Regularized Modal Regression on Markov-Dependent Observations: A Theoretical Assessment

Tieliang Gong[1,2], Yuxin Dong[1,2], Hong Chen[3], Wei Feng[1,2], Bo Dong[2,4], Chen Li[1,2]

[1]School of Computer Science and Technology, Xi'an Jiaotong University, Xi'an 710049, China
[2]Key Laboratory of Intelligent Networks and Network Security, Ministry of Education, Xi'an 710049, China
[3]College of Science, Huazhong Agriculture University, Wuhan 430070, China
[4]School of Continuing Education, Xi'an Jiaotong University, Xi'an,710049, China
adidasgtl@gmail.com, dongyuxin@stu.xjtu.edu.cn, chenh@mail.hzau.edu.cn, weifeng.ft@foxmail.com,
dong.bo@mail.xjtu.edu.cn, cli@xjtu.edu.cn

Abstract

Modal regression, a widely used regression protocol, has been extensively investigated in statistical and machine learning communities due to its robustness to outliers and heavy-tailed noises. Understanding modal regression's theoretical behavior can be fundamental in learning theory. Despite significant progress in characterizing its statistical property, the majority of the results are based on the assumption that samples are independent and identical distributed (i.i.d.), which is too restrictive for real-world applications. This paper concerns the statistical property of regularized modal regression (RMR) within an important dependence structure - Markov dependent. Specifically, we establish the upper bound for RMR estimator under moderate conditions and give an explicit learning rate. Our results show that the Markov dependence impacts on the generalization error in the way that sample size would be discounted by a multiplicative factor depending on the spectral gap of underlying Markov chain. This result shed a new light on characterizing the theoretical underpinning for robust regression.

Introduction

In this paper, we consider the non-parametric regression problem which aims at inferring the relationship between input and output. To formulate this problem, denote X as the covariate variable that takes values in a compared metric space $\mathcal{X} \subset \mathbb{R}^d$ and Y that take values in $\mathcal{Y} = \mathbb{R}$. The sample pair (X, Y) is generated from the following model :

$$Y = f^*(X) + \epsilon,$$

where ϵ is the noise term. The goal of non-parametric regression is to find the unknown function $f^\star$ in a non-parametric manner while some certain assumptions on noise term are imposed. This problem can be boiled down to learn a characterization of the conditional distribution, given a set of observations. Some commonly used characterizations include the conditional mean (Tibshirani 1996), the conditional quantile (Yu, Lu, and Stander 2003; Meinshausen and Ridgeway 2006) and the conditional mode (Chen et al. 2016; Feng, Fan, and Suykens 2020), which correspond to mean regression, quantile regression and modal regression (MR), respectively. Each regression protocol has its own benefits in modeling the noise. For instance, conditional mean regression can achieve satisfactory effect when the noise is Gaussian or sub-Gaussian, while regression towards the conditional quantile and conditional mode can be more robust in complex noise cases. In practice, selecting an appropriate regression protocol usually depends on the data type.

Modal regression is an appropriate regression protocol when facing heavy-tailed noises and outliers. Different from the conventional mean regression, which aims to estimate the conditional mean, modal regression seeks for the unknown truth $f^\star$ by regressing towards to the conditional mode function. For a set of observations, the mode denotes the value that appears most frequently. In the context of density estimation, the mode is the value at which the density function achieves its peak value. Hence, conditional mode can reveal the structure of outputs and the trends of observations. Research on modal regression can be broadly classified into two categories: (semi-) parametric and nonparametric approaches. For parametric approaches, a parametric form of the global conditional mode function is required. To name a few, studies in (Lee 1989; Yu and Aristodemou 2012; Yao and Li 2014; Lv, Zhu, and Yu 2014; Khardani and Yao 2017) fall in this setting. For non-parametric approaches, the conditional mode is sought by maximizing a conditional density or a joint density which is typically estimated in a non-parametric manner. Typical works include (Chen et al. 2016; Feng, Fan, and Suykens 2020; Yao and Xiang 2016; Zhou, Huang et al. 2016; Wang et al. 2017). Great progress on understanding the theoretical property of modal regression estimator have been made during the last two decades (we refer the reader to (Feng, Fan, and Suykens 2020)). In particular, Chen et al. (Chen et al. 2016) derived asymptotic error bounds for local modal regression within the framework of kernel density estimation. Feng et al. (Feng, Fan, and Suykens 2020) established the statistical consistency for modal regression estimator by assuming the existence of global conditional mode function.

All the works mentioned above are based on the assumption that data are independent and identical distributed (i.i.d.). Nevertheless, this assumption is too restrictive in a broad range of real datasets. As a matter of fact, a considerable number of real datasets are tempera in nature. For example, functional magnetic resonance imaging (fMRI) data (Ryali et al. 2012; Smith 2012) are usually collected from different regions over a time period; the macroeconomic data (McCracken and Ng 2016) span the time periods of decades and are kept updating till now. It poses challenges for researchers applying modal regression to these time-series data. Therefore, understanding the statistical behavior of modal regression estimator for time-series data can be one of the most important issues.

This paper aims to close the gap between theories of modal regression and practical requirements in addressing dependent observation of real data. Inspired by the statistical guarantees of modal regression dealing with heavy-tailed errors in the independent setup, we consider extending MR to cope with dependent structure of observations. Albeit convergence rates on modal regression are given in (Feng, Fan, and Suykens 2020; Wang et al. 2017), it is still unclear whether these results work for dependent observations. As an initial exploration on this topic, this paper narrows down to Markov chain, an important and widely used dependence structure, investigating the generalization performance of regularized modal regression (RMR) on Markov-dependence data. Within the Markov-dependence setup, we first show that RMR estimator is statistical consistent under moderate conditions, and establish its explicit convergence rates with order $\mathcal{O}\big((1-\gamma^2)^{-\frac{1}{5}}m^{-\frac{1}{5}}\big)$ under appropriate parameter selection, where m is number of Markov-dependent observations, and γ is the absolute spectral gap of the underlying Markov chain.

The rest of the paper is organized as follows. Section 2 introduces the necessary notions and notations. Section 3 presents the assumptions and the main theorems. Section 4 sketches the proofs of the main theorems. Finally, a brief discussion is concluded in Section 5.

Model and Methodology

Model Setup

Let $\mathcal{X} \in \mathbb{R}^d$ and $\mathcal{Y} \in \mathbb{R}$ be the input and output spaces respectively. In the modal regression setting, training samples $\mathbf{z} = \{(\mathbf{x}_i, y_i)\}_{i=1}^m \subset \mathcal{X} \times \mathcal{Y}$ are generated independently by

$$Y = f^*(X) + \epsilon, \tag{1}$$

where the mode of the conditional distribution of ϵ at any $\mathbf{x} \in \mathcal{X}$ is assumed to be zero, i.e.

$$\mathbf{mode}(\varepsilon|X=\mathbf{x}) = \arg\max_t p_{\epsilon|X}(t|X=\mathbf{x}) = 0, \ \forall \mathbf{x} \in \mathcal{X}, \tag{2}$$

where $p_{\epsilon|X}$ be the conditional density of ε on X. Then, the target function of modal regression can be represented by

$$f^*(\mathbf{x}) = \mathbf{mode}(Y|X=\mathbf{x}) = \arg\max_t p_{Y|X}(t|X=\mathbf{x}). \tag{3}$$

Throughout this paper, we assume that for any $\mathbf{x} \in \mathcal{X}$, $\arg\max_t p_{Y|X}(t|X=\mathbf{x})$ is well defined, which is equivalent to the existence and uniqueness of the global mode of the conditional density $p_{Y|X}$. Moreover, we assume that f^* is bounded, i.e $\|f^*\|_\infty \le M$ for some $M > 0$.

Denote ρ on $\mathcal{X} \times \mathcal{Y}$ as the intrinsic distribution for data generated by (1) and denote $\rho_{\mathcal{X}}$ as the corresponding marginal distribution on $\mathcal{X}$. For any measurable function $f : \mathcal{X} \to \mathbb{R}$, the modal regression performance can be characterized by

$$\mathcal{R}(f) = \int_{\mathcal{X}} p_{Y|X}\big(f(\mathbf{x})|X=\mathbf{x}\big)\,\mathrm{d}\rho_{\mathcal{X}}(\mathbf{x}). \tag{4}$$

It has been proved that f^* is the maximizer of (4) over all measurable functions (Feng, Fan, and Suykens 2020). Since $\rho_{\mathcal{X}}$ and $p_{Y|X}$ are usually unknown, we can not calculate the estimator directly by maximizing (4). Feng et al. (Feng, Fan, and Suykens 2020) proved $\mathcal{R}(f) = p_{\epsilon_f}(0)$, where p_{ε_f} is the density function of random variable $\epsilon_f = Y - f(X)$. This implies that maximizing $\mathcal{R}(f)$ over some hypothesis spaces is equivalent to maximizing the density of ε_f at 0, which can be estimated by non-parametric kernel density estimation.

Let $K_\sigma : \mathbb{R} \times \mathbb{R} \to \mathbb{R}_+$ be a kernel function, and $\phi(\frac{u-u'}{\sigma}) = K_\sigma(u,u')$ be a representing function which satisfies $\phi(u) = \phi(-u)$, $\phi(u) \le \phi(0)$ for any $u \in \mathbb{R}$ and $\int_{\mathbb{R}} \phi(u)\,\mathrm{d}u = 1$. With the help of K_σ, we can obtain the empirical estimation of $\mathcal{R}(f)$ by kernel density estimation, given by

$$\mathcal{R}^\sigma_{\mathbf{z}}(f) = \frac{1}{m\sigma}\sum_{i=1}^m K_\sigma(y_i - f(\mathbf{x}_i), 0) = \frac{1}{m\sigma}\sum_{i=1}^m \phi\Big(\frac{y_i - f(\mathbf{x}_i)}{\sigma}\Big).$$

For any $f : \mathcal{X} \to \mathbb{R}$, the expectation version of $\mathcal{R}^\sigma_{\mathbf{z}}(f)$ is

$$\mathcal{R}^\sigma(f) = \frac{1}{\sigma}\int_{\mathcal{X}\times\mathcal{Y}} \phi\Big(\frac{y - f(\mathbf{x})}{\sigma}\Big)\,\mathrm{d}\rho(\mathbf{x}, y),$$

which can be viewed as a surrogate of the true modal regression risk $\mathcal{R}(f)$ since $\mathcal{R}(f) - \mathcal{R}^\sigma(f) \to 0$ when $\sigma \to 0$ (Feng, Fan, and Suykens 2020).

Markovian Process

Let $\{X_i\}_{i\ge 1}$ be a Markov chain on a general space $\mathcal{X}$ with invariant probability distribution π. Let $P(x, \mathrm{d}y)$ be a Markov transition kernel on a general space $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$ and P^* be its adjoint, i.e. $P^*(x, \mathrm{d}y) := \frac{P(y,\mathrm{d}x)}{\pi(\mathrm{d}x)} \cdot \pi(\mathrm{d}y)$. For a reversible Markov chain, P is self-adjoint and coincides with P^* and $(P + P^*)/2$. For a non-reversible Markov chain, P is not self-adjoint, but $(P + P^*)/2$ is self-adjoint and associates with a reversible transition kernel (Rudolf 2011). Denote $\mathcal{L}_2(\pi)$ by the Hilbert space consisting of square integrable functions with respect to π. For any function $h : \mathcal{X} \to \mathbb{R}$, we write $\pi(h) := \int h(x)\pi(\mathrm{d}x)$. Define the norm of $h \in \mathcal{L}_2(\pi)$ as $\|h\|_\pi = \sqrt{\langle h, h\rangle}$. Let $P^t(x, \mathrm{d}y), (t \in \mathbb{N})$ be the t-step Markov transition kernel corresponding to P, then $P^t(x, S) = \Pr(X_{t+i} \in S | X_i = x)$ for $i \in \mathbb{N}, x \in \mathcal{X}$ and a measurable set S.

Following the above notations, we introduce the definitions of ergodicity and spectral gap for a Markov chain.

Definition 1 *Let $M(x)$ be a non-negative function. For an initial probability measure $\rho(\cdot)$ on $\mathcal{B}(\mathcal{X})$, a Markov chain is uniformly ergodic if*

$$\|P^t(\rho, \cdot) - \pi(\cdot)\|_{TV} \le M(x)\rho^t \tag{5}$$

for some $M(x) < \infty$ and $\rho < 1$, where $\|\cdot\|_{TV}$ denotes total variation norm.

A Markov chain is geometrically ergodic if (5) holds for some $t < 1$, which eliminates the bounded assumption on $M(x)$.

For a Markov chain with stationary distribution π, the spectrum of the chain is defined as $\mathcal{S} := \{\bar{\lambda} \in \mathbb{C} \setminus 0 : (\bar{\lambda} I - P)^{-1}$ does not exist as a bounded linear operator on $\mathcal{L}_2(\pi)\}$. For reversible chains, $\mathcal{S}$ lies on the real line.

Definition 2 *(Spectral gap, absolute spectral gap and pseudo spectral gap) (Paulin 2015) The spectral gap for reversible chains is*

$$\gamma = \begin{cases} 1 - \sup\{\bar{\lambda} : \bar{\lambda} \in \mathcal{S}, \bar{\lambda} \neq 1\}, & \textit{if eigenvalue 1 has multiplicity 1,} \\ 0 & \textit{otherwise} \end{cases}$$

For both reversible and non-reversible chains, the absolute spectral gap is

$$\gamma_a = \begin{cases} 1 - \sup\{|\bar{\lambda}| : \bar{\lambda} \in \mathcal{S}, \bar{\lambda} \neq 1\}, & \textit{if eigenvalue 1 has multiplicity 1,} \\ 0, & \textit{otherwise} \end{cases}$$

The pseudo spectral gap of a Markov operator P is

$$\gamma_p := \max_{k \geq 1} \{\gamma((P^*)^k P^k)/k\},$$

where $\gamma((P^*)^k P^k)$ denotes the spectral gap of the self-adjoint operator $(P^*)^k P^k$.

Remark 1 *The dependence of a Markov chain can be characterized by the spectral gap. A small $\bar{\lambda}$ usually implies a fast convergence of the Markov chain towards its stationary distribution from a non-stationary initial distribution (Rudolf 2011). Note that in the reversible case, $\gamma \geq \gamma_a$. The pseudo spectral gap is similar to the spectral gap in the sense that it allows to derive concentration bounds on MCMC empirical averages and is closely related to the mixing time (Paulin 2015).*

Regularized Modal Regression with Markov-Dependent Observations

Define an integral operator $L_K : \mathcal{L}_2 \to \mathcal{L}_2$ associated with the kernel $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ by

$$L_K f(\mathbf{x}) = \int_{\mathcal{X}} K(\mathbf{x}, \cdot) f(\cdot) \, \mathrm{d}\rho_X, \ \mathbf{x} \in \mathcal{X}.$$

Suppose $\mathcal{X}$ is compact and K is continuous, then $L_K L_K^\top$ is a self-adjoint positive operator with decreasing eigenvalues $\{\lambda_k^2\}_{k=1}^\infty$ with $\lambda_k \geq 0$ and eigenfunctions $\{\psi_k\}_{k=1}^\infty$ forms an orthonormal basis of $\mathcal{L}_2$. With this setup, we further define $|L_K|^\beta = |L_K L_K^\top|^{\frac{\beta}{2}}$ with $|L_K|^\beta (\sum_{k=1}^\infty c_k \psi_k) = \sum_{k=1}^\infty c_k \lambda_k^\beta \psi_k$, $\{c_k\}_k \in \ell_2$.

Given samples $\mathbf{z}$ and a continuous K, the sample dependent hypothesis space (SDHS) is defined as

$$\mathcal{H}_{K,\mathbf{z}} = \Big\{ f = \sum_{i=1}^m \alpha_i K(\mathbf{x}_i, \cdot), \boldsymbol{\alpha} = (\alpha_1, \alpha_2, \cdots, \alpha_m)^\top \in \mathbb{R}^m \Big\}, \tag{6}$$

which has been extensively used in generalization analysis of regression and classification. SDHS does not require the kernel to be symmetric and semi-definite, hence provides much flexibility and adaptivity for learning problems. It should be noted that the hypothesis $\mathcal{H}_{K,\mathbf{z}}$ can be expressed as the span of $K(\mathbf{x}, \cdot)$ over the inputs $\{\mathbf{x}_i\}_{i=1}^m$, which further implies that the hypothesis is determined by the coefficient $\alpha_i, i = 1, 2, \cdots, m$ once the kernel function is specified. Therefore, regularized modal regression aims to solve the following optimization problem

$$f_{\mathbf{z}} = \arg \max_{f \in \mathcal{H}_{K,\mathbf{z}}} \{\mathcal{R}_{\mathbf{z}}^\sigma(f) - \lambda \Omega_q(f)\}, \tag{7}$$

where $\lambda > 0$ is a regularization parameter and $\Omega_q(f)$ is the coefficient regularizer, defined by

$$\Omega_q(f) = \inf \Big\{ \sum_{i=1}^m |\alpha_i|^q : f = \sum_{i=1}^m \alpha_i K_{\mathbf{x}_i} \subset \mathcal{H}_1 \Big\}$$

with $q = 1, 2$, where $\mathcal{H}_1$ is given in Definition 3. Let $\mathbf{K}_i = (K(\mathbf{x}_1, \mathbf{x}_i), K(\mathbf{x}_2, \mathbf{x}_i), \cdots, K(\mathbf{x}_m, \mathbf{x}_i))$, then optimization model (7) can be reformulated as

$$\boldsymbol{\alpha}^{\mathbf{z}} = \arg\max_{\boldsymbol{\alpha} \in \mathbb{R}^m} \Big\{ \frac{1}{m\sigma} \sum_{i=1}^m \phi\Big(\frac{y_i - \mathbf{K}_i^\top \boldsymbol{\alpha}}{\sigma} \Big) - \lambda \|\boldsymbol{\alpha}\|_q^q \Big\} \tag{8}$$

with

$$f_{\mathbf{z}} = \sum_{i=1}^m \alpha_j^{\mathbf{z}} K(\mathbf{x}_i, \cdot).$$

Note that model (8) is reduced to a robust kernel machine to achieve sparseness when $q = 1$, which is a natural extension of sparse kernel regression (Chen and Wang 2018; Shi et al. 2019). When $q = 2$, it is closely related to kernel ridge regression by replacing modal regression criterion with the mean square error criterion. In particular, when Gaussian kernel is employed for kernel density function, (8) can be rewritten as

$$\boldsymbol{\alpha}^{\mathbf{z}} = \arg\max_{\boldsymbol{\alpha} \in \mathbb{R}^m} \Big\{ \frac{1}{m\sigma} \sum_{i=1}^m \exp\Big\{ - \frac{(y_i - \mathbf{K}_i^\top \boldsymbol{\alpha})^2}{\sigma} \Big\} - \lambda \|\boldsymbol{\alpha}\|_q^q \Big\},$$

which is consistent with the sparse correntropy regression with coefficient-based regularization (Chen and Wang 2018). This problem can be solved efficiently through the Half Quadratic (HQ) (Nikolova and Ng 2005) optimization strategy.

Theoretical Assessments

This section mainly concerns the theoretical property of regularized modal regression for Markov-dependent observations. Specifically, our goal is to bound the excess generalization error $\mathcal{R}(f^*) - \mathcal{R}(f_{\mathbf{z}})$ in the context of general Markov chain. To this end, we first introduce a Banach space $\mathcal{H}_1$, which contains all possible SDHS $\mathcal{H}_{K,\mathbf{z}}$ in (6).

Definition 3 *Define a Banach space $\mathcal{H}_1 = \big\{ f : f = \sum_{j=1}^\infty \alpha_j K(\mathbf{x}_j), \alpha_j \in \mathbb{R}, \{\mathbf{x}_j\} \subset X \big\}$ with the norm*

$$\|f\| = \inf \Big\{ \sum_{j=1}^\infty |\alpha_j| : f = \sum_{j=1}^\infty \alpha_j K_{\mathbf{x}_j} \Big\}, \tag{9}$$

It can be observed that $\mathcal{H}_1$ consists of continuous functions due to the continuity of K. As an important measurement of capacity of a hypothesis space, covering number have been extensively studied in the work (Zhou 2002, 2003; Steinwart and Christmann 2008). We adopt empirical covering number involved with $\mathcal{H}_1$ to get a tight bound for RMR estimator.

Definition 4 *(Empirical Covering Number (Wu, Ying, and Zhou 2007)) Let $\mathcal{H}$ be a set of functions on $\mathcal{Z}$ and samples $\mathbf{z} = \{z_1, z_2, \cdots, z_m\} \subset \mathcal{Z}$. The metric on $\mathcal{H}$ is denoted by*

$$d_{2,\mathbf{z}}(f,g) = \Big\{\frac{1}{m}\sum_{i=1}^{m}\big(f(z_i) - g(z_i)\big)^2\Big\}^{1/2} \textit{for} \quad f, g \in \mathcal{H}.$$

For any $\varepsilon > 0$, the empirical covering number of $\mathcal{H}$ with respect to $d_{2,\mathbf{z}}(f,g)$ is

$$\mathcal{N}_2(\mathcal{H},\varepsilon) = \sup_{m\in\mathbb{N}}\sup_{\mathbf{z}}\mathcal{N}_{2,\mathbf{z}}(\mathcal{H},\varepsilon) > 0,$$

where

$$\mathcal{N}_{2,\mathbf{z}}(\mathcal{H},\varepsilon) := \inf\Big\{l \in \mathbb{N} : \exists\{f_i\}_{i=1}^{l} \subset \mathcal{H} \textit{ such that } \mathcal{H} = \bigcup_{i=1}^{l}\{f \in \mathcal{H} : d_{2,\mathbf{z}}(f, f_i) \le \varepsilon\}\Big\}.$$

Note that for any function set $\mathcal{H} \subset \mathcal{C}(\mathcal{X})$, the empirical covering number $\mathcal{N}_{2,\mathbf{z}}(\mathcal{H},\varepsilon)$ can be bounded by $\mathcal{N}(\mathcal{H},\varepsilon)$, the uniform covering number of $\mathcal{H}$ with the metric $\|\cdot\|_\infty$, due to the fact $d_{2,\mathbf{z}}(f,g) \le \|f - g\|_\infty$. The function sets in our situation are balls of the SDHS in the form of $\mathcal{B}_R = \{f \in \mathcal{H} : \Omega_q^{\frac{1}{q}}(f) \le R\}$.

Assumption 1 *(Complexity) For any $\eta > 0$, there exists an exponent s with $0 < s < 2$ and $c_s > 0$ such that*

$$\log\mathcal{N}_2(\mathcal{H}_{K,\mathbf{z}},\eta) \le c_s\eta^{-s}, \ \ \forall\eta > 0. \qquad (10)$$

Assumption 2 *(Non-zero spectral gap) The underlying Markov chain $\{X_i\}_{i=1}^{n}$ is stationary with unique invariant measure π and admits an absolute spectral gap $1 - \gamma$.*

Assumption 3 *(Density) The conditional density of ϵ given X, i.e. $p_{\epsilon|X}$ is second-order continuous differentiable and $\|p''_{\epsilon|X}\|_\infty$ is bounded.*

Assumption 4 *(Calibrated Modal Regression Kernel) The representing function ϕ satisfies: 1) $\forall u \in \mathbb{R}, \phi(u) \le \phi(0) < \infty$; 2) ϕ is Lipschitz continuous with constant L_ϕ; 3) $\int_{\mathbb{R}}\phi(u)\,\mathrm{d}u = 1$ with $\int_{\mathbb{R}}u^2\phi(u)\,\mathrm{d}u < \infty$.*

Assumption 1 is a fairly standard assumption on describing the complexity of hypothesis space. It has been extensively studied in learning theory (Zhou 2002, 2003; Cucker and Smale 2001), from which we know for a C^∞ kernel, (10) holds for any $s > 0$. Assumption 2 requires the underlying Markov chain admits an absolute spectral gap, which quantifies the converge speed of Markov chain towards its invariant distribution π. Assumption 3 is a general condition on conditional density of $p''_{\epsilon|X}$ and conventional noise distributions satisfy this requirement. Assumption 4 requires the represent function to be bounded and Lipschitz continuous. Typical examples include the Gaussian kernel, Epanechnikov kernel, quadratic kernel and Triangular kernel. The following comparison theorem (Feng, Fan, and Suykens 2020) characterizes the relationship between excess modal risk and excess generalization risk.

Lemma 1 *(Feng, Fan, and Suykens 2020) Under assumption 3, for any measurable function $f : \mathcal{X} \to \mathbb{R}$, it holds that*

$$|\mathcal{R}(f^*) - \mathcal{R}(f) - (\mathcal{R}^\sigma(f^*) - \mathcal{R}^\sigma(f))| \le C_1\sigma^2,$$

where $C_1 = \|p''_{\epsilon|X}\|_\infty\int_{\mathbb{R}}u^2\phi(u)\,\mathrm{d}u$.

A well-established approach for conducting error analysis of learning algorithms is error decomposition, where the generalization error is usually decomposed into sample error and approximation error. Considering the characteristic of SDHS, we formulate the error decomposition of RMR by introducing the stepping stone functionf_λ, defined by

$$f_\lambda = \arg\max_{f\in\mathcal{H}_{K,\mathbf{z}}}\{\mathcal{R}^\sigma(f) - \lambda\Omega_q(f)\},$$

where $\lambda > 0$ is the regularization parameter.

Proposition 1 *Suppose $f_{\mathbf{z}}$ is produced by (7) based on Markov-dependent observations, and $f^* \in \mathcal{H}_{K,\mathbf{z}}$.Then*

$$\mathcal{R}(f^*) - \mathcal{R}(f_{\mathbf{z}}) \le \mathcal{S}_1(\mathbf{z}) + \mathcal{S}_2(\mathbf{z}) + C_1\sigma^2 + \lambda\Omega_q(f^*),$$

where $C_1 = \|p''_{\epsilon|X}\|_\infty\int_{\mathbb{R}}u^2\phi(u)\,\mathrm{d}u$ and

$$\mathcal{S}_1(\mathbf{z}) = \mathcal{R}^\sigma_{\mathbf{z}}(f^*) - \mathcal{R}^\sigma_{\mathbf{z}}(f_\lambda) - \{\mathcal{R}^\sigma(f^*) - \mathcal{R}^\sigma(f_\lambda)\},$$
$$\mathcal{S}_2(\mathbf{z}) = \mathcal{R}^\sigma(f^*) - \mathcal{R}^\sigma(f_{\mathbf{z}}) - \{\mathcal{R}^\sigma_{\mathbf{z}}(f^*) - \mathcal{R}^\sigma_{\mathbf{z}}(f_{\mathbf{z}})\}.$$

With these settings, we now present theoretical results for RMR with Markov-dependent observations.

Theorem 1 *Let the Markov-dependent observations $\mathbf{z}$ be generated by (1) with invariant distribution π and non-zero absolute spectral gap $\gamma_a > 0$. Suppose that **Assumptions** 1-4 are satisfied. Let f^* lies in the range of L_K^β for some $\beta \in (0, 2]$. Then for any $0 < \delta < 1$, the following inequality*

$$\mathcal{R}^\sigma(f^*) - \mathcal{R}^\sigma(f_{\mathbf{z}}) \le C\log(2/\delta)\Big((2\gamma_a - \gamma_a^2)^{-\frac{1}{2}}m^{-\frac{1}{2}}\sigma^{-\frac{1}{2}} + (2\gamma_a - \gamma_a^2)^{-\frac{1}{1+s}}m^{-\frac{1}{1+s}}\sigma^{-\frac{4+2s}{1+s}}R^{\frac{s+2}{s+1}} + \sigma^2 + \lambda^{\frac{2\beta}{2+\beta}}\Big)$$

holds with confidence at least $1 - \delta$, where C is a positive constant independent of m, σ, δ.

Remark 2 *Theorem 1 establishes the upper bound for regularized modal regression in Markov-dependent setup. As far as we can tell, this is the first work in the literature. It can be observed that the corresponding generalization error relies on the spectral gap of underlying Markov chain, the capacity of hypothesis space, the regularization parameter λ and the bandwidth parameter σ. The dependence of the Markov chain is measured by a quantity $\gamma_a \in [0, 1]$, denoting the norm of Markov operator (induced by transition kernel) acting on the $\mathcal{L}_2$ space with respect to the invariant distribution. It has been involved as constants in mean square error bound for Markov chain Monte Carlo (Rudolf 2011), Hoeffding-type (Fan, Jiang, and Sun 2018) and Bernstein-type inequalities for Markov chains (Paulin 2015). A non-zero spectral gap is closely related to other convergence* *criterion of Markov chains, e.g. geometrically ergodic, uniformly ergodic(Meyn and Tweedie 2012). Note that such a Markov chain can actually be generated by the so-called Markov sampling strategy (Gong, Zou, and Xu 2015; Gong, Xi, and Xu 2020), where a uniformly ergodic Markov chain can be generated from a given dataset without temporal relation.*

Theorem 2 *Under the same conditions in Theorem 1, take* $\theta = \frac{2\beta}{8\beta+5s\beta+2s+4}$, $\lambda = (2\gamma_a - \gamma_a^2)^{-\frac{\theta}{\beta}} m^{-\frac{\theta}{\beta}}$ *and* $\sigma = (2\gamma_a - \gamma_a^2)^{-\frac{\theta}{2\beta}} m^{-\frac{\theta}{2\beta}}$. *For any* $0 < \delta < 1$, *the excess risk of RMR estimator* $f_{\mathbf{z}}$ *satisfies*

$$\mathcal{R}(f^*) - \mathcal{R}(f_{\mathbf{z}}) \leq \hat{C} \log(2/\delta)(2\gamma_a - \gamma_a^2)^{-\frac{\theta}{\beta}} m^{-\theta}$$

with confidence at least $1-\delta$, *where* $\hat{C}$ *is a positive constant independent of* m, σ, δ.

Remark 3 *Theorem 2 implies the estimation consistency of RMR when* λ, σ *are properly specified. In particular, when* $s \to 0$, $\beta = 2$ *we see that the learning rate in Theorem 2 is* $\mathcal{O}\big((2\gamma_a - \gamma_a^2)^{-\frac{1}{5}} m^{-\frac{1}{5}}\big)$, *which is faster than the result in (Wang et al. 2017), whose learning is* $\mathcal{O}(m^{-\frac{1}{7}})$. *It is worth noting that the learning rate of RMR in Markov-dependent samples would be discounted by a multiplicative coefficient* $(2\gamma_a - \gamma_a^2)^{-\frac{1}{5}}$, *which is determined by the convergence property of the underlying Markov chain. Generally, a small* γ *will lead to a small coefficient, which means a Markov chain with fast converging speed has small generalization error. Note that the absolute spectral gap assumption can be relaxed to the pseudo spectral gap, the corresponding learning rate established in Theorem 2 remains the same order but the multiplicative coefficient* $2\gamma_a - \gamma_a^2$ *is replaced by* γ_p.

Remark 4 *It is well known that any bounded independent random variables* $Z_i \in [a_i, b_i]$ $(a_i \leq b_i, a_i, b_i \in \mathbb{R})$ *can be seen as the transformations of i.i.d. random variables* $U_i \sim \mathbf{Unif}[0,1]$ *via the inverse cumulative distribution functions* $F_{Z_i}^{-}1 : [0,1] \to [a_i, b_i]$, *i.e.* $Z_i = F_{Z_i}^{-1}(U_i)$. *Hence, the i.i.d. sequence* $\{U_i\}_{i\geq 1}$ *can be regarded as a stationary Markov chain on the state space* $[0,1]$ *with invariant measure* $\pi(dy) = dy$ *and transition kernel* $P(x, dy) = dy$. *This Markov chain has* $\gamma = 1$. *In this case, the generalization error in Theorem 2 reduces to the classical i.i.d. case, i.e.* $\mathcal{O}(m^{-\frac{1}{5}})$. *Note that such a learning rate is still better than the result in (Wang et al. 2017). The main reason is that we use empirical covering number to carefully characterize the capacity of function space while Wang et al. (Wang et al. 2017) adopts the Rademacher complexity as the measurement. Some regularity conditions can be imposed on the kernel function to further improve the learning rate.*

To evaluate the robustness of RMR within Markov-dependent observations, we introduce the concept of breakdown point (Donoho 1982), which measures the proportion of bad data in a dataset that an estimator can tolerate before returning arbitrary value.

Given a sample set $\mathbf{z} = \{(\mathbf{x}_i, y_i)\}_{i=1}^m$, the corrupted sample set $\mathbf{z} \cup \mathbf{z}'$ is constructed by adding n arbitrary points $\mathbf{z}' = \{(\mathbf{x}_{m+j}, y_{m+j})\}_{j=1}^n$, which contain a fraction $\frac{n}{m+n}$ of bad values. The finite sample contamination breakdown point $\epsilon(\boldsymbol{\alpha}_{\mathbf{z}})$ is defined as

$$\varepsilon^*(\boldsymbol{\alpha}_{\mathbf{z}}) = \min_{1\leq n\leq m} \Big\{ \frac{n}{m+n} : \sup_{\mathbf{z}'} \|\boldsymbol{\alpha}_{\mathbf{z}\cup\mathbf{z}'}\|_2 = \infty \Big\}, \quad (11)$$

Theorem 3 *Suppose* $\phi(u) = \phi(-u)$ *and* $\phi(t) \to 0$ *when* $|t| \to \infty$. *For a given Markov-dependent observations* $\mathbf{z}$, *and* λ, σ, *let*

$$N = \phi(0)^{-1} \sum_{i=1}^m \phi\Big(\frac{y_i - \mathbf{K}_i^\top \boldsymbol{\alpha}_{\mathbf{z}}}{\sigma}\Big) - \lambda \phi(0)^{-1} m \sigma \|\boldsymbol{\alpha}_{\mathbf{z}}\|_q^q.$$

Then the finite sample contamination breakdown point of $\boldsymbol{\alpha}_{\mathbf{z}}$ *in (8) is*

$$\varepsilon^*(\boldsymbol{\alpha}_{\mathbf{z}}) = \frac{n^*}{m+n^*},$$

where n^* *is an integer satisfying* $\lceil N \rceil \leq n^* \leq \lfloor N \rfloor + 1$, $\lceil a \rceil$ *denotes the largest integer not greater than* a *and* $\lfloor a \rfloor$ *denotes the smallest integer not less than* a.

Remark 5 *Theorem 3 indicates that the breakdown point relies on* $\phi(\cdot)$, *the turning parameter* λ, σ *and the sample configuration. As pointed out in (Huber 1992), the breakdown point can be quite high if the bandwidth parameter is only determined by training samples. However, with appropriate choice of* λ *and* σ *through some data driven strategies, RMR can still achieve a satisfactory learning rate and robustness.*

Proofs

This section presents the proof details of the main theorems. To be clear, we first list several useful lemmas which will be used in the proofs.

Lemma 2 *(Paulin 2015) (Bernstein inequality for reversible Markov Chains) Let* $X_1, X_2, \cdots, X_m$ *be a stationary reversible Markov chain with invariant distribution* π *and absolute spectral gap* γ_a. *Suppose that* $f_1, f_2, \cdots, f_m \in \mathcal{L}_2(\pi)$ *with* $|f_i - \mathbb{E}_\pi(f_i)| \leq C$, *denote* $S := \sum_{i=1}^m f_i(X_i)$ *and* $V_S := \sum_{i=1}^m$ *Var*$_\pi(f_i)$, *then for any* $t > 0$,

$$\mathbb{P}_\pi\Big(|S - \mathbb{E}_\pi(S)| \geq t\Big) \leq 2\exp\Big(-\frac{t^2(2\gamma_a - \gamma_a^2)}{8V_S + 20Ct}\Big). \quad (12)$$

Lemma 3 *(Cucker and Smale 2002) Let* $c_1, c_2 > 0$, *and* $p_1 > p_2 > 0$. *Then, the equation* $x^{p_1} - c_1 x^{p_2} - c_2 = 0$ *has unique positive zero* x^*. *In addition* $x^* \leq \max\{(2c_1)^{1/(p_1-p_2)}, (2c_2)^{1/p_1}\}$.

Proof of Theorem 1

PROOF. The proof of Theorem 1 consists of three steps below.

Step I: Bounding $\mathcal{S}_1(\mathbf{z})$. Define a random variable

$$\xi_1 := \sigma^{-1}\phi\Big(\frac{y - f^*(\mathbf{x})}{\sigma}\Big) - \sigma^{-1}\phi\Big(\frac{y - f_\lambda(\mathbf{x})}{\sigma}\Big), \mathbf{z} \in \mathcal{Z}.$$

According to the boundedness assumption of ϕ, it is easy to check that $|\xi_1(\mathbf{z})| \leq 2\|\phi\|_\infty/\sigma$. Furthermore, we see that

$$\begin{aligned}\mathrm{Var}(\xi_1) &= \mathbb{E}\Big[\sigma^{-1}\phi\Big(\frac{y - f^*(\mathbf{x})}{\sigma}\Big) - \sigma^{-1}\phi\Big(\frac{y - f_\lambda(\mathbf{x})}{\sigma}\Big)\Big]^2 \\ &\leq 2\frac{\|\phi\|_\infty}{\sigma}(\mathcal{R}^\sigma(f^*) + \mathcal{R}^\sigma(f_\lambda)).\end{aligned}$$

By Theorem 9 in (Feng, Fan, and Suykens 2020), we have

$$|\mathcal{R}^\sigma(f) - \mathcal{R}(f)| \le \frac{C_1\sigma^2}{2},$$

which implies $\mathcal{R}^\sigma(f^*) \le \mathcal{R}(f^*) + \frac{C_1}{2}\sigma^2$ and $\mathcal{R}^\sigma(f_\lambda) \le \mathcal{R}(f_\lambda) + \frac{C_1}{2}\sigma^2$, where C_1 is given in Lemma 1. These two inequalities together with the fact $\sigma \le 1$ yield

$$\begin{aligned}\mathrm{Var}(\xi_1) &\le \frac{2\|\phi\|_\infty}{\sigma}(\mathcal{R}(f^*) + \mathcal{R}(f_\lambda) + C_1\sigma^2)\\ &\le \frac{2\|\phi\|_\infty}{\sigma}(p_{f^*}(0) + p_{f_\lambda}(0) + C_1\sigma^2)\\ &\le C_2\sigma^{-1},\end{aligned}$$

where $C_2 = 2\|\phi\|_\infty(p_{f^*}(0) + p_{f_\lambda}(0) + C_1)$. Now applying Lemma 2 to the random variable ξ_1, we have

$$\mathcal{S}_1 \le \frac{20\|\phi\|_\infty \log(2/\delta)}{m\sigma(2\gamma_a - \gamma_a^2)} + 2\sqrt{\frac{2C_2\log(2/\delta)}{m\sigma(2\gamma_a - \gamma_a^2)}}.$$

with confidence at least $1 - \delta$.

Step II: Bounding $\mathcal{S}_2(\mathbf{z})$. To this end, we first prove that under assumptions 1 and 2, for any $f \in \mathcal{B}_R$ with $R \ge 1$ and $\varepsilon \ge C_1\sigma^2$, with confidence at least $1 - \delta$, it holds

$$\begin{aligned}&\mathbb{P}_{\mathbf{z}\in\mathcal{Z}^m}\Big\{\frac{\mathcal{R}^\sigma(f^*) - \mathcal{R}^\sigma(f) - (\mathcal{R}^\sigma_{\mathbf{z}}(f^*) - \mathcal{R}^\sigma_{\mathbf{z}}(f))}{\sqrt{\mathcal{R}^\sigma(f^*) - \mathcal{R}^\sigma(f) + 2\varepsilon}} > 4\sqrt{\varepsilon}\Big\}\\ &\le \mathcal{N}_2\left(\mathcal{B}_R, r\right)\exp\Big\{-\frac{(2\gamma_a - \gamma_a^2)m\varepsilon}{40(M+1)^2R^2(L_\phi\sigma^{-4} + L_\phi\sigma^{-2})}\Big\},\end{aligned} \tag{13}$$

where $r = \frac{\sigma^2\varepsilon}{L_\phi(M+1)R}$. To this end, we introduce a random variable defined by

$$\xi_2 = \sigma^{-1}\phi\Big(\frac{y - f^*(\mathbf{x})}{\sigma}\Big) - \sigma^{-1}\phi\Big(\frac{y - f(\mathbf{x})}{\sigma}\Big),$$

then it is easy to verify that $\mathbb{E}\xi_2 = \mathcal{R}^\sigma(f^*) - \mathcal{R}^\sigma(f)$ and for $\forall f \in \mathcal{B}_R$, $|\xi_2| \le \frac{L_\phi}{\sigma^2}\|f^* - f\|_\infty \le \frac{L_\phi}{\sigma^2}(M+1)R$, $|\xi_2 - \mathbb{E}\xi_2| \le \frac{2L_\phi}{\sigma^2}(M+1)R$ and $\mathrm{Var}(\xi_2) \le \mathbb{E}\xi_2^2 \le \frac{L_\phi^2}{\sigma^4}\|f^* - f\|_\infty^2$. Let $\{f_j\}_{j=1}^J$ be an r-net of the set $\mathcal{B}_R$ with J being the covering number of $\mathcal{N}_2(\mathcal{B}_R, r)$, and define

$$\mu = \sqrt{\mathcal{R}^\sigma(f^*) - \mathcal{R}^\sigma(f_j) + 2\varepsilon}.$$

According to Lemma 2, we get the following conclusion

$$\begin{aligned}&\mathbb{P}_{\mathbf{z}\in\mathcal{Z}^m}\Big\{\frac{\mathcal{R}^\sigma(f^*) - \mathcal{R}^\sigma(f_j) - (\mathcal{R}^\sigma_{\mathbf{z}}(f^*) - \mathcal{R}^\sigma_{\mathbf{z}}(f_j))}{\sqrt{\mathcal{R}^\sigma(f^*) - \mathcal{R}^\sigma(f_j) + 2\varepsilon}} > \sqrt{\varepsilon}\Big\}\\ &\le \exp\Big\{-\frac{(2\gamma_a - \gamma_a^2)m\mu^2\varepsilon}{8L_\phi^2\sigma^{-4}\|f^* - f_j\|_\infty^2 + 40L_\phi\sigma^{-2}(M+1)R\mu\sqrt{\varepsilon}}\Big\}\\ &\le \exp\Big\{-\frac{(2\gamma_a - \gamma_a^2)m\mu^2\varepsilon}{8L_\phi^2\sigma^{-4}(M+1)^2R^2\mu^2 + 40L_\phi\sigma^{-2}(M+1)^2R^2\mu\sqrt{\varepsilon}}\Big\}\\ &\le \exp\Big\{-\frac{(2\gamma_a - \gamma_a^2)m\varepsilon}{40(M+1)^2R^2(L_\phi^2\sigma^{-4} + L_\phi\sigma^{-2})}\Big\}.\end{aligned}$$

Since

$$\mu^2 = \mathcal{R}^\sigma(f^*) - \mathcal{R}^\sigma(f_j) + 2\varepsilon > \mathcal{R}^\sigma(f^*) - \mathcal{R}^\sigma(f_j) + \varepsilon \ge \varepsilon,$$

there exists some j such that $\|f - f_j\|_\infty \le \frac{\sigma^2\varepsilon}{L_\phi(M+1)R}$ for any $f \in \mathcal{B}_R$, hence both $|\mathcal{R}^\sigma(f) - \mathcal{R}^\sigma(f_j)|$ and $|\mathcal{R}^\sigma_{\mathbf{z}}(f) - \mathcal{R}^\sigma_{\mathbf{z}}(f_j)|$ can be bounded by ε, then we have the following inequalities

$$\frac{|\mathcal{R}^\sigma_{\mathbf{z}}(f^*) - \mathcal{R}^\sigma_{\mathbf{z}}(f) - (\mathcal{R}^\sigma_{\mathbf{z}}(f^*) - \mathcal{R}^\sigma_{\mathbf{z}}(f_j))|}{\sqrt{\mathcal{R}^\sigma(f^*) - \mathcal{R}^\sigma(f) + 2\varepsilon}} \le \sqrt{\varepsilon},$$

$$\frac{|\mathcal{R}^\sigma(f^*) - \mathcal{R}^\sigma(f) - (\mathcal{R}^\sigma(f^*) - \mathcal{R}^\sigma(f_j))|}{\sqrt{\mathcal{R}^\sigma(f^*) - \mathcal{R}^\sigma(f) + 2\varepsilon}} \le \sqrt{\varepsilon}.$$

These two inequalities together with the fact $\varepsilon < \mathcal{R}^\sigma(f^*) - \mathcal{R}^\sigma(f) + 2\varepsilon$ yield the following inequality

$$\begin{aligned}&\mathcal{R}^\sigma(f^*) - \mathcal{R}^\sigma(f_j) + 2\varepsilon = \mathcal{R}^\sigma(f^*) - \mathcal{R}^\sigma(f_j) -\\ &\qquad(\mathcal{R}^\sigma(f^*) - \mathcal{R}^\sigma(f)) + \mathcal{R}^\sigma(f^*) - \mathcal{R}^\sigma(f) + 2\varepsilon\\ &\le \sqrt{\varepsilon}\sqrt{\mathcal{R}^\sigma(f^*) - \mathcal{R}^\sigma(f) + 2\varepsilon} + \mathcal{R}^\sigma(f^*) - \mathcal{R}^\sigma(f) + 2\varepsilon\\ &\le 2(\mathcal{R}^\sigma(f^*) - \mathcal{R}^\sigma(f) + 2\varepsilon),\end{aligned}$$

hence

$$\frac{\mathcal{R}^\sigma(f^*) - \mathcal{R}^\sigma(f) - (\mathcal{R}^\sigma_{\mathbf{z}}(f^*) - \mathcal{R}^\sigma_{\mathbf{z}}(f))}{\sqrt{\mathcal{R}^\sigma(f^*) - \mathcal{R}^\sigma(f) + 2\varepsilon}} > 4\sqrt{\varepsilon}$$

for $\forall f \in \mathcal{B}_R$. We further get

$$\frac{\mathcal{R}^\sigma(f^*) - \mathcal{R}^\sigma(f_j) - (\mathcal{R}^\sigma_{\mathbf{z}}(f^*) - \mathcal{R}^\sigma_{\mathbf{z}}(f_j))}{\sqrt{\mathcal{R}^\sigma(f^*) - \mathcal{R}^\sigma(f_j) + 2\varepsilon}} > \sqrt{\varepsilon},$$

which implies

$$\begin{aligned}&\mathbb{P}_{\mathbf{z}\in\mathcal{Z}^m}\Big\{\frac{\mathcal{R}^\sigma(f^*) - \mathcal{R}^\sigma(f) - (\mathcal{R}^\sigma_{\mathbf{z}}(f^*) - \mathcal{R}^\sigma_{\mathbf{z}}(f))}{\sqrt{\mathcal{R}^\sigma(f^*) - \mathcal{R}^\sigma(f) + 2\varepsilon}} > 4\sqrt{\varepsilon}\Big\}\\ &\le \sum_{i=1}^{J}\mathbb{P}_{\mathbf{z}\in\mathcal{Z}^m}\Big\{\frac{\mathcal{R}^\sigma(f^*) - \mathcal{R}^\sigma(f_j) - (\mathcal{R}^\sigma_{\mathbf{z}}(f^*) - \mathcal{R}^\sigma_{\mathbf{z}}(f_j))}{\sqrt{\mathcal{R}^\sigma(f^*) - \mathcal{R}^\sigma(f_j) + 2\varepsilon}} > \sqrt{\varepsilon}\Big\}\\ &\le \mathcal{N}_2\left(\mathcal{B}_R, r\right)\exp\Big\{-\frac{(2\gamma_a - \gamma_a^2)m\varepsilon}{40(M+1)^2R^2(L_\phi^2\sigma^{-4} + L_\phi\sigma^{-2})}\Big\}.\end{aligned}$$

We know from (13) that

$$\begin{aligned}&\mathbb{P}_{\mathbf{z}\in\mathcal{Z}^m}\Big\{\sup_{f_{\mathbf{z}}\in\mathcal{B}_R}\frac{\mathcal{R}^\sigma(f^*) - \mathcal{R}^\sigma(f_{\mathbf{z}}) - (\mathcal{R}^\sigma_{\mathbf{z}}(f^*) - \mathcal{R}^\sigma_{\mathbf{z}}(f_{\mathbf{z}}))}{\sqrt{\mathcal{R}^\sigma(f^*) - \mathcal{R}^\sigma(f_{\mathbf{z}}) + 2\varepsilon}} > 4\sqrt{\varepsilon}\Big\}\\ &\le \mathcal{N}_2\left(\mathcal{B}_R, r\right)\exp\Big\{-\frac{(2\gamma_a - \gamma_a^2)m\varepsilon}{40(M+1)^2R^2(L_\phi^2\sigma^{-4} + L_\phi\sigma^{-2})}\Big\}\\ &\le \exp\Big\{c_s\Big(\frac{1}{r}\Big)^s - \frac{(2\gamma_a - \gamma_a^2)m\varepsilon}{40(M+1)^2R^2(L_\phi^2\sigma^{-4} + L_\phi\sigma^{-2})}\Big\},\end{aligned} \tag{14}$$

Set the last term of inequality (14) equal to δ, and we get

$$\begin{aligned}&\varepsilon^{s+1} - \frac{40(M+1)^2R^2(L_\phi^2\sigma^{-4} + L_\phi\sigma^{-2})\log(2/\delta)}{(2\gamma_a - \gamma_a^2)m}\cdot\varepsilon^s\\ &\quad - \frac{40c_sL_\phi^{s+2}(\sigma^{-4-2s} + \sigma^{-2-2s})(M+1)^{s+2}R^{s+2}}{(2\gamma_a - \gamma_a^2)m} = 0.\end{aligned}$$

By Lemma 3, we obtain the upper bound of the smallest positive solution ε^Δ for the above equation, i.e.

$$\varepsilon^\Delta := C_3(2\gamma_a - \gamma_a^2)^{-\frac{1}{1+s}}\sigma^{-\frac{4+2s}{1+s}}m^{-\frac{1}{1+s}}R^{\frac{s+2}{s+1}}\log(2/\delta),$$

where $C_3 := \max\{80(M+1)L_\phi, (80c_s)^{\frac{1}{1+s}}(M+1)^{\frac{s+2}{s+1}}L_\phi^{\frac{s+2}{s+1}}\}$. Then, we have for $f_{\mathbf{z}} \in \mathcal{B}_R$

$$\begin{aligned}\mathcal{S}_2(\mathbf{z}) &= \mathcal{R}^\sigma(f^*) - \mathcal{R}^\sigma(f_{\mathbf{z}}) - (\mathcal{R}^\sigma_{\mathbf{z}}(f^*) - \mathcal{R}^\sigma_{\mathbf{z}}(f_{\mathbf{z}})) \\ &\leq 4\sqrt{\varepsilon^\Delta} \cdot \sqrt{\mathcal{R}^\sigma(f^*) - \mathcal{R}^\sigma(f_{\mathbf{z}}) + 2\varepsilon^\Delta} \\ &\leq \frac{1}{2}(\mathcal{R}^\sigma(f^*) - \mathcal{R}^\sigma(f_{\mathbf{z}})) + 9\varepsilon^\Delta.\end{aligned} \tag{15}$$

Step III: Let $\{\psi_1, \psi_2, \cdots\}$ be an orthonormal basis of $\mathcal{L}^2_{\rho_X}(X)$ and $\{\lambda_1, \lambda_2, \cdots\}$ be the corresponding eigenvalue with descending order. Recall that $f^* \in L_K^\beta g$ for some $0 < \beta \leq 2$ and $g \in \mathcal{L}^2_{\rho_X}$, then $f^* = \sum_{\lambda_k \geq 0} \alpha_k \lambda_k^\beta \psi_k$ and g can be uniquely written as $g = \sum_{\lambda_k \geq 0} \alpha_k \psi_k$ with $\|g\|^q_{\mathcal{L}^2_{\rho_X}} = \sum_{\lambda_k \geq 0} |\alpha_k|^q \leq \infty$. Assume that $\lambda_1 \leq \lambda^{\frac{\beta-2}{2+\beta}}$, we get $\Omega_q(f^*) \leq \sum_{\lambda_k \geq 0} |\alpha_k|^q \lambda_k^\beta \leq \|g\|^q_{\mathcal{L}^2_{\rho_X}} \lambda^{\frac{\beta-2}{2+\beta}}$. Based on the estimations in **Step I** and **II**, we see that with confidence at least $1-\delta$,

$$\begin{aligned}\mathcal{R}(f^*) - \mathcal{R}(f_{\mathbf{z}}) \leq{}& \frac{20\|\phi\|_\infty \log(2/\delta)}{m\sigma(2\gamma_a - \gamma_a^2)} + 2\sqrt{\frac{2C_2\log(2/\delta)}{m\sigma(2\gamma_a - \gamma_a^2)}} \\ &+ 18C_3(2\gamma_a - \gamma_a^2)^{-\frac{1}{1+s}}\sigma^{-\frac{4+2s}{1+s}}m^{-\frac{1}{1+s}}R^{\frac{s+2}{s+1}}\log(2/\delta) \\ &+ 2\lambda^{\frac{2\beta}{2+\beta}} + 2C_1\sigma^2 \\ \leq{}& C\log(2/\delta)\Big((2\gamma_a - \gamma_a^2)^{-\frac{1}{2}}m^{-\frac{1}{2}}\sigma^{-\frac{1}{2}} + (2\gamma_a - \gamma_a^2)^{-\frac{1}{1+s}} \\ &\cdot m^{-\frac{1}{1+s}}\sigma^{-\frac{4+2s}{1+s}}R^{\frac{s+2}{s+1}} + \sigma^2 + \lambda^{\frac{2\beta}{2+\beta}}\Big),\end{aligned}$$

where $C = 180C_1C_3M^2\|\phi\|_\infty\|g\|^q_{\mathcal{L}^2_{\rho_X}}$ is a constant independent of m, δ, σ and λ. We complete the proof. ■

Proof of Theorem 2

PROOF. From the definition of $f_{\mathbf{z}}$, we know that $\mathcal{R}^\sigma_{\mathbf{z}}(f_{\mathbf{z}}) - \lambda\Omega_q(f_{\mathbf{z}}) \geq \mathcal{R}^\sigma_{\mathbf{z}}(0)$, then $\lambda\Omega_q(f_{\mathbf{z}}) \leq \mathcal{R}^\sigma_{\mathbf{z}}(f_{\mathbf{z}}) - \mathcal{R}^\sigma_{\mathbf{z}}(0) \leq \frac{2\|\phi\|_\infty}{\sigma}$, which implies $\Omega_q(f_{\mathbf{z}}) \leq 2\|\phi\|_\infty\lambda^{-1}\sigma^{-1}$. Hence taking $R = 2\|\phi\|_\infty\lambda^{-1}\sigma^{-1}$ together with Theorem 1 yield

$$\begin{aligned}\mathcal{R}(f^*) - \mathcal{R}(f_{\mathbf{z}}) &\leq \hat{C}\log(2/\delta)\Big((2\gamma_a - \gamma_a^2)^{-\frac{1}{2}}m^{-\frac{1}{2}}\sigma^{-\frac{1}{2}} \\ &+ (2\gamma_a - \gamma_a^2)^{-\frac{1}{1+s}}m^{-\frac{1}{1+s}}\sigma^{-\frac{6+3s}{1+s}}\lambda^{-\frac{s+2}{s+1}} + \sigma^2 + \lambda^{\frac{2\beta}{\beta+2}}\Big).\end{aligned}$$

By taking $\lambda = (2\gamma_a - \gamma_a^2)^{-\frac{\theta}{\beta}}m^{-\frac{\theta}{\beta}}$, $\sigma = (2\gamma_a - \gamma_a^2)^{-\frac{\theta}{2\beta}}m^{-\frac{\theta}{2\beta}}$ and $\theta = \frac{2\beta}{8\beta+5s\beta+2s+4}$ with confidence at least $1-\delta$, it holds

$$\mathcal{R}(f^*) - \mathcal{R}(f_{\mathbf{z}}) \leq \hat{C}\log(2/\delta)(2\gamma_a - \gamma_a^2)^{-\frac{\theta}{\beta}}m^{-\theta}.$$

This completes the proof. ■

Proof of Theorem 3

PROOF. Observe that the RMR optimization problem (8) is equivalent to

$$\max_{\boldsymbol{\alpha}\in\mathbb{R}^m}\Big\{\phi(0)^{-1}\sum_{i=1}^m \phi\Big(\frac{y_i - \mathbf{K}_i^\top\boldsymbol{\alpha}_{\mathbf{z}}}{\sigma}\Big) - \lambda\phi(0)^{-1}m\sigma\|\boldsymbol{\alpha}_{\mathbf{z}}\|_q^q\Big\}. \tag{16}$$

Let $\phi^*(t) = \phi(t)/\phi(0)$. Then for any t, $\phi^*(t) \leq \phi^*(0) = 1$ and $\phi^*(\cdot)$ decreases monotonely toward both sides and that $\phi^*(t) = 0$ for $|t| \to \infty$.

We first show that $\boldsymbol{\alpha}_{\mathbf{z}\cup\mathbf{z}'}$ is bounded when $n < N$. To this end, suppose there exists a $\nu > 0$ such that $n + m\nu < N$. Let $\phi^*(t) \leq \nu$ for $t \geq C$ and $\boldsymbol{\alpha}$ be any real vector such that $|y - \mathbf{K}^\top\boldsymbol{\alpha}| \geq C$ for all $(\mathbf{x}, y) \in \mathbf{z}$. Then we have

$$\sum_{i=1}^{m+n}\phi^*(y_i - \mathbf{K}_i^\top\boldsymbol{\alpha}_{\mathbf{z}}) - \lambda m\sigma\|\boldsymbol{\alpha}_{\mathbf{z}}\|_q^q \geq N, \tag{17}$$

and

$$\begin{aligned}&\sum_{i=1}^{m+n}\phi^*(y_i - \mathbf{K}_i^\top\boldsymbol{\alpha}) - \lambda m\sigma\|\boldsymbol{\alpha}\|_q^q \\ &\leq \sum_{i=m+1}^{m+n}\phi^*(y_i - \mathbf{K}_i^\top\boldsymbol{\alpha}) + \sum_{i=1}^m\phi^*(y_i - \mathbf{K}_i^\top\boldsymbol{\alpha}) \\ &\leq n + m\nu.\end{aligned} \tag{18}$$

From (17) and (18), we know that $\boldsymbol{\alpha}_{\mathbf{z}\cup\mathbf{z}'}$ must satisfies $|y - \mathbf{K}^\top\boldsymbol{\alpha}_{\mathbf{z}\cup\mathbf{z}'}| < C$ for a sample in $\mathbf{z}$.

On the other hand, if $n > N$, let $\nu > 0$, such that $n - n\nu > N$, and let C be such that $\phi^*(t) \leq \nu$ for $|t| \geq C$. Assume that all points in $\mathbf{z}'$ are the same and satisfy $y = \mathbf{K}_i^\top\boldsymbol{\alpha}^*$. Let $\boldsymbol{\alpha}$ be any vector such that $|y_{m+1} - \mathbf{K}_{m+1}^\top\boldsymbol{\alpha}| < C$. Then $\sum_{i=1}^{m+n}\phi^*(y_i - \mathbf{K}_i^\top\boldsymbol{\alpha}) - \lambda m\sigma\|\boldsymbol{\alpha}\|_q^q \leq N + n\nu$ and $\sum_{i=1}^{m+n}\phi^*(y_i - \mathbf{K}_i^\top\boldsymbol{\alpha}^*) - \lambda m\sigma\|\boldsymbol{\alpha}^*\|_q^q \geq n$. These inequalities imply that $|y_{m+1} - \mathbf{K}_{m+1}^\top\boldsymbol{\alpha}_{\mathbf{z}\cup\mathbf{z}'}| \leq C$. Hence $\boldsymbol{\alpha}_{\mathbf{z}\cup\mathbf{z}'}$ is bounded as $n < N$. Observe that $\|\boldsymbol{\alpha}_{\mathbf{z}\cup\mathbf{z}'}\| \to \infty$ when $y_{m+1} \to \infty$ with $\mathbf{K}_{m+1}$ fixed, and we have the breakdown. ■

Conclusions

In this paper, we investigate the generalization performance of regularized modal regression under Markov-dependence setup. The statistical consistency is established and an explicit learning rate is given as well. Our results show that the Markov dependence impacts on the generalization error in the way that sample size would be discounted by a multiplicative factor depending on the spectral gap of underlying Markov chain. Moreover, the study brings us some insights into robust regression within Markov-dependent setup. It will be interesting to improve the learning rate obtained in current study by imposing some regularity conditions on the kernel function.

Acknowledgments

This work was supported by National Key Research and Development Program of China(2020AAA0108800), National Natural Science Foundation of China (62106191, 12071166, 61772409, 62050194, 61721002), Innovation Research Team of Ministry of Education (IRT_17R86), Project of China Knowledge Centre for Engineering Science and Technology and Project of Chinese academy of engineering "The Online and Offline Mixed Educational ServiceSystem for 'The Belt and Road' Training in MOOC China".

References

Chen, H.; and Wang, Y. 2018. Kernel-based sparse regression with the correntropy-induced loss. *Applied and Computational Harmonic Analysis*, 44(1): 144–164.

Chen, Y.-C.; Genovese, C. R.; Tibshirani, R. J.; Wasserman, L.; et al. 2016. Nonparametric modal regression. *The Annals of Statistics*, 44(2): 489–514.

Cucker, F.; and Smale, S. 2001. On the mathematical foundations of learning. *Bull. Amer. Math. Soc.*, 39(1): 1–49.

Cucker, F.; and Smale, S. 2002. Best choices for regularization parameters in learning theory: on the bias-variance problem. *Found. Comput. Math.*, 2(4): 413–428.

Donoho, D. L. 1982. Breakdown properties of multivariate location estimators. Technical report, Ph. D. Qualifying paper, Department of Statistics, Harvard University.

Fan, J.; Jiang, B.; and Sun, Q. 2018. Hoeffding's lemma for Markov Chains and its applications to statistical learning. *arXiv preprint arXiv:1802.00211*.

Feng, Y.; Fan, J.; and Suykens, J. 2020. A statistical learning approach to modal regression. *J. Mach. Learn. Res.*, 21: 1–35.

Gong, T.; Xi, Q.; and Xu, C. 2020. Robust Gradient-Based Markov Subsampling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 4004–4011.

Gong, T.; Zou, B.; and Xu, Z. 2015. Learning With ℓ_1-Regularizer Based on Markov Resampling. *IEEE Transactions on Cybernetics*, 46(5): 1189–1201.

Huber, P. J. 1992. Robust estimation of a location parameter. In *Breakthroughs in statistics*, 492–518. Springer.

Khardani, S.; and Yao, A. F. 2017. Non linear parametric mode regression. *Communications in Statistics-Theory and Methods*, 46(6): 3006–3024.

Lee, M.-j. 1989. Mode regression. *Journal of Econometrics*, 42(3): 337–349.

Lv, Z.; Zhu, H.; and Yu, K. 2014. Robust variable selection for nonlinear models with diverging number of parameters. *Statistics & Probability Letters*, 91: 90–97.

McCracken, M. W.; and Ng, S. 2016. FRED-MD: A monthly database for macroeconomic research. *Journal of Business & Economic Statistics*, 34(4): 574–589.

Meinshausen, N.; and Ridgeway, G. 2006. Quantile regression forests. *Journal of Machine Learning Research*, 7(6).

Meyn, S. P.; and Tweedie, R. L. 2012. *Markov chains and stochastic stability*. Springer Science & Business Media.

Nikolova, M.; and Ng, M. K. 2005. Analysis of half-quadratic minimization methods for signal and image recovery. *SIAM Journal on Scientific computing*, 27(3): 937–966.

Paulin, D. 2015. Concentration inequalities for Markov chains by Marton couplings and spectral methods. *Electronic Journal of Probability*, 20(79): 1–32.

Rudolf, D. 2011. Explicit error bounds for Markov chain Monte Carlo. *arXiv preprint arXiv:1108.3201*.

Ryali, S.; Chen, T.; Supekar, K.; and Menon, V. 2012. Estimation of functional connectivity in fMRI data using stability selection-based sparse partial correlation with elastic net penalty. *NeuroImage*, 59(4): 3852–3861.

Shi, L.; Huang, X.; Feng, Y.; and Suykens, J. 2019. Sparse Kernel Regression with Coefficient-based lq-Regularization. *Journal of Machine Learning Research*, 20.

Smith, S. M. 2012. The future of FMRI connectivity. *Neuroimage*, 62(2): 1257–1266.

Steinwart, I.; and Christmann, A. 2008. *Support vector machines*. Springer Science & Business Media.

Tibshirani, R. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1): 267–288.

Wang, X.; Chen, H.; Cai, W.; Shen, D.; and Huang, H. 2017. Regularized modal regression with applications in cognitive impairment prediction. *Advances in neural information processing systems*, 30: 1448–1458.

Wu, Q.; Ying, Y.; and Zhou, D.-X. 2007. Multi-kernel regularized classifiers. *J.Complexity*, 23(1): 108–134.

Yao, W.; and Li, L. 2014. A new regression model: modal linear regression. *Scandinavian Journal of Statistics*, 41(3): 656–671.

Yao, W.; and Xiang, S. 2016. Nonparametric and varying coefficient modal regression. *arXiv preprint arXiv:1602.06609*.

Yu, K.; and Aristodemou, K. 2012. Bayesian mode regression. *arXiv preprint arXiv:1208.0579*.

Yu, K.; Lu, Z.; and Stander, J. 2003. Quantile regression: applications and current research areas. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 52(3): 331–350.

Zhou, D. 2002. The covering number in learning theory. *J. Complexity*, 18: 739–767.

Zhou, D. 2003. Capacity of reproducing kernel space in learning theory. *IEEE Trans. Inf. Theory.*, 49(7): 1743–1752.

Zhou, H.; Huang, X.; et al. 2016. Nonparametric modal regression in the presence of measurement error. *Electronic Journal of Statistics*, 10(2): 3579–3620.