2023 62nd IEEE Conference on Decision and Control (CDC)
December 13-15, 2023. Marina Bay Sands, Singapore

Online Learning in Reproducing Kernel Hilbert Space With Non-IID Data

Xiwei Zhang,[1] and Tao Li[*,2]

Abstract— **We analyze the convergence of online regularized learning algorithm based on dependent and non-stationary online data streams for the nonparametric regression problem in reproducing kernel Hilbert space (RKHS). We show that the algorithm achieves mean-square convergence if the algorithm gain and regularization parameter are chosen appropriately, the online data streams are weakly dependent and satisfy the *eigenvalue-wise persistence of excitation* condition. Especially, for the case with independent but non-identically distributed online data streams, we give more intuitive convergence conditions on the drifts of the probability measures induced by the data.**

I. INTRODUCTION

Supervised statistical learning aims to reveal the fundamental laws of the learning process by training or learning datasets to efficiently approximate the mapping relationship between the input and output in a suitable hypothesis space, where the key concern is to control the complexity of the hypothesis space ([1]). The reproducing kernel Hilbert space (RKHS) provides a uniform processing framework for nonparametric regressions containing generalized smooth spline functions, real analytic functions with bounded bandwidth, and Gaussian processes ([2]-[3]). The convergence and optimal convergence rate of the batch learning algorithms with independent and identically distributed (i.i.d.) datasets under RKHS framework have been systematically investigated ([4]).

In recent years, with the widespread use of online learning in training deep neural networks ([5]), online learning has gained the attention of many scholars. Compared with batch learning, which requires processing the entire dataset at once, online learning only needs to process a single piece of data at a time and update the output in real time. Therefore, investigating online learning algorithms under the RKHS framework has gradually become a hot topic in supervised statistical learning. For the nonparametric regression problem in statistical learning, [6]-[9] have obtained rich research results on online learning algorithms based on i.i.d. data streams.

In fact, i.i.d. datasets are difficult to be obtained in many scenarios for which machine learning algorithms are applied, such as market prediction, system diagnosis, and speech recognition, which are all inherently temporal ([10]). Therefore, machine learning and statistics community have been devoted to the development of learning theory by weakening the special assumption of i.i.d. data for a long time, however, most valuable results in this direction are concentrated on batch supervised learning ([10]-[14]), and up till now, the statistical online learning with dependent sampling data is a valuable but unexplored area and the study of online learning algorithms based on non-i.i.d. data streams still remains open. On the one hand, unlike batch learning which can process all data at once, online learning algorithms receive very little information and can only process a single piece of data at a time. On the other hand, unlike online learning assuming i.i.d. data streams, dependent observations contain less information and therefore lead to more unstable learning errors as well as the performance degradation of learning compared with i.i.d. data ([13]-[14]). The best results at present remains the analysis of the convergence rate of online learning algorithms with samples drawn according to a non-identical sequence of probability distributions while maintaining independence ([15]-[16]). Based on independent but non-identically distributed online data streams, Smale and Zhou [15] first proposed the exponential convergence condition of marginal distribution and analyzed the performance of the online regularized learning algorithm. Subsequently, Hu and Zhou [16] proposed the polynomial convergence condition of marginal distribution and gave an analysis of the convergence rate of the online regularization algorithm with general loss function.

In this paper, an online regularized learning algorithm based on dependent and non-stationary online data streams is proposed under the RKHS framework for the nonparametric regression problem. Removing the assumption of independence poses an essential difficulty for the convergence analysis of online algorithms. Firstly, online learning algorithms are intrinsically obtained by the stochastic gradient descent method, and the dependent and non-stationary data cannot provide sufficient information about the true gradient, which leads to poor stability of the learning error. Secondly, since the observation data at different moments are no longer independent, the existing methods in [6]-[9] and [15]-[16], which construct martingale difference sequences by means of the property of independence and separates the operator products, thereby separating the information among coupled terms, are no longer applicable. It is worth noting that, in the past several decades, many scholars have proposed the persistence of excitation condition based on

[*]Author for correspondence. This work was supported in part by the National Natural Science Foundation of China under Grant 62261136550 and in part by the Basic Research Project of Shanghai Science and Technology Commission under Grant 20JC1414000.

[1]Xiwei Zhang is with the School of Mathematical Sciences, East China Normal University, Shanghai 200241, China. Email: xwzhangmath@sina.com

[2]Tao Li is with the Key Laboratory of Pure Mathematical Practice, School of Mathematical Sciences, East China Normal University, Shanghai 200241, China. Email: tli@math.ecnu.edu.cn

979-8-3503-0123-6/23/$31.00 ©2023 IEEE

the minimum eigenvalues of the conditional expectations of the information matrices in finite-dimensional space ([17]). The stochastic persistence of excitation condition was first proposed in the analysis of the Kalman filter algorithm by Guo in [18] and then refined in [19]-[22], which was proved to be necessary and sufficient for exponential stability. However, the above excitation conditions all require to some extent that the information matrix is positive definite, i.e. all the eigenvalues of the matrix have a common strictly positive lower bounds. This is not applicable for the statistical learning problems in RKHS, which is usually infinite-dimensional. The information operator induced by the data in RKHS is self-adjoint, however, even for strictly positive compact operators, the infimum of the infinite eigenvalues of the compact operator is zero.

To overcome the aforementioned difficulties caused by removing the i.i.d. assumption, by probability theory in Banach space, measure theory and stochastic time-varying system theories, we construct a class of sequences of martingale differences consisting of regularization paths without relying on the independence assumption. We develop a more general theory of regularization paths than those in [8] and [23] by proving the compactness and invertibility of the information operators induced by data, and establish the *eigenvalue-wise persistence of excitation* condition. Furthermore, the sufficient conditions for mean-square convergence of online learning algorithms based on dependent and non-stationary data streams are obtained for the first time. We prove that if the algorithm gain and regularization parameter are chosen appropriately and the online data streams satisfy the *eigenvalue-wise persistence of excitation* condition, i.e. each component of the sequence formed by the decreasing order of the eigenvalues of the covariance operators over a fixed-length time period has a positive lower bound, then the algorithm's estimation and regularization path asymptotically coincide in mean square sense, thus showing the algorithm converges in mean square. Especially, for the case with independent but non-identically distributed online data streams, we obtain more intuitive convergence conditions, where the convergence condition of the marginal distributions in [15]-[16] is relaxed to a restriction on the drifts of the probability measures induced by the data.

Notation and symbols: $\mathbb{R}^n$ denotes n dimensional real vector space. Let $\mathscr{L}(X)$ be a linear space consisting of all bounded linear operators mapping from the Banach space X to X. The eigenvalues of a compact operator A are denoted by $\{\lambda_i(A), i=1,2,...\}$, where $\lambda_i(A)$ is the i-th largest eigenvalue of A. For any random element ξ, $\mathbb{E}[\xi]$ denotes its mathematical expectation. The notation $b_n=\mathcal{O}(r_n)$ denotes $\lim_{n\to\infty}\sup\frac{|b_n|}{r_n}<\infty$, where $\{b_n, n\geq 0\}$ is a sequence of real numbers, $\{r_n, n\geq 0\}$ is a sequence of real positive numbers; $b_n=o(r_n)$ denotes $\lim_{n\to\infty}\frac{b_n}{r_n}=0$.

II. Online Regularized Learning in RKHS

A. Reproducing Kernel Hilbert Space

Let $\mathscr{X}$ be a subset of $\mathbb{R}^n$ and $K:\mathscr{X}\times\mathscr{X}\to\mathbb{R}$ be a *Mercer kernel*, i.e. a continuous symmetric real function which is *positive semi-definite* in the sense that $\sum_{i=1}^m\sum_{j=1}^m c_ic_jK(x_i,x_j)\geq 0$ for any $m\geq 1$ and any choice of $x_i\in\mathscr{X}$ and $c_i\in\mathbb{R}$ $(i=1,\cdots,m)$. A Mercer kernel K induces a function $K_x:\mathscr{X}\to\mathbb{R}$ $(x\in\mathscr{X})$ defined by $K_x(x')=K(x,x')$. Let $\mathscr{H}_K$ be the *reproducing kernel Hilbert space* (RKHS) associated with a Mercer kernel K, i.e. the completion of *Span*$\{K_x, x\in\mathscr{X}\}$ with respect to the inner product, defined as the linear extension of the bilinear form $\langle K_x, K_{x'}\rangle_K=K(x,x')$, $\forall x,x'\in\mathscr{X}$. The norm of $\mathscr{H}_K$ is denoted by $\|f\|_K=\sqrt{\langle f,f\rangle_K}$ for each $f\in\mathscr{H}_K$. The most important property of RKHS is the *reproducing property*: for all $f\in\mathscr{H}_K$ and $x\in\mathscr{X}$, $f(x)=\langle f,K_x\rangle_K$.

B. Problem Formulation

Throughout this paper, $(\Omega,\mathcal{F},\mathbb{P})$ is assumed to be a complete probability space. Suppose that $f^\star:\mathscr{X}\to\mathbb{R}$ is the unknown function in $\mathscr{H}_K$. The nonparametric regression model at instant k is given by

$$y(k)=f^\star(x(k))+v(k),\ k\geq 0, \tag{1}$$

where $x(k):\Omega\to\mathscr{X}$ is a random vector at instant k, called the random input data, and the observation noise $v(k):\Omega\to\mathbb{R}$ is a random variable at instant k. Online learning aims to construct the approximation of the unknown function $f^\star$ using only the current observation data $(x(k),y(k))$.

Denote the σ-field $\mathcal{F}(k)=\sigma(x(i),v(i),0\leq i\leq k)$, $k\geq 0$ with $\mathcal{F}(-1)=\{\Omega,\emptyset\}$. For the regression model (1) and the kernel function K which determines the Hilbert space $\mathscr{H}_K$, we need the following assumptions.

Assumption 1: The sequence $\{v(k),\mathcal{F}(k),k\geq 0\}$ is a martingale difference sequence, which is independent of the sequence $\{x(k),k\geq 0\}$, and there exists a constant $\beta>0$, such that $\sup_{k\geq 0}\mathbb{E}[v^2(k)|\mathcal{F}(k-1)]\leq\beta$ a.s.

Assumption 2: $\sup_{x\in\mathscr{X}}K(x,x)<\infty$.

Let $I\in\mathscr{L}(\mathscr{H}_K)$ be the identical operator. Define operator $g\otimes h: f\mapsto\langle f,h\rangle_K g$. It follows from Assumption 2 that $K_{x(k)}\otimes K_{x(k)}$ is a random element with values in $\mathscr{L}(\mathscr{H}_K)$, $\forall k\geq 0$, which thus is Bochner integrable.

Definition 1: The linear operator $\Sigma_k: g\mapsto\Sigma_k g$, $\forall g\in\mathscr{H}_K$, is called the *covariance operator* of data $x(k)$, where

$$\langle f,\Sigma_k g\rangle_K=\int_\Omega f(x(k))g(x(k))\mathrm{d}\mathbb{P},\ \forall f,g\in\mathscr{H}_K,\ \forall k\geq 0.$$

Remark 1: By Riesz' representation theorem, Σ_k is well-defined. Using the reproducing property, we have $\Sigma_k=\mathbb{E}[K_{x(k)}\otimes K_{x(k)}]$, $\forall k\geq 0$, where the mathematical expectation is formally defined as a Bochner integration. In finite-dimensional space $\mathscr{H}_K=\mathbb{R}^n$, for $g,h\in\mathbb{R}^n$, we have $g\otimes h=gh^\top\in\mathbb{R}^{n\times n}$ since for any f, $(gh^\top)f=(h^\top f)g=\langle f,h\rangle_K g$. Thus in finite-dimensional space, $\Sigma_k=\mathbb{E}[x(k)x^\top(k)]$ is the auto-correlation matrix of $x(k)$.

It follows from Assumption 2 and [24] that the conditional mathematical expectation of $K_{x(k)}\otimes K_{x(k)}$ with respect to the σ-field $\mathcal{F}(m)$ exists uniquely, which is denoted by

$$\Sigma_{k|m}\triangleq\mathbb{E}\left[K_{x(k)}\otimes K_{x(k)}\middle|\mathcal{F}(m)\right],\ \forall k\geq 0,\ \forall m\geq -1.$$

Note that $\Sigma_{k|m} = \Sigma_k$ if $x(k)$ is independent of $\mathcal{F}(m)$. By Arzela-Ascoli theorem and the spectral decomposition of compact operator, we have the following propositions of $\Sigma_{k|m}$, whose proofs are omitted.

Proposition 1: If Assumption 2 holds, then $\Sigma_{k|m} : \Omega \to \mathscr{L}(\mathscr{H}_K)$ is a self-adjoint, positive and compact operator a.s., $\forall\ k \geq 0$, $\forall\ m \geq -1$.

Proposition 2: If Assumption 2 holds, then $\Sigma_{k|m} + \lambda I$ is invertible a.s., $\forall\ \lambda > 0$, $\forall\ k \geq 0$, $\forall\ m \geq -1$.

C. Regularization Path and Stochastic Gradient Algorithms

The Hilbert space $\mathscr{H}_K$ is generally an infinite-dimensional functional space with high complexity, regularization is necessary and the following Tikhonov regularization is widely adopted ([4]). Let, for all $k \geq 0$, $f_\lambda(k)$ be the solution of the regularized least square problem in $\mathscr{H}_K$,

$$\arg \min_{f \in \mathscr{H}_K} \int_\Omega (y(k) - f(x(k))^2 \mathrm{d}\mathbb{P} + \lambda(k) \|f\|_K^2, \tag{2}$$

where $y(k)$ is given in (1) and $\lambda(k) > 0$ is the regularization parameter. Depending on the assumptions on the covariance operator Σ_k, we will show that $f_\lambda(k)$ converges to $f^\star$ in mean square as $k \to \infty$.

Definition 2: The map $P_{f^\star} : k \mapsto f_\lambda(k)$ is called the *regularization path* of $f^\star$ at instant k in $\mathscr{H}_K$.

Remark 2: For online learning with i.i.d. data, Tarrès and Yao [8] studied the regularization path concerning with time-invariant covariance operator. It is worth noting that the covariance operator in the regularization path defined by Definition 2 can be time-varying, and is applicable to the non-i.i.d. case consequently.

The following proposition gives an explicit form of the regularization path, whose proof is omitted.

Proposition 3: If Assumptions 1-2 hold, then $f_\lambda(k) = (\Sigma_k + \lambda(k)I)^{-1}\Sigma_k f^\star$, $\forall k \geq 0$.

To solve the regularized least square problem (2), by the reproducing property and Assumptions 1-2, we can see that

$$\begin{aligned} &\operatorname{grad} \int_\Omega (y(k) - f(x(k))^2 \mathrm{d}\mathbb{P} + \lambda(k)\|f\|_K^2 \\ &= 2 \int_\Omega \left[(f(x(k)) - y(k)) K_{x(k)} + \lambda(k) f \right] \mathrm{d}\mathbb{P},\ \forall k \geq 0. \end{aligned}$$

Through the stochastic gradient descent (SGD) method, we obtain the online regularized learning algorithm in $\mathscr{H}_K$,

$$f_{k+1} = f_k - a(k) \left((f_k(x(k)) - y(k)) K_{x(k)} + \lambda(k) f_k \right) \tag{3}$$

with deterministic initial value $f_0 \in \mathscr{H}_K$, where $\{a(k), k \geq 0\}$ and $\{\lambda(k), k \geq 0\}$ are the *gain sequence* and *regularization sequence*, respectively. The following conditions may be needed later.

Condition 1: The gain sequence $\{a(k), k \geq 0\}$ and regularization sequence $\{\lambda(k), k \geq 0\}$ are all positive sequences monotonically decreasing to zero.

Condition 2: The gain sequence $\{a(k), k \geq 0\}$ and regularization sequence $\{\lambda(k), k \geq 0\}$ satisfy $\sum_{k=0}^\infty a(k)\lambda(k) = \infty$ and $a(k) = o(\lambda(k))$.

Condition 3: The gain sequence $\{a(k), k \geq 0\}$ and regularization sequence $\{\lambda(k), k \geq 0\}$ satisfy $\lambda(k) - \lambda(k+1) = \mathcal{O}(a(k)\lambda^2(k))$.

Remark 3: If $a(k) = (k+1)^{-\tau}$ and $\lambda(k) = (k+1)^{\tau-1}$, $k \geq 0$, $\tau \in (0.5, 1)$, then Conditions 1-3 hold since

$$\frac{\lambda(k) - \lambda(k+1)}{a(k)\lambda^2(k)} = (k+1)\left(1 - \left(\frac{k+1}{k+2}\right)^{1-\tau}\right) \to 1 - \tau,$$

as $k \to \infty$.

III. Main Results

To establish the online learning theory with non-i.i.d. data, we introduce the following *eigenvalue-wise persistence of excitation* (P.E.) condition, which is an indispensable part of the convergence analysis of the algorithm.

Condition 4 (eigenvalue-wise P.E. condition): There exists an integer $h \geq 0$, such that

$$\inf_{k \geq 0} \lambda_i \left(\sum_{j=k}^{k+h} \Sigma_j \right) > 0,\ i = 1, 2, ...$$

Remark 4: Condition 4 ensures that the regularization path $f_\lambda(k)$ converges to $f^\star$ in mean square, which does not require that the data streams be stationary. To learn the unknown element $f^\star$ under valid measurement information, Condition 4 requires the distributions of data streams to have the *persistence of excitation* property: for any given positive integer i, the i-th eigenvalues of the covariance operators over a fixed length time period have a positive lower bound, i.e. $\inf_{k\geq 0} \lambda_i(\sum_{j=k}^{k+h} \Sigma_j) > 0$, where the i-th eigenvalue of the covariance operator at each instant is not necessarily required to have a positive lower bound, i.e. $\inf_{k \geq 0} \lambda_i(\Sigma_k) > 0$. If the online data streams are independent and identically distributed, then the eigenvalue-wise P.E. condition degenerates to requiring $\Sigma_0 = \mathbb{E}[K_{x(0)} \otimes K_{x(0)}]$ to be strictly positive, i.e. $\lambda_i(\Sigma_0) > 0$, $\forall\ i \geq 1$, which is exactly the case in [8].

In the past decades, to solve the problems of parameter estimation and signal tracking with non-stationary and non-independent observation matrices, many scholars have proposed persistence of excitation conditions based on the minimum eigenvalues of the conditional expectations of the information matrices in finite-dimensional space. The stochastic persistence of excitation condition was first proposed by Guo [18] in the analysis of the Kalman filtering algorithm and then refined in [19]-[22], which was proved to be necessary and sufficient for exponential stability. However, the stochastic persistence of excitation conditions proposed for finite-dimensional systems all require to some extent that the information matrix is positive definite, i.e. the eigenvalues of the matrix have strictly positive lower bounds. This is not applicable for the statistical learning problems in infinite-dimensional RKHS. The information operator induced by the data in RKHS is self-adjoint, even for strictly positive compact operators, all the persistence of excitation conditions in finite-dimensional space cannot hold, due to the infimum of the infinite eigenvalues of the compact operator is zero.

Based on the algorithm, assumptions and conditions established above, the convergence analysis of the algorithm (3) are presented in this section. The proofs of Lemma 1 and Theorem 1 are given in Section IV, and the proof of Corollary 1 is omitted.

Denote

$$\varphi(k) \triangleq \sup_{\substack{u(k)\in\mathcal{F}(k-1)\\ \|u(k)\|_K=1}} \mathbb{E}\left[\left\|\left(\Sigma_k-\Sigma_{k|k-1}\right)u(k)\right\|_K^2\right]^{\frac{1}{2}}.$$

Let $\delta(k)=f_k-f_\lambda(k)$ be the tracking error, which is used to measure the deviation of the algorithm from following the regularization path. The asymptotic analysis of the tracking error is obtained in the following lemma, which plays a key role in the convergence analysis of the algorithm.

Lemma 1: Suppose that Assumptions 1-2 hold. For the algorithm (3), assume that Conditions 1 and 4 hold. If $\varphi(k)=\mathcal{O}(\lambda^2(k))$ and

(A) $\displaystyle\sum_{i=0}^{\infty} a(i)\lambda(i)=\infty$;

(B) $\displaystyle\lim_{k\to\infty}\sum_{i=0}^{k} a^2(i)\prod_{j=i+1}^{k}(1-a(i)\lambda(i))^2=0$;

(C) $\displaystyle\lim_{k\to\infty}\sum_{i=0}^{k}\|f_\lambda(i+1)-f_\lambda(i)\|_K\prod_{j=i+1}^{k}(1-a(i)\lambda(i))=0$,

then $\lim_{k\to\infty}\mathbb{E}[\|\delta(k)\|_K^2]=0$.

Remark 5: Online learning with dependent data usually lead to poor stability of learning error due to lack of information. According to the *no free lunch* principle, it is necessary to make some assumptions on the dependence among data. The condition $\varphi(k)=\mathcal{O}(\lambda^2(k))$ in Lemma 1 requires the data streams to satisfy only a certain weak dependence condition instead of the temporal-independent condition. Intuitively, for the purpose of obtaining enough information from the data as time going, the prediction of the "future" $\Sigma_{k|k-1}$ given the "past" data $\{x(i),0\le i\le k-1\}$ is required to be consistent with Σ_k in mean-square sense. Especially, the independent sequence $\{x(k),k\ge 0\}$ contains sufficiently enough information, which satisfies $\varphi(k)\equiv 0,\ \forall k\ge 0$.

Theorem 1: Suppose that Assumptions 1-2 hold. For the algorithm (3), assume that Conditions 1-2 and 4 hold. If $\varphi(k)=\mathcal{O}(\lambda^2(k))$ and

(D) $\displaystyle\lim_{k\to\infty}\frac{a(k)\lambda(k)}{\|f_\lambda(k+1)-f_\lambda(k)\|_K}=\infty$,

then $\lim_{k\to\infty}\mathbb{E}[\|f_k-f^\star\|_K^2]=0$.

Remark 6: The choice of the gain sequence and regularization sequence is crucial for the algorithm to successfully learn the unknown element $f^\star$. To reduce the effect of measurement noise, Condition 1 requires the algorithm avoid making excessive changes to the current estimate when acquiring new noisy data. Condition 2 requires $\{a(k)\lambda(k),k\ge 0\}$ not to be too small for the convergence of the algorithm, which is often used in the stochastic approximation algorithms to drive the estimations to the unknown true element from arbitrary initial conditions. Condition (D) in Theorem 1 implies that the drifts along regularization path drop faster than $a(k)\lambda(k)$, under which the learning sequence $\{f_k,k\ge 0\}$ can follow the regularization path $\{f_\lambda(k),k\ge 0\}$.

We are now in position to consider a special setting where the assumption of i.i.d. data is weakened by keeping the independence but abandoning the identical restriction. Denote $\rho_{\mathscr{X}}(k)\triangleq\mathbb{P}\circ x^{-1}(k)$ the probability measure induced by the random data $x(k):\Omega\to\mathscr{X},\ \forall k\ge 0$. The Hölder space $C^s(\mathscr{X})$, $0\le s\le 1$ is defined by

$$C^s(\mathscr{X})\triangleq\left\{f\in C(\mathscr{X}):\|f\|_\infty+\|f\|_{C^s(\mathscr{X})}<\infty\right\},\quad (4)$$

where $\|f\|_\infty=\sup_{x\in\mathscr{X}}|f(x)|$ and $|f|_{C^s(\mathscr{X})}=\sup_{x\ne y\in\mathscr{X}}\frac{|f(x)-f(y)|}{\|x-y\|_{\mathbb{R}^n}^s}$, which is a Banach space. For any given probability measure ρ on $\mathscr{X}$, it follows from [15] that $\rho\in(C^s(\mathscr{X}))^*$, i.e. ρ is a bounded linear functional on $C^s(\mathscr{X})$. With the above settings, we need the following assumption on the kernel function.

Assumption 3: The kernel function $K\in C^2(\mathscr{X}\times\mathscr{X})$.

In fact, Assumption 3 guarantees the fact that $\mathscr{H}_K$ is included in $C^s(\mathscr{X})$ ([15]), under which we have the following corollary with independent but non-identical distributed data.

Corollary 1: Suppose that Assumptions 1 and 3 hold. For the algorithm (3), assume that Conditions 1-4 hold. If the online data stream is an independent sequence satisfying

(E) $\|\rho_{\mathscr{X}}(k+1)-\rho_{\mathscr{X}}(k)\|_{(C^s(\mathscr{X}))^*}=\mathcal{O}\left(a(k)\lambda^2(k)\right)$,

then $\lim_{k\to\infty}\mathbb{E}[\|f_k-f^\star\|_K^2]=0$.

Remark 7: For online learning with independent but non-stationary data, the convergence depends largely on the measure sequence $\{\rho_{\mathscr{X}}(k),k\ge 0\}$ induced by the data. For this reason, Smale and Zhou [15] proposed the exponential convergence condition of marginal distribution: there exists a probability measure $\rho_{\mathscr{X}}$ and a constant $\alpha\in(0,1)$, such that $\|\rho_{\mathscr{X}}(k)-\rho_{\mathscr{X}}\|_{(C^s(\mathscr{X}))^*}=\mathcal{O}(\alpha^k)$. Subsequently, Hu and Zhou [16] improved above condition by proposing the polynomial convergence condition of marginal distribution: there exists a probability measure $\rho_{\mathscr{X}}$ and a constant $b>0$ such that $\|\rho_{\mathscr{X}}(k)-\rho_{\mathscr{X}}\|_{(C^s(\mathscr{X}))^*}=\mathcal{O}(k^{-b})$. In Corollary 1, neither the exponential convergence condition nor the polynomial convergence condition of marginal distribution is needed, instead of that, Condition (E) in Corollary 1 only requires the drifts of the measures induced by the data drop faster than $a(k)\lambda^2(k)$. To our best knowledge, we have obtained the most general results ever, even for online learning with independent and non-stationary data.

IV. Proofs of Main Results

The proofs of main results need the following lemmas, whose proofs are omitted.

Lemma 2: ([25]) Assume that $\{s_1(k), k \geq 0\}$ and $\{s_2(k), k \geq 0\}$ are real sequences satisfying $0 \leq s_2(k) < 1$, $\sum_{k=0}^{\infty} s_2(k) = \infty$ and $\lim_{k\to\infty} \frac{s_1(k)}{s_2(k)}$ exists. Then $\lim_{k\to\infty} \sum_{i=1}^{k} s_1(k) \prod_{j=i+1}^{k} (1 - s_2(k)) = \lim_{k\to\infty} \frac{s_1(k)}{s_2(k)}$.

Lemma 3: Suppose that Assumptions 1-2 hold. For the algorithm (3), assume that Condition 1 holds. If the gain sequence and regularization sequence satisfy

$$\lim_{k\to\infty} \sum_{i=0}^{k} a^2(i) \prod_{j=i+1}^{k} (1 - a(j)\lambda(j))^2 = 0,$$

then $\sup_{k\geq 0} \mathbb{E}[\|f_k - f^\star\|_K^2] < \infty$.

Proof of Lemma 1. Denote $H_k \triangleq K_{x(k)} \otimes K_{x(k)}$. Subtracting $f_\lambda(k)$ from both sides of (3) at the same time gives

$$\begin{aligned} &\delta(k+1) \\ &= (I - a(k)(H_k + \lambda(k)I))\delta(k) - (f_\lambda(k+1) - f_\lambda(k)) \\ &\quad - a(k)((H_k + \lambda(k)I)f_\lambda(k) - H_k f^\star) + a(k)v(k)K_{x(k)}, \end{aligned} \quad (5)$$

Denote $g_\lambda(k) \triangleq (\Sigma_{k|k-1} + \lambda(k)I)^{-1}\Sigma_{k|k-1} f^\star$. Noting that $\delta(0) = f_0 - f_\lambda(0) \in \mathscr{H}_K$, it follows from the tracking error equation (5) and Cauchy-Schwarz inequality that

$$\begin{aligned} &\frac{1}{7}\mathbb{E}\left[\|\delta(k+1)\|_K^2\right] \\ &\leq \|\Phi(k,0)\delta(0)\|_K^2 \\ &\quad + \mathbb{E}\left[\left\|\sum_{i=0}^{k} a(i)\Phi(k,i+1)\left(\Sigma_{i|i-1} - H_i\right)\delta(i)\right\|_K^2\right] \\ &\quad + \mathbb{E}\left[\left\|\sum_{i=0}^{k} a(i)\Phi(k,i+1)\left(\Sigma_i - \Sigma_{i|i-1}\right)\delta(i)\right\|_K^2\right] \\ &\quad + \mathbb{E}\left[\left\|\sum_{i=0}^{k} a(i)\Phi(k,i+1)v(i)K_{x(i)}\right\|_K^2\right] + \mathbb{E}\left[\left\|\sum_{i=0}^{k} a(i)\right.\right. \\ &\qquad \left.\left. \times \Phi(k,i+1)\left((H_i + \lambda(i)I)\, g_\lambda(i) - H_i f^\star\right)\right\|_K^2\right] \\ &\quad + \mathbb{E}\left[\left\|\sum_{i=0}^{k} a(i)\Phi(k,i+1)\right.\right. \\ &\qquad \left.\left. \times \left(H_i + \lambda(i)I\right)\left(g_\lambda(i) - f_\lambda(i)\right)\right\|_K^2\right] \\ &\quad + \left\|\sum_{i=0}^{k} \Phi(k,i+1)(f_\lambda(i+1) - f_\lambda(i))\right\|_K^2, \ \forall\, k \geq 0, \end{aligned} \quad (6)$$

where $\Phi(i,j) = \prod_{k=j}^{i}(I - a(k)(\Sigma_k + \lambda(k)I))$ if $i \geq j$ and $\Phi(i,j) = I$ if $i < j$. For notational convenience, the operator norm $\|\cdot\|_{\mathscr{L}(\mathscr{H}_K)}$ will be abbreviated as $\|\cdot\|$ in the sequel. Denote the terms on the right-hand side of the inequality (6) in turn as $A_i(k),\ i = 1, \cdots, 7$, which will be analyzed term by term. By Condition 1 and Assumption 2, there exists a constant $C_1 > 0$, such that

$$\|A_1(k)\|_K^2 \leq C_1 \prod_{i=i_0}^{k}(1 - a(i)\lambda(i)),\ \forall k \geq 0, \quad (7)$$

where $i_0 = \min\{k \geq 0 : a(k)\lambda(k) + a(k)\sup_{x\in\mathscr{X}} K(x,x) < 1\}$. Noting that $\delta(k) \in \mathcal{F}(k-1)$, it follows from the properties of conditional expectations that $\mathbb{E}[\langle \Phi(k,i+1)(\Sigma_{i|i-1} - H_i)\delta(i), \Phi(k,j+1)(\Sigma_{j|j-1} - H_j)\delta(j)\rangle_K] = 0,\ \forall 0 \leq i < j \leq k$, which together with Assumption 2 shows that there exist constants $C_2 > 0$ and $C_3 > 0$, such that

$$\begin{aligned} A_2(k) &\leq C_2 \prod_{j=i_0}^{k}(1 - a(j)\lambda(j)) + C_3 \sum_{i=i_0}^{k} a^2(i)\mathbb{E}\left[\|\delta(i)\|_K^2\right] \\ &\quad \times \prod_{j=i+1}^{k}(1 - a(j)\lambda(j))^2,\ \forall\, k \geq 0. \end{aligned} \quad (8)$$

We can see from Lemma 3 and Cauchy-Schwarz inequality that $\sup_{k\geq 0}\mathbb{E}[\|\delta(k)\|_K^2] < \infty$. Noting that $\varphi(k) = \mathcal{O}(\lambda^2(k))$, by Condition 1 and Minkowski inequality, we know that there exists a constant $C_4 > 0$, such that

$$\begin{aligned} A_3(k) &\leq C_4 \prod_{j=i_0}^{k}(1 - a(j)\lambda(j)) + \sum_{i=i_0}^{k} a(i)\varphi(i) \\ &\quad \times \left(\mathbb{E}\left[\|\delta(i)\|_K^2\right]\right)^{\frac{1}{2}} \prod_{j=i+1}^{k}(1 - a(j)\lambda(j)),\ \forall\, k \geq 0. \end{aligned} \quad (9)$$

By Assumption 1 and the properties of conditional expectations, we get $\mathbb{E}[\langle \Phi(k,i+1)v(i)K_{x(i)}, \Phi(k,j+1)v(j)K_{x(j)}\rangle_K] = 0$, where $0 \leq i < j \leq k$. Thus, by Assumptions 1-2, we know that there exist constants $C_5 > 0$ and $C_6 > 0$ satisfying

$$\begin{aligned} A_4(k) &\leq C_5 \prod_{j=i_0}^{k}(1 - a(j)\lambda(j)) \\ &\quad + C_6 \sum_{i=i_0}^{k} a^2(i) \prod_{j=i+1}^{k}(1 - a(j)\lambda(j))^2,\ \forall k \geq 0. \end{aligned} \quad (10)$$

Noting that $\lambda(k)g_\lambda(k) = \Sigma_{k|k-1}(f^\star - g_\lambda(k))$ a.s., $\forall k \geq 0$ and $g_\lambda(k) \in \mathcal{F}(k-1)$, by the properties of conditional expectations, we get $\mathbb{E}[\langle g(i), g(j)\rangle_K] = 0,\ \forall 0 \leq i < j \leq k$, where $g(k) \triangleq (H_k + \lambda(k)I)g_\lambda(k) - H_k f^\star$. It follows from Assumption 2 that there exist constants $C_7 > 0$ and $C_8 > 0$, such that

$$\begin{aligned} A_5(k) &\leq C_7 \prod_{j=i_0}^{k}(1 - a(j)\lambda(j)) \\ &\quad + C_8 \sum_{i=i_0}^{k} a^2(i) \prod_{j=i+1}^{k}(1 - a(j)\lambda(j))^2,\ \forall\, k \geq 0. \end{aligned} \quad (11)$$

It follows from Assumption 2 that $\sup_{k\geq 0}\|f^\star - f_\lambda(k)\| < \infty$, which together with $\varphi(k) = \mathcal{O}(\lambda^2(k))$ gives $\mathbb{E}[\|g_\lambda(k) - f_\lambda(k)\|_K^2]^{\frac{1}{2}} \leq \lambda^{-1}(k)\mathbb{E}[\|(\Sigma_{k|k-1} - \Sigma_k)(f^\star - f_\lambda(k))\|_K^2]^{\frac{1}{2}} \leq 2\varphi(k)\lambda^{-1}(k)\|f^\star\|_K$ a.s., $\forall k \geq 0$. Thus, by Condition 1 and Minkowski inequality, there exist constants $C_9 > 0$ and $C_{10} > 0$ satisfying

$$A_6(k) \leq C_9 \prod_{j=i_0}^{k}(1 - a(j)\lambda(j))$$

$$+C_{10}\sum_{i=i_0}^{k}\frac{a(i)\varphi(i)\|f_\lambda(i)-f^\star\|_K}{\lambda(i)}\prod_{j=i+1}^{k}(1-a(j)\lambda(j)).\tag{12}$$

By Condition 1 and Assumption 2, there exists a constant $C_{11} > 0$, such that

$$\begin{aligned}&\left\|\sum_{i=0}^{k}\Phi(k,i+1)(f_\lambda(i+1)-f_\lambda(i))\right\|_K\\ &\le C_{11}\prod_{j=i_0}^{k}(1-a(j)\lambda(j))\\ &+\sum_{i=i_0}^{k}\|f_\lambda(i+1)-f_\lambda(i)\|_K\prod_{j=i+1}^{k}(1-a(j)\lambda(j)).\end{aligned}\tag{13}$$

Hence, by (6)-(13) and Lemma 3, we obtain

$$\begin{aligned}&\mathbb{E}\left[\|\delta(k+1)\|_K^2\right]\\ \le{}& 7C_{12}\prod_{j=i_0}^{k}(1-a(j)\lambda(j))+7\sum_{i=i_0}^{k}a^2(i)\left(C_3\mathbb{E}\left[\|\delta(i)\|_K^2\right]\right.\\ &\left.+C_6+C_8\right)\prod_{j=i+1}^{k}(1-a(j)\lambda(j))^2+7\sum_{i=i_0}^{k}\left(a(i)\varphi(i)\right.\\ &\times\left(\mathbb{E}\left[\|\delta(i)\|_K^2\right]\right)^{\frac{1}{2}}+C_{10}\frac{a(i)\varphi(i)\|f_\lambda(i)-f^\star\|_K}{\lambda(i)}\\ &\left.+\|f_\lambda(i+1)-f_\lambda(i)\|_K\right)\prod_{j=i+1}^{k}(1-a(j)\lambda(j)),\end{aligned}\tag{14}$$

where $C_{12} \triangleq C_1 + C_2 + C_4 + C_5 + C_7 + C_9 + C_{11}$. Noting that $\varphi(k) = \mathcal{O}(\lambda^2(k))$, by Condition 4 and Lemma 2, it can be proved that $\sum_{i=i_0}^{k}\frac{a(i)\varphi(i)\|f_\lambda(i)-f^\star\|_K}{\lambda(i)}\prod_{j=i+1}^{k}(1-a(j)\lambda(j)) \to 0$, as $k \to \infty$. It follows from Condition (A) that $\prod_{j=i_0}^{k}(1-a(j)\lambda(j)) \to \infty$ as $k \to \infty$. By Assumption 2 and Lemma 3, we have $\sup_{k\ge 0}\mathbb{E}[\|\delta(k)\|_K^2] < \infty$, which together with Conditions (B)-(C) and (14) gives $\lim_{k\to\infty}\mathbb{E}[\|\delta(k)\|_K^2] = 0$.

Proof of Theorem 1. It follows from Conditions 1-2, Condition (D) in Theorem 1 and Lemma 2 that Conditions (A)-(C) in Lemma 1 hold, thus by Lemma 1, we get $\lim_{k\to\infty}\mathbb{E}[\|f_k - f_\lambda(k)\|_K^2] = 0$. It follows from Conditions 2, 4 and Condition (D) in Theorem 1 that $\lim_{k\to\infty}\|f_\lambda(k) - f^\star\|_K^2 = 0$. Thus, we have $\lim_{k\to\infty}\mathbb{E}[\|f_k - f^\star\|_K^2] = 0$.

V. Conclusion

We analyze the convergence of online regularized learning algorithm based on dependent and non-stationary online data streams for the nonparametric regression problem in reproducing kernel Hilbert space (RKHS). We show that the algorithm achieves mean-square convergence if the algorithm gain and regularization parameter are chosen appropriately, the online data streams are weakly dependent and satisfy the *eigenvalue-wise persistence of excitation* condition. Especially, for the case with independent but non-identically distributed online data streams, we give more intuitive convergence conditions on the drifts of the probability measures induced by the data.

References

[1] V. N. Vapnik, *Adaptive and Learning Systems for Signal Processing, Communications, and Control*. John Wiley & Sons Inc., New York, 1998.

[2] E. Parzen, "An approach to time series analysis," *Ann. Math. Statist.*, vol. 32, no. 4, pp. 951-989, 1961.

[3] G. Wahba, *Spline Models for Observational Data*. PA, USA: SIAM, 1990.

[4] S. Smale and F. Cucker, "On the mathematical foundations of learning," *Bulletin of the American Mathematical Society*, vol. 39, no. 1, pp. 1-49, 2001.

[5] I. Sutskever, J. Martens, G. Dahl, and G. Hinton, "On the importance of initialization and momentum in deep learning", *Proceedings of* 30^{th} *International Conference on Machine Learning, PMLR*, vol. 28, no. 3, pp. 1139-1147, Atlanta, Georgia, USA, May 26, 2013.

[6] S. Smale and Y. Yao, "Online learning algorithms", *Found. Comput. Math.*, vol. 6, no. 2, pp. 145-170, 2006.

[7] Y. Ying and M. Pontil, "Online gradient descent learning algorithms", *Found. Comput. Math.*, vol. 5, no. 5, pp. 561-596, 2008.

[8] P. Tarrès and Y. Yao, "Online learning as stochastic approximation of regularization paths", *IEEE Trans. Information Theory*, vol. 60, no. 99, pp. 5716-5735, 2014.

[9] A. Dieuleveut and F. Bach, "Nonparametric stochastic approximation with large step-sezes", *The Annals of Statistics*, vol. 44, no. 4, pp. 1363-1399, 2016.

[10] I. Steinwart, D. Hush, and C. Scovel, "Learning from dependent observations", *J. Multivariate Anal.*, vol. 100, no. 1, pp. 175-194, 2009.

[11] A. Agarwal and J. C. Duchi, "The generalization ability of online algorithms for dependent data", *IEEE Trans. Information Theory*, vol. 59, no. 1, pp. 573-587, 2013.

[12] M. J. Zhang and H. W. Sun, "Regression learning with non-identically and non-independently sampling", *International Journal of Wavelets, Multiresolution and Information Processing*, vol. 15, no. 1, 2017.

[13] H. W. Sun and Q. Wu, "Regularized least square regression with dependent samples", *Advances in Computational Mathematics*, vol. 32, no. 2, pp. 175-189, 2010.

[14] Z. Zhang, Y. Zhang, D. Guo, S. Zhao, and X. L. Zhu, "Communication-efficient federated continual learning for distributed learning system with Non-IID data", *SCIENCE CHINA Information Sciences*, vol. 66, no. 2, pp. 122102:1-122102:20, 2023.

[15] S. Smale and D. -X. Zhou, "Online learning with Markov sampling", *Analysis and Applications*, vol. 7, no. 1, pp. 87-113, 2009.

[16] T. Hu and D. -X. Zhou, "Online learning with samples drawn from non-identical distributions", *Journal of Machine Learning Research*, vol. 10, no. 12, pp. 2873-2898, 2009.

[17] Green, M., and J. B. Moore, "Persistency of excitation in linear systems", *Syst. Control Lett.*, vol 7. no. 5, pp. 351-360, 1986.

[18] L. Guo, "Estimating time-varying parameters by Kalman filter based algorithm: Stability and convergence", *IEEE Trans. Autom. Control*, vol. 35, no. 2, pp. 141-147, 1990.

[19] J. F. Zhang, L. Guo, and H. F. Chen, "Lp-stability of estimation errors of Kalman filter for tracking time-varying parameters", *Int. J. Adaptive Control and Signal Processing*, vol. 5, pp. 155-174, 1991.

[20] L. Guo, "Stability of recursive stochastic tracking algorithms", *SIAM J. Control Optim.*, vol. 32, no. 5, pp. 1195-1225, 1994.

[21] L. Guo and L. Ljung, "Performance analysis of general tracking algorithms", *IEEE Trans. Autom. Control*, vol. 40, no. 8, pp. 1388-1402, 1995.

[22] L. Guo and L. Ljung, and G. J. Wang, "Necessary and sufficient conditions for stability of LMS", *IEEE. Trans. Autom. Control*, vol. 42, no. 6, pp. 761-770, 1997.

[23] Y. Yao, "A dynamic theory of learning", Ph.D dissertation, Dept. Math., Univ. Calfornia, Berkeley, CA, USA, 2006.

[24] T. Hytönen, J. V. Neerven, M. Veraar, and L. Weis, *Analysis in Banach Spaces*, Berlin: Springer, 2016.

[25] L. Guo, *Time-varying Stochastic Systems: Stability, Estimation and Control*. Jilin, China: Jilin Science and Technology Press, 1990.