Electronic Journal of Statistics
Vol. 11 (2017) 1326–1350
ISSN: 1935-7524
DOI: 10.1214/17-EJS1265

Divide and conquer local average regression*

Xiangyu Chang†

Center of Data Science and Information Quality
School of Management
Xi'an Jiaotong University, Xi'an, China
e-mail: xiangyuchang@xjtu.edu.cn

Shao-Bo Lin‡

Department of Statistics
College of Mathematics and Information Science
Wenzhou University, Wenzhou, China
e-mail: sblin1983@gmail.com

and

Yao Wang§

Department of Statistics
School of Mathematics and Statistics
Xi'an Jiaotong University, Xi'an, China
e-mail: yao.s.wang@gmail.com

Abstract: The divide and conquer strategy, which breaks a massive data set into a series of manageable data blocks, and combines the independent results of data blocks to obtain a final decision, has been recognized as a state-of-the-art method to overcome challenges of massive data analysis. In this paper, we equip the classical local average regression with some divide and conquer strategies to infer the regressive relationship of input-output pairs from a massive data set. When the average mixture, a widely used divide and conquer approach, is adopted, we prove that the optimal learning rate can be achieved under some restrictive conditions on the number of data blocks. We then propose two variants to relax (or remove) these conditions and derive the same optimal learning rates as that for the average mixture local average regression. Our theoretical assertions are verified by a series of experimental studies.

MSC 2010 subject classifications: 62G08.

*Authors make an equal contribution to this paper. Lin is the corresponding author.

†Chang was partially supported by the National Natural Science Foundation of China (Project No. 11401462, 61603162) and the China Postdoctoral Science Foundation (Project No. 2015M582630).

‡Lin was partially supported by the National Natural Science Foundation of China (Project No. 61502342).

§Wang was partially supported by the National Natural Science Foundation of China (Project No. 11501440).

Keywords and phrases: Divide and conquer strategy, local average regression, Nadaraya-Watson estimate, k nearest neighbor estimate.

Received March 2016.

Contents

1. Introduction

Rapid expansion of capacity in the automatic data acquisition has made a profound impact on statistics and machine learning, as it brings data of unprecedented size and complexity. These data are generally called as the *massive data* or *big data* [28]. Massive data bring new opportunities of discovering subtle population patterns and heterogeneities which are believed to embody rich values and are impossible to be found in relatively small data sets. They, however, simultaneously lead to a series of challenges such as the storage bottleneck, efficient computation, and so on [32].

To handle the aforementioned challenges, some divide and conquer strategies were suggested and widely used in statistical and machine learning communities [17, 7, 14, 30, 29, 1, 26, 15, 4]. These approaches firstly decompose a massive data set into m data blocks, then run some specified learning algorithm on each data block independently to get a *local estimate* $\hat{f}_j, j = 1, \ldots, m$ and finally transmit the m local estimates into one machine to synthesize a *global estimate* $\bar{f}$, which is expected to model the structure of original massive data. A practical and exclusive synthesizing method is the *average mixture* (AVM) [17, 14, 30, 29, 15], i.e., $\bar{f} = \frac{1}{m}\sum_{j=1}^{m} \hat{f}_j$.

In practice, divide and conquer strategies have many applicable scenarios. We show the following three situations as motivating examples. The first one focuses on using limited primary memory to handle a massive data set. In this situation, the divide and conquer strategy is regarded as a two-stage procedure. In the first stage, it reads the whole data set sequentially block by block with manageable sample size and derives a local estimate based on each block. In the second stage, it synthesizes local estimates to build up a global estimate [14]. The second motivating example refers to using distributed data management systems to tackle massive data. In this situation, distributed data management systems (e.g., Hadoop) are designed by some divide and conquer strategies. They can load the whole data set into the systems and tackle computational tasks separably and automatically. Guha et al. [10] have developed an integrated programing environment of R and Hadoop (called RHIPE) for expedient and efficient statistical computing. The third example is the massive data privacy. In this situation, it divides a massive data set into several small pieces and combines the estimates derived from these pieces for keeping the data privacy [7, 4].

For nonparametric regression, the average mixture has been shown to be efficient and feasible for global modeling methods such as conditional maximum entropy model [17], kernel ridge regression [29, 15, 4], kernel-based gradient descent [16] and kernel-based spectral algorithms [3, 11]. Compared with these global modeling methods, local average regression (LAR) [12, 8, 25], such as the Nadaraya-Watson kernel (NWK) and k nearest neighbor (KNN) estimates, which is by definition a learning scheme that averages outputs whose corresponding inputs satisfy certain localization assumptions, is recognized in the literature [12] to possess lower computational burden and therefore, is widely used in image processing [24], recommendation system [2] and financial engineering [13]. A natural idea to use LAR on massive data is to combine it with the average mixture strategy to produce a new learning scheme called average mixture local average regression (AVM-LAR), just as [29] did for the kernel ridge regression.

Our first purpose is to analyze the performance of AVM-LAR. We show that AVM-LAR can achieve the optimal learning rate of LAR on the whole data set under some strong restrictions on m, the number of data blocks. We prove that these restrictions cannot be essentially relaxed, which makes AVM-LAR feasible only for small m. Therefore, different from the AVM version of the global modeling in the literature [29, 15, 3, 4, 11, 16], AVM-LAR does not bring essential improvements over LAR, since we must pay much attention to determine an appropriate m.

Our second purpose is to pursue other divide and conquer strategies to equip LAR efficiently. In particular, we provide two concrete variants of AVM-LAR in this paper. The first variant is motivated by the difference between KNN and NWK, since the range of m to guarantee the optimal learning rate of AVM-KNN is much larger than AVM-NWK in our simulation studies. We attribute the reason to the data dependent property of localization parameter of KNN. Therefore, we slightly modify AVM-LAR by adopting a data dependent localization parameters of each data block. We establish optimal learning rates of

this variant under mild restriction on m and verify its feasibility by numerical simulations. The second variant is based on the definitions of AVM and LAR. It follows from the definition of LAR that the predicted value of a new input depends on samples near the input. If there are no such samples in a specified data block, then this data block doesn't affect the prediction of LAR. However, AVM averages local estimates directly, neglecting the concrete value of a specified local estimate, which usually leads to an inaccurate prediction. Based on this observation, we propose another variant of AVM-LAR by distinguishing whether a specified data block affects the prediction. We provide the optimal learning rate of this variant without any restriction on m and also present the experimental verifications.

To complete the above missions, the rest of paper is organized as follows. In Section 2, we present optimal learning rates of LAR and AVM-LAR and analyze the pros and cons of AVM-LAR. In Section 3, we propose two new modified AVM-LARs to improve the performance of AVM-LAR. A set of simulation studies to support the correctness of our assertions are given in Section 4. In Section 5, we prove all the theorems detailedly. In Section 6, we present the conclusion and some useful remarks.

2. Divide and conquer local average regression

In this section, after introducing some basic concepts of LAR, we present optimal learning rates of LAR. Then we derive optimal learning rates of AVM-LAR and analyze its pros and cons.

2.1. Local average regression

Let $D^N = \{(X_i, Y_i)\}_{i=1}^N$ be the data set where $X_i \in \mathcal{X} \subseteq \mathbb{R}^d$ is a explanatory variable and $Y_i \in [-M, M]$ is the real-valued response for some $0 < M < \infty$. We always assume $\mathcal{X}$ is a compact set. Suppose that samples are drawn independently according to an unknown joint distribution ρ over $\mathcal{X} \times [-M, M]$. The main aim of nonparametric regression is to construct a function $\hat{f} : \mathcal{X} \to [-M, M]$ that can describe future responses based on new inputs. The quality of the estimate $\hat{f}$ is measured in terms of the *mean-squared prediction error* $\mathbf{E}\{\hat{f}(X) - Y\}^2$, which is minimized by the so-called *regression function* $f_\rho(x) = \mathbf{E}\{Y|X = x\}$.

LAR, as one of the most widely used nonparametric regression approaches, constructs an estimate formed as

$$\hat{f}_{D^N,h}(x) = \sum_{i=1}^{N} W_{h,X_i}(x) Y_i, \tag{2.1}$$

where the localization weight W_{h,X_i} satisfies $W_{h,X_i}(x) > 0$ and $\sum_{i=1}^N W_{h,X_i}(x) = 1$. Here, $h > 0$ is the so-called localization parameter reflecting the extent of localization. Its value may depend on the data and the query point x. Generally

speaking, $W_{h,X_i}(x)$ is *small* if X_i is *far* from x. Two widely used examples of LAR are the Nadaraya-Watson kernel (NWK) and k nearest neighbor (KNN) estimates.

Example 2.1. *(NWK estimate) Let $K : \mathcal{X} \to \mathbb{R}_+$ be a kernel function [12], and $h > 0$ be its localization parameter. The NWK estimate is defined by*

$$\hat{f}_h(x) = \frac{\sum_{i=1}^N K\left(\frac{x-X_i}{h}\right) Y_i}{\sum_{i=1}^N K\left(\frac{x-X_i}{h}\right)}, \tag{2.2}$$

and therefore,

$$W_{h,X_i}(x) = \frac{K\left(\frac{x-X_i}{h}\right)}{\sum_{i=1}^N K\left(\frac{x-X_i}{h}\right)}.$$

It is worth noting that we use the convention $\frac{0}{0} = 0$ throughout this paper. Two popular kernel functions are the naive kernel, $K(x) = I_{\{\|x\|\leq 1\}}$ and Gaussian kernel $K(x) = \exp\left(-\|x\|^2\right)$, where I_A is an indicator function with the feasible domain $A \subset \mathcal{X}$ and $\|\cdot\|$ denotes the Euclidean norm. In the NWK estimate, the localization parameter depends only on the size of data.

Example 2.2. *(KNN estimate) For $x \in \mathcal{X}$, let $\{(X_{(i)}(x), Y_{(i)}(x))\}_{i=1}^N$ be a permutation of $\{(X_i, Y_i)\}_{i=1}^N$ such that*

$$\|x - X_{(1)}(x)\| \leq \cdots \leq \|x - X_{(N)}(x)\|.$$

Then the KNN estimate is defined by

$$\hat{f}_k(x) = \frac{1}{k}\sum_{i=1}^k Y_{(i)}(x). \tag{2.3}$$

According to (2.1), we have

$$W_{h,X_i}(x) = \begin{cases} 1/k, & \text{if } X_i \in \{X_{(1)}, \ldots, X_{(k)}\}, \\ 0, & \text{otherwise.} \end{cases}$$

Here we denote the weight of KNN as W_{h,X_i} instead of W_{k,X_i} for the sake of unity and $h = \|x - X_{(k)}(x)\|$ depends on the distribution of data and the query point x.

2.2. Optimal learning rate of LAR

The weakly universal consistency and optimal learning rates of some specified examples of LAR have been justified by [21, 22, 23] and summarized in [12]. In particular, Theorem 4.1 in [12] presented a sufficient condition to guarantee the weakly universal consistency of LAR. Theorem 5.2 and 6.2 in [12] deduced optimal learning rates of NWK and KNN. The aim of this subsection is to present some sufficient conditions to guarantee optimal learning rates of general LAR.

Generally, it is impossible to obtain a nontrivial rate of convergence result for arbitrary learning algorithm without imposing strong restrictions on ρ [12, Theorem 3.1], even when the output is bounded, i.e., $Y \in [-M, M]$ almost surely. A large portion of statistical learning theory [6, 20, 31, 18, 19] proceeds under the assumption that f_ρ is in a known set possessing some regularity. For $r, c_0 > 0$, let $\mathcal{F}^{c_0,r} = \{f | f : \mathcal{X} \to \mathcal{Y}, |f(x) - f(x')| \le c_0 \|x - x'\|^r, \forall x, x' \in \mathcal{X}\}$. We suppose in this paper that $f_\rho \in \mathcal{F}^{c_0,r}$, since this prior assumption has been employed in [12, 25, 29]. In this way, we present a baseline of our analysis in the following proposition, in terms of providing optimal learning rates for LAR under some mild conditions on the weights $\{W_{h,X_i}(x)\}_{i=1}^N$. Throughout the paper, we assume that h is data-independent for the sake of brevity. Similar results of data-dependent parameter can be derived by using the similar approach as that in the proof of Theorem 3.1 in Section 5.

Proposition 2.1. *Let $\hat{f}_{D^N,h}$ be defined by (2.1) and the localization weight W_{h,X_i} satisfy $W_{h,X_i}(x) > 0$ and $\sum_{i=1}^N W_{h,X_i}(x) = 1$. Assume further that:*

(A) there exists a positive number c_1 such that

$$\mathbf{E}\left\{\sum_{i=1}^N W_{h,X_i}^2(X)\right\} \le \frac{c_1}{Nh^d};$$

(B) there exists a positive number c_2 such that

$$\mathbf{E}\left\{\sum_{i=1}^N W_{h,X_i}(X) I_{\{\|X - X_i\| > h\}}\right\} \le \frac{c_2}{\sqrt{Nh^d}}.$$

If $h \sim N^{-1/(2r+d)}$, then there exist constants C_0 and C_1 depending only on d, r, c_0, c_1 and c_2 such that

$$C_0 N^{-2r/(2r+d)} \le \sup_{f_\rho \in \mathcal{F}^{c_0,r}} \mathbf{E}\{\|\hat{f}_{D^N,h} - f_\rho\|_\rho^2\} \le C_1 N^{-2r/(2r+d)} \tag{2.4}$$

where $\|f\|_\rho = (\int_{\mathcal{X}} |f(x)|^2 d\rho_X)^{1/2}$ and ρ_X is the marginal distribution of ρ.

Proposition 2.1 presents sufficient conditions of the localization weights to ensure the optimal learning rate of LAR. There are totally four constraints of the localization weight $W_{h,X_i}(\cdot)$. The first one is the averaging constraint $\sum_{i=1}^N W_{h,X_i}(x) = 1$, for all $X_i, x \in \mathcal{X}$. It essentially reflects the word *average* in LAR. The second one is the non-negative constraint. We regard it as a mild constraint as it holds for all the widely used LAR such as NWK and KNN. The third constraint is condition (A), which devotes to controlling the scope of the weights. It aims at avoiding the extreme case that there is a very large weight near 1 and others are almost 0. The last constraint is condition (B), which implies the localization property of LAR.

Proposition 2.1 is a direct generalization of Theorems 4.1, 5.2 and 6.2 in [12] and is important for our analysis, since it provides a sanity-check that an efficient AVM-LAR estimate should possess the similar learning rate as (2.4).

Furthermore, equipping LAR with some specified divide and conquer strategies may yield a new LAR estimate (such as Algorithm 3 in this paper). Thus, Proposition 2.1 provides a theoretical tool to derive optimal learning rates for this type of algorithms.

2.3. AVM-LAR

The AVM-LAR estimate, which is a marriage of the classical AVM strategy [17, 30, 29] and LAR, can be formulated in the following Algorithm 1.

Algorithm 1 AVM-LAR

Initialization: Let $D^N = \{(X_i, Y_i)\}_{i=1}^N$ be a data set of size N, m be the number of data blocks, and h be the localization parameter.
Output: The global estimate $\overline{f}_h$.
Division: Randomly divide D^N into m data blocks $D_1, D_2, \ldots, D_m$ such that $D^N = \bigcup_{j=1}^m D_j, D_i \cap D_j = \varnothing, i \neq j$ and $|D_1| = \cdots = |D_m| = n = N/m$.
Local processing: For $j = 1, 2, \ldots, m$, implement LAR on the data block D_j to get the jth *local estimate*

$$f_{j,h}(x) = \sum_{(X_i,Y_i)\in D_j} W_{X_i,h}(x) Y_i.$$

Synthesization: Obtain a *global estimate* defined by

$$\overline{f}_h = \frac{1}{m}\sum_{j=1}^m f_{j,h}. \tag{2.5}$$

In the following Theorem 2.1, we show that this simple generalization of LAR achieves the optimal learning rate with a condition concerning m. We also show that this condition is essential.

Theorem 2.1. *Let $\overline{f}_h$ be defined by (2.5) and h_{D_j} be the mesh norm of the data block D_j defined by $h_{D_j} := \max_{X\in\mathcal{X}} \min_{X_i\in D_j} \|X - X_i\|$. Suppose that*

(C) for all $D_1, \ldots, D_m$, there exists a positive number c_3 such that

$$\mathbf{E}\left\{\sum_{(X_i,Y_i)\in D_j} W^2_{h,X_i}(X)\right\} \le \frac{c_3}{nh^d};$$

(D) for all $D_1, \ldots, D_m$, there holds almost surely

$$W_{X_i,h} I_{\{\|x-X_i\|>h\}} = 0.$$

If $h \sim N^{-1/(2r+d)}$, and the event $\{h_{D_j} \le h \text{ for all } D_j\}$ holds with probability $\delta \in (0,1)$, then there exists a constant C_2 depending only on d, r, M, c_0 and c_3 such that

$$C_0 N^{-2r/(2r+d)} \le \sup_{f_\rho\in\mathcal{F}^{c_0,r}} \mathbf{E}\{\|\overline{f}_h - f_\rho\|_\rho^2\} \le C_2 N^{-2r/(2r+d)} \tag{2.6}$$

holds with confidence δ. Otherwise, for arbitrary $h \geq \frac{1}{2}(n+2)^{-1/d}$, with confidence at least $1-\delta$, there exists a distribution ρ such that

$$\sup_{f_\rho \in \mathcal{F}^{c_0,r}} \mathbf{E}\{\|\overline{f}_h - f_\rho\|_\rho^2\} \geq \frac{M^2\{(2h)^{-d}-2\}}{3n}. \tag{2.7}$$

Remark 2.1. *We formulate our condition in a probability way in order to highlight the importance of the event $\{h_{D_j} \leq h \text{ for all } D_j\}$, although the probability that the event holds can be derived directly, as shown in (2.8) below. In particular, as exhibited in simulations in Section 4, the mentioned event plays a crucial role in controlling m in AVM-LAR.*

The assertions in Theorem 2.1 can be divided into two parts. The first one is the positive assertion, showing that if some conditions on the weights and an extra constraint $\{h_{D_j} \leq h \text{ for all } D_j\}$ are imposed, then the AVM-LAR estimate (2.5) possesses the same learning rate as that in (2.4) by taking the same localization parameter h (ignoring constants). Thus, the average mixture doesn't degrade the learning performance of LAR.

We then explain the conditions of Theorem 2.1 and compare them with those of Proposition 2.1. To get an error estimate like (2.6), it can be found in the proof that (D) can be relaxed to the following condition (D*).

(D*) For all $D_1, \ldots, D_m$, there exists a positive number c_4 such that

$$\mathbf{E}\left\{\sum_{(X_i,Y_i)\in D_j} W_{h,X_i}(X) I_{\{\|X-X_i\|>h\}}\right\} \leq \frac{c_4}{\sqrt{Nh^d}}.$$

Condition (C) is the same as condition (A) by noting that there are only n samples in each D_j. Condition (D*) is stronger than condition (B) as there are totally n samples in D_j but the localization bound is $c_4/(\sqrt{Nh^d})$. However, we should point out that such a restriction is also mild, since in almost all widely used LAR, the localization bound either is 0 (see NWK with naive kernel, and KNN) or decreases exponentially (such as NWK with Gaussian kernel). All the above methods satisfy conditions (C) and (D*). The most important restriction in Theorem 2.1 is the requirement that the event $\{h_{D_j} \leq h \text{ for all } D_j\}$ holds. Since $\mathbf{P}\{h_{D_j} \leq h \text{ for all } D_j\} = 1 - m\mathbf{P}\{h_{D_1} > h\}$, and it can be found in [12, Lemma 6.4] that $\mathbf{P}\{h_{D_1} > h\} \leq \frac{c}{nh^d}$, we have $\mathbf{P}\{h_{D_j} \leq h \text{ for all } D_j\} \geq 1 - \frac{m}{nh^d}$. When $h \sim N^{-1/(2r+d)} = (mn)^{-1/(2r+d)}$, we have

$$\mathbf{P}\{h_{D_j} \leq h \text{ for all } D_j\} \geq 1 - c' \frac{m^{(2r+2d)/(2r+d)}}{n^{2r/(2r+d)}}. \tag{2.8}$$

The above quantity is small when m is large, which means that the probability that the event $\{h_{D_j} \leq h \text{ for all } D_j\}$ holds is very small. Noticing [12, Problem 2.4], it is easy to prove that the above probability estimate is essential in the sense that for the uniform distribution, the equality holds for some constant c'.

Once the event $\{h_{D_j} \leq h \text{ for all } D_j\}$ does not hold, our theorem drives to a negative direction, saying that for any $h \geq \frac{1}{2}(n+2)^{-1/d}$, the learning rate of AVM-LAR isn't faster than $\frac{1}{nh^d}$. It follows from Theorem 2.1 that the

best localization parameter to guarantee the optimal learning rate satisfies $h \sim N^{-1/(2r+d)}$. The condition $h \geq \frac{1}{2}(n+2)^{-1/d}$ implies that if the best parameter is selected, then m should satisfy $m \leq \mathcal{O}(N^{2r/(2r+d)})$. Under this condition, from (2.7), we have

$$\sup_{f_\rho \in \mathcal{F}^{c_0,r}} \mathbf{E}\{\|\overline{f}_h - f_\rho\|_\rho^2\} \geq \frac{C}{nh^d}.$$

This means, if we select $h \sim N^{-1/(2r+d)}$ and $m \leq \mathcal{O}(N^{2r/(2r+d)})$, then the learning rate of AVM-LAR is essentially slower than that in (2.4). If we select a smaller h, then the above inequality yields the similar conclusion. If we select a larger h, however, the approximation error (see the proof of Proposition 2.1) is $\mathcal{O}(h^{2r})$ which is larger than the learning rate in (2.4). In short, if the event $\{h_{D_j} \leq h \text{ for all } D_j\}$ does not hold, then the average mixture essentially degrades the learning performance of LAR for all selection of h.

In the previous studies on implementing average mixture on some global modeling strategies such as the kernel ridge regression [29, 15] and kernel-based spectral algorithms [3, 11], the range of m to guarantee the optimal learning rates depends only on N, the capacity of the corresponding reproducing kernel Hilbert space and the regularity of the regression function. Our results in Theorem 3.2 shows that LAR involves an event $\{h_{D_j} \leq h \text{ for all } D_j\}$ which may not hold even when m is a constant. This phenomenon makes AVM-LAR fairly instable and urges us to develop stable divide and conquer strategies for LAR.

3. Modified AVM-LAR

In this section, we propose two stable variants of AVM-LAR in the sense that they achieve the optimal learning rates under mild conditions.

3.1. AVM-LAR with data-dependent parameters

The event $\{h_{D_j} \leq h \text{ for all } D_j\}$ essentially implies that for arbitrary x, there is at least one sample in the ball $B_h(x) := \{x' \in \mathbb{R}^d : \|x - x'\| \leq h\}$. This condition holds automatically for KNN since $h = \|x - X_{(k)}(x)\|$. However, for NWK and other local average methods (e.g., partition estimation [12]), this condition has a high probability to be broken down. Motivated by KNN, it is natural to select a localization parameter h to ensure the event $\{h_{D_j} \leq h \text{ for all } D_j\}$. Therefore, we propose a variant of AVM-LAR with data-dependent parameters in Algorithm 2.

Compared with AVM-LAR in Algorithm 1, the only difference of Algorithm 2 is the division step, where we select the localization parameter to be larger than all $h_{D_j}, j = 1, \ldots, m$. The following Theorem 3.1 states the theoretical merit of AVM-LAR with data-dependent localization parameters.

Theorem 3.1. *Let $r < d/2$ and $\hat{f}_{\tilde{h}}$ be defined by (3.1). Assume (C) and (D^*) hold for arbitrary $h > 0$. Suppose*

$$\tilde{h} = \max\{m^{-1/(2r+d)} \max_j \{h_{D_j}^{d/(2r+d)}\}, \max_j \{h_{D_j}\}\}, \tag{3.2}$$

Algorithm 2 AVM-LAR with data-dependent parameters

Initialization: Let $D^N = \{(X_i, Y_i)\}_{i=1}^N$ be a data set of size N and m be the number of data blocks.

Output: The global estimate $\hat{f}_{\tilde{h}}$.

Division: Randomly divide D^N into m data blocks $D_1, D_2, \ldots, D_m$ such that $D^N = \bigcup_{j=1}^m D_j, D_i \cap D_j = \varnothing, i \neq j$ and $|D_1| = \cdots = |D_m| = n = N/m$. Compute the mesh norms $h_{D_1}, \ldots, h_{D_m}$, and select $\tilde{h} \geq h_{D_j}$, $j = 1, 2, \ldots, m$.

Local processing: For any $j = 1, 2, \ldots, m$, implement LAR with the localization parameter $\tilde{h}$ for the data block D_j to get the jth *local estimate*

$$f_{j,\tilde{h}}(x) = \sum_{(X_i,Y_i)\in D_j} W_{X_i,\tilde{h}}(x) Y_i.$$

Synthesization: Transmit m *local estimates* $f_{j,\tilde{h}}$ to a machine, getting a *global estimate* defined by

$$\hat{f}_{\tilde{h}} = \frac{1}{m}\sum_{j=1}^m f_{j,\tilde{h}}. \tag{3.1}$$

and $m \leq \left(\frac{c_0^2(2r+d)+8d(c_3+2c_4^2)M^2}{4r(c_0^2+2)}\right)^{d/(2r)} N^{2r/(2r+d)}$, *then there exists a constant* C_3 *depending only on* c_0, c_3, c_4, r, d *and* M *such that*

$$C_0 N^{-2r/(2r+d)} \leq \sup_{f_\rho \in \mathcal{F}^{c_0,r}} \mathbf{E}\{\|\hat{f}_{\tilde{h}} - f_\rho\|_\rho^2\} \leq C_3 N^{-2r/(2r+d)}. \tag{3.3}$$

Theorem 3.1 shows that if the localization parameter is selected elaborately, then AVM-LAR can achieve the optimal learning rate under mild conditions concerning m. It should be noted that there is an additional restriction on the smoothness degree, $r < d/2$. We highlight that this condition cannot be removed. In fact, without this condition, (3.3) may not hold for some marginal distribution ρ_X. For example, let $d = 1$, it can be deduced from [12, Problem 6.1] that there exists a ρ_X such that (3.3) doesn't hold. However, if we don't aim at deriving a distribution free result, we can remove this condition by using the technique in [12, Problem 6.7]. Actually, if there exist $\varepsilon_0 > 0$, a nonnegative function g such that for all $x \in \mathcal{X}$, and $0 < \varepsilon \leq \varepsilon_0$ there holds $\rho_X(B_\varepsilon(x)) > g(x)\varepsilon^d$, and $\int_{\mathcal{X}} \frac{1}{g^{2/d}(x)} d\rho_X < \infty$, then (3.3) holds for arbitrary r and d. It is obvious that the uniform distribution satisfies the above conditions. In this case, Theorem 3.1 exhibits an advantage of Algorithm 2 over the average mixture versions of some global modeling strategies in terms that it requires larger range of m to guarantee the optimal learning rate, since the largest m for AVM versions of spectral algorithms [3, 11] and kernel ridge regression [29, 15, 4] are $\mathcal{O}(m^{-\frac{2r-d}{2r+d}})$ when $r > d/2$. We encourage the readers to compare our results with results in [29, 15, 3, 11, 4].

Instead of imposing a restriction on h_{D_j}, Theorem 3.1 states that after using the data-dependent parameter $\tilde{h}$, AVM-LAR doesn't degrade the learning performance of LAR for a wide range of m. We declare that the derived bound of m cannot be improved further. Indeed, Our proof in Section 5 shows

that the bias of AVM-LAR is bounded by $C\mathbf{E}\{\tilde{h}^{2r}\}$. Under the conditions of Theorem 3.1, if $m \sim N^{(2r+\varepsilon)/(2r+d)}$, then for arbitrary D_j, there holds $\mathbf{E}\{h_{D_j}\} \leq Cn^{-1/d} = C(N/m)^{-1/d} \leq CN^{(d-\varepsilon)/(2r+d)}$. Thus, it is easy to check that $\mathbf{E}\{\tilde{h}^{2r}\} \leq CN^{(-2r+\varepsilon)/(2r+d)}$, which implies a learning rate slower than $N^{-2r/(2r+d)}$.

We conclude this subsection with the sober note: the results in Theorem 3.1 are built upon delicate selection of $\tilde{h}$ in (3.2), requiring the smoothness information of the regression function and computations of the mesh norm of each data block. By definition, the mesh norm h_{D_j} measures the maximum distance that any points on $\mathcal{X}$ can be from $D_j(x) := \{x \in X : (x, y) \in D_j\}$. It reflects the denseness of $D_j(x)$ in $\mathcal{X}$. Computing h_{D_j} requires at least $\mathcal{O}(n^2)$ computational complexity and makes the computational cost of Algorithm 2 be higher than that of LAR.

3.2. Qualified AVM-LAR

Algorithm 2 provided an intuitive way to improve the performance of AVM-LAR. However, Algorithm 2 increases the computational complexity of AVM-LAR, because we have to compute the mesh norm $h_{D_j}, j = 1, \ldots, m$. A natural question is whether we can avoid this procedure while maintaining the learning performance. The following Algorithm 3 provides a possible way to answer this question.

Algorithm 3 Qualified AVM-LAR

Initialization: Let $D^N = \{(X_i, Y_i)\}_{i=1}^N$ be a data set of size N, m be the number of data blocks, and h be the localization parameter.
Output: The global estimate $\hat{f}_h$.
Division: Randomly divide D^N into m data blocks, i.e. $D^N = \cup_{j=1}^m D_j$ with $D_j \cap D_k = \varnothing$ for $k \neq j$ and $|D_1| = \cdots = |D_m| = n$.
Qualification: For a test input x, if there exists an $X_0^j \in D_j$ such that $|x - X_0^j| \leq h$, then we qualify D_j as an active data block for the *local estimate*. Rewrite all the active data blocks as $T_1, \ldots, T_{m_0}$.
Local processing : For arbitrary data block T_j, $j = 1, \ldots, m_0$, define

$$f_{j,h}(x) = \sum_{(X_i, Y_i) \in T_j} W_{X_i,h}(x) Y_i.$$

Synthesization: Transmit m_0 *local estimates* $f_{j,h}$ to a machine, getting a *global estimate* defined by

$$\hat{f}_h = \frac{1}{m_0} \sum_{j=1}^{m_0} f_{j,h}. \tag{3.4}$$

Comparing with Algorithms 1 and 2, the only difference of Algorithm 3 is the qualification step which essentially does not need extra computation. In fact, the qualification and local processing steps can be implemented simultaneously. It should be further mentioned that the qualification step actually eliminates the data blocks which have a chance to break down the event $\{h_{D_j} \leq h$ for

all $D_j\}$. We show in the following theorem that the qualified AVM-LAR can achieve the optimal learning rate of LAR without any restriction on m.

Theorem 3.2. *Let $\hat{f}_h$ be defined by (3.4). Assume (C) holds and*

(E) for all $D_1, \ldots, D_m$, there exists a positive number c_5 such that

$$\mathbf{E}\left\{\sum_{i=1}^{n} |W_{h,X_i}(X)| I_{\{\|X-X_i\|>h\}}\right\} \le \frac{c_5}{m\sqrt{nh^d}}.$$

If $h \sim N^{-1/(2r+d)}$, then there exists a constant C_4 depending only on c_0, c_1, c_3, c_5, r, d and M such that

$$C_0 N^{-2r/(2r+d)} \le \sup_{f_\rho \in \mathcal{F}^{c_0,r}} \mathbf{E}\{\|\hat{f}_h - f_\rho\|_\rho^2\} \le C_4 N^{-2r/(2r+d)}. \tag{3.5}$$

In Theorem 3.1, we declare that AVM-LAR with data-dependent parameter does not slow down the learning rate of LAR. However, the bound of m in Theorem 3.1 depends on the smoothness of the regression function, which is usually unknown in the real world applications. This makes m be a potential parameter in AVM-LAR with data-dependent parameter, as we do not know which m definitely works. However, Theorem 3.2 states that we can avoid this problem by introducing a qualification step. The theoretical price of such an improvement is only to use condition (E) to replace condition (D*). As shown above, all the widely used LARs such as the partition estimate, NWK with naive kernel, NWK with Gaussian kernel and KNN satisfy condition (E) (with a logarithmic term for NWK with Gaussian kernel).

4. Experiments

In this section, we conduct experimental studies on synthetic data sets to demonstrate the performances of AVM-LAR and its variants.

4.1. Simulation 1

We use a fixed total number of samples $N = 10,000$, but assume that the number of data blocks m (the data block size $n = N/m$) and dimensionality d are varied. The simulation results are based on the average values of 20 trials.

We generate data from the following regression models $y = g_j(x) + \varepsilon,\ j = 1, 2$, where ε is the Gaussian noise $\mathcal{N}(0, 0.1)$,

$$g_1(x) = \begin{cases} (1-2x)_+^3(1+6x), & 0 < x \le 0.5 \\ 0 & x > 0.5 \end{cases}, \tag{4.1}$$

and

$$g_2(x) = \begin{cases} (1-\|x\|)_+^5(1+5\|x\|) + \frac{1}{5}\|x\|^2, & 0 < \|x\| \le 1, x \in \mathbb{R}^5 \\ \frac{1}{5}\|x\|^2 & \|x\| > 1 \end{cases}. \tag{4.2}$$

Actually g_1 and g_2 are the so-called Wendland functions [27] with the property $g_1, g_2 \in \mathcal{F}^{c_0,1}$ for some absolute constant c_0 and $g_1, g_2 \notin \mathcal{F}^{c_1,2}$ for all $0 <$

$c_1 < \infty$. In simulated N samples, $X_i, i = 1, \ldots, N$, are drawn i.i.d. according to the uniform distribution on the (hyper-)cube $[0,1]^d$ with $d = 1, 5$ and $Y_i = g_j(X_i) + \epsilon_i, j = 1, 2$. We also generate $N' = 1,000$ test samples (X_i', Y_i') with X_i' drawn i.i.d. according to the uniform distribution and $Y_i' = g_j(X_i'), j = 1, 2$.

We employ three criteria for the comparison purpose. The first criterion is the *global error* (GE) with GE:= $\frac{1}{N'}\sum_{i=1}^{N'}(Y_i' - \hat{f}_{D^N,h}(X_i'))^2$. The second criterion is the *local error* (LE) with LE:= $\min_{j=1,\ldots,m} \frac{1}{N'}\sum_{i=1}^{N'}(Y_i' - f_{j,h}(X_i'))^2$ and the third one is the *average error* (AE) satisfying AE:= $\frac{1}{N'}\sum_{i=1}^{N'}(Y_i' - \bar{f}(X_i'))^2$ with $\bar{f} = \bar{f}_h$ for Algorithm 1, $\bar{f} = \hat{f}_{\tilde{h}}$ for Algorithm 2 and $\bar{f} = \hat{f}_h$ for Algorithm 3. It is easy to see that GE reflects the learning performance of LAR processing the whole data in one single machine and provides a baseline for our analysis. It is obvious that GE is independent of m. LE refers to the learning performance of LAR processing N/m data in one single machine. It provides a lower limit for the divide and conquer algorithms in the sense that LE must be larger than AE to show the necessity of synthesization. AE concerns the learning performance of the divide and conquer algorithm and is the subject in our experiments.

The aims of Simulation 1 are two folds. The one is to compare the learning performance between Algorithm 1 and LAR processing the whole data and the other is to show the outperformance of Algorithms 2 and 3. We employ AVM-NWK and AVM-KNN for the first purpose and NWK for the second purpose, since Algorithms 1, 2 and 3 are the same for KNN. The detailed implementation of NWK and KNN is specified as follows.

- NWK: In Algorithm 1 and Algorithm 3, for each $m \in \{5, 10, \ldots, 350\}$, the localization parameter satisfies $h \sim N^{-\frac{1}{2r+d}}$ according to Theorem 2.1 and Theorem 3.2. In Algorithm 2, we set $\tilde{h} \sim \max\left\{\frac{\max_j\{h_{D_j}^{d/(2r+d)}\}}{m^{2r+d}}, \max_j\{h_{D_j}\}\right\}$ according to Theorem 3.1.
- KNN: According to Theorem 2.1, the parameter k is set to $k \sim \frac{N^{\frac{2r}{2r+d}}}{m}$. However, as $k \geq 1$, the range of m should satisfy $m \in \{1, 2, \ldots, N^{\frac{2r}{2r+d}}\}$.

We use 5-fold cross-validation to provide an appropriate constant in selecting the localization parameters. The simulation results are reported in the Figures 1, 2 and 3.

As shown in Figure 1, AEs are smaller than LEs, implying that AVM-NWK outperforms NWK with only one data block. Furthermore, AEs of NWK are comparable with GEs when m is not too large, which means AVM-NWK does not degrade the learning performance of NWK with the whole data and verifies (2.6) in Theorem 3.1. It is also shown in Figure 1 that there exists an m', the upper bound of the number of data blocks, to guarantee the optimal learning rate, larger than which the curve of AE increases dramatically. Moreover, m' decreases when d increases as shown in (2.8). All these verifies the positive part of Theorem 3.1.

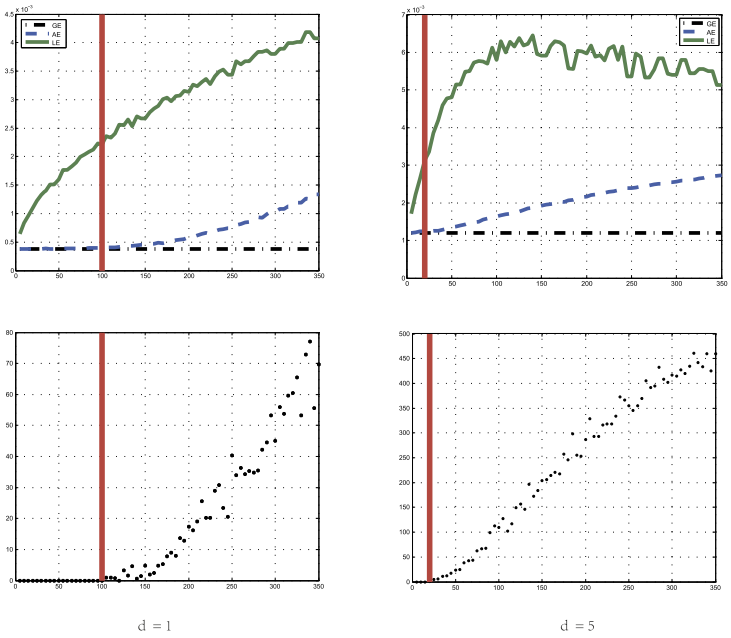

FIG 1. The first row shows AEs, LEs and GEs of NWK for different m. The second row shows the number of inactive machines which satisfy $h_{D_j} > h$. The vertical axis of the second row of Figure 1 is the number of inactive data blocks which break down the condition $h_{D_j} \leq h$.

Before verifying the negative part of Theorem 3.1, we present some explanations for the phenomenon exhibited in Figure 1. If only one data block is utilized, then it follows from Proposition 2.1 that $\min_{j=1,\ldots,m} \mathbf{E}\{\|f_{j,h} - f_\rho\|_\rho^2\} = \mathcal{O}(n^{-\frac{2r}{2r+d}})$, which is far larger than $\mathcal{O}(N^{-\frac{2r}{2r+d}})$ for AVM-NWK due to Theorem 2.1. Thus, AEs are smaller than LEs. Moreover, Theorem 2.1 shows that AEs are comparable with GE as long as the event $\{h_{D_j} \leq h \text{ for all } D_j, j = 1, \ldots, m\}$ holds. Once this event does not hold, Theorem 3.1 drives a totally different direction and implies the drawback for AVM-NWK. To verify this assertion, we record the number of data blocks with $h_{D_j} > h$ for different m in the second row of Figure 1 and use a bar to highlight a crucial m^*, larger than which there exist inactive data blocks. Figure 1 exhibits $m' \approx m^*$, which is extremely consistent with the negative part of Theorem 2.1.

Different from AVM-NWK, there is not an upper bound for the number of data blocks in AVM-KNN to guarantee the comparability between AEs and GE. The reason is that KNN adapts a data-dependent localization parameter h that makes the event $\{h_{D_j} \leq h \text{ for all } D_j, j = 1, \ldots, m\}$ always holds. This result is also consistent with the positive part of Theorem 2.1. However, by definition, the restriction $k \geq 1$ in each data block restricts the range of m in AVM-KNN to be $\{1, 2, \ldots, N^{\frac{2r}{2r+d}}\}$, showing a design deficiency. All these numerical

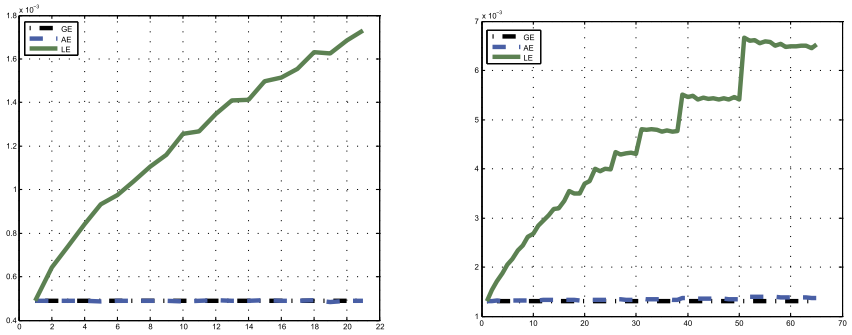

FIG 2. AEs, LEs and GEs of KNN for different m.

results finish our first purpose in Simulation 1 and verifies the correctness of Theorem 2.1.

For the second purpose, we denote AE-A1, AE-A2 and AE-A3 as AEs of Algorithms 1, 2 and 3 respectively and record the values of them with different m in Figure 3. When $m \leq m'$, AE-A1, AE-A2 and AE-A3 have similar values which are comparable with GE in Figure 1. The reason is the occurrence of the event $\{h_{D_j} \leq h \text{ for all } D_j, j = 1, \ldots, m\}$. When m increases, the event $\{h_{D_j} > h \text{ for some } j\}$ inevitably happens, then Algorithm 1 fails according to the negative part of Theorem 2.1, making AE-A1 increase dramatically. As Algorithms 2 and 3 are designed to avoid the weakness of Algorithm 1, AE-A2 and AE-A3 are always smaller than AE-A1 when $m > m'$, which verifies the correctness of Theorems 3.1 and 3.2. An interesting phenomenon exhibited in Figure 3 is that AE-A3 is always smaller than AE-A2. We guess the reason is an inaccurate computation of the mesh norm since we set $\tilde{h} = c \cdot \max_{i,j \in 1,2\ldots,N} \|X_i - X_j\|$ and then use 5-fold cross-validation to select c in Algorithm 2, which only provides an upper bound of $\max\{h_{D_j} : j = 1, 2, \ldots, m\}$. We believe that the performance of Algorithm 2 can be improved if the mesh norm is efficiently and accurately computed.

4.2. Simulation 2

We make use of the same simulation study conducted by [29] for comparing the learning performance of Algorithms 1, 2, 3 and the divide and conquer kernel ridge regression (DKRR for short).

We generate data from the regression model $y = g_3(x) + \epsilon$, where $g_3(x) = \min(x, 1-x)$, the noise variable ϵ is normally distributed with mean 0 and variance $\sigma^2 = 1/5$, and $X_i, i = 1, \ldots, N$ are simulated from the uniform distribution in $[0, 1]$ independently. In the simulation of [29], DKRR used the kernel function $K(x, x') = 1 + \min\{x, x'\}$, and regularization parameter $\lambda = N^{-2/3}$ due to

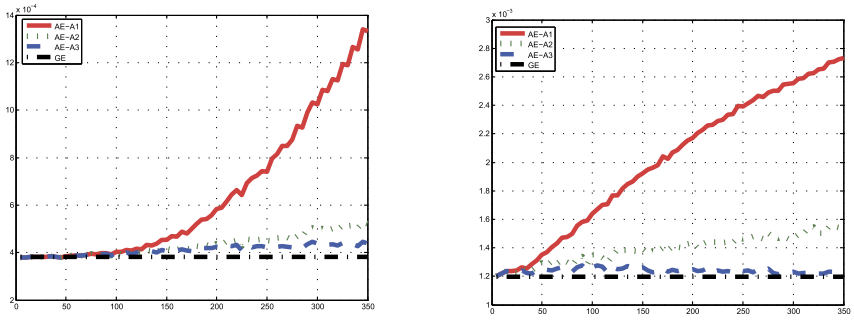

FIG 3. AE-A1, AE-A2, AE-A3 and GE of the simulation. The curves of AE-A2 and AE-A3 are always below AE-A1's to illustrate the improved capability of modified AVM-LARs.

$g_3 \in \mathcal{F}^{c_0,1}$ for some absolute constant c_0. We gather $N = 10,000$ training samples, and 1,000 test samples. The parameter selection strategies of Algorithms 1, 2 and 3 is the same as those in Simulation 1.

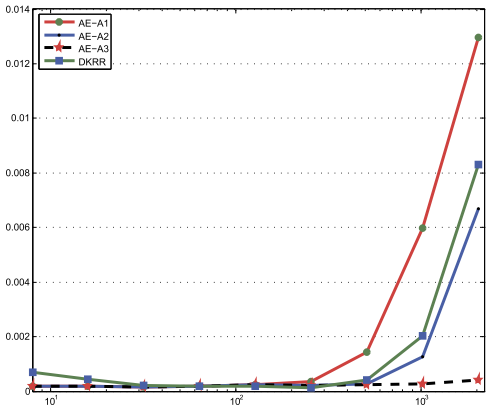

FIG 4. AEs of Algorithm 1, 2, 3 and DKRR.

In Figure 4, we plot AEs of Algorithms 1, 2, 3 and DKRR. Here $m \in \{2^3, 2^4, \ldots, 2^{11}\}$. Figure 4 shows that AEs of Algorithms 1, 2, 3 and DKRR are comparable when $m < 256$. For larger m, AEs of Algorithms 1, 2 and DKRR increase dramatically. Differently, AEs of Algorithm 3 are stable. This phenomenon is consistent with the theoretical assertions in Theorem 3.1, 3.2 and Theorem 1 in [29], showing that the largest m to keep the optimal learning rates of divide and conquer algorithms are $\mathcal{O}(N^{1/3})$ for DKRR, $\mathcal{O}(N^{2/3})$ for Algorithm 2 and $\mathcal{O}(N)$ for Algorithm 3.

5. Proofs

5.1. Proof of Proposition 2.1

Let $f_{\rho,h}(x) = \sum_{i=1}^{N} W_{h,X_i}(x) f_\rho(X_i)$. Then, it is obvious that $f_{\rho,h}(x) = \mathbf{E}^*\{\hat{f}_{D^N,h}(x)\}$, where $\mathbf{E}^*\{\cdot\} = \mathbf{E}\{\cdot|X_1, X_2, \ldots, X_n\}$. Therefore, we can deduce

$$\mathbf{E}^*\{(\hat{f}_{D^N,h}(x) - f_\rho(x))^2\} = \mathbf{E}^*\{(\hat{f}_{D^N,h}(x) - f_{\rho,h}(x))^2\} + (f_{\rho,h}(x) - f_\rho(x))^2.$$

That is,

$$\begin{aligned}\mathbf{E}\{\|\hat{f}_{D^N,h} - f_\rho\|_\rho^2\} &= \int_{\mathcal{X}} \mathbf{E}\{\mathbf{E}^*\{(\hat{f}_{D^N,h}(X) - f_{\rho,h}(X))^2\}\} d\rho_X \\ &+ \int_{\mathcal{X}} \mathbf{E}\{(f_{\rho,h}(X) - f_\rho(X))^2\} d\rho_X.\end{aligned}$$

The first and second terms are referred to the *sample error* and *approximation error*, respectively. To bound the sample error, noting $\mathbf{E}^*\{Y_i\} = f_\rho(X_i)$, we have

$$\begin{aligned}&\mathbf{E}^*\{(\hat{f}_{D^N,h}(x) - f_{\rho,h}(x))^2\} = \mathbf{E}^*\left\{\left(\sum_{i=1}^{N} W_{h,X_i}(x)(Y_i - f_\rho(X_i))\right)^2\right\} \\ \leq\ &\mathbf{E}^*\left\{\sum_{i=1}^{N} (W_{h,X_i}(x)(Y_i - f_\rho(X_i)))^2\right\} \leq 4M^2 \sum_{i=1}^{N} W_{h,X_i}^2(x).\end{aligned}$$

Therefore we can use (A) to bound the sample error as

$$\mathbf{E}\{(\hat{f}_{D^N,h}(X) - f_{\rho,h}(X))^2\} \leq 4M^2 \mathbf{E}\left\{\sum_{i=1}^{N} W_{h,X_i}^2(X)\right\} \leq \frac{4c_1 M^2}{Nh^d}.$$

Now, we turn to bound the approximation error. Let $B_h(x)$ be the l^2 ball with center x and radius h, we have

$$\begin{aligned}&\mathbf{E}\{(f_{\rho,h}(X) - f_\rho(X))^2\} = \mathbf{E}\left\{\left(\sum_{i=1}^{N} W_{h,X_i}(X) f_\rho(X_i) - f_\rho(X)\right)^2\right\} \\ =\ &\mathbf{E}\left\{\left(\sum_{i=1}^{N} W_{h,X_i}(X)(f_\rho(X_i) - f_\rho(X))\right)^2\right\} \\ =\ &\mathbf{E}\left\{\left(\sum_{i=1}^{N} W_{h,X_i}(X)(f_\rho(X_i) - f_\rho(X))\right)^2 I_{\{B_h(X)\cap D=\varnothing\}}\right\} \\ +\ &\mathbf{E}\left\{\left(\sum_{i=1}^{N} W_{h,X_i}(X)(f_\rho(X_i) - f_\rho(X))\right)^2 I_{\{B_h(X)\cap D\neq\varnothing\}}\right\}.\end{aligned}$$

It follows from [12, Theorem 4.3] and $\sum_{i=1}^{N} W_{h,X_i}(X) = 1$ that

$$\mathbf{E}\left\{\left(\sum_{i=1}^{N} W_{h,X_i}(X)(f_\rho(X_i) - f_\rho(X))\right)^2 I_{\{B_h(X)\cap D=\varnothing\}}\right\} \le \frac{16M^2}{Nh^d}.$$

Furthermore,

$$\begin{aligned}
&\mathbf{E}\left\{\left(\sum_{i=1}^{N} W_{h,X_i}(X)(f_\rho(X_i) - f_\rho(X))\right)^2 I_{\{B_h(X)\cap D\neq\varnothing\}}\right\} \\
\le\ &\mathbf{E}\left\{\left(\sum_{\|X_i - X\|\le h} W_{h,X_i}(X)|f_\rho(X_i) - f_\rho(X)|\right)^2 I_{\{B_h(X)\cap D\neq\varnothing\}}\right\} \\
+\ &\mathbf{E}\left\{\left(\sum_{\|X_i - X\|> h} W_{h,X_i}(X)|f_\rho(X_i) - f_\rho(X)|\right)^2\right\} \\
\le\ &c_0^2 h^{2r} + \frac{4c_2^2 M^2}{Nh^d},
\end{aligned}$$

where the last inequality is deduced by $f_\rho \in \mathcal{F}^{c_0,r}$, condition (B) and Jensen's inequality. In this way, we get

$$\mathbf{E}\{\|\hat{f}_{D^N,h} - f_\rho\|_\rho^2\} \le c_0^2 h^{2r} + \frac{4(c_1 + c_2^2 + 4)M^2}{Nh^d}.$$

If we set $h = \left(\frac{4(c_1+c_2^2+4)M^2}{c_0^2 N}\right)^{-1/(2r+d)}$, then

$$\mathbf{E}\{\|\hat{f}_{D^N,h} - f_\rho\|_\rho^2\} \le c_0^{2d/(2r+d)}(4(c_1 + c_2^2 + 4)M^2)^{2r/(2r+d)} N^{-2r/(2r+d)}.$$

This together with [12, Theorem 3.2] finishes the proof of Proposition 2.1. □

5.2. Proof of Theorem 2.1

Since $\mathbf{E}\left\{\|\overline{f}_h - f_\rho\|_\rho^2\right\} = \mathbf{E}\left\{\|\overline{f}_h - \mathbf{E}\{\overline{f}_h\} + \mathbf{E}\{\overline{f}_h\} - f_\rho\|_\rho^2\right\}$ and $\mathbf{E}\{\overline{f}_h\} = \mathbf{E}\{f_{j,h}\}$ for all $j = 1, \ldots, m$, we get

$$\begin{aligned}
\mathbf{E}\left\{\|\overline{f}_h - f_\rho\|_\rho^2\right\} &= \frac{1}{m^2}\mathbf{E}\left\{\sum_{j=1}^{m}\left(\|f_{j,h} - \mathbf{E}\{f_{j,h}\}\|_\rho^2 + \|\mathbf{E}\{f_{j,h}\} - f_\rho\|_\rho^2\right)\right. \\
&\quad \left. + 2\sum_{j=1}^{m}\sum_{k\neq j}\langle f_{j,h} - \mathbf{E}\{f_{j,h}\}, f_{k,h} - \mathbf{E}\{f_{k,h}\}\rangle_\rho\right\} \\
&= \frac{1}{m}\mathbf{E}\{\|f_{1,h} - \mathbf{E}\{f_{1,h}\}\|_\rho^2\} + \|\mathbf{E}\{f_{1,h}\} - f_\rho\|_\rho^2 \qquad (5.1)
\end{aligned}$$

$$\leq \quad \frac{2}{m}\mathbf{E}\{\|f_{1,h}-f_\rho\}\|_\rho^2\} + 2\|\mathbf{E}\{f_{1,h}\}-f_\rho\|_\rho^2.$$

As $h \geq h_{D_j}$ for all $1 \leq j \leq m$ with probability δ, we obtain with confidence δ that $B_h(x) \cap D_j \neq \varnothing$ for all $x \in \mathcal{X}$ and $1 \leq j \leq m$. Then, using the same method as that in the proof of Proposition 2.1, (C) and (D*) yield that with confidence δ

$$\mathbf{E}\{\|f_{1,h}-f_\rho\}\|_\rho^2\} \leq c_0^2 h^{2r} + \frac{4(c_3+c_4^2)M^2}{nh^d}.$$

Due to Jensen's inequality, with confidence δ

$$\mathbf{E}\left\{\|\overline{f}_h - f_\rho\|_\rho^2\right\} \leq \frac{2c_0^2h^{2r}}{m} + \frac{8(c_3+c_4^2)M^2}{mnh^d} + 2\mathbf{E}\left\{\|\mathbf{E}^*\{f_{1,h}\}-f_\rho\|_\rho^2\right\}.$$

Noting $B_h(X) \cap D_j \neq \varnothing$ with confidence δ, the same method as that in the proof of Proposition 2.1 together with (D*) yields that with confidence δ

$$\mathbf{E}\left\{\|\mathbf{E}^*\{f_{1,h}\}-f_\rho\|_\rho^2\right\} \leq c_0^2h^{2r} + \frac{8c_4^2M^2}{mnh^d}.$$

Thus,

$$\mathbf{E}\left\{\|\overline{f}_h - f_\rho\|_\rho^2\right\} \leq \frac{16(c_3+c_4^2)M^2}{mnh^d} + 3c_0^2h^{2r}.$$

This finishes the proof of (2.6) by taking $h = \left(\frac{16(c_3+c_4^2)M^2}{3c_0^2nm}\right)^{-1/(2r+d)}$ into account.

Now, we turn to prove (2.7). According to (5.1), we have

$$\begin{aligned}\mathbf{E}\left\{\|\overline{f}_h - f_\rho\|_\rho^2\right\} &\geq \|\mathbf{E}\{f_{1,h}\}-f_\rho\|_\rho^2 \\ &= \int_{\mathcal{X}}\left(\mathbf{E}\left\{\sum_{i=1}^n W_{X_i,h}(X)f_\rho(X_i)-f_\rho(X)\right\}\right)^2 d\rho_X.\end{aligned}$$

Since with confidence $1-\delta$, the event $\{h_{D_j} \leq h \text{ for all } D_j\}$ does not holds. Without loss of generality, we assume that $h < h_{D_1}$ holds in a probabilistic setting. It then follows from the definition of the mesh norm that there exits an $X \in \mathcal{X}$ which is not in $B_h(X_i)$, $X_i \in D_1$. Define the separation radius of a set of points $S = \{\zeta_i\}_{i=1}^n \subset \mathcal{X}$ via

$$r_{_S} := \frac{1}{2}\min_{j\neq k}\|\zeta_j - \zeta_k\|.$$

The mesh ratio $\tau_{_S} := \frac{h_S}{r_S} \geq 1$ provides a measure of how uniformly points in S are distributed on $\mathcal{X}$. If $\tau_{_S} \leq 2$, we then call S as the quasi-uniform point set. Let $\Xi_l = \{\xi_1, \ldots, \xi_l\}$ be $l = \lfloor (2h)^{-d} \rfloor$ quasi-uniform points [27] in $\mathcal{X}$. That is $\tau_{\Xi_l} = \frac{h_{\Xi_l}}{r_{\Xi_l}} \leq 2$. Since $h_{\Xi_l} \geq l^{-1/d}$, we have $r_{\Xi_l} \geq \frac{1}{2l^{1/d}} \geq h$. Then,

$$\mathbf{E}\left\{\|\overline{f}_h - f_\rho\|_\rho^2\right\} \quad = \quad \|\mathbf{E}\{f_{1,h}\}-f_\rho\|_\rho^2$$

$$\geq \sum_{k=1}^{l} \int_{B_{r_{\Xi_l}}(\xi_k)} \left(\mathbf{E} \left\{ \sum_{i=1}^{n} W_{X_i,h}(X) f_\rho(X_i) - f_\rho(X) \right\} \right)^2 d\rho_X$$

holds with confidence $1-\delta$. If $f_\rho(x) = M$, then with confidence $1-\delta$

$$\begin{aligned}
\mathbf{E}\left\{\|\overline{f}_h - f_\rho\|_\rho^2\right\} &\geq M^2 \sum_{k=1}^{l} \int_{B_{r_{\Xi_l}}(\xi_k)} \left(\mathbf{E}\left\{ I_{\{D_1 \cap B_{r_{\Xi_l}}(\xi_k) = \varnothing\}} \right\} \right)^2 d\rho_X \\
&\geq M^2 \sum_{k=1}^{l} \rho_X(B_{r_{\Xi_l}}(\xi_k)) \mathbf{P}\{D_1 \cap B_{r_{\Xi_l}}(\xi_k) = \varnothing\} \\
&= M^2 \sum_{k=1}^{l} \rho_X(B_{r_{\Xi_l}}(\xi_k))(1 - \rho_X(B_{r_{\Xi_l}}(\xi_k)))^n.
\end{aligned}$$

Since $h \geq \frac{1}{2}(n+2)^{-1/d}$, we can let ρ_X be the marginal distribution satisfying

$$\rho_X(B_{r_{\Xi_l}}(\xi_k)) = 1/n, \ k = 1, 2, \ldots, l-1.$$

Then with confidence $1-\delta$

$$\mathbf{E}\left\{\|\overline{f}_h - f_\rho\|_\rho^2\right\} \geq M^2 \sum_{k=1}^{l-1} \frac{1}{n}(1-1/n)^n \geq \frac{M^2((2h)^{-d} - 2)}{3n}.$$

This finishes the proof of Theorem 2.1. □

5.3. Proof of Theorem 3.1

Without loss of generality, we assume $h_{D_1} = \max_j\{h_{D_j}\}$. It follows from (5.1) that

$$\mathbf{E}\left\{\|\hat{f}_{\tilde{h}} - f_\rho\|_\rho^2\right\} \leq \frac{2}{m}\mathbf{E}\{\|f_{1,\tilde{h}} - f_\rho\}\|_\rho^2\} + 2\|\mathbf{E}\{f_{1,\tilde{h}}\} - f_\rho\|_\rho^2.$$

We first bound $\mathbf{E}\{\|f_{1,\tilde{h}} - f_\rho\}\|_\rho^2\}$. As $\tilde{h} \geq h_{D_1}$ almost surely, the same method as that in the proof of Theorem 2.1 yields that

$$\mathbf{E}\{\|f_{1,\tilde{h}} - f_\rho\}\|_\rho^2\} \leq c_0^2 \mathbf{E}\{\tilde{h}^{2r}\} + \mathbf{E}\left\{\frac{4M^2(c_3 + c_4^2)}{n\tilde{h}^d}\right\}.$$

To bound $\|\mathbf{E}\{f_{1,\tilde{h}}\} - f_\rho\|_\rho^2$, we use the same method as that in the proof of Theorem 2.1 again. As $\tilde{h} \geq m^{-1/(2r+d)} h_{D_1}^{d/(2r+d)}$ holds almost surely, it is easy to deduce that

$$\begin{aligned}
\|\mathbf{E}\{f_{1,\tilde{h}}\} - f_\rho\|_\rho^2 &\leq \mathbf{E}\left\{\|\mathbf{E}^*\{f_{1,\tilde{h}}\} - f_\rho\|_\rho^2\right\} \leq c_0^2 \mathbf{E}\{\tilde{h}^{2r}\} + \mathbf{E}\left\{\frac{8c_4^2 M^2}{mn\tilde{h}^d}\right\} \\
&\leq c_0^2 m^{-2r/(2r+d)} \mathbf{E}\{h_{D_1}^{2rd/(2r+d)}\} + c_0^2 \mathbf{E}\{h_{D_1}^{2r}\}
\end{aligned}$$

$$+ \quad 8c_4^2M^2(mn)^{-1}\mathbf{E}\{m^{d/(2r+d)}h_{D_1}^{-d^2/(2r+d)}\}.$$

Thus

$$\begin{aligned}\mathbf{E}\{\|\hat{f}_{\tilde{h}} - f_\rho\}\|_\rho^2\} &\leq c_0^2 m^{-2r/(2r+d)}\mathbf{E}\{h_{D_1}^{2rd/(2r+d)}\} + (c_0^2+2)\mathbf{E}\{h_{D_1}^{2r}\} \\ &+ 8(c_3+2c_4^2)M^2(mn)^{-1}m^{d/(2r+d)}\mathbf{E}\{h_{D_1}^{-d^2/(2r+d)}\}.\end{aligned}$$

To bound $\mathbf{E}\{h_{D_1}^{2rd/(2r+d)}\}$, we note that for arbitrary $\varepsilon > 0$, there holds

$$\mathbf{P}\{h_{D_1} > \varepsilon\} = \mathbf{P}\{\max_{x\in\mathcal{X}}\min_{X_i\in D_1}\|x - X_i\| > \varepsilon\} \leq \max_{x\in\mathcal{X}}\mathbf{E}\{(1-\rho_X(B_\varepsilon(x)))^n\}.$$

Let $t_1, \ldots, t_l$ be the quasi-uniform points of $\mathcal{X}$. Then it follows from [12, P.93] that $\mathbf{P}\{h_{D_1} > \varepsilon\} \leq \frac{1}{n\varepsilon^d}$. Then, we have

$$\begin{aligned}\mathbf{E}\{h_{D_1}^{2rd/(2r+d)}\} &= \int_0^\infty \mathbf{P}\{h_{D_1}^{2rd/(2r+d)} > \varepsilon\}d\varepsilon = \int_0^\infty \mathbf{P}\{h_{D_1} > \varepsilon^{(2r+d)/(2rd)}\}d\varepsilon \\ &\leq \int_0^{n^{-2r/(2r+d)}} 1d\varepsilon + \int_{n^{-2r/(2r+d)}}^\infty \mathbf{P}\{h_{D_1} > \varepsilon^{(2r+d)/(2rd)}\}d\varepsilon \\ &\leq n^{-2r/(2r+d)} + \frac{1}{n}\int_{n^{-2r/(2r+d)}}^\infty \varepsilon^{-(2r+d)/(2r)}d\varepsilon \\ &\leq \frac{2r+d}{d}n^{-2r/(2r+d)}.\end{aligned}$$

To bound $\mathbf{E}\{h_{D_1}^{2r}\}$, we can use the above method again and $r < d/2$ to derive $\mathbf{E}\{h_{D_1}^{2r}\} \leq 4rd^{-1}n^{-2r/d}$. To bound $\mathbf{E}\{h_{D_1}^{-d^2/(2r+d)}\}$, we use the fact $h_{D_1} \geq n^{-1/d}$ almost surely to obtain $\mathbf{E}\{h_{D_1}^{-d^2/(2r+d)}\} \leq n^{d/(2r+d)}$. Hence

$$\begin{aligned}&\mathbf{E}\{\|\hat{f}_{\tilde{h}} - f_\rho\}\|_\rho^2\} \\ &\quad\leq \left(\frac{c_0^2(2r+d)}{d} + 8(c_3+2c_4^2)M^2\right)N^{-2r/(2r+d)} + \frac{4r(c_0^2+2)}{d}n^{-2r/d}.\end{aligned}$$

Since

$$m \leq \left(\frac{c_0^2(2r+d) + 8d(c_3+2c_4^2)M^2}{4r(c_0^2+2)}\right)^{d/(2r)} N^{2r/(2r+d)},$$

we have

$$\mathbf{E}\{\|\hat{f}_{\tilde{h}} - f_\rho\}\|_\rho^2\} \leq 2\left(\frac{c_0^2(2r+d)}{d} + 8(c_3+2c_4^2)M^2\right)N^{-2r/(2r+d)}$$

which finishes the proof of (3.3). □

5.4. *Proof of Theorem 3.2*

Proof. From the definition, it follows that

$$\hat{f}_h(x) = \sum_{j=1}^m \frac{I_{\{B_h(x)\cap D_j\neq\varnothing\}}}{\sum_{j=1}^m I_{\{B_h(x)\cap D_j\neq\varnothing\}}} \sum_{(X_i^j,Y_i^j)\in D_j} W_{h,X_i^j}(x)Y_i^j.$$

We then use Proposition 2.1 to consider a new local estimate with

$$W^*_{h,X_i^j}(x) = \frac{I_{\{B_h(x)\cap D_j\neq\varnothing\}}W_{h,X_i^j}(x)}{\sum_{j=1}^m I_{\{B_h(x)\cap D_j\neq\varnothing\}}}.$$

We first prove (A) holds. To this end, we have

$$\begin{aligned}
&\mathbf{E}\left\{\sum_{j=1}^m \sum_{(X_i^j,Y_i^j)\in D_j}(W^*_{h,X_i^j}(X))^2\right\}\\
&\quad\le \mathbf{E}\left\{\sum_{j=1}^m \sum_{(X_i^j,Y_i^j)\in D_j, X_i^j\in B_h(X)}(W^*_{h,X_i^j}(X))^2\right\}\\
&\qquad+\mathbf{E}\left\{\sum_{j=1}^m \sum_{(X_i^j,Y_i^j)\in D_j, X_i^j\notin B_h(X)}(W^*_{h,X_i^j}(X))^2\right\},
\end{aligned}$$

where we define $\sum_{\varnothing}=0$. To bound the first term in the right part of the above inequality, it is easy to see that if $I_{\{X_i^j\in B_h(X)\}}=1$, then $I_{\{B_h(X)\cap D_j\neq\varnothing\}}=1$, vice versa. Thus, it follows from (C) that

$$\begin{aligned}
&\mathbf{E}\left\{\sum_{j=1}^m \sum_{(X_i^j,Y_i^j)\in D_j, X_i^j\in B_h(X)}(W^*_{h,X_i^j}(X))^2\right\}\\
=\;&\frac{1}{m^2}\mathbf{E}\left\{\sum_{j=1}^m \sum_{(X_i^j,Y_i^j)\in D_j, X_i^j\in B_h(X)}(W_{h,X_i^j}(X))^2\right\}\\
\le\;&\frac{1}{m}\max_{1\le j\le m}\mathbf{E}\left\{\sum_{(X_i^j,Y_i^j)\in D_j, X_i^j\in B_h(X)}(W_{h,X_i^j}(X))^2\right\}\\
\le\;&\frac{1}{m}\max_{1\le j\le m}\mathbf{E}\left\{\sum_{(X_i^j,Y_i^j)\in D_j}(W_{h,X_i^j}(X))^2\right\}\le\frac{c_3}{Nh^d}
\end{aligned}$$

To bound the second term, we have

$$\begin{aligned}
&\mathbf{E}\left\{\sum_{j=1}^m \sum_{(X_i^j,Y_i^j)\in D_j, X_i^j\notin B_h(X)}(W^*_{h,X_i^j}(X))^2\right\}\\
&=\mathbf{E}\left\{\sum_{j=1}^m \sum_{(X_i^j,Y_i^j)\in D_j, X_i^j\notin B_h(X)}\left(\frac{I_{\{B_h(X)\cap D_j\neq\varnothing\}}W_{h,X_i^j}(X)}{\sum_{j=1}^m I_{\{B_h(X)\cap D_j\neq\varnothing\}}}\right)^2\right\}
\end{aligned}$$

At first, the same method as that in the proof of Proposition 2.1 yields that $\mathbf{E}\{B_h(X)\cap D=\varnothing\}\le\frac{4}{Nh^d}$. Therefore, we have

$$\begin{aligned}
&\mathbf{E}\left\{\sum_{j=1}^{m}\sum_{(X_i^j,Y_i^j)\in D_j, X_i^j\notin B_h(X)}\left(\frac{I_{\{B_h(X)\cap D_j\neq\varnothing\}}W_{h,X_i^j}(X)}{\sum_{j=1}^{m}I_{\{B_h(X)\cap D_j\neq\varnothing\}}}\right)^2\right\}\\
\leq\ & \frac{4}{Nh^d}+m\max_{1\leq j\leq m}\mathbf{E}\left\{\sum_{(X_i^j,Y_i^j)\in D_j}\left(W_{h,X_i^j}(X)I_{\|X-X_i^j\|>h}\right)\right\}\\
\leq\ & \frac{4+c_3+c_5}{Nh^d}.
\end{aligned}$$

Now, we turn to prove (B) holds. This can be deduced directly by using the similar method as the last inequality and the condition (E). That is,

$$\mathbf{E}\left\{\sum_{j=1}^{m}\sum_{(X_i^j,Y_i^j)\in D_j}|W^*_{h,X_i^j}(X)|I_{\{\|X-X_i\|>h\}}\right\}\leq\frac{c_5}{\sqrt{Nh^d}}.$$

Then Theorem 3.2 follows from Proposition 2.1. □

6. Conclusion

In this paper, we combined the divide and conquer strategy with local average regression to provide a new method called average-mixture local average regression (AVM-LAR) to handle the massive data regression problems. We found that the estimate obtained by AVM-LAR can achieve the minimax learning rate, but under a fairly strict restriction on m. We then proposed two variants of AVM-LAR to either relax the restriction or remove it. Theoretical analysis and simulation studies confirmed our assertions.

We discuss here three interesting topics for future study. Firstly, LAR cannot handle the high-dimensional data due to the curse of dimensionality [12, 8]. How to design variants of AVM-LAR to overcome this hurdle can be accommodated as a desirable research topic. Secondly, we have justified that applying the divide and conquer strategy on the LARs does not degenerate the order of learning rate under some conditions. However, we did not show there is no loss in the constant factor. Discussing the constant factor of the optimal learning rate is an interesting project. Finally, equipping other nonparametric methods [9, 12, 25] with the divide and conquer strategy can be taken into consideration for massive data analysis. For example, Cheng and Shang [5] have discussed that how to appropriately apply the divide and conquer strategy to the smoothing spline method.

Acknowledgements

We would like to thank the associate editor and the two anonymous reviewers for constructive comments, which greatly help us to improve the paper quality.

References

[1] Battey, H., Fan, J., Liu, H., Lu, J., and Zhu, Z. (2015). Distributed estimation and inference with statistical guarantees. https://arxiv.org/abs/1509.05457

[2] Biau, G., Cadre, B., Rouviere, L., et al. (2010). Statistical analysis of k-nearest neighbor collaborative recommendation. *Ann. Stat.* **38** 1568–1592. MR2662352 (2011c:62094)

[3] Blanchard, G. and Mücke, N. (2016). Parallelizing spectral algorithms for kernel learning. https://arxiv.org/abs/1610.07487

[4] Chang, X., Lin, S.-B., and Zhou, D.-X. (2017). Distributed semi-supervised learning with kernel ridge regression. *J. Mach. Learn. Res.* To appear.

[5] Cheng, G. and Shang, Z. (2015). Computational limits of divide-and-conquer method. https://arxiv.org/abs/1512.09226

[6] Cucker, F. and Zhou, D.-X. (2007). *Learning Theory: An Approximation Theory Viewpoint*, Cambridge University Press, Cambridge. MR2354721 (2009a:41001)

[7] Dwork, C. and Smith, A. (2009). Differential privacy for statistics: What we know and what we want to learn. *J. Priv. Confid.* **1** 135–154.

[8] Fan, J. (2000). Prospects of nonparametric modeling. *J. Am. Stat. Assoc.* **95** 1296–1300. MR1825280

[9] Fan, J. and Gijbels, I. (1994). Censored regression: local linear approximations and their applications. *J. Am. Stat. Assoc. 89* 560–570. MR1294083 (95f:62099)

[10] Guha, S., Hafen, R., Rounds, J., Xia, J., Li, J., Xi, B., and Cleveland, W. S. (2012). Large complex data: divide and recombine (d&r) with rhipe. *Stat.* **1** 53–67.

[11] Guo, Z.-C., Lin, S.-B., and Zhou, D.-X. (2017). Learning theory of distributed spectral algorithms. *Inverse Probl.* Minor Revision Under Review.

[12] Györfi, L., Kohler, M., Krzyzak, A., and Walk, H. (2002). *A Distribution-Free Theory of Nonparametric Regression*, Springer, New York. MR1920390 (2003g:62006)

[13] Kato, K. (2012). Weighted Nadaraya–Watson estimation of conditional expected shortfall. *J. Financ. Econ.* **10** 265–291.

[14] Li, R., Lin, D. K., and Li, B. (2013). Statistical inference in massive data sets. *Appl. Stoch. Models Bus. Ind.* **29** 399–409. MR3117826

[15] Lin, S.-B., Guo, X., and Zhou, D.-X. (2016). Distributed learning with regularized least squares. *J. Mach. Learn. Res.*. Revision Under Review.

[16] Lin, S.-B. and Zhou, D.-X. (2017). Distributed kernel gradient descent algorithms. *Constr. Approx.*. To appear.

[17] Mcdonald, R., Mohri, M., Silberman, N., Walker, D., and Mann, G. S. (2009). Efficient large-scale distributed training of conditional maximum entropy models. in *Adv. Neural Inf. Process. Syst.* 1231–1239.

[18] Shi, L., Feng, Y.-L., and Zhou, D.-X. (2011). Concentration estimates for learning with ℓ_1-regularizer and data depen-

dent hypothesis spaces. *Appl. Comput. Harmon. Anal.* **31** 286–302. MR2806485 (2012e:62130)

[19] Shi, L. (2013). Learning theory estimates for coefficient-based regularized regression. *Appl. Comput. Harmon. Anal.* **34** 252–265. MR3008565

[20] Steinwart, I. and Christmann, A. (2008). *Support Vector Machines*, Springer, New York. MR2450103 (2010f:62002)

[21] Stone, C. J. (1977). Consistent nonparametric regression. *Ann. Stat.* 595–620. MR0443204 (56 #1574)

[22] Stone, C. J. (1980). Optimal rates of convergence for nonparametric estimators. *Ann. Stat.* 1348–1360. MR0594650 (83d:62068)

[23] Stone, C. J. (1982). Optimal global rates of convergence for nonparametric regression. *Ann. Stat.* 1040–1053. MR0673642 (84b:62058)

[24] Takeda, H., Farsiu, S., and Milanfar, P. (2007). Kernel regression for image processing and reconstruction. *IEEE Trans. Image Process.* **16** 349–366. MR2462728 (2009j:94021)

[25] Tsybakov, A. B. (2008). *Introduction to Nonparametric Estimation*, Springer, New York. MR2724359 (2011g:62006)

[26] Wang, C., Chen, M.-H., Schifano, E., Wu, J., and Yan, J. (2015). Statistical methods and computing for big data. https://arxiv.org/abs/1502.07989

[27] Wendland, H. (2004). *Scattered Data Approximation*, Cambridge university press. MR2131724 (2006i:41002)

[28] Wu, X., Zhu, X., Wu, G.-Q., and Ding, W. (2014). Data mining with big data. *IEEE Trans. Knowl. Data Eng.* **26** 97–107.

[29] Zhang, Y., Duchi, J., and Wainwright, M. (2015). Divide and conquer kernel ridge regression: A distributed algorithm with minimax optimal rates. *J. Mach. Learn. Res.* **16** 3299–3340.MR3450540

[30] Zhang, Y., Wainwright, M., and Duchi, J. (2013). Communication-efficient algorithms for statistical optimization. *J. Mach. Learn. Res.* **14** 3321–3363. MR3144464

[31] Zhou, D.-X. and Jetter, K. (2006). Approximation with polynomial kernels and SVM classifiers. *Adv. Comput. Math.* **25** 323–344. MR2231707 (2008k:62129)

[32] Zhou, Z., Chawla, N., Jin, Y., and Williams, G. (2014). Big data opportunities and challenges: Discussions from data analytics perspectives. *IEEE Comput. Intell. Mag.* **9** 62–74.