Journal of Inequalities and Applications
a SpringerOpen Journal

RESEARCH Open Access

The learning performance of the weak rescaled pure greedy algorithms

Qin Guo[1*], Xianghua Liu[2†] and Peixin Ye[2†]

*Correspondence:
guoqin_1985@163.com
†School of Science, Shandong Jianzhu University, Jinan, 250101, China
Full list of author information is available at the end of the article
†Equal contributors

Abstract

We investigate the regression problem in supervised learning by means of the weak rescaled pure greedy algorithm (WRPGA). We construct learning estimator by applying the WRPGA and deduce the tight upper bounds of the K-functional error estimate for the corresponding greedy learning algorithms in Hilbert spaces. Satisfactory learning rates are obtained under two prior assumptions on the regression function. The application of the WRPGA in supervised learning considerably reduces the computational cost while maintaining its powerful generalization capability when compared with other greedy learning algorithms.

Keywords: Least squares regression; Rescaled pure greedy algorithm; Weakness sequence; K-functional; Learning rate

1 Introduction

The applications of greedy algorithms to supervised learning have sparked great research interest because they have appealing generalization capability with lower computing burden than typical regularized methods, particularly in large-scale dictionary learning problem [1–6]. Big data sets for the most traditional learning algorithms frequently cause slow machine performance. To tackle this problem, many researchers [1–3, 7, 8] advocate greedy learning algorithms, which have greatly improved learning performance.

The approximation abilities of greedy-type algorithms for frames or more dictionaries $\mathcal{D}$ were investigated in [7, 9–12], as well as various applications, see [3, 7, 13–19]. The pure greedy algorithm (PGA) can realize the best bilinear approximation, see [20, 21]. Although the PGA is outstanding at computing, the main problem is that it lacks optimal convergence properties for a general dictionary, and consequently the slower convergence rate than the best nonlinear approximation [11, 21–23] corrupts its learning performance. To improve the approximation rate, the orthogonal greedy algorithm (OGA), the relaxed greedy algorithm (RGA), the stepwise projection algorithm (SPA), and their weak versions have been proposed. It was shown that these greedy algorithms all achieved the optimal rate $\mathcal{O}(m^{-\frac{1}{2}})$ for approximating the elements in the class $\mathcal{A}_1(\mathcal{D})$, which will be defined in (14), where m is the iteration number, see [9, 11].

Both the OGA and the RGA have recently been employed successfully in machine learning [1–3, 7, 8]. For example, Barron et al. [7] established the optimal convergence rate

Springer

$\mathcal{O}(n/\log n)^{-\frac{1}{2}})$, where n is the sample size. To reduce the OGA's computational load, Fang et al. [1] investigated the learning performance of the orthogonal super greedy algorithm (OSGA) and derived the almost same rate as the orthogonal greedy learning algorithm (OGLA). All these results demonstrate that each greedy learning algorithm has its advantages and disadvantages.

We study the applications of weak greedy algorithms to least squares regression in supervised learning. It is well known that the weak type are easier to implement than the usual greedy algorithms, see [12]. Specifically, the weak rescaled pure greedy algorithm (WRPGA), one fairly simple modification of the PGA, is the goal of our investigation, see [24, 25]. When compared to the OGA and the RGA, the WRPGA can also furthermore reduce the computational load. The best rate $\mathcal{O}(m^{-\frac{1}{2}})$ for functions in the basic sparse class has been proved [24]. Motivated by research results of [24], we proceed to use the same method employed for the RPGA in [24] to deduce the error bound of the K-functional estimate in the Hilbert space $\mathcal{H}$ for the WRPGA. The WRPGA is a simple greedy algorithm with good approximation ability. Based on this, we propose the weak rescaled pure greedy learning algorithm (WRPGLA) for solving the kernel-based regression problems in supervised learning. Using the WRPGA's proven approximation result, we can derive that the WRPGLA has the almost same learning rate as the OGLA. Our results show that the WRPGLA further cuts down the computational complexity even more without reducing generalization capabilities.

The paper is organized as follows. In Sect. 2, we review least squares regression learning theory and the WRPGA. In Sect. 3, we propose the WRPGLA and state the main theorems on the error estimates. Section 4 is devoted to proofs of the main results. We present the convergence rates under two smoothness assumptions on the regression function f_ρ in the last section.

2 Preliminaries

Some preliminaries are presented in this section. Sections 2.1 and 2.2 provide a fast overview of least squares regression learning and the WRPGA, respectively.

2.1 Least squares regression

In this paper, the approximation problem is addressed in the following statistical learning context. Let X be a compact metric space and $Y = \mathbb{R}$. Let ρ be a Borel probability measure on $Z = X \times Y$. The generalization error for a function $f : X \to Y$ is defined by

$$\mathcal{E}(f) = \int_Z \big(f(x) - y\big)^2 \, d\rho, \tag{1}$$

which is minimized by the following regression function:

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

where $\rho(\cdot|x)$ is the conditional distribution induced by ρ at $x \in X$. In regression learning, ρ is unknown, and what one can know is a set of samples $\mathbf{z} = \{z_i\}_{i=1}^n = \{(x_i, y_i)\}_{i=1}^n \in Z^n$ that are drawn independently and identically according to ρ. The goal of learning is to find a

good approximation $f_{\mathbf{z}}$ of f_ρ, which minimizes the empirical error

$$\mathcal{E}_{\mathbf{z}}(f) = \|y - f\|_n^2 := \frac{1}{n}\sum_{i=1}^{n}\big(f(x_i) - y_i\big)^2. \tag{2}$$

Denote the Hilbert space of the square integrable functions defined on X with respect to the measure ρ_X by $L^2_{\rho_X}(X)$, where ρ_X is the marginal measure of ρ on X. It is clear from the definition of $f_\rho(x)$ that for each $x \in X$, $\int_Y (f_\rho(x) - y)\, d\rho(y|x) = 0$. For any $f \in L^2_{\rho_X}(X)$, it holds that

$$\begin{aligned}\mathcal{E}(f) &= \int_Z \big(f(x) - f_\rho(x) + f_\rho(x) - y\big)^2\, d\rho \\ &= \int_X \big(f(x) - f_\rho(x)\big)^2\, d\rho_X + \int_Z \big(f_\rho(x) - y\big)^2\, d\rho \\ &\quad + 2\int_X \big(f(x) - f_\rho(x)\big)\, d\rho_X \int_Y \big(f_\rho(x) - y\big)\, d\rho(y|x) \\ &= \int_X \big(f(x) - f_\rho(x)\big)^2\, d\rho_X + \mathcal{E}(f_\rho).\end{aligned}$$

Therefore,

$$\mathcal{E}(f) - \mathcal{E}(f_\rho) = \|f - f_\rho\|^2 \tag{3}$$

with the norm $\|\cdot\|$

$$\|f\| = \left(\int_X |f(x)|^2 d\rho_X\right)^{\frac{1}{2}}. \tag{4}$$

The prediction accuracy of learning algorithms is measured by $E(\|f_{\mathbf{z}} - f_\rho\|^2)$.

We will assume $|y| \le B$ for a positive real number $B < \infty$ almost surely. In this paper, we construct the learning estimator $f_{\mathbf{z}}$ by applying the WRPGA and estimate $E(\|f_{\mathbf{z}} - f_\rho\|^2)$. So, in the following subsection, we recall this algorithm.

2.2 Weak rescaled pure greedy algorithm

We shall restrict our analysis to the situation in which approximation takes place in a real, separable Hilbert space $\mathcal{H}$ with the inner product $\langle\cdot,\cdot\rangle_{\mathcal{H}}$ and the norm $\|\cdot\| := \|\cdot\|_{\mathcal{H}} = \langle\cdot,\cdot\rangle_{\mathcal{H}}^{\frac{1}{2}}$. Let $\mathcal{D} \subset \mathcal{H}$ be a given dictionary satisfying $\|g\| = 1$ for every $g \in \mathcal{D}$, $g \in \mathcal{D}$ implies $-g \in \mathcal{D}$ and $\overline{\mathrm{Span}(\mathcal{D})} = \mathcal{H}$.

Petrova developed the rescaled pure greedy algorithm (RPGA) to enhance the PGA's convergence rate, which simply rescales f_m at the mth greedy step, see [24]. We begin by describing the weak rescaled pure greedy algorithm (WRPGA) also introduced by Petrova in [24].

WRPGA$(\{t_m\}, \mathcal{D})$:

Step 0: Let $f_0 := 0$.

Step m $(m \ge 1)$:

(1) If $f = f_{m-1}$, then terminate the iterative process and define $f_k = f_{m-1} = f$ for $k \ge m$.

(2) If $f \neq f_{m-1}$, then choose a direction $\varphi_m \in \mathcal{D}$ such that

$$\left|\langle f - f_{m-1}, \varphi_m \rangle\right| \geq t_m \sup_{\varphi \in \mathcal{D}} \left|\langle f - f_{m-1}, \varphi \rangle\right|, \tag{5}$$

where $\{t_m\}_{m=1}^{\infty}$ is a weakness sequence and $t_m \in (0, 1]$.

Let

$$\lambda_m := \langle f - f_{m-1}, \varphi_m \rangle, \tag{6}$$

$$\hat{f}_m := f_{m-1} + \lambda_m \varphi_m, \tag{7}$$

$$s_m := \frac{\langle f, \hat{f}_m \rangle}{\|\hat{f}_m\|^2}. \tag{8}$$

The m step approximation f_m is defined as

$$f_m = s_m \hat{f}_m, \tag{9}$$

and proceed to Step $m + 1$.

Remark 1 When $t_m = 1$, this algorithm is the RPGA. Note that if the supremum is not attained, one can select $t_m < 1$ and proceed with the algorithm. In this case, it is easier to choose φ_m. If the output at the mth greedy step was $\hat{f}_m$ rather than $f_m = s_m \hat{f}_m$, this would be the PGA. The WRPGA uses $s_m \hat{f}_m$, which is just suitable scaling of $\hat{f}_m$, and thus increases the rate to $\mathcal{O}(m^{-\frac{1}{2}})$ for functions in the closure of the convex hull of $\mathcal{D}$.

3 Weak rescaled pure greedy learning

We shall provide the WRPGLA for regression. From the definition of the WRPGA, computing $\sup_{\varphi \in \mathcal{D}} |\langle f - f_{m-1}, \varphi \rangle|$ may result in computation difficulty. Therefore we compute only over the truncation of the dictionary, which is a finite subset of $\mathcal{D}$. Let $\mathcal{D}_1 \subset \mathcal{D}_2 \subset \cdots \subset \mathcal{D}$. Then $\mathcal{D}_m$ is the truncation of $\mathcal{D}$ with the cardinality $\#(\mathcal{D}_m) = m$. Here we assume that

$$m \leq m(n) := \lfloor n^a \rfloor \quad \text{for some fixed } a \geq 1. \tag{10}$$

Then the WRPGLA is defined by the following simple processes.

WRPGLA:

Step 1: We apply the WRPGA for $\mathcal{D}_m$ to the function $y(x_i) = y_i$ by utilizing the norm $\|\cdot\|_n$ associated with the empirical inner product, that is,

$$\|f\|_n := \left(\frac{1}{n} \sum_{i=1}^{n} |f(x_i)|^2\right)^{\frac{1}{2}}.$$

Step 2: The algorithms establish the approximation $f_{\mathbf{z},k} := f_k$ to the data at the kth greedy step. Then, we define our estimator as $f_{\mathbf{z}} := T f_{\mathbf{z},k^*}$, where $Tu := T_B \min\{B, |u|\} \operatorname{sgn}(u)$ and

$$k^* := \arg\min_{k>0} \left\{ \|y - T f_{\mathbf{z},k}\|_n^2 + \kappa \frac{k \log n}{n} \right\}, \tag{11}$$

where the constant $\kappa \geq \kappa_0 = 2568 B^4 (a + 5)$, which will be discussed in proof of Theorem 1.

Remark 2 First, when $k = 0$, it follows from $f_0 = 0$ and $|y| \le B$ that $\kappa \frac{k \log n}{n} \le B^2$. This suggests that k^* is not larger than $\frac{Bn}{\kappa}$. Second, from the definition of the estimator, we observe further that the computing cost of the kth greedy step is less than $O(n^a)$. For the WRPGA, it only requires an additional computation of s_m.

To discuss the approximation properties of WRPGLA, we introduce the class of functions

$$\mathcal{A}_1^0(\mathcal{D}, M) := \left\{ f = \sum_{k \in \Lambda} c_k(f) \varphi_k : \varphi_k \in D, \#(\Lambda) < \infty, \sum_{k \in \Lambda} \left| c_k(f) \right| \le M \right\}, \tag{12}$$

and

$$\mathcal{A}_1(\mathcal{D}, M) = \overline{\mathcal{A}_1^0(\mathcal{D}, M)}. \tag{13}$$

Then

$$\mathcal{A}_1(\mathcal{D}) = \bigcup_{M>0} \mathcal{A}_1(\mathcal{D}, M) \tag{14}$$

and

$$\|f\|_{\mathcal{A}_1(\mathcal{D})} := \inf\left\{ M : f \in \mathcal{A}_1(\mathcal{D}, M) \right\}. \tag{15}$$

We also use the following K-functional:

$$K(f, t) := K\left(f, t, \mathcal{H}, \mathcal{A}_1(\mathcal{D})\right) := \inf_{h \in \mathcal{A}_1(\mathcal{D})} \left\{ \|f - h\|_{\mathcal{H}} + t \|h\|_{\mathcal{A}_1(\mathcal{D})} \right\}, \quad t > 0. \tag{16}$$

Since all the constants in this work depend at most on κ_0, B, and a, we denote all of them by C for simplicity of notation. Now we take $\mathcal{H} = L^2_{\rho_X}(X)$ with the norm defined by (4).

Then, we provide our main results on the generalization error bounds for the WRPGLA.

Theorem 1 *There exists κ_0 depending only on B and a such that if $\kappa \ge \kappa_0$, then for all $k > 0$ and $h \in \operatorname{Span}(\mathcal{D}_m)$, the learning estimator by applying the WRPGA satisfies*

$$E\left(\|f_{\mathbf{z}} - f_\rho\|^2 \right) \le 8 \frac{\|h\|^2_{\mathcal{A}_1(\mathcal{D}_m)}}{\sum_{i=1}^k t_i^2} + 2\|f_\rho - h\|^2 + C \frac{k \log n}{n}. \tag{17}$$

Furthermore, we have

$$E\left(\|f_{\mathbf{z}} - f_\rho\| \right) \le 2K\left(f_\rho, 2\left(\sum_{i=1}^k t_i^2 \right)^{-\frac{1}{2}} \right) + C \frac{k \log n}{n}. \tag{18}$$

Applying Theorem 1 with $t_i = t_0$ for all $i \ge 1$ and $0 < t_0 \le 1$, we get the following theorem.

Theorem 2 *Under the assumptions of Theorem 1, if $t_i = t_0$ for all $i \ge 1$ and $0 < t_0 \le 1$, then we have*

$$E\left(\|f_{\mathbf{z}} - f_\rho\|^2 \right) \le 8 \frac{\|h\|^2_{\mathcal{A}_1(\mathcal{D}_m)}}{k t_0^2} + 2\|f_\rho - h\|^2 + C \frac{k \log n}{n}. \tag{19}$$

Furthermore, we have

$$E\big(\|f_{\mathbf{z}}-f_\rho\|\big)\le 2K\big(f_\rho,2k^{-\frac{1}{2}}t_0^{-1}\big)+C\frac{k\log n}{n}. \tag{20}$$

4 Proofs of the main results

To prove Theorem 1, we establish a lemma on the upper error bound for the WRPGA.

Lemma 4.1 *If $f\in\mathcal{H}$, $h\in\mathcal{A}_1(\mathcal{D})$, then the output $(f_m)_{m\ge0}$ of the WRPGA satisfies*

$$e_m:=\|f-f_m\|\le 2K\left(f,\left(\sum_{k=1}^{m}t_k^2\right)^{-1/2}\right),\quad m=0,1,2,\ldots. \tag{21}$$

Proof In terms of the definition of K-functional, we just need to prove that for $f\in\mathcal{H}$ and $h\in\mathcal{A}_1(\mathcal{D})$,

$$e_m^2\le\|f-h\|^2+\frac{4}{\sum_{k=1}^{m}t_k^2}\|h\|_{\mathcal{A}_1(\mathcal{D})}^2,\quad m=1,2,\ldots. \tag{22}$$

Since $\mathcal{A}_1^0(\mathcal{D},M)$ is dense in $\mathcal{A}_1(\mathcal{D},M)$, it suffices to prove (22) for functions h that are finite sums $\sum_j c_j\varphi_j$ with $\sum_j|c_j|\le M$. We fix $\epsilon>0$ and select a representation for $h=\sum_{\varphi\in\mathcal{D}}c_\varphi\varphi$, such that

$$\sum_{\varphi\in\mathcal{D}}|c_\varphi|<M+\epsilon. \tag{23}$$

Denote

$$a_m:=e_m^2-\|f-h\|^2,\quad m=1,2,\ldots. \tag{24}$$

The nonincreasing of $\{e_m\}_{m=0}^{\infty}$ implies that $\{a_m\}_{m=0}^{\infty}$ is also a nonincreasing sequence.

Then we discuss these two cases separately.

Case 1: $a_0:=\|f\|^2-\|f-h\|^2\le0$. Then, for every $m\ge1$, we have $a_m\le0$. Therefore inequality (22) holds true.

Case 2: $a_0>0$. Assume that $a_{m-1}>0$, $m\ge1$. Note that f_m is the orthogonal projection of f onto the linear space spanned by $\hat{f}_m$, it implies

$$\langle f-f_m,f_m\rangle=0,\quad m\ge0. \tag{25}$$

This together with the selection of φ_m implies

$$\begin{aligned}e_{m-1}^2&=\langle f-f_{m-1},f-f_{m-1}\rangle\\&=\langle f-f_{m-1},f\rangle\\&=\langle f-f_{m-1},f-h\rangle+\langle f-f_{m-1},h\rangle\\&\le e_{m-1}\|f-h\|+\sum_{\varphi\in\mathcal{D}}c_\varphi\langle f-f_{m-1},\varphi\rangle\end{aligned}$$

$$\leq e_{m-1}\|f-h\| + t_m^{-1}\big|\langle f - f_{m-1}, \varphi_m\rangle\big| \sum_{\varphi\in\mathcal{D}} |c_\varphi|. \tag{26}$$

By (23), we get

$$e_{m-1}^2 \leq \frac{1}{2}\big(e_{m-1}^2 + \|f-h\|^2\big) + t_m^{-1}\big|\langle f - f_{m-1}, \varphi_m\rangle\big|(M+\epsilon). \tag{27}$$

Let $\epsilon \to 0$. Therefore

$$\big|\langle f - f_{m-1}, \varphi_m\rangle\big| \geq \frac{t_m(e_{m-1}^2 - \|f-h\|^2)}{2M}. \tag{28}$$

It has been proved in [24] that

$$e_m^2 \leq e_{m-1}^2 - \langle f - f_{m-1}, \varphi_m\rangle^2, \quad m = 1, 2, \ldots. \tag{29}$$

Then, using the assumption that $a_{m-1} > 0$, we have

$$e_m^2 \leq e_{m-1}^2 - \frac{t_m^2 a_{m-1}^2}{4M^2}. \tag{30}$$

It yields

$$a_m \leq a_{m-1}\left(1 - \frac{t_m^2 a_{m-1}}{4M^2}\right). \tag{31}$$

In particular, for $m = 1$, we have

$$a_1 \leq a_0\left(1 - \frac{t_1^2 a_0}{4M^2}\right). \tag{32}$$

Case 2.1: $0 < a_0 < \frac{4M^2}{t_1^2}$. Since $\psi(t) := t(1 - \frac{t_1^2 t}{4M^2})$ on $(0, \frac{4M^2}{t_1^2})$ has maximum $\frac{M^2}{t_1^2}$, it follows that

$$a_m \leq \psi(a_0) \leq \frac{M^2}{t_1^2} \leq \frac{4M^2}{t_1^2}.$$

Therefore, either all $\{a_m\}_{m=0}^\infty \subset (0, \frac{4M^2}{t_1^2})$ and then satisfy (31), or we know that $a_{m^*} \leq 0$ for some $m^* \geq 1$. The analysis for $m \geq m^*$ is therefore the same as in Case 1. For the positive elements in $\{a_m\}_{m=0}^\infty$, by applying Lemma 2.2 from [24] with $l = 1$, $r_m = t_m^2$, $B = \frac{4M^2}{t_1^2}$, $J = 0$, and $r = 4M^2$, we obtain

$$a_m \leq \frac{4M^2}{t_1^2 + \sum_{k=1}^m t_k^2} \leq \frac{4M^2}{\sum_{k=1}^m t_k^2}, \tag{33}$$

which gives inequality (22).

Case 2.2: $a_0 \geq \frac{4M^2}{t_1^2}$. It follows from (32) that $a_1 < 0$. That is, $e_1^2 < \|f-h\|^2$, which yields (22) due to monotonicity. Lemma 4.1 is proved. □

Now we prove Theorem 1.

Proof of Theorem 1 As shown in [5], $\|f_{\mathbf{z}} - f_\rho\|^2$ can be decomposed as

$$\begin{aligned}\|f_{\mathbf{z}} - f_\rho\|^2 \le{}& \mathcal{S}_1 + \mathcal{S}_2 + \mathcal{S}_3 \\ &+ 2\left(\|y - f_{\mathbf{z}}\|_n^2 + \kappa\frac{k^*\log n}{n} - \|y - Tf_{\mathbf{z},k}\|_n^2 - \kappa\frac{k\log n}{n}\right),\end{aligned} \tag{34}$$

where

$$\begin{aligned}\mathcal{S}_1 &:= \|f_{\mathbf{z}} - f_\rho\|^2 - 2\left(\|y - f_{\mathbf{z}}\|_n^2 - \|y - f_\rho\|_n^2 + \kappa\frac{k^*\log n}{n}\right), \\ \mathcal{S}_2 &:= 2\left(\|y - f_{\mathbf{z},k}\|_n^2 - \|y - h\|_n^2\right), \\ \mathcal{S}_3 &:= 2\left(\|y - h\|_n^2 - \|y - f_\rho\|_n^2 + \kappa\frac{k\log n}{n}\right),\end{aligned} \tag{35}$$

and $h \in \operatorname{Span}\{\mathcal{D}_m\}$.

We firstly estimate the bound of $\mathcal{S}_1$. To do this, we introduce Ω,

$$\Omega = \left\{\mathbf{z} : \mathbf{z} \in Z^n, \|f_{\mathbf{z}} - f_\rho\|^2 \ge 2\left(\|y - f_{\mathbf{z}}\|_n^2 - \|y - f_\rho\|_n^2 + \kappa\frac{k^*\log n}{n}\right)\right\}. \tag{36}$$

Let $\operatorname{Prob}(\Omega)$ be the probability that the sample point is a member of the set Ω. Then from $|y| \le B$ and the definition of f_ρ and $f_{\mathbf{z}}$, we have

$$E(\mathcal{S}_1) \le 6B^2 \operatorname{Prob}(\Omega). \tag{37}$$

For $\mathcal{S}_2$, according to Lemma 4.1, we get

$$\|y - f_{\mathbf{z},k}\|_n^2 - \|y - h\|_n^2 \le 4\frac{\|h\|_{\mathcal{A}_1^n}^2}{\sum_{k=1}^m t_k^2}, \tag{38}$$

where

$$\mathcal{A}_1^n(\mathcal{D}) := \left\{h : h = \sum_{i\in\Lambda} c_i^n \|g_i\|_n \frac{g_i}{\|g_i\|_n}, h \in \mathcal{A}_1(\mathcal{D})\right\} \tag{39}$$

and

$$\|h\|_{\mathcal{A}_1^n(\mathcal{D})} := \inf_h\left\{\sum_{i\in\Lambda}\left|c_i^n\right| \cdot \|g_i\|_n, h \in \mathcal{A}_1^n(\mathcal{D})\right\}. \tag{40}$$

It has been proved in Lemma 3.4 of [7] that

$$E\left(\|h\|_{\mathcal{A}_1^n}^2\right) \le \|h\|_{\mathcal{A}_1}^2, \tag{41}$$

which implies

$$E(\mathcal{S}_2) \le 8\frac{\|h\|_{\mathcal{A}_1}^2}{\sum_{k=1}^m t_k^2}. \tag{42}$$

For $\mathcal{S}_3$, from the property of mathematical expectation and (1), we have

$$E\big(\|y-h\|_n^2-\|y-f_\rho\|_n^2\big)=E\big(|y-h(x)|^2\big)-E\big(|y-f_\rho(x)|^2\big)$$
$$=\mathcal{E}(h)-\mathcal{E}(f_\rho). \tag{43}$$

This together with (3) yields

$$E(\mathcal{S}_3)=2\|f_\rho-h\|^2+2\kappa\frac{k\log n}{n}. \tag{44}$$

Combining (37), (42), with (44), we obtain

$$E\big(\|f_{\mathbf{z}}-f_\rho\|^2\big)\le 6B^2\,\mathrm{Prob}(\Omega)+8\frac{\|h\|_{\mathcal{A}_1}^2}{\sum_{k=1}^m t_k^2}+2\|f_\rho-h\|^2+2\kappa\frac{k\log n}{n}. \tag{45}$$

Next we bound Prob(Ω). To this end, we need the following known result in [10].

Lemma 4.2 *Let $\mathcal{F}$ be the class of functions $\mathcal{F}=\{|f|\le B\}$ for some fixed constant B. For all n and $\alpha,\beta>0$, we have*

$$\mathrm{Prob}\big\{\exists f\in\mathcal{F}:\|f-f_\rho\|_{\rho_X}^2\ge 2\big(\|y-f\|_n^2-\|y-f_\rho\|_n^2\big)+\alpha+\beta\big\}$$
$$\le 14\sup_{\boldsymbol{x}}\mathcal{N}\left(\frac{\beta}{40B},\mathcal{F},L_1(\vec{\nu}_{\boldsymbol{x}})\right)\exp\left(-\frac{\alpha n}{2568B^4}\right), \tag{46}$$

where $\boldsymbol{x}=(x_1,\dots,x_n)\in X^n$ and $\mathcal{N}(t,\mathcal{F},L_1(\vec{\nu}_{\boldsymbol{x}}))$ is the covering number for the class $\mathcal{F}$ by balls of radius t in $L_1(\vec{\nu}_{\boldsymbol{x}})$, with $\vec{\nu}_{\boldsymbol{x}}:=\frac{1}{n}\sum_{i=1}^n\delta_{x_i}$ the empirical discrete measure.

We define $\mathcal{G}_\Lambda:=\mathrm{Span}\{g:g\in\Lambda\subset\mathcal{D}\}$ and $\mathcal{F}_k:=\bigcup_{\Lambda\subset\mathcal{D}_m,\#(\Lambda)\le k}\{Tf:f\in\mathcal{G}_\Lambda\}$. Consider the probability

$$p_k=\mathrm{Prob}\left\{\exists f\in\mathcal{F}_k:\|f-f_\rho\|^2\ge 2\left(\|y-f\|_n^2-\|y-f_\rho\|_n^2+\kappa\frac{k\log n}{n}\right)\right\}.$$

Applying Lemma 4.2 to $\mathcal{F}_k$ with $\alpha=\kappa\frac{k\log n}{n}$, $\beta=\frac{1}{n}$, and $\kappa>1$, we get

$$p_k\le 14\sup_{\mathbf{x}}\mathcal{N}\left(\frac{1}{40Bn},\mathcal{F}_k,L_1(\vec{\nu}_{\mathbf{x}})\right)\exp\left(-\kappa\frac{k\log n}{2568B^4}\right)$$
$$=14\sup_{\mathbf{x}}\mathcal{N}\left(\frac{1}{40Bn},\mathcal{F}_k,L_1(\vec{\nu}_{\mathbf{x}})\right)n^{-\frac{\kappa k}{2568B^4}}. \tag{47}$$

Lemma 3.3 of [7] provides the upper bound for $\mathcal{N}(t,\mathcal{F}_k,L_1(\vec{\nu}_{\mathbf{x}}))$, which implies

$$p_k\le Cn^{ak}n^{2(k+1)}n^{-\frac{\kappa k}{2568B^4}}. \tag{48}$$

Let $\kappa\ge\kappa_0=2568B^4(a+5)$. Then the above inequality yields

$$p_k\le Cn^{-3k+2}\le Cn^{-2}. \tag{49}$$

So we have

$$\operatorname{Prob}(\Omega) \leq \sum_{1\leq k\leq \frac{Bn}{\kappa}} p_k \leq \frac{C}{n}. \tag{50}$$

By substituting the bound (50) of $\operatorname{Prob}(\Omega)$ into (45), we get

$$E\big(\|f_{\mathbf{z}} - f_\rho\|^2\big) \leq 8\frac{\|h\|^2_{\mathcal{A}_1}}{\sum_{k=1}^m t_k^2} + 2\|f_\rho - h\|^2 + C\frac{k\log n}{n}. \tag{51}$$

Next we derive the K-functional result of the upper bound (51). It is known from the property of variance that

$$E^2\big(\|f_{\mathbf{z}} - f_\rho\|\big) \leq E\big(\|f_{\mathbf{z}} - f_\rho\|\big)^2. \tag{52}$$

Combining (51) with (52), we have

$$\begin{aligned} E(\|f_{\mathbf{z}} - f_\rho\|) &\leq \sqrt{8\frac{\|h\|^2_{\mathcal{A}_1}}{\sum_{k=1}^m t_k^2} + 2\|f_\rho - h\|^2 + C\frac{k\log n}{n}} \\ &\leq 2\left(\frac{2\|h\|_{\mathcal{A}_1}}{(\sum_{k=1}^m t_k^2)^{1/2}} + \|f_\rho - h\|\right) + C\frac{k\log n}{n} \\ &\leq 2K\left(f_\rho, 2\left(\sum_{k=1}^m t_k^2\right)^{-1/2}\right) + C\frac{k\log n}{n}. \end{aligned} \tag{53}$$

This completes the proof of Theorem 1. □

5 Convergence rate and universal consistency

In this section, we analyze Theorem 2 under two different prior assumptions on f_ρ. We begin with the definitions of $\mathcal{A}_1(\mathcal{D}_m)$, $\mathcal{A}_{1,r}$, and $\mathcal{B}_{p,r}$.

We define the space $\mathcal{A}_1(\mathcal{D}_m)$ to be the space $\operatorname{Span}\{\mathcal{D}_m\}$ with the norm $\|\cdot\|_{\mathcal{A}_1(\mathcal{D}_m)}$ defined by (15). Note that now $\mathcal{D}$ is replaced by $\mathcal{D}_m$.

For $r > 0$, we then introduce the space

$$\mathcal{A}_{1,r} = \big\{f : \forall m, \exists h = h(m) \in \operatorname{Span}\{\mathcal{D}_m\}, \|h\|_{\mathcal{A}_1(\mathcal{D}_m)} \leq C, \|f - h\| \leq Cm^{-r}\big\}, \tag{54}$$

where $\|\cdot\|_{\mathcal{A}_{1,r}}$ is the minimum value of C such that (54) holds.

Furthermore, we present the following space:

$$\mathcal{B}_{p,r} := [\mathcal{H}, \mathcal{A}_{1,r}]_{\theta,\infty}, \quad 0 < \theta < 1, \tag{55}$$

with $\frac{1}{p} = \frac{1+\theta}{2}$. From the definition of interpolation spaces in [26], we know that $f \in [\mathcal{H}, \mathcal{A}_{1,r}]_{\theta,\infty}$ if and only if for any $t > 0$,

$$K(f, t, \mathcal{H}, \mathcal{A}_{1,r}) := \inf_{h\in\mathcal{A}_{1,r}} \big\{\|f - h\|_{\mathcal{H}} + t\|h\|_{\mathcal{A}_{1,r}}\big\} \leq Ct^\theta. \tag{56}$$

The minimum C such that (56) holds true is defined as the norm on $\mathcal{B}_{p,r}$.

Now we first consider $f_\rho \in \mathcal{A}_{1,r}$.

Corollary 5.1 *Under the assumptions of Theorem 2, if $f_\rho \in \mathcal{A}_{1,r}$ with $r > \frac{1}{2a}$, then we have*

$$E\big(\|f_{\mathbf{z}} - f_\rho\|^2\big) \le C\big(1 + \|f_\rho\|_{\mathcal{A}_{1,r}}\big)t_0^{-1}\left(\frac{n}{\log n}\right)^{-\frac{1}{2}}. \tag{57}$$

Proof From the definition of $\mathcal{A}_{1,r}$, there exists $h := h(m) \in \operatorname{Span}\{\mathcal{D}_m\}$ for every m that satisfies

$$\|h\|_{\mathcal{A}_1(\mathcal{D}_m)} \le M$$

and

$$\|f_\rho - h\| \le Mm^{-r},$$

where $M := \|f_\rho\|_{\mathcal{A}_{1,r}}$.

Theorem 2 thus implies

$$E\big(\|f_{\mathbf{z}} - f_\rho\|^2\big) \le C\min_{k>0}\left(\frac{M^2}{kt_0^2} + M^2n^{-2ar} + \frac{k\log n}{n}\right). \tag{58}$$

Moreover, the mild restriction $2ar \ge 1$ with a arbitrarily large allows us to remove the term M^2n^{-2ar} in (58). To balance the errors in (58), we take $k := \lceil \frac{(M+1)^2}{t_0^2}\frac{n}{\log n}\rceil^{\frac{1}{2}}$. Then the desired result (57) can be obtained. □

Next we consider $f_\rho \in \mathcal{B}_{p,r}$.

Corollary 5.2 *Under the assumptions of Theorem 2, if $f_\rho \in \mathcal{B}_{p,r}$ with $r > \frac{1}{2a}$, then we have*

$$E\big(\|f_{\mathbf{z}} - f_\rho\|^2\big) \le Ct_0^{-p}\big(1 + \|f_\rho\|_{\mathcal{B}_{p,r}}\big)^p\left(\frac{n}{\log n}\right)^{-1+\frac{p}{2}}. \tag{59}$$

Proof By (56), if $f \in \mathcal{B}_{p,r}$, then for any $t > 0$, we can find a function $\tilde{f} \in \mathcal{A}_{1,r}$ that satisfies

$$\|\tilde{f}\|_{\mathcal{A}_{1,r}} \le \|f\|_{\mathcal{B}_{p,r}}t^{\theta-1} \tag{60}$$

and

$$\|f - \tilde{f}\| \le \|f\|_{\mathcal{B}_{p,r}}t^{\theta}. \tag{61}$$

For $\tilde{f} \in \mathcal{A}_{1,r}$, according to (54), there exists $h := h(m) \in \operatorname{Span}\{\mathcal{D}_m\}$ for every m that satisfies

$$\|h\|_{\mathcal{A}_1(\mathcal{D}_m)} \le \|\tilde{f}\|_{\mathcal{A}_{1,r}} \tag{62}$$

and

$$\|\tilde{f} - h\| \le \|\tilde{f}\|_{\mathcal{A}_{1,r}}m^{-r}. \tag{63}$$

The relations (60), (62), and (63) imply

$$\|h\|_{\mathcal{A}_1(\mathcal{D}_m)} \le \|f\|_{\mathcal{B}_{p,r}} t^{\theta-1} \tag{64}$$

and

$$\|\tilde{f} - h\| \le \|f\|_{\mathcal{B}_{p,r}} t^{\theta-1} m^{-r}. \tag{65}$$

Then combining (61) with (65), we obtain

$$\|f - h\| \le \|f\|_{\mathcal{B}_{p,r}} \left(t^{\theta} + t^{\theta-1} m^{-r}\right). \tag{66}$$

From (64) and (66), there exists $h := h(m) \in \operatorname{Span}\{\mathcal{D}_m\}$ for every m and $t > 0$ that satisfies

$$\|h\|_{\mathcal{A}_1(\mathcal{D}_m)} \le M t^{\theta-1}$$

and

$$\|f_\rho - h\| \le M\left(t^{\theta} + t^{\theta-1} m^{-r}\right),$$

where $M = \|f_\rho\|_{\mathcal{B}_{p,r}}$.

Therefore, Theorem 2 with $t = k^{-\frac{1}{2}}$ implies

$$E\left(\|f_{\mathbf{z}} - f_\rho\|^2\right) \le C \min_{k>0} \left(M^2 t_0^{-2} k^{1-\frac{2}{p}} + M^2 \left(k^{\frac{1}{2}-\frac{1}{p}} + k^{1-\frac{1}{p}} n^{-ar}\right)^2 + \frac{k \log n}{n} \right). \tag{67}$$

The condition $2ar \ge 1$ also enables us to eliminate the term involving n^{-ar}. Then, by taking $k := \lceil \frac{(M+1)^2}{t_0^2} \frac{n}{\log n} \rceil^{\frac{p}{2}}$ in (67), we obtain the desired result (59). □

Then we show the universal consistency of the WRPGLA.

Theorem 3 *Under the assumptions of Theorem 2, if the dictionary $\mathcal{D}$ is complete in $L^2_{\rho_X}(X)$, for any f_ρ, we have*

$$\lim_{n \to +\infty} E\left(\|f_{\mathbf{z}} - f_\rho\|^2\right) = 0. \tag{68}$$

Proof Since $\mathcal{D}$ is complete in $L^2_{\rho_X}(X)$, we can find $h \in \operatorname{Span}\{\mathcal{D}_m\}$ satisfying $\|f_\rho - h\| \le \varepsilon$, where $\varepsilon > 0$ and n is big enough. It follows from Theorem 2 that

$$E\left(\|f_{\mathbf{z}} - f_\rho\|^2\right) \le C \min_{k>0} \left(\frac{\|h\|^2_{\mathcal{A}_1(\mathcal{D}_m)}}{k t_0^2} + \varepsilon^2 + \frac{k \log n}{n} \right). \tag{69}$$

To balance the first and third error term, we choose $k := n^{\frac{1}{2}} t_0^{-1}$, which implies

$$E\left(\|f_{\mathbf{z}} - f_\rho\|^2\right) \le C\left(\varepsilon^2 + t_0^{-1} n^{-\frac{1}{2}} \log n\right). \tag{70}$$

Thus, for n sufficiently large,

$$E\left(\|f_{\mathbf{z}} - f_\rho\|^2\right) \le 2C\varepsilon^2. \tag{71}$$

This completes the proof of Theorem 3. □

Remark 3 It is known from [11] that the OGA and the RGA can achieve the optimal convergence rate $\mathcal{O}(m^{-\frac{1}{2}})$ on $\mathcal{A}_1(\mathcal{D})$. When $t_k = 1$, Lemma 4.1 shows that the WRPGA also attains the best rate. Meanwhile, we compare the WRPGLA with the OGLA and the relaxed greedy learning algorithm (RGLA). For $f_\rho \in \mathcal{A}_{1,r}$, we derive the same convergence rate $\mathcal{O}((n\log n)^{-1/2})$ of the WRPGLA as that of the OGLA and the RGLA in Ref. [7]. For $f_\rho \in \mathcal{B}_{p,r}$, when $p \to 1$, the rate $\mathcal{O}((n\log n)^{-1+\frac{p}{2}})$ of the WRPGLA can be arbitrarily close to $\mathcal{O}((n\log n)^{-1/2})$.

Moreover, from the viewpoint of the computational complexity, for the WRPGLA, the approximant f_k is constructed by solving a one-dimensional optimization problem since f_k is an orthogonal projection of f onto $\mathrm{Span}\{\hat{f}_k\}$. On the other hand, the OGLA is more expensive to implement since at each step, the algorithm requires the evaluation of orthogonal projection on a k-dimensional space, and the output is constructed by solving a k-dimensional optimization problem. And it is clear that the WRPGLA is simpler than the RGLA. Thus, the WRPGLA should essentially reduce the complexity and make the learning process accelerated.

In future research, it would be an interesting project to deduce the error bound of the WRPGLA in Banach spaces with modulus of smoothness $\rho(u) \le \gamma u^q$, $1 < q \le 2$ as [24, 27]. Furthermore, Guo and Ye [28, 29] derived the convergence rates of the moving least-squares learning algorithm for the weakly dependent and nonidentical samples. It remains open to explore the greedy learning algorithms in the non-i.i.d. and nonidentical sampling setting.

Funding
This research is supported by the National Science Foundation for Young Scientists of China (Grant No. 12001328), the National Natural Science Foundation of China (Grant No. 11671213), the Development Plan of Youth Innovation Team of University in Shandong Province (No. 2021KJ067), the Shandong Provincial Natural Science Foundation of China (No. ZR2022MF223), and the Qilu University of Technology (Shandong Academy of Sciences) Talent Research Project (No. 2023RCKY133).

Data availability
All data, models, and code generated or used during the study appear in the submitted article.

Declarations

Competing interests
The authors declare no competing interests.

Author contributions
All authors contributed substantially to this paper, participated in drafting and checking the manuscript. All authors read and approved the final manuscript.

Author details
[1]School of Science, Shandong Jianzhu University, Jinan, 250101, China. [2]School of Mathematical Sciences and LPMC, Nankai University, Tianjin, 300071, China.

Received: 2 September 2023 Accepted: 2 January 2024 Published online: 04 March 2024

References

1. Fang, J., Lin, S.B., Xu, Z.B.: Learning and approximation capabilities of orthogonal super greedy algorithm. Knowl.-Based Syst. **95**, 86–98 (2016)
2. Chen, H., Li, L.Q., Pan, Z.B.: Learning rates of multi-kernel regression by orthogonal greedy algorithm. J. Stat. Plan. Inference **143**, 276–282 (2013)
3. Lin, S.B., Rong, Y.H., Sun, X.P., Xu, Z.B.: Learning capability of relaxed greedy algorithms. IEEE Trans. Neural Netw. Learn. Syst. **24**(10), 1598–1608 (2013)
4. Xu, L., Lin, S.B., Xu, Z.B.: Learning capability of the truncated greedy algorithm. Sci. China Inf. Sci. **59**(5), 052103 (2016). https://doi.org/10.1007/s11432-016-5536-6
5. Barron, A.R.: Universal approximation bounds for superposition of a sigmoidal function. IEEE Trans. Inf. Theory **39**(3), 930–945 (1993)
6. Guo, Q.: Distributed semi-supervised regression learning with coefficient regularization. Results Math. **77**, 1–19 (2022)
7. Barron, A.R., Cohen, A., Dahmen, W., DeVore, R.A.: Approximation and learning by greedy algorithms. Ann. Stat. **36**(1), 64–94 (2008)
8. Xu, L., Lin, S.B., Zeng, J.S., Liu, X., Xu, Z.B.: Greedy criterion in orthogonal greedy learning. IEEE Trans. Cybern. **48**(3), 955–966 (2018)
9. Jones, L.K.: A simple lemma on greedy approximation in Hilbert spaces and convergence rates for projection pursuit regression and neural network training. Ann. Stat. **20**(1), 608–613 (1992)
10. Lee, W.S., Bartlett, P.L., Williamson, R.C.: Efficient agnostic learning of neural networks with bounded fan-in. IEEE Trans. Inf. Theory **42**(6), 2118–2132 (1996)
11. DeVore, R.A., Temlyakov, V.N.: Some remarks on greedy algorithms. Adv. Comput. Math. **5**, 173–187 (1996)
12. Temlyakov, V.N.: Greedy Approximation. Cambridge University Press, Cambridge (2011)
13. Dai, W., Milenkovic, O.: Subspace pursuit for compressive sensing signal recontruction. IEEE Trans. Inf. Theory **55**(5), 2230–2249 (2009)
14. Kunis, S., Rauhut, H.: Random sampling of sparse trigonometric polynomials ii-orthogonal matching pursuit versus basis pursuit. Found. Comput. Math. **8**, 737–763 (2008)
15. Donoho, D.L., Tsaig, Y., Drori, I., Starck, J.L.: Sparse solution of underdetermined systems of linear equations by stagewise orthogonal matching pursuit. IEEE Trans. Inf. Theory **58**(2), 1094–1121 (2012)
16. Tropp, J.A., Wright, S.: Computational methods for sparse solution of linear inverse problems. Proc. IEEE **98**(6), 948–958 (2010)
17. Temlyakov, V.N., Zheltov, P.: On performance of greedy algorithms. J. Approx. Theory **163**(9), 1134–1145 (2011)
18. Donoho, D.L., Elad, M., Temlyakov, V.N.: On Lebesgue-type inequalities for greedy approximation. J. Approx. Theory **147**(2), 185–195 (2007)
19. Chen, H., Zhou, Y.C., Tang, Y.Y., Li, L.Q., Pan, Z.B.: Convergence rate of the semi-supervised greedy algorithm. Neural Netw. **44**, 44–50 (2013)
20. Schmidt, E.: Zur theorie der linearen und nicht linearen integralgleichungen zweite abhandlung. Math. Ann. **64**, 161–174 (1907)
21. Temlyakov, V.N.: Greedy approximation. Acta Numer. **17**, 235–409 (2008)
22. Livshitz, E.D., Temlyakov, V.N.: Two lower estimates in greedy approximation. Constr. Approx. **19**, 509–523 (2003)
23. Livshits, E.D.: Lower bounds for the rate of convergence of greedy algorithms. Izv. Math. **73**, 1197–1215 (2009)
24. Petrova, G.: Rescaled pure greedy algorithm for Hilbert and Banach spaces. Appl. Comput. Harmon. Anal. **41**, 852–866 (2016)
25. Jiang, B., Ye, P.X.: Efficiency of the weak rescaled pure greedy algorithm. Int. J. Wavelets Multiresolut. Inf. Process. **19**(4), 2150001 (2021)
26. Bergh, J., Lofstrom, J.: Interpolation Spaces. Springer, Berlin (1976)
27. Temlyakov, V.N.: Greedy algorithms in Banach spaces. Adv. Comput. Math. **14**, 277–292 (2001)
28. Guo, Q., Ye, P.X.: Error analysis of the moving least-squares method with non-identical sampling. Int. J. Comput. Math. **96**(4), 767–781 (2019)
29. Guo, Q., Ye, P.X.: Error analysis of the moving least-squares regression learning algorithm with β-mixing and non-identical sampling. Int. J. Comput. Math. **97**(8), 1586–1602 (2020)

Publisher's Note