquad + \left| \int_{\mathcal{Z}} \left(L_{g_\varepsilon} - L_{f_{\frac{1}{n_{k_l}}\sum P^i,\lambda_{n_{k_l}}}} \right) dP \right| \\
&\overset{L \text{ Lipschitz}}{\leq} \frac{1}{n_{k_l}} \sum_{i=1}^{n_{k_l}} \int_{\mathcal{Z}} |L|_1 \| f_{\frac{1}{n_{k_l}}\sum P^i,\lambda_{n_{k_l}}} - g_\varepsilon \|_\infty\, dP^i + \left| \frac{1}{n_{k_l}} \sum_{i=1}^{n_{k_l}} \int_{\mathcal{Z}} L_{g_\varepsilon}\, dP^i - \int_{\mathcal{Z}} L_{g_\varepsilon}\, dP \right| \\
&\qquad + \int_{\mathcal{Z}} |L|_1 \| g_\varepsilon - f_{\frac{1}{n_{k_l}}\sum P^i,\lambda_{n_{k_l}}} \|_\infty\, dP
\end{aligned}$$

$$\begin{aligned} &\overset{\|f-g_\varepsilon\|_\infty \leq \varepsilon}{\leq} && 2|L|_1\varepsilon + \left| \frac{1}{n_{k_l}} \sum_{i=1}^{n_{k_l}} \int_{\mathcal{Z}} L_{g_\varepsilon} \, dP^i - \int_{\mathcal{Z}} L_{g_\varepsilon} \, dP \right| \\ &\leq && 2|L|_1\varepsilon + \max_{g \in U} \left| \frac{1}{n_{k_l}} \sum_{i=1}^{n_{k_l}} \int_{\mathcal{Z}} L_g \, dP^i - \int_{\mathcal{Z}} L_g \, dP \right|. \end{aligned} \tag{4.35}$$

By assumption L is continuous in $(x, y) \in \mathcal{X} \times \mathcal{Y}$ and $g_\varepsilon \in U$ is continuous by construction. Hence $L_{g_\varepsilon} : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ is continuous in $(x, y) \in \mathcal{X} \times \mathcal{Y}$, and due to (4.20) L_{g_ε} is bounded, even uniformly bounded, as $\mathcal{U}$ is. Since additionally $(Z_i)_{i \in \mathbb{N}}$ is asymptotically mean stationary by assumption, Lemma 3.2.7 yields:

$$\left| \frac{1}{n_{k_l}} \sum_{i=1}^{n_{k_l}} \int_{\mathcal{Z}} L_{g_\varepsilon} \, dP^i - \int_{\mathcal{Z}} L_{g_\varepsilon} \, dP \right| \longrightarrow 0, \quad n_{k_l} \to \infty.$$

And as U is a finite set:

$$\max_{g \in U} \left| \frac{1}{n_{k_l}} \sum_{i=1}^{n_{k_l}} \int_{\mathcal{Z}} L_g \, dP^i - \int_{\mathcal{Z}} L_g \, dP \right| \longrightarrow 0, \quad n_{k_l} \to \infty.$$

Applying this to (4.35) shows:

$$\left| R_{L, \frac{1}{n_k} \sum P^i}(f_{\frac{1}{n_{k_l}} \sum P^i, \lambda_{n_{k_l}}}) - R_{L,P}(f_{\frac{1}{n_{k_l}} \sum P^i, \lambda_{n_{k_l}}}) \right| \longrightarrow 0, \; n_{k_l} \longrightarrow \infty. \tag{4.36}$$

Now we consider the minimal risk $R^*_{L,P,H}$ over functions f in H. By definition of $R^*_{L,P,H}$ we have for all $\tilde{f} \in H$:

$$\begin{aligned} 0 \leq R_{L,P}(\tilde{f}) - R^*_{L,P,H} &\overset{(4.32)}{\leq} \lambda c^2 + R_{L,P}(\tilde{f}) - R^*_{L,P,H} \\ &\overset{(4.31),(4.34)}{=} \lim_{n_{k_l} \to \infty} \lambda_{n_{k_l}} \| f_{\frac{1}{n_{k_l}} \sum P^i, \lambda_{n_{k_l}}} \|^2_H + R_{L,P}(f_{\frac{1}{n_{k_l}} \sum P^i, \lambda_{n_{k_l}}}) - R^*_{L,P,H} \\ &\overset{(4.36)}{=} \lim_{n_{k_l} \to \infty} \lambda_{n_{k_l}} \| f_{\frac{1}{n_{k_l}} \sum P^i, \lambda_{n_{k_l}}} \|^2_H + R_{L, \frac{1}{n_{k_l}} \sum P^i}(f_{\frac{1}{n_{k_l}} \sum P^i, \lambda_{n_{k_l}}}) - R^*_{L,P,H} \\ &= \lim_{n_{k_l} \to \infty} \inf_{f \in H} \lambda_{n_{k_l}} \|f\|^2_H + R_{L, \frac{1}{n_{k_l}} \sum P^i}(f) - R^*_{L,P,H}. \end{aligned}$$

For fixed $f \in H$, regard the functions $\lambda \mapsto \lambda \|f\|^2_H + R_{L, \frac{1}{n_{k_l}} \sum P^i}(f)$, $\lambda > 0$ and $Q \mapsto \lambda_{n_{k_l}} \|f\|^2_H + R_{L,Q}(f)$, $Q \in \mathcal{M}(\mathcal{Z})$. As $\lambda \mapsto \lambda \|f\|^2_H + R_{L, \frac{1}{n_{k_l}} \sum P^i}(f)$ is a linear function in λ, it is continuous. Moreover for every sequence $(Q_n)_{n \in \mathbb{N}} \subset (\mathcal{M}(\mathcal{Z}^n))_{n \in \mathbb{N}}$, $Q_n \rightsquigarrow Q$ is equivalent to $\int g \, dQ_n \to \int g \, dQ$ for every continuous and bounded function g by definition. Hence $Q \mapsto \lambda_{n_{k_l}} \|f\|^2_H + R_{L,Q}(f)$ is continuous for fixed $f \in H$, since L is continuous by

assumption and bounded by (4.20).
Therefore $\lim_{n_{k_l}\to\infty} \inf_{f\in H} \left(\lambda_{n_{k_l}} \|f\|_H^2 + R_{L,\frac{1}{n_{k_l}}\sum P^i}(f)\right) - R^*_{L,P,H}$ is upper semicontinuous, see Denkowski et al. (2003, Theorem 1.1.36).
Then,

$$\begin{aligned}
&\lim_{n_{k_l}\to\infty} \inf_{f\in H} \left(\lambda_{n_{k_l}} \|f\|_H^2 + R_{L,\frac{1}{n_{k_l}}\sum P^i}(f)\right) - R^*_{L,P,H} \\
&\qquad = \limsup_{n_{k_l}\to\infty} \inf_{f\in H} \left(\lambda_{n_{k_l}} \|f\|_H^2 + R_{L,\frac{1}{n_{k_l}}\sum P^i}(f)\right) - R^*_{L,P,H} \\
&\qquad \leq \inf_{f\in H} \left(\lambda \|f\|_H^2 + R_{L,P}(f)\right) - R^*_{L,P,H},
\end{aligned}$$

as $\lambda_{n_{k_l}} \longrightarrow \lambda$, $n_{k_l} \to \infty$, and due to the AMS property of the process and Lemma 3.2.7, which implies $\frac{1}{n_{k_l}} \sum_{i=1}^{n_{k_l}} \int f\, dP^i \longrightarrow \int f\, dP$, for f bounded and continuous.
Now $\lambda = 0$ yields:

$$0 \leq R_{L,P}(\tilde{f}) - R^*_{L,P,H} \leq \inf_{f\in H} R_{L,P}(f) - R^*_{L,P,H}.$$

Hence $\tilde{f} = \arg\inf_{f\in H} R_{L,P}(f)$, i. e. $\tilde{f}$ is a minimizer of $R_{L,P,H}$. Then Steinwart and Christmann (2008, Lemma 5.16) yields $\|\tilde{f}\|_H \geq \|f^*\|_H$, where f^* is the Bayes decision function in H.

With $\lambda > 0$ we can conclude:

$$0 \leq R_{L,P}(\tilde{f}) - R^*_{L,P,H} \leq \inf_{f\in H} \lambda \|f\|_H^2 + R_{L,P}(f) - R^*_{L,P,H},$$

that is $\tilde{f}$ is a minimizer of $\lambda\|f\|_H^2 + R_{L,P}(f)$ and therefore $\|\tilde{f}\|_H \leq \|f^*\|_H$.

Combining these two observations, we have: $\|\tilde{f}\|_H = \|f^*\|_H$ and due to the uniqueness of the Bayes decision function in H, see Steinwart and Christmann (2008, Lemma 5.16): $\tilde{f} = f^*$.

Furthermore, the preliminary considerations show,

$$\begin{aligned}
0 \leq \lambda c^2 + R_{L,P}(\tilde{f}) - R^*_{L,P,H} &\leq \inf_{f\in H} \lambda \|f\|_H^2 + R_{L,P}(f) - R^*_{L,P,H} \\
&= \lambda \|\tilde{f}\|_H^2 + R_{L,P}(\tilde{f}) - R^*_{L,P,H}.
\end{aligned}$$

Thus $\|\tilde{f}\|_H^2 \geq c^2$, with (4.32) actually equality is given. The convergence in (4.31) then

yields $\lim_{n_{k_l}\to\infty} \|f_{\frac{1}{n_{k_l}}\sum P^i\lambda_{n_{k_l}}}\|_H \longrightarrow \|\tilde{f}\|_H = c$, $n_{k_l} \to \infty$. Convergence of the norm and the weak convergence in (4.29) imply:

$$\begin{aligned}\left\|f_{\frac{1}{n_{k_l}}\sum P^i,\lambda_{n_{k_l}}} - \tilde{f}\right\|_H^2 &= \|\tilde{f}\|_H^2 - 2\left\langle f_{\frac{1}{n_{k_l}}\sum P^i,\lambda_{n_{k_l}}}, \tilde{f}\right\rangle_H + \|f_{\frac{1}{n_{k_l}}\sum P^i,\lambda_{n_{k_l}}}\|_H^2 \\ &\overset{(4.29)}{\longrightarrow} \|\tilde{f}\|_H^2 - 2\left\langle \tilde{f}, \tilde{f}\right\rangle_H + \|\tilde{f}\|_H^2 = 0,\ n_{k_l} \to \infty.\end{aligned}$$

Therefore, $f_{\frac{1}{n_{k_l}}\sum P^i,\lambda_{n_{k_l}}} \longrightarrow \tilde{f} = f^*$, $n_{k_l} \longrightarrow \infty$ in H. Assume that $f_{\frac{1}{n}\sum P^i,\lambda_n}$ does not converge to f^*, hence there exists a subsequence which does not converge to f^*. As this subsequence is bounded, the result above shows the existence of a sub-subsequence which converges strongly to f^*. This leads to a contradiction. Hence,

$$\begin{aligned}\left|R_{L,P}(f_{\frac{1}{n}\sum P^i,\lambda_n}) - R^*_{L,P,H}\right| &= \left|\int_{\mathcal{Z}} L\circ f_{\frac{1}{n}\sum P^i,\lambda_n}\, dP - \int_{\mathcal{Z}} L\circ f^*\, dP\right| \\ &\leq |L|_1\|k\|_\infty \int_{\mathcal{Z}} \|f_{\frac{1}{n}\sum P^i,\lambda_n} - f^*\|_H\, dP \\ &\longrightarrow 0,\ n\to\infty.\end{aligned}$$

Using part I and part II yields the assertion of Theorem 4.4.4. □

The next section links assumptions on the dependence structure of a stochastic process to (4.17). If such stochastic processes are additionally asymptotically mean stationary, the SVM estimator is consistent. For weakly dependent processes and $\mathcal{C}$-mixing processes, the speed of the decay of the dependence coefficients does not influence the choice of the sequence $(\lambda_n)_{n\in\mathbb{N}}$ directly. If the coefficients are summable, the consistency is ensured for every $0 < r < \frac{1}{2}$, as long as $\lambda_n n^r \longrightarrow \infty$, $n \to \infty$. In all cases the condition on the sequence $(\lambda_n)_{n\in\mathbb{N}}$ nearly equals the condition for the i.i.d. case, which is $r = \frac{1}{2}$, that is the SVM estimator is rather robust against violations of the i.i.d. assumption.

4.4.1 Weakly dependent processes

The first example are weakly dependent processes, introduced by Doukhan and Louhichi (1999) and Bickel and Bühlmann (1999). They satisfy the almost sure convergence in (4.17) as long as their dependence coefficient $\varepsilon(\ell)$, $\ell \in \mathbb{N}$, decreases fast enough to be summable.

Theorem 4.4.6 *Let $(\Omega, \mathcal{A}, \mu)$ be a probability space, let $(\mathcal{X}, d_{\mathcal{X}})$ be compact and $(\mathcal{Y}, |\cdot|) \subset \mathbb{R}$, $\mathcal{Y}$ closed, and let $(\mathcal{Z}, d_{\mathcal{Z}}) = (\mathcal{X}\times\mathcal{Y}, d_{\mathcal{X}\times\mathcal{Y}})$, with $d_{\mathcal{X}\times\mathcal{Y}}((x,y),(x',y')) = d(x,x') + |y-y'|$*

be a separable, metric space. Let $L\colon \mathcal{X} \times \mathcal{Y} \times \mathbb{R} \to [0,\infty)$ *be a distance-based, Lipschitz continuous loss function with* $\sup_{(x,y)\in\mathcal{X}\times\mathcal{Y}} L(x,y,0) \leq S$*, for a constant* $S \in (0,\infty)$*, and* $|L|_1 > 0$*. Let* k *be a continuous, bounded kernel with corresponding RKHS* H*. Further let* $(Z_i)_{i\in\mathbb{N}}$*,* $Z_i : \Omega \to \mathcal{X} \times \mathcal{Y}$*, be a* η*-,* λ*-,* ζ*-,* κ*- or* θ*-weakly dependent stochastic process with* $\sum_{l=1}^{\infty} \varepsilon(\ell) < \infty$*. Then for* $0 < r < \frac{1}{2}$

$$\frac{1}{n^{1-r}} \sum_{i=1}^{n} \left(L_{f_n} \circ Z_i - \int L(x,y,f_n(x))\, dP^i(x,y) \right) \longrightarrow 0 \quad \text{almost surely, } n \to \infty,\ f_n \in \mathcal{G},$$

where $\mathcal{G} \subset H$ *is any uniformly bounded subset of functions* $f \in H$*, i. e. there is a constant* $M > 0$ *such that* $\|f\|_H \leq M$ *for all* $f \in \mathcal{G}$*.*

The metric $d_{\mathcal{X}\times\mathcal{Y}}$ on the space $\mathcal{X}\times\mathcal{Y}$ is chosen for technical reasons. Due to the definition of weak dependence, Lipschitz continuous functions are needed. In the proof of Theorem 4.4.6 the Lipschitz continuity of a distance-based loss function L_f with respect to $d_{\mathcal{X}\times\mathcal{Y}}$ is shown if L, respectively ψ, and f are Lipschitz continuous. It is tempting to expect the p-product metric $d((x,y),(x',y')) = \sqrt{d_{\mathcal{X}}(x,x')^2 + d_{\mathcal{Y}}(y,y')^2}$ instead of $d_{\mathcal{X}\times\mathcal{Y}} = d(x,x') + |y-y'|$, but we need to choose a metric for which we can guarantee the Lipschitz continuity of $L(\cdot,\cdot,f(\cdot))\colon \mathcal{X}\times\mathcal{Y} \to \mathbb{R}$. If $\mathcal{X} \subset \mathbb{R}$ we can for example use the Euclidean metric, due to the strong equivalence of the metrics on $\mathbb{R}$ or, for $\mathcal{X} \subset \mathbb{R}^d$, we can choose a p-product metric on $\mathbb{R}^{d+1}$.

Without loss of generality $|L|_1 > 0$ is assumed. $|L|_1 = 0$ implies the function L to be constant with respect to the last argument, hence $L(x,y,t) = L(x,y,t')$ for all $t,t' \in \mathbb{R}$. This leads to a risk which does not depend on the prediction $f(x)$ and therefore is not useful for practical purposes.

In order to prove Theorem 4.4.6 the following technical Lemmata are needed. As we are going to use a moment inequality of the maximum of a sum of random variables by Serfling (1970) (see Theorem A10) we use Lemma 4.4.7 to introduce a certain function h, depending on the joint distribution of arbitrary random variables Z_i, $i \in \mathbb{N}$. Let $P_{a,n}$ be the joint distribution of $(Z_{a+1},\ldots,Z_{a+n})$, $a,n \in \mathbb{N}$, $n > 1$.

Lemma 4.4.7 *Let* $Z_1,\ldots,Z_{a+n}$ *be square integrable random variables and* $f : \mathcal{Z} \to \mathbb{R}$ *a measurable, square integrable function. Then the function*

$$h_{a,n}(P_{a.n}) := \sum_{i=a+1}^{a+n} \mathrm{Var}(f\circ Z_i) + 2\sum_{i=a+1}^{a+n-1}\sum_{j=i+1}^{a+n} |\mathrm{Cov}(f\circ Z_i, f\circ Z_j)|,\ a,n\in\mathbb{N},\ n>1, \quad (4.37)$$

has the following properties for $a, k, n \in \mathbb{N}$, $n, k > 1$*:*

$$h_{a,k}(P_{a,k}) + h_{a+k,n}(P_{a+k,n}) \leq h_{a,k+n}(P_{a,k+n}),$$

and

$$\mathbb{E}_\mu \left(\sum_{i=a+1}^{a+n} (f \circ Z_i - \mathbb{E}_\mu f \circ Z_i) \right)^2 \leq h_{a,n}(P_{a,n}).$$

Proof: The proof of both properties is straightforward. Let $a, k, n \in \mathbb{N}$ and $n, k > 1$. Then

$$\begin{aligned} h_{a,k}(P_{a,k}) + h_{a+k,n}(P_{a+k,n}) &= \sum_{i=a+1}^{a+k} \mathrm{Var}(f \circ Z_i) + 2 \sum_{i=a+1}^{a+k-1} \sum_{j=i+1}^{a+k} |\mathrm{Cov}(f \circ Z_i, f \circ Z_j)| \\ &\quad + \sum_{i=a+k+1}^{a+k+n} \mathrm{Var}(f \circ Z_i) + 2 \sum_{i=a+k+1}^{a+k+n-1} \sum_{j=i+1}^{a+k+n} |\mathrm{Cov}(f \circ Z_i, f \circ Z_j)| \\ &\leq \sum_{i=a+1}^{a+k+n} \mathrm{Var}(f \circ Z_i) + 2 \sum_{i=a+1}^{a+k+n-1} \sum_{j=i+1}^{a+k+n} |\mathrm{Cov}(f \circ Z_i, f \circ Z_j)| \\ &= h_{a,k+n}(P_{a,k+n}). \end{aligned}$$

And

$$\begin{aligned} \mathbb{E}_\mu \left(\sum_{i=a+1}^{a+n} (f \circ Z_i - \mathbb{E}_\mu f \circ Z_i) \right)^2 &= \sum_{i=a+1}^{a+n} \mathrm{Var}(f \circ Z_i) + 2 \sum_{i=a+1}^{a+n-1} \sum_{j=i+1}^{a+n} \mathrm{Cov}(f \circ Z_i, f \circ Z_j) \\ &\leq h_{a,n}(P_{a,n}). \end{aligned}$$

□

Lemma 4.4.7 is also used to prove the almost sure convergence in (4.17) of $\mathcal{C}$-mixing and α-mixing random variables in Theorem 4.4.12 and 4.4.10.
Lemma 4.4.8 gives a bound on the Lipschitz constant of a family of Lipschitz continuous and equicontinuous (see Definition A9) functions.

Lemma 4.4.8 *Let* $\mathcal{G}$ *be a family of equicontinuous and Lipschitz continuous functions* $f_i : \mathcal{X} \to \mathbb{R}$, $i \in \mathbb{N}$, *where* $(\mathcal{X}, d_\mathcal{X})$ *is a metric space. Then*

$$\sup_{f_i \in \mathcal{G}} \{|f_i|_1\} < \infty.$$

Proof: By assumption f_i is Lipschitz continuous. Hence for every $f_i \in \mathcal{G}$ there is $|f_i|_1 := \sup_{x,x' \in \mathcal{X}} \frac{|f_i(x)-f_i(x')|}{d_{\mathcal{X}}(x,x')}$, $x \neq x'$, such that

$$|f_i(x) - f_i(x')| \leq |f_i|_1 d_{\mathcal{X}}(x, x'), \quad \text{for all } x, x' \in \mathcal{X}.$$

Note that we do not need to consider functions with Lipschitz constant $|f_i|_1 = 0$, as they do not change the supremum in Lemma 4.4.8. Due to the Lipschitz continuity of f_i the function is also uniformly continuous, as for every $\varepsilon > 0$, $\delta_i := \frac{\varepsilon}{|f_i|_1}$ gives:

$$d_{\mathcal{X}}(x, x') \leq \frac{\varepsilon}{|f_i|_1} \quad \Rightarrow \quad |f_i(x) - f_i(x')| \leq |f_i|_1 d_{\mathcal{X}}(x, x') \leq \varepsilon, \quad \text{for all } x, x' \in \mathcal{X}.$$

In particular the definition of the Lipschitz constant as smallest upper bound on $\frac{|f_i(x)-f_i(x')|}{d_{\mathcal{X}}(x,x')}$, $x \neq x'$, implies, that there is no $\delta > \frac{\varepsilon}{|f_i|_1}$ such that the above equation applies.

Moreover the set $\mathcal{G}$ is equicontinuous by assumption, hence for every $\varepsilon > 0$, for every $x \in \mathcal{X}$, there is $\tilde{\delta} > 0$ such that for all $x' \in \mathcal{X}$ with:

$$|x - x'| \leq \tilde{\delta} \quad \Rightarrow \quad |f_i(x) - f_i(x')| \leq \varepsilon, \text{ for every } f_i \in \mathcal{G}.$$

Due to the uniform continuity the family of functions $\mathcal{G}$ is uniformly equicontinuous. In particular $\tilde{\delta} \leq \delta_i$, $i \in \mathbb{N}$. Assume that the sequence $|f_i|_1$ is unbounded. Then we obtain $\delta = 0$, which is a contradiction to the equicontinuity of $\mathcal{G}$. Hence the set $\{|f_i|_1 \mid f_i \in \mathcal{G}\}$ is bounded. □

Now we can prove Theorem 4.4.6.

Proof of Theorem 4.4.6: We have:

$$\frac{1}{n^{1-r}} \sum_{i=1}^{n} \left(L_{f_n} \circ Z_i - \int L_{f_n} \circ Z_i \, d\mu \right) = \frac{1}{n^{1-r}} \sum_{i=1}^{n} \left(L_{f_n} \circ Z_i - \int L(x, y, f_n(x)) \, dP^i(x, y) \right). \tag{4.38}$$

Let $\mathcal{G} \subset H$ be a set of uniformly bounded functions $f \in H$. Since $\mathcal{X}$ is a compact space by assumption, Dudley (1989, Theorem 11.2.4) states, that the space of bounded Lipschitz functions $\mathrm{BL}(\mathcal{X}) = \{f \colon \mathcal{X} \to \mathbb{R} \mid f \text{ Lipschitz and } \|f\|_{\mathrm{BL}} < \infty\}$ is dense in $C(\mathcal{X})$ with respect to $\|\cdot\|_\infty$. Moreover Dudley (1989, Corollary 11.2.5) states the separability of $(C(\mathcal{X}), \|\cdot\|_\infty)$. As $\mathcal{G} \subset H \subset C(\mathcal{X})$ and as $(C(\mathcal{X}), \|\cdot\|_\infty)$ is a metric space, $\mathcal{G}$ is separable with respect to $\|\cdot\|_\infty$, see Denkowski et al. (2003, Corollary 1.4.12). Therefore the set $\mathrm{BL}(\mathcal{X}) \cap \mathcal{G}$ is dense in $\mathcal{G}$ with respect to $\|\cdot\|_\infty$. Then, for every $\rho > 0$ and for every $f_n \in \mathcal{G}$ there is

$g_{\rho,n} \in \mathrm{BL}(\mathcal{X}) \cap \mathcal{G}$ such that:

$$\|f_n - g_{\rho,n}\|_\infty \leq \rho. \tag{4.39}$$

Now, for any fixed $n \in \mathbb{N}$ and for $f_n \in \mathcal{G}$, the triangle inequality and the approximation (4.39) above yield:

$$\begin{aligned}
\left| \frac{1}{n^{1-r}} \sum_{i=1}^n \left(L_{f_n} \circ Z_i - \int L_{f_n}\, dP^i \right) \right| &\leq \frac{1}{n^{1-r}} \sum_{i=1}^n \left| L_{f_n} \circ Z_i - L_{g_{\rho,n}} \circ Z_i \right| \\
&\quad + \left| \frac{1}{n^{1-r}} \sum_{i=1}^n \left(L_{g_{\rho,n}} \circ Z_i - \int L_{g_{\rho,n}}\, dP^i \right) \right| + \frac{1}{n^{1-r}} \sum_{i=1}^n \left| \int L_{g_{\rho,n}}\, dP^i - \int L_{f_n}\, dP^i \right| \\
&\overset{L \text{ Lipschitz}}{\leq} \frac{1}{n^{1-r}} \sum_{i=1}^n |L|_1 \|f_n - g_{\rho,n}\|_\infty + \left| \frac{1}{n^{1-r}} \sum_{i=1}^n \left(L_{g_{\rho,n}} \circ Z_i - \int L_{g_{\rho,n}}\, dP^i \right) \right| \\
&\quad + \frac{1}{n^{1-r}} \sum_{i=1}^n \int |L|_1 \|f - g_{\rho,n}\|_\infty\, dP^i \\
&\overset{(4.39)}{\leq} 2n^r |L|_1 \cdot \rho + \left| \frac{1}{n^{1-r}} \sum_{i=1}^n \left(L_{g_{\rho,n}} \circ Z_i - \int L_{g_{\rho,n}}\, dP^i \right) \right|.
\end{aligned}$$

Define $\rho(n) = \frac{1}{n^{1+r}}$. Therefore the above computation leads the following bound on (4.38):

$$\begin{aligned}
&\left| \frac{1}{n^{1-r}} \sum_{i=1}^n \left(L_{f_n} \circ Z_i - \int L_{f_n}\, dP^i \right) \right| \\
&\qquad \leq \quad 2n^r |L|_1 \rho(n) + \left| \frac{1}{n} \sum_{i=1}^n \left(L_{g_{\rho(n),n}} \circ Z_i - \int L_{g_{\rho(n),n}}\, dP^i \right) \right| \\
&\qquad \overset{\rho(n)=1/n^{1+r}}{\leq} \quad \frac{2|L|_1}{n} + \left| \frac{1}{n^{1-r}} \sum_{i=1}^n \left(L_{g_{\rho(n),n}} \circ Z_i - \int L_{g_{\rho(n),n}}\, dP^i \right) \right|.
\end{aligned} \tag{4.40}$$

To show the almost sure convergence of the last part we follow the same lines as the proof of Hu et al. (2008, Theorem 1). We split the sum in two parts and show that both parts converge almost surely. For every $n > 1$ choose $s \in \mathbb{N}$ such hat $2^{s-1} < n \leq 2^s$. Then,

$$\begin{aligned}
&\left| \frac{1}{n^{1-r}} \sum_{i=1}^n \left(L_{g_{\rho(n),n}} \circ Z_i - \int L_{g_{\rho(n),n}}\, dP^i \right) \right| \\
&= \quad \frac{1}{n^{1-r}} \left| \sum_{i=1}^{2^{s-1}} \left(L_{g_{\rho(n),n}} \circ Z_i - \int L_{g_{\rho(n),n}}\, dP^i \right) + \sum_{i=2^{s-1}+1}^{n} \left(L_{g_{\rho(n),n}} \circ Z_i - \int L_{g_{\rho(n),n}}\, dP^i \right) \right|
\end{aligned}$$

$$
\begin{aligned}
\leq \quad & \frac{1}{n^{1-r}} \left| \sum_{i=1}^{2^{s-1}} \left(L_{g_{\rho(n),n}} \circ Z_i - \int L_{g_{\rho(n),n}} \, dP^i \right) \right. \\
& \left. + \max_{1 \leq q \leq 2^{s-1}} \sum_{i=2^{s-1}+1}^{2^{s-1}+q} \left(L_{g_{\rho(n),n}} \circ Z_i - \int L_{g_{\rho(n),n}} \, dP^i \right) \right| \\
\overset{n>2^{s-1}}{\leq} \quad & \underbrace{\frac{1}{2^{(s-1)\cdot(1-r)}} \left| \sum_{i=1}^{2^{s-1}} \left(L_{g_{\rho(n),n}} \circ Z_i - \int L_{g_{\rho(n),n}} \, dP^i \right) \right|}_{I} \\
& + \underbrace{\frac{1}{2^{(s-1)\cdot(1-r)}} \left| \max_{1 \leq q \leq 2^{s-1}} \sum_{i=2^{s-1}+1}^{2^{s-1}+q} \left(L_{g_{\rho(n),n}} \circ Z_i - \int L_{g_{\rho(n),n}} \, dP^i \right) \right|}_{II}. \qquad (4.41)
\end{aligned}
$$

The almost sure convergence of the terms in part I and II is shown via the boundedness of the sum of covariances $\sum_{i=a+1}^{b-1} \sum_{j=i+1}^{b} \operatorname{Cov}(L_{g_{\rho(n),n}} \circ Z_i, L_{g_{\rho(n),n}} \circ Z_j)$ for $a, b \in \mathbb{N},\ a < b-1$. In particular we show that the bound does not depend on the function $g_{\rho(n),n}$, respectively n. The next part of the proof leads to this bound.

By assumption the loss function L is Lipschitz continuous and distance-based, i. e. there exists a function $\psi : \mathcal{Y} \times \mathbb{R} \to [0, \infty)$ such that $L(x, y, t) = \psi(y - t)$, for all $(x, y, t) \in \mathcal{X} \times \mathcal{Y} \times \mathbb{R}$ and $\psi(0) = 0$. The Lipschitz continuity of L is equivalent to the Lipschitz continuity of ψ in t, see Steinwart and Christmann (2008, Lemma 2.33). Hence, for all $(x, y, t) \in \mathcal{X} \times \mathcal{Y} \times \mathbb{R}$:

$$
\begin{aligned}
\left| \psi(y - t) - \psi(y' - t') \right| = \ & \left| \psi(y - t) - \psi(y - (y - y' + t')) \right| \leq |\psi|_1 \left| t - (y - y' + t') \right| \\
\leq \ & |\psi|_1 \left| y' - y + t - t' \right| \leq |\psi|_1 \left(\left| y - y' \right| + \left| t - t' \right| \right).
\end{aligned}
$$

That is ψ is Lipschitz continuous with respect to the metric given by $d_{\mathcal{Y} \times \mathbb{R}}((y, t), (y', t')) = |y - y'| + |t - t'|$. Using the Lipschitz continuity of $g_{\rho(n),n} \in \mathrm{BL}(\mathcal{X})$ with respect to $d_{\mathcal{X}}$, we have for all $x, x' \in \mathcal{X},\ y, y' \in \mathcal{Y}$:

$$
\begin{aligned}
\left| L(x, y, g_{\rho(n),n}(x)) - L(x', y', g_{\rho(n),n}(x')) \right| &= \left| \psi\left(y - g_{\rho(n),n}(x)\right) - \psi\left(y' - g_{\rho(n),n}(x')\right) \right| \\
&\leq |\psi|_1 \left(|y - y'| + |g_{\rho(n),n}(x) - g_{\rho(n),n}(x')| \right) \\
&\overset{g_{\rho(n),n} \in \mathrm{BL}(\mathcal{X})}{\leq} |\psi|_1 \left(|y - y'| + |g_{\rho(n),n}|_1 d_{\mathcal{X}}(x, x') \right) \\
&\leq \max\{|\psi|_1 \cdot |g_{\rho(n),n}|_1, |\psi|_1\} \cdot \left(|y - y'| + d_{\mathcal{X}}(x, x') \right). \qquad (4.42)
\end{aligned}
$$

Hence the function $L_{g_{\rho(n),n}}$ is Lipschitz continuous with respect to $d_{\mathcal{X}\times\mathcal{Y}}((x,y),(x',y')) = d_{\mathcal{X}}(x,x') + |y-y'|$. Therefore the function $L_{g_{\rho(n),n}}$ is an element of $\mathcal{F}_1$, where $\mathcal{F}_1$ is the function class defined for the λ-, η-, ζ-, and κ-dependence coefficients in Section 2.1, respectively, in case of θ-dependence, the function $L_{g_{\rho(n),n}}$ belongs to both required function classes $\mathcal{F}_1$ and $\mathcal{G}_1$.
Due to the uniform boundedness of $\mathcal{G}$ and due to $g_{\rho(n),n} \in \mathrm{BL}(\mathcal{X}) \cap \mathcal{G} \subset H$ we have $\|g_{\rho(n),n}\|_H \leq M$. Now Inequality (4.20), leads to the boundedness of $L_{g_{\rho(n),n}}$ by a constant $C_L > 0$:

$$\begin{aligned} \|L_{g_{\rho(n),n}}\|_\infty \leq S + |L|_1 \|g_{\rho(n),n}\|_\infty &\overset{(4.20)}{\leq} S + |L|_1 \|g_{\rho(n),n}\|_H \|k\|_\infty \\ &\leq S + |L|_1 M \|k\|_\infty \leq C_L. \end{aligned} \tag{4.43}$$

Furthermore,

$$\mathrm{Var}(L_{g_{\rho(n),n}} \circ Z_i) \;\leq\; \|L_{g_{\rho(n),n}}\|_\infty^2 \;\leq C_L^2 \tag{4.44}$$

and $\sum_{i=1}^n \mathrm{Var}(L_{g_{\rho(n),n}} \circ Z_i) \leq C_L^2 n$. Note that the constant C_L does not depend on n and that the boundedness of $\|L_{g_{\rho(n),n}}\|_\infty$ by C_L implies the boundedness of the first element $(n=1)$ of the sequence $\left|\frac{1}{n^{1-r}} \sum_{i=1}^n \left(L_{g_{\rho(n),n}} \circ Z_i - \int L_{g_{\rho(n),n}}\, dP^i\right)\right|$ by $2C_L$.

In order to relate the covariances to the different dependence coefficients, we need to regard the function $\Psi\colon \mathcal{F}_1 \times \mathcal{F}_1 \to \mathbb{R}$, which depends on the type of weak dependence, i. e. on the dependence coefficients, see 2.1.

The function $\Psi(f,f)$ varies for the different dependence coefficients $\varepsilon(\ell)$, but always depends on $\|f\|_\infty$ and on the Lipschitz constant $|f|_1$ of f, see Doukhan and Louhichi (1999, page 12) and Section 2.1. As $\|L_{g_{\rho(n),n}}\|_\infty \leq C_L$, for all $g_{\rho(n),n} \in \mathcal{G}$, we get that, for every considered dependence coefficient, the function $\Psi(L_{g_{\rho(n),n}}, L_{g_{\rho(n),n}})$ is bounded by a constant C, depending on M, $\|k\|_\infty$ and $|L_{g_{\rho(n),n}}|_1$:

for η-weakly dependent processes we have for $f = L_{g_{\rho(n),n}}$:

$$\Psi(f,f) = 2\|f\|_\infty |f|_1 \leq 2C_L |L_{g_{\rho(n),n}}|_1;$$

for λ-weakly dependent processes we have for $f = L_{g_{\rho(n),n}}$:

$$\Psi(f,f) = 2\|f\|_\infty |f|_1 + |f|_1 |f|_1 \leq 2C_L |L_{g_{\rho(n),n}}|_1 + |L_{g_{\rho(n),n}}|_1^2;$$

for κ- and ζ-weakly dependent processes we have for $f = L_{g_{\rho(n),n}}$:

$$\Psi(f,f) = |f|_1^2 \leq |L_{g_{\rho(n),n}}|_1^2.$$

for θ-weakly dependent processes we have for $f = L_{g_{\rho(n),n}}$:

$$\Psi(f,f) = \|f\|_\infty |f|_1 \leq C_L |L_{g_{\rho(n),n}}|_1.$$

Similar to the proof of Theorem 4.4.4, the reproducing property of the kernel yields the equicontinuity of the functions $f \in \mathcal{G}$: Let $d_\mathcal{X}$ be the metric on $\mathcal{X}$. By assumption the kernel k is continuous, that is, in particular, for every $\varepsilon > 0$, there is $\delta > 0$ such that for all $x' \in \mathcal{X}$

$$d_\mathcal{X}(x,x') \leq \delta \quad \Rightarrow \quad \|k(\cdot,x) - k(\cdot,x')\|_H \leq \varepsilon.$$

Due to the reproducing property of the kernel, (4.24), for all $x' \in \mathcal{X}$ with $d_\mathcal{X}(x,x') \leq \delta$:

$$\begin{aligned} \left|f(x) - f(x')\right| &\overset{(4.24)}{=} \left|\langle f, k(\cdot,x)\rangle_H - \left\langle f, k(\cdot,x')\right\rangle_H\right| = \left|\left\langle f, k(\cdot,x) - k(\cdot,x')\right\rangle_H\right| \\ &\leq \|f\|_H \|k(\cdot,x) - k(\cdot,x')\|_H \leq M\varepsilon. \end{aligned}$$

Hence $\mathrm{BL}(\mathcal{X}) \cap \mathcal{G}$ is equicontinuous. As $(\mathcal{X}, d_\mathcal{X})$ is a compact metric space by assumption, Dudley (1989, Theorem 2.4.5) yields the uniform equicontinuity of $\mathrm{BL}(\mathcal{X}) \cap \mathcal{G}$ with respect to $\|\cdot\|_\infty$.

Due to Lemma 4.4.8 the set $\{|g_{\rho(n),n}|_1 \mid g_{\rho(n),n} \in \mathrm{BL}(\mathcal{X}) \cap \mathcal{G}\}$ of Lipschitz constants of the functions $g_{\rho(n),n}$ is uniformly bounded. Hence $\left\{|L_{g_{\rho(n),n}}|_1,\ g_{\rho(n),n} \in \mathrm{BL}(\mathcal{X}) \cap \mathcal{G}\right\}$ is uniformly bounded, see (4.42). Therefore there exists, separately for every dependence coefficient, a constant C_Ψ, depending on the kernel and the function class $\mathcal{G}$ such that $\Psi(L_{g_{\rho(n),n}}, L_{g_{\rho(n),n}}) \leq C_\Psi$ for all $n \in \mathbb{N}$. In particular C_Ψ does not depend on the choice of $g_{\rho(n),n}$, respectively of n.

Without loss of generality we assume for the next calculations that there is $(x,y) \in \mathcal{X} \times \mathcal{Y}$ such that $L_{g_{\rho(n),n}}(x,y) \neq 0$, i. e. $L_{g_{\rho(n),n}} \neq 0$. Together with the assumption $|L|_1 > 0$, this implies that $\Psi(L_{g_{\rho(n),n}}, L_{g_{\rho(n),n}}) > 0$ for all $n \in \mathbb{N}$. If $L_{g_{\rho(n),n}}$ equals the null-function, which is denoted by $L_{g_{\rho(n),n}} = 0$, the calculations in (4.48) and (4.49) on the next page are trivial.

Hence we have for $a < b-1,\ a,b \in \mathbb{N}$, and for all $n \in \mathbb{N}$ such that $L_{g_{\rho(n),n}} \neq 0$:

$$\begin{aligned}
&\sum_{i=a+1}^{b-1} \sum_{j=i+1}^{b} \operatorname{Cov}\left(L_{g_{\rho(n),n}} \circ Z_i, L_{g_{\rho(n),n}} \circ Z_j\right) \\
&\qquad \overset{\Psi>0}{\leq} \sum_{i=a+1}^{b-1} \sum_{j=i+1}^{b} \Psi(L_{g_{\rho(n),n}}, L_{g_{\rho(n),n}}) \frac{\operatorname{Cov}(L_{g_{\rho(n),n}} \circ Z_i, L_{g_{\rho(n),n}} \circ Z_j)}{\Psi(L_{g_{\rho(n),n}}, L_{g_{\rho(n),n}})} \\
&\qquad \leq \quad \Psi(L_{g_{\rho(n),n}}, L_{g_{\rho(n),n}}) \sum_{i=a+1}^{b-1} \sum_{j=i+1}^{b} \frac{\operatorname{Cov}(L_{g_{\rho(n),n}} \circ Z_i, L_{g_{\rho(n),n}} \circ Z_j)}{\Psi(L_{g_{\rho(n),n}}, L_{g_{\rho(n),n}})}. \qquad (4.45)
\end{aligned}$$

Now (4.45), the assumption on the dependence coefficients $\sum_{\ell=1}^{\infty} \varepsilon(\ell) \leq \tilde{C}$ for a constant $\tilde{C} < \infty$, and the Lipschitz continuity of $L_{g_{\rho(n),n}}$ yield for $a < b-1,\ a,b \in \mathbb{N}$, and for all $n \in \mathbb{N}$ such that $L_{g_{\rho(n),n}} \neq 0$:

$$\begin{aligned}
&\sum_{i=a+1}^{b-1} \sum_{j=i+1}^{b} \operatorname{Cov}(L_{g_{\rho(n),n}} \circ Z_i, L_{g_{\rho(n),n}} \circ Z_j) \\
&\qquad \overset{(4.45)}{\leq} \quad \Psi(L_{g_{\rho(n)}}, L_{g_{\rho(n)}}) \sum_{i=a+1}^{b-1} \sum_{j=i+1}^{b} \frac{\operatorname{Cov}(L_{g_{\rho(n),n}} \circ Z_i, L_{g_{\rho(n),n}} \circ Z_j)}{\Psi(L_{g_{\rho(n),n}}, L_{g_{\rho(n),n}})} \\
&\qquad \leq \quad C_\Psi \sum_{i=a+1}^{b-1} \sum_{j=i+1}^{b} \sup_{f \in \mathcal{F}_1} \frac{|\operatorname{Cov}(f \circ Z_i, f \circ Z_j)|}{\Psi(f,f)} \\
&\qquad \overset{(2.1)}{\leq} \quad C_\Psi \sum_{\ell=1}^{b-a-1} (b-a-\ell)\varepsilon(\ell) \\
&\qquad \overset{b-a-\ell \leq b-a}{\leq} \quad C_\Psi (b-a) \sum_{\ell=1}^{\infty} \varepsilon(\ell) \qquad (4.46) \\
&\qquad \leq \quad \tilde{C} C_\Psi (b-a). \qquad (4.47)
\end{aligned}$$

To show almost sure convergence of the term in (4.41) part I, we show for all $\varepsilon > 0$:

$$\sum_{s=1}^{\infty} \mu\left(\left\{\omega \in \Omega \;\middle|\; \frac{1}{2^{(s-1)\cdot(1-r)}} \left|\sum_{i=1}^{2^{s-1}} \left(L_{g_{\rho(n),n}} \circ Z_i(\omega) - \int L_{g_{\rho(n),n}}\, dP^i\right)\right| > \varepsilon\right\}\right) < \infty. \qquad (4.48)$$

Then the Lemma of Borel-Cantelli, see e. g. Hoffmann-Jørgensen (1994, Theorem 2.11), yields $\frac{1}{2^{(s-1)\cdot(1-r)}} \left|\sum_{i=1}^{2^{s-1}} \left(L_{g_{\rho(n),n}} \circ Z_i - \int L_{g_{\rho(n),n}}\, dP^i\right)\right|$ to 0, almost surely, $n \to \infty$.

A short note should be done on the argumentation. Remember that s is chosen such that $2^{s-1} < n \leq 2^s$. Regarding the sum over s, we do not cover every element of the sequence $\frac{1}{n^{1-r}} \sum_{i=1}^{n} \left(L_{f_n} \circ Z_i - \int L_{f_n} \, dP^i \right)$. The last computation shows that the sum of covariances does not depend on n, but only on the number of summands. To get the sequence for $n \in \mathbb{N}$, you only add, for every $s \in \mathbb{N}$, at most countable many elements, which are bounded by the given element for $s \in \mathbb{N}$. Hence, if the almost sure convergence for the sequence in s is shown, the almost sure convergence of $\frac{1}{n^{1-r}} \sum_{i=1}^{n} \left(L_{f_n} \circ Z_i - \int L_{f_n} \, dP^i \right)$ is still implied.

By Markov's inequality, see for example Hoffmann-Jørgensen (1994, Theorem 3.9), we have, for all $\varepsilon > 0$, $s > 1$ and for all $n \in \mathbb{N}$ such that $L_{g_{\rho(n),n}} \neq 0$:

$$
\begin{aligned}
&\mu \left(\left\{ \omega \in \Omega \;\middle|\; \frac{1}{2^{(s-1)\cdot(1-r)}} \left| \sum_{i=1}^{2^{s-1}} \left(L_{g_{\rho(n),n}} \circ Z_i(\omega) - \int L_{g_{\rho(n),n}} \, dP^i \right) \right| > \varepsilon \right\} \right) \\
&\overset{Markov}{\leq} \frac{1}{\varepsilon^2} \mathbb{E}_\mu \left(\frac{1}{2^{(s-1)\cdot(1-r)}} \sum_{i=1}^{2^{s-1}} \left[L_{g_{\rho(n),n}} \circ Z_i - \int L_{g_{\rho(n),n}} \, dP^i \right] \right)^2 \\
&= \frac{1}{\varepsilon^2} \left(\frac{1}{2^{(s-1)\cdot(1-r)}} \right)^2 \left[\sum_{i=1}^{2^{s-1}} \mathbb{E}_\mu (L_{g_{\rho(n),n}} \circ Z_i - \int L_{g_{\rho(n),n}} \, dP^i)^2 \right. \\
&\quad \left. + 2 \sum_{i=1}^{2^{s-1}-1} \sum_{j=i+1}^{2^{s-1}} \mathbb{E}_\mu \left((L_{g_{\rho(n),n}} \circ Z_i - \int L_{g_{\rho(n),n}} \, dP^i)(L_{g_{\rho(n),n}} \circ Z_j - \int L_{g_{\rho(n),n}} \, dP^j) \right) \right] \\
&= \frac{1}{\varepsilon^2} \left(\frac{1}{2^{(s-1)\cdot(1-r)}} \right)^2 \left[\sum_{i=1}^{2^{s-1}} \mathrm{Var}(L_{g_{\rho(n),n}} \circ Z_i) + 2 \sum_{i=1}^{2^{s-1}-1} \sum_{j=i+1}^{2^{s-1}} \mathrm{Cov}(L_{g_{\rho(n),n}} \circ Z_i, L_{g_{\rho(n),n}} \circ Z_j) \right] \\
&\overset{(4.44),(4.47)}{\leq} \frac{1}{\varepsilon^2} \left[\frac{1}{2^{(s-1)\cdot(1-2r)}} C_L^2 + 2 \frac{1}{2^{(s-1)\cdot(2-2r)}} C_\Psi \tilde{C} 2^{s-1} \right] \\
&\leq \frac{1}{\varepsilon^2} \frac{1}{2^{(s-1)\cdot(1-2r)}} \tilde{\tilde{C}},
\end{aligned}
\tag{4.49}
$$

for a constant $\tilde{\tilde{C}} := C_L^2 + 2 C_\Psi \tilde{C} > 0$.
If there exists $n \in \mathbb{N}$ such that $L_{g_{\rho(n),n}} = 0$, the calculation above easily yields

$$
\begin{aligned}
&\mu \left(\left\{ \omega \in \Omega \;\middle|\; \frac{1}{2^{(s-1)\cdot(1-r)}} \left| \sum_{i=1}^{2^{s-1}} \left(L_{g_{\rho(n),n}} \circ Z_i(\omega) - \int L_{g_{\rho(n),n}} \, dP^i \right) \right| > \varepsilon \right\} \right) \\
&\overset{L_{g_{\rho(n),n}}=0}{=} 0 \;\leq\; \frac{1}{\varepsilon^2} \frac{1}{2^{(s-1)\cdot(1-2r)}} \tilde{\tilde{C}}.
\end{aligned}
$$

For $s = 1$ we obtain

$$\mu\left(\left\{\omega \in \Omega \;\middle|\; \left|L_{g_{\rho(n),n}} \circ Z_1(\omega) - \int L_{g_{\rho(n),n}}\, dP^1\right| > \varepsilon\right\}\right) \leq \frac{1}{\varepsilon^2}\mathrm{Var}(L_{g_{\rho(n),n}} \circ Z_1) \overset{(4.44)}{\leq} \frac{C_L^2}{\varepsilon^2}. \tag{4.50}$$

As $\frac{1}{2^{1-2r}} < 1$ for all $0 < r < \frac{1}{2}$, the series above equals a geometric series and therefore is convergent:

$$\sum_{s=1}^{\infty}\left(\frac{1}{\varepsilon^2}\frac{1}{2^{(s-1)\cdot(1-2r)}}\tilde{\tilde{C}}\right) = \frac{1}{\varepsilon^2}\tilde{\tilde{C}}\sum_{s=0}^{\infty}\left(\frac{1}{2^{(1-2r)}}\right)^s < \frac{1}{\varepsilon^2}\tilde{\tilde{C}}\frac{1}{1-\frac{1}{2^{1-2r}}} \;<\; \infty.$$

Hence the term in part I in (4.41) converges almost surely to zero.

The almost sure convergence of the second part in (4.41) is shown via a maximal inequality and again the application of the Borel-Cantelli Lemma. It is to show, that for all $\varepsilon > 0$

$$\sum_{s=1}^{\infty}\mu\left(\left\{\omega \in \Omega \mid \frac{1}{2^{(s-1)\cdot(1-r)}}\max_{1\leq q\leq 2^{s-1}}\left|\sum_{i=2^{s-1}+1}^{2^{s-1}+q}\left(L_{g_{\rho(n),n}} \circ Z_i(\omega) - \int L_{g_{\rho(n),n}}\, dP^i\right)\right| > \varepsilon\right\}\right) \tag{4.51}$$

is finite.
Again Markov's inequality yields:

$$\begin{aligned}&\mu\left(\left\{\omega \in \Omega \;\middle|\; \frac{1}{2^{(s-1)\cdot(1-r)}}\left|\max_{1\leq q\leq 2^{s-1}}\sum_{i=2^{s-1}+1}^{2^{s-1}+q}\left(L_{g_{\rho(n),n}} \circ Z_i(\omega) - \int L_{g_{\rho(n),n}}\, dP^i\right)\right| > \varepsilon\right\}\right)\\ &\leq \frac{1}{\varepsilon^2}\left(\frac{1}{2^{(s-1)\cdot(1-r)}}\right)^2 \mathbb{E}_\mu\left(\max_{1\leq q\leq 2^{s-1}}\sum_{i=2^{s-1}+1}^{2^{s-1}+q}\left(L_{g_{\rho(n),n}} \circ Z_i - \int L_{g_{\rho(n),n}}\, dP^i\right)\right)^2. \end{aligned} \tag{4.52}$$

Moreover we assume $L_{g_{\rho(n),n}} \neq 0$, $n \in \mathbb{N}$, similar to the first part.

Now we can use a generalization of the Rademacher-Mensov-Inequality in Serfling (1970, Theorem A). We choose the function

$$h_{a,m}(P_{a,m}) := \sum_{a+1}^{a+m}\mathrm{Var}(f \circ Z_i) + 2\sum_{i=a+1}^{a+m-1}\sum_{j=i+1}^{a+m}|\mathrm{Cov}(f \circ Z_i, f \circ Z_j)|,$$

$a \in \mathbb{N}$, $m > 1$, which has due to Lemma 4.4.7 the required properties for Serfling (1970, Theorem A).

Hence, we have, for all $n \in \mathbb{N}$ and $s > 1$ such that $L_{g_{\rho(n),n}} \neq 0$,

$$\begin{aligned}
&\mathbb{E}_\mu \left(\max_{1 \le q \le 2^{s-1}} \sum_{i=2^{s-1}+1}^{2^{s-1}+q} \left(L_{g_{\rho(n),n}} \circ Z_i - \int L_{g_{\rho(n),n}} \, dP^i \right) \right)^2 \\
&\overset{Serfling}{\le} (\log_2(2 \cdot 2^{s-1}))^2 h_{2^{s-1},2^{s-1}}(P_{2^{s-1},2^{s-1}}) \\
&= \quad (\log_2(2 \cdot 2^{s-1}))^2 \left[\sum_{2^{s-1}+1}^{2^s} \mathrm{Var}(L_{g_{\rho(n),n}} \circ Z_i) \right. \\
&\qquad \left. + 2 \sum_{i=2^{s-1}+1}^{2^s-1} \sum_{j=i+1}^{2^s} |\mathrm{Cov}(L_{g_{\rho(n),n}} \circ Z_i, L_{g_{\rho(n),n}} \circ Z_j)| \right]. \qquad (4.53)
\end{aligned}$$

Now,

$$\log_2(2 \cdot 2^{s-1})^2 \le (1 + \log_2 2^{s-1})^2 \le C(\log_2 2^{s-1})^2, \ s > 1, \ C := 4. \qquad (4.54)$$

For $s = 1$, we have $\log_2(2 \cdot 2^{s-1}) = 1$ and

$$\begin{aligned}
&\mathbb{E}_\mu \left(\max_{1 \le q \le 2^{s-1}} \sum_{i=2^{s-1}+1}^{2^{s-1}+q} \left(L_{g_{\rho(n),n}} \circ Z_i - \int L_{g_{\rho(n),n}} \, dP^i \right) \right)^2 \\
&\qquad = \mathrm{Var}(L_{g_{\rho(2),2}} \circ Z_2) \ \le \|L_{g_{\rho(2),2}}\|_\infty^2 \overset{(4.43)}{=} C_L^2 < \infty. \qquad (4.55)
\end{aligned}$$

Then, we get with Markov's inequality (4.52) and (4.53), for all $L_{g_{\rho(n),n}} \neq 0$:

$$\begin{aligned}
&\sum_{s=1}^{\infty} \mu \left(\left\{ \omega \in \Omega \ \middle| \ \frac{1}{2^{(s-1)\cdot(1-r)}} \left| \max_{1 \le q \le 2^{s-1}} \sum_{i=2^{s-1}+1}^{2^{s-1}+q} \left(L_{g_{\rho(n),n}} \circ Z_i(\omega) - \int L_{g_{\rho(n),n}} \, dP^i \right) \right| > \varepsilon \right\} \right) \\
&\overset{(4.52),(4.53),(4.55)}{\le} \frac{C_L^2}{\varepsilon^2} + \sum_{s>1}^{\infty} \frac{1}{\varepsilon^2} \left(\frac{1}{2^{(s-1)\cdot(1-r)}} \right)^2 (\log_2(2 \cdot 2^{s-1}))^2 \cdot \left[\sum_{2^{s-1}+1}^{2^s} \mathrm{Var}(L_{g_{\rho(n),n}} \circ Z_i) \right. \\
&\qquad \left. + 2 \sum_{i=2^{s-1}+1}^{2^s-1} \sum_{j=i+1}^{2^s} |\mathrm{Cov}(L_{g_{\rho(n),n}} \circ Z_i, L_{g_{\rho(n),n}} \circ Z_j)| \right] \\
&\overset{(4.54)}{\le} \frac{C_L^2}{\varepsilon^2} + \sum_{s>1}^{\infty} \frac{C}{\varepsilon^2} \left(\frac{1}{2^{(s-1)\cdot(1-r)}} \right)^2 (\log_2 2^{s-1})^2 \left[\sum_{2^{s-1}+1}^{2^s} \mathrm{Var}(L_{g_{\rho(n),n}} \circ Z_i) \right. \\
&\qquad \left. + 2 \sum_{i=2^{s-1}+1}^{2^s-1} \sum_{j=i+1}^{2^s} |\mathrm{Cov}(L_{g_{\rho(n),n}} \circ Z_i, L_{g_{\rho(n),n}} \circ Z_j)| \right]
\end{aligned}$$

$$\begin{aligned}
&\overset{(4.44),(4.46)}{\leq} \quad \frac{C_L^2}{\varepsilon^2} + \sum_{s>1}^{\infty} \frac{C}{\varepsilon^2} \left(\frac{1}{2^{(s-1)\cdot(1-r)}} \right)^2 (\log_2 2^{s-1})^2 \left[2^{s-1} C_L^2 + 2(2^s - 2^{s-1}) \sum_{\ell=1}^{\infty} C_\Psi \varepsilon(\ell) \right] \\
&\overset{(4.47)}{\leq} \quad \frac{C_L^2}{\varepsilon^2} + \frac{C}{\varepsilon^2} \sum_{s>1}^{\infty} \frac{(\log_2 2^{s-1})^2}{(2^{(s-1)\cdot(1-r)})^2} 2^{s-1} \tilde{\tilde{C}} \;\leq\; \frac{C_L^2}{\varepsilon^2} + \frac{C'}{\varepsilon^2} \sum_{s>1}^{\infty} \frac{(\log_2 2^{s-1})^2}{2^{(s-1)\cdot(1-2r)}} \\
&\leq \quad \frac{C_L^2}{\varepsilon^2} + \frac{C'}{\varepsilon^2} \sum_{s>1}^{\infty} \frac{(s-1)^2}{2^{(s-1)\cdot(1-2r)}} \;<\; \infty \qquad (4.56)
\end{aligned}$$

for a constant $C' := C_L^2 + 2C_\Psi \sum_{\ell=1}^{\infty} \varepsilon(\ell) > 0$.

Note that the same argumentation as in part I yields for $L_{g_{\rho(n),n}} = 0$:

$$\begin{aligned}
&\mu \left(\left\{ \omega \in \Omega \;\middle|\; \frac{1}{2^{(s-1)\cdot(1-r)}} \left| \max_{1 \leq q \leq 2^{s-1}} \sum_{i=2^{s-1}+1}^{2^{s-1}+q} \left(L_{g_{\rho(n),n}} \circ Z_i(\omega) - \int L_{g_{\rho(n),n}} \, dP^i \right) \right| > \varepsilon \right\} \right) \\
&\overset{L_{g_{\rho(n),n}}=0}{=} 0 \;\leq\; \frac{C'}{\varepsilon^2} \frac{(s-1)^2}{2^{(s-1)\cdot(1-2r)}},
\end{aligned}$$

respectively for $s = 1$:

$$\mu \left(\left\{ \omega \in \Omega \;\middle|\; \left| \left(L_{g_{\rho(n),n}} \circ Z_i(\omega) - \int L_{g_{\rho(n),n}} \, dP^i \right) \right| > \varepsilon \right\} \right) \overset{L_{g_{\rho(n),n}}=0}{=} 0.$$

The convergence of the last series in (4.56) follows directly via the ratio test:

$$\frac{\frac{(s+1)^2}{(2^{s+1})^{1-2r}}}{\frac{s^2}{(2^s)^{(1-2r)}}} \leq \left(1 + \frac{1}{s} \right)^2 \frac{1}{2^{(1-2r)}}, \; 0 < r < \frac{1}{2}.$$

As $(1 + \frac{1}{s})^2 \longrightarrow 1$, $s \to \infty$ and $\frac{1}{2^{(1-2r)}} < 1$, for every $\frac{1}{2^{(1-2r)}} < a < 1$, there exists $s \in \mathbb{N}$ such that $(1 + \frac{1}{s})^2 \cdot \frac{1}{2^{(1-2r)}} < a$. Hence the series converges and we have the almost sure convergence of the term in part II in (4.41).

Then, the almost sure convergence in (4.38) is implied, $n > 1$:

$$\begin{aligned}
&\frac{1}{n^{1-r}} \sum_{i=1}^{n} \left(L_{f_n} \circ Z_i - \int L(x, y, f_n(x)) \, dP^i(x, y) \right) \\
&\qquad \overset{(4.40)}{\leq} \quad \frac{2|L|_1}{n} + \left| \frac{1}{n^{1-r}} \sum_{i=1}^{n} \left(L_{g_{\rho(n),n}} \circ Z_i - \int L_{g_{\rho(n),n}} \, dP^i \right) \right|
\end{aligned}$$

$$\overset{(4.41)}{\leq} \frac{2|L|_1}{n} + \underbrace{\frac{1}{2^{(s-1)\cdot(1-r)}} \left| \sum_{i=1}^{2^{s-1}} \left(L_{g_{\rho(n),n}} \circ Z_i - \int L_{g_{\rho(n),n}} \, dP^i \right) \right|}_{I \longrightarrow 0 \text{ almost surely}}$$

$$+ \underbrace{\frac{1}{2^{(s-1)\cdot(1-r)}} \left| \max_{1 \leq q \leq 2^{s-1}} \sum_{i=2^{s-1}+1}^{2^{s-1}+q} \left(L_{g_{\rho(n),n}} \circ Z_i - \int L_{g_{\rho(n),n}} \, dP^i \right) \right|}_{II \longrightarrow 0 \text{ almost surely}}$$

$$\longrightarrow 0 \quad \text{almost surely.}$$

This proves the assertion. □

Corollary 4.4.9 *Let $(\Omega, \mathcal{A}, \mu)$ be a probability space, let $(\mathcal{X}, d_X)$ be compact and $(\mathcal{Y}, |\cdot|) \subset \mathbb{R}$, $\mathcal{Y}$ closed, and let $(\mathcal{Z}, d_{\mathcal{Z}}) = (\mathcal{X} \times \mathcal{Y}, d_{\mathcal{X}\times\mathcal{Y}})$, $d_{\mathcal{X}\times\mathcal{Y}}((x,y),(x',y')) = d(x,x') + |y-y'|$ be a separable, metric space. Let $L\colon \mathcal{X}\times\mathcal{Y}\times\mathbb{R} \to [0,\infty)$ be a convex, distance-based and Lipschitz continuous loss function, which is additionally continuous in (x,y) for all $(x,y) \in \mathcal{X}\times\mathcal{Y}$, with $\sup_{(x,y)\in\mathcal{X}\times\mathcal{Y}} L(x,y,0) \leq S$ for some constant $S \in (0,\infty)$, and $|L|_1 > 0$. Moreover let H be a reproducing kernel Hilbert space of an universal, bounded and continuous kernel $k : \mathcal{X}\times\mathcal{X} \to \mathbb{R}$. Let $(Z_i)_{i\in\mathbb{N}}$, $Z_i : \Omega \to \mathcal{Z}$, $i \in \mathbb{N}$, be an asymptotically mean stationary, η-, λ-, ζ-, κ- or θ-weakly dependent stochastic process with dependence coefficients $\varepsilon(\ell)$ such that $\sum_{\ell=1}^{\infty} \varepsilon(\ell) < \infty$. Let $(\lambda_n)_{n\in\mathbb{N}} \subset (0,\infty)$ such that $\lambda_n \to 0$ and $\lambda_n n^r \to \infty$, for some $0 < r < \frac{1}{2}$, and let the sequences $(f_{\frac{1}{n}\sum P^i,\lambda_n})_{n\in\mathbb{N}}$ and $(f_{\mathbb{P}_{\mathbf{W}_n(\omega)},\lambda_n})_{n\in\mathbb{N}}$ be bounded for all $\omega \in \Omega$, i. e. there are constants $M, \tilde{M} > 0$ such that $\|f_{\frac{1}{n}\sum P^i,\lambda_n}\|_H \leq M$ and $\|f_{\mathbb{P}_{\mathbf{W}_n(\omega)},\lambda_n}\|_H \leq \tilde{M}$, $n \in \mathbb{N}$.*

Then:

$$R_{L,P}(f_{\mathbb{P}_{\mathbf{W}_n},\lambda_n}) \to R^*_{L,P} \quad \text{in probability, } n \to \infty.$$

That is, the SVM estimator is L-risk-consistent for asymptotically mean stationary weakly dependent processes, which have summable dependence coefficients, given the assumptions on k, L, and $\mathcal{X}$. The sequence $(\lambda_n)_{n\in\mathbb{N}}$ has to satisfy $\lambda_n n^r \to \infty$, $n \to \infty$, for some $0 < r < \frac{1}{2}$, which is stronger than the assumption $\lambda_n^2 n \to \infty$ for the i.i.d. case. In particular the proof shows that, given the assumptions, (4.17) is fulfilled for every $0 < r < \frac{1}{2}$. For $r = \frac{1}{2} - \varepsilon$, $\frac{1}{2} > \varepsilon > 0$, the assumptions on the sequence $(\lambda_n)_{n\in\mathbb{N}}$ is only slightly stronger than the assumption for the i.i.d. case. We can still weaken the assumptions on the stochastic process. As long as $\sum_{s=1}^{\infty} \frac{(s-1)^2}{2^{1-2r}} \sum \varepsilon(\ell) < \infty$, the proof can easily be adapted. Moreover, the

smaller the constant r the weaker the assumption on the process, but the sequence $(\lambda_n)_{n\in\mathbb{N}}$ has to converge appropriately slow, such that $\lambda_n^r n \to \infty$, which is a stronger assumption.

Theorem 4.4.4 shows the convergence of the empirical risk to the minimal risk with respect to the function space H. Since k is a universal kernel, even the convergence to the Bayes risk $R^*_{L,P}$ over all measurable functions $f : \mathcal{X} \to \mathbb{R}$ is ensured, see Steinwart and Christmann (2008, Corollary 5.29).

4.4.2 α-mixing processes

The next example are α-mixing processes. In Steinwart et al. (2009, Theorem 3.3) L-risk-consistency of SVMs for α-mixing processes under some assumptions on the dependence coefficient is shown. The process is assumed to be asymptotically mean stationary and α-*bi*-mixing with a special rate and needs to fulfil a stability assumption. For a compact input space $\mathcal{X}$, the next theorem shows that, to ensure consistency of SVMs, the assumptions on the stochastic process can be reduced to the AMS property and an assumption on the α-*bi*-mixing. Of course the compactness of $\mathcal{X}$ assumed in Theorem 4.4.4 is restrictive, however this assumption is easy to check. Note that α-*bi*-mixing is a slightly weaker assumption on a stochastic process, than the commonly used α-mixing assumption.

Theorem 4.4.10 *Let $(\Omega, \mathcal{A}, \mu)$ be a probability space and $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ a Polish space, $\mathcal{Y} \subset \mathbb{R}$ closed. Let $L\colon \mathcal{X} \times \mathcal{Y} \times \mathbb{R} \to [0, \infty)$ be a Lipschitz continuous loss function such that $\sup_{(x,y)\in\mathcal{X}\times\mathcal{Y}} L(x, y, 0) \leq S$, for some constant $S \in (0, \infty)$. Let H be a Hilbert space consisting of bounded measurable functions $f\colon \mathcal{X} \to \mathbb{R}$. Moreover let $(Z_i)_{i\in\mathbb{N}}$, $Z_i : \Omega \to \mathcal{Z}$, $i \in \mathbb{N}$, be a stochastic process such that there is a constant $C_\alpha > 0$ with:*

$$\frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{i-1} \alpha(Z, \mu, i, j) \leq \frac{C_\alpha}{n},\ n \in \mathbb{N}. \tag{4.57}$$

Then, for $0 < r < \frac{1}{2}$,

$$\frac{1}{n^{1-r}} \sum_{i=1}^{n} \left(L_{f_n} \circ Z_i - \int L(x, y, f_n(x))\, dP^i(x, y) \right) \longrightarrow 0 \quad \text{almost surely, } n \to \infty,\ f_n \in \mathcal{G},$$

where $\mathcal{G} \subset H$ is any uniformly bounded subset of functions $f \in H$, i. e. there is a constant $M > 0$ such that $\|f\|_H \leq M$ for all $f \in \mathcal{G}$.

The assumption on the α-mixing process in Theorem 4.4.10 can be weakened, depending on the constant r. For every $a \in (2r, 2)$, $r \in (0, \frac{1}{2})$, the almost sure convergence of the sequence $\frac{1}{n^{1-r}} \sum_{i=1}^{n} \left(L_{f_n} \circ Z_i - \int L(x, y, f_n(x))\, dP^i(x, y) \right) \longrightarrow 0,\ n \to \infty,\ f_n \in \mathcal{G}$, can be shown in the same way as below if

$$\frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n-i} \alpha((Z, \mu, i, j) \leq \frac{C_\alpha}{n^a},\ n \in \mathbb{N}.$$

Only the exponents have to be adapted. Again, the choice of r can weaken the assumptions on the process but then strengthens the assumption on the sequence $(\lambda_n)_{n \in \mathbb{N}}$. For $a = 1$ we get $r = \frac{1}{2} - \varepsilon, \varepsilon > 0$. Compared to Steinwart et al. (2009, Theorem 3.3), this results in almost the same assumptions on the convergence rate of the sequence $(\lambda_n)_{n \in \mathbb{N}}$, $\lambda_n n^{\frac{1}{2}-\varepsilon} \to \infty$, although we do not require a stability assumption.

Proof of Theorem 4.4.10: The proof follows the same lines as the proof of the consistency for weakly dependent processes, see Theorem 4.4.10. Therefore some calculations are shortened.
Let $\mathcal{G} \subset H$ be a set of uniformly bounded functions $f \in H$. Similar to the proof of Theorem 4.4.6, (4.41), we split the sequence $\frac{1}{n^{1-r}} \sum_{i=1}^{n} \left(L_{f_n} \circ Z_i - \int L(x, y, f_n(x))\, dP^i(x, y) \right)$ in two parts for $n > 1$:

$$\begin{aligned}
&\frac{1}{n^{1-r}} \sum_{i=1}^{n} \left(L_{f_n} \circ Z_i - \int L(x, y, f_n(x))\, dP^i(x, y) \right) \\
&\quad \leq \underbrace{\frac{1}{2^{(s-1)\cdot(1-r)}} \left| \sum_{i=1}^{2^{s-1}} \left(L_{f_n} \circ Z_i - \int L_{f_n}\, dP^i \right) \right|}_{I} \\
&\qquad + \underbrace{\frac{1}{2^{(s-1)\cdot(1-r)}} \left| \max_{1 \leq q \leq 2^{s-1}} \sum_{i=2^{s-1}+1}^{2^{s-1}+q} \left(L_{f_n} \circ Z_i - \int L_{f_n}\, dP^i \right) \right|}_{II}. \qquad (4.58)
\end{aligned}$$

Again the Lemma of Borel-Cantelli is used to show the almost sure convergence of part I, that is we show that for all $\varepsilon > 0$:

$$\sum_{s=1}^{\infty} \mu \left(\left\{ \omega \in \Omega \;\middle|\; \frac{1}{2^{(s-1)\cdot(1-r)}} \left| \sum_{i=1}^{2^{s-1}} \left(L_{f_n} \circ Z_i(\omega) - \int L_{f_n}\, dP^i \right) \right| > \varepsilon \right\} \right) < \infty.$$

Since $\sup_{(x,y)\in\mathcal{X}\times\mathcal{Y}} L(x,y,0)$ is bounded by assumption and $\mathcal{G}$ is uniformly bounded, there is a constant $C_L > 0$ such that

$$\|L_{f_n}\|_\infty := \sup_{(x,y)\in\mathcal{X}\times\mathcal{Y}} |L(x,y,f(x)| \leq C_L, \tag{4.59}$$

for all $f_n \in \mathcal{G}$, see (4.44). For $n = 1$ this yields the boundedness of the first element of the sequence:

$$\frac{1}{n^{1-r}} \sum_{i=1}^{n} \left(L_{f_n} \circ Z_i - \int L(x,y,f_n(x))\, dP^i(x,y) \right) \overset{n=1}{\leq} 2C_L.$$

By Markov's inequality, see for example Hoffmann-Jørgensen (1994, Theorem 3.9), we have:

for $s = 1$

$$\mu\left(\left\{ \omega \in \Omega \;\middle|\; \frac{1}{2^{(s-1)\cdot(1-r)}} \left| \sum_{i=1}^{2^{s-1}} \left(L_{f_n} \circ Z_i(\omega) - \int L_{f_n}\, dP^i \right) \right| > \varepsilon \right\} \right) < \frac{C_L^2}{\varepsilon^2}$$

and

$$\begin{aligned}
&\sum_{s=1}^{\infty} \mu\left(\left\{ \omega \in \Omega \;\middle|\; \frac{1}{2^{(s-1)\cdot(1-r)}} \left| \sum_{i=1}^{2^{s-1}} \left(L_{f_n} \circ Z_i(\omega) - \int L_{f_n}\, dP^i \right) \right| > \varepsilon \right\} \right) \\
&\quad \leq \frac{C_L^2}{\varepsilon^2} + \sum_{s>1}^{\infty} \frac{1}{\varepsilon^2} \mathbb{E}_\mu \left(\frac{1}{2^{(s-1)\cdot(1-r)}} \sum_{i=1}^{2^{s-1}} \left(L_{f_n} \circ Z_i - \int L_{f_n}\, dP^i \right) \right)^2 \\
&\quad = \frac{C_L^2}{\varepsilon^2} + \sum_{s>1}^{\infty} \frac{1}{\varepsilon^2} \left(\frac{1}{2^{(s-1)\cdot(1-r)}} \right)^2 \left[\sum_{i=1}^{2^{s-1}} \mathbb{E}_\mu (L_{f_n} \circ Z_i - \int L_{f_n}\, dP^i)^2 \right. \\
&\qquad \left. + 2 \sum_{i=1}^{2^{s-1}} \sum_{j=1}^{i-1} \mathbb{E}_\mu \left((L_{f_n} \circ Z_i - \int L_{f_n}\, dP^i)(L_{f_n} \circ Z_j - \int L_{f_n}\, dP^j) \right) \right].
\end{aligned} \tag{4.60}$$

Without loss of generality we assume $\|L_{f_n}\|_\infty > 0$, $n \in \mathbb{N}$. $\|L_{f_n}\|_\infty = 0$ implies that $L_{f_n} = 0$, as $L : \mathcal{X} \times \mathcal{Y} \times \mathbb{R} \to [0,\infty)$. Then we easily obtain

$$\mu\left(\left\{ \omega \in \Omega \;\middle|\; \frac{1}{2^{(s-1)\cdot(1-r)}} \left| \sum_{i=1}^{2^{s-1}} \left(L_{f_n} \circ Z_i(\omega) - \int L_{f_n}\, dP^i \right) \right| > \varepsilon \right\} \right) = 0.$$

Moreover we use that the covariance of a stochastic process is related to the $R_\infty^{\mathbb{R}}$-mixing coefficient, see Definition 2.3, which is on the other hand related to α-mixing, see (2.10). Then for all $n \in \mathbb{N}$ such that $\|L_{f_n}\|_\infty \neq 0$:

$$\begin{aligned}
&\frac{C_L^2}{\varepsilon^2} + \sum_{s>1}^{\infty} \frac{1}{\varepsilon^2} \left(\frac{1}{2^{(s-1)\cdot(1-r)}} \right)^2 \left[\sum_{i=1}^{2^{s-1}} \mathbb{E}_\mu (L_{f_n} \circ Z_i - \int L_{f_n}\, dP^i)^2 \right. \\
&\qquad \left. + 2 \sum_{i=1}^{2^{s-1}} \sum_{j=1}^{i-1} \mathbb{E}_\mu \left((L_{f_n} \circ Z_i - \int L_{f_n}\, dP^i)(L_{f_n} \circ Z_j - \int L_{f_n}\, dP^j) \right) \right] \\
&\overset{(4.59)}{\leq} \quad \frac{C_L^2}{\varepsilon^2} + \frac{1}{\varepsilon^2} \sum_{s>1}^{\infty} \left(\frac{1}{2^{(s-1)\cdot(1-r)}} \right)^2 \left[\sum_{i=1}^{2^{s-1}} C_L^2 \right. \\
&\qquad \left. + 2 \sum_{i=1}^{2^{s-1}} \sum_{j=1}^{i-1} \|L_{f_n}\|_\infty^2 \frac{\mathbb{E}_\mu (L_{f_n} \circ Z_i - \int L_{f_n} \circ Z_i\, d\mu)(L_{f_n} \circ Z_j - \int L_{f_n} \circ Z_j\, d\mu)}{\|L_{f_n}\|_\infty \|L_{f_n}\|_\infty} \right] \\
&\overset{(4.59),(2.3)}{\leq} \quad \frac{C_L^2}{\varepsilon^2} + \frac{1}{\varepsilon^2} \sum_{s>1}^{\infty} \left[\left(\frac{1}{2^{(s-1)\cdot(1-r)}} \right)^2 2^{s-1} C_L^2 + 2 \left(\frac{C_L}{2^{(s-1)\cdot(1-r)}} \right)^2 \sum_{i=1}^{2^{s-1}} \sum_{j=1}^{i-1} R_\infty^{\mathbb{R}} \right] \\
&\overset{(2.10)}{\leq} \quad \frac{C_L^2}{\varepsilon^2} + \frac{1}{\varepsilon^2} \sum_{s>1}^{\infty} \left[\frac{1}{2^{(s-1)\cdot(1-2r)}} C_L^2 + 4\pi \left(\frac{C_L}{2^{(s-1)\cdot(1-r)}} \right)^2 \sum_{i=1}^{2^{s-1}} \sum_{j=1}^{i-1} \alpha(Z, \mu, i, j) \right].
\end{aligned}$$

Assumption (4.57) gives $\frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{i-1} \alpha(Z, \mu, i, j) \leq \frac{C_\alpha}{n}$. Hence $\sum_{i=1}^{n} \sum_{j=1}^{i-1} \alpha(Z, \mu, i, j) \leq Cn$. Therefore,

$$\begin{aligned}
&\frac{C_L^2}{\varepsilon^2} + \frac{1}{\varepsilon^2} \sum_{s>1}^{\infty} \left[\frac{1}{2^{(s-1)\cdot(1-2r)}} C_L^2 + 4\pi \left(\frac{C_L}{2^{(s-1)\cdot(1-r)}} \right)^2 \sum_{i=1}^{2^{s-1}} \sum_{j=1}^{i-1} \alpha(Z, \mu, i, j) \right] \\
&\overset{(4.57)}{\leq} \quad \frac{C_L^2}{\varepsilon^2} + \frac{1}{\varepsilon^2} \sum_{s>1}^{\infty} \left[\frac{1}{2^{(s-1)\cdot(1-2r)}} C_L^2 + 4\pi \left(\frac{C_L}{2^{(s-1)\cdot(1-r)}} \right)^2 C_\alpha \cdot 2^{s-1} \right] \\
&\leq \quad \frac{C_L^2}{\varepsilon^2} + \frac{1}{\varepsilon^2} \sum_{s>1}^{\infty} \frac{1}{2^{(s-1)\cdot(1-2r)}} \tilde{C},
\end{aligned}$$

where $\tilde{C} := (1 + 4\pi C_\alpha) C_L^2 > 0$. As $\frac{1}{2^{1-2r}} < 1$ for all $0 < r < \frac{1}{2}$, the series equals a geometric series and therefore is convergent:

$$\frac{1}{\varepsilon^2} \sum_{s=1}^{\infty} \frac{1}{2^{(s-1)\cdot(1-2r)}} \tilde{C} = \frac{1}{\varepsilon^2} \tilde{C} \sum_{s=0}^{\infty} \left(\frac{1}{2^{(1-2r)}} \right)^s < \infty.$$

This implies the almost sure convergence of the term in part I.

For the second term we show that again the generalization of the Rademacher-Menosv-Inequality by Serfling (1970, Theorem A) leads the almost sure convergence.

First Markov's inequality, see Hoffmann-Jørgensen (1994, Theorem 3.9), gives:

$$\sum_{s=1}^{\infty} \mu\left(\left\{\omega \in \Omega \left| \frac{1}{2^{(s-1)\cdot(1-r)}} \right| \max_{1\le q\le 2^{s-1}} \sum_{i=2^{s-1}+1}^{2^{s-1}+q} \left(L_{f_n}\circ Z_i(\omega) - \int L_{f_n}\, dP^i\right)\right| > \varepsilon\right\}\right)$$
$$\le \frac{1}{\varepsilon^2} \sum_{s=1}^{\infty} \left(\frac{1}{2^{(s-1)\cdot(1-r)}}\right)^2 \mathbb{E}_\mu \left(\max_{1\le q\le 2^{s-1}} \sum_{i=2^{s-1}+1}^{2^{s-1}+q} \left(L_{f_n}\circ Z_i - \int L_{f_n}\, dP^i\right)\right)^2.$$

For $s = 1$ we obtain

$$\mathbb{E}_\mu \left(\max_{1\le q\le 2^{s-1}} \sum_{i=2^{s-1}+1}^{2^{s-1}+q} \left(L_{f_n}\circ Z_i - \int L_{f_n}\, dP^i\right)\right)^2 = \mathbb{E}_\mu \left(\sum_{i=2}^{2} L_{f_n}\circ Z_i - \int L_{f_n}\, dP^i\right)^2$$
$$\le \|L_{f_n}\|_\infty^2 \overset{(4.59)}{\le} C_L^2. \tag{4.61}$$

Again we assume $\|L_{f_n}\|_\infty \neq 0$, $n \in \mathbb{N}$. If there exists $n \in \mathbb{N}$ such that $\|L_{f_n}\|_\infty = 0$ the calculations are again trivial. The maximal inequality by Serfling (1970) for the function $h_{2^{s-1},2^{s-1}}$, see Lemma 4.4.7, and the definition of the mixing coefficient $R_\infty^{\mathbb{R}}$, Definition (2.3), and Inequality (2.10), yield:

$$\sum_{s=1}^{\infty} \mu\left(\left\{\omega \in \Omega \left| \frac{1}{2^{(s-1)\cdot(1-r)}} \right| \max_{1\le q\le 2^{s-1}} \sum_{i=2^{s-1}+1}^{2^{s-1}+q} \left(L_{f_n}\circ Z_i(\omega) - \int L_{f_n}\, dP^i\right)\right| > \varepsilon\right\}\right)$$
$$\overset{\text{Lem (4.4.7),(4.61)}}{\le} \frac{C_L^2}{\varepsilon^2} + \sum_{s>1}^{\infty} \left(\frac{1}{\varepsilon 2^{(s-1)\cdot(1-r)}}\right)^2 (\log_2(2\cdot 2^{s-1}))^2 \left[\sum_{2^{s-1}+1}^{2^s} \mathrm{Var}(L_{f_n}\circ Z_i)\right.$$
$$\left. +2 \sum_{i=2^{s-1}+1}^{2^s-1} \sum_{j=i+1}^{2^s} |\mathrm{Cov}(L_{f_n}\circ Z_i, L_{f_n}\circ Z_j)|\right]$$
$$\overset{(4.59)}{\le} \frac{C_L^2}{\varepsilon^2} + \sum_{s>1}^{\infty} C \left(\frac{1}{2^{(s-1)\cdot(1-r)}}\right)^2 (\log_2 2^{s-1})^2 \left[2^{s-1} C_L^2 + \right.$$
$$\left. + 2 \sum_{i=2^{s-1}+1}^{2^s} \sum_{j=2^{s-1}+1}^{i-1} \|L_{f_n}\|_\infty^2 \frac{\mathbb{E}_\mu(L_{f_n}\circ Z_i - \int L_{f_n}\circ Z_i\, d\mu)(L_{f_n}\circ Z_j - \int L_{f_n}\circ Z_j\, d\mu)}{\|L_{f_n}\|_\infty \|L_{f_n}\|_\infty}\right]$$

$$
\begin{aligned}
&\overset{(2.3),(2.10),(4.54)}{\leq} && \frac{C_L^2}{\varepsilon^2} + C \sum_{s>1}^{\infty} \left(\frac{\log_2 2^{s-1}}{2^{(s-1)\cdot(1-r)}} \right)^2 \left[2^{s-1} C_L^2 + 4\pi C_L^2 \sum_{i=2^{s-1}+1}^{2^s} \sum_{j=2^{s-1}+1}^{i-1} \alpha(Z, \mu, i, j) \right] \\
&\overset{(4.57)}{\leq} && \frac{C_L^2}{\varepsilon^2} + C \sum_{s>1}^{\infty} \frac{(\log_2 2^{s-1})^2}{(2^{(s-1)\cdot(1-r)})^2} \left[2^{s-1} C_L^2 + 4\pi C_L^2 C_\alpha (2^s - 2^{s-1}) \right] \\
&\leq && \frac{C_L^2}{\varepsilon^2} + C' \sum_{s>1}^{\infty} \frac{(s-1)^2}{2^{(s-1)\cdot(1-2r)}} \; < \infty,
\end{aligned}
$$

for a constant $C' := (1 + 4\pi C_\alpha) C C_L^2 > 0$. The convergence again follows via the ratio test, similar to (4.56).

Hence,

$$
\begin{aligned}
&\frac{1}{n^{1-r}} \sum_{i=1}^{n} \left(L_{f_n} \circ Z_i - \int L(x, y, f_n(x)) \, dP^i(x, y) \right) \\
&\qquad \overset{(4.58)}{\leq} \underbrace{\frac{1}{2^{(s-1)\cdot(1-r)}} \left| \sum_{i=1}^{2^{s-1}} \left(L_{f_n} \circ Z_i - \int L_{f_n} \, dP^i \right) \right|}_{I \longrightarrow 0 \text{ almost surely}} \\
&\qquad\quad + \underbrace{\frac{1}{2^{(s-1)\cdot(1-r)}} \left| \max_{1 \leq q \leq 2^{s-1}} \sum_{i=2^{s-1}+1}^{2^{s-1}+q} \left(L_{f_n} \circ Z_i - \int L_{f_n} \, dP^i \right) \right|}_{II \longrightarrow 0 \text{ almost surely}} \\
&\qquad \longrightarrow \; 0 \quad \text{almost surely.}
\end{aligned}
$$

□

The L-risk-consistency is ensured for the following assumptions:

Corollary 4.4.11 *Let $(\Omega, \mathcal{A}, \mu)$ be a probability space, let $(\mathcal{Z}, d_{\mathcal{Z}}) = (\mathcal{X} \times \mathcal{Y}, d_{\mathcal{X}\times\mathcal{Y}})$ be a separable, metric space and let $\mathcal{X}$ be compact and $\mathcal{Y} \subseteq \mathbb{R}$ closed. Let $L \colon \mathcal{X} \times \mathcal{Y} \times \mathbb{R} \to [0, \infty)$ be a loss function which is convex and Lipschitz continuous in the last argument, continuous in all $(x, y) \in \mathcal{X} \times \mathcal{Y}$ and $\sup_{(x,y)\in\mathcal{X}\times\mathcal{Y}} L(x, y, 0) \leq S$, for a constant $S \in (0, \infty)$. Let H be a reproducing kernel Hilbert space of an universal, bounded and continuous kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. Let $(Z_i)_{i\in\mathbb{N}}$, $Z_i : \Omega \to \mathcal{Z}$, be an asymptotically mean stationary, α-mixing stochastic process such that there is a constant $C > 0$ with:*

$$
\frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{i-1} \alpha(Z, \mu, i, j) \leq \frac{C}{n}, \; n \in \mathbb{N}.
$$

Let $(\lambda_n)_{n\in\mathbb{N}} \subset (0,\infty)$ *such that* $\lambda_n \to 0$ *and* $\lambda_n n^r \to \infty$ *for some* $0 < r < \frac{1}{2}$. *Let the sequences* $(f_{\frac{1}{n}\sum P^i,\lambda_n})_{n\in\mathbb{N}}$ *and* $(f_{\mathbb{P}_{\mathbf{W}_n(\omega)},\lambda_n})_{n\in\mathbb{N}}$ *be bounded for all* $\omega \in \Omega$, *i. e. there are constants* $M, \tilde{M} > 0$ *such that* $\|f_{\frac{1}{n}\sum P^i,\lambda_n}\|_H \leq M$ *and* $\|f_{\mathbb{P}_{\mathbf{W}_n(\omega)},\lambda_n}\|_H \leq \tilde{M}$, $n \in \mathbb{N}$.

Then:

$$R_{L,P}(f_{\mathbb{P}_{\mathbf{W}_n},\lambda_n}) \to R^*_{L,P} \quad \textit{in probability, } n \to \infty.$$

The L-risk-consistency in H follows directly from Theorem 4.4.4 and 4.4.10. As k is a universal kernel by assumption and $\mathcal{X}$ is compact, the convergence to the Bayes risk $R^*_{L,P}$ follows by Steinwart and Christmann (2008, Corollary 5.28).

4.4.3 $\mathcal{C}$-mixing processes

Another example for processes which guarantee almost sure convergence in (4.17) are certain $\mathcal{C}$-mixing processes. The next theorem shows that $\mathcal{C}$-mixing processes on the space of Lipschitz continuous, bounded functions comply with (4.17). That is the class $\mathcal{C}$ of functions equals the set of bounded Lipschitz functions $\mathrm{BL}(\mathcal{Z}) := \{f : \mathcal{Z} \to \mathbb{R} \mid ||f||_{\mathrm{BL}} < \infty\}$ equipped with semi-norm $\|f\|_{\mathcal{C}} := \|f\|_{\mathrm{BL}} = \|f\|_\infty + |f|_1$. Hang and Steinwart (2015, Theorem 4.7) show a Bernstein-type inequality for strongly stationary (time reversed) geometrically $\mathcal{C}$-mixing processes, that is $\Phi_{\mathcal{C}} \leq c\exp(-bn^\gamma)$, $\gamma, b, c > 0$. Moreover learning rates for support vector machines for the least squares loss and for the pinball loss are achieved. This implies the L-risk-consistency of the SVM estimator under this $\mathcal{C}$-mixing condition. Hence, concerning L-risk-consistency, we regard other loss functions. Contrary to Hang and Steinwart (2015), Theorem 4.4.4 does not cover the least squares loss, as it is not Lipschitz continuous, but the pinball loss and other Lipschitz continuous losses. The theorem below shows that (4.17) covers more processes than the Bernstein-type inequality in Hang and Steinwart (2015). We do not need an exponential decay of the mixing coefficients, but require that $\Phi_{\mathcal{C}}$ is summable. Furthermore we require the AMS property, see (4.15), instead of the stationarity of the stochastic process. That is we require the existence of a probability measure $P \in \mathcal{M}(\mathcal{Z})$ such that $P(B) = \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^n \mathbb{E}_\mu I_B \circ Z_i$, for all $B \in \mathcal{B}$. The AMS property is not necessary if the process is strongly stationary. Due to the weaker assumptions on the process it covers more processes than the Bernstein-type inequality in Hang and Steinwart (2015), but we do not achieve learning rates or a concentration inequality.

Theorem 4.4.12 *Let* $(\Omega, \mathcal{A}, \mu)$ *be a probability space and let* $(\mathcal{Z}, d_{\mathcal{Z}}) = (\mathcal{X}\times\mathcal{Y}, d_{\mathcal{X}\times\mathcal{Y}})$ *be a measurable space,* $(\mathcal{X}, d_{\mathcal{X}})$ *compact and* $\mathcal{Y} \subset \mathbb{R}$ *closed. Let* $L\colon \mathcal{X}\times\mathcal{Y}\times\mathbb{R} \to [0,\infty)$ *be a*

distance-based Lipschitz continuous loss function with $\sup_{(x,y)\in\mathcal{X}\times\mathcal{Y}} L(x,y,0) \leq S$, *for some constant* $S \in (0,\infty)$, *and* $|L|_1 > 0$. *Let* $k : \mathcal{X}\times\mathcal{X} \to \mathbb{R}$ *be a continuous and bounded kernel with RKHS* H. *Moreover let* $\mathcal{C} = \mathrm{BL}(\mathcal{Z})$ *be the space of Lipschitz continuous, bounded functions* $\mathcal{Z} \to \mathbb{R}$. *Let* $(Z_i)_{i\in\mathbb{N}}$, $Z_i : \Omega \to \mathcal{Z}$, *be a* $\mathcal{C}$*-mixing stochastic process.*
Then, for $0 < r < \frac{1}{2}$,

$$\frac{1}{n^{1-r}} \sum_{i=1}^{n} L_{f_n} \circ Z_i - \int L_{f_n} \circ Z_i \, d\mu \longrightarrow 0 \text{ almost surely, } n \to \infty, \ f_n \in \mathcal{G}, \tag{4.62}$$

where $\mathcal{G} \subset H$ *is any uniformly bounded subset of functions* $f \in H$, *i. e. there is a constant* $M > 0$ *such that* $\|f\|_H \leq M$ *for all* $f \in \mathcal{G}$.

Proof of Theorem 4.4.12: The proof follows the same lines as the proof of Theorem 4.4.6 and 4.4.10, Therefore some calculations are again shortened. We have

$$\frac{1}{n^{1-r}} \sum_{i=1}^{n} \left(L_{f_n} \circ Z_i - \int L_{f_n} \circ Z_i \right) = \frac{1}{n^{1-r}} \sum_{i=1}^{n} \left(L_{f_n} \circ Z_i - \int L_{f_n} \, dP^i \right). \tag{4.63}$$

Analogously to the proof of Theorem 4.4.6, (4.39) and (4.38), for any $\rho(n) = \frac{1}{n^{1+r}}$, $n \in \mathbb{N}$, there is a function $g_{\rho(n),n} \in \mathrm{BL}(\mathcal{X}) \cap \mathcal{G}$ such that

$$\frac{1}{n^{1-r}} \sum_{i=1}^{n} \left(L_{f_n} \circ Z_i - \int L_{f_n} \, dP^i \right) \ \leq \ \frac{2|L|_1}{n} + \frac{1}{n^{1-r}} \sum_{i=1}^{n} \left| L_{g_{\rho(n),n}} \circ Z_i - \int L_{g_{\rho(n),n}} \, dP^i \right|.$$

The last term on the right hand side can be split up for $n > 1$, see (4.41):

$$\begin{aligned} &\frac{1}{n^{1-r}} \left| \sum_{i=1}^{n} \left(L_{g_{\rho(n),n}} \circ Z_i - \int L_{g_{\rho(n),n}} \, dP^i \right) \right| \\ &\quad \leq \underbrace{\frac{1}{2^{(s-1)\cdot(1-r)}} \left| \sum_{i=1}^{2^{s-1}} \left(L_{g_{\rho(n),n}} \circ Z_i - \int L_{g_{\rho(n),n}} \, dP^i \right) \right|}_{I} \\ &\qquad + \underbrace{\frac{1}{2^{(s-1)\cdot(1-r)}} \left| \max_{1\leq q\leq 2^{s-1}} \sum_{i=2^{s-1}+1}^{2^{s-1}+q} \left(L_{g_{\rho(n),n}} \circ Z_i - \int L_{g_{\rho(n),n}} \, dP^i \right) \right|}_{II}. \end{aligned} \tag{4.64}$$

Now the almost sure convergence of the terms on the right hand side follows, as the sum of covariances

$$\mathbb{E}_\mu \left(\sum_{i=a+1}^{b-1} \sum_{j=i+1}^{b} \left(L_{g_{\rho(n),n}} \circ Z_i - \mathbb{E}_\mu L_{g_{\rho(n),n}} \circ Z_i \right) \left(L_{g_{\rho(n),n}} \circ Z_j - \mathbb{E}_\mu L_{g_{\rho(n),n}} \circ Z_j \right) \right),$$

$a, b \in \mathbb{N}$, $a < b-1$, is bounded under the assumptions on the $\mathcal{C}$-mixing coefficients. For $n = 1$ the boundedness of $L_{g_{\rho(n),n}}$, see the proof of Theorem 4.4.6, (4.43), yields

$$\frac{1}{n^{1-r}} \left| \sum_{i=1}^{n} \left(L_{g_{\rho(n),n}} \circ Z_i - \int L_{g_{\rho(n),n}} \, dP^i \right) \right|^{n=1} \leq 2C_L,$$

i. e. the boundedness of the first element of the sequence.

Calculation (4.42) in the proof of Theorem 4.4.6 shows that the function $L_{g_{\rho(n),n}}$ is Lipschitz continuous with respect to $d_{\mathcal{X}\times\mathcal{Y}}((x,y),(x',y')) = d_{\mathcal{X}}(x,x') + |y-y'|$ as L is distance-based and Lipschitz continuous. Again the same argumentation as in the proof of Theorem 4.4.6 and Lemma 4.4.8 ensures the existence of a constant $M' > 0$ such that $|g_{\rho(n),n}|_1 \leq M'$ and therefore $|L_{g_{\rho(n),n}}|_1 \leq \tilde{\tilde{M}}$ for a non-negative constant $\tilde{\tilde{M}} := \max\{|L|_1, |L|_1 \cdot |g_{\rho(n),n}|_1\} = \max\{|L|_1, |L|_1 \cdot M'\}$, see (4.42). In particular this constant does not depend on n respectively on $g_{\rho(n),n} \in \mathrm{BL}(\mathcal{X}) \cap \mathcal{G}$. Further

$$\|L_{g_{\rho(n),n}}\|_{\mathcal{C}} := \|L_{g_{\rho(n),n}}\|_{\mathrm{BL}} = \|L_{g_{\rho(n),n}}\|_\infty + |L_{g_{\rho(n),n}}|_1 \leq \tilde{M}, \tag{4.65}$$

where

$$\tilde{M} := S + |L|_1 M \|k\|_\infty + \tilde{\tilde{M}} > 0 \tag{4.66}$$

is a constant, which again does not dependent on n. Also $\|L_{g_{\rho(n),n}}\|_{\mathrm{BL}} > 0$, as $|L|_1 > 0$ by assumption. Hence $L_{g_{\rho(n),n}} \in \mathrm{BL}(\mathcal{X} \times \mathcal{Y})$ and

$$\frac{L_{g_{\rho(n),n}}}{\|L_{g_{\rho(n),n}}\|_{\mathrm{BL}}} \in \mathrm{BL}_1(\mathcal{X} \times \mathcal{Y}). \tag{4.67}$$

Moreover $\left\| \frac{L_{g_{\rho(n),n}}}{\|L_{g_{\rho(n),n}}\|_{\mathrm{BL}}} \right\|_1 \leq 1$, as $\left\| \frac{L_{g_{\rho(n),n}}}{\|L_{g_{\rho(n),n}}\|_{\mathrm{BL}}} \right\|_\infty \leq 1$.

Now $L_{g_{\rho(n),n}} : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ is continuous by assumption and therefore measurable with respect to the Borel σ-algebras on $\mathcal{X} \times \mathcal{Y}$ and $\mathbb{R}$, hence $L_{g_{\rho(n),n}} \circ Z_i$ is measurable with respect to $(\mathcal{A}_i^i, \mathcal{B})$, $i \in \mathbb{N}$, where $\mathcal{A}_i^i$ is the σ-algebra generated by Z_i, $i \in \mathbb{N}$, on Ω. In particular each function $L_{g_{\rho(n),n}} \circ Z_i$ is measurable with respect to $(\mathcal{A}_1^i, \mathcal{B})$, $i \in \mathbb{N}$, where

$\mathcal{A}_1^i = \sigma(Z_1, \dots, Z_i)$. Hence, for all $a, b \in \mathbb{N}$, $a < b - 1$,

$$\begin{aligned}
&\mathbb{E}_\mu \left[\sum_{i=a+1}^{b-1} \sum_{j=i+1}^{b} \left(L_{g_{\rho(n),n}} \circ Z_i - \mathbb{E}_\mu L_{g_{\rho(n),n}} \circ Z_i \right) \left(L_{g_{\rho(n),n}} \circ Z_j - \mathbb{E}_\mu L_{g_{\rho(n),n}} \circ Z_j \right) \right] \\
&\quad = \sum_{\ell=1}^{b-a-1} \sum_{i=a+1}^{b-\ell} \mathbb{E}_\mu (L_{g_{\rho(n),n}} \circ Z_i) \cdot (L_{g_{\rho(n),n}} \circ Z_{i+\ell}) - \mathbb{E}_\mu (L_{g_{\rho(n),n}} \circ Z_i) \mathbb{E}_\mu (L_{g_{\rho(n),n}} \circ Z_{i+\ell}) \\
&\quad \leq \sum_{\ell=1}^{b-a-1} (b - \ell - a) \sup_{i=a+1,\dots,b-\ell} \Big| \mathbb{E}_\mu (L_{g_{\rho(n),n}} \circ Z_i) \cdot (L_{g_{\rho(n),n}} \circ Z_{i+\ell}) \\
&\qquad\qquad - \mathbb{E}_\mu (L_{g_{\rho(n),n}} \circ Z_i) \mathbb{E}_\mu (L_{g_{\rho(n),n}} \circ Z_{i+\ell}) \Big| \\
&\quad \overset{(2.12),(4.67)}{\leq} \sum_{\ell=1}^{b-a-1} (b - \ell - a) (\|L_{g_{\rho(n),n}}\|_{\mathrm{BL}})^2 \Phi_{\mathcal{C}}(Z, \ell) \\
&\quad \overset{(4.65),(4.66)}{\leq} \sum_{\ell=1}^{b-a-1} (b - \ell - a) \tilde{M}^2 \Phi_{\mathcal{C}}(Z, \ell). \qquad (4.68)
\end{aligned}$$

Now, we can use the same argumentation as in the proof of Theorem 4.4.6 to show the almost sure convergence of the terms in part I and II. For part I we have, see (4.49), for every $\varepsilon > 0$ and $s > 1$,

$$\begin{aligned}
&\mu \left(\left\{ \omega \in \Omega \;\middle|\; \frac{1}{2^{(s-1)\cdot(1-r)}} \left| \sum_{i=1}^{2^{s-1}} \left(L_{g_{\rho(n),n}} \circ Z_i(\omega) - \int L_{g_{\rho(n),n}} \, dP^i \right) \right| > \varepsilon \right\} \right) \\
&\overset{Markov}{\leq} \frac{1}{\varepsilon^2} \left(\frac{1}{2^{(s-1)\cdot(1-r)}} \right)^2 \left[\sum_{i=1}^{2^{s-1}} \mathbb{E}_\mu (L_{g_{\rho(n),n}} \circ Z_i - \int L_{g_{\rho(n),n}} \, dP^i)^2 \right. \\
&\qquad \left. + 2 \sum_{i=1}^{2^{s-1}-1} \sum_{j=i+1}^{2^{s-1}} \mathbb{E}_\mu \left(L_{g_{\rho(n),n}} \circ Z_i - \int L_{g_{\rho(n),n}} \, dP^i \right) \left(L_{g_{\rho(n),n}} \circ Z_j - \int L_{g_{\rho(n),n}} \, dP^j \right) \right] \\
&\overset{(4.68)}{\leq} \frac{1}{\varepsilon^2} \left(\frac{1}{2^{(s-1)\cdot(1-r)}} \right)^2 \left[\sum_{i=1}^{2^{s-1}} \mathrm{Var}(L_{g_{\rho(n),n}} \circ Z_i) + 2 \sum_{\ell=1}^{2^{s-1}-1} (2^{s-1} - \ell) \tilde{M}^2 \Phi_{\mathcal{C}}(Z, \ell) \right]. \qquad (4.69)
\end{aligned}$$

For $s = 1$ we obtain

$$\mu \left(\left\{ \omega \in \Omega \;\middle|\; \left| \left(L_{g_{\rho(n),n}} \circ Z_1(\omega) - \int L_{g_{\rho(n),n}} \, dP^i \right) \right| > \varepsilon \right\} \right) \leq \frac{1}{\varepsilon^2} \mathrm{Var}(L_{g_{\rho(n),n}} \circ Z_1). \quad (4.70)$$

As $(Z_i)_{i\in\mathbb{N}}$ is $\mathcal{C}$-mixing, i. e.

$$\sum_{\ell=1}^{\infty}\Phi_{\mathcal{C}}(Z,\ell)\leq C_\Phi<\infty, \tag{4.71}$$

and due to the uniform bound on the variance, see (4.44), we have:

$$\begin{aligned}
&\sum_{s=1}^{\infty}\mu\left(\left\{\omega\in\Omega\;\middle|\;\frac{1}{2^{(s-1)\cdot(1-r)}}\left|\sum_{i=1}^{2^{s-1}}\left(L_{g_{\rho(n),n}}\circ Z_i(\omega)-\int L_{g_{\rho(n),n}}\,dP^i\right)\right|>\varepsilon\right\}\right)\\
&\overset{(4.69),(4.70)}{\leq}\quad\frac{1}{\varepsilon^2}\mathrm{Var}(L_{g_{\rho(n),n}}\circ Z_i)\\
&\qquad\qquad+\frac{1}{\varepsilon^2}\sum_{s>1}^{\infty}\left(\frac{1}{2^{(s-1)\cdot(1-r)}}\right)^2\left[\sum_{i=1}^{2^{s-1}}\mathrm{Var}(L_{g_{\rho(n),n}}\circ Z_i)+2\sum_{\ell=1}^{2^{s-1}-1}(2^{s-1}-\ell)\tilde{M}^2\Phi_{\mathcal{C}}(Z,\ell)\right]\\
&\overset{(4.44)(4.71)}{\leq}\quad\frac{C_L^2}{\varepsilon^2}+\frac{1}{\varepsilon^2}\sum_{s>1}^{\infty}\left[\left(\frac{1}{2^{(s-1)\cdot(1-r)}}\right)^2\left[2^{s-1}C_L^2+2\cdot 2^{s-1}\tilde{M}^2C_\Phi\right]\right]\\
&\leq\quad\frac{C_L^2}{\varepsilon^2}+\frac{1}{\varepsilon^2}\sum_{s=1}^{\infty}\left[\frac{1}{2^{(s-1)\cdot(1-2r)}}\left[C_L^2+2\tilde{M}^2C_\Phi\right]\right]<\infty,
\end{aligned}$$

where the last sum again is a geometric series and therefore finite for $0<r<\frac{1}{2}$. Hence the almost sure convergence of the term in part I follows.

For part II, again Markov's inequality, see Hoffmann-Jørgensen (1994, Theorem 3.9), and the maximal inequality by Serfling (1970) for the function $h_{2^{s-1},2^{s-1}}$, see Lemma 4.4.7, are used. For $s=1$ we have, similar to (4.55),

$$\mathbb{E}_\mu\left(\max_{1\leq q\leq 2^{s-1}}\sum_{i=2^{s-1}+1}^{2^{s-1}+q}\left(L_{g_{\rho(2),2}}\circ Z_i-\int L_{g_{\rho(2),2}}\,dP^i\right)\right)^2\leq\|L_{g_{\rho(2),2}}\|_\infty^2\overset{(4.44)}{=}C_L^2<\infty. \tag{4.72}$$

Then,

$$\begin{aligned}
&\sum_{s=1}^{\infty}\mu\left(\left\{\omega\in\Omega\;\middle|\;\frac{1}{2^{(s-1)\cdot(1-r)}}\left|\max_{1\leq q\leq 2^{s-1}}\sum_{i=2^{s-1}+1}^{2^{s-1}+q}\left(L_{g_{\rho(n),n}}\circ Z_i(\omega)-\int L_{g_{\rho(n),n}}\,dP^i\right)\right|>\varepsilon\right\}\right)\\
&\overset{Markov}{\leq}\quad\frac{1}{\varepsilon^2}\sum_{s=1}^{\infty}\left(\frac{1}{2^{(s-1)\cdot(1-r)}}\right)^2\mathbb{E}_\mu\left(\max_{1\leq q\leq 2^{s-1}}\sum_{i=2^{s-1}+1}^{2^{s-1}+q}\left(L_{g_{\rho(n),n}}\circ Z_i-\int L_{g_{\rho(n),n}}\,dP^i\right)\right)^2
\end{aligned}$$

$$
\begin{aligned}
&\overset{Serfling,(4.72)}{\leq} && \frac{C_L^2}{\varepsilon^2} + \frac{1}{\varepsilon^2}\sum_{s>1}^{\infty}\left(\frac{1}{2^{(s-1)\cdot(1-r)}}\right)^2 (\log_2(2\cdot 2^{s-1}))^2 \left[\sum_{2^{s-1}+1}^{2^s} \mathrm{Var}(L_{g_{\rho(n),n}} \circ Z_i)\right. \\
&&& \qquad \left. + 2\sum_{i=2^{s-1}+1}^{2^s-1}\sum_{j=i+1}^{2^s} |\mathrm{Cov}(L_{g_{\rho(n),n}} \circ Z_i, L_{g_{\rho(n),n}} \circ Z_j)|\right] \\
&\overset{(4.54)}{\leq} && \frac{C_L^2}{\varepsilon^2} + \frac{C}{\varepsilon^2}\sum_{s>1}^{\infty}\left(\frac{1}{2^{(s-1)\cdot(1-r)}}\right)^2 (\log_2 2^{s-1})^2 \left[\sum_{2^{s-1}+1}^{2^s} \mathrm{Var}(L_{g_{\rho(n),n}} \circ Z_i)\right. \\
&&& \qquad \left. + 2\sum_{i=2^{s-1}+1}^{2^s-1}\sum_{j=i+1}^{2^s} |\mathrm{Cov}(L_{g_{\rho(n),n}} \circ Z_i, L_{g_{\rho(n),n}} \circ Z_j)|\right] \\
&\overset{(4.44),(4.68)}{\leq} && \frac{C_L^2}{\varepsilon^2} + \frac{C}{\varepsilon^2}\sum_{s>1}^{\infty}\left(\frac{1}{2^{(s-1)\cdot(1-r)}}\right)^2 (\log_2 2^{s-1})^2 \left[2^{s-1}C_L^2 + 2^{s-1}\cdot 2\tilde{M}\sum_{\ell=1}^{\infty}\Phi_{\mathcal{C}}(Z,\ell)\right] \\
&\leq && \frac{C_L^2}{\varepsilon^2} + \frac{C}{\varepsilon^2}\sum_{s>1}^{\infty}\left(\frac{\log_2 2^{s-1}}{2^{(s-1)\cdot(1-r)}}\right)^2 2^{s-1}\left[2\tilde{M}C_\Phi + C_L^2\right] \\
&\leq && \frac{C_L^2}{\varepsilon^2} + \frac{C'}{\varepsilon^2}\sum_{s>1}^{\infty}\frac{(s-1)^2}{2^{(s-1)\cdot(1-2r)}} \;<\; \infty,
\end{aligned}
$$

for a constant $C' := 2\tilde{M}C_\Phi + C_L^2 > 0$. The convergence again follows via the ratio test, similar to (4.56), as $0 < r < \frac{1}{2}$.

Combining these results, the assertion follows:

$$
\begin{aligned}
&\frac{1}{n^{1-r}}\sum_{i=1}^{n}\left(L_{f_n}\circ Z_i - \int L(x,y,f_n(x))\,dP^i(x,y)\right) \\
&\qquad \overset{(4.63)}{\leq} \frac{2|L|_1}{n} + \left|\frac{1}{n^{1-r}}\sum_{i=1}^{n}\left(L_{g_{\rho(n),n}}\circ Z_i - \int L_{g_{\rho(n),n}}\,dP^i\right)\right| \\
&\qquad \overset{(4.64)}{\leq} \frac{2|L|_1}{n} + \underbrace{\frac{1}{2^{(s-1)\cdot(1-r)}}\left|\sum_{i=1}^{2^{s-1}}\left(L_{g_{\rho(n),n}}\circ Z_i - \int L_{g_{\rho(n),n}}\,dP^i\right)\right|}_{I \longrightarrow 0 \text{ almost surely}} \\
&\qquad\quad + \underbrace{\frac{1}{2^{(s-1)\cdot(1-r)}}\left|\max_{1\leq q\leq 2^{s-1}}\sum_{i=2^{s-1}+1}^{2^{s-1}+q}\left(L_{g_{\rho(n),n}}\circ Z_i - \int L_{g_{\rho(n),n}}\,dP^i\right)\right|}_{II \longrightarrow 0 \text{ almost surely}} \\
&\qquad \longrightarrow\; 0 \quad \text{almost surely, } n\to\infty. \qquad \square
\end{aligned}
$$

Therefore, the SVM is consistent also for $\mathcal{C}$-mixing stochastic processes.

Corollary 4.4.13 *Let $(\Omega, \mathcal{A}, \mu)$ be a probability space and let $(\mathcal{Z}, d_{\mathcal{Z}}) = (\mathcal{X} \times \mathcal{Y}, d_{\mathcal{X}\times\mathcal{Y}})$ be a separable, metric space, let $(\mathcal{X}, d_{\mathcal{X}})$ be compact and $\mathcal{Y} \subset \mathbb{R}$ closed. Let $L\colon \mathcal{X} \times \mathcal{Y} \times \mathbb{R} \to [0, \infty)$ be a distance-based loss function which is convex and Lipschitz continuous in the last argument, continuous in (x, y) for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$ and $\sup_{(x,y)\in\mathcal{X}\times\mathcal{Y}} L(x, y, 0) \leq S$, for some constant $S \in (0, \infty)$, and $|L|_1 > 0$. Moreover let H be the reproducing kernel Hilbert space of an universal, bounded and continuous kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. Let $\mathcal{C}$ be the space of Lipschitz continuous functions $\mathcal{Z} \to \mathbb{R}$ and let $(Z_i)_{i\in\mathbb{N}}$, $Z_i : \Omega \to \mathcal{Z}$, be an asymptotically mean stationary and $\mathcal{C}$-mixing stochastic process.*
Let $(\lambda_n)_{n\in\mathbb{N}} \subset (0, \infty)$ such that $\lambda_n \to 0$ and $\lambda_n n^r \to \infty$, for some $0 < r < \frac{1}{2}$, and let the sequences $(f_{\frac{1}{n}\sum P^i, \lambda_n})_{n\in\mathbb{N}}$ and $(f_{\mathbb{P}_{\mathbf{W}_n(\omega)}, \lambda_n})_{n\in\mathbb{N}}$ be bounded for all $\omega \in \Omega$, i. e. there are constants $M, \tilde{M} > 0$ such that $\|f_{\frac{1}{n}\sum P^i, \lambda_n}\|_H \leq M$ and $\|f_{\mathbb{P}_{\mathbf{W}_n(\omega)}, \lambda_n}\|_H \leq \tilde{M}$, $n \in \mathbb{N}$.

Then:

$$R_{L,P}(f_{\mathbb{P}_{\mathbf{W}_n}\lambda_n}) \to R^*_{L,P} \quad \text{in probability, } n \to \infty.$$

Again the proof of this corollary follows directly from Theorem 4.4.12 and 4.4.4 and the assumption, that k is a universal kernel.

Chapter 5

Conclusion and outlook

Throughout this thesis we generalize properties of support vector machines (SVMs), in particular robustness and consistency, to data generating stochastic processes which are not necessarily independent and identically distributed. In case of qualitative robustness our results are more general and can be applied to a larger class of estimators than just SVMs.

To operate the dependence of the data generating stochastic process, we introduce strong respectively weak Varadarajan processes in Chapter 3. These are stochastic processes which provide almost sure convergence, respectively convergence in probability, of their empirical measures $\mathbb{P}_{\mathbf{W}_n}$, $n \in \mathbb{N}$, to a limiting distribution P with respect to the Prohorov metric or with respect to the bounded Lipschitz metric. Examples are stochastic processes which fulfil a law of large numbers for events, for example many Markov chains, many α-mixing processes or some strongly stationary ergodic processes, as well as several weakly dependent processes or some $\mathcal{C}$-mixing processes. Both properties, statistical robustness as well as consistency, rely on the empirical distribution of the data generating stochastic process, which justifies the above definition.

For the i.i.d. case a lot of theory on robustness properties, consistency, and learning rates of SVMs is available, see e. g. Christmann and Steinwart (2004), Hable and Christmann (2011) for robustness of SVMs and Koltchinskii and Beznosova (2005), Christmann and Steinwart (2007), and Eberts and Steinwart (2011) for consistency and learning rates. Also in the non-i.i.d. case, some effort has been done in order to find concentration inequalities for different kinds of dependence structures and hence to obtain consistency and learning rates, see e. g. Xu and Chen (2008) and Pan and Xiao (2009).

Concerning qualitative robustness, a lot of generalizations of the original definition in Hampel (1968), which also apply for non-i.i.d. cases, have been proposed, but there is not so much literature which deals with these cases. Papantoni-Kazakos and Gray (1979) and Bustos (1980) for example introduce different kinds of qualitative robustness. Some generalizations of Hampel's theorem for qualitative robustness can be found in Cox (1981), Boente et al. (1982), and Zähle (2015). Qualitative robustness of the bootstrap approximation is also introduced in the i.i.d. case, see Cuevas and Romo (1993), but, to my knowledge, not generalized to non-i.i.d. observations.

In this thesis we generalize Hampel's theorem for qualitative robustness of estimators to Varadarajan processes. That is, Theorem 3.1.3 shows that a sequence of continuous estimators $(S_n)_{n\in\mathbb{N}}$ which can be represented by a statistical operator S, which is continuous in the limiting distribution P, is qualitatively robust for weak Varadarajan processes. Regarding support vector machines, we show that the sequence of estimators which maps the given data set $\mathbf{w}_n$ to the SVM $f_{L,\mathbb{P}_{\mathbf{w}_n},\lambda_n}$ is qualitatively robust under common assumptions on the kernel and the loss function, as long as the sequence of regularization parameters $(\lambda_n)_{n\in\mathbb{N}} \subset (0,\infty)$ converges to $\lambda_0 \neq 0$, $n \to \infty$, see Theorem 4.2.1. Compared to consistency, where $\lambda_n \searrow 0$ is required, we can not achieve qualitative robustness in this case. This is due to the problem, which is a so-called ill-posed problem. This implies that consistency and qualitative robustness can not be achieved simultaneously, see Hable and Christmann (2013). Therefore we regard qualitative robustness for the sequence of estimators where the sequence $(\lambda_n)_{n\in\mathbb{N}} \subset (0,\infty)$ converges to a positive but small value.

Moreover we generalize qualitative robustness to bootstrap approximations in Theorem 3.4.2, 3.4.5, and 3.4.6. We have to strengthen the assumptions on the stochastic process and the sample space $\mathcal{Z}$ and for the last two results the statistical operator is assumed to be uniformly continuous on the space of probability measures on $\mathcal{Z}$. The first theorem refers to the case of independent, but not necessarily identically distributed random variables, the last two results cover some α-mixing processes.

The second part of this thesis, i.e. Chapter 4, focusses on consistency of support vector machines for data generating stochastic processes with different dependence structures. We show that SVMs are L-risk-consistent for such processes in Theorem 4.4.4, under common assumptions on the loss function and on the kernel and under the assumptions that the stochastic process is asymptotically mean stationary and fulfils a convergence condition similar to a law of large numbers. Moreover we assume the sequences of empirical estimates $f_{L,\mathbb{P}_{\mathbf{W}_n(\omega)},\lambda_n}$, $\omega \in \Omega,\ n \in \mathbb{N}$, as well as the sequence of theoretical estimates $f_{L,\frac{1}{n}\sum_{i=1}^n P^i,\lambda_n}$,

$n \in \mathbb{N}$, to be uniformly bounded. Consistency is achieved for many $\mathcal{C}$-mixing, α-mixing and η-, λ-, ζ-, κ- and θ-weakly dependent stochastic processes.

Hence, statistical robustness and consistency can also be shown for non-i.i.d. observations, which enlarges the applicability of SVMs to a broader class of stochastic processes. Of course there are still several open questions concerning consistency and robustness of support vector machines or more general of estimators for non-i.i.d. observations.

A first one is the generalization to other dependence structures, for example, some martingales or other mixing structures might also be Varadarajan processes. In Steinwart et al. (2009) it is shown that some martingales fulfil a law of large numbers for events and therefore Theorem 3.2.1 shows that they are Varadarajan processes, but, from my point of view, the assumption on the process is very strong. So the question in case of martingales is, weather these assumptions can be considerably weakened. We have not been working with these dependence structures as their properties are hard to transfer from the original stochastic process $(Z_i)_{i\in\mathbb{N}}$ to the stochastic process $(f \circ Z_i)_{i\in\mathbb{N}}$, if f is a continuous function for example.

Qualitative robustness of the bootstrap approximation for α-mixing processes is achieved if the statistical operator is uniformly continuous. For independent not necessarily identically distributed random variables this assumption was weakened, see Theorem 3.4.2. It would also be of interest to weaken the assumption of uniform continuity for α-mixing processes, one way might be to achieve a uniform continuity of the bootstrap approximation. Moreover the assumptions on the input space to be totally bounded or compact are strong. These assumptions should also be weakened if possible.

Our proof of consistency of support vector machines is based on two convergence properties of the stochastic process: the AMS property and the almost sure convergence of $\frac{1}{n^{1-r}} \sum_{i=1}^{n} \left(L_{f_n} \circ Z_i - \int L_{f_n} \circ Z_i \, d\mu \right)$ to 0, $n \to \infty$, $0 < r < \frac{1}{2}$, where we do not assume any convergence rates. Hence, we do not achieve learning rates. Trying to assume rates of convergence for the AMS property and for the almost sure convergence could lead to a learning rate. But probably the learning rates would be very bad compared to the i.i.d. case and those which are achieved via other concentration inequalities, see for example Sun and Wu (2009) and Hang and Steinwart (2015).

Moreover both parts, statistical robustness and consistency require the stochastic process to be either asymptotically mean stationary or a Varadarajan process. These properties yield convergence of the mixture distribution $\frac{1}{n} \sum_{i=1}^{n} P^i$, respectively convergence of the empirical distribution of the stochastic process to a limiting distribution. It would be interesting if

these properties are in general implied by the dependence structure of the process, without assuming stationarity or identical distributions.

Also some numerical simulations should be done in order to illustrate qualitative robustness and consistency of SVMs for finite $n \in \mathbb{N}$, under different dependence assumptions on the data generating stochastic process.

Appendix A

On the following pages some definitions which are used in different meanings in the literature or might not be immediately remembered by the reader, can be found.

Definition A 1 (strong equivalence of metrics, Sutherland (1975), p.39) *Two metrics $d_{\mathcal{X}}$ and $d'_{\mathcal{X}}$ on a topological space are called strongly or Lipschitz equivalent if there exist strictly positive constants m, M such that for all $x, x' \in \mathcal{X}$*

$$m d'_{\mathcal{X}}(x, x') \leq d_{\mathcal{X}}(x, x') \leq M d'_{\mathcal{X}}(x, x').$$

Definition A 2 (strong stationarity, Krengel (1985) p.25) *Let $(\Omega, \mathcal{A}, P)$ be a probability space. A stochastic process $(X_i)_{i\in\mathbb{N}}$ on $(\Omega, \mathcal{A}, P)$ is called stationary or strongly stationary, if the distribution of $(X_{s+i})_{i\in\mathbb{N}}$ does not depend on the shift s. That is*

$$P(X_{t_1} \in A_1, X_{t_2} \in A_2, \ldots, X_{t_n} \in A_n) = P(X_{t_1+s} \in A_1, X_{t_2+s} \in A_2, \ldots, X_{t_n+s} \in A_n),$$

applies for all $n, s \in \mathbb{N}$, for all $A_1, \ldots . A_n \in \mathcal{A}$ and all $t_i \in \mathbb{N}$, $i \in \mathbb{N}$.

Definition A 3 (uniform Glivenko-Cantelli class, Dudley et al. (1991) p.2) *Let $\mathcal{Z}$ be a separable metric space and $P \in \mathcal{M}(\mathcal{Z})$. A class of functions $\mathcal{F} := \{f \colon \mathcal{Z} \to \mathbb{R}\}$ is called a uniform Glivenko-Cantelli class if*

$$\lim_{n\to\infty} \sup_{P\in\mathcal{M}(\mathcal{Z})} Pr\left\{ \sup_{m\geq n} \|\mathbb{P}_{\mathbf{w}_m} - P\|_{\mathcal{F}} > \varepsilon \right\} = 0.$$

Where Pr denotes the outer probability and $\mathbb{P}_{\mathbf{w}_n} = \frac{1}{n}\sum_{i=1}^{n} \delta_{z_i}$ an empirical measure of i.i.d. random variables $Z_i \sim P$. $\|G\|_{\mathcal{F}} := \sup\{G(f),\ f \in \mathcal{F}\}$, $G \colon \mathcal{F} \to \mathbb{R}$, here $\|\mathbb{P}_{\mathbf{w}_m} - P\|_{\mathcal{F}} = \sup_{f\in\mathcal{F}} |\int f\, d\mathbb{P}_{\mathbf{w}_m} - \int f\ dP|$.

Definition A 4 (Suslin space, Dudley (1989) p.229) *A separable and measurable space $(\mathcal{Y}, S)$ is called a Suslin space if there is a Polish space $\mathcal{X}$ and a Borel measurable map from $\mathcal{X}$ to $\mathcal{Y}$.*

Definition A 5 (image admissible Suslin, Dudley (1989) p.229) *Let $(\Omega, \mathcal{A})$ be a measurable space and $\mathcal{F}$ a set. Then a real valued function $X\colon (f, \omega) \mapsto X(f, \omega)$ is called image admissible Suslin via $(\mathcal{Y}, S, T)$ if $(\mathcal{Y}, S)$ is a Suslin space, T is a function from $\mathcal{Y}$ to $\mathcal{F}$, and $(y, \omega) \mapsto X(T(y), \omega)$ is jointly measurable on $\mathcal{Y} \times \Omega$.*

Definition A 6 (tight, Billingsley (1999), p.8, p.59) *A probability measure P on a metric space $(\mathcal{X}, d_{\mathcal{X}})$ is tight if for each $\varepsilon > 0$ there exists a compact set $K \subset \mathcal{X}$ such that $P(K) > 1 - \varepsilon$.*
A family $\mathcal{P}$ of probability measures is tight if for every $\varepsilon > 0$ there exists a compact set $K \subset \mathcal{X}$ such that $P(K) > 1 - \varepsilon$ for every $P \in \mathcal{P}$.

Definition A 7 (totally bounded, Dudley (1989) p.35) *Let $(\mathcal{X}, d_{\mathcal{X}})$ be a metric space. The space $(\mathcal{X}, d_{\mathcal{X}})$ is called totally bounded if for every $\varepsilon > 0$ there is a finite set $\mathcal{Y} \subset \mathcal{X}$ such that for every $x \in \mathcal{X}$, there is some $y \in \mathcal{Y}$ with $d(x, y) \leq \varepsilon$.*

Definition A 8 (subdifferential, Denkowski et al. (2003) Definition 5.3.20) *Let X be a Banach space and X^* the topological dual, $f\colon X \to \mathbb{R} \cup \{\infty\}$ be a convex function, and let $x \in X$ with $f(x) < \infty$. Then the subdifferential ∂f of f at x is defined by*

$$\partial f(x) = \{x^* \in X \,, \langle x^*, y - x\rangle \leq f(y) - f(x) \text{ for all } y \in \mathcal{X}\}\,.$$

Definition A 9 (equicontinuous, Dudley (1989) p.39/40) *A collection of functions $\mathcal{F}$ from a topological space $\mathcal{X}$ into $\mathcal{Y}$, where $(\mathcal{Y}, d_{\mathcal{Y}})$ is a metric space, is called equicontinuous if for every $x \in \mathcal{X}$ there is a neighborhood U of x such that $d_{\mathcal{Y}}(f(x), f(x')) \leq \varepsilon$ for all $x' \in U$ and all $f \in \mathcal{F}$.*
If $(\mathcal{X}, d_{\mathcal{X}})$ is a metric space and for every $\varepsilon > 0$ there is a $\delta > 0$ such that $d_{\mathcal{X}}(x, x') \leq \delta$ implies $d_{\mathcal{Y}}(f(x), f(x')) \leq \varepsilon$ for all x and x' in $\mathcal{X}$ and all $f \in \mathcal{F}$ is called uniformly equicontinuous.

Theorem A 10 (maximal inequality Serfling (1970)) *Let $\nu \geq 2$. Let $P_{a,n}$ be the joint distribution of random variables $(Z_{a+1}, \ldots, Z_{a+n})$, $a, n \in \mathbb{N}$. Suppose that there exists a*

function $h(P_{a.n})$*, such that*

$$\begin{aligned} & h(P_{a,k}) + h(P_{a+k,n}) \quad \leq \; h(P_{a,k+n}),\; a,k,n \in \mathbb{N},\; 1 \leq k < k+n \\ \textit{and} \qquad & \mathbb{E}\left|\textstyle\sum_{i=a+1}^{a+n} Z_i\right|^{\nu} \quad \leq \; h^{\frac{1}{2}\nu}(P_{a,n}),\; a,n \in \mathbb{N},\; n \geq 1. \end{aligned}$$

Then,

$$\mathbb{E}\left(\max_{1\leq q\leq n}\left|\sum_{i=a+1}^{a+q} Z_i\right|\right)^{\nu} \;\leq\; (log_2(2n))^{\nu} h^{\frac{1}{2}\nu}(P_{a,n}).$$

The space $D^p[0,1]$(see Bickel and Wichura (1971) p. 1662)

The following descriptions and definition of the space $D^d(T)$ can be found in Bickel and Wichura (1971, Chapter 3, p. 1662):
Let T denote the unit cube $[0,1]^d$. Call a function $X : T \to \mathbb{R}$ a step function if x is a linear combination of functions of the form

$$t \mapsto I_{E_1 \times E_2 \times \ldots \times E_p}(t),$$

where each E_p is either a left-closed, right-open subinterval of $[0,1]$, or the singleton $\{1\}$ and where I_E denotes the indicator of the set E. Let D^d be the uniform closure, in the space of all bounded functions from T to $\mathbb{R}$, of the vector subspace of simple functions. The functions in D^d may be characterized by their continuity properties, as follows. If $t \in T$ and if, for $1 \leq p \leq d$, R_p is one of the relations $<$ and $\geq$, let $Q_{R_1,\ldots,R_d}(t)$ denote the quadrant

$$\{(s_1,\ldots,s_d) \in T; s_p R_p t_p,\; 1 \leq p \leq d\}$$

Then (see Neuhaus (1969), Straf (1969b)), page 29) $x \in D^d$ iff for each $t \in T$

(a) $x_Q = \lim_{s\to t,\, s\in Q} x(s)$ exists for each of the 2^d quadrants $Q_{R_1,\ldots,R_d}(t)$, and

(b) $x(t) = x_{Q_{\geq,\ldots,\geq}}$.

In this sense, the functions of D^d are "continuous from above, with limits from below". One can introduce a metric topology on D^d which for $d = 1$ coincides with Skorohod's well-known and useful J_1-topology (see Billingsley (1999), for example). For this, let Λ be the group of all transformations $\lambda : T \to T$ of the form $\lambda(t_1,\ldots,t_d) = (\lambda(t_1),\ldots,\lambda(t_d))$, where each $\lambda_p : [0,1] \to [0,1]$ is continuous, strictly increasing, and fixes zero and one. Define the "Skorohod" distance between x and y in D^d to be

$$d(x,y) = \inf\{\min(\|x - y\lambda\|, \|\lambda\|) :\; \lambda \in \Lambda\},$$

where $\|x - y\lambda\| = \sup\{|x(t) - y(\lambda(t))|,\ t \in T\}$ and $\|\lambda\| = \sup\{|\lambda(t) - t|,\ t \in T\}$. With respect to the corresponding metric topology (S-topology), D^d is separable and topologically complete, and the Borel σ-algebra $\mathcal{D}^d$ coincides with the σ-algebra generated by the coordinate mappings (Billingsley (1995), Neuhaus (1969), Straf (1969b). Consequently, a stochastic process $(X(t))_{t\in T}$ taking values in D^d is $\mathcal{D}^d$-measurable.

Bibliography

N. Aronszajn. Theory of reproducing kernels. *Transactions of the American mathematical society*, 68(3):337–404, 1950.

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

E. Beutner and H. Zähle. Functional delta-method for the bootstrap of quasi-Hadamard differentiable functionals. *Electron. J. Stat.*, 10, 2016.

P. J. Bickel and P. Bühlmann. A new mixing notion and functional central limit theorems for a sieve bootstrap in time series. *Bernoulli*, 5(3):413–446, 1999.

P. J. Bickel and M. J. Wichura. Convergence criteria for multiparameter stochastic processes and some applications. *Ann. Math. Statist.*, 42:1656–1670, 1971.

P. Billingsley. *Probability and measure*. Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, Inc., New York, third edition, 1995.

P. Billingsley. *Convergence of probability measures*. Wiley Series in Probability and Statistics: Probability and Statistics. John Wiley & Sons, Inc., New York, second edition, 1999.

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

G. Boente, R. Fraiman, and V. J. Yohai. Qualitative robustness for general stochastic processes. Technical report, Department of Statistics, University of Washington, 1982.

G. Boente, R. Fraiman, and V. J. Yohai. Qualitative robustness for stochastic processes. *The Annals of Statistics*, 15(3):1293–1312, 1987.

B. E. Boser, I. M. Guyon, and V. N. Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on computational learning theory*, pages 144–152, 1992.

S. Boucheron, G. Lugosi, and P. Massart. *Concentration inequalities*. Oxford University Press, Oxford, 2013.

R. C. Bradley. Basic properties of strong mixing conditions. Technical report, DTIC Document, 1985.

R. C. Bradley. Basic properties of strong mixing conditions. A survey and some open questions. *Probab. Surv.*, 2:107–144, 2005.

R. C. Bradley. *Introduction to strong mixing conditions. Vol. 1*. Kendrick Press, Heber City, UT, 2007a.

R. C. Bradley. *Introduction to strong mixing conditions. Vol. 2*. Kendrick Press, Heber City, UT, 2007b.

R. C. Bradley. *Introduction to strong mixing conditions. Vol. 3*. Kendrick Press, Heber City, UT, 2007c.

L. Breiman. *Probability*. Addison-Wesley Publishing Company, Reading, Mass., 1968.

P. Bühlmann. Blockwise bootstrapped empirical process for stationary sequences. *Ann. Statist.*, 22(2):995–1012, 1994.

P. Bühlmann. The blockwise bootstrap for general empirical processes of stationary sequences. *Stochastic Process. Appl.*, 58(2):247–265, 1995.

O. Bustos. On qualitative robustness for general processes. unpublished manuscript, 1980.

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

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

A. Christmann, A. Van Messem, and I. Steinwart. On consistency and robustness properties of support vector machines for heavy-tailed distributions. *Stat. Interface*, 2(3):311–327, 2009.

A. Christmann, M. Salibian-Barrera, and S. Van Aelst. On the stability of bootstrap estimators. *arXiv preprint arXiv:1111.1876*, 2011.

A. Christmann, M. Salibián-Barrera, and S. Van Aelst. Qualitative robustness of bootstrap approximations for kernel based methods. In *Robustness and complex data structures*, pages 263–278. Springer, Heidelberg, 2013.

J. B. Conway. *A course in functional analysis*, volume 96 of *Graduate Texts in Mathematics*. Springer, New York, 1985.

D. D. Cox. metrics on stochastic processes and qualitative robustness. Technical report, Department of Statistics, University of Washington, 1981.

N. Cristianini and J. Shawe-Taylor. *An introduction to support vector machines and other kernel-based learning methods*. Cambridge university press, 2000.

F. Cucker and D. X. Zhou. *Learning theory: an approximation theory viewpoint*, volume 24. Cambridge University Press, 2007.

A. Cuevas. Qualitative robustness in abstract inference. *Journal of Statistical Planning and Inference*, 18(3):277–289, 1988.

A. Cuevas and J. Romo. On robustness properties of bootstrap approximations. *J. Statist. Plann. Inference*, 37(2):181–191, 1993.

E. De Vito, L. Rosasco, A. Caponnetto, M. Piana, and A. Verri. Some properties of regularized kernel methods. *J. Mach. Learn. Res.*, 5:1363–1390, 2003/04.

E. De Vito, A. Caponnetto, and L. Rosasco. Model selection for regularized least-squares algorithm in learning theory. *Found. Comput. Math.*, 5(1):59–85, 2005.

J. Dedecker and C. Prieur. New dependence coefficients. examples and applications to statistics. *Probability Theory and Related Fields*, 132(2):203–236, 2005.

J. Dedecker, P. Doukhan, G. Lang, J. R. León R., S. Louhichi, and C. Prieur. *Weak dependence: with examples and applications*, volume 190 of *Lecture Notes in Statistics*. Springer, New York, 2007.

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

P. Doukhan. *Mixing*. Springer, New York, 1994.

P. Doukhan and S. Louhichi. A new weak dependence condition and applications to moment inequalities. *Stochastic Processes and their Applications*, 84(2):313–342, 1999.

R. M. Dudley. *Real Analysis and Probability.* Chapman&Hall, New York, 1989.

R. M. Dudley. *Uniform central limit theorems*, volume 63 of *Cambridge Studies in Advanced Mathematics.* Cambridge University Press, Cambridge, 2014.

R. M. Dudley, E. Giné, and J. Zinn. Uniform and universal Glivenko-Cantelli classes. *J. Theoret. Probab.*, 4(3):485–510, 1991.

N. Dunford and J. T. Schwartz. *Linear Operators. I. General Theory.* With the assistance of W. G. Bade and R. G. Bartle. Pure and Applied Mathematics, Vol. 7. Interscience Publishers, Inc., New York; Interscience Publishers, Ltd., London, 1958.

M. Eberts and I. Steinwart. Optimal learning rates for least squares svms using gaussian kernels. In J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 24*, pages 1539–1547. Curran Associates, Inc., 2011.

B. Efron. Bootstrap methods: another look at the jackknife. *Ann. Statist.*, 7(1):1–26, 1979.

B. Efron and R. J. Tibshirani. *An introduction to the bootstrap*, volume 57 of *Monographs on Statistics and Applied Probability.* Chapman and Hall, New York, 1993.

T. Fender. *Empirische Risiko-Minimierung für dynamische Datenstrukturen.* PhD thesis, Universität Dortmund, 2003. URL `http://hdl.handle.net/2003/2788`.

R. M. Gray. *Probability, random processes, and ergodic properties.* Springer, New York, 1988.

R. Hable. Universal consistency of localized versions of regularized kernel methods. *Journal of Machine Learning Research (JMLR)*, 14:153–186, 2013.

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

R. Hable and A. Christmann. Robustness versus consistency in ill-posed classification and regression problems. In A. Giusti, G. Ritter, and M. Vichi, editors, *Classification and Data Mining*, pages 27–35. Springer, 2013.

F. R. Hampel. *Contributions to the theory of robust estimation.* PhD thesis, Univ. California, Berkeley, 1968.

F. R. Hampel. A general qualitative definition of robustness. *Annals of Mathematical Statistics*, 42:1887–1896, 1971.

F. R. Hampel, E. M. Ronchetti, P. J. Rousseeuw, and W. A. Stahel. *Robust statistics*. John Wiley & Sons, Inc., New York, 1986.

H. Hang and I. Steinwart. A bernstein-type inequality for some mixing processes and dynamical systems with an application to learning. 2015.

T. Hastie, R. Tibshirani, and J. Friedman. *The elements of statistical learning*. Springer Series in Statistics. Springer-Verlag, New York, 2001.

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

T.-C. Hu, A. Rosalsky, and A. Volodin. On convergence properties of sums of dependent random variables under second moment and covariance restrictions. *Statist. Probab. Lett.*, 78(14):1999–2005, 2008.

P. J. Huber. *Robust statistics*. John Wiley & Sons Inc., New York, 1981.

J. Jurečková and J. Picek. *Robust statistical methods with R*. Chapman & Hall/CRC, Boca Raton, FL, 2006.

A. Klenke. *Probability theory: a comprehensive course*. Springer Science & Business Media, 2013.

V. Koltchinskii and O. Beznosova. Exponential convergence rates in classification. In *Learning theory*, volume 3559 of *Lecture Notes in Comput. Sci.*, pages 295–307. Springer, Berlin, 2005.

V. Krätschmer, A. Schied, and H. Zähle. Domains of weak continuity of statistical functionals with a view toward robust statistics. *J. Multivariate Anal.*, 158:1–19, 2017.

U. Krengel. *Ergodic theorems*, volume 6 of *de Gruyter Studies in Mathematics*. Walter de Gruyter & Co., Berlin, 1985.

S. Kulkarni, A. C. Lozano, and R. E. Schapire. Convergence and consistency of regularized boosting algorithms with stationary *beta*-mixing observations. In *Advances in neural information processing systems*, pages 819–826, 2005.

H. R. Künsch. The jackknife and the bootstrap for general stationary observations. *Ann. Statist.*, 17(3):1217–1241, 1989.

S. N. Lahiri. *Resampling methods for dependent data*. Springer Series in Statistics. Springer, New York, 2003.

R. Y. Liu and K. Singh. Moving blocks jackknife and bootstrap capture weak dependence. In *Exploring the limits of bootstrap (East Lansing, MI, 1990)*, Wiley Ser. Probab. Math. Statist. Probab. Math. Statist., pages 225–248. Wiley, New York, 1992.

R. A. Maronna, R. D. Martin, and V. J. Yohai. *Robust statistics*. Wiley Series in Probability and Statistics. John Wiley & Sons Ltd., Chichester, 2006.

V. Maume-Deschamps. Exponential inequalities and functional estimations for weak dependent data; applications to dynamical systems. *Stoch. Dyn.*, 6(4):535–560, 2006.

S. Mukherjee, E. Osuna, and F. Girosi. Nonlinear prediction of chaotic time series using support vector machines. In *Neural Networks for Signal Processing [1997] VII. Proceedings of the 1997 IEEE Workshop*, pages 511–520, 1997.

K.-R. Müller, A. J. Smola, G. Rätsch, B. Schölkopf, J. Kohlmorgen, and V. Vapnik. Predicting time series with support vector machines. In *International Conference on Artificial Neural Networks*, pages 999–1004, 1997.

U. V. Naik-Nimbalkar and M. B. Rajarshi. Validity of blockwise bootstrap for empirical processes with stationary observations. *Ann. Statist.*, 22(2):980–994, 1994.

G. Neuhaus. *Zur Theorie der Konvergenz stochastischer Prozesse mit mehrdimensionalem Zeitparameter*. PhD thesis, 1969.

Z.-W. Pan and Q.-W. Xiao. Least-square regularized regression with non-iid sampling. *Journal of Statistical Planning and Inference*, 139(10):3579–3587, 2009.

P. Papantoni-Kazakos and R. M. Gray. Robustness of estimators on stationary observations. *The Annals of Probability*, 7(6):989–1002, 1979.

K. R. Parthasarathy. *Probability measures on metric spaces*, volume 352. American Mathematical Soc., 1967.

M. Peligrad. On the blockwise bootstrap for empirical processes for stationary sequences. *Ann. Probab.*, 26(2):877–901, 1998.

T. Poggio and F. Girosi. A sparse representation for function approximation. *Neural computation*, 10(6):1445–1454, 1998.

D. N. Politis and J. P. Romano. A circular block-resampling procedure for stationary data. In *Exploring the limits of bootstrap (East Lansing, MI, 1990)*, Wiley Ser. Probab. Math. Statist. Probab. Math. Statist., pages 263–270. 1990.

D. Radulović. The bootstrap for empirical processes based on stationary observations. *Stochastic Process. Appl.*, 65, 1996.

M. Rosenblatt. A central limit theorem and a strong mixing condition. *Proc. Nat. Acad. Sci. U. S. A.*, 42:43–47, 1956.

A. L. Samuel. Some studies in machine learning using the game of checkers. *IBM Journal of research and development*, 3(3):210–229, 1959.

B. Schölkopf and A. J. Smola. *Learning with Kernels*. Massachusetts Institute of Technology, Cambridge, 2002.

R. J. Serfling. Moment inequalities for the maximum cumulative sum. *Ann. Math. Statist.*, 41:1227–1234, 1970.

J. Shao and D. S. Tu. *The jackknife and bootstrap*. Springer Series in Statistics. Springer New York, 1995.

Q. M. Shao and H. Yu. Bootstrapping the sample means for stationary mixing sequences. *Stochastic Process. Appl.*, 48(1):175–190, 1993.

K. Singh. On the asymptotic accuracy of Efron's bootstrap. *Ann. Statist.*, 9(6):1187–1195, 1981.

S. Smale and D.-X. Zhou. Online learning with markov sampling. *Analysis and Applications*, 7(01):87–113, 2009.

I. Steinwart. Support vector machines are universally consistent. *J. Complexity*, 18(3): 768–791, 2002.

I. Steinwart. Consistency of support vector machines and other regularized kernel classifiers. *IEEE Trans. Inform. Theory*, 51(1):128–142, 2005.

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

I. Steinwart and A. Christmann. Estimating conditional quantiles with the help of the pinball loss. *Bernoulli*, 17(1):211–225, 2011.

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

I. Steinwart, D. Hush, and C. Scovel. Learning from dependent observations. *Journal of Multivariate Analysis*, 100:175–194, 2009.

M. L. Straf. A general skorohod space, 1969a.

M. L. Straf. *A general Skorohod space and its application to the weak convergence of stochastic processes with several parameters.* PhD thesis, University of Chicago, Department of Statistics, 1969b.

M. L. Straf. Weak convergence of stochastic processes with several parameters. pages 187–221, 1972.

K. Strohriegl. Qualitative robustness for bootstrap approximations. *arXiv preprint arXiv:1702.05933*, 2017.

K. Strohriegl and R. Hable. On qualitative robustness for stochastic processes. *Metrika*, pages 895–917, 2016.

H. Sun and Q. Wu. A note on application of integral operator in learning theory. *Applied and Computational Harmonic Analysis*, 26(3):416–421, 2009.

W. Sutherland. Introduction to metric and topological spaces. 1975.

A. W. van der Vaart. *Asymptotic statistics*, volume 3 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, 1998.

V. N. Vapnik. *The nature of statistical learning theory.* Springer, New York, 1995.

V. N. Vapnik. *Statistical learning theory.* Adaptive and Learning Systems for Signal Processing, Communications, and Control. John Wiley & Sons Inc., New York, 1998.

Y.-L. Xu and D.-R. Chen. Learning rates of regularized regression for exponentially strongly mixing sequence. *Journal of Statistical Planning and Inference*, 138(7):2180–2189, 2008.

H. Zähle. Qualitative robustness of statistical functionals under strong mixing. *Bernoulli*, 21(3):1412–1434, 2015.

H. Zähle. A definition of qualitative robustness for general point estimators, and examples. *Journal of Multivariate Analysis*, 143:12–31, 2016.

T. Zhang. Statistical behavior and consistency of classification methods based on convex risk minimization. *Ann. Statist.*, 32(1):56–85, 2004.

B. Zou, L. Li, and Z. Xu. The generalization performance of erm algorithm with strongly mixing observations. *Machine learning*, 75(3):275–295, 2009a.

B. Zou, H. Zhang, and Z. Xu. Learning from uniformly ergodic markov chains. *Journal of Complexity*, 25(2):188–200, 2009b.

Publications:

K. Strohriegl and R. Hable. On qualitative robustness for stochastic processes. Metrika, pages 895-917, 2016.

K. Strohriegl. Qualitative robustness for bootstrap approximations. arXiv:1702.05933 [math.ST]. 20.02.2017.

Eidesstattliche Versicherung

Hiermit versichere ich an Eides statt, dass ich die vorliegende Arbeit selbstständig verfasst und keine anderen als die von mir angegebenen Quellen und Hilfsmittel verwendet habe.

Weiterhin erkläre ich, dass ich die Hilfe von gewerblichen Promotionsberatern bzw. Promotionsvermittlern oder ähnlichen Dienstleistern weder bisher in Anspruch genommen habe, noch künftig in Anspruch nehmen werde.

Zusätzlich erkläre ich hiermit, dass ich keinerlei frühere Promotionsversuche unternommen habe.

Bayreuth, den