

Full length article

# Metric entropy and sparse linear approximation of $\ell_q$ -hulls for $0 < q \leq 1$

 Fuchang Gao<sup>a</sup>, Ching-Kang Ing<sup>b</sup>, Yuhong Yang<sup>c,\*</sup>
<sup>a</sup> *University of Idaho, United States*
<sup>b</sup> *Academia Sinica and National Taiwan University, Taiwan*
<sup>c</sup> *University of Minnesota, United States*

Received 4 May 2012; received in revised form 16 September 2012; accepted 3 October 2012

Available online 26 October 2012

Communicated by Allan Pinkus

---

## Abstract

Consider  $\ell_q$ -hulls,  $0 < q \leq 1$ , from a dictionary of  $M$  functions in  $L^p$  space for  $1 \leq p < \infty$ . Their precise metric entropy orders are derived. Sparse linear approximation bounds are obtained to characterize the number of terms needed to achieve accurate approximation of the best function in a  $\ell_q$ -hull that is closest to a target function. Furthermore, in the special case of  $p = 2$ , it is shown that a weak orthogonal greedy algorithm achieves the optimal approximation under an additional condition.

© 2012 Elsevier Inc. All rights reserved.

---

In recent years, sparse linear combinations of given functions (or variables) have played important roles in statistical learning theories and methodologies that deal with a large number of predictors (often more than the number of observations). Let  $F = \{f_1, \dots, f_M\}$  be a collection of  $M$  functions defined on a measurable space taking values in  $\mathcal{X}$  with a  $\sigma$ -finite dominating measure  $\mu$ . For any  $\theta = (\theta_1, \dots, \theta_M)' \in \mathbb{R}^M$ , define the  $\ell_0$ -norm and the  $\ell_q$ -norm ( $0 < q \leq 1$ ) by

$$\|\theta\|_0 = \sum_{j=1}^M I(\theta_j \neq 0), \quad \text{and} \quad \|\theta\|_q = \left( \sum_{j=1}^M |\theta_j|^q \right)^{1/q},$$

---

\* Corresponding author.

E-mail address: [yyang@stat.umn.edu](mailto:yyang@stat.umn.edu) (Y. Yang).

where  $I(\cdot)$  is the indicator function (note that for  $0 \leq q < 1$ ,  $\|\cdot\|_q$  is not a real norm). Define the  $\ell_{q,t}$ -hull of  $F$  to be the class of linear combinations of functions in  $F$  with the  $\ell_q$ -constraint

$$\mathcal{F}_q(t) = \mathcal{F}_q(t; M; F) = \left\{ f_\theta = \sum_{j=1}^M \theta_j f_j : \|\theta\|_q \leq t, f_j \in F \right\}, \quad 0 \leq q \leq 1, t > 0.$$

In statistical learning theories, the functions in  $F$  are either some initial estimates or observable variables. Much of the current statistical research interest on function estimation focuses on the case of a large dictionary  $F$  (often with a small or moderate number of observations of pairs of response and values of the  $f_j$ 's with noise in response). To this goal of efficient estimation of the regression function (the conditional expectation of the response given the predictors), an understanding of sparse representation or approximation of the functions in  $\mathcal{F}_q(t)$  is essential. Note that for  $0 < q \leq 1$ , which is the focused case in this paper, a bound on the  $\ell_q$ -norm implies that there can be a small number of coefficients that are relatively large. Consequently the function classes  $\mathcal{F}_q(t)$  can have good sparse linear approximations. One notable feature here is that no restrictive assumptions on the relationships between functions in the dictionary  $F$  are necessary for our upper bound results.

The  $\ell_q$ -ball of  $\mathbb{R}^M$  (often denoted  $\ell_q^M$ ) for  $q > 0$  is well studied, with the metric entropy order and Gelfand widths understood. Recently, [13] have derived the metric entropy order for  $\mathcal{F}_q(t)$  with  $0 < q \leq 1$  (see [11] for earlier but less precise results) and further showed that any function in  $\mathcal{F}_q(t)$  can be well approximated by linear combinations of a relatively few terms in  $F$ . Their results deal only with the  $L^2$  norm on the function classes. In this work, we complete the result for a general  $L^p$  ( $p \geq 1$ ).

The rest of the paper is organized as follows. In Section 1, the metric entropy order of the  $\ell_q$ -hull ( $0 < q \leq 1$ ) under the  $L^p$  norm is determined. In Section 2, for any target function in the  $\ell_q$ -hull, the order of the best linear approximation error in the  $L^p$  norm using only a sparse number of terms is obtained. In Section 3, in the special case of  $p = 2$  and under an additional condition, it is shown that a greedy approximation achieves the optimal sparse linear approximation. An implication on recovery of sparse vectors is also given. The results in the different sections complement each other: Theorem 1 characterizes the massiveness of the  $\ell_q$ -hull ( $0 < q \leq 1$ ); Theorem 2 addresses the capability of sparse linear approximation of functions in the  $\ell_q$ -hull; and Theorem 3 deals with a practically efficient term-after-term approximation of the same functions in a way that does not require searching over a large combinatorial number of terms.

## 1. Metric entropy of $\ell_q$ -hull under the $L^p$ norm

Throughout the paper, let  $r = \min(2, p)$ .

**Theorem 1.** Suppose  $F = \{f_1, f_2, \dots, f_M\}$  with  $\|f_j\|_{L^p(v)} \leq 1$ ,  $1 \leq j \leq M$  for some  $p \geq 1$ , where  $v$  is a  $\sigma$ -finite measure. For  $0 < q \leq 1$ , there exists a positive constant  $c_{p,q}$  depending only on  $q$  and  $p$  such that for any  $0 < \varepsilon < 1$ ,  $\mathcal{F}_q(1)$  contains an  $\varepsilon$ -net  $\{e_j\}_{j=1}^{N_\varepsilon}$  in the  $L^p$  norm for  $j = 1, 2, \dots, N_\varepsilon$ , where  $N_\varepsilon$  satisfies

$$\log N_\varepsilon \leq \begin{cases} c_{p,q} \varepsilon^{-\frac{rq}{r-q}} \log \left( 1 + M^{\frac{1}{q} - \frac{1}{r}} \varepsilon \right) & \text{if } \varepsilon > M^{\frac{1}{r} - \frac{1}{q}}, \\ c_{p,q} M \log \left( 1 + M^{\frac{1}{r} - \frac{1}{q}} \varepsilon^{-1} \right) & \text{if } \varepsilon \leq M^{\frac{1}{r} - \frac{1}{q}}. \end{cases} \quad (1)$$

Furthermore, the estimates are best possible up to a constant.

**Remark.** Metric entropy plays a central role in determining how well functions in a class can be estimated based on contaminated observations; see e.g. [14] for references.

**Proof.** From classical results (cf. [5, p. 98]), for any positive integer  $k$ , the unit ball of  $\ell_q^M$  can be covered by  $2^{k-1}$  balls of radius  $\varepsilon_k$  in the  $\ell_1$  norm, where

$$\varepsilon_k \leq c \begin{cases} 1 & 1 \leq k < \log_2(2M) \\ \left( \frac{\log_2 \left( 1 + \frac{2M}{k} \right)}{k} \right)^{\frac{1}{q}-1} & \log_2(2M) \leq k \leq 2M \\ 2^{-\frac{k}{2M}} (2M)^{1-\frac{1}{q}} & k > 2M, \end{cases}$$

which is also known to be sharp [9]. Thus, we can have  $2^{k-1}$  functions  $g_j$ ,  $1 \leq j \leq 2^{k-1}$ , such that

$$\mathcal{F}_q(1) \subset \bigcup_{j=1}^{2^{k-1}} (g_j + \mathcal{F}_1(\varepsilon_k)). \quad (2)$$

For any  $g \in \mathcal{F}_1(\varepsilon_k)$ ,  $g$  can be expressed as  $g = \sum_{i=1}^M c_i f_i$  with  $\sum_{i=1}^M |c_i| \leq \varepsilon_k$ . Define a random function  $U$  by

$$\mathbb{P}(U = \text{sign}(c_i) \varepsilon_k f_i) = |c_i| / \varepsilon_k, \quad \mathbb{P}(U = 0) = 1 - \sum_{i=1}^M |c_i| / \varepsilon_k.$$

Then we have  $\|U\|_p \leq \varepsilon_k$  a.s. and  $\mathbb{E}U = g$  under the randomness just introduced. Let  $U_1, \dots, U_m, U'_1, \dots, U'_m$ , be i.i.d. copies of  $U$ , and let  $V = \frac{1}{m} \sum_{i=1}^m U_i$ . We have

$$\begin{aligned} \mathbb{E}\|V - g\|_p &= \mathbb{E} \left\| \mathbb{E}' \left( \frac{1}{m} \sum_{i=1}^m U'_i \right) - \frac{1}{m} \sum_{i=1}^m U_i \right\|_p \\ &\leq \mathbb{E} \mathbb{E}' \left\| \frac{1}{m} \sum_{i=1}^m (U_i - U'_i) \right\|_p \\ &= \frac{1}{m} \mathbb{E} \mathbb{E}' \left\| \sum_{i=1}^m (U_i - U'_i) \right\|_p \\ &= \frac{1}{m} \mathbb{E} \mathbb{E}' \left\| \sum_{i=1}^m \xi_i (U_i - U'_i) \right\|_p \\ &\leq \frac{2}{m} \mathbb{E} \left\| \sum_{i=1}^m \xi_i U_i \right\|_p, \end{aligned}$$

where  $\xi_i$  are i.i.d. Rademacher variables and are independent of all  $U_j$  and  $U'_j$ . Taking expectation with respect to  $\xi_i$ , and using the fact that  $L^p$  space is of type  $r = \min(p, 2)$  (see e.g., Section 9.2 of [10]), we have

$$\mathbb{E}\|V - g\|_p \leq \frac{2T_p}{m} \mathbb{E} \left( \sum_{i=1}^m \|U_i\|_p^r \right)^{1/r} \leq 2T_p m^{\frac{1}{r}-1} \varepsilon_k,$$

where  $T_p$  is the type- $\min(p, 2)$  constant of  $L^p$ . Therefore, there must exist a realization of  $V$  such that  $\|V - g\|_p \leq 2T_p m^{\frac{1}{r}-1} \varepsilon_k$ . Clearly,  $V$  can be expressed as  $\varepsilon_k m^{-1}(k_1 f_1 + k_2 f_2 + \cdots + k_M f_M)$ , where  $k_1, k_2, \dots, k_M$  are integers, and  $|k_1| + |k_2| + \cdots + |k_M| \leq m$ . The total number of different realizations of  $V$  is therefore bounded above by  $\binom{2M+m}{m}$ . Together with (2), we conclude that  $\mathcal{F}_q(1)$  can be covered by  $2^{k-1} \binom{2M+m}{m}$  balls of radius  $2T_p m^{\frac{1}{r}-1} \varepsilon_k$ .

We have the freedom to choose integers  $m$  and  $k$ . We choose  $m$  according to  $k$  as follows.

(i) If  $\log_2(2M) \leq k \leq 2M$ , we choose  $m$  to be the largest integer such that  $\binom{2M+m}{m} \leq 2^k$ . Then we have

$$\frac{1}{m} \leq \frac{c'}{k} \log_2 \left( 1 + \frac{2M}{k} \right)$$

for some positive absolute constant  $c'$ . Thus,  $\mathcal{F}_q(1)$  can be covered by  $2^{2k-1}$  balls of radius

$$2T_p m^{\frac{1}{r}-1} \varepsilon_k \leq c_2 \left( k^{-1} \log_2 \left( 1 + \frac{2M}{k} \right) \right)^{1/q-1/r}$$

in the  $L^p$  norm, where  $c_2$  depends only on  $p$ .

(ii) If  $k > 2M$ , we choose  $m = M$ . Then  $\mathcal{F}_q(1)$  can be covered by  $2^{k-1} \binom{3M}{M}$  balls of radius

$$2T_p m^{\frac{1}{r}-1} \varepsilon_k \leq c_3 2^{-\frac{k}{2M}} M^{\frac{1}{r}-\frac{1}{q}}$$

in the  $L^p$  norm, where  $c_3$  depends only on  $p$ .

Now, we finish the proof of (1). If  $M^{\frac{1}{r}-\frac{1}{q}} \leq \varepsilon < 1$ , we choose  $k$  to be the smallest integer such that  $k \geq \log_2(2M)$ , and

$$\begin{cases} c_2 \left( k^{-1} \log_2 \left( 1 + \frac{2M}{k} \right) \right)^{1/q-1/r} \leq \varepsilon, \\ c_3 2^{-\frac{k}{2M}} M^{\frac{1}{r}-\frac{1}{q}} \leq \varepsilon. \end{cases}$$

It is not difficult to see that  $k \leq CM$  for some constant  $C$  depending only on  $p$ . Indeed, the set of inequalities above is equivalent to the following set

$$\begin{cases} c_2 \left( \frac{M}{k} \log_2 \left( 1 + \frac{2M}{k} \right) \right)^{1/q-1/r} \leq M^{\frac{1}{q}-\frac{1}{r}} \varepsilon, \\ c_3 2^{-\frac{k}{2M}} \leq M^{\frac{1}{q}-\frac{1}{r}} \varepsilon. \end{cases}$$

The right-hand sides of the two inequalities are not less than 1, while the left-hand sides go to 0 as  $\frac{M}{k} \rightarrow 0$ . Hence, the smallest integer  $k$  satisfying both inequalities is bounded by  $CM$  for some constant  $C$  depending only on  $p$ .

If  $k < 2M$ , then by (i),  $\mathcal{F}_q(1)$  can be covered by  $2^{2k-1}$  balls of radius  $\varepsilon$  in the  $L^p$  norm. Note that for such  $k$ , we have

$$2^{2k-1} \leq \exp \left( c_4 \varepsilon^{-\frac{rq}{r-q}} \log \left( 1 + M^{\frac{1}{q}-\frac{1}{r}} \varepsilon \right) \right),$$

where  $c_4$  is a constant depending only on  $p$ . If  $2M \leq k < CM$ , then by (ii)  $\mathcal{F}_q(1)$  can be covered by  $2^{k-1} \binom{3M}{M}$  balls of radius  $\varepsilon$  in the  $L^p$  norm. Note that for such  $k$ , we have

$$2^{k-1} \binom{3M}{M} \leq 2^{k-1+3M} \leq 2^{(C+3)M-1} \leq \exp\left(c_5 \varepsilon^{-\frac{rq}{r-q}} \log\left(1 + M^{\frac{1}{q}-\frac{1}{r}} \varepsilon\right)\right),$$

where  $c_5$  is a constant depending only on  $p$ . In either case, the first inequality in (1) follows.

If  $\varepsilon < M^{\frac{1}{r}-\frac{1}{q}}$ , we choose  $k$  to be the smallest integer such that  $k \geq 2M$ , and  $c_3 2^{-\frac{k}{2M}} M^{\frac{1}{r}-\frac{1}{q}} \leq \varepsilon$ . It is easy to check that

$$2^k \leq 2\left(c_3 M^{\frac{1}{r}-\frac{1}{q}} \varepsilon^{-1}\right)^{2M}.$$

By (ii),  $\mathcal{F}_q(1)$  can be covered by  $2^{k-1} \binom{3M}{M}$  balls of radius  $\varepsilon$  in the  $L^p$  norm. Note that for such  $k$ , we have

$$2^{k-1} \binom{3M}{M} \leq 2^{k-1+3M} \leq 2^{3k} \leq \exp\left(c_6 M \log\left(1 + M^{\frac{1}{r}-\frac{1}{q}} \varepsilon^{-1}\right)\right),$$

where  $c_6$  is a constant depending only on  $p$ . This proves the second inequality in (1).

Finally, we show that the estimates in the theorem are best possible up to a constant. We construct functions  $\{f_1, f_2, \dots, f_M\} \subset L^p[0, 1]$ , so that the reverse inequality holds. We need to consider the cases  $1 \leq p \leq 2$  and  $2 < p < \infty$  separately.

When  $1 \leq p \leq 2$ , we choose  $f_j(x) = M^{1/p}$  if  $x \in [(j-1)/M, j/M)$ , and  $f_j = 0$  otherwise. Thus,  $\|f\|_p = 1$ . Because

$$\sum_{j=1}^M |c_j|^q \leq \left(\sum_{j=1}^M |c_j|^p\right)^{q/p} M^{1-q/p},$$

we have

$$\mathcal{F}_q(1) \supset \mathcal{H} := \left\{ \sum_{j=1}^M c_j f_j : \left(\sum_{j=1}^M |c_j|^p\right)^{1/p} \leq M^{\frac{1}{p}-\frac{1}{q}} \right\}.$$

But

$$\left\| \sum_{j=1}^M c_j f_j \right\|_p = \left(\sum_{j=1}^M |c_j|^p\right)^{1/p},$$

and  $\mathcal{H}$  is isometric to the closed ball of  $l_p^M$  with radius  $M^{\frac{1}{p}-\frac{1}{q}}$ . It is known that when  $\varepsilon < M^{\frac{1}{p}-\frac{1}{q}}$  the metric entropy of the latter has a lower bound that is, up to a constant, of the same order as the upper bound given in the theorem. Hence the metric entropy estimate is sharp when  $1 \leq p \leq 2$ , and  $\varepsilon < M^{\frac{1}{p}-\frac{1}{q}}$ .

For  $\varepsilon \geq M^{\frac{1}{p}-\frac{1}{q}}$ , we let  $\delta = 5^{1/(p-q)} \varepsilon^{p/(p-q)}$ . Then  $d := \lfloor \delta^{-q} \rfloor \leq M/10$ . Note that

$$\mathcal{F}_q(1) \supset \mathcal{G} := \left\{ \sum_{j \in I} \delta f_j : I \subset \{1, 2, \dots, M\}, |I| = d \right\},$$

and  $\mathcal{G}$  contains  $\binom{M}{d}$  functions. For each  $g = \sum_{j \in I} \delta f_j \in \mathcal{G}$ , we define

$$\mathcal{N}(g, 9d/10) = \left\{ h = \sum_{j \in J} \delta f_j \in \mathcal{G} : |I \cap J| \leq 9d/10 \right\}.$$

Then  $\mathcal{N}(g, 9d/10)$  contains not more than

$$\binom{d}{\lfloor 9d/10 \rfloor} \binom{M}{\lceil d/10 \rceil} \leq \sqrt{\binom{M}{d}}$$

functions. Therefore, we can find at least  $\sqrt{\binom{M}{d}}$  disjoint such sets. Note that if  $g = \sum_{j \in I} \delta f_j$  and  $g' = \sum_{j \in I'} \delta f_j$  are chosen from different sets, then  $|I \Delta I'| \geq d/5$ . Thus,  $\|g - g'\|_p = |I \Delta I'|^{1/p} \delta \geq (d/5)^{1/p} \delta \geq \varepsilon$ . Therefore, the  $\varepsilon$  covering number of  $\mathcal{F}_q(1)$  is at least

$$\sqrt{\binom{M}{d}} \geq \exp\left(c\varepsilon^{-\frac{rq}{r-q}} \log\left(1 + M^{\frac{1}{q} - \frac{1}{r}}\right)\right).$$

When  $2 < p < \infty$ , we define  $f_j(t) = \text{sgn}(\sin(2^j \pi t))$ ,  $1 \leq j \leq M$ . It is clear that  $\|f\|_p = 1$ . Because

$$\sum_{j=1}^M |c_j|^q \leq \left( \sum_{j=1}^M |c_j|^2 \right)^{q/2} M^{1-q/2},$$

we have

$$\mathcal{F}_q(1) \supset \mathcal{U} := \left\{ \sum_{j=1}^M c_j f_j : \sum_{j=1}^M |c_j|^2 \leq M^{1-2/q} \right\}.$$

We define a linear operator  $P$  from  $\mathcal{F}_q(1)$  to  $l_2^M$ , so that

$$P\left(\sum_{j=1}^M c_j f_j\right) = (c_1, c_2, \dots, c_M).$$

Let  $B$  be the closed ball of  $l_2^M$  with radius  $M^{1-2/q}$ . Then  $P^{-1}B \subset \mathcal{F}_q(1)$ . By our construction of  $f_j$ , for any  $g \in P^{-1}B$ ,

$$\|g\|_p = \left( \int_0^1 \left| \sum_{j=1}^M c_j f_j(t) \right|^p dt \right)^{1/p} = \left( \int_0^1 \left| \sum_{j=1}^M \xi_j c_j f_j(t) \right|^p dt \right)^{1/p},$$

where  $\xi_j$  are independent Rademacher random variables. Taking expectation, and using the fact that  $|f_j| = 1$ , we obtain

$$\|g\|_p \geq \left( \sum_{j=1}^M c_j^2 \right)^{1/2} = \|P^{-1}g\|_{l_2^M}.$$

Thus, restricted on  $P^{-1}B$ ,  $P$  is a contraction map. Hence the metric entropy of  $\mathcal{F}_q(1)$  under  $\|\cdot\|_p$  is bounded below by the metric entropy of  $B$  under the  $\|\cdot\|_{l_1^M}$  norm. If  $\varepsilon < M^{\frac{1}{2} - \frac{1}{q}}$ , the latter

has a lower bound which is, up to a constant, the same as the upper bound given in the theorem. Hence, the estimate is best possible when  $2 < p < \infty$  and  $\varepsilon < M^{\frac{1}{2}-\frac{1}{q}}$ .

If  $\varepsilon \geq M^{\frac{1}{2}-\frac{1}{q}}$ , then by the same argument as for the case  $1 \leq p \leq 2$ , we can find at least  $\sqrt{\binom{M}{d}}$  such sets that are disjoint. Note that if  $g = \sum_{j \in I} \delta f_j$  and  $g' = \sum_{j \in I'} \delta f_j$  are chosen from different sets, then  $|I \Delta I'| \geq d/5$ . Note that  $\|g - g'\|_p \geq \|Pg - Pg'\|_{l_2^M} \geq \delta |I \Delta I'|^{1/2} \geq (d/5)^{1/2} \delta \geq \varepsilon$ . Therefore, the  $\varepsilon$  covering number of  $\mathcal{F}_q(1)$  is at least

$$\sqrt{\binom{M}{d}} \geq \exp \left( c \varepsilon^{-\frac{2q}{2-q}} \log \left( 1 + M^{\frac{1}{q}-\frac{1}{2}} \right) \right).$$

This completes the proof of the theorem.  $\square$

## 2. Sparse linear approximation bounds

Under the assumption that the functions in the dictionary  $F$  have finite  $L^p$  norms, we give below an upper bound on the approximation error by the best linear combination of a given number of members in the dictionary.

**Theorem 2.** Let  $f_0$  be any function with  $\|f_0\|_p < \infty$  for some  $p \geq 1$ . Suppose  $F = \{f_1, \dots, f_M\}$  with  $\max_{1 \leq j \leq M} \|f_j\|_p < \infty$  for  $1 \leq j \leq M$ . For any  $1 \leq m \leq M$ ,  $0 < q \leq 1$ ,  $t > 0$ , there exist a subset  $J_m$  of  $\{1, \dots, M\}$  of cardinality  $m$  and  $f_{\theta^m} \in \mathcal{F}_{J_m} = \text{span of } f_j \text{ in } J_m$  with  $\|\theta^m\|_1 \leq t$  such that

$$\|f_0 - f_{\theta^m}\|_p \leq \|f_0 - f_{\theta^*}\|_p + 2T_p \max_{1 \leq j \leq M} \|f_j\|_p \cdot t m^{\frac{1}{r}-\frac{1}{q}},$$

where  $f_{\theta^*} = \arg \min_{f_{\theta} \in \mathcal{F}_q(t)} \|f_0 - f_{\theta}\|_p$ ,  $r = \min(p, 2)$ , and  $T_p$  depending only on  $p$  is the type- $\min(p, 2)$  constant of  $L^p$ . If  $\mu$  is a probability measure, then for any  $1 \leq m \leq M$ ,  $0 < q \leq 1$ ,  $t > 0$ ,  $1 \leq p' \leq p$ , there exist a subset  $J'_m$  and  $\tilde{f}_{\theta^m} \in \mathcal{F}_{J'_m} = \text{span of } f_j \text{ in } J'_m$  with  $\|\tilde{\theta}^m\|_1 \leq t$  such that

$$\|f_0 - \tilde{f}_{\theta^m}\|_{p'} \leq \|f_0 - \tilde{f}_{\theta^*}\|_{p'} + 2T_p \max_{1 \leq j \leq M} \|f_j\|_p \cdot t m^{\frac{1}{r}-\frac{1}{q}}$$

where  $\tilde{f}_{\theta^*} = \arg \min_{f_{\theta} \in \mathcal{F}_q(t)} \|f_0 - f_{\theta}\|_{p'}$ . Furthermore, the estimates are best possible up to a constant.

**Remarks.** 1. If  $M = \infty$ , the approximation bounds in the theorem continue to hold for each  $m \geq 1$  with the obvious change of  $\max_{1 \leq j \leq M} \|f_j\|_p < \infty$  to  $\max_{1 \leq j < \infty} \|f_j\|_p < \infty$ . This is seen from the proof of the theorem.

2. A currently very active research in statistics and machine learning is on learning when the number of predictors is huge relative to the number of observations. A particular setting is high-dimensional linear modeling under the so called soft sparsity assumption, i.e., the  $\ell_q^M$  norm of the coefficients of the best (or a good) linear approximation of the target function  $f_0$  by the functions in the dictionary  $F$  is small for some  $q \leq 1$ . This corresponds to when  $M$  is very large and one can only afford a sparse model with  $m \ll M$ . The theorem characterizes the capability of sparse models in approximation, which together with statistical estimation theory can lead to a precise understanding on potential gain of sparse linear modeling (see, e.g., [13]).

**Proof.** We first prove the first result. Without loss of generality, assume  $\max_{1 \leq j \leq M} \|f_j\|_p \leq 1$  (otherwise consider  $f'_j = f_j / \max_{1 \leq j \leq M} \|f_j\|_p$  and observe that  $\left\{ \sum_{j=1}^M \theta_j f_j : \|\theta\|_q \leq t, f_j \in F \right\} \subset \left\{ \sum_{j=1}^M \theta'_j f'_j : \|\theta'\|_q \leq t \max_{1 \leq j \leq M} \|f_j\|_p, f_j \in F \right\}$ ). Let  $f_{\theta^*} = \sum_{j=1}^M c_j f_j = \arg \inf_{f_{\theta} \in \mathcal{F}_q(t)} \|f_{\theta} - f_0\|_p$ . For any  $1 \leq m \leq M$ , let  $L^* = \{j : |c_j| > tm^{-1/q}\}$ . Because  $\sum_{j=1}^M |c_j|^q \leq t^q$ , we have

$$|L^*| t^q m^{-1} < \sum |c_j|^q \leq t^q.$$

Thus  $|L^*| < m$  (so there are not too many large coefficients). Also

$$\begin{aligned} \sum_{j \notin L^*} |c_j| &= \sum_{j \notin L^*} |c_j|^q |c_j|^{1-q} \\ &\leq \sum_{j \notin L^*} |c_j|^q \left( t m^{-\frac{1}{q}} \right)^{1-q} \\ &= \sum_{j \notin L^*} |c_j|^q t^{1-q} m^{1-\frac{1}{q}} \\ &\leq t m^{1-\frac{1}{q}} \equiv D. \end{aligned}$$

Define  $v^* = \sum_{j \in L^*} c_j f_j$  and  $\omega^* = \sum_{j \notin L^*} c_j f_j$ . Clearly,  $\omega^* \in \mathcal{F}_1(D)$ . Define a random function  $U$  so that

$$P(U = D \operatorname{sign}(c_j) f_j) = \frac{|c_j|}{D} \quad \text{for } j \notin L^*$$

and

$$P(U = 0) = 1 - \sum_{j \notin L^*} \frac{|c_j|}{D}.$$

Note that  $E(U) = \omega^*$  and  $\|U\|_p \leq D \max_{1 \leq j \leq M} \|f_j\|_p \leq D$ .

Let  $U_1, \dots, U_m, U'_1, \dots, U'_m$ , be i.i.d. copy of  $U$ . Then,

$$E \left\| f_0 - v^* - \frac{1}{m} \sum_{i=1}^m U_i \right\|_p \leq \|f_0 - f_{\theta^*}\|_p + E \left\| \omega^* - \frac{1}{m} \sum_{i=1}^m U_i \right\|_p.$$

Now, for  $p \geq 1$ ,

$$\begin{aligned} E \left\| \omega^* - \frac{1}{m} \sum_{i=1}^m U_i \right\|_p &= E \left\| E' \left( \frac{1}{m} \sum_{i=1}^m U'_i \right) - \frac{1}{m} \sum_{i=1}^m U_i \right\|_p \\ &\leq E E' \left\| \frac{1}{m} \sum_{i=1}^m (U_i - U'_i) \right\|_p \\ &= \frac{1}{m} E E' \left\| \sum_{i=1}^m (U_i - U'_i) \right\|_p \\ &= \frac{1}{m} E E' \left\| \sum_{i=1}^m \xi_i (U_i - U'_i) \right\|_p \end{aligned}$$



$$\leq \frac{2}{m} \mathbb{E} \left\| \sum_{i=1}^m \xi_i U_i \right\|_p$$

where  $\xi_i$  are i.i.d. Bernoulli random variables, independent of all  $U_j$  and  $U'_j$ . Taking expectation with respect to  $\xi_i$ , and using the fact that the  $L^p$  space is of type  $r = \min(p, 2)$ , we have

$$\mathbb{E} \left\| \omega^* - \frac{1}{m} \sum_{i=1}^m U_i \right\|_p \leq \frac{2T_p}{m} \mathbb{E} \left( \sum_{i=1}^m \|U_i\|_p^r \right)^{1/r} \leq 2T_p m^{\frac{1}{r}-1} D = 2T_p t m^{\frac{1}{r}-\frac{1}{q}}.$$

So there exists a realization of  $U_1, \dots, U_m$ , such that

$$\left\| f_0 - v^* - \frac{1}{m} \sum_{i=1}^m u_i \right\|_p \leq \|f_0 - f_{\theta^*}\|_p + 2T_p t m^{\frac{1}{r}-\frac{1}{q}}.$$

Note that  $\left\| v^* + \frac{1}{m} \sum_{i=1}^m u_i \right\|_0 \leq (m-1) + m = 2m-1$ . Consider  $\tilde{m} = \lfloor (m+1)/2 \rfloor$ . Then  $2\tilde{m}-1 \leq m$  and  $\tilde{m} \geq m/2$ . The first result of [Theorem 2](#) then follows from above.

Now for the second result, from the earlier derivation, with  $\tilde{v}^*$ ,  $\tilde{\omega}^*$ ,  $\tilde{U}_i$  and  $\tilde{\theta}^*$  defined under  $L^{p'}$ , we have

$$\begin{aligned} \mathbb{E} \left\| f_0 - \tilde{v}^* - \frac{1}{m} \sum_{i=1}^m \tilde{U}_i \right\|_{p'} &\leq \|f_0 - f_{\tilde{\theta}^*}\|_{p'} + \mathbb{E} \left\| \tilde{\omega}^* - \frac{1}{m} \sum_{i=1}^m \tilde{U}_i \right\|_{p'} \\ &\leq \|f_0 - f_{\tilde{\theta}^*}\|_{p'} + \mathbb{E} \left\| \tilde{\omega}^* - \frac{1}{m} \sum_{i=1}^m \tilde{U}_i \right\|_p. \end{aligned}$$

The second sparse approximation error bound in the theorem then follows similarly. Also, the examples of the functions constructed in the proof of [Theorem 1](#) prove the sharpness of the estimates. This completes the proof of the theorem.  $\square$

### 3. A greedy approximation for $p = 2$

Greedy approximation, among nonlinear approximations, has received an increasing attention in research and application as a means of providing accurate and computationally fast approximation. See [\[12\]](#) for convergence results for various versions of greedy approximation and earlier works, and some recent statistical applications of orthogonal greedy approximations are in e.g., [\[1,7\]](#), and [\[8\]](#). In particular, Theorem 3 of [\[12\]](#) shows that an orthogonal greedy algorithm leads to  $m$ -term approximation error of order  $m^{-1/2}$  for functions in the  $\ell_1$ -hull of a dictionary in a Hilbert space.

In the case of  $p = 2$ , for a general  $0 < q \leq 1$ , we show below that, under an extra condition on the functions in the dictionary, the convergence rate  $m^{\frac{1}{r}-\frac{1}{q}} = m^{\frac{1}{2}-\frac{1}{q}}$  in the first expression of [Theorem 2](#) is attainable by an  $m$ -term approximation of  $f_{\theta^*}$  based on the weak orthogonal greedy algorithm (WOGA) defined below. An implication on recovering a hard or strong sparse function is given later.

#### 3.1. Optimal approximation by WOGA

*Weak Orthogonal Greedy Algorithm.* Let  $0 < \xi \leq 1$  and  $f$  be any function with  $\|f\|_2 < \infty$ . Define  $f_0^{o,\xi} := f$ . Then for each  $m \geq 1$ , we inductively define the following.

(1)  $\varphi_m^{o,\xi} \in F$  is any choice that satisfies

$$|\langle f_{m-1}^{o,\xi}, \varphi_m^{o,\xi} / \|\varphi_m^{o,\xi}\|_2 \rangle| \geq \xi \sup_{g \in F} |\langle f_{m-1}^{o,\xi}, g / \|g\|_2 \rangle|,$$

where  $\langle \cdot, \cdot \rangle$  denotes the inner product in  $L^2$ .

(2)  $G_m^{o,\xi}(f, F) := P_{H_m^\xi}(f)$ , where  $H_m^\xi = \text{span}(\varphi_1^{o,\xi}, \dots, \varphi_m^{o,\xi})$  and  $P_{H_m^\xi}$  is the orthogonal projector on  $H_m^\xi$ .

(3)  $f_m^{o,\xi} := f - G_m^{o,\xi}(f, F)$ .

**Theorem 3.** Let  $f_0$  be any function with  $\|f_0\|_2 < \infty$ . Suppose  $F = \{f_1, \dots, f_M\}$  with  $\max_{1 \leq j \leq M} \|f_j\|_2 < \infty$ , and  $\lambda := \lambda_{\min}(\Sigma) > 0$ , where  $\lambda_{\min}(A)$  denotes the minimum eigenvalue of a symmetric matrix  $A$  and  $\Sigma = (\gamma_{ij})_{1 \leq i, j \leq M}$  with  $\gamma_{ij} = \langle f_i / \|f_i\|_2, f_j / \|f_j\|_2 \rangle$ . Then, with  $f_{\theta^*}$  defined as in Theorem 2, for any  $m \geq 1$ ,  $0 < q \leq 1$ ,  $t > 0$ ,

$$\|f_0 - G_m^{o,\xi}(f_{\theta^*}, F)\|_2 \leq \|f_0 - f_{\theta^*}\|_2 + Sm^{\frac{1}{2} - \frac{1}{q}},$$

where  $S$  is some positive number which depends on  $\xi, t, q, \lambda, \max_{1 \leq j \leq M} \|f_j\|_2, \|f_{\theta^*}\|_2$  and whose precise form will be clear in the proof.

The minimum eigenvalue assumption in Theorem 3 seems to be commonly used in the contexts of time series and high-dimensional data analysis. A further discussion of this condition will also be given at the end of this section.

To prove Theorem 3, we need the following ancillary lemma, which is an extension of Lemma 3.4 of [4].

**Lemma 1.** Let  $\{a_n\}$  be a sequence of non-negative numbers. Suppose that there exist  $A, C > 0$  and  $0 < \alpha \leq 1$  such that

$$a_1 \leq A, \quad a_{n+1} \leq a_n(1 - Ca_n^\alpha), \quad n = 1, 2, \dots$$

Then, for each  $n \geq 1$ ,

$$a_n \leq Kn^{-\alpha^{-1}},$$

where  $K = K(C, \alpha, A) = \max\{2^{\alpha^{-2}}(C\alpha)^{-\alpha^{-1}}, A\}$ .

**Proof.** We proceed by induction on  $n$ . The statement being obvious when  $n = 1$ , suppose  $a_n \leq Kn^{-1/\alpha}$  for  $n = 1, \dots, N$ . We will show that  $a_{N+1} \leq K(N+1)^{-1/\alpha}$ . If  $a_{N+1} = 0$ , then the desired conclusion holds trivially. If  $a_{N+1} > 0$ , then  $a_i > 0$  for  $i = 1, \dots, N$ . By the assumption of the lemma and using the inductive hypothesis, it follows that for each  $i = 1, \dots, N$ ,

$$a_{i+1}^{-1} \geq a_i^{-1}(1 - Ca_i^\alpha)^{-1} \geq a_i^{-1}(1 + Ca_i^\alpha) \geq a_i^{-1} + CK^{\alpha-1}i^{-1+\alpha^{-1}}.$$

Summing both sides of the expression over  $1 \leq i \leq N$  yields

$$\begin{aligned} a_{N+1}^{-1} &\geq C\alpha K^{\alpha-1} \{N/(N+1)\}^{\alpha^{-1}} (N+1)^{\alpha^{-1}} \\ &\geq C\alpha \{2^{\alpha^{-2}}(C\alpha)^{-\alpha^{-1}}\}^\alpha K^{-1} 2^{-\alpha^{-1}} (N+1)^{\alpha^{-1}} \\ &= K^{-1} (N+1)^{\alpha^{-1}}. \end{aligned}$$

Therefore,  $a_{N+1} \leq K(N+1)^{-\alpha^{-1}}$ .  $\square$

**Proof of Theorem 3.** Let  $D = (t \max_{1 \leq j \leq M} \|f_j\|_2)^{q/(2-q)} \lambda^{(q-1)/(2-q)}$ . It suffices to show that

$$\mu_m := \|f_{\theta^*} - G_m^{o,\xi}(f_{\theta^*}, F)\|_2^2 \leq K(\xi^2 D^{-2}, q/(2-q), \|f_{\theta^*}\|_2^2)^m^{1-(2/q)},$$

where  $K$  is defined as in Lemma 1. Denote by  $J_{\xi,m}$  the set of positive integers consisting of the subscripts of the functions in  $F$  chosen by the WOGA (with  $f = f_{\theta^*}$ ) after  $m$  iterations. Then, it follows that

$$\begin{aligned} \mu_m &= \left\langle f_{\theta^*} - G_m^{o,\xi}(f_{\theta^*}, F), \sum_{1 \leq j \leq M, j \notin J_{\xi,m}} c_j f_j \right\rangle \\ &\leq \sup_{1 \leq j \leq M} |\langle f_j / \|f_j\|_2, f_{\theta^*} - G_m^{o,\xi}(f_{\theta^*}, F) \rangle| \left| \sum_{j=1, j \notin J_{\xi,m}}^M c_j \right| \|f_j\|_2, \end{aligned}$$

recalling that  $f_{\theta^*} = \sum_{j=1}^M c_j f_j$ . In addition, one has  $\mu_m \geq \lambda \sum_{j=1, j \notin J_{\xi,m}}^M c_j^2 \|f_j\|_2^2$ . By making use of these two inequalities, Hölder's inequality and  $(\sum_{j=1}^M |c_j|^q)^{1/q} \leq t$ , we know that  $\mu_m$  is bounded above by

$$\begin{aligned} &\sup_{1 \leq j \leq M} |\langle f_j / \|f_j\|_2, f_{\theta^*} - G_m^{o,\xi}(f_{\theta^*}, F) \rangle| \left( \sum_{1 \leq j \leq M, j \notin J_{\xi,m}} |c_j|^q \|f_j\|_2^q \right)^{\frac{1}{2-q}} \\ &\quad \times \left( \sum_{1 \leq j \leq M, j \notin J_{\xi,m}} c_j^2 \|f_j\|_2^2 \right)^{\frac{1-q}{2-q}} \\ &\leq \sup_{1 \leq j \leq M} |\langle f_j / \|f_j\|_2, f_{\theta^*} - G_m^{o,\xi}(f_{\theta^*}, F) \rangle| D \mu_m^{(1-q)/(2-q)}, \end{aligned}$$

which gives  $\mu_m^{1/(2-q)} \leq D \sup_{1 \leq j \leq M} |\langle f_j / \|f_j\|_2, f_{\theta^*} - G_m^{o,\xi}(f_{\theta^*}, F) \rangle|$ . As a result,

$$\begin{aligned} \mu_{m+1} &\leq \left\| f_{\theta^*} - G_m^{o,\xi}(f_{\theta^*}, F) - \langle f_{\theta^*} - G_m^{o,\xi}(f_{\theta^*}, F), \varphi_{m+1}^{o,\xi} \rangle \varphi_{m+1}^{o,\xi} / \|\varphi_{m+1}^{o,\xi}\|_2^2 \right\|_2^2 \\ &= \mu_m - |\langle f_{\theta^*} - G_m^{o,\xi}(f_{\theta^*}, F), \varphi_{m+1}^{o,\xi} / \|\varphi_{m+1}^{o,\xi}\|_2 \rangle|^2 \\ &\leq \mu_m - \xi^2 \sup_{1 \leq j \leq M} |\langle f_{\theta^*} - G_m^{o,\xi}(f_{\theta^*}, F), f_j / \|f_j\|_2 \rangle|^2 \\ &\leq \mu_m - \xi^2 D^{-2} \mu_m^{2/(2-q)} = \mu_m (1 - \xi^2 D^{-2} \mu_m^{q/(2-q)}). \end{aligned}$$

This, together with Lemma 1 and  $\mu_1 \leq \|f_{\theta^*}\|_2^2$ , gives the desired conclusion.  $\square$

### 3.2. An implication on recovering a hard sparse function

Let  $f_{\theta_0} \in \mathcal{F}_q(t)$  for some  $0 < q \leq 1$ , and  $\theta_0 = (c_1, \dots, c_M)^\top$  be an  $s$ -sparse vector, namely,  $1 \leq |J_{\theta_0}| \leq s \ll M$ , where  $J_{\theta_0} = \{i : 1 \leq i \leq M, c_i \neq 0\}$ . In the next corollary, making use of Theorem 3, we show that the index set determined by the WOGA, with  $f = f_{\theta_0}$  and  $0 < \xi \leq 1$ , includes  $J_{\theta_0}$ , provided a minimum eigenvalue condition is fulfilled and the number of iterations is reasonably large. To state the result, for  $J \subset \{1, \dots, M\}$ , define  $\Sigma_J = (\gamma_{i,j})_{i,j \in J}$ . In addition, let  $W_{q,\xi} = 2^{(2\beta^2)^{-1}} (\xi^2 \beta)^{-(2\beta)^{-1}}$  and  $B_q(x) = (\lfloor x / \min_{i \in J_{\theta_0}} |c_i| \rfloor + 1)^{2\beta}$ , where

$x \geq 0$ ,  $\beta = \beta(q) = q/(2-q)$  and  $\lfloor a \rfloor$  denotes the largest integer  $\leq a$ . Fix  $0 < \xi \leq 1$ , denote by  $J_{\xi,m}(\theta_0)$  the index set chosen by the WOGA after  $m$  iterations.

**Corollary 1.** Suppose that for some  $\bar{c} > \max \{W_{q,\xi} t \max_{1 \leq j \leq M} \|f_j\|_2, \|f_{\theta_0}\|_2\}$ , we have  $s + B_q(\bar{c}) + 1 \leq M$  and

$$\underline{\lambda}(\bar{c}) := \inf_{|J|=s+B_q(\bar{c})+1} \lambda_{\min}(\Sigma_J) > \max \left\{ \left( \frac{W_{q,\xi} t \max_{1 \leq j \leq M} \|f_j\|_2}{\bar{c}} \right)^{\frac{2q}{2-q}}, \left( \frac{\|f_{\theta_0}\|}{\bar{c}} \right)^2 \right\}.$$

Then for  $m \geq B_q(\bar{c}) + 1$ , we must have  $J_{\theta_0} \subset J_{\xi,m}(\theta_0)$ .

Suppose the smallest eigenvalue of  $\Sigma$  is bounded below by  $0 < \psi < 1$ . Then, the magnitude of  $B_q(\bar{c})$  in Corollary 1 depends on  $q$ ,  $t$  and  $\psi$ . As a simple example, consider the special case that the  $s$  nonzero elements of  $\theta_0$  are of the same order, and  $\max_{1 \leq j \leq M} \|f_j\|_2$  and  $\|f_{\theta_0}\|_2$  are bounded above by 1. Then for  $\bar{c} = \max\{W_{q,\xi} t \psi^{(q-2)/(2q)}, \psi^{-1/2}\}$ ,  $\underline{\lambda}(\bar{c})$  satisfies the above lower bound

condition, and hence  $B_q(\bar{c})$  is of order  $\frac{g(q)s^{\frac{2}{2-q}}}{\psi}$ , where  $g(q) \rightarrow \infty$  as  $q \rightarrow 0$ . With  $q$  close to 0, this is  $\frac{g(q)s^{1+\varpi}}{\psi}$  for some  $\varpi$  close to zero (but  $g(q)$  can be large). Thus, when  $s$  is small compared to  $M$ , with the number of iterations of a slightly higher order than  $s$ , the WOGA is guaranteed to include all the nonzero terms. From Corollary 1, it is clear that the minimum eigenvalue lower bound condition (bounded below by  $\psi$ ) is only needed on all  $\Sigma_J$  with  $|J| \leq Cs^{\frac{2}{2-q}}$  for some large enough constant  $C > 0$ .

**Proof.** Let  $\mu_m = \|f_{\theta_0} - G_m^{o,\xi}(f_{\theta_0}, F)\|_2^2$  if  $1 \leq m \leq B_q(\bar{c}) + 1$ , and 0 if  $m > B_q(\bar{c}) + 1$ . Since for  $1 \leq m \leq B_q(\bar{c})$ ,  $|J_{\theta_0} \cup J_{\xi,m}(\theta_0)| \leq s + B_q(\bar{c})$ , one obtains from the definition of  $\underline{\lambda}(\bar{c})$  and an argument similar to that used in the proof of Theorem 3 that for  $1 \leq m \leq B_q(\bar{c})$ ,

$$\begin{aligned} \mu_m &\leq \sup_{1 \leq j \leq M} |\langle f_j / \|f_j\|_2, f_{\theta_0} - G_m^{o,\xi}(f_{\theta_0}, F) \rangle| \sum_{j \in J_{\theta_0} \setminus J_{\xi,m}(\theta_0)} |c_j| \|f_j\|_2, \\ \mu_m &\geq \underline{\lambda}(\bar{c}) \sum_{j \in J_{\theta_0} \setminus J_{\xi,m}(\theta_0)} c_j^2 \|f_j\|_2^2, \end{aligned}$$

and

$$\mu_{m+1} \leq \mu_m \left( 1 - \xi^2 \underline{D}^{-2} \mu_m^\beta \right),$$

where  $\underline{D} = (t \max_{1 \leq j \leq M} \|f_j\|_2)^\beta \{\underline{\lambda}(\bar{c})\}^{(q-1)/(2-q)}$ . Therefore, by Lemma 1,

$$\mu_m \leq m^{-\beta^{-1}} \max \left\{ 2^{\beta-2} \left( \xi^2 \underline{D}^{-2} \beta \right)^{-\beta^{-1}}, \|f_{\theta_0}\|_2^2 \right\}, \quad m \geq 1.$$

Now, if  $J_{\theta_0} \not\subset J_{\xi,B_q(\bar{c})+1}(\theta_0)$ , then

$$\underline{\lambda}^{1/2}(\bar{c}) \min_{i \in J_{\theta_0}} |c_i| \leq \mu_{B_q(\bar{c})+1}^{1/2} \leq (B_q(\bar{c}) + 1)^{(q-2)/(2q)} \max\{W_{q,\xi} \underline{D}^{(2-q)/q}, \|f_{\theta_0}\|_2\},$$

yielding  $\underline{\lambda}^{1/2}(\bar{c}) \leq \max\{W_{q,\xi} t \max_{1 \leq j \leq M} \|f_j\|_2 \underline{\lambda}^{1-\frac{1}{q}}(\bar{c})/\bar{c}, \|f_{\theta_0}\|_2/\bar{c}\}$ , which contradicts the hypothesis on  $\underline{\lambda}(\bar{c})$  of the corollary. This completes the proof.  $\square$

**Remark.** Under the assumptions of Corollary 1, it is possible for  $J_{\xi,m}(\theta_0)$  to include some indices  $i$  whose corresponding coefficients  $c_i$  are zero. However, one can still exactly recover  $\theta_0$  through  $\theta^{(m)} = (c_i^{(m)})_{1 \leq i \leq M}$ , where  $c_i^{(m)} = 0$  if  $i \notin J_{\xi,m}(\theta_0)$  and  $(c_i^{(m)})_{i \in J_{\xi,m}(\theta_0)} = \Sigma_{J_{\xi,m}(\theta_0)}^{-1} E(\mathbf{f}_{J_{\xi,m}(\theta_0)} f_{\theta_0})$ , with  $\mathbf{f}_{J_{\xi,m}(\theta_0)} = (f_i)_{i \in J_{\xi,m}(\theta_0)}$  being a  $|J_{\xi,m}(\theta_0)|$ -dimensional vector. The relevance of this to real applications of high-dimensional regression is that when  $B_q(\bar{c}) + 1$  is much smaller than the sample size, then both  $E(\mathbf{f}_{J_{\xi,m}(\theta_0)} f_{\theta_0})$  and  $\Sigma_{J_{\xi,m}(\theta_0)}$  may be accurately estimated based on data for  $m$  slightly larger than  $B_q(\bar{c}) + 1$ . Therefore, assuming  $J_{\theta_0}$  is sparse relative to the sample size, after a suitable number of data-driven iterations of WOGA and a subsequent statistical determination of the zero-coefficient terms as suggested above, the true set of predictors  $J_{\theta_0}$  can be obtained with a high probability.

We close this section by noting that the minimum eigenvalue condition used in Corollary 1 is related to the restricted isometry property (RIP), introduced in [2] and defined for the  $n \times M$  matrix  $A = (a_{i,j})_{1 \leq i \leq n, 1 \leq j \leq M}$  in the linear system  $A\theta_0 = \mathbf{y}$ . More specifically, RIP of order  $k \geq 1$  requires that there exists the smallest constant  $0 \leq \delta_k < 1$  for which  $1 - \delta_k \leq \inf_{|J|=k} \lambda_{\min}(\widehat{\Sigma}_J) \leq \sup_{|J|=k} \lambda_{\max}(\widehat{\Sigma}_J) \leq 1 + \delta_k$ , where  $\widehat{\Sigma}_J = A_J^\top A_J / n$ ,  $A_J = (a_{i,j})_{1 \leq i \leq n, j \in J}$  and  $\lambda_{\max}(A)$  denotes the maximum eigenvalue of a symmetric matrix  $A$ . When  $(a_{i,1}, \dots, a_{i,M})^\top$ ,  $1 \leq i \leq n$ , are independent and identically distributed copies of  $(f_1, \dots, f_M)^\top$ ,  $\widehat{\Sigma}_J$  can be viewed as a “sample version” of  $\Sigma_J$  and the difference between the two matrices is negligible uniformly over all  $|J| \ll n$  as  $n$  is sufficiently large (see [8] for more details). It is shown in [3] that when  $\delta_{3s} + 3\delta_{4s} < 2$ , any  $s$ -sparse vector  $\theta_0$  is exactly recovered via solving the  $l_1$ -minimization problem associated with the linear system mentioned above. Under a more flexible assumption,  $r_{2k} - 1 < 4(\sqrt{2} - 1)(k/s)^{(1/q)-(1/2)}$  for some  $0 < q \leq 1$  and some  $k \geq s$ , where  $r_{2k} = \sup_{|J|=2k} \lambda_{\max}(\widehat{\Sigma}_J) / \inf_{|J|=2k} \lambda_{\min}(\widehat{\Sigma}_J)$ , [6] further showed that the same exact recovery result is achieved through solving the corresponding  $l_q$ -minimization problem.

While the minimum eigenvalue assumption described in Corollary 1 is not necessarily weaker than those in [3,6], the maximum eigenvalue assumption is dropped in this corollary, thereby substantially expanding its applicability in particular to cases where the correlations between  $f_i$ 's are large. To see this, assume that for any  $J \subseteq \{1, \dots, M\}$ ,  $\Sigma_J$  satisfies  $\gamma_{f_{ij}} = 1$  if  $i = j$ , and  $\gamma_{ij} = \rho$  for some  $0 < \rho < 1$  if  $i \neq j$ . It is straightforward to show that  $\lambda_{\min}(\Sigma_J) = 1 - \rho$ , and hence by Corollary 1 and the remark given after it,  $J_{\theta_0} \subset J_{\xi,m}(\theta_0)$  and  $\theta_0$  is exactly recovered by  $\theta^{(m)}$  if  $m \geq B_q(\bar{c}) + 1$  with

$$\bar{c} > \max \left\{ W_{q,\xi,t} \max_{1 \leq j \leq M} \|f_j\|_2 (1 - \rho)^{-\frac{2-q}{2q}}, \|f_{\theta_0}\| (1 - \rho)^{-1/2} \right\}.$$

In contrast, if  $\widehat{\Sigma}_J$  is the same as  $\Sigma_J$ , then  $\lambda_{\max}(\widehat{\Sigma}_J) = 1 + (|J| - 1)\rho$  for each  $J$ , which implies that unless  $\rho$  is very small,  $\delta_{3s} + 3\delta_{4s} < 2$  can fail to hold even for a moderate value of  $s$ , say  $s = 5$ . This simple example illustrates a typical challenge that  $l_1$ -minimization faces while pursuing exact recovery in highly correlated dictionaries. Alternatively, since for any  $0 < \rho < 1$ , there exists a sufficiently small  $q$  such that  $r_{2s+2} - 1 = 2s\rho(1 - \rho)^{-1} < 4(\sqrt{2} - 1)(1 + s^{-1})^{(1/q)-(1/2)}$ , the above difficulty encountered by  $l_1$ -minimization is alleviated by  $l_q$ -minimization, provided  $q$  is sufficiently small. However, when  $0 < q < 1$ , this  $l_q$ -minimization problem is nonconvex, and therefore, very difficult to solve globally. For an approximate sparse solution to this problem, see Section 4 of [6]. Finally, we point out an inevitable difficulty with WOGA (besides the obvious challenge with  $\rho$  being close to 1) that when  $\min_{i \in J_{\theta_0}} |c_i|$  is close to 0, WOGA may require a large number of iterations to include all the non-zero terms, as disclosed by Corollary 1.

## Acknowledgments

The authors thank Chiao-Yi Yang and Wei-Ying Wu for helpful discussions. The work of Fuchang Gao was supported in part by a grant from the Simons Foundation (#246211). The work of Ching-Kang Ing was supported in part by the Academia Sinica Investigator Award. The work of Yuhong Yang was supported by funding from NSC of Taiwan, the University of Minnesota and NSF of USA. He thanks the Institute of Statistical Science at the Academia Sinica for their hospitality during his visit. Comments and suggestions by three referees are greatly appreciated for improving the paper.

## References

- [1] A.R. Barron, A. Cohen, W. Dahmen, R.A. DeVore, Approximation and learning by greedy algorithms, *The Annals of Statistics* 36 (2008) 64–94.
- [2] E.J. Candes, T. Tao, Decoding by linear programming, *IEEE Transactions on Information Theory* 51 (2005) 4203–4215.
- [3] E.J. Candes, T. Tao, Near-optimal signal recovery from random projections: universal encoding strategies, *IEEE Transactions on Information Theory* 52 (2006) 5406–5425.
- [4] R.A. DeVore, V.N. Temlyakov, Some remarks on greedy algorithms, *Advances in Computational Mathematics* 5 (1996) 173–187.
- [5] D.E. Edmunds, H. Triebel, Function Spaces, Entropy Numbers, and Differential Operators, in: *Cambridge Tracts in Mathematics*, vol. 120, Cambridge University Press, 1998.
- [6] S. Foucart, M.-J. Lai, Sparsest solutions of underdetermined linear systems via  $l_q$ -minimization for  $0 < q \leq 1$ , *Applied and Computational Harmonic Analysis* 26 (2009) 395–407.
- [7] C. Huang, G.H.L. Cheang, A.R. Barron, Risk of penalized least squares, greedy selection and  $L_1$ -penalization for flexible function libraries, *Manuscript*, 2008.
- [8] C.-K. Ing, T.L. Lai, A stepwise regression method and consistent model selection for high-dimensional sparse linear models, *Statistica Sinica* 21 (2011) 1473–1513.
- [9] T. Kühn, A lower estimate for entropy numbers, *Journal of Approximation Theory* 110 (2001) 120–124.
- [10] M. Ledoux, M. Talagrand, *Probability in Banach Spaces: Isoperimetry and Processes*, Springer, New York, 1991.
- [11] G. Raskutti, M. Wainwright, B. Yu, Minimax rates of estimation for high-dimensional linear regression over  $\ell_q$ -balls, *IEEE Transactions on Information Theory* 57 (2011) 6976–6994.
- [12] V.N. Temlyakov, Weak greedy algorithms, *Advances in Computational Mathematics* 12 (2000) 213–227.
- [13] Z. Wang, S. Paterlini, F. Gao, Y. Yang, Adaptive minimax estimation over sparse  $\ell_q$ -hulls, 2011. Arxiv Preprint. [arXiv:1108.1961](https://arxiv.org/abs/1108.1961).
- [14] Y. Yang, A.R. Barron, Information theoretic determination of minimax rates of convergence, *Annals of Statistics* 27 (1999) 1564–1599.