# Learning Capability of Relaxed Greedy Algorithms

Shaobo Lin, Yuanhua Rong, Xingping Sun, and Zongben Xu

***Abstract*— In the practice of machine learning, one often encounters problems in which noisy data are abundant while the learning targets are imprecise and elusive. To these challenges, most of the traditional learning algorithms employ hypothesis spaces of large capacity. This has inevitably led to high computational burdens and caused considerable machine sluggishness. Utilizing greedy algorithms in this kind of learning environment has greatly improved machine performance. The best existing learning rate of various greedy algorithms is proved to achieve the order of $(m/\log m)^{-1/2}$, where $m$ is the sample size. In this paper, we provide a relaxed greedy algorithm and study its learning capability. We prove that the learning rate of the new relaxed greedy algorithm is faster than the order $m^{-1/2}$. Unlike many other greedy algorithms, which are often indecisive issuing a stopping order to the iteration process, our algorithm has a clearly established stopping criteria.**

***Index Terms*— Algorithm, generalization error, learning theory, orthogonal greedy algorithm, relaxed greedy algorithm.**

## I. Introduction

MACHINE learning refers to training a computer system to perform a task with available data of the form $(x_i, y_i)_{i=1}^m$. The data are also known as examples. In mathematical terms, training means synthesizing a function $f$ that best represents the relation between inputs $x$ and the corresponding outputs $y$. The function $f$ is chosen from a suitable class of functions, called hypothesis space, encoding prior knowledge on the relation between $x$ and $y$. Statistically, a learning algorithm is an inference process from the (often unruly) data to a reasonable decision-making model based on the capacity of the hypothesis space. One of the main goals of learning theory is to design an efficient algorithm such that the synthesized function can approximate the best possible relation between $x$ and $y$, as the number of available data increases.

The core of learning theory is a quantitative assessment of the inference property of a learning algorithm. The central question is and will always be: how well does a synthesized function generalize to reflect the reality that the prior examples purport to show us? The celebrated representation theorem [5] in machine learning, in a certain sense, asserts that an original minimization problem in a usually infinite dimensional reproducing kernel Hilbert spaces (RKHS) can be reduced to finding finitely many coefficients in a linear combination of kernel function. This has greatly propelled the popular use of RKHS as hypothesis spaces in support vector machine (SVM) algorithms and regularized least square (RLS) algorithms [3]–[6], [9], [11], [13], [15], [20], [22]. Many of these algorithms are easy to implement as long as the number of samples $m$ is not too large. However, memory requirements for storing the symmetric positive semidefinite matrix increase quadratically with $m$. As such, large data sets pose a serious problem for the above-mentioned algorithms, and often cause sluggishness in machine performance. Further complications can evolve when dealing with problems in which data are noisy and learning targets imprecise and elusive.

To these challenges, most traditional algorithms employ hypothesis spaces of large capacities. While learning theorists cheer the philosophy behind the no pain, no gain strategy, practitioners are often at their wit's end in handling the high and sometimes insurmountable computation burdens brought forth by the introduction of super-large hypothesis spaces. To tackle this dichotomy, Barron *et al.* [2] advocate greedy algorithms in this kind of learning environment. They show that the learning rates of both orthogonal greedy algorithms (OGA) and relax greedy algorithms (RGA) are in the order $(m/\log m)^{-1/2}$. In achieving this goal, they have essentially used a complexity regularization principle as the stopping criterion. Literature abounds in the analysis of pros and cons of OGA and RGA. A consensus in the machine learning community is that OGA are inherently sensitive to noise, a hallmark of many machine learning problems. Our numerical experiments in Section 4 also support this viewpoint. Nevertheless, OGA have been widely used in signal processing [1], [12], [17] and approximation theory [8], [16], [18].

Two of the authors of the current paper have been avid practitioners of the RGA proposed in [2]. A long time period of exposure to and experimentation with the RGA has motivated them to improve the algorithms' efficiency and user-friendliness. In this paper, we propose a prior knowledge-dependent RGA and study its generalization capability. Using an alternating projection method, we are able to obtain a representation of the RGA estimator (in closed form) in every step of the iterations, and reduce the computation complexity to a level comparable to that of the pure greedy algorithm (PGA). Furthermore, we find that if a truncation operator is applied to the estimator in every step, then the learning capability is monotonically nonincreasing with respect to the number of iterations, which essentially establishes a new and easy-to-code stopping criteria. This feature is well-received by the community of computer programmers we have been working with. Using a widely used $l^2$ empirical covering number technique, we show that the learning rate of the

Manuscript received November 20, 2012; revised March 14, 2013 and May 20, 2013; accepted May 23, 2013. Date of publication June 13, 2013; date of current version September 27, 2013. This work was supported in part by the National 973 Program under Grant 2013CB329404, the Key Program of National Natural Science Foundation of China, under Grant 11131006, and the National Natural Science Foundations of China under Grant 61075054.

S. Lin, Y. Rong, and Z. Xu are with the Institute for Information and System Sciences, School of Mathematics and Statistics, Xiàn Jiaotong University, Xiàn 710049, China (e-mail: sblin1983@gmail.com; yuanhuarong@stu.xjtu.edu.cn; zbxu@mail.xjtu.edu.cn).

X. Sun is with the Department of Mathematics, Missouri State University, Springfield, MO 65897 USA (e-mail: xsun@missouristate.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TNNLS.2013.2265397

algorithm is faster than the order $m^{-1/2}$, which, to the best of our knowledge, is a new record ever achieved by any classical RGAs. Numerous numerical simulation results confirm that the algorithm is more stable in dealing with noisy machine learning problems than OGAs. In comparison to RLS and SVM algorithms, we witness noticeable machine performance enhancement made possible by the new RGA.

The rest of the paper is organized as follows. In Section II, we review notations and preliminary results in machine learning theory and greedy algorithms that will be frequently referred to throughout this paper. In Section III, we introduce our new relaxed greedy algorithms and state the main results of the paper. In Section IV, we present numerical simulation results that compare the performance of our new RGA with several other algorithms. Section V is devoted to proofs of the main results and Section VI concluding remarks.

## II. Preliminaries

### A. Review of Learning Theory

In most of the machine learning problems, data are taken from two sets: the input space $X \subseteq \mathbf{R}^d$ and the output space $Y \subseteq \mathbf{R}$. The relation between the variable $x \in X$ and the variable $y \in Y$ is not deterministic, and is described by a probability distribution $\rho$ on $Z := X \times Y$ that admits the decomposition

$$\rho(x, y) = \rho_X(x)\rho(y|x)$$

in which $\rho(y|x)$ denotes the conditional (given $x$) probability measure on $Y$, and $\rho_X(x)$ the marginal probability measure on $X$. Let $\mathbf{z} = (x_i, y_i)_{i=1}^m$ be a set of finite random samples of size $m$, $m \in \mathbf{N}$, drawn identically, independently according to $\rho$ from $Z$. The set of examples $\mathbf{z}$ is called a training set. Without loss of generality, we assume that $|y_i| \leq M$ for a prescribed (and fixed) $M > 0$.

The major goal in a machine learning problem is to derive from a training set a function $f : X \to Y$ such that $f(x)$ is an effective and reliable estimate of $y$ when $x$ is given. One natural measurement of the error incurred by using $f(x)$ for this purpose is the generalization error, given by

$$\mathcal{E}(f) := \int_Z (f(x) - y)^2 d\rho$$

which is minimized by the regression function [7], defined by

$$f_\rho(x) := \int_Y y d\rho(y|x).$$

This ideal minimizer $f_\rho$ exists in theory only. In practice, we do not know $\rho$, and we can only access random examples from $X \times Y$ sampled according to $\rho$.

Let $L^2_{\rho_X}$ be the Hilbert space of $\rho_X$ square integrable functions on $X$, with norm denoted by $\|\cdot\|_\rho$. With the assumption that $f_\rho \in L^2_{\rho_X}$, it is well-known [5] that, for every $f \in L^2_{\rho_X}$, there holds

$$\mathcal{E}(f) - \mathcal{E}(f_\rho) = \|f - f_\rho\|_\rho^2. \tag{1}$$

The task of the least square regression problem is then to construct functions $f_{\mathbf{z}}$ that approximates $f_\rho$, in the norm $\|\cdot\|_\rho$, using finite samples.

### B. Greedy Algorithm

There exist several types of greedy algorithms [16]. The four most commonly used are the pure greedy, orthogonal greedy, relax greedy, and stepwise projection algorithms, which are often denoted by their acronyms PGA, OGA, RGA, and SPA, respectively.

Let $H$ be a Hilbert space with norm and inner product $\|\cdot\|_H$ and $\langle\cdot,\cdot,\rangle_H$. Let $D_n := \{g_i\}_{i=1}^n$ be a given dictionary. In all the above greedy algorithms, we begin by setting $f_0 := 0$. The new approximation $f_k$ ($k \geq 1$) is defined based on $f_{k-1}$ and its residual $r_{k-1} := f - f_{k-1}$. In relaxed greedy algorithms, $f_k$ is defined as

$$f_k = \alpha_k f_{k-1} + \beta_k g_k$$

where $(\alpha_k, \beta_k) \in \mathbf{R}^2$ and $g_k \in D_n$. There exist many methods to choose $(\alpha_k, \beta_k, g_k)$, and the most greedy approach is

$$(\alpha_k, \beta_k, g_k) := \arg\min_{(\alpha,\beta,g)\in\mathbf{R}^2\times D_n} \|f - \alpha f_{k-1} - \beta g\|_H.$$

To reduce the computational burden, one chooses (judiciously) the first parameter $\alpha_k$, and then determine $\beta_k, g_k$ using the following optimization [2], [8]:

$$(\beta_k, g_k) := \arg\min_{(\beta,g)\in\mathbf{R}\times D_n} \|f - \alpha_k f_{k-1} - \beta g\|_H. \tag{2}$$

Given a training sample $\mathbf{z} = (x_i, y_i)_{i=1}^m$, the empirical inner product and norm are defined by

$$\langle f, g\rangle_m := \frac{1}{m}\sum_{i=1}^m f(x_i)g(x_i), \quad \|f\|_m^2 := \frac{1}{m}\sum_{i=1}^m |f(x_i)|^2.$$

Barron *et al.* [2] studied the RGA as described in (2) with $\alpha_k = 1 - 1/k$, $\|\cdot\|_H = \|\cdot\|_m$, and $\langle\cdot,\cdot,\rangle_H = \langle\cdot,\cdot\rangle_m$. To rephrase their result precisely, we need to reintroduce some of their notations. Let $D_n \subset D := \{g_i\}_{i=1}^\infty$, $\mathcal{L}_1(D) := \{f : f = \sum_{g\in D} a_g g\}$. The norm of $\mathcal{L}_1(D)$ is defined by $\|f\|_{\mathcal{L}_1} := \inf\left\{\sum_{g\in D}|a_g| : f = \sum_{g\in D} a_g g\right\}$. For $r > 0$, the space $\mathcal{L}_1^r$ is defined to be the set of all functions $f$ such that, for all $n$, there exists $h \in \text{span}\{D_n\}$ such that

$$\|h\|_{\mathcal{L}_1} \leq \mathcal{B}, \text{ and } \|f - h\| \leq \mathcal{B}n^{-r} \tag{3}$$

where $\|\cdot\|$ denotes the uniform norm for $C(X)$ and $C(X)$ denotes the spaces of continuous functions defined on $X$. The infimum of all such $\mathcal{B}$ defines a norm (for $f$ ) on $\mathcal{L}_1^r$.

One of the main results in [2] can be stated as follows. Let $\hat{f} := \Pi_M f_{k^*}$, where $\Pi_M f(x) := \min\{M, |f(x)|\}\text{sgn}(f(x))$ is the truncation operator at level $M$ and

$$k^* := \arg\min_{1\leq k\leq n}\left\{\|y - \Pi_M f_k\|_m^2 + \kappa\frac{k\log m}{m}\right\} \tag{4}$$

with $\kappa \geq 12840M^4$. Then for each $f_\rho \in \mathcal{L}_1^r$, we have

$$\mathbf{E}(\|\hat{f} - f_\rho\|_\rho^2) \leq C\left((1 + \mathcal{B}^2)\left(\frac{m}{\log m}\right)^{-1/2} + n^{-2r}\right) \tag{5}$$

where $C$ is a constant depending only on $\kappa$ and $M$. If $n$ is sufficiently large, then (5) shows that the learning rate of the RGA with the number of iterations satisfying (4) is $\mathcal{O}(m/\log m)^{-1/2}$. Barron *et al.* [2, Theorem 3.8] also showed that the RGA is weakly universally consistent.

## III. Prior Dependent RGA and its Learning Capability

### A. Motivation

Two of the authors (of the present paper) have had an extensive experience in the numerical implementation of RGA defined in (4). We have run simulations of problems from both virtual and real world: from weather forecasting to video game design. To a large extent, we have met with success. However, we have also encountered glitches, stemming mostly from the difficulty of communicating the sophisticated mathematical ideas in RGA to programmers who coded the numerical simulations. In particular, we had a hard time trying to convince the programmers that there is a certain mathematical necessity to use the truncating operator in the algorithm.

Programmers believed that the truncation operators have made the estimator $\hat{f}$ difficult to code. They also confirmed a remark in [2, Remark 3.5] that the estimate on the parameter $k$ is too pessimistic. More precisely, it follows from (4) and $|y| \le M$ that

$$\begin{aligned}\kappa\frac{k^*\log m}{m} &\le \|y-\Pi_M f_{k^*}\|_m^2+\kappa\frac{k^*\log m}{m}\\ &\le \|y-\Pi_M f_0\|_m^2 \le M^2.\end{aligned}$$

This implies that $k^*$ is not larger than $M^2m/(\kappa\log m)$. If $\kappa \ge 12840M^4$ and $m \le 12840M^2$, then only one iteration can be done in implementing the RGA. Many of the programmers' spirit are dampened by this restriction, and shy away from running the RGA for large $k$.

These experiences have motivated us to modify the above RGA for the purpose of making it more user-friendly and efficient. In the following subsection, we introduce a prior dependent RGA. Granted, the RGA in [2] are built independently of any prior knowledge and are weakly universal consistent. Theoretically, these algorithms are applicable in a variety of learning environment. The RGA we are proposing is prior dependent in nature. More practically efficient and noise-resistant as they manifest, we are unable to show that the new RGA is weakly universal consistent.

### B. Algorithm

In this part, we design a new RGA depending on the prior knowledge (3) and analyze the computational feasibility of the new algorithm.

Given a dictionary $D_n := (g_i)_{i=1}^n$, let

$$D_n^* := \left\{\pm\frac{g_i}{\|g_i\|_m} : i = 1,\dots,n\right\}.$$

We introduce our new RGA as follows:

$$f_0 = 0,\ \ f_{\mathbf{z}}^k := \alpha_k f_{\mathbf{z}}^{k-1} + \beta_{\mathbf{z}}^k g_{\mathbf{z}}^k \tag{6}$$

in which

$$\begin{aligned}&\alpha_0 = 0, \alpha_k := 1-\frac{1}{k}, \text{for } k \ge 1\\ &g_{\mathbf{z}}^k := \arg\max_{g\in D_n^*}\sum_{i=1}^m\left(y_i-\alpha_k f_{\mathbf{z}}^{k-1}(x_i)\right)g(x_i)\end{aligned}$$

and

$$\begin{aligned}\beta_{\mathbf{z}}^k := \arg\min_{\beta\in[0,\mathcal{B}/k]}(&\beta^2\\ &-2\beta\frac{1}{m}\sum_{i=1}^m(y_i-\alpha_k f_{\mathbf{z}}^{k-1}(x_i))g_{\mathbf{z}}^k(x_i)).\end{aligned}\tag{7}$$

For each $g \in D_n^*$, we have

$$\begin{aligned}&\frac{1}{m}\sum_{i=1}^m\left|y_i-\alpha_k f_{\mathbf{z}}^{k-1}(x_i)-\beta g(x_i)\right|^2\\ &\quad=\frac{1}{m}\sum_{i=1}^m\left|y_i-\alpha_k f_{\mathbf{z}}^{k-1}(x_i)\right|^2+\beta^2\frac{1}{m}\sum_{i=1}^m g^2(x_i)\\ &\qquad-2\beta\frac{1}{m}\sum_{i=1}^m\left(y_i-\alpha_k f_{\mathbf{z}}^{k-1}(x_i)\right)g(x_i)\\ &\quad=\frac{1}{m}\sum_{i=1}^m\left|y_i-\alpha_k f_{\mathbf{z}}^{k-1}(x_i)\right|^2+\beta^2\\ &\qquad-2\beta\frac{1}{m}\sum_{i=1}^m\left(y_i-\alpha_k f_{\mathbf{z}}^{k-1}(x_i)\right)g(x_i).\end{aligned}$$

It follows from the definition of $g_{\mathbf{z}}^k$ that for each given $\beta \in [0, \mathcal{B}/k]$, we have:

$$g_{\mathbf{z}}^k = \arg\min_{g\in D_n^*}\frac{1}{m}\sum_{i=1}^m\left|y_i-\alpha_k f_{\mathbf{z}}^{k-1}(x_i)-\beta g(x_i)\right|^2. \tag{8}$$

Similarly, for each given $g \in D_n^*$, we have

$$\beta_{\mathbf{z}}^k = \arg\min_{\beta\in[0,\mathcal{B}/k]}\frac{1}{m}\sum_{i=1}^m\left|y_i-\alpha_k f_{\mathbf{z}}^{k-1}(x_i)-\beta g(x_i)\right|^2. \tag{9}$$

Thus, the proposed algorithm (6) is a RGA with its parameter $\beta^k$ and $g_k$ chosen according to an alternating projection strategy. Furthermore, the definitions of $g_{\mathbf{z}}^k$ and $D_n^*$ yield that

$$\sum_{i=1}^m\left(y_i-\alpha_k f_{\mathbf{z}}^{k-1}(x_i)\right)g_{\mathbf{z}}^k(x_i) \ge 0$$

which implies that the solution of (7) is

$$\beta_{\mathbf{z}}^k = \min\left\{\frac{\sum_{i=1}^m\left(y_i-\alpha_k f_{\mathbf{z}}^{k-1}(x_i)\right)g_{\mathbf{z}}^k(x_i)}{\sum_{j=1}^m g_{\mathbf{z}}^k(x_j)}, \frac{\mathcal{B}}{k}\right\}.$$

From the above statement, it can be found that the coefficient $\beta_{\mathbf{z}}^k$ depends on the prior knowledge $\mathcal{B}$ defined in (3). So the new RGA we presented is a prior dependent algorithm, and $\mathcal{B}$ is called as the prior parameter. We observe further that $f_k \in \text{span}(D_n)$ for all $k \in \mathbf{N}$ and the computational complexity of algorithm (6) is similar to that of PGA. Moreover, it is easy to see that the derived estimator $f_{\mathbf{z}}^k$ belongs to $\text{span}(D_k)$ since the truncation operator is only implemented upon $\beta_{\mathbf{z}}^k$ in every step.

### C. Learning Rate

Before giving the main result, we need to introduce few notations. Let $s \in (0, 1]$ and $\phi(x, y)$ be a continuous function on $X \times X$ such that for all $y, x, x' \in X$, there holds

$$|\phi(y,x)-\phi(y,x')| \le C_s|x-x'|^s, \quad \max|\phi(y,x)| \le 1. \tag{10}$$

Here $C_s$ is a positive constant depending only on $s$. We remark that the first inequality in (10) is simply the Lipchitz continuity of $\phi$ with respect to $x$, and that the second inequality is essentially the boundedness of $\phi$ with respect to both $x$ and $y$. There is an abundant supply of such functions. For example, for each fixed $a > 0$, the widely used Gausssian kernel $G(x, y) = \exp\{-|x - y|^2/a\}$ fulfills the assumption (10) with $s = 1$, as are the inverse multiquadrics and most of the Wendland functions ( [19]). For a given set of $n$ distinct points $z_1, \ldots, z_n$, let

$$D_n := \{\phi(z_i, \cdot), z_i \in X, i = 1, 2, \ldots, n\} \tag{11}$$

and $D := \{\phi(x, \cdot) : x \in X\}$.

We state the main result of this paper and give the proof in Section V.

*Theorem 1:* Let $\delta \in (0, 1)$, $D_n$ and $f_{\mathbf{z}}^k$ be defined in (11) and (6), respectively. If $f_\rho \in \mathcal{L}_1^r$, then the inequality

$$\mathcal{E}(f_{\mathbf{z}}^k) - \mathcal{E}(f_\rho) \le C\mathcal{B}^2 \left( m^{-\frac{2s+d}{2d+2s}} \log \frac{2}{\delta} + k^{-1} + n^{-2r} \right) \tag{12}$$

holds with probability at least $1 - \delta$, where $C$ is a positive constant depending only on $\phi, d$, and $f_\rho$.

*Remark 1:* In a certain sense, the learning rate established in (12) is faster than that in (5). Indeed, if we choose $n$ and $k$ large enough, then the learning rate of (12) is asymptotically $m^{-2s+d/2d+2s}$, which is faster than $m^{-1/2}$.

*Remark 2:* The presence of the term $k^{-1}$ in the error estimate of Theorem 1 is very much to the liking of programmers who think it gives a well-indicated stopping criteria for iteration. If $k^{-1}$ is smaller than $n^{-2r}$, then the generalization error do not increase when $k$ does. Roughly speaking, the larger the number of iteration, the better is the generalization error we get. This is in sharp contrast to a related feature manifested by most of OGA.

*Remark 3:* In (12), we give a probabilistic estimate rather than expectation estimate. The latter can be derived from the former with a standard probabilistic manipulation.

## IV. Simulation Results

In this section, we present the results of two numerical experiments[1] in which we test the performance of the new RGA as described in Section III. In each experiment, we run the new RGA and several other algorithms (SVM, OGA, and RLS) to the same machine learning problem. Needless to say, we want to have a level competing ground to compare these algorithms. In the first experiment we randomly sampled 500 data from the function

$$x \mapsto \frac{\sin x}{x} \quad x \in [-1, 1].$$

Let $x_0 = -1$, and let $x_i (0 \le i < 200)$ be the 200 equally spaced points in $[-1, 1]$. We use $\left\{e^{-\|x-x_i\|^2} : i = 0, \ldots, 199\right\}$ as dictionary. To show the stability of RGA (6), we also added a Gaussian noise $N(0, \delta^2)$ with $\delta^2 = 0.2$. Recalling (7), there is also a prior parameter $\mathcal{B}$ in the new RGA. In this simulation,

[1]All the numerical simulations are carried out in MATLAB2009b environment running Windows 7, 2.66 HZ CPU.

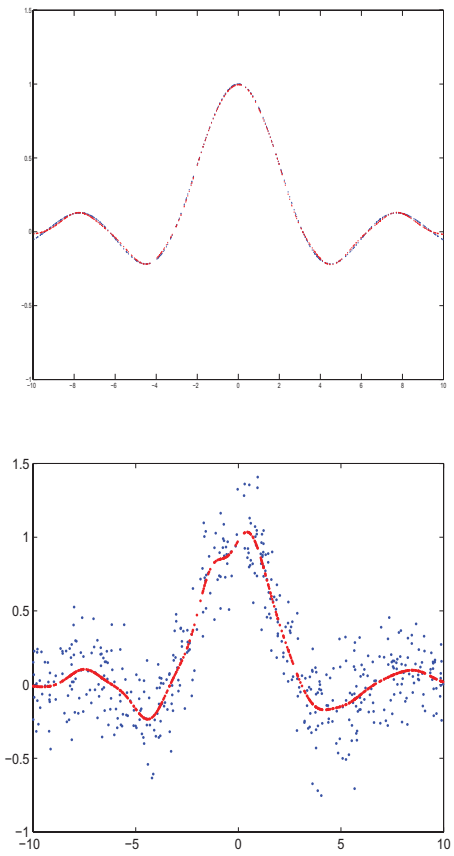



Fig. 1. RGA with noiseless data and noisy data.

we use the so-called cross validation strategy [10, Chap. 8] to select an optimal $\mathcal{B}$ from the set $\{1, 10, 100, 1000\}$. In the numerical experiments, RMSETest is the rooted mean square error (RMSE) of the testing data. RMSETrain is the RMSE of the training data. Sparsity shows the average number of functions in the dictionary used to construct the estimator and Time represents the average time (in seconds) of running one simulation (We run 20 simulations). Fig. 1 shows an intuitive effects of RGA learning. Table I summarizes the performance of the algorithms handling non-noisy data. We see that all the algorithms demonstrate good generalization capability. OGA and RGA are better-behaved than SVM and RLS in terms of the training time. Table 2 shows the results of 20 simulations when noise is added in. We observe that RGA is much more robust than OGA in dealing with noisy data. Both Tables II and III show that the average number of functions used in the dictionary in constructing the estimator of RGA and OGA are much smaller than that in RLS, which shows that RGA and OGA are capable of producing sparse yet efficient estimators. Overall, the numerical experiments underscore the fact that the new RGA (6) is stable, fast, and efficient.

In Fig. 2, we describe the relationships between test error and number of iteration (the upper left figure), training time, and number of iteration (the upper right figure), number of functions used in the dictionary and number of iteration (the lower left figure). The lower right figure shows test error for $k = 5$ (blue line), $k = 50$ (green line), and $k = 2000$ (red line). All of these support the result of Theorem 1.

TABLE I

RESULTS OF 20 SIMULATIONS (NO NOISE ADDED). NUMBER OF ITERATION IN RGA IS 2000 AND PRIOR PARAMETER $\mathcal{B}$ IS 10

| Methods | RMSETest | RMSETrain | Sparsity | Time |
|---|---|---|---|---|
| OGA | 0.0094 | 0.0091 | 18.700 | 0.8479 |
| RGA | 0.0102 | 0.0097 | 83.300 | 3.2674 |
| RLS | 0.0104 | 0.0098 | 500 | 84.700 |
| SVM | 0.0070 | 0.0071 | 161.300 | 70.481 |

TABLE II

RESULTS OF 20 SIMULATIONS WITH NOISY DATA. NUMBER OF ITERATION IN RGA IS 2000 AND PRIOR PARAMETER $\mathcal{B}$ IS 10

| Methods | RMSETest | RMSETrain | Sparsity | Time |
|---|---|---|---|---|
| OGA | 0.0786 | 0.1871 | 60.7000 | 2.0607 |
| RGA | 0.0391 | 0.1981 | 79.4500 | 3.2667 |
| RLS | 0.0412 | 0.1962 | 500 | 74.3499 |
| SVM | 0.0382 | 0.1965 | 43.6000 | 90.3201 |

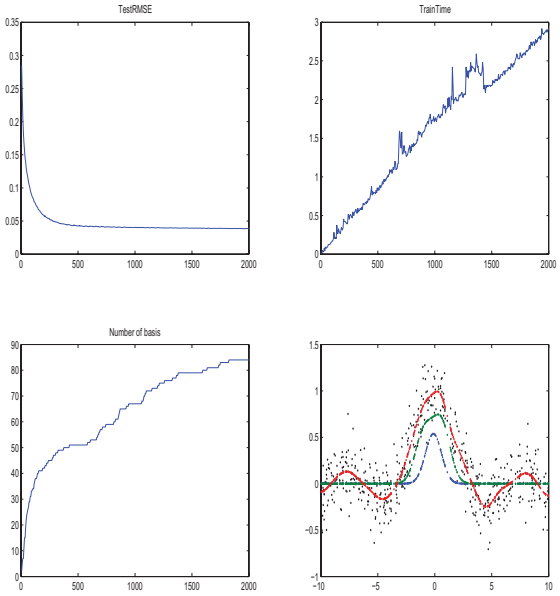



Fig. 2. Criteria to choose parameter $k$ for the univariate case.

In the second numerical experiment, we sampled 1000 data from

$$f(x) = 0.1e^{-\|x-z_1\|^2} + 0.2e^{-\|x-z_2\|^2} + 0.3e^{-\|x-z_3\|^2} + 0.4e^{-\|x-z_4\|^2}$$

according to uniform distribution on $[0,1]^4$, where $\{z_i\}_{i=1}^4$ are arbitrarily chosen from $[0,1]^4$. Again we added in Gaussian noise $N(0,\delta^2)$ with $\delta^2 = 0.1$. It can be easily found that the prior parameter $\mathcal{B} = 1$ in this simulation. We use as our dictionary 400 functions of the form $e^{-\|x-x_i\|^2}$, where the centers $x_i'$s are drawn from $[0,1]^4$ according to uniform distribution on $[0,1]^4$. Table III shows the results of this numerical experiment. Similar to Fig. 2, Fig 3. depicts the relationships between test error and number of iteration (the left figure), training time and number of iteration (the middle figure), number of functions used in the dictionary and number of iteration (the right figure).

TABLE III

NUMERICAL RESULTS FOR 20 MULTIVARIATE SIMULATIONS WITH NOISY DATA. NUMBER OF ITERATION IN RGA IS 2000 AND THE PRIOR PARAMETER $\mathcal{B}$ IS 1

| Methods | RMSETest | RMSETrain | Sparsity | Time |
|---|---|---|---|---|
| RGA | 0.0103 | 0.0992 | 19.2503 | 13.952 |
| OGA | 0.0192 | 0.0997 | 15.7000 | 3.0601 |
| RLS | 0.0191 | 0.0973 | 1000 | 19.775 |
| SVM | 0.0163 | 0.0985 | 319.4520 | 43.663 |

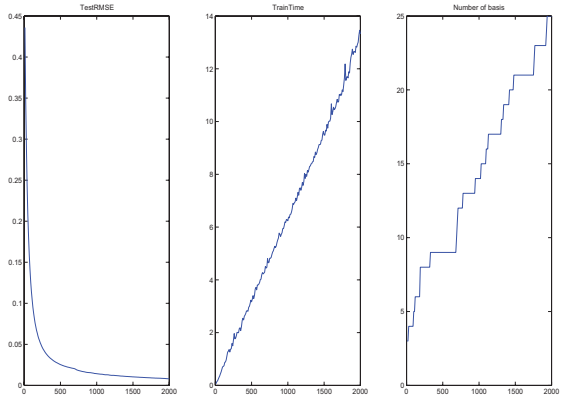



Fig. 3. Criteria to choose parameter $k$ for the multivariate case.

## V. PROOF OF THEOREM 1

In this section, we prove our main result (Theorem 1). The proof is divided into five parts, which contains the error decomposition strategy, approximation error estimate, sample error estimate, hypothesis error estimate, and learning rate analysis.

### A. Error Decomposition Strategy

To give an error decomposition strategy for $\mathcal{E}(f_{\mathbf{z}}^k) - \mathcal{E}(f_\rho)$, we need to construct a function $f_k^* \in \text{span}(D_n)$ as follows. As $f_\rho \in \mathcal{L}_1^r$, there exists a $h_\rho := \sum_{i=1}^n a_i g_i \in \text{span}(D_n)$ such that

$$\|h_\rho\|_{\mathcal{L}_1} \le \mathcal{B}, \text{ and } \|f_\rho - h_\rho\| \le \mathcal{B}n^{-r}. \tag{13}$$

Define

$$f_0^* = 0, \quad f_k^* = \left(1 - \frac{1}{k}\right) f_{k-1}^* + \frac{\sum_{i=1}^n |a_i| \|g_i\|_\rho}{k} g_k^* \tag{14}$$

where

$$g_k^* := \arg\max_{g \in D_n'} \left\langle h_\rho - \left(1 - \frac{1}{k}\right) f_{k-1}^*, g \right\rangle_\rho$$

and

$$D_n' := \left\{g_i(x)/\|g_i\|_\rho\right\}_{i=1}^n \bigcup \left\{-g_i(x)/\|g_i\|_\rho\right\}_{i=1}^n$$

with $g_i \in D_n$.

Let $f_{\mathbf{z}}^k$ and $f_k^*$ be defined as in (6) and (14), respectively, then we have

$$\begin{aligned}&\mathcal{E}(f_{\mathbf{z}}^k)-\mathcal{E}(f_\rho)\\&\quad\le \mathcal{E}(f_k^*)-\mathcal{E}(f_\rho)+\mathcal{E}_{\mathbf{z}}(f_{\mathbf{z}}^k)-\mathcal{E}_{\mathbf{z}}(f_k^*)\\&\qquad+\mathcal{E}_{\mathbf{z}}(f_k^*)-\mathcal{E}(f_k^*)+\mathcal{E}(f_{\mathbf{z}}^k)-\mathcal{E}_{\mathbf{z}}(f_{\mathbf{z}}^k)\end{aligned}$$

where $\mathcal{E}_{\mathbf{z}}(f)=(1/m)\sum_{i=1}^m(y_i-f(x_i))^2$.

Upon making the short-hand notations

$$\begin{aligned}\mathcal{D}(k)&:=\mathcal{E}(f_k^*)-\mathcal{E}(f_\rho)\\\mathcal{S}(\mathbf{z},k)&:=\mathcal{E}_{\mathbf{z}}(f_k^*)-\mathcal{E}(f_k^*)+\mathcal{E}(f_{\mathbf{z}}^k)-\mathcal{E}_{\mathbf{z}}(f_{\mathbf{z}}^k)\end{aligned}$$

and

$$\mathcal{P}(\mathbf{z},k):=\mathcal{E}_{\mathbf{z}}(f_{\mathbf{z}}^k)-\mathcal{E}_{\mathbf{z}}(f_k^*)$$

respectively for the approximation error, the sample error and the hypothesis error, we have

$$\mathcal{E}(f_{\mathbf{z}}^k)-\mathcal{E}(f_\rho)=\mathcal{D}(k)+\mathcal{S}(\mathbf{z},k)+\mathcal{P}(\mathbf{z},k). \tag{15}$$

This completes our error decomposition strategy.

### *B. Approximation Error*

In this subsection, we give an upper bound estimate for $\mathcal{D}(k)$.

*Proposition 1:* Let $f_k^*$ be defined in (14). If $f_\rho\in\mathcal{L}_1^r$, then

$$\mathcal{D}(k)\le\mathcal{B}^2(k^{-1/2}+n^{-r})^2. \tag{16}$$

*Proof:* From the definition of $\mathcal{D}(k)$ and (1), it follows that for arbitrary $h\in\operatorname{span}(D_n)$, there holds:

$$\begin{aligned}\mathcal{D}(k)&=\mathcal{E}(f_k^*)-\mathcal{E}(f_\rho)=\|f_k^*-f_\rho\|_\rho^2\\&\le\left(\|f_k^*-h\|_\rho+\|h-f_\rho\|_\rho\right)^2.\end{aligned}$$

As $f_\rho\in\mathcal{L}_1^r$ and $\|f\|_\rho\le\|f\|$, (13) and (14) imply

$$\begin{aligned}\mathcal{D}(k)&\le\left(\|f_k^*-h_\rho\|_\rho+\|h-f_\rho\|_\rho\right)^2\\&\le\left(\|f_k^*-h_\rho\|_\rho+\|h-f_\rho\|\right)^2\\&\le\left(\|f_k^*-h_\rho\|_\rho+\mathcal{B}n^{-r}\right)^2.\end{aligned}$$

To bound $\|f_k^*-h_\rho\|_\rho$, we note that for arbitrary $\varphi\in D_n'$

$$\begin{aligned}&\left\|h_\rho-\alpha_k f_{k-1}^*-\frac{\sum_{i=1}^n|a_i|\|g_i\|_\rho}{k}\varphi\right\|_\rho^2\\&\quad=\left\|h_\rho-\alpha_k f_{k-1}^*\right\|_\rho^2+\frac{\left(\sum_{i=1}^n|a_i|\|g_i\|_\rho\right)^2}{k^2}\\&\qquad-2\frac{\sum_{i=1}^n|a_i|\|g_i\|_\rho}{k}\left\langle h_\rho-\alpha_k f_{k-1}^*,\varphi\right\rangle_\rho.\end{aligned}$$

It follows from the definition of $g_k^*$ that:

$$\begin{aligned}g_k^*=\arg\min_{\varphi\in D_n'}&\left\|h_\rho-\left(1-\frac{1}{k}\right)f_{k-1}^*\right.\\&\left.-\frac{\sum_{i=1}^n|a_i|\|g_i\|_\rho}{k}\varphi\right\|_\rho^2.\end{aligned} \tag{17}$$

Using the fact that $h_\rho$ is a linear combination of the $g_i$s, we write

$$h_\rho=\sum_{i=1}^n a_i g_i=\sum_{i=1}^n a_i\|g_i\|_\rho\frac{g_i}{\|g_i\|_\rho}=\sum_{i=1}^n a_i\|g_i\|_\rho\varphi_i.$$

Thus we obtain $h_\rho(x)=\sum_{i=1}^n|a_i|\|g_i\|_\rho\varphi_i^*(x)$ where $\varphi_i:=g_i/\|g_i\|_\rho\in D_n'$ and $\varphi_i^*:=\begin{cases}\varphi_i, & a_i\ge 0\\-\varphi_i, & a_i<0\end{cases}$. It follows from (17) and (14) that, for all $\varphi\in D_n'$, there holds:

$$\begin{aligned}&\|h_\rho-f_k^*\|_\rho^2\\&\quad\le\left\|\alpha_k f_{k-1}^*+\frac{\sum_{i=1}^n|a_i|\|g_i\|_\rho}{k}\varphi-h_\rho\right\|_\rho^2\\&\quad=\left\|\alpha_k(f_{k-1}^*-h_\rho)+\frac{\sum_{i=1}^n|a_i|\|g_i\|_\rho}{k}\varphi-\frac{h_\rho}{k}\right\|_\rho^2\\&\quad=\alpha_k^2\|f_{k-1}^*-h_\rho\|_\rho^2+\frac{\left(\sum_{i=1}^n|a_i|\|g_i\|_\rho\right)^2}{k^2}\\&\qquad-\frac{2}{k^2}\left\langle\sum_{i=1}^n|a_i|\|g_i\|_\rho\varphi,h_\rho\right\rangle+\frac{1}{k^2}\|h_\rho\|_\rho^2\\&\qquad+2\alpha_k\left\langle f_{k-1}^*-h_\rho,\frac{\sum_{i=1}^n|a_i|\|g_i\|_\rho}{k}\varphi-\frac{h_\rho}{k}\right\rangle.\end{aligned}$$

The above inequality holds true for all $\varphi_1^*,\ldots,\varphi_n^*\in D_n'$. Thus we have

$$\begin{aligned}&\|h_\rho-f_k^*\|_\rho^2\\&\le\frac{1}{k}\left[\left(\alpha_k^2\|f_{k-1}^*-h_\rho\|_\rho^2+\frac{\left(\sum_{i=1}^n|a_i|\|g_i\|_\rho\right)^2}{k^2}\right.\right.\\&\qquad-\frac{2}{k^2}\left\langle\sum_{i=1}^n|a_i|\|g_i\|_\rho\varphi_1^*,h_\rho\right\rangle_\rho+\frac{1}{k^2}\|h_\rho\|_\rho^2\\&\qquad\left.+2\alpha_k\left\langle f_{k-1}^*-h_\rho,\frac{\sum_{i=1}^n|a_i|\|g_i\|_\rho}{k}\varphi_1^*-\frac{h_\rho}{k}\right\rangle_\rho\right)\\&\qquad+\cdots+\left(\alpha_k^2\|f_{k-1}^*-h_\rho\|_\rho^2+\frac{\left(\sum_{i=1}^n|a_i|\|g_i\|_\rho\right)^2}{k^2}\right.\\&\qquad-\frac{2}{k^2}\left\langle\sum_{i=1}^n|a_i|\|g_i\|_\rho\varphi_n^*,h_\rho\right\rangle_\rho+\frac{1}{k^2}\|h_\rho\|_\rho^2\\&\qquad\left.\left.+2\alpha_k\left\langle f_{k-1}^*-h_\rho,\frac{\sum_{i=1}^n|a_i|\|g_i\|_\rho}{k}\varphi_n^*-\frac{h_\rho}{k}\right\rangle_\rho\right)\right]\\&\quad=\alpha_k^2\|f_{k-1}^*-h_\rho\|_\rho^2-\frac{1}{k^2}\|h_\rho\|_\rho^2\\&\qquad+\left(\frac{\sum_{i=1}^n|a_i|\|g_i\|_\rho}{k}\right).\end{aligned}$$

Noting that $\|h_\rho\|_{\mathcal{L}_1}=\sum_{i=1}^n|a_i|\|g_i\|_\rho$, we therefore, obtain

$$\|h_\rho-f_k^*\|_\rho^2\le\alpha_k^2\|h_\rho-f_{k-1}^*\|_\rho^2+\frac{1}{k^2}(\|h_\rho\|_{\mathcal{L}_1}^2-\|h_\rho\|_\rho^2).$$

Therefore, similar method as that in the proof of [2, Theorem 2.2] yields the estimation

$$\|h_\rho-f_k^*\|_\rho^2\le(\|h_\rho\|_{\mathcal{L}_1}^2-\|h_\rho\|_\rho^2)^{1/2}k^{-1/2}.$$

Hence, the inequality $\|h_\rho\|_\rho\le\|h_\rho\|_{\mathcal{L}_1}\le\mathcal{B}$ implies

$$\begin{aligned}\mathcal{D}(k)&\le\left(\|h_k^*-h_\rho\|_\rho+\mathcal{B}n^{-r}\right)^2\\&\le\mathcal{B}^2(k^{-1/2}+n^{-r})^2.\end{aligned}$$

This completes the proof of Proposition 1. ■

### C. Sample Error

In this subsection, we will bound the sample error $\mathcal{S}(\mathbf{z}, k)$. Upon using the short-hand notations

$$S_1(\mathbf{z}, k) := \{\mathcal{E}_{\mathbf{z}}(f_k^*) - \mathcal{E}_{\mathbf{z}}(f_\rho)\} - \{\mathcal{E}(f_k^*) - \mathcal{E}(f_\rho)\}$$

and

$$S_2(\mathbf{z}, k) := \{\mathcal{E}(f_{\mathbf{z}}^k) - \mathcal{E}(f_\rho)\} - \{\mathcal{E}_{\mathbf{z}}(f_{\mathbf{z}}^k) - \mathcal{E}_{\mathbf{z}}(f_\rho)\}$$

we write

$$\mathcal{S}(\mathbf{z}, k) = \mathcal{S}_1(\mathbf{z}, k) + \mathcal{S}_2(\mathbf{z}, k). \tag{18}$$

To bound $S_1(\mathbf{z}, k)$, we need the following two lemmas. The first one gives an upper-bound estimate for $\|f_k^*\|$. The second one is the well-known Bernstein inequality that can be found in [14].

*Lemma 1:* Let $f_k^*$ be defined in (14), then there holds

$$\|f_k^*\| \le \mathcal{B}. \tag{19}$$

*Proof:* It follows from the definition of $f_k^*$ that

$$f_k = \alpha_k\alpha_{k-1}\cdots\alpha_2\beta_1 g_1^* + \alpha_k\alpha_k\alpha_{k-1}\cdots\alpha_3\beta_2 g_2^* + \cdots + \alpha_k\beta_{k-1}g_{k-1}^* + \beta_k g_k^*.$$

As $\alpha_k = 1 - 1/k$, $\beta_k = (1/k)\sum_{i=1}^n |a_i|\|g_i\|_\rho \le \mathcal{B}/k$, we get

$$\begin{aligned}\|f_k^*\|_{\mathcal{L}^1} &\le \left|\left(1-\frac{1}{k}\right)\cdots\left(1-\frac{1}{k-(k-2)}\right)\mathcal{B}\right| \\ &\quad + \left|\left(1-\frac{1}{k}\right)\cdots\left(1-\frac{1}{k-(k-3)}\right)\frac{\mathcal{B}}{2}\right| \\ &\quad + \cdots + \left|\left(1-\frac{1}{k}\right)\frac{\mathcal{B}}{k-1}\right| + \frac{\mathcal{B}}{k} \\ &= k\times\frac{\mathcal{B}}{k} \le \mathcal{B}.\end{aligned}$$

We then use the assumptions made in (10) to finish the proof of Lemma 1. ∎

*Lemma 2:* Let $\xi$ be a random variable on a probability space $Z$ with variance $\sigma^2$ satisfying $|\xi - \mathbf{E}\xi| \le M_\xi$ for some constant $M_\xi$. Then for any $0 < \delta < 1$, with confidence $1 - \delta$, we have

$$\frac{1}{m}\sum_{i=1}^m \xi(z_i) - \mathbf{E}\xi \le \frac{2M_\xi \log\frac{1}{\delta}}{3m} + \sqrt{\frac{2\sigma^2\log\frac{1}{\delta}}{m}}.$$

*Proposition 2:* For any $0 < \delta < 1$, with confidence $1 - \delta/2$

$$S_1(\mathbf{z}, k) \le \frac{7(3M + \mathcal{B}\log\frac{2}{\delta})}{3m} + \frac{1}{2}\mathcal{D}(k).$$

*Proof:* Let the random variable $\xi$ on $Z$ be defined by

$$\xi(\mathbf{z}) = (y - f_k^*(x))^2 - (y - f_\rho(x))^2 \quad \mathbf{z} = (x, y) \in Z.$$

As $|f_\rho(x)| \le M$ almost everywhere, it follows from Lemma 1 that

$$\begin{aligned}|\xi(\mathbf{z})| &= (f_\rho(x) - f_k^*(x))(2y - f_k^*(x) - f_\rho(x)) \\ &\le (M + \mathcal{B})(3M + \mathcal{B}) \\ &\le M_\xi := (3M + \mathcal{B})^2\end{aligned}$$

and almost surely

$$|\xi - \mathbf{E}\xi| \le 2M_\xi.$$

Moreover, we have

$$\begin{aligned}&\mathbf{E}(\xi^2) \\ &= \int_Z (f_k^*(x) + f_\rho(x) - 2y)^2(f_k^*(x) - f_\rho(x))^2 d\rho \\ &\le M_\xi\|f_\rho - f_k^*\|_\rho^2\end{aligned}$$

which implies that the variance $\sigma^2$ of $\xi$ can be bounded as $\sigma^2 \le \mathbf{E}(\xi^2) \le M_\xi\mathcal{D}(k)$. Now applying Lemma 2, with confidence $1 - \delta/2$, we have

$$\begin{aligned}\mathcal{S}_1(\mathbf{z}, k) &= \frac{1}{m}\sum_{i=1}^m \xi(z_i) - E\xi \\ &\le \frac{4M_\xi\log\frac{2}{\delta}}{3m} + \sqrt{\frac{2M_\xi\mathcal{D}(k)\log\frac{2}{\delta}}{m}} \\ &\le \frac{7(3M + \mathcal{B})^2\log\frac{2}{\delta}}{3m} + \frac{1}{2}\mathcal{D}(k).\end{aligned}$$

∎

To bound $\mathcal{S}_2(\mathbf{z}, k)$, we need the concept of an empirical covering number.

*Definition 1:* Let $(\mathcal{M}, d)$ be a pseudo-metric space and $T \subset \mathcal{M}$ a subset. For every $\varepsilon > 0$, the covering number $\mathcal{N}(T, \varepsilon, d)$ of $T$ with respect to $\varepsilon$ and $d$ is defined as the minimal number of balls of radius $\varepsilon$ whose union covers $T$, that is

$$\mathcal{N}(T, \varepsilon, d) := \min\left\{l \in \mathbf{N} : T \subset \bigcup_{j=1}^l B(t_j, \varepsilon)\right\}$$

for some $\{t_j\}_{j=1}^l \subset \mathcal{M}$, where $B(t_j, \varepsilon) = \{t \in \mathcal{M} : d(t, t_j) \le \varepsilon\}$.

The $l^2$-empirical covering number of a function set is defined by means of the normalized $l^2$-metric $d_2$ on the Euclidean space $\mathbf{R}^d$ given in [14] with $d_2(\mathbf{a}, \mathbf{b}) = ((1/m)\sum_{i=1}^m |a_i - b_i|^2)^{1/2}$ for $\mathbf{a} = (a_i)_{i=1}^m$, $\mathbf{b} = (b_i)_{i=1}^m \in \mathbf{R}^m$.

*Definition 2:* Let $\mathcal{F}$ be a set of functions on $X$, $\mathbf{x} = (x_i)_{i=1}^m \subset X^m$, and let

$$\mathcal{F}|_{\mathbf{x}} := \{(f(x_i))_{i=1}^m : f \in \mathcal{F}\} \subset R^m.$$

Set $\mathcal{N}_{2,\mathbf{x}}(\mathcal{F}, \varepsilon) = \mathcal{N}(\mathcal{F}|_{\mathbf{x}}, \varepsilon, d_2)$. The $l^2$-empirical covering number of $\mathcal{F}$ is defined by

$$\mathcal{N}_2(\mathcal{F}, \varepsilon) := \sup_{m\in\mathbf{N}}\sup_{\mathbf{x}\in S^m}\mathcal{N}_{2,\mathbf{x}}(\mathcal{F}, \varepsilon), \quad \varepsilon > 0.$$

The following two lemmas can be found, respectively, in [14, Theorem 2] and [21].

*Lemma 3:* If $\phi$ satisfies (10), then for arbitrary $\varepsilon > 0$

$$\log\mathcal{N}_2(B_1, \varepsilon) \le C_1\varepsilon^{-\frac{2d}{d+2s}}$$

where $B_R$ is the ball in $\mathcal{L}_1$ with radius $R$, and $C_1$ is a constant depending only on $s$, $C_s$, and $X$.

*Lemma 4:* Let $\mathcal{F}$ be a class of measurable functions on $Z$. Assume that there are constants $B, c > 0$ and $\alpha \in [0, 1]$ such that $\|f\|_\infty \le B$ and $\mathbf{E}f^2 \le c(\mathbf{E}f)^\alpha$ for every $f \in \mathcal{F}$. If for some $a > 0$ and $p \in (0, 2)$

$$\log\mathcal{N}_2(\mathcal{F}, \varepsilon) \le a\varepsilon^{-p} \quad \forall\varepsilon > 0 \tag{20}$$

then there exists a constant $c'_p$ depending only on $p$ such that for any $t > 0$, with probability at least $1 - e^{-t}$, there holds

$$\mathbf{E}f - \frac{1}{m}\sum_{i=1}^{m} f(z_i) \le \frac{1}{2}\eta^{1-\alpha}(\mathbf{E}f)^{\alpha} + c'_p\eta + 2\left(\frac{ct}{m}\right)^{\frac{1}{2-\alpha}} + \frac{18Bt}{m} \ \forall f \in \mathcal{F} \tag{21}$$

where

$$\eta := \max\left\{c^{\frac{2-p}{4-2\alpha+p\alpha}}\left(\frac{a}{m}\right)^{\frac{2}{4-2\alpha+p\alpha}}, B^{\frac{2-p}{2+p}}\left(\frac{a}{m}\right)^{\frac{2}{2+p}}\right\}.$$

By using the same method as that in the proof of Lemma 1, we obtain the following result.

*Lemma 5:* Let $f_{\mathbf{z}}^k$ be defined in (6), then there holds

$$\|f_{\mathbf{z}}^k\|_{\mathcal{L}_1} \le \mathcal{B}. \tag{22}$$

*Proof:* It follows from the definition of $f_{\mathbf{z}}^k$ that

$$f_{\mathbf{z}}^k = \alpha_k\alpha_{k-1}\cdots\alpha_2\beta_{\mathbf{z}}^1 g_{\mathbf{z}}^k + \alpha_k\alpha_k\alpha_{k-1}\cdots\alpha_3\beta_{\mathbf{z}}^2 g_{\mathbf{z}}^2 + \cdots + \alpha_k\beta_{\mathbf{z}}^{k-1} g_{\mathbf{z}}^{k-1} + \beta_{\mathbf{z}}^k g_{\mathbf{z}}^k.$$

Since $\alpha_k = 1 - 1/k$ and $\beta_{\mathbf{z}}^k \le \mathcal{B}/k$, we get

$$\begin{aligned}\|f_{\mathbf{z}}^k\|_{\mathcal{L}^1} &\le \left|\left(1-\frac{1}{k}\right)\cdots\left(1-\frac{1}{k-(k-2)}\right)\mathcal{B}\right| + \left|\left(1-\frac{1}{k}\right)\cdots\left(1-\frac{1}{k-(k-3)}\right)\frac{\mathcal{B}}{2}\right| + \cdots + \left|\left(1-\frac{1}{k}\right)\frac{\mathcal{B}}{k-1}\right| + \frac{\mathcal{B}}{k} \\ &= k \times \frac{\mathcal{B}}{k} \le \mathcal{B}.\end{aligned}$$

We then use the assumptions made in (10) to finish the proof of Lemma 5. ∎

We are now in a position to establish an upper bound estimate for $\mathcal{S}_2(\mathbf{z}, k)$.

*Proposition 3:* Let $f_{\mathbf{z}}^k$ be defined as in (6) and $0 < \delta < 1$, then with confidence $1 - \delta/2$, there holds

$$\mathcal{S}_2(\mathbf{z}, k) \le \frac{1}{2}\{\mathcal{E}(f_{\mathbf{z}}^k) - \mathcal{E}(f_\rho)\} + C_3 \log\frac{2}{\delta} m^{-\frac{d+2s}{2d+2s}}$$

where $C_3$ is a constant depending only on $d$, $X$, $\phi$ and $M$.

*Proof:* We apply Lemma 4 to the set of functions $\mathcal{F}_R$, where

$$\mathcal{F}_R := \left\{(y - f(x))^2 - (y - f_\rho(x))^2 : f \in B_R\right\}. \tag{23}$$

Each function $g \in \mathcal{F}_R$ has the form

$$g(z) = (y - f(x))^2 - (y - f_\rho(x))^2, \quad f \in B_R,$$

and is automatically a function on $Z$. Hence

$$\mathbf{E}g = \mathcal{E}(f) - \mathcal{E}(f_\rho) = \|f - f_\rho\|_\rho^2$$

and

$$\frac{1}{m}\sum_{i=1}^{m} g(z_i) = \mathcal{E}_{\mathbf{z}}(f) - \mathcal{E}_{\mathbf{z}}(f_\rho).$$

Observe that

$$g(z) = (f(x) - f_\rho(x))((f(x) - y) + (f_\rho(x) - y)).$$

Using the obvious inequalities $\|f\|_\infty \le R$ a.e. $|f_\rho| \le M$ a.e., we get the inequalities

$$|g(z)| \le (R + M)(R + 3M) \le (R + 3M)^2$$

and

$$\begin{aligned}\mathbf{E}g^2 &= \int_Z (2y - f(x) - f_\rho(x))^2(f(x) - f_\rho(x))^2 d\rho \\ &\le (R + 3M)^2\mathbf{E}g.\end{aligned}$$

For $g_1, g_2 \in \mathcal{F}_R$, we have

$$\begin{aligned}|g_1(z) - g_2(z)| &= |(y - f_1(x))^2 - (y - f_2(x))^2| \\ &\le (2M + 2R)|f_1(x) - f_2(x)|.\end{aligned}$$

It follows that:

$$\begin{aligned}\mathcal{N}_{2,\mathbf{z}}(\mathcal{F}_R, \varepsilon) &\le \mathcal{N}_{2,\mathbf{x}}\left(B_R, \frac{\varepsilon}{2M + 2R}\right) \\ &\le \mathcal{N}_{2,\mathbf{x}}\left(B_1, \frac{\varepsilon}{R(2M + 2R)}\right).\end{aligned}$$

Using the above inequality and Lemma 3, we have

$$\log\mathcal{N}_{2,\mathbf{z}}(\mathcal{F}_R, \varepsilon) \le C_1(2MR + 2R^2)^{\frac{2d}{d+2s}}\varepsilon^{-\frac{2d}{d+2s}}.$$

By Lemma 4 with $B = c = (3M + R)^2$, $\alpha = 1$ and $a = C_1(2MR + 2R^2)^{2d/d+2s}$, we know that for any $\delta \in (0, 1)$, with confidence $1 - \delta/2$, there exists a constant $C$ depending only on $d$, $X$, and $\phi$ such that for all $g \in \mathcal{F}_R$

$$\mathbf{E}g - \frac{1}{m}\sum_{i=1}^{m} g(z_i) \le \frac{1}{2}\mathbf{E}g + C\eta + C(M + 1)^2\frac{\log(4/\delta)}{m}.$$

Here

$$\eta = \{(3M + R)^2\}^{\frac{s}{s+d}}\left(\frac{(2RM + 2R^2)^{\frac{2d}{d+2s}}}{m}\right)^{\frac{d+2s}{2d+2s}}.$$

Therefore, there exists a constant $C_2$ depending only on $d$, $X$, $\phi$ and $M$ such that

$$\eta \le C_2R^2m^{-\frac{d+2s}{2d+2s}}$$

which implies

$$\mathbf{E}g - \frac{1}{m}\sum_{i=1}^{m} g(z_i) \le \frac{1}{2}\mathbf{E}g + CC_2R^2\log\frac{\delta}{2}m^{-\frac{d+2s}{2d+2s}}.$$

By Lemma 5, we know that $\|f_{\mathbf{z}}^k\|_{\mathcal{L}_1} \le \mathcal{B}$. It follows that there exists a constant $C_3$ depending only on $d$, $X$, $\phi$ and $M$ such that

$$\mathcal{S}_2(\mathbf{z}, k) \le \frac{1}{2}\{\mathcal{E}(f_{\mathbf{z}}^k) - \mathcal{E}(f_\rho)\} + C_3\mathcal{B}^2\log\frac{\delta}{4}m^{-\frac{2s+d}{2d+2s}}.$$

∎

### D. Hypothesis Error

In this subsection, we give an error estimate for $\mathcal{P}(\mathbf{z}, k)$.

*Proposition 4:* If $f_{\mathbf{z}}^k$ and $f_k^*$ are defined in (6) and (14), then we have

$$\mathcal{P}(\mathbf{z}, k) \leq \mathcal{B}^2 k^{-1}.$$

*Proof:* For arbitrary $h \in \text{span}(D_n)$ satisfying $\|h\|_{\mathcal{L}_1} \leq \mathcal{B}$, there exists a set of real numbers $\{b_i\}_{i=1}^n$ such that

$$\begin{aligned} h(x) &= \sum_{i=1}^n b_i g_i(x) = \sum_{i=1}^n b_i \|g_i\|_m \frac{g_i(x)}{\|g_i\|_m} \\ &= \sum_{i=1}^n b_i \|g_i\|_m \varphi_i(x) = \sum_{i=1}^n |b_i| \|g_i\|_m \varphi_i^*(x) \end{aligned}$$

where $\varphi_i := g_i/\|g_i\|_m \in D_n^*$ and $\varphi_i^* := \begin{cases} \varphi_i, & b_i \geq 0 \\ -\varphi_i, & b_i < 0 \end{cases}$. Now we prove

$$\|f_{\mathbf{z}}^k - y\|_m^2 - \|h - y\|_m^2 \leq \frac{\mathcal{B}}{k}. \tag{24}$$

From (8) and (9), it follows that for arbitrary fixed $\beta \in [0, \mathcal{B}/k]$, the inequalities

$$\begin{aligned} \|y - f_{\mathbf{z}}^k\|_m^2 &\leq \left\| y - \alpha_k f_{\mathbf{z}}^{k-1} - \beta\varphi \right\|_m^2 \\ &= \left\| \alpha_k (y - f_{\mathbf{z}}^{k-1}) + \frac{1}{k} y - \beta\varphi \right\|_m^2 \\ &= \alpha_k^2 \|y - f_{\mathbf{z}}^{k-1}\|_m^2 + 2\alpha_k \left\langle y - f_{\mathbf{z}}^{k-1}, \frac{1}{k} y - \beta\varphi \right\rangle_m \\ &\quad + \frac{1}{k^2} \|y - h\|_m^2 + \frac{1}{k} \left\langle y - h, \frac{1}{k} h - \beta\varphi \right\rangle_m \\ &\quad + \frac{1}{k^2} \|h\|_m^2 + \beta^2 - \frac{2}{k} \langle h, \beta\varphi \rangle_m \end{aligned}$$

hold for all $\varphi \in D_n^*$. As the same as the proof of Proposition 1, it also holds true for every function in the convex hull of $D_n^*$. Setting $\beta = (1/k) \sum_{i=1}^n |b_i| \|g_i\|_m$, and applying Hölder's inequality, we get

$$\begin{aligned} \|y - f_{\mathbf{z}}^k\|_m^2 &\leq \alpha_k^2 \|y - f_{\mathbf{z}}^{k-1}\|_m^2 \\ &\quad + 2\alpha_k \left\langle y - f_{\mathbf{z}}^{k-1}, \frac{1}{k} y - \frac{1}{k} h \right\rangle_m \\ &\quad + \frac{1}{k^2} \|y - h\|_m^2 - \frac{1}{k^2} \|h\|_m^2 + \beta^2 \\ &= \left\| \alpha_k (y - f_{\mathbf{z}}^{k-1}) + \frac{1}{k} (y - h) \right\|_m^2 - \frac{1}{k^2} \|h\|_m^2 + \beta^2 \end{aligned}$$

As $\beta = \|h\|_{\mathcal{L}_1}/k$, we obtain

$$\begin{aligned} \|y - f_{\mathbf{z}}^k\|_m^2 \leq \left( \alpha_k \|y - f_{\mathbf{z}}^{k-1}\|_m + \frac{1}{k} \|y - h\|_m \right)^2 \\ + \frac{\|h\|_{\mathcal{L}_1}^2 - \|h\|_m^2}{k^2}. \end{aligned}$$

This is similar to that in [2, (2.45)]. Thus, using the similar method as that in the proof of [2, Theorem 2.4], we can deduce

$$\mathcal{P}(\mathbf{z}, k) \leq \mathcal{B}^2 k^{-1} \tag{25}$$

directly. This finishes the proof of Proposition 4. ∎

### E. Final Derivation of the Error Estimate

*Proof of Theorem 1:* We assemble the results in Propositions 1 through 4 and (15) to write

$$\begin{aligned} \mathcal{E}(f_{\mathbf{z}}^k) - \mathcal{E}(f_\rho) &\leq \mathcal{D}(k) + \mathcal{S}(\mathbf{z}, k) + \mathcal{P}(\mathbf{z}, k) \\ &= \mathcal{D}(k) + \mathcal{S}_1(\mathbf{z}, k) + \mathcal{S}_2(\mathbf{z}, k) + \mathcal{P}(\mathbf{z}, k) \\ &\leq \frac{3}{2} (\mathcal{B}^2 (k^{-1/2} + n^{-r})^2) + \frac{7(3M + \mathcal{B} \log \frac{2}{\delta})}{3m} \\ &\quad + \frac{1}{2} \{\mathcal{E}(f_{\mathbf{z}}^k) - \mathcal{E}(f_\rho)\} + C_3 \log \frac{2}{\delta} m^{-\frac{d+2s}{2d+2s}} \\ &\quad + \mathcal{B}^2 k^{-1} \end{aligned}$$

holds with confidence at least $1 - \delta$. Therefore

$$\mathcal{E}(f_{\mathbf{z}}^k) - \mathcal{E}(f_\rho) \leq C\mathcal{B}^2 \left( \log \frac{2}{\delta} m^{-\frac{2s+d}{2d+2s}} + k^{-1} + n^{-2r} \right) \tag{26}$$

holds with confidence at least $1 - \delta$, where $C$ is a constant depending only on $\phi$, $d$, $X$, and $M$. This completes the proof of Theorem 1.

## VI. Concluding Remarks

The main contributions of the present paper can be summarized as follows. Firstly, one important tool used in [2] is the truncation operators that act as a deportation vehicle. Namely, they send elements in the original dictionary out of the dictionary. This has added a layer of difficulty in the coding process. We have modified this process by truncating in every iteration step. We have succeeded in getting an estimator within the dictionary without compromising its generalization capability. Secondly, most practitioners of greedy algorithms prefer to have clear cut stopping criteria for the iteration which most of the existing error estimates for greedy algorithms are lacking. We have made a stride in this direction. Our error estimate contains only one term involving the number of iteration. Furthermore, this term is inversely proportional to the number of iteration. This type of error estimate for greedy algorithms has been well-received in the programming community. Finally, our new RGA has improved the previously established learning rates for greedy algorithms. Precisely, our error estimate yields a learning rate (in probability terms) that is faster than $m^{-1/2}$.

To make sense of the RGA presented in this paper, the following two remarks are required.

*Remark 4:* Generally speaking, the tug of war between bias and variance dictates that a small hypothesis space gives rise to a large approximation error (or hypothesis error) while a large one gives rise to a large sample error. To reach and stay in the happy middle, one needs to carefully adjust her balance act in each step of the iteration. This is the basic strategy we followed in the proof of Theorem 1. A prevailing conception is that the error estimate should have had more terms containing $k$, besides the term $k^{-1}$. However, this is not what our proof has witnessed. We succeeded in deriving an error estimate in the present form largely because we realize that applying the truncation operator in each step of iteration does not essentially increase the capacity of the hypothesis space. To elaborate, we show first (in the proof of Proposition 3) that the $l^2$ empirical covering number of the $\mathcal{F}_R$ defined as in (23) depends only

on $R$. We then show (in the proof of Lemma 5) that $R$ does not increase with the number of iteration. A desirable consequence of this process is that generalization error does not increase when $k$ increases.

*Remark 5:* Practitioners have frequently asked us how to choose the prior parameter $\beta$ in the new RGA algorithm (6). This is a very good question. Admittedly, it is often unlikely to have enough prior knowledge as required by (3) in any machine learning problems. Thus, judiciously choosing a value for $\mathcal{B}$ is crucial. If $\mathcal{B}$ is chosen to be too large, then Theorem 1 gives a relatively weaker generalization error estimate. If $\mathcal{B}$ is chosen to be too small, then the algorithm can only be applicable to a limited variety of real world machine learning problems. In a certain sense, this is a reflection of the general Bias-Variance problem we alluded to in Remark 4. We have been advising our programmers to use a cross validation scheme to select an appropriate prior parameter $\mathcal{B}$.

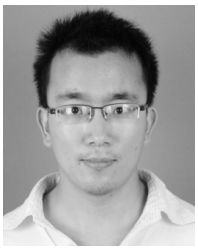
We conclude this paper with the sober note: there is still much room for improvement. When applying the algorithm to big data real world problems, we have been seeing mixed results. Our programmers are still fretting about the frequent improvising and supervising work required in the trial-and-error stage of selecting the parameter $\mathcal{B}$. Admittedly, most of the real world data contain a lot of raw material that come in a variety of shapes and sizes. In many circumstances, we know what have caused the less-than-ideal system performance. But we do not know why. We will keep working on this interesting project, and report our progress in a future publication.

## ACKNOWLEDGMENT

Three anonymous editors have carefully read this paper and have given us numerous constructive suggestions. As a result, the overall quality of this paper has been noticeably enhanced, for which we are very grateful.

## REFERENCES

[1] M. Avellaneda, G. Davis, and S. Mallat, "Adaptive greedy approximations," *Constructive Approximation*, vol. 13, no. 1, pp. 57–89, 1997.
[2] A. R. Barron, A. Cohen, W. Dahmen, and R. A. Devore, "Approximation and learning by greedy algorithms," *Ann. Stat.*, vol. 36, no. 1, pp. 64–94, Feb. 2008.
[3] P. Boudoulis, K. Slavakis, and S. Theodoridis, "Adaptive learning in complex reproducing kernel Hilbert spaces employing Writinger's. subgradients," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 23, no. 1, pp. 425–438, Jan. 2012.
[4] A. Caponnetto and E. DeVito, "Optimal rates for the regularized least squares algorithm," *Found. Comput. Math.*, vol. 7, no. 3, pp. 331–368, Jul. 2007.
[5] F. Cucker and S. Smale, "On the mathematical foundations of learning," *Bull. Amer. Math. Soc.*, vol. 39, no. 1, pp. 1–49, Oct. 2001.
[6] F. Cucker and S. Smale, "Best choices for regularization parameters in learning theory: On the bias–variance problem," *Found. Comput. Math.*, vol. 2, no. 4, pp. 413–428, Oct. 2002.
[7] F. Cucker and D. X. Zhou, *Learning Theory: An Approximation Theory Viewpoint*. Cambridge, U.K.: Cambridge Univ. Press, 2007.
[8] R. Devore and V. Temlyakov, "Some remarks on greedy algorithms," *Adv. Comput. Math.*, vol. 5, no. 1, pp. 173–187, 1996.
[9] T. Evgeniou, M. Pontil, and T. Poggio, "Regularization networks and support vector machines," *Adv. Comput. Math.*, vol. 13, no. 1, pp. 1–50, Apr. 2000.
[10] L. Györfy, M. Kohler, A. Krzyzak, and H. Walk, *A Distribution-Free Theory of Nonparametric Regression*. Berlin, Germany: Springer-Verlag, 2002.
[11] M. Krejnik and A. Tyutin, "Reproducing kernel Hilbert spaces with odd kernels in price prediction," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 23, no. 10, pp. 1564–1573, Oct. 2012.
[12] S. Kunis and H. Rauhut, "Random sampling of sparse trigonometric polynomials, II. Orthogonal matching pursuit versus basis pursit," *Found. Comput. Math.*, vol. 8, no. 6, pp. 737–763, Nov. 2008.
[13] B. Schölkopf, R. Herbrich, and A. J. Smola, "A generalized representer theorem," in *Proc. 14th Ann. Conf. Comput. Learn. Theory*, 2001, pp. 416–426.
[14] L. Shi, Y. L. Feng, and D. X. Zhou, "Concentration estimates for learning with $\ell^1$-regularizer and data dependent hypothesis spaces," *Appl. Comput. Harmon. Anal.*, vol. 31, no. 2, pp. 286–302, Sep. 2011.
[15] K. Slavakis, P. Bouboulis, and S. Theodoridis, "Adaptive multiregression in reproducing kernel Hilbert spaces: The multiaccess MIMO channel case," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 23, no. 2, pp. 260–276, Feb. 2012.
[16] V. Temlyakov," Nonlinear methods of approximation," *Found. Comput. Math.*, vol. 3, no. 1, pp. 33–107, Jan. 2003.
[17] V. Temlyakov and P. Zheltov, "On performance of greedy algorithms," *J. Approx. Theory*, vol. 163, no. 9, pp. 1134–1145, Sep. 2011.
[18] J. A. Tropp, "Greed is good: Algorithmic results for sparse approximation," *IEEE Trans. Inf. Theory*, vol. 50, no. 10, pp. 2231–2242, Oct. 2004.
[19] H. Wendland, "Piecewise polynomial, positive definite and compactly supported radial functions of minimal degree," *Adv. Comput. Math.*, vol. 4, no. 4, pp. 389–396, Dec. 1995.
[20] Q. Wu, Y. M. Ying, and D. X. Zhou, "Learning rates of least square regularized regression," *Found. Comput. Math.*, vol. 6, no. 2, pp. 171–192, Apr. 2006.
[21] Q. Wu, Y. Ying, and D. X. Zhou, "Multi-kernel regularized classifiers," *J. Complex.*, vol. 23, no. 1, pp. 108–134, 2007.
[22] Y. L. Xu, D. R. Chen, H. X. Li, and L. Liu, "Least square regularized regression in sum space," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 24, no. 4, pp. 635–646, Apr. 2013.
[23] D. X. Zhou and K. Jetter, "Approximation with polynomial kernels and SVM classifiers," *Adv. Comp. Math.*, vol. 25, nos. 1–3, pp. 323–344, Oct. 2006.



**Shaobo Lin** received the B.S. degree in mathematics and the M.S. degree in basic mathematics from Hangzhou Normal University, Hangzhou, China. He is currently pursuing the Ph.D. degree with Xi'an Jiaotong University, Xi'an, China.

His current research interests include machine learning and scattered data fitting.

**Yuanhua Rong** received the B.S. degree in information and computing science from the University of Electronic Science and Technology of China, Sichuan, China, in 2010. He is currently pursuing the Masters degree with Xi'an Jiaotong University, Xi'an, China.

His current research interests include learning theory, data mining, and compressed sensing.

**Xingping Sun** is a Professor with the Department of Mathematics, Missouri State University, Springfield, MO, USA. He has authored or co-authored over 40 papers in the areas of computational and applied harmonic analysis, approximation theory and numerical analysis. He is keen on applying classical analysis to these new areas of active research. His current research interests include statistical machine learning and data mining.

**Zongben Xu** was born in 1955. He received the Ph.D. degree in mathematics from Xi'an Jiaotong University, Xi'an, China, in 1987.

He is currently a Vice President with Xi'an Jiaotong University, the Chief Scientist of the National Basic Research Program of China (973 Project), and the Director of the Institute for Information and System Sciences of the University. He delivered a 45-minute talk at the International Congress of Mathematicians 2010. He was elected as a member of the Chinese Academy of Science in 2011. His current research interests include intelligent information processing and applied mathematics.

Dr. Xu received the National Natural Science Award of China in 2007 and the CSIAM Su Buchin Applied Mathematics Prize in 2008.