Coefficient Dependent Regularization Regression with Indefinite Kernels for Streaming Data

Weishan Pan, Hongwei Sun

***Abstract*—Regularization regression learning scheme finds its powerful applications in many areas, such as Computer science, manufacturing engineering, economic decision making, etc. In this paper, the indefinite kernel based coefficient dependent regularization algorithm for block-wise streaming data is proposed, and learning performance of this algorithm is studied by bounding the learning error. Our learning scheme works in an online and weighted average manner. The total error is decomposed into weighted average of local variance error and weighted average of local bias error. By the kernel decompose and integral operator method, satisfied error bound and learning rates are derived. Our error analysis shows that mild growth of the sizes of data block and the underregularization strategy can guarantee the convergence of the learning scheme.**

***Index Terms*—learning theory, kernel regularization, streaming data, learning rates, adaptive underregularization**

I. Introduction

IN the digital age, the processing and application of massive information is particularly critical. Machine learning [1] has shown outstanding advantages in data processing and analysis. Classical machine learning mainly includes regression learning [2]–[4], classification [5], [6], cluster analysis and dimension reduction. In recent years, the complexity, variability and fluidity of data sources have aroused widespread concern and research on whether the usual learning methods can be applied to other different kind data. Some researchers have studied the learning performance of algorithms related to streaming data, such as distributed least square regularized algorithms for streaming data [3] and big data mining about streaming data [7], etc.

Kernel-based learning algorithms (also known as kernel methods) for a single data set have been discussed and studied in detail [8], including Support Vector Machine (SVM) [5], [6], Kernel-based Fisher Discriminant Analysis (KFD) [9], Kernel Principal Component Analysis (KPCA) [10], etc. For a large data set, adopting the divide-and-conquer method [11]–[14], can also achieve pretty well learning effect.

Stream data is a dynamic data set, which can be received instance-wise or block-wise. From the point of instance-wise, online learning by gradient descent method has been widely applied in learning systems. From the point of block-wise, the divide-and-conquer approach is available, and distributed kernel regularized learning systems have been explored [3].

The main purpose of this paper is to study the error bound and the asymptotic convergence rate of coefficient dependent regularization systems with indefinite kernels for streaming data. For this goal, we firstly recall coefficient and kernel based regularization regression (CKRR) learning.

Let X be a compact metric space and $Y = \mathbb{R}$, ρ be an unknown Borel probability distribution on $Z = X \times Y$. The regression function is defined by

$$f_\rho(x) = \int_Y y d\rho(y|x). \tag{1}$$

where $\rho(y|x)$ is the condition distribution of y for given x. In fact, $f_\rho(x)$ is the condition expectation of y when x, i.e. $f_\rho(x) = \mathbf{E}(y|x)$. As we all know, regression learning is to learn an approximation of f_ρ by the sample set

$$D = \{(x_i, y_i)\}_{i=1}^N \in Z^N$$

drawn independently and randomly according to ρ.

Let $K : X \times X \longrightarrow \mathbb{R}$ be continuous and bounded real function called $kernel$. The CKRR associated with kernel K minimizes the regularized empirical loss

$$f_{D,\lambda} = f_{\alpha_D},$$

$$\alpha_D = \arg\min_{\alpha \in \mathbb{R}^N} \frac{1}{|D|} \sum_{(x,y)\in D} (f_\alpha(x) - y)^2 + \lambda N \sum_{i=1}^N a_i^2.$$

Here the coefficient dependent function f_α is defined by

$$f_\alpha = \sum_{i=1}^N a_i K(\cdot, x_i),$$

and the hypothesis space in CKRR is

$$\mathscr{H}_{K,\mathbf{x}} = \Big\{ f_\alpha = \sum_{i=1}^N a_i K(\cdot, x_i) : \alpha = (a_1, \cdots, a_N) \in \mathbb{R}^N \Big\}.$$

The divide and conquer learning scheme divides a single data set $D = \{(x_i, y_i)\}_{i=1}^N \in Z^N$ into m data subsets $\left\{D^{(j)}\right\}_{j=1}^m$, and some basic learning algorithm is applied to each subset to learn a local regression model, finally the local models are averaged to generate the final learning model [13]. The streaming data may be received block-wise or instance-wise. Even an instance-wise data are often dealt in a block-wise manner, for instance, in the dynamic pricing problems and other economic and financial data analysis. Let $D_s = \{(x_{s,i}, y_{s,i})\}_{i=1}^{n_s} \in Z^{n_s}$ be the data block receiving at time s with the size n_s. The set of all available data until time t is

$$\widetilde{D}_t \doteq \bigcup_{s=1}^t D_s,$$

the size of $\widetilde{D}_t$ is

$$|\widetilde{D}_t| = \sum_{s=1}^t |D_s| = \sum_{s=1}^t n_s \doteq N_t.$$

Manuscript received November 21, 2022; revised July 11, 2023.

This work was supported in part by the National Natural Science Foundation of China under Grants No. 11671171 and 11871167.

W. S. Pan is a postgraduate student of University of Jinan, Shandong 250000 China. (e-mail: panweishan98@163.com).

H. W. Sun is a professor of University of Jinan, Shandong 250000 China (corresponding author to provide phone: 86-15069142316 ; e-mail: ss_sunhw@ujn.edu.cn).

Our regression learning scheme for streaming data takes the divide and conquer approach, and indefinite kernel based CKRR as the base algorithm. The CKRR for local data subset D_s at time s can be stated as

$$f_s = f_{D_s,\lambda_s} = f_{\alpha_{D_s}},$$
$$\alpha_{D_s} = \arg\min_{\alpha\in\mathbb{R}^{n_s}} \frac{1}{|D_s|}\sum_{(x,y)\in D_s}(f_\alpha(x)-y)^2 + \lambda n_s \sum_{i=1}^{n_s} a_i^2. \tag{2}$$

The global estimator F_t is obtained by weighted average of these local estimators $\{f_s\}_{s=1}^t$,

$$F_t = \frac{n_t}{N_t}f_t + \frac{N_{t-1}}{N_t}F_{t-1} = \sum_{s=1}^{t}\frac{n_s}{N_t}f_s. \tag{3}$$

II. Assumptions and main results

For simplicity, we assume the uniform boundness of the output data.

Assumption II.1. *Assume that $|y| \le M$ for some constant $M > 0$ almost surely.*

The above assumption indicates that the regression function f_ρ is bounded and $f_\rho \in L^2_{\rho_X}$, where $L^2_{\rho_X}$ is the Hilbert space of square integrable functions related to the marginal distribution ρ_X. In addition, the variance of distribution ρ is finite, i.e.

$$\sigma^2 = \mathbf{E}\left[(y - f_\rho(x))^2\right] < \infty$$

Our error analysis is accomplished mainly by the technique of integral operator. Hence we should recall the structure theory of indefinite kernels, for more details see references [4], [15].

Let L_K be the integral operator corresponding to kernel K, which is defined by

$$L_K : L^2_{\rho_X} \longrightarrow L^2_{\rho_X};\quad L_K f = \int_X K(\cdot,t)f(t)d\rho_X(t).$$

L_K is continuous and compact operator, but it may not be self-adjoint and positive without the positive semidefinite assumption of kernel K. Although K may not be a Mercer kernel, it can induce two Mercer kernels defined as

$$\widetilde{K}(x,t) = \int_X K(x,u)K(t,u)d\rho_X(u),$$
$$\widehat{K}(x,t) = \int_X K(v,x)K(v,t)d\rho_X(v).$$

It is easy to see that

$$L_{\widehat{K}} = L_K^* L_K\ ,\ L_{\widetilde{K}} = L_K L_K^*. \tag{4}$$

Thus, $L_{\widetilde{K}}$ and $L_{\widehat{K}}$ have the same positive eigenvalue sequence $\sigma_l^2, l\in\mathbb{N}$, and we can assume these eigenvalues are in a non-increasing order. Let $\varphi_l\ , l\in\mathbb{N}$ be associated orthonormal and continuous eigenfunctions of $L_{\widetilde{K}}$ and $\psi_l\ , l\in\mathbb{N}$ be associated orthonormal and continuous eigenfunctions of $L_{\widehat{K}}$ respectively. It is proved in [4] that

$$L_{\widetilde{K}} = \sum_{l=1}^{\infty}\sigma_l^2\varphi_l\otimes\varphi_l,\ L_{\widehat{K}} = \sum_{l=1}^{\infty}\sigma_l^2\psi_l\otimes\psi_l,$$

$$L_K = \sum_{l=1}^{\infty}\sigma_l\varphi_l\otimes\psi_l.$$

Our second assumption called *kernel condition* is proposed in [15], which ensures that $K(x,\cdot)$ and $K(\cdot,t)$ can belong to some reproduce kernel Hilbert space (RKHS).

Assumption II.2.

$$\kappa_0^2 = \sup_{x\in X}\sum_{l=1}^{\infty}\sigma_l\varphi_l^2(x) < \infty,\ \kappa_1^2 = \sup_{t\in X}\sum_{l=1}^{\infty}\sigma_l\psi_l^2(t) < \infty.$$

Denote $\kappa = \max\{\kappa_0,\kappa_1\}$. By Assumption II.2, we have two Mercer kernels

$$K_0(x,t) = \sum_{l=1}^{\infty}\sigma_l\varphi_l(x)\varphi_l(t),\ K_1(x,t) = \sum_{l=1}^{\infty}\sigma_l\psi_l(x)\psi_l(t)$$

The corresponding RKHS $\mathscr{H}_0$ and $\mathscr{H}_1$ are function spaces,

$$\mathscr{H}_0 = \left\{f = \sum_{l=1}^{\infty}c_l\varphi_l : \sum_{l=1}^{\infty}\frac{c_l^2}{\sigma_l} < \infty\right\};$$

$$\mathscr{H}_1 = \left\{f = \sum_{l=1}^{\infty}d_l\psi_l : \sum_{l=1}^{\infty}\frac{d_l^2}{\sigma_l} < \infty\right\}.$$

Let U be the partial isometry operator from $L^2_{\rho_X}$ to $L^2_{\rho_X}$, satisfying that $\psi_l = U\varphi_l, l\in\mathbb{N}$. Proposition II.1 proved in [4] summarizes some properties of kernel K and integral operator L_K.

Proposition II.1. *Under the Kernel Condition, we have*

(i) $L_K = L_{K_0}U^* = U^*L_{K_1}$;
(ii) $K(\cdot,x)\in\mathscr{H}_0$ *and* $K(x,\cdot)\in\mathscr{H}_1$ *for any* $x\in X$;
(iii) U *is an isometry operator from* $\mathscr{H}_0$ *to* $\mathscr{H}_1$ *and* $UK(\cdot,x) = K_1(\cdot,x)$; U^* *is an isometry operator from* $\mathscr{H}_1$ *to* $\mathscr{H}_0$ *and* $U^*K(x,\cdot) = K_0(\cdot,x)$;
(iv) L_K *is bounded from* $L^2_{\rho_X}$ *to* $\mathscr{H}_0$ *and from* $\mathscr{H}_1$ *to* $\mathscr{H}_0$ *with both operator norms bounded by* κ^2.

In order to deduce the error bound, the following *prior condition* is needed to depict the approximation ability of hypothesis space to the regression function f_ρ.

Assumption II.3. *There holds* $f_\rho = L_{K_0}^{\beta}g_\rho$ *for some* $g_\rho\in L^2_{\rho_X}(X)$ *and* $0 < \beta \le 2$.

Now we can state our main results on error analysis.

Theorem II.1. *Under Assumption II.1, Assumption II.2 and Assumption II.3, and suppose that for any $s\in\mathbb{N}$, the sample size at time s satisfies $n_s \ge a_0 s^p$ with absolute constants $a_0 > 0$ and $p > 0$. When $0 < \beta \le \frac{3}{2}$, by taking*

$$\lambda_s = n_s^{-\min\left\{\frac{2(1+p)}{3(1+p)+2p\beta},\frac{1}{2}\right\}},$$

there exists constant c_1 independent of n_s, λ_s or N_s such that

$$\mathbf{E}\|F_t - f_\rho\|^2_{L^2_{\rho_X}} \le c_1 N_t^{-\min\left\{\frac{2p\beta}{3(1+p)+2p\beta},\frac{p\beta}{2(1+p)}\right\}}.$$

When $\frac{3}{2} < \beta \le 2$, by taking

$$\lambda_s = n_s^{-\min\left\{\frac{2(1+p)}{3(1+p)+2p\beta},\frac{2}{2\beta+1}\right\}},$$

there exists constant c_2 independent of n_s, λ_s or N_s such that

$$\mathbf{E}\|F_t - f_\rho\|^2_{L^2_{\rho_X}} \leq c_2 N_t^{-\min\left\{\frac{2p\beta}{3(1+p)+2p\beta}, \frac{2p\beta}{(2\beta+1)(1+p)}\right\}}.$$

Theorem II.2. *Under Assumption II.1, Assumption II.2 and Assumption II.3, and suppose that for any $s \in \mathbb{N}$, the sample size at time s satisfies $a_1 s^p \leq n_s \leq a_2 s^p$ for some absolute constants $0 < a_1 < a_2$ and $p > 0$. When $0 < \beta \leq \frac{3}{2}$, by taking*

$$\lambda_s = n_s^{-\min\left\{\frac{1}{2}, \frac{2(1+p)}{(3+2\beta)p}\right\}},$$

there exists constant c_3 independent of n_s, λ_s or N_s such that

$$\mathbf{E}\|F_t - f_\rho\|^2_{L^2_{\rho_X}} \leq c_3 N_t^{-\min\left\{\frac{\beta p}{2(1+p)}, \frac{2\beta}{3+2\beta}\right\}}.$$

When $\frac{3}{2} < \beta \leq 2$, by taking

$$\lambda_s = n_s^{-\min\left\{\frac{2}{1+2\beta}, \frac{2(1+p)}{(3+2\beta)p}\right\}},$$

there exists constant c_4 independent of n_s, λ_s or N_s such that

$$\mathbf{E}\|F_t - f_\rho\|^2_{L^2_{\rho_X}} \leq c_4 N_t^{-\min\left\{\frac{2\beta p}{(1+p)(1+2\beta)}, \frac{2\beta}{3+2\beta}\right\}}.$$

The proof of Theorem II.1 and Theorem II.2 will be given in Section IV. From these derived learning rates, we notice the following facts:

1) p is the parameter which reflects the growth rate of the sizes of data block, the bigger p gives faster convergence rate. Under the conditions of Theorem II.1, our conclusion shows that when $0 < \beta \leq \frac{1}{2}$, let p turn to infinity, the convergence rate grows upward to $\frac{\beta}{2}$; when $\frac{1}{2} < \beta \leq 2$, let p turn to infinity, the convergence rate grows upward to $\frac{2\beta}{3+2\beta}$.

2) It is well known that the rate $O(N_t^{-\frac{2\beta}{2\beta+1}})$ is minimax optimal in a capacity independent sense, see references [16], [17]. Under the conditions of Theorem II.2, when $\frac{3}{2} < \beta \leq 2$ and $p \geq \frac{1}{2} + \beta$, we have the following suboptimal rate for streaming data regression learning,

$$\mathbf{E}\|F_t - f_\rho\|^2_{L^2_{\rho_X}} \leq c_4 N_t^{-\frac{2\beta}{3+2\beta}}.$$

The difference between $\frac{2\beta}{2\beta+1}$ and $\frac{2\beta}{2\beta+3}$ is mainly caused by the more general kernels (without symmetric and positive semi-definite) and the coefficient dependent regularization.

3) Mild growth of the sizes of data block can guarantee the convergence of BSD-AKRR. Also, our $a_1 s^p \leq n_s \leq a_2 s^p$ assumption can ensure that

$$\lambda_s = n_s^{-\theta} \sim N_s^{-\frac{\theta}{p+1}},$$

this underregularization strategy means that regularization parameters are in fact selected according to the total sample size of all data blocks available at the time of processing an incoming block.

III. Error bound for local estimators

From formula (3), the global estimator F_t is the weighted average of local estimator functions f_s , so we firstly consider local input data set $D_s(x) = \{x : (x, y) \in D_s\}$ and associated sampling operators. Without causing confusion, the associated sampling operator

$$S_{D_s(x)} : \mathscr{H}_0(\mathscr{H}_1) \longrightarrow R^{n_s},$$

are simplicity expressed as

$$S : \mathscr{H}_0(\mathscr{H}_1) \longrightarrow R^{n_s},$$

that is

$$Sf = (f(x_{s,1}), f(x_{s,2}), \cdots, f(x_{s,n_s}))^\top.$$

Operators T and T_* are from R^{n_s} to $\mathscr{H}_0$ and $\mathscr{H}_1$ respectively, and for any $\mathbf{c} = (c_1, \cdots, c_{n_s}) \in R^{n_s}$,

$$T\mathbf{c} = \frac{1}{n_s}\sum_{i=1}^{n_s} c_i K(\cdot, x_{s,i}),\ T_*\mathbf{c} = \frac{1}{n_s}\sum_{i=1}^{n_s} c_i K(x_{s,i}, \cdot).$$

The definition of sampling operator S and T, T_* asserts that

$$TS = \frac{1}{n_s}\sum_{i=1}^{n_s} K(\cdot, x_{s,i}) \otimes K_1(\cdot, x_{s,i}), \tag{5}$$

$$T_*S = \frac{1}{n_s}\sum_{i=1}^{n_s} K(x_{s,i}, \cdot) \otimes K_0(\cdot, x_{s,i}). \tag{6}$$

By Proposition II.1, L_K is also an operator from $\mathscr{H}_1$ to $\mathscr{H}_0$, and

$$\begin{aligned} L_K = U^* L_{K_1} &= U^* \mathbf{E}\big(K_1(\cdot, x) \otimes K_1(\cdot, x)\big) \\ &= \mathbf{E}\big(K(\cdot, x) \otimes K_1(\cdot, x)\big); \end{aligned}$$

similarly we have

$$L_K^* = \mathbf{E}\big(K(x, \cdot) \otimes K_0(\cdot, x)\big).$$

Combined these expressions with (5) and (6), the integral operator L_K can be approximated by the sampling operator TS, and L_K^* by T_*S. And then, sampling operator TST_*S can approximate integral operator $L_{\widetilde{K}}$ by equation (4). The following Lemma III.1 and III.2 are proven in [4]. In the sequel, for $i, j = 0, 1$, $\|\cdot\|_{ij}$ indicates operator norm from $\mathscr{H}_i$ to $\mathscr{H}_j$.

Lemma III.1. *L_K is considered to be operator from $\mathscr{H}_1$ to $\mathscr{H}_0$. Then there hold*

$$\mathbf{E}\|L_K^* - T_*S\|_{01}^2 \leq \frac{\kappa^4}{n_s},\ \mathbf{E}\|L_K - TS\|_{10}^2 \leq \frac{\kappa^4}{n_s},$$
$$\mathbf{E}\|L_K^* - T_*S\|_{01}^4 \leq \frac{12\kappa^8}{n_s^2},\ \mathbf{E}\|L_K - TS\|_{10}^4 \leq \frac{12\kappa^8}{n_s^2}.$$

Lemma III.2. *We have*

$$\mathbf{E}\|TST_*S - L_{\widetilde{K}}\|_{00}^2 \leq \frac{4\kappa^8}{n_s},\ \mathbf{E}\|TST_*S - L_{\widetilde{K}}\|_{00}^4 \leq \frac{192\kappa^{16}}{n_s^2}.$$

From formula (3), the global estimator F_t is the weighted average of local estimator functions, so we firstly deduce the error bound of the local estimator function f_s in this section. The following operator expression of f_s is proposed in [18],

$$f_s = T(\lambda_s I + ST_*ST)^{-1}ST_*y = (\lambda_s I + TST_*S)^{-1}TST_*y.$$

By that $TS \approx L_K$, $TST_*S \approx L_{\widetilde{K}}$, and $\mathbf{E}T_*y = L_K^* f_\rho$, we introduce the noise free estimator f_{λ_s} as follows

$$f_{\lambda_s} \doteq (\lambda_s I + L_{\widetilde{K}})^{-1} L_{\widetilde{K}} f_\rho \Longrightarrow \lambda_s f_{\lambda_s} = L_{\widetilde{K}}(f_\rho - f_{\lambda_s}).$$

Now we naturally have the decomposition of sample error,

$$\begin{aligned} f_s - f_{\lambda_s} = &(\lambda_s I + TST_*S)^{-1}(TS - L_K)L_K^*(f_\rho - f_{\lambda_s}) \\ &+ (\lambda_s I + TST_*S)^{-1} TS\Delta_s, \end{aligned} \tag{7}$$

where

$$\begin{aligned} \Delta_s &= T_*y - T_*Sf_{\lambda_s} - L_K^*(f_\rho - f_{\lambda_s}) \\ &= \frac{1}{n_s}\sum_{i=1}^{n_s}(y_{s,i} - f_{\lambda_s}(x_{s,i}))K(x_{s,i},\cdot) - L_K^*(f_\rho - f_{\lambda_s}). \end{aligned}$$

Recall the variance of the output data, $\sigma^2 = \mathbf{E}(y - f_\rho(x))^2$. For simplicity, we assume that

$$0 \le \lambda_s \le 1, \quad \kappa \ge 1, \quad \text{and } \max\{1, \sigma, \|g_\rho\|_{L^2_{\rho_X}}\} \le M.$$

The following two lemmas proved in [15] are preliminary knowledge of our error analysis for local estimator.

Lemma III.3. *Under Assumption II.2 and II.3, there holds*

$$\begin{aligned} &\|f_{\lambda_s} - f_\rho\|^2_{L^2_{\rho_X}} \le \lambda_s^\beta \|g_\rho\|^2_{L^2_{\rho_X}}, \\ &\|L_K^*(f_{\lambda_s} - f_\rho)\|_1^2 \le \kappa^2 \lambda_s^{\min\{\frac{1}{2}+\beta,\, 2\}} \|g_\rho\|^2_{L^2_{\rho_X}}. \end{aligned}$$

Lemma III.4. *Under Assumption II.2, there holds*

$$\begin{aligned} &\|(\lambda_s I + TST_*S)^{-1}TS\|_{10} \le \frac{\kappa^2}{\lambda_s}, \\ &\|T_*S(\lambda_s I + TST_*S)^{-1}\|_{01} \le \frac{\kappa^2}{\lambda_s}, \\ &\|(\lambda_s I + TST_*S)^{-1}\|_{00} \le \frac{1}{\lambda_s}\left(1 + \frac{\kappa^2}{\sqrt{\lambda_s}}\right), \\ &\|T_*S(\lambda_s I + TST_*S)^{-1}TS\|_{10} \le \kappa^2 \lambda_s^{-\frac{1}{2}}. \end{aligned}$$

Proof: We only give the proof of the last inequality. Notice that $(ST)^* = ST_*$, then

$$\begin{aligned} &\|T_*S(\lambda_s I + TST_*S)^{-1}TS\|_{10} \\ =&\|T_*ST(\lambda_s I + ST_*ST)^{-1}S\|_{10} \\ \le&\kappa^2\|ST(\lambda_s I + ST_*ST)^{-1}\| \le \kappa^2\lambda_s^{-\frac{1}{2}}. \end{aligned}$$

■

The conditional expectation of Δ_s is denoted by

$$\widetilde{\Delta}_s = \frac{1}{n_s}\sum_{i=1}^{n_s}(f_\rho(x_{s,i}) - f_{\lambda_s}(x_{s,i}))K(x_{s,i},\cdot) - L_K^*(f_\rho - f_{\lambda_s}).$$

Lemma III.5. *Under Assumption II.1, II.2 and II.3. There hold that*

$$\begin{aligned} &\mathbf{E}\|\Delta_s\|_1^2 \le \frac{\kappa^2 M^2}{n_s}\left(1 + \lambda_s^\beta\right), \\ &\mathbf{E}\|\Delta_s\|_1^4 \le 112\kappa^{12}M^4\left(n_s^{-3}\lambda_s^{\min\{2\beta-\frac{1}{2},\beta\}} + n_s^{-2}\right), \\ &\mathbf{E}\|\widetilde{\Delta}_s\|_1^2 \le \frac{\kappa^2\lambda_s^\beta}{n_s}\|g_\rho\|^2_{L^2_{\rho_X}}. \end{aligned}$$

Proof: The first inequality is proved in [14], and the third inequality is proved in [3]. So we only prove the upper bound of $\mathbf{E}\|\Delta_s\|_1^4$.

Denote the random variable

$$\eta(z) = (y - f_{\lambda_s}(x))K(x,\cdot),$$

and referring the proof of lemma 4 in [14], we obtain that

$$\begin{aligned} &\mathbf{E}\big\|\Delta_s\big\|_1^4 \\ &\le \frac{1}{n_s^3}\mathbf{E}\big\|\eta(z) - \mathbf{E}\eta\big\|_1^4 + \frac{3}{n_s^2}\Big(\mathbf{E}\big\|\eta(z) - \mathbf{E}\eta\big\|_1^2\Big)^2. \end{aligned} \tag{8}$$

Introducing the random variable

$$\xi(x) = (f_\rho(x) - f_{\lambda_s}(x))K(x,\cdot),$$

and by that

$$\eta(z) - \mathbf{E}\eta = (y - f_\rho(x))K(x,\cdot) + (\xi(x) - \mathbf{E}\xi),$$

we have

$$\begin{aligned} &\mathbf{E}\big\|\eta(z) - \mathbf{E}\eta\big\|_1^4 \\ &= \mathbf{E}\Big[\|\xi(x) - \mathbf{E}\xi\|_1^2 + (y - f_\rho(x))^2\langle K(x,\cdot),\, K(x,\cdot)\rangle_{\mathscr{H}_1} \\ &+ 2(y - f_\rho(x))\langle K(x,\cdot),\, \xi(x) - \mathbf{E}\xi\rangle_{\mathscr{H}_1}\Big]^2. \end{aligned}$$

By that $|y - f_\rho(x)| \le 2M$ almost surely, and

$$\langle K(x,\cdot),\, K(x,\cdot)\rangle_{\mathscr{H}_1} = K_0(x,x) \le \kappa^2,$$

The square term can be spread as

$$\begin{aligned} &\mathbf{E}\big\|\eta(z) - \mathbf{E}\eta\big\|_1^4 \\ \le&\mathbf{E}\|\xi(x) - \mathbf{E}\xi\|_1^4 + 4M^2\kappa^4\sigma^2 + 24M^2\kappa^2\mathbf{E}\|\xi(x) - \mathbf{E}\xi\|_1^2 \\ &+4\kappa\sqrt{2M}\mathbf{E}|y - f_\rho(x)|^{\frac{1}{2}}\|\xi(x) - \mathbf{E}\xi\|_1^3 \\ &+16M^2\kappa^3\mathbf{E}|y - f_\rho(x)|\|\xi(x) - \mathbf{E}\xi\|_1. \end{aligned}$$

Now by Cauchy-Schwarz Inequality, and $\sigma \le M$,

$$\begin{aligned} \mathbf{E}\big\|\eta(z) - \mathbf{E}\eta\big\|_1^4 \le\ &\mathbf{E}\|\xi(x) - \mathbf{E}\xi\|_1^4 + 4M^4\kappa^4 \\ &+ 24M^2\kappa^2\mathbf{E}\|\xi(x) - \mathbf{E}\xi\|_1^2 + 4\sqrt{2}\kappa M\big(\mathbf{E}\|\xi(x) - \mathbf{E}\xi\|_1^4\big)^{\frac{3}{4}} \\ &+ 16M^3\kappa^3\big(\mathbf{E}\|\xi(x) - \mathbf{E}\xi\|_1^2\big)^{\frac{1}{2}}. \end{aligned} \tag{9}$$

By Lemma III.3, we have that

$$\begin{aligned} &\mathbf{E}\|\xi(x) - \mathbf{E}\xi\|_1^2 \\ &\le \mathbf{E}\|\xi(x)\|_1^2 \le \kappa^2\|f_\rho - f_{\lambda_s}\|^2_{L^2_{\rho_X}} \le \kappa^2\lambda_s^\beta\|g_\rho\|^2_{L^2_{\rho_X}}. \end{aligned} \tag{10}$$

To bound $\mathbf{E}\big\|\xi(x) - \mathbf{E}\xi\big\|_1^4$, we first derive the upper bound of $\big\|\xi(x) - \mathbf{E}\xi\big\|_1^2$,

$$\begin{aligned} \big\|\xi(x) - \mathbf{E}\xi\big\|_1^2 &\le 2(f_\rho(x) - f_{\lambda_s}(x))^2 K_0(x,x) + 2\|\mathbf{E}\xi\|_1^2 \\ &\le 4\kappa^2\big(M^2 + \kappa^2\|f_{\lambda_s}\|_0^2\big) + 2\|L_K^*(f_\rho - f_{\lambda_s})\|_1^2 \\ &\le 6\kappa^2M^2 + 4\kappa^{10}M^2\lambda_s^{\min\{\beta-\frac{1}{2},\,0\}} \doteq b^2. \end{aligned}$$

The last inequality holds by Lemma III.3 and

$$\begin{aligned} \|f_{\lambda_s}\|_0 &= \|(\lambda_s I + L_{\widetilde{K}})^{-1}L_{\widetilde{K}}^{\frac{3+2\beta}{4}} g_\rho\|_{L^2_{\rho_X}} \\ &\le \kappa^3\lambda_s^{\min\{\frac{2\beta-1}{4},\,0\}}\|g_\rho\|_{L^2_{\rho_X}}. \end{aligned}$$

By Chebyshev's Inequality and (10), for any $0 \le t \le b$,

$$\begin{aligned} F(t) &\doteq \mathbb{P}rob\big\{\|\xi(x) - \mathbf{E}\xi\|_1 \ge t\big\} \\ &\le \frac{\mathbf{E}\|\xi(x) - \mathbf{E}\xi\|_1^2}{t^2} \le \frac{\kappa^2\lambda_s^\beta\|g_\rho\|^2_{L^2_{\rho_X}}}{t^2}. \end{aligned}$$

Now we can deduce that

$$\begin{aligned}&\mathbf{E}\big\|\xi(x)-\mathbf{E}\xi\big\|_1^4=\int_0^b -t^4dF(t)=4\int_0^b t^3F(t)dt\\&\le 2\kappa^2M^2\lambda_s^\beta b^2=12\kappa^4M^4\lambda_s^\beta+8\kappa^{12}M^4\lambda_s^{\min\{2\beta-\frac{1}{2},\beta\}}\\&\le 20\kappa^{12}M^4\lambda_s^{\min\{2\beta-\frac{1}{2},\beta\}}.\end{aligned}\tag{11}$$

Plugging (10) and (11) into (9), we can get

$$\mathbf{E}\big\|\eta(z)-\mathbf{E}\eta\big\|_1^4\le 100M^4\kappa^{12}\Big(1+\lambda_s^{\min\{2\beta-\frac{1}{2},\beta\}}\Big).\tag{12}$$

By (10) and that

$$\mathbf{E}\langle(y-f_\rho(x))K(x,\cdot),\,\xi(x)-\mathbf{E}\xi\rangle_1=0,$$

there holds

$$\begin{aligned}\mathbf{E}\big\|\eta(z)-\mathbf{E}\eta\big\|_1^2&=\mathbf{E}\big\|(y-f_\rho(x))K(x,\cdot)+(\xi(x)-\mathbf{E}\xi)\big\|_1^2\\&\le\kappa^2\sigma^2+\kappa^2M^2\lambda_s^\beta\le 2\kappa^2M^2.\end{aligned}\tag{13}$$

Combing (12) and (13) with (8), the second inequality of Lemma III.5 is proved. ∎

The goal of this paper is to explore the approximation ability of the global empirical function F_t to f_ρ. We recall the error decomposition as follows from [3],

$$\begin{aligned}\mathbf{E}\|F_t-f_\rho\|_{L^2_{\rho_X}}^2\le&\sum_{s=1}^t\frac{n_s^2}{N_t^2}\mathbf{E}\|f_s-\mathbf{E}f_s\|_{L^2_{\rho_X}}^2\\&+\sum_{s=1}^t\frac{n_s}{N_t}\|\mathbf{E}f_s-f_\rho\|_{L^2_{\rho_X}}^2.\end{aligned}\tag{14}$$

From the above inequality (14), we will concentrate on the estimate of the local variance $\mathbf{E}\|f_s-\mathbf{E}f_s\|_{L^2_{\rho_X}}^2$ and the local bias$\|\mathbf{E}f_s-f_\rho\|_{L^2_{\rho_X}}^2$ in Propositions III.1 and III.2 respectively. As we all know, for any $f\in L^2_{\rho_X}$, there holds

$$\mathbf{E}\|f_s-\mathbf{E}f_s\|_{L^2_{\rho_X}}^2\le\mathbf{E}\|f_s-f\|_{L^2_{\rho_X}}^2.$$

Taking $f=f_{\lambda_s}$, the local variance can be bounded by the approximation error, i.e.

$$\mathbf{E}\|f_s-\mathbf{E}f_s\|_{L^2_{\rho_X}}^2\le\mathbf{E}\|f_s-f_{\lambda_s}\|_{L^2_{\rho_X}}^2.$$

Proposition III.1. *Under Assumption II.1, II.2 and II.3, there holds*

$$\begin{aligned}&\mathbf{E}\|f_s-f_{\lambda_s}\|_{L^2_{\rho_X}}^2\\&\le 72\kappa^{12}M^2\Big(n_s^{-\frac{1}{2}}\lambda_s^{-\frac{1}{2}}+1\Big)\lambda_s^{\min\{\beta-2,-\frac{3}{2}\}}n_s^{-1}.\end{aligned}$$

Proof: Notice that both f_s and f_{λ_s} are in $\mathscr{H}_0$. By the error decompose (7),

$$\begin{aligned}&\mathbf{E}\|f_s-f_{\lambda_s}\|_{L^2_{\rho_X}}^2\\=&\mathbf{E}\|L_{K_0}^{\frac{1}{2}}(f_s-f_{\lambda_s})\|_0^2\\\le&2\mathbf{E}\|L_{K_0}^{\frac{1}{2}}(\lambda_sI+TST_*S)^{-1}TS\Delta_s\|_0^2\\&+2\mathbf{E}\|L_{K_0}^{\frac{1}{2}}(\lambda_sI+TST_*S)^{-1}(TS-L_K)L_K^*(f_\rho-f_{\lambda_s})\|_0^2\\\doteq&2\mathbf{E}\|A_1\|_0^2+2\mathbf{E}\|A_2\|_0^2\end{aligned}\tag{15}$$

For any $g\in\mathscr{H}_0$, by $UL_{K_0}=L_K^*$ and Cauchy-Schwarz Inequality, there holds

$$\begin{aligned}&\|L_{K_0}^{\frac{1}{2}}(\lambda_sI+TST_*S)^{-1}g\|_0^2\\&=\big\langle(\lambda_sI+TST_*S)^{-1}g,L_{K_0}(\lambda_sI+TST_*S)^{-1}g\big\rangle_{\mathscr{H}_0}\\&\le\|(\lambda_sI+TST_*S)^{-1}g\|_0\|L_K^*(\lambda_sI+TST_*S)^{-1}g\|_1\\&\le\|T_*S(\lambda_sI+TST_*S)^{-1}g\|_1\|(\lambda_sI+TST_*S)^{-1}g\|_0\\&\quad+\|L_K^*-T_*S\|_{01}\|(\lambda_sI+TST_*S)^{-1}g\|_0^2.\end{aligned}\tag{16}$$

Taking $g=TS\Delta_s$, by Lemma III.4,

$$\begin{aligned}&\mathbf{E}\|A_1\|_0^2\\&\le\mathbf{E}\|L_K^*-T_*S\|_{01}\|(\lambda_sI+TST_*S)^{-1}TS\|_{10}^2\|\Delta_s\|_1^2\\&+\mathbf{E}\|T_*S(\lambda_sI+TST_*S)^{-1}TS\|_{10}\|\Delta_s\|_1^2\\&\times\|(\lambda_sI+TST_*S)^{-1}TS\|_{10}\\&\le\kappa^4\lambda_s^{-2}\mathbf{E}\|L_K^*-T_*S\|_{01}\|\Delta_s\|_1^2+\kappa^4\lambda_s^{-\frac{3}{2}}\mathbf{E}\|\Delta_s\|_1^2.\end{aligned}$$

We can continue our estimate by Lemma III.1, Lemma III.5, and Cauchy-Schwarz Inequality,

$$\begin{aligned}\mathbf{E}\|A_1\|_0^2\le&\kappa^4\lambda_s^{-2}\Big(\mathbf{E}\|L_K^*-T_*S\|_{01}^2\Big)^{\frac{1}{2}}\Big(\mathbf{E}\|\Delta_s\|_1^4\Big)^{\frac{1}{2}}\\&+\kappa^4\lambda_s^{-\frac{3}{2}}\mathbf{E}\|\Delta_s\|_1^2\\\le&11\kappa^{12}M^2\Big[\lambda_s^{\min\{\beta-\frac{9}{4},\frac{\beta}{2}-2\}}n_s^{-2}+\lambda_s^{-2}n_s^{-\frac{3}{2}}\Big]\\&+2\kappa^6M^2\lambda_s^{-\frac{3}{2}}n_s^{-1}\\\le&11\kappa^{12}M^2\lambda_s^{-2}n_s^{-\frac{3}{2}}\Big[\lambda_s^{\min\{\beta-\frac{1}{4},\frac{\beta}{2}\}}n_s^{-\frac{1}{2}}+1\Big]\\&+2\kappa^6M^2\lambda_s^{-\frac{3}{2}}n_s^{-1}.\end{aligned}$$

Similarly,taking $g=(TS-L_K)L_K^*(f_\rho-f_{\lambda_s})$, by Lemma III.4, and Lemma III.1, Lemma III.5,

$$\begin{aligned}&\mathbf{E}\|A_2\|_0^2\\\le&4\kappa^4\lambda_s^{-3}\mathbf{E}\|L_K^*-T_*S\|_{01}\|TS-L_K\|_{10}^2\|L_K^*(f_\rho-f_{\lambda_s})\|_1^2\\&+2\kappa^4\lambda_s^{-\frac{5}{2}}\mathbf{E}\|TS-L_K\|_{10}^2\|L_K^*(f_\rho-f_{\lambda_s})\|_1^2\\\le&\|L_K^*(f_\rho-f_{\lambda_s})\|_1^2\times\Big[2\kappa^4\lambda_s^{-\frac{5}{2}}\mathbf{E}\|TS-L_K\|_{10}^2\\&+4\kappa^4\lambda_s^{-3}\Big(\mathbf{E}\|L_K^*-T_*S\|_{01}^2\Big)^{\frac{1}{2}}\Big(\mathbf{E}\|TS-L_K\|_{10}^4\Big)^{\frac{1}{2}}\Big]\\\le&8\sqrt{3}\kappa^{12}M^2\lambda_s^{\min\{\beta-\frac{5}{2},-1\}}n_s^{-\frac{3}{2}}\\&+2\kappa^{10}M^2\lambda_s^{\min\{\beta-2,-\frac{1}{2}\}}n_s^{-1}.\end{aligned}$$

Plugging the above two estimates into (15), and note that $0<\lambda_s\le 1$, thus

$$\begin{aligned}\max\Big\{\lambda_s^{\min\{\beta-2,-\frac{1}{2}\}},\,\lambda_s^{-\frac{3}{2}}\Big\}&\le\lambda_s^{\min\{\beta-2,-\frac{3}{2}\}};\\\max\Big\{\lambda_s^{\min\{\beta-\frac{5}{2},-1\}},\,\lambda_s^{-2}\Big\}&\le\lambda_s^{\min\{\beta-\frac{5}{2},-2\}}.\end{aligned}$$

Then there holds

$$\begin{aligned}&\mathbf{E}\|f_s-f_{\lambda_s}\|_{L^2_{\rho_X}}^2\\&\le 22\kappa^{12}M^2\lambda_s^{-2}n_s^{-\frac{3}{2}}\Big[\lambda_s^{\min\{\beta-\frac{1}{4},\frac{\beta}{2}\}}n_s^{-\frac{1}{2}}+1\Big]+4\kappa^6M^2\lambda_s^{-\frac{3}{2}}n_s^{-1}\\&+28\kappa^{12}M^2\lambda_s^{\min\{\beta-\frac{5}{2},-1\}}n_s^{-\frac{3}{2}}+4\kappa^{10}M^2\lambda_s^{\min\{\beta-2,-\frac{1}{2}\}}n_s^{-1}\\&\le 22\kappa^{12}M^2\lambda_s^{\min\{\beta-\frac{9}{4},-2\}}n_s^{-2}\\&+50\kappa^{12}M^2\lambda_s^{\min\{\beta-\frac{5}{2},-2\}}n_s^{-\frac{3}{2}}+8\kappa^{10}M^2\lambda_s^{\min\{\beta-2,-\frac{3}{2}\}}n_s^{-1}\\&\le 72\kappa^{12}M^2\Big(n_s^{-\frac{1}{2}}\lambda_s^{-\frac{1}{2}}+1\Big)n_s^{-1}\lambda_s^{\min\{\beta-2,-\frac{3}{2}\}}.\end{aligned}$$

■

Proposition III.2. *Under Assumption II.1, II.2 and II.3, there holds*

$$\begin{aligned}&\|\mathbf{E}f_s - f_\rho\|^2_{L^2_{\rho_X}}\\&\leq 30\kappa^{12}M^2\Big[\lambda_s^{\min\{\beta-2,-\frac{1}{2}\}}n_s^{-1}\Big(n_s^{-\frac{1}{2}}\lambda_s^{-\frac{1}{2}}+1\Big)+\lambda_s^{\beta}\Big].\end{aligned}$$

Proof: We only need to bound $\|\mathbf{E}f_s - f_{\lambda_s}\|^2_{L^2_{\rho_X}}$, because

$$\|\mathbf{E}f_s - f_\rho\|^2_{L^2_{\rho_X}} \leq 2\|\mathbf{E}f_s - f_{\lambda_s}\|^2_{L^2_{\rho_X}} + 2\|f_{\lambda_s} - f_\rho\|^2_{L^2_{\rho_X}}.$$

By the decomposition of sample error in formula (7),

$$\begin{aligned}\mathbf{E}f_s - f_{\lambda_s} = \mathbf{E}(\lambda_s I + TST_*S)^{-1}(TS - L_K)L_K^*(f_\rho - f_{\lambda_s})\\ + \mathbf{E}(\lambda_s I + TST_*S)^{-1}TS\widetilde{\Delta}_s.\end{aligned}$$

Hence

$$\begin{aligned}&\|\mathbf{E}f_s - f_{\lambda_s}\|^2_{L^2_{\rho_X}}\\&\leq 2\|L_{K_0}^{\frac{1}{2}}\mathbf{E}(\lambda_s I + TST_*S)^{-1}TS\widetilde{\Delta}_s\|_0^2\\&+ 2\|L_{K_0}^{\frac{1}{2}}\mathbf{E}(\lambda_s I + TST_*S)^{-1}(TS - L_K)L_K^*(f_\rho - f_{\lambda_s})\|_0^2\\&\doteq 2\|J_1\|_0^2 + 2\|J_2\|_0^2\end{aligned}$$

By that $U^*L_K^* = L_{K_0}$, then

$$\begin{aligned}\|J_1\|_0^2 &= \langle U^*L_K^*\mathbf{E}(\lambda_s I + TST_*S)^{-1}TS\widetilde{\Delta}_s,\\&\qquad \mathbf{E}(\lambda_s I + TST_*S)^{-1}TS\widetilde{\Delta}_s\rangle_{\mathscr{H}_0}\\&\leq \|\mathbf{E}L_K^*(\lambda_s I + TST_*S)^{-1}TS\widetilde{\Delta}_s\|_1\\&\qquad\times\|\mathbf{E}(\lambda_s I + TST_*S)^{-1}TS\widetilde{\Delta}_s\|_0\\&\leq \|\mathbf{E}(L_K^* - T_*S)(\lambda_s I + TST_*S)^{-1}TS\widetilde{\Delta}_s\|_1\\&\qquad\times\|\mathbf{E}(\lambda_s I + TST_*S)^{-1}TS\widetilde{\Delta}_s\|_0\\&\qquad+\|\mathbf{E}T_*S(\lambda_s I + TST_*S)^{-1}TS\widetilde{\Delta}_s\|_1\\&\qquad\times\|\mathbf{E}(\lambda_s I + TST_*S)^{-1}TS\widetilde{\Delta}_s\|_0\end{aligned}$$

By Lemma III.4, Lemma III.5, and Jensen Inequality,

$$\begin{aligned}\|J_1\|_0^2 \leq&\kappa^4\lambda_s^{-2}\big(\mathbf{E}\|L_K^* - T_*S\|_{01}\|\widetilde{\Delta}_s\|_1\big)\mathbf{E}\|\widetilde{\Delta}_s\|_1\\&+\kappa^4\lambda_s^{-\frac{3}{2}}\big(\mathbf{E}\|\widetilde{\Delta}_s\|_1\big)^2\\\leq&\kappa^4\lambda_s^{-2}\big(\mathbf{E}\|L_K^* - T_*S\|_{01}^2\big)^{\frac{1}{2}}\mathbf{E}\|\widetilde{\Delta}_s\|_1^2\\&+\kappa^4\lambda_s^{-\frac{3}{2}}\mathbf{E}\|\widetilde{\Delta}_s\|_1^2\\\leq&\kappa^8M^2\Big(n_s^{-\frac{1}{2}}\lambda_s^{-\frac{1}{2}}+1\Big)\lambda_s^{\beta-\frac{3}{2}}n_s^{-1}.\end{aligned}$$

Refer to the estimation of $\mathbf{E}\|A_2\|_0^2$ in the proof of Proposition III.1, we have

$$\begin{aligned}&\|J_2\|_0^2\\\leq&\mathbf{E}\|L_{K_0}^{\frac{1}{2}}(\lambda_s I + TST_*S)^{-1}(TS - L_K)L_K^*(f_\rho - f_{\lambda_s})\|_0^2\\\leq&14\kappa^{12}M^2\lambda_s^{\min\{\beta-2,-\frac{1}{2}\}}n_s^{-1}\Big(n_s^{-\frac{1}{2}}\lambda_s^{-\frac{1}{2}}+1\Big).\end{aligned}$$

Combing the above estimates with Lemma III.3, the proof of Proposition III.2 is completed. ■

IV. Learning Rates of CKRR with Streaming Data

In this section we prove our main conclusions proposed in Section 2. To simplify the expression, we use $f \preceq g$ denote $f \leq cg$ where c is a constant independent of n_s and λ_s, $s \in \mathbb{N}$. Furthermore, we use $f \sim g$ denote $f \preceq g$ and $g \preceq f$ simultaneously. Hence for all $a > -1$, there holds

$$\sum_{s=1}^{t} s^a \sim t^{a+1},\ t \in \mathbb{N}. \tag{17}$$

By the error decomposition (14), and Proposition III.1, Proposition III.2,

$$\begin{aligned}&\mathbf{E}\|F_t - f_\rho\|^2_{L^2_{\rho_X}}\\\preceq&\frac{1}{N_t^2}\sum_{s=1}^{t} n_s\lambda_s^{\min\{\beta-2,-\frac{3}{2}\}}\Big(n_s^{-\frac{1}{2}}\lambda_s^{-\frac{1}{2}}+1\Big)+\frac{1}{N_t}\sum_{s=1}^{t} n_s\lambda_s^{\beta}\\&+\frac{1}{N_t}\sum_{s=1}^{t}\lambda_s^{\min\{\beta-2,-\frac{1}{2}\}}\Big(n_s^{-\frac{1}{2}}\lambda_s^{-\frac{1}{2}}+1\Big).\end{aligned} \tag{18}$$

Lemma IV.1. *Under the condition $n_s \geq a_0 s^p$ for any $s \in \mathbb{N}$, there holds*

$$\sum_{s=1}^{t} n_s^\alpha \preceq \begin{cases} N_t^{\max\{\frac{1+\alpha p}{1+p},0\}}(\log N_t)^{\vartheta(\alpha p)} & \text{if } \alpha \leq 1;\\ N_t^\alpha & \text{if } \alpha \geq 1.\end{cases}$$

Here

$$\vartheta(t) = \begin{cases}1 & \text{if } t = -1;\\ 0 & \text{otherwise.}\end{cases}$$

Proof: Under the condition $n_s \geq a_0 s^p$ for any $s \in \mathbb{N}$, there holds

$$N_t = \sum_{s=1}^{t} n_s \succeq \sum_{s=1}^{t} s^p \succeq t^{p+1}.$$

Hence, for any $t \in \mathbb{N}$, there is

$$t \preceq N_t^{\frac{1}{p+1}}. \tag{19}$$

When $0 < \alpha \leq 1$, by Hölder inequality, we can deduce

$$\sum_{s=1}^{t} n_s^\alpha \leq \left(\sum_{s=1}^{t}(n_s^\alpha)^{\frac{1}{\alpha}}\right)^\alpha t^{1-\alpha} = t^{1-\alpha}N_t^\alpha \preceq N_t^{\frac{1+\alpha p}{1+p}}.$$

When $\alpha \leq 0$, we have

$$\begin{aligned}\sum_{s=1}^{t} n_s^\alpha \preceq \sum_{s=1}^{t} s^{p\alpha} &\preceq t^{\max\{1+\alpha p,\,0\}}(\log t)^{\vartheta(p\alpha)}\\&\preceq N_t^{\max\{\frac{1+\alpha p}{1+p},0\}}(\log N_t)^{\vartheta(p\alpha)}.\end{aligned}$$

For all $\alpha \geq 1$, it is obvious to see

$$\sum_{s=1}^{t} n_s^\alpha \leq \Big(\sum_{s=1}^{t} n_s\Big)^\alpha = N_t^\alpha.$$

The proof of Lemma IV.1 is completed. ■

Proof of Theorem II.1. By taking $\lambda_s = n_s^{-\theta}$ with $0 < \theta \leq 1$, there is

$$n_s^{-\frac{1}{2}}\lambda_s^{-\frac{1}{2}} + 1 \leq 2.$$

Hence the formula (18) can be rewritten as

$$\begin{aligned}&\mathbf{E}\|F_t-f_\rho\|^2_{L^2_{\rho_X}}\\ \preceq&\frac{1}{N_t^2}\sum_{s=1}^t n_s^{1-\theta\min\{\beta-2,-\frac{3}{2}\}}+\frac{1}{N_t}\sum_{s=1}^t n_s^{-\theta\min\{\beta-2,-\frac{1}{2}\}}\\ &+\frac{1}{N_t}\sum_{s=1}^t n_s^{1-\theta\beta}\\ \preceq&\frac{1}{N_t^2}\sum_{s=1}^t n_s^{1+\theta\max\{2-\beta,\frac{3}{2}\}}+\frac{1}{N_t}\sum_{s=1}^t n_s^{\theta\max\{2-\beta,\frac{1}{2}\}}\\ &+\frac{1}{N_t}\sum_{s=1}^t n_s^{1-\theta\beta}.\end{aligned}\tag{20}$$

When $0<\beta\le\frac{1}{2}$ and $0<\theta<\frac{1}{2-\beta}$. By Lemma IV.1,

$$\begin{aligned}&\mathbf{E}\|F_t-f_\rho\|^2_{L^2_{\rho_X}}\\ \preceq&N_t^{-1+\theta(2-\beta)}+N_t^{(-1+\theta(2-\beta))\frac{p}{1+p}}+N_t^{-\frac{\theta\beta p}{1+p}}\\ \preceq&N_t^{(-1+\theta(2-\beta))\frac{p}{1+p}}+N_t^{-\frac{\theta\beta p}{1+p}}.\end{aligned}$$

Hence by choosing $\theta=\frac{1}{2}$, we have

$$\mathbf{E}\|F_t-f_\rho\|^2_{L^2_{\rho_X}}\preceq N_t^{-\frac{\beta p}{2(1+p)}}.$$

When $\frac{1}{2}<\beta\le\frac{3}{2}$ and $0<\theta<\frac{2}{3}$. By Lemma IV.1,

$$\mathbf{E}\|F_t-f_\rho\|^2_{L^2_{\rho_X}}\preceq N_t^{-1+\frac{3}{2}\theta}+N_t^{(-1+\theta(2-\beta))\frac{p}{1+p}}+N_t^{-\frac{\theta\beta p}{1+p}}.$$

Hence by choosing

$$\theta=\min\Big\{\frac{2(1+p)}{3(1+p)+2p\beta},\frac{1}{2}\Big\},$$

we have

$$\mathbf{E}\|F_t-f_\rho\|^2_{L^2_{\rho_X}}\preceq N_t^{-\min\{\frac{2p\beta}{3(1+p)+2p\beta},\frac{p\beta}{2(1+p)}\}}.$$

Note that when $0<\beta\le\frac{1}{2}$, there are

$$\frac{2(1+p)}{3(1+p)+2p\beta}\ge\frac{1}{2},\text{ and }\frac{2p\beta}{3(1+p)+2p\beta}\ge\frac{p\beta}{2(1+p)}.$$

Hence our first conclusion in Theorem II.1 is proved.

When $\frac{3}{2}<\beta\le 2$ and $0<\theta\le 1$. By Lemma IV.1,

$$\begin{aligned}\mathbf{E}\|F_t-f_\rho\|^2_{L^2_{\rho_X}}\preceq&N_t^{-1+\frac{3}{2}\theta}+N_t^{(-1+\frac{1}{2}\theta)\frac{p}{1+p}}\\ &+N_t^{-\min\{\frac{\theta\beta p}{1+p},1\}}(\log N_t)^{\vartheta((1-\theta\beta)p)}\end{aligned}$$

Hence by choosing

$$\theta=\min\Big\{\frac{2(1+p)}{3(1+p)+2p\beta},\frac{2}{2\beta+1}\Big\},$$

we have

$$\mathbf{E}\|F_t-f_\rho\|^2_{L^2_{\rho_X}}\preceq N_t^{-\min\{\frac{2p\beta}{3(1+p)+2p\beta},\frac{2p\beta}{(2\beta+1)(1+p)}\}}.$$

The proof of Theorem II.1 is completed.

Proof of Theorem II.2. Under the condition $a_1s^p\le n_s\le a_2s^p$, $s=1,2,\cdots,t$, we have $N_s\sim s^{p+1}$. Taking $\lambda_s=n_s^{-\theta}$ with $0<\theta\le 1$, by Lemma IV.1 and (17), we can continue our estimate from (20),

$$\begin{aligned}&\mathbf{E}\|F_t-f_\rho\|^2_{L^2_{\rho_X}}\\ \preceq&\frac{1}{N_t^2}\sum_{s=1}^t s^{p+p\theta\max\{2-\beta,\frac{3}{2}\}}+\frac{1}{N_t}\sum_{s=1}^t s^{p\theta\max\{2-\beta,\frac{1}{2}\}}\\ &+\frac{1}{N_t}\sum_{s=1}^t n_s^{1-\theta\beta}\\ \preceq&N_t^{-2}t^{1+p+p\theta\max\{2-\beta,\frac{3}{2}\}}+N_t^{-1}t^{1+p\theta\max\{2-\beta,\frac{1}{2}\}}\\ &+N_t^{-1+\max\{1-\frac{p\theta\beta}{1+p},0\}}(\log t)^{\vartheta(p(1-\theta\beta))}\\ \preceq&N_t^{-1+\frac{p\theta}{1+p}\max\{2-\beta,\frac{3}{2}\}}+N_t^{-\frac{p}{1+p}+\frac{p\theta}{1+p}\max\{2-\beta,\frac{1}{2}\}}\\ &+N_t^{\max\{-\frac{p\theta\beta}{1+p},-1\}}(\log N_t)^{\vartheta(p(1-\theta\beta))}.\end{aligned}$$

When $0<\beta\le\frac{1}{2}$, we have

$$\begin{aligned}&\mathbf{E}\|F_t-f_\rho\|^2_{L^2_{\rho_X}}\\ \preceq&N_t^{-1+\frac{p\theta}{1+p}(2-\beta)}+N_t^{-\frac{p}{1+p}+\frac{p\theta}{1+p}(2-\beta)}+N_t^{-\frac{p\theta\beta}{1+p}}\\ \preceq&N_t^{-\min\left\{\frac{p}{1+p}(1-\theta(2-\beta)),\frac{p\theta\beta}{1+p}\right\}}.\end{aligned}$$

By taking $\theta=\frac{1}{2}$, there holds

$$\mathbf{E}\|F_t-f_\rho\|^2_{L^2_{\rho_X}}\preceq N_t^{-\frac{\beta p}{2(1+p)}}.$$

When $\frac{1}{2}<\beta\le\frac{3}{2}$, we have

$$\begin{aligned}&\mathbf{E}\|F_t-f_\rho\|^2_{L^2_{\rho_X}}\\ \preceq&N_t^{-1+\frac{3p\theta}{2(1+p)}}+N_t^{-\frac{p}{1+p}+\frac{p\theta}{1+p}(2-\beta)}\\ &+N_t^{\max\{-\frac{p\theta\beta}{1+p},-1\}}(\log N_t)^{\vartheta(p(1-\theta\beta))}\\ \preceq&N_t^{-\min\left\{1-\frac{3p\theta}{2(1+p)},\frac{p}{1+p}(1-\theta(2-\beta)),\frac{p\theta\beta}{1+p}\right\}}(\log N_t)^{\vartheta(p(1-\theta\beta))}.\end{aligned}$$

By taking $\theta=\min\left\{\frac{1}{2},\frac{2(1+p)}{(3+2\beta)p}\right\}$, there holds

$$\mathbf{E}\|F_t-f_\rho\|^2_{L^2_{\rho_X}}\preceq N_t^{-\frac{\beta p}{1+p}\min\left\{\frac{1}{2},\frac{2(1+p)}{(3+2\beta)p}\right\}}.$$

When $\frac{3}{2}<\beta\le 2$, we have

$$\begin{aligned}&\mathbf{E}\|F_t-f_\rho\|^2_{L^2_{\rho_X}}\\ \preceq&N_t^{-1+\frac{3p\theta}{2(1+p)}}+N_t^{-\frac{p}{1+p}+\frac{p\theta}{2(1+p)}}\\ &+N_t^{\max\{-\frac{p\theta\beta}{1+p},-1\}}(\log N_t)^{\vartheta(p(1-\theta\beta))}\\ \preceq&N_t^{-\min\left\{1-\frac{3p\theta}{2(1+p)},\frac{p}{1+p}(1-\frac{\theta}{2}),\frac{p\theta\beta}{1+p}\right\}}(\log N_t)^{\vartheta(p(1-\theta\beta))}.\end{aligned}$$

By taking $\theta=\min\left\{\frac{2}{1+2\beta},\frac{2(1+p)}{(3+2\beta)p}\right\}$, there holds

$$\mathbf{E}\|F_t-f_\rho\|^2_{L^2_{\rho_X}}\preceq N_t^{-\frac{\beta p}{1+p}\min\left\{\frac{2}{1+2\beta},\frac{2(1+p)}{(3+2\beta)p}\right\}}.$$

Hence the desired conclusions in Theorem II.2 is proved.

REFERENCES

[1] T. M. Mitchell, *Machine learning*. McGraw-Hill New York, 1997.

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

[3] X. Q. Zheng, H. W. Sun, and Q. Wu, "Regularized least square kernel regression for streaming data," *Communications in Mathematical Sciences*, vol. 19, no. 6, pp. 1533–1548, 2021.

[4] H. W. Sun and Q. Wu, "Indefinite kernel network with dependent sampling," *Analysis and Applications*, vol. 11, no. 5, pp. 880–646, 2013.
[5] T. Evgeniou, M. Pontil, and T. Poggio, "Regularization networks and support vector machines," *Advances in Computational Mathematics*, vol. 13, no. 1, pp. 1–50, 2000.
[6] Q. Wu and D. X. Zhou, "Svm soft margin classifiers: Linear programming versus quadratic programming," *Neural Computation*, vol. 17, no. 5, pp. 1160–1187, 2005.
[7] N. Jiang and L. Gruenwald, "Research issues in data stream association rule mining," *Association for Computing Machinery Sigmod Record*, vol. 35, no. 1, pp. 14–19, 2006.
[8] T. Hofmann, B. Schölkopf, and A. J. Smola, "Kernel methods in machine learning," *Annals of Statistics*, vol. 36, no. 3, pp. 1171–1220, 2008.
[9] S. Mika, "Kernel fisher discriminants," *Proceedings Aistats*, 2003.
[10] T. J. Chin and D. Suter, "Incremental kernel principal component analysis," *Institute of Electrical and Electronics Engineers Transactions on Image Processing*, vol. 16, no. 6, pp. 1662–1674, 2007.
[11] L. S. Bo, X. Chang, and Z. D. Xuan, "Distributed semi-supervised learning with kernel ridge regression," *Journal of Machine Learning Research*, vol. 18, no. 1, pp. 1493–1514, 2017.
[12] S. B. Lin, X. Guo, and D. X. Zhou, "Distributed learning with regularized least squares," *Journal of Machine Learning Research*, vol. 18, no. 1, pp. 3202–3232, 2017.
[13] Y. Zhang, J. Duchi, and M. Wainwright, "Divide and conquer kernel ridge regression," *Journal of Machine Learning Research*, vol. 30, pp. 592–617, 2013.
[14] M. J. Pang and H. W. Sun, "Distributed regression learning with coefficient regularization," *Journal of Mathematical Analysis and Applications*, vol. 466, no. 1, pp. 676–689, 2018.
[15] Q. Wu, "Regularization networks with indefinite kernels," *Journal of Approximation Theory*, vol. 166, pp. 1–18, 2013.
[16] H. W. Sun and Q. Wu, "Optimal rates of distributed regression with imperfect kernels," *Journal of Machine Learning Research*, vol. 22, pp. 171–1, 2021.
[17] T. Zhang, "Leave-one-out bounds for kernel methods," *Neural Computation*, vol. 15, no. 6, pp. 1397–1437, 2003.
[18] H. W. Sun and Q. Wu, "Least square regression with indefinite kernels and coefficient regularization," *Applied and Computational Harmonic Analysis*, vol. 30, no. 1, pp. 96–109, 2011.