Research Article

Learning Bounds of ERM Principle for Sequences of Time-Dependent Samples

Mingchen Yao, Chao Zhang, and Wei Wu

School of Mathematical Sciences, Dalian University of Technology, Dalian, Liaoning 116024, China

Correspondence should be addressed to Chao Zhang; zhangchao1015@gmail.com

Received 17 July 2015; Revised 26 October 2015; Accepted 1 November 2015

Academic Editor: Juan R. Torregrosa

Copyright © 2015 Mingchen Yao et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Many generalization results in learning theory are established under the assumption that samples are independent and identically distributed (i.i.d.). However, numerous learning tasks in practical applications involve the time-dependent data. In this paper, we propose a theoretical framework to analyze the generalization performance of the empirical risk minimization (ERM) principle for sequences of time-dependent samples (TDS). In particular, we first present the generalization bound of ERM principle for TDS. By introducing some auxiliary quantities, we also give a further analysis of the generalization properties and the asymptotical behaviors of ERM principle for TDS.

1. Introduction

Let $\mathcal{X} \subset \mathbb{R}^I$ and $\mathcal{Y} \subset \mathbb{R}^J$ be an input space and the corresponding output space, respectively. Define $\mathcal{Z} := \mathcal{X} \times \mathcal{Y} \subset \mathbb{R}^P$ with $P = I + J$. Many classical results of statistical learning theory are built under the assumption that samples $\{\mathbf{z}_n\} := \{(\mathbf{x}_n, \mathbf{y}_n)\}$ are independently drawn from an identical distribution on $\mathcal{Z}$, that is, the so-called i.i.d.-sample assumption, for example, [1–5]. However, the theoretical results sharing the i.i.d.-sample assumption may not be valid for (or cannot be directly used in) the non-i.i.d. scenario.

There have been many research interests lying in the theoretical analysis of the learning processes for time-dependent samples. Zhang and Tao [6, 7] study the generalization properties of ERM-based learning processes for time-dependent samples drawn from Lévy processes and continuous-time Markov chains, respectively. Zou et al. [8] establish exponential bound on the rate of relative uniform convergence for the ERM algorithm with dependent observations and derived the related generalization bounds with α -mixing sequence. Jiang [9] introduces a triplex inequality based on Hoeffding's inequality (cf. [10]) and then obtains probability bounds for uniform deviations in a very general framework, which is suitable for the cases with unbounded loss functions and dependent samples. Zou et al. [11] present a novel Markov

sampling algorithm to generate uniformly ergodic Markov chain samples from a given dataset. Xu et al. [12] adopt the similar techniques of [11] to develop the error bound for an online SVM classification algorithm, which provides a higher learning performance than that of classical random sampling methods.

In this paper, we are mainly concerned with the generalization performance of ERM principle for the sequences of time-dependent samples (TDS), which are independently and repeatedly observed from an undetermined stochastic process at the fixed time points. We can see that the generalization performance of such learning process is affected by the following factors: the number of independent observation sequences, time-dependence among samples in a sequence, the sample number in a sequence, and the observing time points.

By introducing some auxiliary quantities, we propose a new framework to analyze the generalization properties of the learning process. In particular, we first show that the generalization bound of this learning process can be decomposed into four parts: Φ_1 , Φ_2 , Φ_3 , and Φ_4 , which are related to the aforementioned factors. We then analyze the properties of these quantities. By imposing some mild conditions, we further obtain the upper bounds of the first three types of quantities Φ_1 , Φ_2 , and Φ_3 , respectively. Finally,

some techniques of statistical learning theory are applied to bound the quantity Φ_4 , for example, the uniform entry number, the relevant deviation, and the symmetrization inequalities for independent sequences of TDS.

Different from previous works [6, 7], there is no specific assumption on the distribution of the stochastic process sampled from. In contrast, the previous works require that samples should be observed from Lévy processes and continuous-time Markov chains, respectively. Instead of the only $\mathcal{X}$ -valued function class appearing in the previous works, we impose a function class consisting of functions evaluated on $\mathcal{X}$ and the time interval $[T_1, T_2]$. Moreover, the samples considered in previous works [6–8] are a series of data observed from one certain stochastic process, while this paper discusses the learning process based on quantities of independent sequences of TDS observed from a stochastic process at the fixed time points.

Moreover, the works [9, 11, 12] only consider the case of one single sequence of TDS, while this paper studies the learning process for multiple sequences of TDS, where one sequence corresponds to one trajectory (sample path) of a stochastic process. Therefore, our results are more general than previous ones.

The rest of paper is organized as follows. In Section 2, we first introduce some notions and notations used in the paper and then exhibit the decomposition of the generalization bounds of ERM principle for TDS. In Section 3, we bound the four auxiliary quantities Φ_1 , Φ_2 , Φ_3 , and Φ_4 and present the main results of the paper. The last section concludes the paper.

2. Problem Setup

In this section, we formalize the main research issue of this paper and then show the decomposition of the generalization error of ERM principle for TDS.

2.1. ERM Principle and Generalization Bounds. For any $t \geq 0$, let $\mathbf{x}_t \in \mathcal{X}$ and $\mathbf{y}_t \in \mathcal{Y}$ be the t -time inputs and the corresponding outputs, respectively. Denote $\mathbf{z}_t := (\mathbf{x}_t, \mathbf{y}_t) \in \mathbb{R}^P$ ($t \geq 0$) and assume that $\{\mathbf{z}_t\}_{t \geq 0}$ is an undetermined stochastic process with a countable space. Consider a function class $\mathcal{G} \subset \mathcal{Y}^{\mathcal{X} \times \mathcal{T}}$ consisting of functions evaluated at the input space $\mathcal{X}$ and the time interval $\mathcal{T} = [T_1, T_2]$. We would like to find a function $g^* \in \mathcal{G}$ such that, for any input $\mathbf{x}_t \in \mathcal{X}$ ($t \in \mathcal{T}$), the corresponding output $\mathbf{y}_t$ can be predicted as accurately as possible.

A natural criterion to choose the function g^* is the lowest *expected risk* caused by some function in $\mathcal{G}$:

$$\begin{aligned} g^* &= \arg \min_{g \in \mathcal{G}} E(\ell \circ g) \\ &= \arg \min_{g \in \mathcal{G}} \frac{1}{T} \int_{T_1}^{T_2} \int \ell(g(\mathbf{x}_t, t), \mathbf{y}_t) dP_t dt, \end{aligned} \quad (1)$$

where $g(\mathbf{x}_t, t)$ is a function with respect to the t -time input $\mathbf{x}_t$ and the time point t (Here, $\mathbf{x}_t$ in the function $g(\mathbf{x}_t, t)$ is just an input value of the functional $g(\cdot, t)$), $\ell \circ g$ is the composite

function ℓ after g , and P_t stands for the t -time distribution of the stochastic process $\{\mathbf{z}_t\}_{t \geq 0}$ on $\mathcal{X}$. However, the distribution of $\{\mathbf{z}_t\}_{t \geq 0}$ is unknown, and it is difficult to directly obtain the target function g^* by minimizing the expected risk $E(\ell \circ g)$.

Instead, the empirical risk minimization (ERM) principle provides a solution scheme to this issue. For $1 \leq m \leq M$, let $\mathbf{S}_N^{[m]} := \{\mathbf{z}_{t_n}^{[m]}\}_{n=1}^N$ be the m th sequence of samples observed from a certain stochastic process $\{\mathbf{z}_t\}_{t \geq 0}$ at the fixed time points $T_1 \leq t_1 < t_2 < \dots < t_n \leq T_2$ and let the sequences $\mathbf{S}_N^{[m]}$ ($1 \leq m \leq M$) be independent of each other. The ERM principle aims to minimize the *empirical risk* over $\mathcal{G}$:

$$\begin{aligned} \hat{g} &= \arg \min_{g \in \mathcal{G}} \hat{E}(\ell \circ g) \\ &= \arg \min_{g \in \mathcal{G}} \frac{1}{MN} \sum_{m=1}^M \sum_{n=1}^N \ell(g(\mathbf{x}_{t_n}^{[m]}, t_n), \mathbf{y}_{t_n}^{[m]}), \end{aligned} \quad (2)$$

and the solution $\hat{g}$ is regarded as an estimate to the expected solution g^* with respect to the sample sequences $\{\mathbf{S}_N^{[m]}\}_{m=1}^M$.

For convenience, we further define the loss function class

$$\mathcal{F} := \{(\mathbf{z}_t, t) \mapsto \ell(g(\mathbf{x}_t, t), \mathbf{y}_t) : g \in \mathcal{G}\} \quad (3)$$

and call $\mathcal{F}$ the function class in the rest of this paper. Given M sample sequences $\{\mathbf{S}_N^{[m]}\}_{m=1}^M$, taking $f = \ell \circ g$ provides the following brief notations of expected risk (1) and empirical risk (2):

$$Ef = E(\ell \circ g) = \frac{1}{T} \int_{T_1}^{T_2} \int f(\mathbf{z}_t, t) dP_t dt, \quad (4)$$

$$\hat{E}f = \hat{E}(\ell \circ g) = \frac{1}{MN} \sum_{m=1}^M \sum_{n=1}^N f(\mathbf{z}_{t_n}^{[m]}, t_n). \quad (5)$$

Similar to the classical statistical learning theory [13], the main issue of this paper is to discuss whether the empirical solution $\hat{g}$ provided by ERM principle will perform as well as the expected solution g^* . Then, the supremum

$$\sup_{f \in \mathcal{F}} |Ef - \hat{E}f|, \quad (6)$$

called the generalization bound of ERM principle for TDS $\{\mathbf{S}_N^{[m]}\}_{m=1}^M$, will play an important role in analyzing the generalization performance of the above learning process as well.

2.2. Relationship with Some Time-Dependent Problem. At the end of this section, we will show that many time-dependent problems can come down to the aforementioned learning process based on ERM principle, for example, the estimation of information channel and functional linear models.

2.2.1. Estimation of Information Channel. Since the functions in $\mathcal{F}$ are evaluated at both the real input space $\mathcal{X}$ and the time interval $\mathcal{T}$, the framework proposed in the paper can describe the inherent characteristics of the time-dependent problems more precisely than the learning framework given in the previous works [6, 7], where the function class is only evaluated

at $\mathcal{X}$. Certainly, the time-dependent problem mentioned in [14], for example, the estimation of information channel, can also be included in this learning setting. Different from the previous one, the setting considered in this paper is also suitable for analyzing the performance of estimating dynamic information channel.

The estimation of dynamic information channel is of the following model: $\mathbf{y}_t = \mathbf{H}(t)\mathbf{x}_t + \mathbf{n}(t)$, where $\mathbf{H}(t)$ and $\mathbf{n}(t)$, changing status with time varying, are the channel matrix and the noise vector, respectively. The corresponding function class $\mathcal{H}$ is formalized as

$$\mathcal{H} := \left\{ (\mathbf{x}, t) \mapsto \mathbf{H}(t)\mathbf{x} + \mathbf{n}(t) : \mathbf{H}(t) \in \mathbb{R}^J \times \mathbb{R}^I, \mathbf{n}(t) \in \mathbb{R}^J \right\}. \quad (7)$$

2.2.2. Functional Linear Models. Moreover, the functional data classification or regression with functional linear models is also in accordance with the learning setting mentioned in this paper. By denoting

$$\begin{aligned} \mathbf{B}(t) &:= (\beta_{ji}(t))_{J \times I}, \\ \boldsymbol{\rho}(t) &:= (\rho_1(t), \dots, \rho_J(t))^T, \\ \boldsymbol{\tau}(t) &:= (\tau_1(t), \dots, \tau_J(t))^T, \end{aligned} \quad (8)$$

the following are the most frequently used functional linear models mentioned by [15]:

- (i) The model with a scalar input and a functional output: $\mathbf{y}(t) = \mathbf{B}(t)\mathbf{x} + \boldsymbol{\rho}(t)$, which corresponds to the function class for any $t \in \mathcal{T}$:

$$\mathcal{G}(\cdot, t) = \left\{ \mathbf{x} \mapsto \mathbf{B}(t)\mathbf{x} + \boldsymbol{\rho}(t) : \beta_{ji}(t), \rho_j(t) \in \mathbb{R}^{\mathcal{T}} \right\}. \quad (9)$$

- (ii) The model with a functional input and a scalar output: $y = \boldsymbol{\tau} + \int_{\Theta_t} \mathbf{B}(t)\mathbf{x}(t)dt + \boldsymbol{\rho}$, which corresponds to the function class for any $t \in \mathcal{T}$:

$$\begin{aligned} \mathcal{G}(\cdot, t) &= \left\{ \mathbf{x}(t) \mapsto \boldsymbol{\tau} + \int_{\Theta_t} \mathbf{B}(t)\mathbf{x}(t)dt + \boldsymbol{\rho} : \beta_{ji}(t) \in \mathbb{R}^{\mathcal{T}}; \boldsymbol{\tau}, \boldsymbol{\rho} \in \mathbb{R}^J \right\}. \end{aligned} \quad (10)$$

- (iii) The two models with a functional input and a functional output:

- (a) one model is $\mathbf{y}(t) = \boldsymbol{\tau}(t) + \mathbf{B}(t)\mathbf{x}(t) + \boldsymbol{\rho}(t)$, which corresponds to the function class for any $t \in \mathcal{T}$:

$$\begin{aligned} \mathcal{G}(\cdot, t) &= \left\{ \mathbf{x}(t) \mapsto \boldsymbol{\tau}(t) + \mathbf{B}(t)\mathbf{x}(t) \right. \\ &\quad \left. + \boldsymbol{\rho}(t) : \beta_{ji}(t), \rho_j(t), \tau_j(t) \in \mathbb{R}^{\mathcal{T}} \right\}; \end{aligned} \quad (11)$$

- (b) the other is $\mathbf{y}(t) = \boldsymbol{\tau}(t) + \int_{\Theta_t} \mathbf{B}(s, t)\mathbf{x}(t)ds + \boldsymbol{\rho}(t)$, which corresponds to the function class for any $t \in \mathcal{T}$:

$$\begin{aligned} \mathcal{G}(\cdot, t) &= \left\{ \mathbf{x}(t) \mapsto \boldsymbol{\tau}(t) + \int_{\Theta_t} \mathbf{B}(s, t)\mathbf{x}(t)ds \right. \\ &\quad \left. + \boldsymbol{\rho}(t) : \beta_{ji}(s, t) \in \mathbb{R}^{\mathcal{T} \times \mathcal{T}}; \rho_j(t), \tau_j(t) \in \mathbb{R}^{\mathcal{T}} \right\}. \end{aligned} \quad (12)$$

In the above models, the loss function ℓ is usually selected as the mean square error function and then some functional data algorithms are used to find the function that minimizes empirical risk (2) over the function class $\mathcal{F}$. We refer to [16] for more details on functional linear models.

2.3. Analysis of Generalization Performance. Different from the i.i.d. learning setting, regardless of the distribution characteristics of $\{\mathbf{z}_t\}_{t \geq 0}$, the generalization performance of the aforementioned learning process is also affected by the following factors:

- (i) Time-dependence among samples in one sequence.
- (ii) The number M of independent sequences of samples.
- (iii) The number N of samples in one sequence.
- (iv) The fixed time points $t_1, \dots, t_N$, at which the samples $\mathbf{z}_{t_1}, \dots, \mathbf{z}_{t_N}$ are observed.

The main concern of this paper is to quantitatively analyze the influence to the generalization performance caused by these factors and then achieve the generalization bound of ERM principle for TDS.

Proposition 1. Denote $\Gamma^* := \{\tau_0^*, \tau_1^*, \dots, \tau_N^*\} \subset [T_1, T_2]$ as an ordered set with $T_1 = \tau_0^* < \tau_1^* < \dots < \tau_N^* < T_2$ and $\Delta\tau_n^* := \tau_n^* - \tau_{n-1}^*$ ($1 \leq n \leq N$). If the set Γ^* achieves the supremum

$$\begin{aligned} \sup_{\substack{f \in \mathcal{F} \\ \Gamma \subset [T_1, T_2]}} & \left| \sum_{n=1}^N \frac{\Delta\tau_n^*}{T} \int f(\mathbf{z}_{\tau_n^*}, \tau_n^*) dP_{\tau_n^*} \right. \\ & \left. - \frac{1}{NM} \sum_{m=1}^M \sum_{n=1}^N f(\mathbf{z}_{t_n}^{[m]}, t_n) \right|, \end{aligned} \quad (13)$$

then there holds that

$$\sup_{f \in \mathcal{F}} |Ef - \widehat{Ef}| \leq \Phi_1 + \Phi_2 + \Phi_3 + \Phi_4, \quad (14)$$

where the quantities Φ_1 , Φ_2 , Φ_3 , and Φ_4 are, respectively, defined by

$$\Phi_1 := \sum_{n=1}^N \left(\left| \frac{\Delta\tau_n^*}{T} - \frac{1}{N} \right| \sup_{f \in \mathcal{F}} \left| \int f(\mathbf{z}_{\tau_n^*}, \tau_n^*) dP_{\tau_n^*} \right| \right); \quad (15)$$

$$\Phi_2$$

$$:= \sup_{f \in \mathcal{F}} \frac{1}{N} \sum_{n=1}^N \left| \int f(\mathbf{z}_{\tau_n^*}, \tau_n^*) dP_{\tau_n^*} - \int f(\mathbf{z}_{t_n}, \tau_n^*) dP_{t_n} \right|; \quad (16)$$

$$\Phi_3 := \sup_{f \in \mathcal{F}} \frac{1}{N} \sum_{n=1}^N \left| \int f(\mathbf{z}_{t_n}, \tau_n^*) dP_{t_n} - \int f(\mathbf{z}_{t_n}, t_n) dP_{t_n} \right|; \quad (17)$$

$$\Phi_4 := \sup_{f \in \mathcal{F}} \left| \frac{1}{N} \sum_{n=1}^N \int f_{t_n}(\mathbf{z}) dP_{t_n} - \frac{1}{MN} \sum_{n=1}^N \sum_{m=1}^M f_{t_n}(\mathbf{z}_{t_n}^{[m]}) \right|. \quad (18)$$

This result implies that the behavior of the generalization bound can be described by using the summation of the quantities Φ_1 , Φ_2 , Φ_3 , and Φ_4 . In the next section, we will discuss the properties of the quantities Φ_1 , Φ_2 , Φ_3 , and Φ_4 , respectively.

3. Analysis of Relevant Quantities

As mentioned above, regardless of the distribution characteristics of $\{\mathbf{z}_t\}_{t \geq 0}$, there are four factors affecting the generalization performance of the ERM learning process for TDS: time-dependence, the number of TDS sequences, the sample number, and the observing time points, which are actually related to the quantities Φ_1 , Φ_2 , Φ_3 , and Φ_4 , respectively.

Moreover, two mild conditions will be imposed for the following discussion:

- (C1) Assume that there exists a constant C_1 such that, for all $t \in \mathcal{T}$ and $f \in \mathcal{F}$, there holds that

$$\int f(\mathbf{z}_t, t) dP_t < C_1. \quad (19)$$

- (C2) Assume that each $f \in \mathcal{F}$ is differentiable with respect to the time t and there exists a constant C_2 such that, for all $t \geq 0$ and $f \in \mathcal{F}$, there holds that, for any $\mathbf{z} \in \mathcal{Z}$ and $t \in \mathcal{T}$,

$$\frac{\partial f}{\partial t}(\mathbf{z}, t) \leq C_2. \quad (20)$$

The former requires that any function $f \in \mathcal{F}$ should have the bounded expectation with respect to the distribution of $\{\mathbf{z}_t\}_{t \geq 0}$ at any time $t \in \mathcal{T}$, and the latter implies that all functions in $\mathcal{F}$ should have bounded first-order partial derivative with respect to t .

3.1. Upper Bound of Quantity Φ_1 . Under Condition (C1) and recalling (15), we can bound the quantity Φ_1 as follows:

$$\Phi_1 \leq C_1 \sum_{n=1}^N \left| \frac{\Delta \tau_n^*}{T} - \frac{1}{N} \right|, \quad (21)$$

which implies that Φ_1 is affected by the choice of the specific time sequence $\Gamma^* = \{\tau_0^*, \tau_1^*, \dots, \tau_N^*\}$ which achieves supremum (13). Note that it is possible that the time sequence

Γ^* is not the unique one which can achieve supremum (13); then we define

$$\Gamma_N^* := \arg \min_{\Gamma} \sum_{n=1}^N \left| \frac{\Delta \tau_n}{T} - \frac{1}{N} \right|. \quad (22)$$

It can be observed that if the time points $\tau_1^*, \dots, \tau_N^*$ satisfy that the length of each $\Delta \tau_n^*$ ($1 \leq n \leq N$) is equivalent, the sampling-time error Φ_1 is equal to zero. From the probabilistic perspective, it means that Γ_N^* is the one that is closest to uniform distribution among the candidate sequences. In other words, more uniformly the time points in Γ_N^* are distributed and the quantity Φ_1 is closer to zero. On the other hand, the uniform distribution of time points $\tau_0^*, \tau_1^*, \dots, \tau_N^*$ implies that, for the stochastic process $\{\mathbf{z}_t\}_{t \geq 0}$, the distributions of $\mathbf{z}_t$ are identical for all $t \leq 0$, which is actually a probability distribution.

3.2. Upper Bound of Quantity Φ_2 . Denote

$$\begin{aligned} \mathcal{F}_t &:= \{f(\cdot, t) : f \in \mathcal{F}\}, \\ D_{\mathcal{F}_t}(P_t, P_\tau) &= \sup_{f(\cdot, t) \in \mathcal{F}_t} \left| \int f(\mathbf{z}_t, t) dP_t - \int f(\mathbf{z}_\tau, t) dP_\tau \right|, \end{aligned} \quad (23)$$

which is the so-called integral probability metric (IPM) between the two distributions P_t and P_τ with respect to the function class $\mathcal{F}_t$. IPM plays an important role in measuring the discrepancy between two distributions in probability theory, and we refer to [17, 18] for more details on IPM.

According to (16), we then obtain that

$$\Phi_2 \leq \frac{1}{N} \sum_{n=1}^N D_{\mathcal{F}_{\tau_n^*}}(P_{t_n}, P_{\tau_n^*}). \quad (24)$$

Particularly, according to (23), if the stochastic process $\{\mathbf{z}_t\}_{t \geq 0}$ has an identical distribution at any time, the quantity Φ_2 is equal to zero.

3.3. Upper Bound of Quantity Φ_3 . According to (17), it follows from the condition of (C2) that

$$\begin{aligned} \Phi_3 &= \sup_{f \in \mathcal{F}} \frac{1}{N} \sum_{n=1}^N \left| \int f(\mathbf{z}_{t_n}, \tau_n^*) dP_{t_n} - \int f(\mathbf{z}_{t_n}, t_n) dP_{t_n} \right| \\ &= \sup_{f \in \mathcal{F}} \frac{1}{N} \sum_{n=1}^N \left| \int \frac{\partial f}{\partial t}(\mathbf{z}_{t_n}, \rho_n) (\tau_n^* - t_n) P_{t_n} \right| \\ &\leq \frac{C_2}{N} \sum_{n=1}^N |\tau_n^* - t_n|, \end{aligned} \quad (25)$$

where ρ_n lies between τ_n^* and t_n . According to (25), quantity Φ_3 can be bounded by the summation of the discrepancies between the supremum time points τ_n^* and the fixed observation time point t_n . Recalling (13), the time points $\{\tau_n^*\}_{n=1}^N$

are determined by the selection of $\{t_n\}_{n=1}^N$. It is noteworthy that the time points $t_1, \dots, t_n$ usually cannot coincide with $\tau_1^*, \dots, \tau_N^*$.

3.4. Upper Bound of Quantity Φ_4 . Similar to the classical statistical learning theory, the task to bound Φ_4 can be divided into three steps: complexity measure of function classes and the deviation and the symmetrization inequalities. Based on Azuma's inequality, we will obtain a suitable deviation inequality for TDS and then the relevant symmetrization inequality. By using the uniform entropy number, we finally bound the quantity Φ_4 in the sense of probability.

3.4.1. Deviation Inequality. Let $\mathbf{S}_N = \{\mathbf{z}_{t_n}\}_{n=1}^N$ be the observation sequence of the stochastic process $\{\mathbf{z}_t\}_{t \geq 0}$ with a countable state space $\mathcal{Z}$. By the way presented in [19], a filtration associated with $\mathbf{S}_N$ can be built as follows:

- (i) Let Ω_N be all subset of $\mathcal{Z}^N$.
- (ii) Set $\Omega_0 = \{\emptyset, \mathcal{Z}^N\}$.
- (iii) For any $1 \leq n \leq N-1$, let Ω_n be the σ -algebra generated by $\{\mathbf{z}_{t_1}, \dots, \mathbf{z}_{t_{n-1}}\}$.

Then, there naturally holds that

$$\Omega_0 \subset \Omega_1 \subset \dots \subset \Omega_N. \quad (26)$$

According to Azuma's inequality [19, 20], since the observation sequences $\mathbf{S}_N^{[1]}, \dots, \mathbf{S}_N^{[M]}$ are independent of each other, we obtain the following result.

Lemma 2. Given a function $F : \mathcal{Z}^N \rightarrow \mathbb{R}$, set $F := \sum_{n=1}^N f(\mathbf{z}_{t_n}, t_n)$ and for any $1 \leq n \leq N$, define the associated martingale differences as follows:

$$U_n(F) = E\{F(\mathbf{S}_N) \mid \Omega_n\} - E\{F(\mathbf{S}_N) \mid \Omega_{n-1}\}. \quad (27)$$

Then, there holds that, for any $\xi > 0$,

$$\Pr \left\{ \left| \frac{1}{M} \sum_{m=1}^M F(\mathbf{S}_N^{[m]}) - E\{F(\mathbf{S}_N)\} \right| > \xi \right\} \leq 2 \exp \left(- \frac{M\xi^2}{2 \sum_{n=1}^N |U_n(F)|^2} \right). \quad (28)$$

3.4.2. Symmetrization Inequality. Symmetrization inequalities are mainly used to replace the expected risk by an empirical risk computed on another sample set that is independent of the given sample set but has the identical distribution. In this manner, some kinds of complexity measures can be incorporated in the resulting bounds.

For clarity of presentation, we give a useful notation for the following discussion. Given a sample sequence $\mathbf{S}_N = \{\mathbf{z}_{t_n}\}_{n=1}^N$, consider another sample sequence $\mathbf{S}'_N = \{\mathbf{z}'_{t_n}\}_{n=1}^N$ independent of $\mathbf{S}_N$. If $\mathbf{z}'_{t_n}$ has the same distribution as $\mathbf{z}_{t_n}$ for any $1 \leq n \leq N$, the sequence $\mathbf{S}'_N$ is called the ghost sample sequence of $\mathbf{S}_N$. The following result is a direct extension of the classical symmetrization results of Lemma 2 in [21].

Theorem 3. Assume that $\mathcal{F}$ is a function class consisting of the functions with the range $[a, b]$ and $\{\mathbf{z}_t\}_{t \geq 0}$ is a stochastic process with a countable state space $\Omega \subset \mathbb{R}^K$. Let $\{\mathbf{S}_N^{[m]}\}_{m=1}^M$ and $\{\mathbf{S}'_N^{[m]}\}_{m=1}^M$ be drawn from $\{\mathbf{z}_t\}_{t \geq 0}$ in the time interval $[T_1, T_2]$. Then given any $\xi > 0$, one has, for any $M \geq 8(b-a)^2/(\xi)^2$,

$$\Pr \left\{ \sup_{f \in \mathcal{F}} \left| \frac{1}{M} \sum_{m=1}^M F(\mathbf{S}_N^{[m]}) - E\{F(\mathbf{S}_N)\} \right| > \xi \right\} \leq 2 \Pr \left\{ \sup_{f \in \mathcal{F}} \frac{1}{M} \left| \sum_{m=1}^M F(\mathbf{S}_N^{[m]}) - \sum_{m=1}^M F(\mathbf{S}'_N^{[m]}) \right| > \frac{\xi}{2} \right\}, \quad (29)$$

where $F(\mathbf{S}_N^{[m]}) := \sum_{n=1}^N f(\mathbf{z}_{t_n}^{[m]}, t_n)$.

3.4.3. Approximate Error Φ_4 . The following is the definition of the covering number and we refer to [22] for details.

Definition 4. Let $\mathcal{F}$ be a function class and let d be a metric on $\mathcal{F}$. For any $\xi > 0$, the covering number of $\mathcal{F}$ at radius ξ with respect to the metric d , denoted by $\mathcal{N}(\mathcal{F}, \xi, d)$, is the minimum size of a cover of radius ξ .

The uniform entropy number (UEN) is a variant of the covering number and we refer to [22] for details as well. By setting the metric $\ell_p(\mathbf{S}_N)$ ($p > 0$), the UEN is defined as follows:

$$\ln \mathcal{N}_p(\mathcal{F}, \xi, N) := \sup_{\mathbf{z}_1^N} \ln \mathcal{N}(\mathcal{F}, \xi, \ell_p(\mathbf{S}_N)). \quad (30)$$

Based on the uniform entropy number, the upper bound of Φ_4 can be obtained as follows.

Theorem 5. Assume that $\mathcal{F} \subset \mathcal{Y}^{\mathcal{X} \times \mathcal{T}}$ with $\mathcal{T} = [T_1, T_2]$ is a function class such that, for any $f \in \mathcal{F}$ and any $t \in \mathcal{T}$, the functional $f(\cdot; t)$ is bounded with the range $[a, b]$. Let $\{\mathbf{z}_t\}_{t \geq 0}$ be a stochastic process with a countable state space. Let $\{\mathbf{S}_N^{[m]}\}_{m=1}^M$ be M independent observation sequences of $\{\mathbf{z}_t\}_{t \geq 0}$ at the time points $t_1, t_2, \dots, t_N$. Let $\{\mathbf{S}'_N^{[m]}\}_{m=1}^M$ be the ghost samples of $\{\mathbf{S}_N^{[m]}\}_{m=1}^M$ and denote $\mathbf{S}_{2N}^{[m]} := \{\mathbf{S}_N^{[m]}, \mathbf{S}'_N^{[m]}\}$ for any $1 \leq m \leq M$. Then, given any $\xi > 0$, one has, for any $M \geq 8(b-a)^2/(\xi)^2$,

$$\Pr \left\{ \sup_{f \in \mathcal{F}} \left| \frac{1}{N} \sum_{n=1}^N \int f(\mathbf{z}_{t_n}, t_n) dP_{t_n} - \frac{1}{MN} \sum_{n=1}^N \sum_{m=1}^M f(\mathbf{z}_{t_n}^{[m]}, t_n) \right| > \xi \right\} \leq 8\mathcal{N}_1 \left(\mathcal{F}, \frac{\xi}{8} \right) \exp \left(- \frac{MN^2\xi^2}{128 \sum_{n=1}^N |U_n(F)|^2} \right). \quad (31)$$

Compared to the classical generalization results for i.i.d. samples (cf. Theorem 2.3 of [22]), time-dependent data are involved in the analysis of upper bound (31). This result is also different from the generalization bound of [7] because more than one sequence of time-dependent samples are

investigated. Denote the right-hand side of (31) by ϵ , and under notations of Theorem 5 we immediately obtain that, with probability of at least $1 - \epsilon$,

$$\sup_{f \in \mathcal{F}} \left| E \{F(\mathbf{S}_N)\} - \frac{1}{M} \sum_{m=1}^M F(\mathbf{S}_N^{[m]}) \right| \leq \sqrt{\frac{128 (\ln 8\mathcal{N}_1(\mathcal{F}, \xi/8, 2MN) - \ln \epsilon) \sum_{n=1}^N |U_n(F)|^2}{MN^2}}, \quad (32)$$

which explicitly shows the upper bound of Φ_4 .

Finally, according to the upper bounds of Φ_i ($i = 1, 2, 3, 4$) and Proposition 1, we can bound $\sup_{f \in \mathcal{F}} |Ef - \hat{E}f|$ given in (6) as follows: with probability of at least $1 - \epsilon$,

$$\begin{aligned} \sup_{f \in \mathcal{F}} |Ef - \hat{E}f| &\leq C_1 \sum_{n=1}^N \left| \frac{\Delta \tau_n^*}{T} - \frac{1}{N} \right| + \frac{1}{N} \sum_{n=1}^N D_{\mathcal{F}_{\tau_n^*}}(P_{t_n}, P_{\tau_n^*}) \\ &\quad + \frac{C_2}{N} \sum_{n=1}^N |\tau_n^* - t_n| \\ &\quad + \sqrt{\frac{128 (\ln 8\mathcal{N}_1(\mathcal{F}, \xi/8, 2MN) - \ln \epsilon) \sum_{n=1}^N |U_n(F)|^2}{MN^2}}. \end{aligned} \quad (33)$$

From result (33), we can find that the generalization performance of ERM principle for TDS is affected by the following factors:

- (i) The choice of the observation time points $t_1, t_2, \dots, t_N$.
- (ii) The number M of TDS sequences.
- (iii) The length N of sequences.
- (iv) The properties of the martingale differences $\{U_n(F)\}$.
- (v) The complexity of function classes.

And meanwhile, the quantitative relationship among these factors is explicitly given in the above result as well.

4. Conclusion

In this paper, we propose a new framework to analyze the generalization properties of ERM principle for the sequences of TDS. By introducing four auxiliary quantities ϕ_1 , ϕ_2 , ϕ_3 , and ϕ_4 , we give a detailed insight to the interactions among the complexity of function classes, the size of samples, and the dependence among samples. To achieve the upper bound of ϕ_4 , we develop the relevant deviation inequality and symmetrization inequality for the sequences of TDS. This work is an extension of the classical techniques of statistical learning theory. In our future works, we will consider to relax conditions (C1) and (C2) and further investigate the properties of ERM-based learning processes for the TDS sequences, which are not observed at the fixed observation time points.

Appendices

A. Proof of Proposition 1

According to (3), (4), and (5), we have

$$\begin{aligned} \sup_{f \in \mathcal{F}} |Ef - \hat{E}f| &= \sup_{f \in \mathcal{F}} \left| \frac{1}{T} \int_{T_1}^{T_2} f(\mathbf{z}_t, t) dP_t dt \right. \\ &\quad \left. - \frac{1}{NM} \sum_{m=1}^M \sum_{n=1}^N f(\mathbf{z}_{t_n}^{[m]}, t_n) \right| \\ &\leq \sup_{f \in \mathcal{F}} \left| \sum_{n=1}^M \frac{\Delta \tau_n}{T} \int f(\mathbf{z}_{\tau_n}, \tau_n) dP_{\tau_n} \right. \\ &\quad \left. - \frac{1}{NM} \sum_{m=1}^M \sum_{n=1}^N f(\mathbf{z}_{t_n}^{[m]}, t_n) \right| \\ &\leq \sup_{f \in \mathcal{F}} \left| \sum_{n=1}^N \frac{\Delta \tau_n^*}{T} \int f(\mathbf{z}_{\tau_n^*}, \tau_n^*) dP_{\tau_n^*} \right. \\ &\quad \left. - \sum_{n=1}^N \frac{1}{N} \int f(\mathbf{z}_{\tau_n^*}, \tau_n^*) dP_{\tau_n^*} \right| \\ &\quad + \sup_{f \in \mathcal{F}} \left| \sum_{n=1}^N \frac{1}{N} \int f(\mathbf{z}_{\tau_n^*}, \tau_n^*) dP_{\tau_n} \right. \\ &\quad \left. - \sum_{n=1}^N \frac{1}{N} \int f(\mathbf{z}_{t_n}, t_n) dP_{t_n} \right| \\ &\quad + \sup_{f \in \mathcal{F}} \left| \sum_{n=1}^N \frac{1}{N} \int f(\mathbf{z}_{t_n}, t_n) dP_{t_n} \right. \\ &\quad \left. - \frac{1}{NM} \sum_{m=1}^M \sum_{n=1}^N f(\mathbf{z}_{t_n}^{[m]}, t_n) \right|, \end{aligned} \quad (A.1)$$

which can lead to result (14). This completes the proof.

B. Proof of Theorem 5

Given a time sequence $t_1, \dots, t_N$, let $\mathbf{S}_N = \{\mathbf{z}_{t_1}, \dots, \mathbf{z}_{t_N}\}$. For any $f \in \mathcal{F}$, define a function $F: \mathcal{Z}^N \rightarrow \mathbb{R}$:

$$F(\mathbf{S}_N) = \sum_{n=1}^N f(\mathbf{z}_{t_n}, t_n). \quad (B.1)$$

Then, consider $\{\epsilon_m\}_{m=1}^M$ as independent Rademacher random variables, that is, independent $\{-1, 1\}$ -valued random variables with equal probability of taking either value. Given $\{\epsilon_m\}_{m=1}^M$ and $\mathbf{S}_{2N}$, we denote

$$\vec{\epsilon} := (\epsilon_1, \dots, \epsilon_M, -\epsilon_1, \dots, -\epsilon_M)^\top, \quad (B.2)$$

and for any $f \in \mathcal{F}$,

$$\begin{aligned} \vec{F}(\mathbf{s}_{2N}^{[1]}, \dots, \mathbf{s}_{2N}^{[M]}) &:= (F(\mathbf{s}_N'^{[1]}), \dots, F(\mathbf{s}_N'^{[M]}), F(\mathbf{s}_N^{[1]}), \\ &\dots, F(\mathbf{s}_N^{[M]}))^T. \end{aligned} \quad (\text{B.3})$$

According to (5) and Theorem 3, given any $\xi > 0$, we have, for any $N \geq 8(b-a)^2/\xi^2$,

$$\begin{aligned} &\Pr \left\{ \sup_{f \in \mathcal{F}} \left| \frac{1}{M} \sum_{m=1}^M F(\mathbf{s}_N^{[m]}) - E\{F(\mathbf{s}_N)\} \right| > \xi \right\} \leq 2 \\ &\cdot \Pr \left\{ \sup_{f \in \mathcal{F}} \frac{1}{M} \left| \sum_{m=1}^M F(\mathbf{s}_N^{[m]}) - \sum_{m=1}^M F(\mathbf{s}_N'^{[m]}) \right| > \frac{\xi}{2} \right\} \\ &= 2 \Pr \left\{ \sup_{f \in \mathcal{F}} \frac{1}{M} \left| \sum_{m=1}^M \epsilon_m (F(\mathbf{s}_N^{[m]}) - F(\mathbf{s}_N'^{[m]})) \right| \right. \\ &> \frac{\xi}{2} \left. \right\} = 2 \Pr \left\{ \sup_{f \in \mathcal{F}} \left| \frac{1}{2M} \langle \vec{\epsilon}, \vec{F}(\mathbf{s}_{2N}^{[1]}, \dots, \mathbf{s}_{2N}^{[M]}) \rangle \right| \right. \\ &> \frac{\xi}{4} \left. \right\}. \end{aligned} \quad (\text{B.4})$$

For any given $\mathbf{s}_{2N}^{[1]}, \dots, \mathbf{s}_{2N}^{[M]}$, set Λ to be a $\xi/8$ -radius cover of $\mathcal{F}$ with respect to the $\ell_1(\mathbf{s}_{2N}^{[1]}, \dots, \mathbf{s}_{2N}^{[M]})$ norm. Since $\mathcal{F}$ is composed of bounded functions with range $[a, b]$, we assume that the same holds for any $h \in \Lambda$. According to the triangle inequality, if f_0 is the function that achieves

$$\sup_{f \in \mathcal{F}} \frac{1}{2M} \left| \langle \vec{\epsilon}, \vec{F}(\mathbf{s}_{2N}^{[1]}, \dots, \mathbf{s}_{2N}^{[M]}) \rangle \right| > \frac{\xi}{4}, \quad (\text{B.5})$$

there must be $h_0 \in \Lambda$ that satisfies

$$\begin{aligned} &\frac{1}{2M} \sum_{m=1}^M (|F_0(\mathbf{s}_N'^{[m]}) - H_0(\mathbf{s}_N'^{[m]})| \\ &+ |F_0(\mathbf{s}_N^{[m]}) - H_0(\mathbf{s}_N^{[m]})|) < \frac{\xi}{8}, \end{aligned} \quad (\text{B.6})$$

and meanwhile

$$\frac{1}{2M} \left| \langle \vec{\epsilon}, \vec{H}_0(\mathbf{s}_{2N}^{[1]}, \dots, \mathbf{s}_{2N}^{[M]}) \rangle \right| > \frac{\xi}{8}. \quad (\text{B.7})$$

Therefore, we arrive at

$$\begin{aligned} &\Pr \left\{ \sup_{f \in \mathcal{F}} \left| \frac{1}{2N} \langle \vec{\epsilon}, \vec{F}(\mathbf{s}_{2N}^{[1]}, \dots, \mathbf{s}_{2N}^{[M]}) \rangle \right| > \frac{\xi}{4} \right\} \\ &\leq \Pr \left\{ \sup_{h \in \Lambda} \left| \frac{1}{2N} \langle \vec{\epsilon}, \vec{H}(\mathbf{s}_{2N}^{[1]}, \dots, \mathbf{s}_{2N}^{[M]}) \rangle \right| > \frac{\xi}{8} \right\}. \end{aligned} \quad (\text{B.8})$$

On the other hand, given any $\xi > 0$, we also have, for any $M \geq 8(b-a)^2/\xi^2$,

$$\begin{aligned} &\Pr \left\{ \sup_{h \in \Lambda} \left| \frac{1}{2M} \langle \vec{\epsilon}, \vec{H}(\mathbf{s}_{2N}^{[1]}, \dots, \mathbf{s}_{2N}^{[M]}) \rangle \right| > \frac{\xi}{8} \right\} \\ &= \Pr \left\{ \sup_{h \in \Lambda} \left| \frac{1}{M} \langle \vec{\epsilon}, \vec{H}(\mathbf{s}_{2N}^{[1]}, \dots, \mathbf{s}_{2N}^{[M]}) \rangle \right| > \frac{\xi}{4} \right\} \\ &= \Pr \left\{ \sup_{h \in \Lambda} \frac{1}{M} \left| \sum_{m=1}^M H(\mathbf{s}_N'^{[m]}) - H(\mathbf{s}_N^{[m]}) \right| > \frac{\xi}{4} \right\} \\ &\leq \Pr \left\{ \frac{1}{M} \sum_{h \in \Lambda} \left| \sum_{m=1}^M (H(\mathbf{s}_N'^{[m]}) - H(\mathbf{s}_N^{[m]})) \right| > \frac{\xi}{4} \right\} \\ &\leq \Pr \left\{ \sum_{h \in \Lambda} \left| E\{H(\mathbf{s}_N)\} - \frac{1}{M} \sum_{m=1}^M H(\mathbf{s}_N^{[m]}) \right| \right. \\ &+ \left| E\{H(\mathbf{s}_N)\} - \frac{1}{M} \sum_{m=1}^M H(\mathbf{s}_N'^{[m]}) \right| > \frac{\xi}{4} \left. \right\} \leq 2 \\ &\cdot \Pr \left\{ \sum_{h \in \Lambda} \left| E\{H(\mathbf{s}_N)\} - \frac{1}{M} \sum_{m=1}^M H(\mathbf{s}_N^{[m]}) \right| > \frac{\xi}{8} \right\} \\ &\leq 4\mathcal{N}_1 \left(\mathcal{F}, \frac{\xi}{8}, 2MN \right) \\ &\cdot \exp \left(-\frac{M\xi^2}{128 \sum_{n=1}^N |U_n(F)|^2} \right). \end{aligned} \quad (\text{B.9})$$

The last inequality of (B.9) is derived from Definition (30) and Lemma 2.

The combination of (B.4), (B.8), and (B.9) leads to the result: given any $\xi > 0$, there holds that, for any $M \geq 8(b-a)^2/\xi^2$,

$$\begin{aligned} &\Pr \left\{ \sup_{f \in \mathcal{F}} \left| \frac{1}{N} \sum_{n=1}^N \int f(\mathbf{z}_{t_n}, t_n) dP_{t_n} - \frac{1}{MN} \sum_{n=1}^N \sum_{m=1}^M f(\mathbf{z}_{t_n}^{[m]}, t_n) \right| \right. \\ &> \xi \left. \right\} \leq 8\mathcal{N}_1 \left(\mathcal{F}, \frac{\xi}{8}, 2MN \right) \exp \left(-\frac{MN^2\xi^2}{128 \sum_{n=1}^N |U_n(F)|^2} \right). \end{aligned} \quad (\text{B.10})$$

This completes the proof.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgment

This work is supported by the National Natural Science Foundation of China under Project 11401076, Project 61473328, Project 11171367, and Project 61473059.

References

- [1] V. N. Vapnik, "An overview of statistical learning theory," *IEEE Transactions on Neural Networks*, vol. 10, no. 5, pp. 988–999, 1999.
- [2] E. Parrado-Hernández, A. Ambroladze, J. Shawe-Taylor, and S. Sun, "PAC-Bayes bounds with data dependent priors," *Journal of Machine Learning Research*, vol. 13, no. 1, pp. 3507–3531, 2012.
- [3] L. Oneto, A. Ghio, S. Ridella, and D. Anguita, "Local rademacher complexity: sharper risk bounds with and without unlabeled samples," *Neural Networks*, vol. 65, pp. 115–125, 2015.
- [4] X. Liu, S. Lin, J. Fang, and Z. Xu, "Is extreme learning machine feasible? A theoretical assessment (Part I)," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 26, no. 1, pp. 7–20, 2015.
- [5] S. Lin, X. Liu, J. Fang, and Z. Xu, "Is extreme learning machine feasible? A theoretical assessment (Part II)," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 26, no. 1, pp. 21–34, 2015.
- [6] C. Zhang and D. Tao, "Risk bounds of learning processes for lévy processes," *Journal of Machine Learning Research*, vol. 14, no. 1, pp. 351–376, 2013.
- [7] C. Zhang and D. Tao, "Generalization bounds of ERM-based learning processes for continuous-time markov chains," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 23, no. 12, pp. 1872–1883, 2012.
- [8] B. Zou, L. Li, and Z. Xu, "The generalization performance of ERM algorithm with strongly mixing observations," *Machine Learning*, vol. 75, no. 3, pp. 275–295, 2009.
- [9] W. Jiang, "On uniform deviations of general empirical risks with unboundedness, dependence, and high dimensionality," *Journal of Machine Learning Research*, vol. 10, pp. 977–996, 2009.
- [10] W. Hoeffding, "Probability inequalities for sums of bounded random variables," *Journal of the American Statistical Association*, vol. 58, no. 301, pp. 13–30, 1963.
- [11] B. Zou, Y. Y. Tang, Z. Xu, L. Li, J. Xu, and Y. Lu, "The generalization performance of regularized regression algorithms based on markov sampling," *IEEE Transactions on Cybernetics*, vol. 44, no. 9, pp. 1497–1507, 2014.
- [12] J. Xu, Y. Y. Tang, B. Zou, Z. Xu, L. Li, and Y. Lu, "The generalization ability of online svm classification based on markov sampling," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 26, no. 3, pp. 628–639, 2015.
- [13] P. L. Bartlett, S. Mendelson, and P. Philips, "Local complexities for empirical risk minimization," in *Learning Theory*, pp. 270–284, Springer, 2004.
- [14] A. Sutivong, M. Chiang, T. M. Cover, and Y.-H. Kim, "Channel capacity and state estimation for state-dependent Gaussian channels," *IEEE Transactions on Information Theory*, vol. 51, no. 4, pp. 1486–1495, 2005.
- [15] J. Ramsay, *Functional Data Analysis*, Wiley, 2005.
- [16] G. M. James, J. Wang, and J. Zhu, "Functional linear regression that's interpretable," *The Annals of Statistics*, vol. 37, no. 5A, pp. 2083–2108, 2009.
- [17] B. K. Sriperumbudur, K. Fukumizu, A. Gretton, B. Schölkopf, and G. R. G. Lanckriet, "On the empirical estimation of integral probability metrics," *Electronic Journal of Statistics*, vol. 6, pp. 1550–1599, 2012.
- [18] A. Müller, "Integral probability metrics and their generating classes of functions," *Advances in Applied Probability*, vol. 29, no. 2, pp. 429–443, 1997.
- [19] L. A. Kontorovich and K. Ramanan, "Concentration inequalities for dependent random variables via the martingale method," *The Annals of Probability*, vol. 36, no. 6, pp. 2126–2158, 2008.
- [20] K. Azuma, "Weighted sums of certain dependent random variables," *Tohoku Mathematical Journal, Second Series*, vol. 19, no. 3, pp. 357–367, 1967.
- [21] O. Bousquet, S. Boucheron, and G. Lugosi, "Introduction to statistical learning theory," in *Advanced Lectures on Machine Learning*, pp. 169–207, Springer, 2004.
- [22] S. Mendelson, "A few notes on statistical learning theory," in *Advanced Lectures on Machine Learning*, vol. 2600 of *Lecture Notes in Computer Science*, pp. 1–40, Springer, Berlin, Germany, 2003.