SUPPORT VECTOR MACHINES FOR FORECASTING THE EVOLUTION OF AN UNKNOWN ERGODIC DYNAMICAL SYSTEM FROM OBSERVATIONS WITH UNKNOWN NOISE

BY INGO STEINWART AND MARIAN ANGHEL

Los Alamos National Laboratory

We consider the problem of forecasting the next (observable) state of an unknown ergodic dynamical system from a noisy observation of the present state. Our main result shows, e.g., that support vector machines (SVMs) using Gaussian RBF kernels can learn the best forecaster from a sequence of noisy observations if a) the unknown observational noise processes is bounded and has a summable $\alpha$-mixing rate and b) the unknown ergodic dynamical system is defined by a Lipschitz continuous function on some compact subset of $\mathbb{R}^d$ and has a summable decay of correlations for Lipschitz continuous functions. In order to prove this result we first establish a general consistency result for SVMs and all stochastic processes that satisfy a mixing notion that is substantially weaker than $\alpha$-mixing.

1. Introduction. Let us assume that we have an ergodic dynamical system described by the sequence $(F^n)_{n \geq 0}$ of iterates of an (essentially) unknown map $F : M \to M$, where $M \subset \mathbb{R}^d$ is compact and the corresponding ergodic measure $\mu$ is assumed to be unique. Furthermore, assume that all observations $\tilde{x}$ of this dynamical system are corrupted by some stationary, $\mathbb{R}^d$-valued, additive noise process $\mathcal{E} = (\varepsilon_n)_{n \geq 0}$ whose distribution $\nu$ we assume to be independent of the state, but otherwise unknown, too. In other words all possible observations of the system at time $n \geq 0$ are of the form

\begin{equation}
\tilde{x}_n = F^n(x_0) + \varepsilon_n,
\end{equation}

where $x_0$ is a true but unknown state at time 0. Now, given an observation of the system at some arbitrary time our goal is to forecast the next observable state (we will see later that under some circumstances this is equivalent to forecasting the next true state) i.e., given $x + \varepsilon$ we want to forecast $F(x) + \varepsilon'$, where $\varepsilon$ and $\varepsilon'$ are the observational errors for $x$ and its successor.

AMS 2000 subject classifications: Primary 62M20; secondary 37D25, 37C99, 37M10, 60K99, 62M10, 62M45, 68Q32, 68T05

Keywords and phrases: Observational noise model, Forecasting dynamical systems, Support vector machines, Consistency
F(x). Of course, if we know neither F nor \( \nu \) then this task is impossible, and hence we assume that we have a finite sequence \( T = (\tilde{x}_0, \ldots, \tilde{x}_{n-1}) \) of noisy observations from a trajectory of the dynamical system, i.e., all \( \tilde{x}_i, i = 0, \ldots, n-1 \), are given by (1) for a conjoint initial state \( x_0 \). Now, informally speaking our goal is to use \( T \) to build a forecaster \( f : \mathbb{R}^d \to \mathbb{R}^d \) whose average forecasting performance on future noisy observations is as small as possible. In order to render this goal more precisely we need a loss function \( L : \mathbb{R}^d \to [0, \infty) \) such that

\[
L(F(x) + \varepsilon' - f(x + \varepsilon))
\]
gives a value for the discrepancy between the forecast \( f(x + \varepsilon) \) and the observed next state \( F(x) + \varepsilon' \). In the following, we always assume implicitly, that small values of \( L(F(x) + \varepsilon' - f(x + \varepsilon)) \) correspond to small values of \( \|F(x) + \varepsilon' - f(x + \varepsilon)\|_2 \), where \( \| \cdot \|_2 \) denotes the Euclidean distance in \( \mathbb{R}^d \).

Now, by the stationarity of \( \mathcal{E} \) the average forecasting performance is given by the \( L \)-risk

\[
\mathcal{R}_{L,P}(f) := \int \int L(F(x) + \varepsilon_1 - f(x + \varepsilon_0)) \nu(\varepsilon_1) \mu(\varepsilon_0) \,
\]

where \( \varepsilon = (\varepsilon_i)_{i \geq 0} \) and \( P := \nu \otimes \mu \). Obviously, the smaller the risk the better the forecaster is, and hence we ideally would like to have a forecaster \( f_L^* : \mathbb{R}^d \to \mathbb{R}^d \) that attains the minimal \( L \)-risk

\[
\mathcal{R}_{L,P}^* := \inf \{ \mathcal{R}_{L,P}(f) \mid f : \mathbb{R}^d \to \mathbb{R}^d \text{ measurable} \} .
\]

Now assume that we have a method \( \mathcal{L} \) that assigns to every training set \( T \) a forecaster \( f_T \). Then the method \( \mathcal{L} \) achieves our goal asymptotically, if it is consistent in the sense of

\[
\lim_{n \to \infty} \mathcal{R}_{L,P}(f_T) = \mathcal{R}_{L,P}^* ,
\]

where the limit is in probability \( P \).

To our best knowledge the forecasting problem (1) – (4) has not been considered in the literature, and even the observational noise model itself has only been considered sporadically, though it clearly “captures important features of many experimental situations”, [27]. Moreover, most of the existing work on the observational noise model deals with the question of denoising [24, 17, 35, 23, 25, 26, 27]. In particular, [25, 26, 27] provide both positive and negative results on the existence of consistent denoising procedures.

In [32] a related forecasting goal is considered for the least squares loss and stochastic processes of the form \( Z_{i+1} := F(Z_i) + \varepsilon_{i+1}, i \geq 0 \), where \( (F) \) is a
dynamical system and \((\varepsilon_i)\) is some additive centered i.i.d. dynamical noise. In particular, consistency of two histogram-based methods is established if

\begin{itemize}
  \item[a)] \(F : M \to M\) is continuous and \((\varepsilon_i)\) is bounded, or
  \item[b)] \(F\) is bounded and \(\varepsilon_i\) is absolutely continuous, respectively.
\end{itemize}

Note that the first case shows that in the absence of dynamical and observational noise there is a method which can learn to identify \(F\) whenever it is continuous but otherwise unknown. However, it is unclear how to extend the methods of [32] to deal with observational noise.

Variants of the forecasting problem for general stationary ergodic processes \((Z_i)\) have been extensively studied in the literature. One often considered variant is static autoregression (see [22] p. 569ff and the references therein) where the goal is to find sequences \(f_m(Z_{-1}, \ldots, Z_{-m})\) of estimators that converge almost surely to \(\mathbb{E}(Z_0 | Z_{-1}, \ldots, Z_{-\infty})\), which is known to be the least squares optimal one-step-ahead forecaster using an infinite past of observations. However, even if forecasters using a longer history of observations are considered in (2) – (4), the goal of static autoregression cannot be compared to our concept of consistency. Indeed, in static autoregression the goal is to find a near optimal prediction for \(\tilde{x}_0\) using the previously observed \(\tilde{x}_{-1}, \ldots, \tilde{x}_{-m}\) of the same trajectory, whereas our goal is to use the observations to build a predictor which predicts near optimal for arbitrary future observations. In machine learning terminology static autoregression is thus an “on-line” learning problem whereas our notion of consistency defines a “batch” learning problem.

Learning methods for estimating \(\mathbb{E}(Z_0 | Z_{-1}, \ldots, Z_{-\infty})\) in a sense similar to (4) are considered by, e.g., [30, 29] but unfortunately these methods require \(\alpha\)- or \(\beta\)-mixing conditions for \((Z_i)\) that cannot be satisfied by non-trivial dynamical systems. Finally, a result by Nobel [31] shows that there is no method that is universally consistent for, e.g., classification and regression problems where the data is generated by an arbitrary stationary ergodic process \((Z_i)\). In particular this result shows that our general consistency Theorem 3.4 cannot be extended to such \((Z_i)\).

If the observational noise process \(\mathcal{E}\) is mixing in the ergodic sense then it is not hard to check that the process described by (1) is ergodic and hence it satisfies a strong law of large numbers by Birkhoff’s theorem. Using the recent results in [39] we then see that there exists a support vector machine (see the next section for a description) depending on \(F\) and \(\mathcal{E}\) which is consistent in the sense of (4). However, [39] does not provide an explicit method for finding a consistent SVM even if both \(F\) and \(\mathcal{E}\) are known. Consequently, it is fair to say that though SVMs do not have principal limitations for the forecasting problem (1) – (4) there is currently no theoretically sound way...
to use them. The goal of this work is to address this issue by showing that certain SVMs are consistent for all Lipschitz continuous $F$ and bounded $E$ that have a sufficiently fast decay of correlations for Lipschitz continuous functions. In particular, we show that these SVMs are consistent for, e.g., all uniformly smooth expanding or hyperbolic dynamics $F$ and all bounded i.i.d. noise processes $E$.

The rest of this work is organized as follows: In Section 2 we recall the definition of support vector machines (SVMs). Then, in Section 3 we present a consistency result for SVMs and general stochastic processes which have a sufficiently fast decay of correlations. This result is then applied to the above forecasting problem in Section 4, where we also briefly review some dynamical systems with sufficiently fast decay of correlations. Possible future extensions of this work are discussed in Section 5. Finally, the proofs of the two main results can be found in Section 6 and Section 7, respectively.

2. Support Vector Machines. The goal of this section is to briefly describe support vector machines which were first introduced by [7,15] as a method for learning binary classification tasks. Since then they were generalized to other problem domains such as regression and anomaly detection, and nowadays they are considered to be one of the state-of-the-art machine learning methods for these problem domains. For an thorough introduction to SVMs we refer the reader to the books [42,16,36].

Let us begin by introducing some notions related to SVMs. To this end let us fix two non-empty closed sets $X \subset \mathbb{R}^d$ and $Y \subset \mathbb{R}$, and a measurable function $L : X \times Y \times \mathbb{R} \rightarrow [0,\infty)$, which in the following is called loss function.

For a finite sequence $T \in (X \times Y)^n$ and a function $f : X \rightarrow \mathbb{R}$ we define the empirical $L$-risk by

$$R_{L,T}(f) := \frac{1}{n} \sum_{i=0}^{n-1} L(x_i, y_i, f(x_i)).$$

Moreover, for a distribution $P$ on $X \times Y$ we write

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

and $R^*_{L,P} := \inf \{ R_{L,P}(f) \mid f : \mathbb{R}^d \rightarrow \mathbb{R}^d \text{ measurable} \}$ for the $L$-risk and minimal $L$-risk associated to $P$. Now, let $H$ be the reproducing kernel Hilbert space (RKHS) of a measurable kernel $k : X \times X \rightarrow \mathbb{R}$ (see [1] for a general

1Note that this is a more general concept of a loss function than the informal notion of a loss function used in the introduction.
theory of such spaces). Given a finite sequence $T \in (X \times Y)^n$ and a regularization parameter $\lambda > 0$, support vector machines construct a function $f_{T,\lambda,H} : X \to \mathbb{R}$ satisfying
\begin{equation}
\lambda \|f_{T,\lambda,H}\|_H^2 + \mathcal{R}_{L,T}(f_{T,\lambda,H}) = \inf_{f \in H} \lambda \|f\|_H^2 + \mathcal{R}_{L,T}(f).
\end{equation}

In the following we are mainly interested in the commonly used Gaussian RBF kernels $k_\sigma : X \times X \to \mathbb{R}$ defined by
\begin{equation}
k_\sigma(x,x') := \exp(-\sigma^2 \|x - x'\|^2_2), \quad x,x' \in X,
\end{equation}
where $X \subset \mathbb{R}^d$ is a non-empty subset and $\sigma > 0$ is a free parameter called the width. We write $H_\sigma(X)$ for the corresponding RKHSs which are described in some detail in [40]. Finally, for SVMs using a Gaussian kernel $k_\sigma$ we write $f_{T,\lambda,\sigma} := f_{T,\lambda,H_\sigma(X)}$ in order to simplify notations.

It is well-known that if $L$ is a convex loss function in the sense that $L(x,y,\cdot) : \mathbb{R} \to [0,\infty)$ is convex for all $(x,y) \in X \times Y$, then there exists a unique $f_{T,\lambda,H}$. Moreover, in this case (5) becomes a strictly convex optimization problem which can be solved by, e.g., simple gradient descent algorithms. However, for specific losses including the least squares loss other, more efficient algorithmic approaches are used in practice, see [43], [42], [36], and [41]. Let us now introduce some additional properties of loss functions:

**Definition 2.1** A loss function $L : X \times Y \times \mathbb{R} \to [0,\infty)$ is called
\begin{enumerate}
  \item[i)] differentiable if $L(x,y,\cdot) : \mathbb{R} \to [0,\infty)$ is differentiable for all $(x,y) \in X \times Y$. In this case the derivative is denoted by $L'(x,y,\cdot)$.
  \item[ii)] locally Lipschitz continuous if for all $a \geq 0$ there exists a constant $c_a \geq 0$ such that for all $x \in X$, $y \in Y$ and all $t,t' \in [-a,a]$ we have
  \begin{equation}
  |L(x,y,t) - L(x,y,t')| \leq c_a |t - t'|.
  \end{equation}
  In this case the smallest possible constant $c_a$ is denoted by $|L|_{a,1}$.
  \item[iii)] Lipschitz continuous if $|L|_1 := \sup_{a \geq 0} |L|_{a,1} < \infty$.
\end{enumerate}

With the help of these definitions we can now summarize assumptions on the loss function $L$ that we will use frequently.

**Assumption L:** The loss $L : X \times Y \times \mathbb{R} \to [0,\infty)$ is convex, differentiable and locally Lipschitz continuous in the above sense, and it also satisfies $L(x,y,0) \leq 1$ for all $(x,y) \in X \times Y$. Moreover, for the derivative $L'$ there exists a constant $c \in [0,\infty)$ such that for all $(x,y,t),(x',y',t') \in X \times Y \times \mathbb{R}$ we have $|L'(x,y,0)| \leq c$ and
\begin{equation}
|L'(x,y,t) - L'(x',y',t')| \leq c \|_{2}(x,y,t) - (x',y',t')\|_2.
\end{equation}
Note that combining the two assumptions on \(L\) yields \(|L'(x, y, t)| \leq c(1 + |t|)\) for all \((x, y, t) \in X \times Y \times \mathbb{R}\), and from this it is not hard to conclude that \(|L|_{a,1} \leq c(1 + a)\) for all \(a > 0\).

Since the Assumption \(L\) is rather complex let us now illustrate it for two particular classes of loss functions used in many SVM variants.

**Example 2.2** A loss \(L : X \times Y \times \mathbb{R} \rightarrow [0, \infty)\) of the form \(L(x, y, t) = \varphi(yt)\) for a suitable function \(\varphi : \mathbb{R} \rightarrow \mathbb{R}\) and all \(x \in X\), \(y \in Y := \{-1, 1\}\) and \(t \in \mathbb{R}\), is called margin-based. Obviously, \(L\) is convex, continuous, (locally) Lipschitz continuous, or differentiable if and only if \(\varphi\) is. In addition, convexity of \(L\) implies local Lipschitz continuity of \(L\). Furthermore recall that \(\psi\) showed that \(L\) is suitable for binary classification tasks if and only if \(\varphi\) is differentiable at 0 with \(\varphi'(0) < 0\).

Let us now consider Assumption \(L\). Obviously, the first part is satisfied if and only if \(\varphi\) is convex and differentiable, and also satisfies \(\varphi(0) \leq 1\). Moreover, the latter can always be ensured by rescaling \(\varphi\). Furthermore, we have \(L'(x, y, t) = y\varphi'(yt)\) and by considering the cases \(y = y'\) and \(y \neq y'\) separately we see that \(\psi\) is satisfied if and only if \(\varphi\) is Lipschitz continuous and satisfies

\[
|\varphi'(t) + \varphi'(t')| \leq c(1 + |t + t'|),
\]

for a constant \(c > 0\). Finally, the condition \(|L'(x, y, 0)| = |\varphi'(0)| \leq c\) is always satisfied for sufficiently large \(c\). From these considerations we conclude that the classical SVM losses \(\varphi(t) = (1 - t)_+^2\) and \(\varphi(t) = (1 - t)^2_+\), where \((x)_+ := \max\{0, x\}\), do not satisfy Assumption \(L\), whereas the least square loss and the logistic loss defined \(\varphi(t) = (1 - t)^2\) and \(\varphi(t) = \ln(1 + \exp(-t))\), respectively, fulfill \(L\).

**Example 2.3** A loss \(L : X \times Y \times \mathbb{R} \rightarrow [0, \infty)\) of the form \(L(y, t) = \psi(y - t)\) for a suitable function \(\psi : \mathbb{R} \rightarrow \mathbb{R}\) and all \(x \in X\), \(y \in Y \subset \mathbb{R}\) and \(t \in \mathbb{R}\), is called distance-based. Recall that distance-based losses such as the least squares loss \(\psi(r) = r^2\), Huber’s insensitive loss \(\psi(r) = \min\{r^2, \max\{1, 2|r| - 1\}\}\), the logistic loss \(\psi(r) = \ln((1 + e^r)^2 e^{-r}) - \ln 4\), or the \(\epsilon\)-insensitive loss \(\psi(r) = (|r| - \epsilon)_+\) are usually used for regression.

In order to consider Assumption \(L\) we assume that \(Y\) is a compact subset of \(\mathbb{R}\). Then it is easy to see that the first part of \(L\) is satisfied if and only if \(\psi\) is convex and differentiable, and also satisfies \(\sup_{y \in Y} \psi(y) \leq 1\). Note that the latter can always be ensured by rescaling \(\psi\) since the convexity of \(\psi\) implies its continuity. Furthermore, we have \(L'(x, y, t) = -\psi'(y - t)\), and hence we see that \(\psi\) is satisfied if and only if \(\psi\) is Lipschitz continuous. Finally, every convex and differentiable function is continuously differentiable and hence we can always ensure \(|L'(x, y, 0)| = |\psi'(y)| \leq c\). From these considerations we immediately see that all of the above distance-based losses besides the \(\epsilon\)-insensitive loss satisfy Assumption \(L\).

**3. Consistency of SVMs for a Class of Stochastic Processes.** The goal of this section is to establish consistency of SVMs for a class of stochastic processes having a uniform decay of correlations for Lipschitz continuous functions.
functions. This result will then be used to establish consistency of SVMs for the forecasting problem and suitable combinations of dynamical systems $F$ and noise processes $E$.

Let us begin with some notations. To this end let us assume that we have a probability space $(\Omega, \mathcal{A}, \mu)$, a measurable space $(Z, \mathcal{B})$, and a measurable map $T : \Omega \to Z$. Then $\sigma(T)$ denotes the smallest $\sigma$-algebra on $\Omega$ for which $T$ is measurable. Moreover, $\mu_T$ denotes the $T$-image measure of $\mu$, which is defined by $\mu_T(B) := \mu(T^{-1}(B))$, $B \subset Z$ measurable. Recall that a stochastic process $Z := (Z_n)_{n \geq 0}$, i.e., a sequence of measurable maps $Z_n : \Omega \to Z$, $n \geq 0$, is called identically distributed if $\mu_{Z_n} = \mu_{Z_m}$ for all $n, m \geq 0$. In this case we write $P := \mu_{Z_0}$ in the following. Moreover, $Z$ is called stationary in the wide sense if $\mu(Z_{i_1+i}, Z_{i_2+i}) = \mu(Z_{i_1}, Z_{i_2})$ for all $i_1, i_2, i \geq 1$, and it is said to be stationary if $\mu(Z_{i_1+i, \ldots, i_n+i}) = \mu(Z_{i_1, \ldots, i_n})$ for all $n, i, i_1, \ldots, i_n \geq 1$.

The following definition introduces the correlation sequence for stochastic processes that will be used throughout this work.

**Definition 3.1** Let $(\Omega, \mathcal{A}, \mu)$ be a probability space, $(Z, \mathcal{B})$ be a measurable space, $Z$ be a $Z$-valued, identically distributed process on $\Omega$ and $P := \mu_{Z_0}$. Then for $\psi, \varphi \in L_2(P)$ the $n$-th correlation, $n \geq 0$, is defined by

$$\text{cor}_{Z,n}(\psi, \varphi) := \int_{\Omega} \psi(Z_0) \cdot \varphi(Z_n) \, d\mu - \int_{Z} \psi \, d\mu \cdot \int_{Z} \varphi \, d\mu.$$

Obviously, if $Z$ is an i.i.d. process we have $\text{cor}_{Z,n}(\psi, \varphi) = 0$ for all $\varphi, \psi \in L_2(P)$ and $n \geq 0$, and this remains true if $\psi \circ Z_0$ and $\varphi \circ Z_n$ are uncorrelated. Consequently, if $\lim_{n \to \infty} \text{cor}_{Z,n}(\psi, \varphi) = 0$ the corresponding speed of convergence provides information about how fast $\psi \circ Z_0$ becomes uncorrelated from $\varphi \circ Z_n$. This idea has been intensively used in the statistical literature in terms of, e.g., the $\alpha$-mixing coefficients

$$\alpha(Z, n) := \sup_{A \in \mathcal{F}_i^0} \left| \mu(A \cap B) - \mu(A)\mu(B) \right|,$$

where $\mathcal{F}_i$ is the initial $\sigma$-algebra of $Z_i, \ldots, Z_j$. These and related (stronger) coefficients together with examples including, e.g., certain Markov chains, ARMA processes, and GARCH processes are discussed in detail in the survey article [10] and the books [8, 21, 11]. Moreover, for processes $Z$ satisfying $\alpha(Z, n) \leq cn^{-\alpha}$ for some constant $c > 0$ and all $n \geq 1$ it was recently described in [39] how to find a regularization sequence $(\lambda_n)$ for which the corresponding SVM is consistent. Unfortunately, however, it is well-known that every non-trivial ergodic dynamical system does not satisfy $\lim_{n \to \infty} \alpha(Z, n) = 0$, and therefore the result of [39] cannot be used.
to investigate consistency for the forecasting problem. On the other hand various dynamical systems enjoy a uniform decay rate over smaller sets of functions such as Lipschitz continuous functions (see Section 4 for some examples). This leads to the following definition.

**Definition 3.2** Let \((\Omega, \mathcal{A}, \mu)\) be a probability space, \(Z \subset \mathbb{R}^d\) be a compact set, \(Z\) be a \(Z\)-valued, identically distributed process on \(\Omega\) and \(P := \mu_{Z_0}\). Moreover, let \((\gamma_i)_{i \geq 0}\) be a strictly positive sequence converging to 0. Then \(Z\) is said to have a decay of correlations of the order \((\gamma_i)\) if for all \(\psi, \varphi \in \text{Lip}(Z)\) there exists a constant \(\kappa_{\psi, \varphi} \in [0, \infty)\) such that

\[
|\text{corr}_{Z,i}(\psi, \varphi)| \leq \kappa_{\psi, \varphi} \gamma_i, \quad i \geq 0,
\]

where \(\text{Lip}(Z)\) denotes the set of all Lipschitz continuous \(f : Z \to \mathbb{R}\).

Let us now summarize our assumptions on the process \(Z\) which we will make in the rest of this section.

**Assumption Z:** The process \(Z = (X_i, Y_i)_{i \geq 0}\) is defined on the probability space \((\Omega, \mathcal{A}, \mu)\) and is \(X \times Y\)-valued, where \(X \subset \mathbb{R}^d\) and \(Y \subset \mathbb{R}\) are compact subsets. Moreover \(Z\) is stationary in the wide sense.

Finally, we will need the following mutually exclusive assumptions on the regularization sequence and the kernel width of SVMs.

**Assumption S1:** For a fixed strictly positive sequence \((\gamma_i)_{i \geq 0}\) converging to \(0\) and a locally Lipschitz continuous loss \(L\) the monotone sequences \((\lambda_n) \subset (0, 1]\) and \((\sigma_n) \subset [1, \infty)\) satisfy \(\lim_{n \to \infty} \lambda_n = 0, \sup_{n \geq 1} e^{-\sigma_n} |L|_{\lambda_n^{-1/2}, 1} < \infty, \)

\[
\sup_{n \geq 1} \frac{\lambda_n \sigma_n^d}{|L|_{\lambda_n^{-1/2}, 1}} < \infty, \quad \text{and} \quad \lim_{n \to \infty} \frac{|L|_{\lambda_{n-1}^{-1/2}, 1}^{3/2} \sigma_n^2}{n \lambda_n^4} \sum_{i=0}^{n-1} \gamma_i = 0.
\]

**Assumption S2:** For a fixed strictly positive sequence \((\gamma_i)_{i \geq 0}\) converging to \(0\) and a locally Lipschitz continuous loss \(L\) the sequences \((\lambda_n) \subset (0, 1]\) and \((\sigma_n) \subset [1, \infty)\) satisfy \(\lim_{n \to \infty} \lambda_n \sigma_n^d = 0, \)

\[
\lim_{n \to \infty} \frac{\lambda_n \sigma_n^d}{|L|_{\lambda_n^{-1/2}, 1}} = \infty, \quad \text{and} \quad \lim_{n \to \infty} \frac{\sigma_n^{2+12d}}{n \lambda_n} \sum_{i=0}^{n-1} \gamma_i = 0.
\]

**Assumption S3:** For a fixed strictly positive sequence \((\gamma_i)_{i \geq 0}\) converging to \(0\) and a locally Lipschitz continuous loss \(L\) the monotone sequences \((\lambda_n) \subset
(0, 1] and (\sigma_n) \subset [1, \infty) satisfy \lim_{n \to \infty} \lambda_n = 0, \lim_{n \to \infty} e^{-\sigma_n} |L|_{\lambda_n^{-1/2}, 1} = \infty, \\
\sup_{n \geq 1} \frac{\lambda_n \sigma_n^{4d}}{|L|_{\lambda_n^{-1/2}, 1}} < \infty, \quad \text{and} \\
\lim_{n \to \infty} \frac{|L|_{\lambda_n^{-1/2}, 1}^6 e^{-2\sigma_n}}{n \lambda_n^4} \sum_{i=0}^{n-1} \gamma_i = 0.

**Remark 3.3** In order to illustrate the assumptions \textbf{S1}, \textbf{S2}, and \textbf{S3} let us assume for simplicity that \(L\) is Lipschitz continuous; the case that \(L\) is the least squares loss will be considered in Remark 4.4. Moreover, since for Lipschitz continuous losses \textbf{S3} cannot be satisfied, it suffices to consider \textbf{S1} and \textbf{S2}.

Let us first assume \(\sum_{i \geq 0} \gamma_i < \infty\) as well as \(\lambda_n := n^{-\alpha}\) and \(\sigma_n := n^\beta\) for \(n \geq 1\) and constants \(\alpha > 0\) and \(\beta \geq 0\). Then Assumption \textbf{S1} is met if \(\alpha \geq 4d\beta\) and \(4\alpha + 2\beta < 1\), whereas Assumption \textbf{S2} is met if \(d\beta < \alpha < 4d\beta\) and \(\alpha + (2 + 12d)\beta < 1\). In particular, for \(\beta = 0\) Assumption \textbf{S1} is met if \(0 < \alpha < 1/4\), whereas Assumption \textbf{S2} cannot be met in this case.

Finally, we consider a milder assumption on the decay of correlations, namely \(n^{-1} \sum_{i \geq 0} \gamma_i \leq c (1 + \ln n)^{-1}\), for a constant \(c > 0\) and all \(n \geq 1\). Obviously, this is satisfied if we assume that \((\gamma_i)\) has some arbitrary polynomial decay. Let us consider the sequences \(\lambda_n := (1 + \ln n)^{-\alpha}\) and \(\sigma_n := (1 + \ln n)^\beta\) for \(n \geq 1\) and \(\alpha > 0\) and \(\beta \geq 0\). Then Assumption \textbf{S1} is met if \(\alpha \geq 4d\beta\) and \(4\alpha + 2\beta < 1\), whereas Assumption \textbf{S2} is met if \(d\beta < \alpha < 4d\beta\) and \(\alpha + (2 + 12d)\beta < 1\). In particular, for \(\beta = 0\) Assumption \textbf{S1} is met if \(0 < \alpha < 1/4\), while Assumption \textbf{S2} cannot be met.

The illustrations above show that both Assumptions \textbf{S1} and \textbf{S2} consist of two contrary conditions, namely one which implies that \(\lambda_n\) tends to 0 and another one which ensures that this speed is not too fast. Roughly speaking the first condition guarantees that the approximation error tends to zero (see Lemma 6.5), but since this simultaneously means that the statistical error becomes larger, the second condition is needed to ensure that the latter error still tends to zero (see the proof of Theorem 3.1). This trade-off between approximation and statistical error is typical for consistent learning algorithms (see the books [19] and [22] for several such examples).

With the help of these assumptions we can now establish the announced consistency of SVMs:

**Theorem 3.4** Let \(Z = (X_i, Y_i)_{i \geq 0}\) be stochastic process satisfying Assumption \textbf{Z}. We write \(P := \mu(X_0, Y_0)\) and assume that \(Z\) has a decay of correlations of some order \((\gamma_i)\). In addition, let \(L : X \times Y \times \mathbb{R} \to [0, \infty)\) be a loss satisfying Assumption \textbf{L}. Then for all sequences \((\lambda_n) \subset (0, 1]\) and \((\sigma_n) \subset [1, \infty]\) satisfying either \textbf{S1}, \textbf{S2}, or \textbf{S3} and all \(\epsilon \in (0, 1]\) we have

\[
\lim_{n \to \infty} \mu \left( \omega \in \Omega : |R_{L,P}(f_{T_n}(\omega), \lambda_n, \sigma_n) - R_{L,P}^*| > \epsilon \right) = 0,
\]

where \(T_n(\omega) := ((X_0(\omega), Y_0(\omega)), \ldots, (X_{n-1}(\omega), Y_{n-1}(\omega)))\) and \(f_{T_n}(\omega), \lambda_n, \sigma_n\) is the SVM forecaster defined by (5).
Theorem 3.4 in particular applies to stochastic processes which are \( \alpha \)-mixing with rate \((\gamma_i)\). However, the Assumptions S1, S2, and S3 ensuring consistency are substantially stronger than the ones obtained in [39] for such processes. On the other hand, there are interesting stochastic processes which are not \( \alpha \)-mixing but still enjoy a reasonably fast decay of correlations. Since we are mainly interested in the forecasting problem we will delay the discussion of such examples to the next section.

4. Consistency of SVMs for the Forecasting Problem. In this section we present our main result which establishes consistency of SVM for the forecasting problem (1) – (4) if the dynamical system enjoys a certain decay of correlations. In addition, we discuss some examples of such systems.

We begin by first revisiting our informal problem description given in the introduction. To this end let \( M \subset \mathbb{R}^d \) be a compact set and \( F : M \to M \) be map such that the dynamical system \( \mathcal{D} := (F^i)_{i \geq 0} \) has a unique ergodic measure \( \mu \). Moreover, let \( \mathcal{E} = (\varepsilon_i)_{i \geq 0} \) be a \( \mathbb{R}^d \)-valued stochastic process which is (stochastically) independent of \( \mathcal{D} \). Then the process that generates the noisy observations \( (\tilde{x}_0, \ldots, \tilde{x}_n) \) generated by this process is of the form (4) for a conjoint initial state. Now recall that given an observation of the system at some arbitrary time our goal is to forecast the next observable state. Consequently, we will use the training set

\[
T_n(x, \varepsilon) := ((\tilde{x}_0, \tilde{x}_1), \ldots, (\tilde{x}_{n-1}, \tilde{x}_n))
\]

\[
= ((x + \varepsilon_0, F(x) + \varepsilon_1), \ldots, (F^{n-1}(x) + \varepsilon_{n-1}, F^n(x) + \varepsilon_n))
\]

whose input/output pairs are consecutive observable states. Now note that a single sample \( (F^{i-1}(x) + \varepsilon_{i-1}, F^i(x) + \varepsilon_i) \) depends on the pair \( (\varepsilon_i, \varepsilon_{i+1}) \) and thus we have to consider the process of such pairs. The following assumption summarizes the needed requirements of the process \( \mathcal{N} := ((\varepsilon_i, \varepsilon_{i+1}))_{i \geq 0} \).

**Assumption N:** For the \( \mathbb{R}^{2d} \)-valued stochastic process \( \mathcal{N} \) there exist a constant \( B > 0 \) and a probability measure \( \nu \) on \([-B, B]^{d\mathbb{N}_0}\) such that the coordinate process \( \mathcal{E} := (\pi_0 \circ S^i)_{i \geq 0} \) is stationary with respect to \( \nu \) and satisfies \( \mathcal{N} = (\pi_0 \circ S^i, \pi_0 \circ S^{i+1})_{i \geq 0} \), where \( S \) denotes the shift operator \((x_i)_{i \geq 0} \mapsto (x_{i+1})_{i \geq 0}\) and \( \pi_0 \) denotes the projection \((x_i)_{i \geq 0} \mapsto x_0\).

Before we state our main result we note that the input variable \( x + \varepsilon \) and the output variable \( F(x) + \varepsilon' \) are \( d \)-dimensional vectors. Consequently, our notion of a loss introduced in Section 2 needs a refinement which captures the ideas of the introduction. To this end we state the following assumption.
Assumption LD: For the function $L : \mathbb{R}^d \to [0, \infty)$ there exists a distance-based loss satisfying Assumption L such that its representing function $\psi : \mathbb{R}^d \to [0, \infty)$ has a unique global minimum at 0 and satisfies

$$L(r_1, \ldots, r_d) = \psi(r_1) + \cdots + \psi(r_d), \quad (r_1, \ldots, r_d) \in \mathbb{R}^d.$$  

Obviously, if $L$ satisfies Assumption LD then $L$ is a loss in the sense of the introduction. Moreover note that the specific form (9) makes it possible to consider the coordinates of the output variable separately. Consequently, we will use the forecaster

$$\tilde{f}_{T, \lambda, \sigma} := (f_{T(1), \lambda}, \sigma, \ldots, f_{T(d), \lambda, \sigma}),$$

where $f_{T(j), \lambda, \sigma}$ is the SVM solution obtained by considering the distance-based loss defined by $\psi$ and $T(j) := ((\tilde{x}_0, \pi_j(\tilde{x}_1)), \ldots, (\tilde{x}_{n-1}, \pi_j(\tilde{x}_n)))$ which is obtained by projecting the output variable of $T$ onto its $j$th-coordinate via the coordinate projection $\pi_j : \mathbb{R}^d \to \mathbb{R}$. In other words we build the forecaster $\tilde{f}_{T, \lambda, \sigma}$ by training $d$ different SVMs on the training sets $T^{(1)}, \ldots, T^{(d)}$.

With the help of these preparations we can now present our main result which establishes consistency for such a forecaster.

**Theorem 4.1** Let $M \subset \mathbb{R}^d$ be a compact set, $F : M \to M$ be a Lipschitz continuous map such that the dynamical system $\mathcal{D} := (F^i)_{i \geq 0}$ has a unique ergodic measure $\mu$, and $\mathcal{N}$ be a stochastic process satisfying Assumption N. Assume that both processes $\mathcal{D}$ and $\mathcal{N}$ have a decay of correlations of the order $(\gamma_i)$. Moreover, let $L : \mathbb{R}^d \to [0, \infty)$ be a function satisfying Assumption LD. Then for all sequences $(\lambda_n) \subset (0, 1]$ and $(\sigma_n) \subset [1, \infty)$ satisfying either S1, S2, or S3 and all $\epsilon \in (0, 1]$ we have

$$\lim_{n \to \infty} \mu \otimes \nu \left( (x, \epsilon) \in M \times [-B, B]^d \cap \mathcal{D}^n : |R_{L,P}(\tilde{f}_{T_n(x, \epsilon), \lambda_n, \sigma_n}) - R^*_L,P | > \epsilon \right) = 0,$$

where $T_n(x, \epsilon)$ is defined by (5) and the risks are given by (2) and (3).

Note that if $\mathcal{E}$ is an i.i.d. process then $\mathcal{N}$ has a decay of correlations of any order. Moreover, if $\mathcal{E}$ is $\alpha$-mixing with mixing rate $(\gamma_i)$ then $\mathcal{N}$ has a decay of correlations of order $(\gamma_i)$. Finally, if $\mathcal{D}$ has a decay of correlations $(\gamma_i')$ and $\mathcal{N}$ has a decay of correlations $(\gamma_i''')$ then they obviously have both a decay of correlations $(\gamma_i)$, where $\gamma_i := \max\{\gamma_i', \gamma_i''\}$. In particular, noise processes having slowly decaying correlations will slow down learning even though the system $\mathcal{D}$ may have a fast decay of correlations.

Let us now discuss some examples of classes of dynamical systems enjoying at least a polynomial decay of correlations. Since the existing literature...
on such systems is vast these examples are only meant to be illustrations for situations where Theorem 4.1 can be applied, and are not intended to provide an overview of known results. However, compilations of known results can be found in the survey articles [3, 28] and the book [2].

Example 4.2 (Smooth expanding dynamics) Let $M$ be a compact connected Riemannian manifold and $F : M \to M$ be $C^{1+\varepsilon}$ for some $\varepsilon > 0$. Furthermore assume that there exists constants $c > 0$ and $\lambda > 1$ such that
\[
\max\{\|DF^nx(v)\| : x \in M, v \in T_xM \text{ with } \|v\| = 1\} \geq c\lambda^n
\]
for all $n \geq 0$, where $T_xM$ denotes the tangent space of $M$ at $x$ and $DF^n_x$ denotes the derivative of $F^n$ at $x$. Then it is a classical result that $F$ possesses a unique ergodic measure which is absolutely continuous with respect the Riemannian volume. Moreover, it is well-known (see, e.g., [33] and the references mentioned in [28, Theorem 5]) that there exists a $\tau > 0$ such that the dynamical system has decay of correlations of the order $(e^{-\tau t})$. Generalizations of this result to piecewise smooth and piecewise (non)-uniformly expanding dynamics are discussed in [3]. Finally, [28, Theorem 10] recalls results (together with references) for non-uniformly expanding dynamics having either exponential or polynomial decay of correlations.

Example 4.3 (Smooth hyperbolic dynamics) If $F$ is a topologically mixing $C^{1+\varepsilon}$ Anosov or Axiom A diffeomorphism then it is well-known (see, e.g., [9, 34]) that there exists a $\tau > 0$ such that the dynamical system has decay of correlations of the order $(e^{-\tau t})$. Moreover, [3] lists various extensions of this result to, e.g., smooth non-uniformly hyperbolic systems and hyperbolic systems with singularities.

Besides these classical results and their extensions, [3] also compiles a list of “parabolic” or “intermittent” systems having a polynomial decay.

Let us now consider the forecasting problem for the least squares loss. To this end we first observe that the function $L(r) := \|r\|^2_2$, $r \in \mathbb{R}^d$, satisfies Assumption LD since the least squares loss satisfies Assumption L as we have discussed in Example 2.3 Let us now additionally assume that the noise is pairwise independent, i.e. $\varepsilon_i$ and $\varepsilon_{i'}$ are independent if $i \neq i'$, and centered, i.e. it satisfies $E_{\varepsilon \sim \nu} \pi_0(\varepsilon) = 0$. For a forecaster $f = (f_1, \ldots, f_d) : \mathbb{R}^d \to \mathbb{R}^d$ we then obtain
\[
R_{L,P}(f) = \int \int \sum_{j=1}^d (\pi_j(F(x) + \varepsilon_1) - f_j(x + \varepsilon_0))^2 \nu(d\varepsilon)\mu(dx)
\]
\[
= \int \int \sum_{j=1}^d (\pi_j(F(x)) - f_j(x + \varepsilon_0))^2 \nu(d\varepsilon)\mu(dx) + \int \|\varepsilon_0\|_2^2 \nu(d\varepsilon)
\]
\[
=: R_{L,P}(f) + \int \|\varepsilon_0\|_2^2 \nu(d\varepsilon),
\]

imsart-aos ver. 2007/04/13 file: master.tex date: April 17, 2009
where \( \pi_j : \mathbb{R}^d \to \mathbb{R} \) denotes the \( j \)th-coordinate projection. Consequently, a forecaster \( f \) which approximately minimizes the \( L \)-risk is also an approximate forecaster of the true next state in the sense of \( \mathbb{R}_{L,P}(\cdot) \). Before we combine this observation with Theorem 4.1 let us first rephrase Assumptions S1, S2, and S3 for the least squares loss.

**Assumption S1-LS:** For a strictly positive sequence \( (\gamma_i)_{i \geq 0} \) converging to 0 the monotone sequences \( (\lambda_n) \subset (0, 1) \) and \( (\sigma_n) \subset [1, \infty) \) satisfy \( \lim_{n \to \infty} \lambda_n = 0 \), \( \sup_{n \geq 1} e^{-\sigma_n} \lambda_n^{-1/2} < \infty \),

\[
\sup_{n \geq 1} \lambda_n \sigma_n^{8d/3} < \infty, \quad \text{and} \quad \lim_{n \to \infty} \frac{\sigma_n^2}{n \lambda_n^{11/2}} \sum_{i=0}^{n-1} \gamma_i = 0.
\]

**Assumption S2-LS:** For a strictly positive sequence \( (\gamma_i)_{i \geq 0} \) converging to 0 the sequences \( (\lambda_n) \subset (0, 1) \) and \( (\sigma_n) \subset [1, \infty) \) satisfy \( \lim_{n \to \infty} \lambda_n = 0 \), \( \lim_{n \to \infty} e^{-\sigma_n} \lambda_n^{-1/2} = \infty \),

\[
\lim_{n \to \infty} \lambda_n \sigma_n^{8d/3} = \infty, \quad \text{and} \quad \lim_{n \to \infty} \frac{\sigma_n^{2+12d}}{n \lambda_n} \sum_{i=0}^{n-1} \gamma_i = 0.
\]

**Assumption S3-LS:** For a strictly positive sequence \( (\gamma_i)_{i \geq 0} \) converging to 0 the monotone sequences \( (\lambda_n) \subset (0, 1) \) and \( (\sigma_n) \subset [1, \infty) \) satisfy \( \lim_{n \to \infty} \lambda_n = 0 \), \( \lim_{n \to \infty} e^{-\sigma_n} \lambda_n^{-1/2} = \infty \),

\[
\sup_{n \geq 1} \lambda_n \sigma_n^{8d/3} < \infty, \quad \text{and} \quad \lim_{n \to \infty} \frac{e^{-\sigma_n}}{n \lambda_n^2} \sum_{i=0}^{n-1} \gamma_i = 0.
\]

**Remark 4.4** In order to illustrate the assumptions S1-LS, S2-LS, and S3-LS let us first assume \( \sum_{i \geq 0} \gamma_i < \infty \) as well as \( \lambda_n := n^{-\alpha} \) and \( \sigma_n := n^\beta \) for \( n \geq 1 \) and constants \( \alpha > 0 \) and \( \beta \geq 0 \). Then Assumption S1-LS is met if \( 3\alpha \geq 8d\beta > 0 \) and \( 11\alpha + 4\beta < 2 \), whereas Assumption S2-LS is met if \( \alpha + (2 + 12d)\beta < 1 \) and \( d\beta < \alpha < \frac{3}{2}d\beta \). Finally Assumption S3-LS is satisfied if \( \beta = 0 \) and \( 0 < \alpha < 1/7 \).

Let us now consider the milder assumption \( n^{-1} \sum_{i=0}^{n-1} \gamma_i \leq c(1+\ln n)^{-1} \) which has already been considered in Remark 5.3 for Lipschitz continuous losses. To this end we again consider the sequences \( \lambda_n := (1 + \ln n)^{-\alpha} \) and \( \sigma_n := (1 + \ln n)^\beta \) for \( n \geq 1 \) and constants \( \alpha > 0 \) and \( \beta \geq 0 \). Then Assumption S1-LS is met if \( 3\alpha \geq 8d\beta > 0 \) and \( 11\alpha + 4\beta < 2 \), whereas Assumption S2-LS is met if \( \alpha + (2 + 12d)\beta < 1 \) and \( d\beta < \alpha < \frac{3}{2}d\beta \). Finally Assumption S3-LS is satisfied if \( \beta = 0 \) and \( 0 < \alpha < 1/7 \).

With these preparations we can now state a result showing that SVMs using a least squares loss can be used to forecast the next true state of the dynamical system if the observational noise is sufficiently benign.
Corollary 4.5 Let $M \subset \mathbb{R}^d$ be a compact set and $F : M \rightarrow M$ be a Lipschitz continuous map such that the dynamical system $D := (F^t)_{t \geq 0}$ has a unique ergodic measure $\mu$. Moreover, let $\mathcal{E} = (\varepsilon_i)_{i \geq 0}$ be an i.i.d. process of $[-B, B]^d$-valued and centered random variables. Assume that $D$ has a decay of correlations of the order $(\gamma_i)$. Moreover, let $L : \mathbb{R}^d \rightarrow [0, \infty)$ be defined by $L(r) := \|r\|_2^2$, $r \in \mathbb{R}^d$. Then for all sequences $(\lambda_n) \subset (0, 1]$ and $(\sigma_n) \subset [1, \infty)$ satisfying either $S_1$-LS, $S_2$-LS, or $S_3$-LS and all $\epsilon \in (0, 1]$ we have
\[
\lim_{n \rightarrow \infty} \mu \otimes \nu((x, \varepsilon) \in M \times [-B, B]^d : |\overline{R}_{L,P}(f_{T_n(x,\varepsilon)},\lambda_n,\sigma_n) - \overline{R}_{L,P}^*| > \epsilon) = 0,
\]
where $\overline{R}_{L,P}^* := \inf\{\overline{R}_{L,P}(f) : f : \mathbb{R}^d \rightarrow \mathbb{R}^d \text{ measurable}\}$.

It is interesting to note that the above corollary does not require the noise to be symmetric. Instead it only requires centered noise, i.e. the observations are not systematically biased in a certain direction.

Let us end this section with the following remark that rephrases Theorem 4.14 and its corollary for situations with summable decays of correlations.

Remark 4.6 (Universal consistency) If the sequence $(\gamma_i)$ bounding the correlation is summable, i.e., $\sum \gamma_i < \infty$ then the Assumptions $S_1$, $S_2$, $S_3$, $S_1$-LS, $S_2$-LS, and $S_3$-LS are independent of both the dynamical system and the observational noise process. Consequently, using sequences satisfying one of these assumptions yields an SVM which is consistent for all such pairs of dynamical systems and observational noise processes. In other words, such an SVM can learn the optimal forecaster without knowing specifics of the dynamical systems and the observational noise. To be a bit more specific, let us assume, e.g., that we use the least squares loss and sequences $\lambda_n := n^{-\alpha}$ and $\sigma_n := n^{\beta}$, $n \geq 1$, for fixed $\alpha$ and $\beta$ satisfying $3\alpha \geq 8d\beta > 0$ and $11\alpha + 4\beta < 2$. Then the corresponding SVM is consistent for all bounded observational noise processes having a summable $\alpha$-mixing rate and all ergodic dynamical systems on $M$ which are defined by a Lipschitz continuous $F : M \rightarrow M$ and have a summable decay of correlations. Note that this class of dynamical systems include, but is not limited to, smooth uniformly expanding or hyperbolic dynamics. Finally, if the noise process is also i.i.d. and centered then this SVM actually learns to forecast the next true state.

It is interesting to note that a similar consistency result holds for all noise processes having a polynomial decay of $\alpha$-mixing coefficients and all ergodic dynamical systems on $M$ which are defined by a Lipschitz continuous $F : M \rightarrow M$ and have a polynomial decay of correlations. Indeed, for such combinations SVMs using sequences $\lambda_n := (1 + \ln n)^{-\alpha}$ and $\sigma_n := (1 + \ln n)^{\beta}$ with, e.g., fixed $\alpha$ and $\beta$ satisfying $3\alpha \geq 8d\beta > 0$ and $11\alpha + 4\beta < 2$ are consistent.

5. Discussion. The goal of this work was to show that in principle support vector machines can learn how to predict one-step-ahead noisy observation of a dynamical system without knowing specifics of the dynamical
system or the observational noise besides a certain, rather general stochasticity. However, there remain several open questions which can be subject to further research:

**More General Losses and Kernels.** In the statistical part of our analysis, we used an approach which is based on a “stability” argument. However, it is also possible to use a “skeleton” argument based on covering numbers, instead. Utilizing the latter it seems possible to relax the assumptions on the loss $L$ by making stronger assumptions on both $(\lambda_n)$ and $(\sigma_n)$. A particular loss which is interesting in this direction would be the $\epsilon$-insensitive loss used in classical SVMs for regression. Another possible extension of our work is considering different kernels, such as the kernels that generate Sobolev spaces. In fact, we only focused on Gaussian RBF kernels since these kernels are the most commonly used in practice.

**Learning Rates.** So far we have only shown that the risk of the SVM solution converges to the smallest possible risk, however, for practical considerations the speed of this convergence is of great importance, too. The proof we utilized already gives such learning rates if a quantitative version of the Approximation Lemma 6.5 is available, which is possible if, e.g., quantitative assumptions on the smoothness of $F$ and the regularity of $\nu$ are made. However, since we conjecture that the statistical part of our analysis is not sharp we did not present a corresponding result. In this regard we note that recently [14] established a concentration result for piecewise regular expanding and topologically mixing maps of the interval $[0, 1]$ which is substantially stronger than our elementary Chebyshev inequality of Lemma 6.9. We believe that such a concentration result can be used to substantially sharpen the statistical part of our analysis.

**Perturbed Dynamics.** Another extension of the current work is to consider systems that are perturbed by some noise. Our general consistency result in Theorem 3.4 suggests that such an extension is possible whenever the perturbed system has a decay of correlations. In this regard we note that for some perturbed systems of expanding maps the decay of correlations has already been bounded in [5], and it would be interesting to investigate whether they can be used to prove consistency of SVMs.

**Longer Past.** So far we only used the present observation to forecast the next observation, but it is not hard to see that in almost any system/noise combination the minimal risk $\mathcal{R}^*_L$ reduces if one allows to use additional past observations. On the other hand it appears that the learning problem becomes harder in this case since we have to approximate a function which lives on a higher dimensional input space, and hence there seems to be a trade-off for finite sample sizes. While investigating this trade-off in more
detail seems to be possible with the techniques developed in this work we again assume that the statistical part of our analysis is not sharp enough to obtain a meaningful picture of this trade-off.

Acknowledgment.
The authors gratefully thank V. Baladi for pointing us to the unpublished note [13] of P. Collet.

6. Proof of Theorem 3.4. The goal of this section is to prove Theorem 3.4. Since the proof requires several preliminary results we divided this section into subsections, which provide these prerequisites.

6.1. Some Basics on the Decay of Correlations. The main goal of this section is to establish some uniform bounds on the correlation sequence. Let us begin with the following lemma which establishes some basic properties.

Lemma 6.1 Let \((\Omega, \mathcal{A}, \mu)\) be a probability space, \((Z, \mathcal{B})\) be a measurable space, \(Z\) be a \(Z\)-valued, identically distributed process on \(\Omega\) and \(P := \mu_{Z_0}\). Then the sequence of correlations defines a bilinear operator \(\text{cor}_Z: L_2(P) \times L_2(P) \to \ell_{\infty}\) called the correlation operator.

Proof: The bi-linearity is obvious and the property \(\text{cor}_Z(\psi, \varphi) \in \ell_{\infty}\) follows from \(\mathbb{E}_\mu \psi(Z_0) \cdot \varphi(Z_n) \leq \|\psi \circ Z_0\|_{L_2(\mu)} \|\varphi \circ Z_n\|_{L_2(\mu)} = \|\psi\|_{L_2(P)} \|\varphi\|_{L_2(P)}\).

The following key theorem which goes back to an unpublished note [13] of P. Collet (see also p. 101 in [4]) can be used to establish continuity of the correlation operator. Before we present this result let us first recall that a Banach space \(E\) is said to be continuously embedded into the Banach space \(F\) if \(E \subset F\) and the natural inclusion map \(\text{id}: E \to F\) is continuous.

Theorem 6.2 Let \((\Omega, \mathcal{A}, \mu)\) be a probability space, \((Z, \mathcal{B})\) be a measurable space, \(Z\) be a \(Z\)-valued, identically distributed process on \(\Omega\) and \(P := \mu_{Z_0}\). Moreover, let \(E_1\) and \(E_2\) be Banach spaces that are continuously distributed process on \(\Omega\) and \(P \subseteq \mu_{Z_0}\). If for all \(\psi \in E_1\) and all \(\varphi \in E_2\) the correlation operator satisfies

\[
\text{cor}_Z(\psi, \varphi) \in F
\]

then there exists a constant \(c \in [0, \infty)\) such that

\[
\|\text{cor}_Z(\psi, \varphi)\|_F \leq c \cdot \|\psi\|_{E_1} \|\varphi\|_{E_2}, \quad \psi \in E_1, \ \varphi \in E_2.
\]

For the sake of completeness the proof of this key result can be found in the Appendix. The most obvious examples of Banach spaces \(F\) in the above
the spaces $\ell_p$. However, in the literature on dynamical systems results on the sequence of correlations are usually stated in the form

$$|\text{cor}_{Z,n}(\psi,\varphi)| \leq \kappa_{\psi,\varphi} \gamma_n, \quad n \geq 0,$$

where $(\gamma_n)$ is a strictly positive sequence converging to 0 and $\kappa_{\psi,\varphi}$ is a constant depending on $\psi$ and $\varphi$. To apply Theorem 6.2 in this situation we obviously need Banach spaces which capture such a behaviour of $\text{cor}_{Z}(\cdot, \cdot)$. Therefore, let us fix a strictly positive sequence $\gamma := (\gamma_n)_{n \geq 0}$ such that $\lim_{n \to \infty} \gamma_n = 0$. For a sequence $b := (b_n) \subset \mathbb{R}$ we define

$$\|b\|_{\Lambda(\gamma)} := \sup_{n \geq 0} \frac{|b_n|}{\gamma_n}.$$

Moreover, we write $\Lambda(\gamma) := \{(b_n) \subset \mathbb{R} : \|b_n\|_{\Lambda(\gamma)} < \infty\}$.

The following lemma establishes some basic properties of $(\Lambda(\gamma), \| \cdot \|_{\Lambda(\gamma)})$.

**Lemma 6.3** The pair $(\Lambda(\gamma), \| \cdot \|_{\Lambda(\gamma)})$ is a Banach space and $\|\text{id} : \Lambda(\gamma) \to \ell_\infty\| \leq \|\gamma\|_\infty$.

**Proof:** The fact that $(\Lambda(\gamma), \| \cdot \|_{\Lambda(\gamma)})$ is a normed space is elementary to prove. Moreover, we have

$$\|b\|_{\Lambda(\gamma)} = \sup_{n \geq 0} \frac{|b_n|}{\gamma_n} \geq \sup_{n \geq 0} \frac{|b_n|}{\|\gamma\|_\infty} = \frac{\|b\|_\infty}{\|\gamma\|_\infty},$$

and hence we find $\|\text{id} : \Lambda(\gamma) \to \ell_\infty\| \leq \|\gamma\|_\infty$. Finally, let $(b^{(i)})_{i \geq 1}$ be a Cauchy sequence in $\Lambda(\gamma)$. The previous step shows that it is also a Cauchy sequence in $\ell_\infty$, and by the completeness of $\ell_\infty$ there consequently exists a sequence $b := (b_n) \in \ell_\infty$ such that $\lim_{n \to \infty} \|b^{(i)} - b\|_\infty = 0$. Let us now fix an $\varepsilon > 0$. Then there exists an index $i_0 \geq 0$ such that for all $i, j \geq i_0$ we have $\|b^{(i)} - b^{(j)}\|_{\Lambda(\gamma)} \leq \varepsilon$. Consequently, for fixed $N \geq 0$ we have

$$\sup_{n=0,...,N} \frac{|b^{(i)}_n - b^{(j)}_n|}{\gamma_n} \leq \|b^{(i)} - b^{(j)}\|_{\Lambda(\gamma)} \leq \varepsilon,$$

and by taking the limit $j \to \infty$ we conclude

$$\sup_{n=0,...,N} \frac{|b^{(i)}_n - b_n|}{\gamma_n} \leq \varepsilon.$$

However, $N$ was arbitrary and hence we find $\|b^{(i)} - b\|_{\Lambda(\gamma)} \leq \varepsilon$ for all $i \geq i_0$. In other words we have shown that $(b^{(i)})_{i \geq 1}$ converges to $b$ in $\| \cdot \|_{\Lambda(\gamma)}$. \qed
Combing the above lemma with Theorem 6.2 we immediately obtain:

Corollary 6.4 Let \((\Omega, \mathcal{A}, \mu)\) be a probability space, \((Z, \mathcal{B})\) be a measurable space, \(Z\) be a \(Z\)-valued, identically distributed process on \(\Omega\) and \(P := \mu_{Z_0}\). Moreover, let \(E_1\) and \(E_2\) be Banach spaces that are continuously embedded into \(L_2(P)\). In addition, let \(\gamma := (\gamma_n)_{n \geq 0}\) be a strictly positive sequence such that \(\lim_{n \to \infty} \gamma_n = 0\). If for all \(\psi \in E_1\) and all \(\varphi \in E_2\) there exists a constant \(\kappa_{\psi, \varphi} \in [0, \infty)\) such that

\[
|\text{cor}_{Z, n}(\psi, \varphi)| \leq \kappa_{\psi, \varphi} \gamma_n
\]

for all \(n \geq 0\), then there exists a constant \(c \in [0, \infty)\) such that

\[
|\text{cor}_{Z, n}(\psi, \varphi)| \leq c \|\psi\|_{E_1} \cdot \|\varphi\|_{E_2} \cdot \gamma_n, \quad \psi \in E_1, \varphi \in E_2, n \geq 0.
\]

6.2. Some Properties of Gaussian RBF Kernels. In this subsection we establish some properties of Gaussian RBF kernels which will be heavily used in the proof of Theorem 3.4. Let us begin with an approximation result.

Lemma 6.5 Let \(X \subset \mathbb{R}^d\) and \(Y \subset \mathbb{R}\) be compact subsets, \(L : X \times Y \times \mathbb{R} \to [0, \infty)\) be a convex locally Lipschitz continuous loss and \(P\) be a probability measure on \(X \times Y\) such that \(\mathcal{R}_{L, P}(0) < \infty\). Then for all sequences \((\lambda_n) \subset (0, 1)\) and \((\sigma_n) \subset [1, \infty)\) satisfying

\[
\lim_{n \to \infty} \lambda_n \sigma_n^d = 0
\]

we have

\[
\lim_{n \to \infty} \left( \inf_{f \in H_{\sigma_n}(X)} \lambda_n \|f\|^2_{H_{\sigma_n}(X)} + \mathcal{R}_{L, P}(f) \right) = \mathcal{R}_{L, P}^*.
\]

Proof: For \(\sigma > 0\) we write \(\mathcal{R}_{L, P, H_{\sigma}(X)}^* := \inf\{\mathcal{R}_{L, P}(f) : f \in H_{\sigma}(X)\}\). Since \(L\) is locally Lipschitz continuous and \(\mathcal{R}_{L, P}(0) < \infty\) the discussion after (4) in [38] shows that it is a \(P\)-integrable Nemitski loss in the sense of [38]. Now recall (see [37]) that \(H_{\sigma}(X)\) is universal, i.e. it is dense in \(C(X)\), and hence [38] Corollary 1] shows \(\mathcal{R}_{L, P, H_{\sigma}(X)}^* = \mathcal{R}_{L, P}^*\) for all \(\sigma > 0\). Let us now fix an \(\varepsilon > 0\). The above discussion then shows that there exists an \(f_{\varepsilon} \in H_1(X)\) such that \(\mathcal{R}_{L, P}(f_{\varepsilon}) \leq \mathcal{R}_{L, P}^* + \varepsilon\). Furthermore, by (11) there exists an \(n_0 \geq 0\) such that

\[
\lambda_n \sigma_n^d \leq \varepsilon \left\|f_{\varepsilon}\right\|_{H_1(X)}^{-2}, \quad n \geq n_0.
\]

Since \(\sigma_n \geq 1\) we also know \(f_{\varepsilon} \in H_{\sigma_n}(X)\) and \(\|f_{\varepsilon}\|_{H_{\sigma_n}(X)}^2 \leq \sigma_n \|f_{\varepsilon}\|_{H_1(X)}^2\) by [40] Corollary 6], and therefore we obtain

\[
\inf_{f \in H_{\sigma_n}(X)} \lambda_n \|f\|^2_{H_{\sigma_n}(X)} + \mathcal{R}_{L, P}(f) \leq \lambda_n \|f_{\varepsilon}\|^2_{H_{\sigma_n}(X)} + \mathcal{R}_{L, P}(f_{\varepsilon}) \leq \mathcal{R}_{L, P}^* + 2\varepsilon
\]

for all \(n \geq n_0\). From this we easily deduce the assertion. ■
Before we establish the next result let us recall that a function $f : X \to \mathbb{R}$ on a subset $X \subset \mathbb{R}^d$ is called Lipschitz continuous if there exist a constant $c \in [0, \infty)$ such that $|f(x) - f(x')| \leq c\|x - x'\|_2$ for all $x, x' \in X$. In the following the smallest such constant is denoted by $|f|_1$ and the set of all Lipschitz continuous functions is denoted by Lip($X$). Moreover, recall that if $X$ is compact then Lip($X$) together with the norm $\|f\|_{\text{Lip}(X)} := \max\{\|f\|_{\infty}, |f|_1\}$ forms a Banach space. Moreover, in this case Lip($X$) is also closed under multiplication. Indeed, for $f,g \in \text{Lip}(X)$ and $x, x' \in X$ we have

$$\|f(x)g(x) - f(x')g(x')\| \leq \|f\|_{\infty} \cdot |g|_1 \|x - x'| + |f|_1 \cdot \|g\|_{\infty} \|x - x'|$$

and hence we obtain $fg \in \text{Lip}(X)$ with $|fg|_1 \leq \|f\|_{\infty} \cdot |g|_1 + |f|_1 \cdot \|g\|_{\infty}$. Our next result shows that every function in $H_\sigma(X)$ is Lipschitz continuous.

**Lemma 6.6** Let $X \subset \mathbb{R}^d$ be a non-empty set and $\sigma > 0$. Then every $f \in H_\sigma(X)$ is Lipschitz continuous with $|f|_1 \leq \sqrt{2}\sigma \|f\|_{H_\sigma(X)}$.

**Proof:** Let us write $\Phi : X \to H_\sigma(X)$ for the canonical feature map defined by $\Phi(x) := k_\sigma(x,.)$. Now recall that $\Phi$ satisfies the reproducing property

$$f(x) = \langle \Phi(x), f \rangle, \quad x \in X, f \in H_\sigma(X),$$

and hence in particular $k_\sigma(x', x) = \langle \Phi(x), \Phi(x') \rangle$ for all $x, x' \in X$. Using these equalities together with $1 - e^{-t} \leq t$ for $t \geq 0$ we obtain

$$|f(x) - f(x')| = |\langle \Phi(x) - \Phi(x'), f \rangle|$$

$$\leq \|f\|_{H_\sigma(X)} \cdot \|\Phi(x) - \Phi(x')\|_{H_\sigma(X)}$$

$$= \|f\|_{H_\sigma(X)} \sqrt{\langle \Phi(x), \Phi(x) \rangle + \langle \Phi(x'), \Phi(x') \rangle - 2 \langle \Phi(x), \Phi(x') \rangle}$$

$$= \|f\|_{H_\sigma(X)} \sqrt{2 - 2 \exp(-\sigma^2 \|x - x'\|_2^2)}$$

$$\leq \sqrt{2}\sigma \|f\|_{H_\sigma(X)} \|x - x'\|_2,$$

i.e., we have proved the assertion. \qed

In the following we consider certain orthonormal bases (ONBs) of $H_\sigma(X)$. To this end let us first recall that in [40] Theorem 5 it was shown that $(e_n)_{n \geq 0}$, where $e_n : \mathbb{R} \to \mathbb{R}$ is defined by

$$e_n(x) := \sqrt{\frac{2^n \sigma^{2n}}{n!}} x^n e^{-\sigma^2 x^2}, \quad x \in \mathbb{R},$$

forms an ONB of $H_\sigma(\mathbb{R})$. Moreover, it was shown that if $X \subset \mathbb{R}$ has a non-empty interior the restrictions of $e_n$ to $X$ form an ONB of $H_\sigma(X)$. The following lemma establishes upper bounds on $\|e_n\|_\infty$ if $X$ is a closed interval.
Lemma 6.7  Let $\sigma > 0$ and $a > 0$ be fixed real numbers and $(e_n)_{n \geq 0}$ be the ONB of $H_\sigma([-a, a])$, where $e_n$ is defined by the restriction of (12) to $[-a, a]$. Then we have $\|e_n\|_\infty \leq (2\pi n)^{-1/4}$ for all $n \geq 1$ and

\begin{equation}
\|e_n\|_\infty \leq \sqrt{\frac{2^n a^{2n} \sigma^{2n}}{n!}} e^{-a^2 \sigma^2}
\end{equation}

for all $n \geq 2a^2 \sigma^2$. In addition, for $n \geq 8a^2 \sigma^2$ we have

\begin{equation}
\left(\sum_{i=n+1}^{\infty} \|e_i\|_\infty^2\right)^{1/2} \leq \left(\frac{2}{\pi(n+1)}\right)^{1/4} 2^{-n/2} e^{-a^2 \sigma^2},
\end{equation}

and for $a \sigma \geq 1$ we also have

\begin{equation}
\left(\sum_{i=0}^{\infty} \|e_i\|_\infty^2\right)^{1/2} \leq \sqrt{6} a \sigma.
\end{equation}

Proof: Elementary calculus shows

$$e_n'(x) = \sqrt{\frac{2^n a^{2n} \sigma^{2n}}{n!}} x^{n-1} e^{-\sigma^2 x^2} (n - 2\sigma^2 x^2)$$

for all $n \geq 1$ and $x \in \mathbb{R}$. From this we conclude $e_n'(x^*) = 0$ if and only if $x^* = \pm \sqrt{\frac{n}{2\sigma^2}}$. Now $n \geq 2a^2 \sigma^2$ implies $|x^*| < a$ and then it is not hard to see that $|e_n|$ attains its maximum at these $x^*$. Consequently, we obtain

$$\|e_n\|_\infty \leq \sqrt{\frac{n^n}{n!}} e^{-n/2} \leq \frac{n^n}{\sqrt{2\pi n} n^n} e^{-n/2} = (2\pi n)^{-1/4}$$

by Stirling’s formula. On the other hand, $n \geq 2a^2 \sigma^2$ implies $|x^*| \geq a$ and in this case it is not hard to see that $|e_n|$ attains its maximum at $\pm a$. From these considerations we conclude (13). For the first part of the lemma it thus remains to show that $\|e_n\|_\infty \leq (2\pi n)^{-1/4}$ for all $n \geq 2a^2 \sigma^2$. To this end we apply Stirling’s formula to the already obtained (13). This yields

$$\|e_n\|_\infty \leq \sqrt{\frac{2^n a^{2n} \sigma^{2n} e^n}{\sqrt{2\pi n} n^n}} e^{-a^2 \sigma^2} = (2\pi n)^{-1/4} \left(\frac{2a^2 \sigma^2 e^{1-2a^2 \sigma^2/n}}{n} \right)^{n/2} \leq (2\pi n)^{-1/4},$$

where in the last step we used that $te^{1-t} \leq 1$ for all $t \geq 0$.

For the proof of (14) we recall that the remainder of the Taylor series of the exponential function satisfies

$$\sum_{i=n+1}^{\infty} \frac{y^i}{i!} \leq 2 \frac{|y|^{n+1}}{(n+1)!}.$$
for $|y| \leq 1 + n/2$. Since $n \geq 8a^2\sigma^2$ implies $2a^2\sigma^2 \leq 1 + n/2$ we consequently obtain

$$\sum_{i=n+1}^{\infty} \|e_i\|_\infty^2 \leq \sum_{i=n+1}^{\infty} \frac{2^2a^2\sigma^2 i^2}{i!} e^{-2a^2\sigma^2} \leq \frac{2^{n+1}a^2(n+1)\sigma^2(n+1)}{(n+1)!} e^{-2a^2\sigma^2}$$

$$\leq \frac{2^{n+1}a^2(n+1)\sigma^2(n+1)\epsilon(n+1)}{\sqrt{2\pi(n+1)(n+1)^{(n+1)}}} e^{-2a^2\sigma^2}$$

$$\leq \left(\frac{2}{\pi(n+1)}\right)^{1/2} 4^{-n(n+1)} e^{-2a^2\sigma^2}.$$

From this we easily deduce (14). Finally, for the proof of (15) we observe

$$\sum_{i=0}^{\infty} \|e_i\|_\infty^2 \leq 1 + (2\pi)^{-1/2} + \sum_{i=2}^{\infty} (2\pi i)^{-1/2}$$

$$\leq 1 + (2\pi)^{-1/2} + (2\pi)^{-1/2} \int_1^{\infty} x^{-1/2} dx$$

$$\leq 1 + (2\pi)^{-1/2} + (e/\pi)^{-1/2} 4a\sigma$$

$$\leq 3/2 + 4a\sigma.$$

Combining this estimate with (14) we then obtain

$$\sum_{i=0}^{\infty} \|e_i\|_\infty^2 = \sum_{i=0}^{\infty} \|e_i\|_\infty^2 + \sum_{i=[8a^2\sigma^2]+1}^{\infty} \|e_i\|_\infty^2$$

$$\leq 3/2 + 4a\sigma + \left(\frac{2}{\pi([8a^2\sigma^2]+1)^{1/2}}\right) e^{-2a^2\sigma^2}$$

$$\leq 3/2 + 4a\sigma + \left(\frac{1}{8e\pi a^2\sigma^2}\right)^{1/2} 4^{-8a^2\sigma^2} e^{-2a^2\sigma^2}$$

$$\leq 2 + 4a\sigma,$$

and from the latter we easily obtain (15).

Our next goal is to generalize the above result to the multi-dimensional case. To this end recall that the tensor product $f \otimes g : X \times X \to \mathbb{R}$ of two functions $f, g : X \to \mathbb{R}$ is defined by $f \otimes g(x, x') := f(x)g(x')$, $x, x' \in X$. Obviously, for bounded functions we have $\|f \otimes g\|_\infty = \|f\|_\infty \|g\|_\infty$.

For a multi-index $\eta = (n_1, \ldots, n_d) \in \mathbb{N}_0^d$ we use the notation $\eta \geq n$ if $n \in \mathbb{N}_0$ and $n_i \geq n$ for all $i = 1, \ldots, d$. Moreover, we write

(16) $e_\eta := e_{n_1} \otimes \cdots \otimes e_{n_d}$, \quad $\eta = (n_1, \ldots, n_d) \in \mathbb{N}_0^d$.
where \( e_{n_i} \) is defined by (12). Then [40, Theorem 5] shows that \((e_{\eta})_{\eta \in N^d_0}\) is an ONB of \(H_\sigma(\mathbb{R}^d)\) and the restrictions of the members of this ONB to \([-a,a]^d\) form an ONB of \(H_\sigma([-a,a]^d)\). The following lemma generalizes the estimates of Lemma 6.7 to this multi-dimensional ONB.

**Corollary 6.8** For \( \sigma > 0 \) and \( a > 0 \) satisfying \( a\sigma \geq 1 \), and \( d \in \mathbb{N} \) let \((e_{\eta})_{\eta \in N^d_0}\) be the restriction of the ONB (16) to \([-a,a]^d\). Then for \( n \geq 8e^2a^2\sigma^2 \) we have

\[
\left( \sum_{\eta \in N^d_0} \|e_\eta\|_\infty^2 \right)^{1/2} \leq \sqrt{d} e^{-a^2\sigma^2} (6a\sigma)^{(d-1)/2} \left( \frac{2}{\pi(n+1)} \right)^{1/4} 2^{-(n+1)}.
\]

**Proof:** Using \( \|e_{i_1} \otimes \cdots \otimes e_{i_d}\|_\infty = \|e_{i_1}\|_\infty \cdots \|e_{i_d}\|_\infty \) we obtain

\[
\sum_{\eta \in N^d_0} \|e_\eta\|_\infty^2 \leq d \sum_{i_1=n+1}^{\infty} \sum_{i_2=0}^{\infty} \cdots \sum_{i_d=0}^{\infty} \prod_{j=1}^{d} \|e_{i_j}\|_\infty^2
\]

\[
= d \left( \sum_{i=n+1}^{\infty} \|e_i\|_\infty^2 \right)^{d-1} \left( \sum_{i=0}^{\infty} \|e_i\|_\infty^2 \right)^{d-1}
\]

\[
\leq d \left( \frac{2}{\pi(n+1)} \right)^{1/2} 2^{-2(n+1)} e^{-2a^2\sigma^2} (6a\sigma)^{d-1}
\]

by Lemma 6.7. From this we immediately obtain the assertion. \( \blacksquare \)

### 6.3. Some concentration inequalities in RKHSs.

In this subsection we will establish a concentration inequality for RKHS-valued functions and for processes which have a certain decay of correlations. This concentration result will then be the key ingredient in the statistical analysis of the proof of Theorem 3.4.

Let us begin by showing a simple concentration result in the form of Chebyshev’s inequality.

**Lemma 6.9** Let \( Z = (Z_i)_{i \geq 0} \) be a \( Z \)-valued process on \((\Omega,A,\mu)\) that is stationary in the wide sense. Then for \( P := \mu_{Z_0}, f \in L_2(P), n \geq 1, \) and \( \delta > 0 \) we have

\[
\mu\left( \omega \in \Omega : \left| \frac{1}{n} \sum_{i=0}^{n-1} f \circ Z_i(\omega) - \mathbb{E}_P f \right| \geq \delta \right) \leq \frac{2}{n\delta^2} \sum_{i=0}^{n-1} \text{cor}_{Z_i}(f,f).
\]
Proof: By Markov’s inequality we immediately obtain

\[ \mu(\omega \in \Omega : \left| \frac{1}{n} \sum_{i=0}^{n-1} f \circ Z_i(\omega) - E_P f \right| \geq \delta) \leq \frac{1}{n^2 \delta^2} E_{\mu} \left( \sum_{i=0}^{n-1} f \circ Z_i - E_P f \right)^2 . \]

Moreover, a simple calculation shows

\[ E_{\mu} \left( \sum_{i=0}^{n-1} f \circ Z_i - E_P f \right)^2 \leq 2n \sum_{i=0}^{n-1} \text{cor}_{Z_i}(f,f). \]

Combining both estimates immediately yields the assertion.

For the following results we have to introduce some more notations: given a bounded measurable kernel \( k : X \times X \rightarrow \mathbb{R} \) with RKHS \( H \) we write \( \Phi : X \rightarrow H, \Phi(x) := k(x, \cdot) \) for the canonical feature map. Moreover, for a bounded measurable function \( h : X \times Y \rightarrow \mathbb{R} \) and a distribution \( P \) on \( X \times Y \) we write \( E_P h \Phi \) for the Bochner integral (see [20]) of the \( H \)-valued function \( h \Phi \). Similarly, given \( T := ((x_1, y_1), \ldots, (x_n, y_n)) \in (X \times Y)^n \) we write

\[ E_T h \Phi := \sum_{i=1}^{n} h \Phi(x_i, y_i) \]

for the empirical counterpart of \( E_P h \Phi \). In order to motivate the following results we further mention that the proof of Theorem 3.4 will heavily rely on the estimate

\[ \| f_{P,\lambda,H} - f_{T,\lambda,H} \|_H \leq \frac{1}{\lambda} \| E_P h_\lambda \Phi - E_T h_\lambda \Phi \|_H, \]

where \( f_{P,\lambda,H} \) is the SVM solution (see Theorem 6.13 for an exact definition) one obtains by replacing the empirical risk \( \mathcal{R}_{L,T}(\cdot) \) with the true risk \( \mathcal{R}_{L,P}(\cdot) \) in [5] and \( h_\lambda \) is a function independent of the training set \( T \). Consequently, our next goal is to estimate terms of the form \( \| E_P h \Phi - E_T h \Phi \|_H \).

To this end we begin with the following lemma which, roughly speaking, will be used to reduce RKHS-valued functions to \( \mathbb{R} \)-valued functions.

Lemma 6.10 Let \( H \) be the separable RKHS of a bounded measurable kernel \( k : X \times X \rightarrow \mathbb{R} \), let \( \Phi : X \rightarrow H \) be the corresponding canonical feature map and \( (e_i)_{i \geq 0} \) be an ONB of \( H \). Moreover, let \( Y \) be another measurable space, \( P \) and \( Q \) be probability measures on \( X \times Y \), and \( h \in L_1(P) \cap L_1(Q) \). Then for all \( n \geq 0 \) we have

\[ \| E_P h \Phi - E_Q h \Phi \|_H \leq \left( \sum_{i=0}^{n} \| E_P h e_i - E_Q h e_i \|_2^2 \right)^{1/2} + \left( \sum_{i=n+1}^{\infty} \| e_i \|_2^2 \right)^{1/2} (\| E_P |h| + E_Q |h|).
\]
Proof: Let us define $S_n : H \rightarrow H$ by $\sum_{i \geq 0} \langle f, e_i \rangle e_i \mapsto \sum_{i=0}^n \langle f, e_i \rangle e_i$. Then we have

$$\|S_n \Phi(x) - \Phi(x)\|_H^2 = \sum_{i=n+1}^{\infty} \|\Phi(x), e_i\|_H^2 = \sum_{i=n+1}^{\infty} |\langle \Phi(x), e_i \rangle|^2 = \sum_{i=n+1}^{\infty} |e_i(x)|^2$$

by the reproducing property and hence we obtain

$$\|E_P h \Phi - E_Q h \Phi\|_H$$

$$\leq \|E_P h \Phi - E_P h S_n \Phi\|_H + \|E_P h S_n \Phi - E_Q h S_n \Phi\|_H + \|E_Q h S_n \Phi - E_Q h \Phi\|_H$$

$$\leq E_P |h| \|\Phi - S_n \Phi\|_H + \|E_P h S_n \Phi - E_Q h S_n \Phi\|_H + E_Q |h| \|\Phi - S_n \Phi\|_H$$

$$\leq \|E_P h S_n \Phi - E_Q h S_n \Phi\|_H + \left( \sum_{i=n+1}^{\infty} \|e_i\|_\infty^2 \right)^{1/2} (E_P |h| + E_Q |h|).$$

Moreover, using the reproducing property we have $\langle E_P h \Phi, e_i \rangle = E_P h e_i$ and $\langle E_Q h \Phi, e_i \rangle = E_Q h e_i$, and thus we conclude

$$\|E_P h S_n \Phi - E_Q h S_n \Phi\|_H^2 = \sum_{i=0}^{n} \|E_P h \Phi - E_Q h \Phi, e_i\|^2 = \sum_{i=0}^{n} |E_P h e_i - E_Q h e_i|^2.$$

Combining this equality with the previous estimate we then obtain the assertion. \hfill \blacksquare

Before we can establish the concentration inequality for RKHS-valued functions we finally need the following simple lemma.

Lemma 6.11 For $d \geq 1$ and $t > 18d \ln(d)$ we have $t^{-1/4} 2^{-t} \leq t^{-2d}$.

Proof: Obviously, it suffices to show

$$t \ln 2 + (1/4 - 2d) \ln t \geq 0.$$  \hspace{1cm} (19)

Let us first prove the case $d = 1$. Then (19) reduces to the assertion $h(t) := t \ln 2 - \frac{7}{4} \ln t \geq 0$. To establish the latter, note that we have $h'(t) = \ln 2 - \frac{7}{4} t^{-1}$ and hence $h'(t^*) = 0$ holds if and only if $t^* = \frac{7}{4 \ln 2}$. Simple considerations then show that $h$ has its only global minimum at $t^*$ and therefore we have $h(t) \geq h(t^*) \geq \frac{7}{4} - \frac{7}{4} \ln \left( \frac{7}{4 \ln 2} \right) > 0$.

Let us now consider the case $d \geq 2$. To this end we fix a $t > 18d \ln(d)$. Then
there exists a unique \( x > 18 \) with \( t = x \ln(d) \), and hence we obtain
\[
\begin{align*}
t \ln 2 + (1/4 - 2d) \ln t &= x \ln(d) \ln 2 + 1/4 \ln(x \ln(d)) - 2d \ln(x \ln(d)) \\
&> x \ln(d) \ln 2 - 2d \ln(x \ln(d)) \\
&= d(x \ln(d) \ln 2 - 2 \ln x - 2 \ln d - 2 \ln(\ln(d)) \\
&> d \left( x \ln(d) \ln 2 - \frac{\ln d}{\ln 2} \ln x - 2 \ln d - 2 \ln d \right) \\
&= d \ln(d) \left( x \ln 2 - \frac{2}{\ln 2} \ln x - 4 \right),
\end{align*}
\]
where in the last estimate we used \( d \geq 2 \). Now it is elementary to check that \( x \mapsto x \ln 2 - \frac{2}{\ln 2} \ln x - 4 \) is increasing on \([2(\ln 2)^{-2}, \infty)\) and since \( 18 \ln 2 - \frac{2}{\ln 2} \ln 18 - 4 > 0 \) we then obtain (19).  

**Theorem 6.12** For \( \sigma > 0 \) and \( a > 0 \) satisfying \( a \sigma \geq 1 \), and \( d \geq 1 \) let \( \Phi : [-a, a]^d \to H_\sigma([-a, a]^d) \) be the canonical feature map of the Gaussian RBF kernel and let \((e_n)_{n \in \mathbb{N}^d}\) be the ONB of \( H_\sigma([-a, a]^d) \) which is considered in Corollary 6.8. In addition, let \( Y \) be a measurable space and let \( Z = (X_i, Y_i)_{i \geq 0} \) be a \([-a, a]^d \times Y\)-valued process on \((\Omega, \mathcal{A}, \mu)\) that is stationary in the wide sense. Furthermore, let \((\gamma_i)_{i \geq 0}\) be a strictly positive null sequence, \( h : [-a, a]^d \times Y \to \mathbb{R} \) be a bounded measurable function, and \( K_h \in [1, \infty) \) be a constant such that

\[(20) \quad \text{cor}_{Z,i}(he_{n}, he_{\eta}) \leq K_h \gamma_i \]

for all \( i \geq 0, \eta \in \mathbb{N}^d \). Then for all \( \epsilon > 0 \) satisfying both \( \epsilon \leq (1 + 8a^2 \sigma^2)^{-2d} \) and \( \epsilon \leq (18d \ln d)^{-2d} \), and all \( n \geq 1 \) we have
\[
\mu \left( \omega \in \Omega : \| \mathbb{E}_p h \Phi - \mathbb{E}_{T_n(\omega)} h \Phi \|_H \leq \epsilon \right) \geq 1 - \frac{2(1 + \frac{1}{8a^2 \sigma^2})^d K_h C_{\sigma,d,h}^3}{\epsilon n^d} \sum_{i=0}^{n-1} \gamma_i,
\]
where \( \mathbb{E}_{T_n(\omega)} h \Phi \) denotes the empirical Bochner integral (18) with respect to the training set \( T_n(\omega) := (Z_0(\omega), \ldots, Z_{n-1}(\omega)) \), and
\[
C_{\sigma, d,h} := \left( 1 + \frac{1}{8a^2 \sigma^2} \right)^{\frac{d}{2}} + 2 \sqrt{d} e^{-a^2 \sigma^2} (6a \sigma)^{d-1} \|h\|_\infty.
\]

**Proof:** Let us write
\[
\delta := \left( \frac{\epsilon}{C_{\sigma, d,h}} \right)^{5/4}.
\]
Using \( C_{\sigma, d,h} \geq (1 + \frac{1}{8a^2 \sigma^2})^\frac{d}{2} \geq 1 \) and \( \epsilon \leq (1 + 8a^2 \sigma^2)^{-2d} \) we then find
\[
\delta \leq (1 + 8a^2 \sigma^2)^{-5d/2},
\]
and consequently, there exists a natural number $m \geq 8 e a^2 \sigma^2$ such that $(m + 1)^{-5d/2} \leq \delta < m^{-5d/2}$. Moreover, for later use we note that using $C_{a, d, h} \geq 1$ and $\epsilon \leq (18d \ln d)^{-2d}$ yields

$$\delta^{-\frac{d}{2}} \geq 18d \ln d.$$ 

Let us now consider an $\omega \in \Omega$ such that

$$|E_{\omega} h \eta - E_{\omega} h \eta| < \delta$$

for all $\eta \in \{0, \ldots, m\}^d$. By Lemma 6.10 and Corollary 6.8 we then obtain

$$\|E_{\omega} h \Phi - E_{\omega} h \Phi\|_H \leq \left( \sum_{\eta \leq m} |E_{\omega} h \eta - E_{\omega} h \eta|^2 \right)^{\frac{1}{2}} + \left( \sum_{\eta \in \Pi_d \setminus \exists \eta \eta \geq m} \|e_\eta\|_\infty^2 \right)^{\frac{1}{2}} (E_{\omega} |h| + E_{\omega} |h|)$$

$$\leq (m + 1)^{d/2} \delta + 2 \sqrt{d} e^{-a^2 \sigma^2} (6a\sigma)^{(d-1)/2} \left( \frac{2}{\pi (m + 1)} \right)^{1/4} \delta^{-\frac{d}{2}} \|h\|_\infty$$

$$\leq \left( 1 + \frac{1}{8 e a^2 \sigma^2} \right)^{d/2} \delta^{4/5} + 2 \sqrt{d} e^{-a^2 \sigma^2} (6a\sigma)^{(d-1)/2} \delta^{1/5} \|h\|_\infty$$

where in the last step we used the inequalities $8 e a^2 \sigma^2 \leq m < \delta^{-\frac{d}{2}} \leq m + 1$. Using Lemma 6.11 for $t := \delta^{-\frac{d}{2}}$ we consequently obtain

$$\|E_{\omega} h \Phi - E_{\omega} h \Phi\|_H \leq \left( \left( 1 + \frac{1}{8 e a^2 \sigma^2} \right)^{d/2} + 2 \sqrt{d} e^{-a^2 \sigma^2} (6a\sigma)^{(d-1)/2} \|h\|_\infty \right) \delta^{4/5} = \epsilon.$$ 

Moreover, by Lemma 6.9 and a simple union bound argument we see that the probability of $\omega$ satisfying (21) for all $\eta \in \{0, \ldots, m\}^d$ simultaneously is not smaller than

$$1 - \sum_{\eta \in \{0, \ldots, m\}^d} \frac{2}{n \delta^2} \sum_{i=0}^{n-1} \text{cor}_{\Pi, i}(h \eta, h \eta).$$

In addition, we have

$$\sum_{\eta \in \{0, \ldots, m\}^d} \frac{2}{n \delta^2} \sum_{i=0}^{n-1} \text{cor}_{\Pi, i}(h \eta, h \eta) \leq \frac{2(m + 1)^d n^{-1}}{n \delta^2} \sum_{i=0}^{n-1} K_i \gamma_i.$$
and since \( 21 < 8e^2 \sigma^2 < m < \delta^{-\frac{2}{5}} \) we further estimate

\[
\frac{2(m + 1)^d}{n\delta^2} \leq \frac{2(1 + \frac{1}{8e^2 \sigma^2})^d m^d}{n\delta^2} \leq \frac{2(1 + \frac{1}{8e^2 \sigma^2})^d}{n\delta^{12/5}} = \frac{2(1 + \frac{1}{8e^2 \sigma^2})^d (1 + \frac{1}{8e^2 \sigma^2}) \frac{1}{2} + 2\sqrt{d} e^{-a^2 \sigma^2} (6a\sigma)^{d-1} \|h\|_\infty^3}{n\epsilon^3}.
\]

Combining these estimates we then obtain the assertion. ■

6.4. Proof of Theorem 3.4. For the proof of Theorem 3.4 we need some final preparations. Let us begin with the following result on the existence and uniqueness of infinite sample SVMs which is a slight extension of similar results established in [18, 12]:

**Theorem 6.13** Let \( L : X \times Y \times \mathbb{R} \to [0, \infty) \) be a convex, locally Lipschitz continuous loss function satisfying \( L(x, y, 0) \leq 1 \) for all \((x, y) \in (X \times Y)\), and let \( P \) be a distribution on \( X \times Y \). Furthermore, let \( H \) be a RKHS of a bounded measurable kernel over \( X \). Then for all \( \lambda > 0 \) there exists exactly one element \( f_{P, \lambda, H} \in H \) such that

\[
\lambda \|f_{P, \lambda, H}\|_H^2 + R_{L, P}(f_{P, \lambda, H}) = \inf_{f \in H} \lambda \|f\|_H^2 + R_{L, P}(f).
\]

Furthermore, we have \( \|f_{P, \lambda, H}\|_H \leq \lambda^{-1/2} \).

Note that the above theorem in particular yields \( \|f_{T, \lambda, H}\|_H \leq \lambda^{-1/2} \) by considering the empirical measure associated to a training set \( T \in (X \times Y)^n \).

The following result which was (essentially) shown in [18] describes the stability of the empirical SVM solutions.

**Theorem 6.14** Let \( X \) be a separable metric space, \( L : X \times Y \times \mathbb{R} \to [0, \infty) \) be a convex, locally Lipschitz continuous loss function satisfying \( L(x, y, 0) \leq 1 \) for all \((x, y) \in (X \times Y)\), and let \( P \) be a distribution on \( X \times Y \). Furthermore, let \( H \) be the RKHS of a bounded continuous kernel \( k : X \times X \to \mathbb{R} \) and let \( \Phi : X \to H \) be the corresponding canonical feature map. Then for all \( \lambda > 0 \) the function \( h_{\lambda} : X \times Y \to \mathbb{R} \) defined by

\[
h_{\lambda}(x, y) := L'(x, y, f_{P, \lambda}(x)), \quad (x, y) \in X \times Y,
\]

is bounded and satisfies and

\[
\|f_{P, \lambda, H} - f_{T, \lambda, H}\|_H \leq \frac{1}{\lambda} \left\| \mathbb{E}_P h_{\lambda} \Phi - \mathbb{E}_T h_{\lambda} \Phi \right\|_H, \quad T \in (X \times Y)^n.
\]
Proof of Theorem 3.4: Obviously, it suffices to consider sets \( X \) of the form \( X = [-a,a]^d \) for some \( a \geq 1 \). For \( \sigma > 0 \) and \( \lambda > 0 \) we write \( h_{\lambda,\sigma} \) for the function we obtain by Theorem 6.14 for \( H := H_\sigma(X) \). By the local Lipschitz continuity of \( L \), \( \|k_\sigma\|_\infty \leq 1 \), Theorem 6.13, and (24) we then have
\[
|\mathcal{R}_{L,P}(f_{T,\lambda,\sigma}) - \mathcal{R}_{L,P}(f_{P,\lambda,\sigma})| \leq |L|_{\lambda^{-1/2},1} \|f_{T,\lambda,\sigma} - f_{P,\lambda,\sigma}\|_\infty \\
\leq |L|_{\lambda^{-1/2},1} \|f_{T,\lambda,\sigma} - f_{P,\lambda,\sigma}\|_{H_\sigma(X)} \\
\leq \frac{|L|_{\lambda^{-1/2},1}}{\lambda} \|E_P h_{\lambda,\sigma} \Phi - E_T h_{\lambda,\sigma} \Phi\|_{H_\sigma(X)}
\]
for all \( \sigma > 0, \lambda > 0 \) and all \( T \in (X \times Y)^n \). Moreover, using (23), (6), and Lemma 6.6 we have
\[
|h_{\lambda,\sigma}(x,y) - h_{\lambda,\sigma}(x',y')| \\
= |L'(x,y,f_{P,\lambda,\sigma}(x)) - L'(x',y',f_{P,\lambda,\sigma}(x'))| \\
\leq c \cdot \left( |x - x'|^2 + |y - y'|^2 + |f_{P,\lambda,\sigma}(x) - f_{P,\lambda,\sigma}(x')|^2 \right)^{1/2} \\
\leq c \cdot \left( |x - x'|^2 + |y - y'|^2 + 2\sigma^2 \|f_{P,\lambda,\sigma}\|_{H_\sigma(X)}^2 |x - x'|^2 \right)^{1/2} \\
\leq c \cdot \left( |x - x'|^2 + |y - y'|^2 + 2\sigma^2 \lambda^{-1} |x - x'|^2 \right)^{1/2} \\
\leq 2c\sigma \lambda^{-1/2} \|(x,y) - (x',y')\|_2
\]
for all \( \sigma \geq 1, \lambda \in (0,1] \) and all \( (x,y),(x',y') \in X \times Y \). Consequently, we find \( |h_{\lambda,\sigma}|_1 \leq 2c\sigma \lambda^{-1/2} \). Moreover, we have
\[
|h_{\lambda,\sigma}(x,y)| = |L'(x,y,f_{P,\lambda,\sigma}(x))| \leq \sup_{|t| \leq \lambda^{-1/2}} |L'(x,y,t)| \leq |L|_{\lambda^{-1/2},1}
\]
for all \( \lambda > 0 \) and all \( (x,y) \in X \times Y \). Let us now write \( e^{(\sigma)}_\eta \) for the \( \eta \)-th element, \( \eta \in \mathbb{N}_0^d \), of the ONB of \( H_\sigma(X) \) considered in Corollary 6.8. Combining the above estimates with Lemma 6.6 and the trivial bound \( \|e^{(\sigma)}_\eta\|_\infty \leq \|e^{(\sigma)}_\eta\|_{H_\sigma(X)} \leq 1 \) we obtain
\[
|h_{\lambda,\sigma} e^{(\sigma)}_\eta|_1 \leq |h_{\lambda,\sigma}|_1 \|e^{(\sigma)}_\eta\|_\infty + \|h_{\lambda,\sigma}\|_\infty \|e^{(\sigma)}_\eta\|_1 \leq 5c\sigma \lambda^{-1/2}
\]
for all \( \lambda \in (0,1] \) and \( \sigma \geq 1 \), where in the last step we used the estimate \( |L|_{a,1} \leq c(1+a), a > 0 \), which we derived after Assumption L. Since we additionally have \( \|h_{\lambda,\sigma} e^{(\sigma)}_\eta\|_\infty \leq \|h_{\lambda,\sigma}\|_\infty \leq 2c\lambda^{-1/2} \) we find \( \|h_{\lambda,\sigma} e^{(\sigma)}_\eta\|_{\text{Lip}(X \times Y)} \leq 5c\sigma \lambda^{-1/2} \). Moreover, by Corollary 6.3 we may assume without loss of generality that \( \kappa_{\psi,\varphi} \) is of the form
\[
\kappa_{\psi,\varphi} = c \varepsilon \|\psi\|_{\text{Lip}(X \times Y)} \|\varphi\|_{\text{Lip}(X \times Y)}
\]
where \( c_Z \) is a constant only depending on \( Z \) and \((\gamma_i)\). Consequently, we obtain

\[
|\text{cor}_{Z,i}(h_{\lambda,\sigma}e_{\eta}(\sigma), h_{\lambda,\sigma}e_{\eta}(\sigma))| \leq 25c_Z \epsilon^2 \lambda^{-1} \sigma^2
\]

for all \( \sigma \geq 1, \lambda \in (0,1], \text{ and } \eta \in N_d^0 \), i.e., \( (28) \) is satisfied for \( K_{h_{\lambda,\sigma}} := \tilde{c} \lambda^{-1} \sigma^2 \), where \( \sigma \geq 1, \lambda \in (0,1] \), and \( \tilde{c} \) is a constant independent of \( \lambda \) and \( \sigma \). For \( n \geq 1 \) and \( \epsilon > 0 \) satisfying both

\[
\epsilon \leq (1 + 8e^2 \sigma^2)^{-2d}|L|_{\lambda^{-1/2},1} \lambda^{-1}
\]

and \( \epsilon \leq (18d \ln d)^{-2d} \), Theorem \( 6.12 \) together with \((25)\) and \((26)\) thus yields

\[
\begin{align*}
\mu & \left( \omega \in \Omega : |(R_{L,P}(f_{T_n(\omega),\lambda,\sigma}) - R_{L,P}(f_{P,\lambda,\sigma})| > \epsilon \right) \\
& \leq \mu \left( \omega \in \Omega : \|E_Ph\Phi - \mathbb{E}_{T_n(\omega)}h\Phi\|_H > \frac{\lambda \epsilon}{|L|_{\lambda^{-1/2},1}} \right) \\
& \leq \frac{2\tilde{c}(1 + \frac{1}{8e^2\sigma^2})^{d/2} \tilde{C}_{\lambda,\sigma,d,a} |L|_{\lambda^{-1/2},1} \sigma^2}{\epsilon^2 n \lambda^4} \sum_{i=0}^{n-1} \gamma_i ,
\end{align*}
\]

where \( \tilde{C}_{\lambda,\sigma,d,a} := (1 + \frac{1}{8e^2\sigma^2})^{d/2} + 2c\sqrt{d}e^{-c^2} \sigma^2 (6a)^{d+1} \epsilon^2 |L|_{\lambda^{-1/2},1} \). Using the fact that the function \( x \mapsto e^{-x^2} x \) is bounded on \([0, \infty)\) we further obtain

\[
(29) \quad \tilde{C}_{\lambda_n,\sigma_n,d,a} \leq C_d (1 + e^{-a\sigma_n|L|_{\lambda_n^{-1/2},1}}),
\]

where \( C_d \) is a constant only depending on \( d \). Let us now consider the case where Assumption \( S_1 \) is fulfilled. Then we have \( C_{d,a} := \sup_{n \geq 1} \tilde{C}_{\lambda_n,\sigma_n,d,a} < \infty \) and

\[
\epsilon_0 := \inf_{n \geq 1} (1 + 8e^2 \sigma_n^2)^{-2d}|L|_{\lambda_n^{-1/2},1} \lambda_n^{-1} > 0 ,
\]

and hence \((27)\) is satisfied for all \( \epsilon \in (0, \epsilon_0) \). Moreover, by the remark after Assumption \( L \) we have \( |L|_{\lambda^{-1/2},1} \leq c (1 + \lambda^{-1/2}) \) for all \( \lambda > 0 \) and hence the first and third assumption of \( S_1 \) together with \( \sigma_n \leq \sigma_{n+1} \) imply \( \lim_{n \to \infty} \lambda_n \sigma_n^d = 0 \). By Lemma \( 5.5 \) we thus find

\[
\lim_{n \to \infty} R_{L,P}(f_{P,\lambda_n,\sigma_n}) = R_{L,P}^*.
\]

Consequently, \((28)\) shows that for sufficiently large \( n \) and \( \epsilon \in (0, \epsilon_0] \) we have

\[
\begin{align*}
\mu & \left( \omega \in \Omega : |(R_{L,P}(f_{T_n(\omega),\lambda_n,\sigma_n}) - R_{L,P}^*)| > 2\epsilon \right) \\
& \leq \mu \left( \omega \in \Omega : |(R_{L,P}(f_{T_n(\omega),\lambda_n,\sigma_n}) - R_{L,P}(f_{P,\lambda_n,\sigma_n})| > \epsilon \right) \\
& \leq 2^{(d+1)/2} \tilde{c} \epsilon^2 n \lambda^4 \sum_{i=0}^{n-1} \gamma_i ,
\end{align*}
\]
and hence we obtain the assertion by the last condition of S1.
Let us now consider the case where Assumption S2 is fulfilled. Then it is easy to see that the second assumption of S2 implies \( \lim_{n \to \infty} \sigma_n^{-4d} |L|_{\lambda_n^{-1/2},1} = 0 \) which in turn yields \( \sup_{n \geq 1} e^{-\sigma_n} |L|_{\lambda_n^{-1/2},1} < \infty \). From this we conclude \( C_{d,a} := \sup_{n \geq 1} \tilde{C}_{\lambda_n,\sigma_n,d,a} < \infty \) by \( (29) \). Moreover, we have
\[
\epsilon_n := (1 + 8 \sigma_n^2 \sigma_n^4) \frac{2d}{\lambda_n^2 \lambda_n^{-1/2}} \leq 2 \frac{d+7}{2} \frac{\tilde{C}_{d,a} \sigma_n^2 + 12d}{\lambda_n n \lambda_n^4} \sum_{i=0}^{n-1} \gamma_i,
\]
and consequently for a fixed \( \epsilon > 0 \) we have \( \epsilon_n \leq \epsilon \) for all sufficiently large \( n \).
Analogously to the first case we thus find
\[
\mu \left( \omega \in \Omega : \left| R_{L,P}(f_{T_n}(\omega),\lambda_n,\sigma_n) - R_{L,P}^* \right| > 2\epsilon \right)
\leq 2^{(d+1)/2} \tilde{C}_{d,a} \frac{3}{\epsilon^3 n \lambda_n^4} \sum_{i=0}^{n-1} \gamma_i
\leq \tilde{C}_{d,a} \frac{\sigma_n^2 + 12d}{\lambda_n n \lambda_n^4} \sum_{i=0}^{n-1} \gamma_i,
\]
for all sufficiently large \( n \), where \( \tilde{C}_{d,a} \) is a constant only depending on \( d \) and \( a \). From this estimate we obtain the assertion by the last condition of S2.
Finally, let us consider the case where Assumption S3 is satisfied. Using \( (29) \) and \( a \geq 1 \) we then obtain for sufficiently large \( n \) and \( \epsilon \in (0, \epsilon_0] \) that
\[
\mu \left( \omega \in \Omega : \left| R_{L,P}(f_{T_n}(\omega),\lambda_n,\sigma_n) - R_{L,P}^* \right| > 2\epsilon \right)
\leq 2^{(d+7)/2} \tilde{C}_{d,a} \frac{e^{-3\sigma_n} |L|_{\lambda_n^{-1/2},1}^6 \sigma_n^2}{\epsilon^3 n \lambda_n^4} \sum_{i=0}^{n-1} \gamma_i.
\]
Again, from this we easily find the assertion.

7. Proof of Theorem 4.1 For the proof of Theorem 4.1 we need to bound the correlation sequences for stochastic processes which are the sum of a dynamical system and an observational noise process. This is the goal of the following results. We begin with a lemma which computes the correlation of a joint process from the correlations of its components.

**Lemma 7.1** Let \( X = (X_i)_{i \geq 0} \) be an \( X \)-valued, identically distributed stochastic process defined on \( (\Omega, A, \mu) \) and \( Y = (Y_i)_{i \geq 0} \) be a \( Y \)-valued, identically distributed stochastic process defined on \( (\Theta, B, \nu) \). Then the stochastic process \( Z = (Z_i)_{i \geq 0} \) defined on \( (\Omega \times \Theta, A \otimes B, \mu \otimes \nu) \) by \( Z_i := (X_i, Y_i) \) is
identically distributed with $P := (\mu \otimes \nu)_{Z_0} = \mu_{X_0} \otimes \nu_{Y_0}$. Moreover, for $\psi, \varphi \in L_2(P)$ the $i$-th coordinate of the correlation sequence of $Z$ is given by
\[
\text{cor}_{Z,i}(\psi, \varphi) = E_\mu \text{cor}_{X,i}(\psi(\cdot, Y_0), \varphi(\cdot, Y_1)) + E_\mu E_\nu \text{cor}_{Y,i}(\psi(X_0, \cdot), \varphi(X'_0, \cdot)),
\]
where $X'_0$ is an independent copy of $X_0$.

**Proof:** The first assertion regarding $P$ is obvious. For the second assertion we fix an independent copy $X' = (X'_i)_{i \geq 0}$ of $X$. Then an easy calculation using the fact that both $X$ and $Y$ are identically distributed yields
\[
\text{cor}_{Z,i}(\psi, \varphi) = E_\mu E_\nu \psi(X_0, Y_0) \varphi(X_i, Y_i) - E_\mu E_\nu \psi(X_0, Y_0) \cdot E_\mu E_\nu \varphi(X_0, Y_0)
\]
\[
= E_\nu (E_\mu \psi(X_0, Y_0) \varphi(X_i, Y_i) - E_\mu E_\nu \psi(X_0, Y_0) \varphi(X'_0, Y_0)) + E_\mu E_\nu (E_\mu \psi(X_0, Y_0) \cdot E_\nu \varphi(X'_0, Y_0))
\]
\[
= E_\nu \text{cor}_{X,i}(\psi(\cdot, Y_0), \varphi(\cdot, Y_i)) + E_\mu E_\nu \text{cor}_{Y,i}(\psi(X_0, \cdot), \varphi(X'_0, \cdot)),
\]
i.e., we have proved the assertion. \hfill \blacksquare

The following elementary lemma establishes the Lipschitz continuity of a certain type of function which is important when considering the process that generates noisy observations of a dynamical system.

**Lemma 7.2** Let $M \subset \mathbb{R}^d$ be a compact subset and $F : M \to M$ be a Lipschitz continuous map. For $B > 0$ and a fixed $j \in \{1, \ldots, d\}$ we write $X := M + [-B, B]^d$ and $Y := \pi_j(X)$, where $\pi_j : \mathbb{R}^d \to \mathbb{R}$ denotes the $j$-th coordinate projection. For $h \in \text{Lip}(X \times Y)$ and $x \in M$, $\varepsilon_0, \varepsilon_1 \in [-B, B]^d$ we define the function $\tilde{h} : M \times [-B, B]^d \to \mathbb{R}$ by
\[
(30) \quad \tilde{h}(x, \varepsilon_0, \varepsilon_1) := h(x + \varepsilon_0, \pi_j(F(x) + \varepsilon_1)).
\]
Then for all $x \in M$ and $\varepsilon_0, \varepsilon_1 \in [-B, B]^d$ we have
\[
\|\tilde{h}(x, \cdot, \cdot)\|_{\text{Lip}([-B, B]^d)} \leq (1 + \|F\|_{\text{Lip}(M)}) \|h\|_{\text{Lip}(X \times Y)}
\]
\[
\|\tilde{h}(\cdot, \varepsilon_0, \varepsilon_1)\|_{\text{Lip}(M)} \leq \|h\|_{\text{Lip}(X \times Y)}.\]
Moreover, we define the process $Y_{\theta} \in \mathbb{B}, B$ valued stochastic process on $\mathbb{B}, B$. Analogously, we have

$$h(x + \varepsilon_0, \pi_j(F(x) + \varepsilon_1)) - h(x + \varepsilon_0', \pi_j(F(x) + \varepsilon_1'))$$

$$\leq \|h\|_{\text{Lip}(X \times Y)} (\|x - x'\|_2^2 + |\pi_j(F(x) + \varepsilon_1) - \pi_j(F(x') + \varepsilon_1')|^2)^{1/2}$$

$$\leq \|h\|_{\text{Lip}(X \times Y)} (\|x - x'\|_2^2 + \|\varepsilon_1 - \varepsilon_1'\|_2^2)^{1/2}$$

$$= \|h\|_{\text{Lip}(X \times Y)} \|(\varepsilon_1, \varepsilon_1) - (\varepsilon_1', \varepsilon_1')\|_2.$$

From these estimates we easily obtain the assertions.

**Proof:** For the first assertion we observe that for $(\varepsilon_0, \varepsilon_1), (\varepsilon_0', \varepsilon_1') \in [-B, B]^d \times [-B, B]^d$ we obviously have

$$h(x + \varepsilon_0, \pi_j(F(x) + \varepsilon_1)) - h(x + \varepsilon_0', \pi_j(F(x) + \varepsilon_1'))$$

$$\leq \|h\|_{\text{Lip}(X \times Y)} (\|x - x'\|_2^2 + |\pi_j(F(x) + \varepsilon_1) - \pi_j(F(x') + \varepsilon_1')|^2)^{1/2}$$

$$\leq \|h\|_{\text{Lip}(X \times Y)} (\|x - x'\|_2^2 + \|\varepsilon_1 - \varepsilon_1'\|_2^2)^{1/2}$$

$$= \|h\|_{\text{Lip}(X \times Y)} \|(\varepsilon_1, \varepsilon_1) - (\varepsilon_1', \varepsilon_1')\|_2.$$

The next theorem bounds the correlation for functions defined by $\psi, \varphi \in \text{Lip}(M)$.\]

**Theorem 7.3** Let $M \subset \mathbb{R}^d$ be a compact subset and $F : M \to M$ be a Lipschitz continuous map such that the dynamical system $X := (F^i)_{i \geq 0}$ has an ergodic measure $\mu$. Moreover, let $\gamma = (\gamma_i)_{i \geq 0}$ be a strictly positive null sequence such that

$$(31) \quad \text{cor}_X(\psi, \varphi) \in \Lambda(\gamma), \quad \psi, \varphi \in \text{Lip}(M).$$

Furthermore, let $E = (E_i)_{i \geq 0}$ be an in the wide sense stationary $[-B, B]^d$-valued stochastic process on $(\Theta, \mathcal{B}, \nu)$ such that the $[-B, B]^d$-valued process $Y = (Y_i)_{i \geq 0}$ on $(\Theta, \mathcal{B}, \nu)$ that is defined by $Y_i(\vartheta) = (\varepsilon_i(\vartheta), \varepsilon_i+1(\vartheta))$, $i \geq 0$, $\vartheta \in \Theta$, satisfies

$$(32) \quad \text{cor}_Y(\psi, \varphi) \in \Lambda(\gamma), \quad \psi, \varphi \in \text{Lip}([-B, B]^{2d}).$$

For a fixed $j \in \{1, \ldots, d\}$ we write $X := M + [-B, B]^d$ and $Y := \pi_j(X)$. Moreover, we define the process $Z = (Z_i)_{i \geq 0}$ on $(\Omega \times \Theta, \mathcal{A} \otimes \mathcal{B}, \mu \otimes \nu)$ by $Z_i = (F_i, \varepsilon_i, \varepsilon_{i+1})$, $i \geq 0$. Then for all $\psi, \varphi \in \text{Lip}(X \times Y)$ we have

$$\text{cor}_Z(\bar{\psi}, \bar{\varphi}) \in \Lambda(\gamma),$$

where $\bar{\psi}$ and $\bar{\varphi}$ are defined by $30$.\]

[imsart-aos ver. 2007/04/13 file: master.tex date: April 17, 2009]
Proof: Let $c_X$ and $c_Y$ be the constants we obtain by applying (31) and (32) to Corollary 6.4. Moreover, since $E$ is stationary in the wide sense we observe that $\mathcal{Y}$ is identically distributed. Applying Lemma 7.1 to the processes $\mathcal{X}$ and $\mathcal{Y}$ then yields

$$|\text{cor}_{x,i}(\psi, \phi)|$$

\[ \leq \left| \mathbb{E}_{\nu} \text{cor}_{x,i}(\psi(\cdot, Y_0), \phi(\cdot, Y_1)) \right| 
+ \left| \mathbb{E}_{x,y} \| \psi(F_0(x), \cdot) \| \mathcal{L}^\nu \cdot \gamma_i \right| 
+ c_Y \left| \mathbb{E}_{x,y} \| \psi(F_0(x), \cdot) \| \mathcal{L}^\nu \cdot \gamma_i \right| 
\]

where in the last step we used Lemma 7.2.

Note that for using the estimate of Theorem 7.3 in Lemma 6.9 we need that the process $\mathcal{Y}$ is stationary in the wide sense. Obviously, the latter is satisfied if the process $\mathcal{E}$ is stationary.

Proof of Theorem 4.1: For a fixed $j \in \{1, \ldots, d\}$ we write $X := M + [-B, B]^d$ and $Y := \pi_j(X)$. Moreover, we define the $X \times Y$-valued process $Z = (X_i, Y_i)_{i \geq 0}$ on $(M \times [-B, B]^d, \mathcal{L}^\nu)$ by $X_i := F^i + \pi_0 \circ S^i$ and $Y_i := \pi_j(F^i + \pi_0 \circ S^i)$, and in addition, we write $P^{(j)} := (\mathcal{L}^\nu)_j := (\pi_0 \circ S^i)^j$. Let us further consider the $M \times [-B, B]^{2d}$-valued stationary process $Z := (F^i, \pi_0 \circ S^i, \pi_0 \circ S^i)$ which is defined on $(M \times [-B, B]^{2d}, \mathcal{L}^\nu)$. For $\psi, \phi \in \text{Lip}(X \times Y)$ Theorem 7.3 together with our decay of correlations assumptions then shows $|\text{cor}_{x,i}(\psi, \phi)| \leq \kappa_{\psi, \phi} \gamma_i$ for all $i \geq 0$, where $\kappa_{\psi, \phi} \in [0, \infty)$ is a constant independent of $i$. Moreover, our construction ensures $\text{cor}_{x,i}(\psi, \phi) = \text{cor}_{x,i}(\psi, \phi)$ for all $i \geq 0$ and hence Theorem 3.2 yields

$$\mu \otimes \nu((x, \varepsilon) \in M \times [-B, B]^{2d}) = |R_{L,P^{(j)}}(f_{T_n^{(j)}(x, \varepsilon)}, \lambda_{n,s}), -R_{L,P^{(j)}}^*| > \varepsilon \rightarrow 0$$

for $n \rightarrow \infty$ and all $\varepsilon > 0$. Using Assumption LD and the definition (10) we then easily obtain the assertion.

Appendix: Proof of Theorem 6.2: In the following $B_E$ denotes the closed unit ball of a Banach space $E$. Recall that a linear operator $S : E \rightarrow F$ acting between two Banach spaces $E$ and $F$ is continuous if and only if it is bounded, i.e., $\|S\| := \sup_{x \in B_E} \|Sx\| < \infty$. Our first goal is to recall another equivalent condition which in practice is often easier to check. To this end we need the following definition.
**Definition 7.4** Let $E$ and $F$ be Banach spaces and $S : E \to F$ be a linear map. Then $S$ is said to have a closed graph if for all $x \in E$, $y \in F$, and all sequences $(x_n) \subset E$ satisfying $x_n \to x$ and $Sx_n \to y$ we have $Sx = y$.

Obviously, every continuous linear operator has a closed graph. The following fundamental result from functional analysis known as the closed graph theorem shows the converse implication.

**Theorem 7.5** Let $E$ and $F$ be Banach spaces and $S : E \to F$ be a linear map that has a closed graph. Then $S$ is continuous.

Our next goal is to establish an analogous result for bilinear maps. To this end we first recall the principle of uniform boundedness which is also known as Banach-Steinhaus Theorem.

**Theorem 7.6** Let $E$ and $F$ be Banach spaces, $A$ be a non-empty set, and $S_\alpha : E \to F$, $\alpha \in A$, be bounded linear operators. If the family $(S_\alpha)_{\alpha \in A}$ satisfies

$$\sup_{\alpha \in A} \|S_\alpha x\| < \infty$$

for all $x \in E$ then we actually have $\sup_{\alpha \in A} \sup_{x \in B_E} \|S_\alpha x\| < \infty$.

Let us now recall that a map $S : E_1 \times E_2 \to F$ between Banach spaces $E_1$, $E_2$, and $F$ is called bilinear if the maps $S(x_1, \cdot) : E_2 \to F$ and $S(\cdot, x_2) : E_1 \to F$ are linear for all $x_1 \in E_1$ and $x_2 \in E_2$. In order to state a closed graph theorem for bilinear maps we also need a notion which describes a closed graph property for bilinear maps:

**Definition 7.7** Let $E_1$, $E_2$, and $F$ be Banach spaces and $S : E_1 \times E_2 \to F$ be a bilinear map. Then $S$ is said to have a partially closed graph if the linear maps $S(x_1, \cdot) : E_2 \to F$ and $S(\cdot, x_2) : E_1 \to F$ have closed graphs for all $x_1 \in E_1$ and $x_2 \in E_2$.

With these preparations we can now state and prove the announced closed graph theorem for bilinear maps:

**Theorem 7.8** Let $E_1$, $E_2$, and $F$ be Banach spaces and $S : E_1 \times E_2 \to F$ be a bilinear map that has a partially closed graph. Then there exists a constant $c \in [0, \infty)$ such that

$$\|S(x_1, x_2)\|_F \leq c \|x_1\|_{E_1} \cdot \|x_2\|_{E_2}, \quad x_1 \in E_1, x_2 \in E_2.$$
Using the bi-linearity of $S$ for all $x_1 \in E_1$ and $x_2 \in E_2$. In particular, the boundedness of the operators $S(\cdot, x_2) : E_1 \rightarrow F$ yields

$$\sup_{x_1 \in B_{E_1}} \|S(x_1, x_2)\| < \infty$$

for all $x_2 \in E_2$. Applying the Principle of Uniform Boundedness to the family of bounded operators $(S(x_1, \cdot))_{x_1 \in B_{E_1}}$ thus shows

$$c := \sup_{x_1 \in B_{E_1}, x_2 \in B_{E_2}} \|S(x_1, x_2)\| < \infty .$$

Using the bi-linearity of $S$ we then obtain the assertion.

Proof: By the Closed Graph Theorem the maps $S(x_1, \cdot) : E_2 \rightarrow F$ and $S(\cdot, x_2) : E_1 \rightarrow F$ are bounded linear operators for all $x_1 \in E_1$ and $x_2 \in E_2$. In particular, the boundedness of the operators $S(\cdot, x_2) : E_1 \rightarrow F$ yields

$$\sup_{x_1 \in B_{E_1}} \|S(x_1, x_2)\| < \infty$$

Proof of Theorem 6.2: Obviously, $\text{cor}_Z : E_1 \times E_2 \rightarrow F$ is a well-defined bilinear operator. In view of Theorem 7.8 it suffices to show that this operator has a partially closed graph. We begin by showing that $\text{cor}_Z(\psi, \cdot) : E_2 \rightarrow F$ has a closed graph for all $\psi \in E_1$. To this end let us fix some $\psi \in E_1, \varphi \in E_2$, a sequence $b := (b_n)_{n \geq 0} \in F$ and a sequence $(\varphi_i)_{i \geq 1} \subset E_2$ such that $\lim_{i \rightarrow \infty} \|\varphi_i - \varphi\|_{E_2} = 0$ and

$$(33) \quad \lim_{i \rightarrow \infty} \|\text{cor}_Z(\psi, \varphi_i) - b\|_F = 0 .$$

Obviously, $\text{cor}_Z(\psi, \cdot) : E_2 \rightarrow F$ has a closed graph if $\text{cor}_Z(\psi, \varphi) = b$. To show this equality we first observe that for fixed $n \geq 0$ and $i \rightarrow \infty$ we have

$$\left| \int_Z \varphi_i \, dP - \int_Z \varphi \, dP \right| \leq \|\varphi - \varphi_i\|_{L_1(P)} \leq \|\text{id} : E_2 \rightarrow L_2(P)\| \cdot \|\varphi - \varphi_i\|_{E_2}$$

and

$$\left| \int_\Omega \psi(Z_0) \cdot \varphi(Z_n) \, d\mu - \int_\Omega \psi(Z_0) \cdot \varphi_i(Z_n) \, d\mu \right| \leq \|\psi\|_{L_2(P)} \cdot \|\varphi - \varphi_i\|_{L_2(P)}$$

and

$$\leq \|\psi\|_{L_2(P)} \cdot \|\text{id} : E_2 \rightarrow L_2(P)\| \cdot \|\varphi - \varphi_i\|_{E_2} .$$

From this we conclude $\lim_{i \rightarrow \infty} \text{cor}_{Z,n}(\psi, \varphi_i) = \text{cor}_{Z,n}(\psi, \varphi)$ for the $n$-th coordinate of sequences of correlations. Moreover, $F$ is continuously included in $\ell_\infty$ and hence (33) implies $\lim_{i \rightarrow \infty} \text{cor}_{Z,n}(\psi, \varphi_i) = b_n$ for all $n \geq 0$. Combining these considerations yields $\text{cor}_{Z,n}(\psi, \varphi) = b_n$ for all $n \geq 0$, i.e. we have shown that $\text{cor}_Z(\psi, \cdot) : E_2 \rightarrow F$ has a closed graph. Since the fact that all $\text{cor}_Z(\cdot, \varphi) : E_1 \rightarrow F$ have a closed graph can be shown completely analogous the proof is completed.
REFERENCES

[1] N. Aronszajn. Theory of reproducing kernels. *Trans. Amer. Math. Soc.*, 68:337–404, 1950.

[2] V. Baladi. *Positive Transfer Operators and Decay of Correlations*. World Scientific, Singapore, 2000.

[3] V. Baladi. Decay of correlations. In *1999 AMS Summer Institute on Smooth Ergodic Theory and Applications*, pages 297–325. AMS, 2001.

[4] V. Baladi, M. Benedicks, and V. Maume-Deschamps. Almost sure rates of mixing for i.i.d. unimodal maps. *Ann. E.N.S.*, 35:77–126, 2002.

[5] V. Baladi, A. Kondah, and B Schmitt. Random correlations for small perturbations of expanding maps. *Random Comput. Dynam.*, 4:179–204, 1996.

[6] P.L. Bartlett, M.I. Jordan, and J.D. McAuliffe. Convexity, classification, and risk bounds. *J. Amer. Statist. Assoc.*, 101:138–156, 2006.

[7] B.E. Boser, I.Guyon, and V. Vapnik. A training algorithm for optimal margin classifiers. In *Computational Learning Theory*, pages 144–152, 1992.

[8] D. Bosq. *Nonparametric Statistics for Stochastic Processes*. Springer, New York, 2nd edition, 1998.

[9] R. Bowen. *Equilibrium States and the Ergodic Theory of Anosov Diffeomorphisms*. Springer, Berlin, 1975.

[10] R.C. Bradley. Basic properties of strong mixing conditions. A survey and some open questions. *Probability Surveys*, 2:107–144, 2005.

[11] R.C. Bradley. *Introduction to Strong Mixing Conditions*, volume 1-3. Technical Report, Department of Mathematics, Indiana University, Bloomington, Custom Publishing of I.U., Bloomington, 2005.

[12] A. Christmann and I. Steinwart. Consistency and robustness of kernel based regression. *Bernoulli*, to appear, 2007. [http://www.c3.lanl.gov/ml/pubs/2005_regression/paper.pdf](http://www.c3.lanl.gov/ml/pubs/2005_regression/paper.pdf).

[13] P. Collet. A remark about uniform de-correlation prefactors. Technical report, 1999.

[14] P. Collet, S. Martinez, and B. Schmitt. Exponential inequalities for dynamical measures of expanding maps of the interval. *Probab. Theory Related Fields*, 123:301–322, 2002.

[15] C. Cortes and V. Vapnik. Support vector networks. *Machine Learning*, 20:273–297, 1995.

[16] N. Cristianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines*. Cambridge University Press, 2000.

[17] M. Davies. Noise reduction schemes for chaotic time series. *Phys. D*, 79:174–192, 1994.

[18] E. DeVito, L. Rosasco, A. Caponnetto, M. Piana, and A. Verri. Some properties of regularized kernel methods. *J. Mach. Learn. Res.*, 5:1363–1390, 2004.

[19] L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer, New York, 1996.

[20] J. Diestel and J.J. Uhl. *Vector Measures*. American Mathematical Society, Providence, 1977.

[21] J. Fan and Q. Yao. *Nonlinear Time Series*. Springer, New York, 2003.

[22] L. Györfi, M. Kohler, A. Krzyżak, and H. Walk. *A Distribution-free Theory of Nonparametric Regression*. Springer, New York, 2002.
E.J. Kostelich and T. Schreiber. Noise reduction in chaotic time-series data: A survey of common methods. Phys. Rev. E, 48:1752–1763, 1993.

E.J. Kostelich and J.A. Yorke. Noise reduction: finding the simplest dynamical system consistent with the data. Phys. D, 41:183–196, 1990.

S.P. Lalley. Beneath the noise, chaos. Ann. Statist., 27:461–479, 1999.

S.P. Lalley. Removing the noise from chaos plus noise. In Nonlinear Dynamics and Statistics, pages 233–244. Birkhäuser, 2001.

S.P. Lalley and A.B. Nobel. Denoising deterministic time series. Dynamics of partial differential equations, 3:259–279, 2006.

S. Luzzatto. Stochastic-like behaviour in nonuniformly expanding maps. In B. Hasselblatt and A. Katok, editors, Handbook of Dynamical Systems 1B, pages 265–326. Elsevier, Amsterdam, 2006.

R. Meir. Nonparametric time series prediction through adaptive model selection. Machine learning, 39:5–34, 2000.

D.S. Modha and E. Masry. Memory-universal prediction of stationary random processes. IEEE Trans. Inform. Theory, 44:117–133, 1998.

A.B. Nobel. Limits to classification and regression estimation from ergodic processes. Ann. Statist., 27:262–273, 1999.

A.B. Nobel. Consistent estimation of a dynamical map. In Nonlinear dynamics and statistics, pages 267–280. Birkhäuser, Boston, 2001.

D. Ruelle. The thermodynamic formalism for expanding maps. Comm. Math. Phys., 125:239–262, 1989.

D. Ruelle. Thermodynamic Formalism. Cambridge University Press, 2nd edition, 2004.

T. Sauer. A noise reduction method for signals from nonlinear systems. Phys. D, 58:193–201, 1992.

B. Schölkopf and A.J. Smola. Learning with Kernels. MIT Press, 2002.

I. Steinwart. On the influence of the kernel on the consistency of support vector machines. J. Mach. Learn. Res., 2:67–93, 2001.

I. Steinwart, D. Hush, and C. Scovel. Function classes that approximate the Bayes risk. In Proceedings of the 19th Annual Conference on Learning Theory, COLT 2006, pages 79–93. Springer, 2006.

I. Steinwart, D. Hush, and C. Scovel. Learning from dependent observations. Technical report, Los Alamos National Laboratory, 2006. http://www.c3.lanl.gov/ml/pubs/2006_dependent/paper.pdf

I. Steinwart, D. Hush, and C. Scovel. The reproducing kernel Hilbert space of the Gaussian RBF kernel. IEEE Trans. Inform. Theory, 52:4635–4643, 2006.

J.A.K. Suykens, T. Van Gestel, J. De Brabanter, B. De Moor, and J. Vandewalle. Least Squares Support Vector Machines. World Scientific, 2002.

V.N. Vapnik. Statistical Learning Theory. Wiley, 1998.

G. Wahba. Spline Models for Observational Data. Series in Applied Mathematics, Vol. 59, SIAM, Philadelphia, 1990.