Prediction-based Termination Rule for Greedy Learning with Massive Data

Chen Xu[1], Shaobo Lin[2], Jian Fang[2] and Runze Li[3]

[1] *Department of Mathematics and Statistics, University of Ottawa Ottawa, ON, Canada K1N 6N5*

[2] *Department of Mathematics and Statistics, Xi'an Jiaotong University Xi'an, Shaanxi, China 710049*

[3] *Department of Statistics, The Pennsylvania State University State College, PA, USA 16801*

Supplementary Material

This supplementary material provides the proofs of Proposition 1 and Theorems 1-2 of the main manuscript. The references cited in this report are listed in the main manuscript.

S1 Technical Lemmas

To facilitate our proofs, we first introduce a few technical lemmas. Specifically, let $\mathcal{G}$ be an arbitrary set of functions (function space). We use $\mathcal{N}_\varepsilon(\mathcal{G},\nu)$ to denote the covering number of $\mathcal{G}$ by balls of radius ε with respect to a measure ν. The lemmas are presented as follows.

Lemma 1. Let $\mathcal{G}$ be a function space defined on a random variable Z. Suppose that, for some constants $C_1, C_2 \geq 0$, we have $|g(Y) - E[g(Y)]| \leq C_1$ and $E[g(Y)^2] \leq C_2 E[g(Y)]$ for any $g \in \mathcal{G}$. Then, for any $\varepsilon > 0$,

$$\mathrm{P}\left\{\sup_{g\in\mathcal{G}} \frac{E[g(Z)] - \frac{1}{n}\sum_{i=1}^n g(z_i)}{\sqrt{E[g(Z)]+\varepsilon}} > \sqrt{\varepsilon}\right\} \leq \mathcal{N}_\varepsilon(\mathcal{G}, \|.\|_\infty) \exp\left\{-\frac{n\varepsilon}{2C_2 + \frac{2C_1}{3}}\right\},$$

where $\{z_1,\ldots,z_n\}$ is an i.i.d sample from Z and $\|.\|_\infty$ is the function L^∞ norm.

Lemma 1 is a direct result from Lemma 2 of Zhou and Jetter (2006), which provides a useful probability concentration inequality to bound a function of random variable.

Lemma 2. Let $\mathcal{V}_k$ be a k-dimensional function space defined on $\mathcal{X}$. Suppose that there exists a constant T such that $|v(\boldsymbol{x})| \leq T$ for any $v \in \mathcal{V}_k$ and $\boldsymbol{x} \in \mathcal{X}$. Then

$$\log \mathcal{N}_\varepsilon(\mathcal{V}_k, \|.\|_2) \leq ck \log\frac{T}{\varepsilon},$$

where c is a positive constant and $\|.\|_2$ denotes the function L^2 norm.

Lemma 2 is implied by Corollary 2 of Mendelson and Vershinin (2003) together with Property 1 of Maiorov and Ratsaby (1999). It shows that the covering number of a bounded functional space can be also bounded properly.

Lemma 3. Let $\boldsymbol{y} = (y_1, \ldots, y_n)^T$ and $\hat{f}_k$ be the k-step estimator defined in Algorithm 1. Then, for any $h \in \text{span}\{D_z^*\}$ and $k \in \mathbb{N}_n$,

$$\|\boldsymbol{y} - \hat{f}_k\|_n^2 \leq \|\boldsymbol{y} - h\|_n^2 + \frac{4\|h\|_{l_1}^2}{k},$$

where $\|h\|_{l_1} = \inf\left\{\sum_{i=1}^n |\theta_i| : h = \sum_{i=1}^n \theta_i K(\boldsymbol{x}_i, \cdot)/\|K(\boldsymbol{x}_i, \cdot)\|_n\right\}$.

The proof of Lemma 3 is similar to Theorem 2.3 of Barron et al. (2008). It shows a nice property of the OGA estimator in terms of the empirical approximation error.

S2 Proof of Proposition 1

Recall that the generalization error of $\hat{f}_k$ is defined as

$$\mathcal{L}(\hat{f}_k) = \mathcal{E}(\hat{f}_k) - \mathcal{E}(f^*),$$

where $\mathcal{E}(f) = E(|f(X) - Y|^2)$ for $f \in \mathcal{F}$. Let $\mathcal{E}_n(f) = \|\boldsymbol{y} - f\|_n^2 = \frac{1}{n}\sum_{i=1}^n (y_i - f(\boldsymbol{x}_i))^2$. Then, for an arbitrary $h \in \text{span}\{D_z^*\}$, $\mathcal{L}(\hat{f}_k)$ can be decomposed by

$$\mathcal{L}(\hat{f}_k) = \mathcal{D} + \mathcal{P} + \mathcal{S}, \tag{S2.1}$$

where

$$\begin{aligned} \mathcal{D} &= \mathcal{E}(h) - \mathcal{E}(f^*) = \|h - f^*\|_{\rho_X}^2, \qquad \text{(S2.2)} \\ \mathcal{P} &= \mathcal{E}_n(\hat{f}_k) - \mathcal{E}_n(h), \\ \mathcal{S} &= \mathcal{E}_n(h) - \mathcal{E}(h) + \mathcal{E}(\hat{f}_k) - \mathcal{E}_n(\hat{f}_k). \end{aligned}$$

By Lemma 3, we readily have

$$\mathcal{P} \leq \frac{4\|h\|_{l_1}^2}{k}. \tag{S2.3}$$

We proceed to prove the theorem by deriving a probability bound for $\mathcal{S}$. Specifically, we further decompose $\mathcal{S}$ by

$$\mathcal{S} = \mathcal{S}_1 + \mathcal{S}_2, \tag{S2.4}$$

where

$$\begin{aligned} \mathcal{S}_1 &= \{\mathcal{E}_n(h) - \mathcal{E}_n(f^*)\} - \{\mathcal{E}(h) - \mathcal{E}(f^*)\}, \\ \mathcal{S}_2 &= \{\mathcal{E}(\hat{f}_k) - \mathcal{E}(f^*)\} - \{\mathcal{E}_n(\hat{f}_k) - \mathcal{E}_n(f^*)\}. \end{aligned}$$

Let us first work on $\mathcal{S}_1$ in (S2.4). Define

$$\begin{aligned} J(Y, X) &= [Y - h(X)]^2 - [Y - f^*(X)]^2 \\ &= [f^*(X) - h(X)][2Y - h(X) - f^*(X)]. \end{aligned}$$

Clearly, we have

$$\mathcal{S}_1 = \frac{1}{n}\sum_{i=1}^{n} J(y_i, \boldsymbol{x}_i) - E[J(Y, X)].$$

In our model setup, we assume $|Y| \le M$, which implies that

$$|J| \le (M + \|h\|_\infty)(3M + \|h\|_\infty) \le (3M + \|h\|_\infty)^2.$$

Let $\xi = (3M + \|h\|_\infty)^2$. It is then easy to show that

$$|J - E(J)| \le 2\xi \quad \text{and} \quad E(J^2) \le \mathcal{D}\xi \tag{S2.5}$$

with $\mathcal{D}$ defined in (S2.2). The bounds in (S2.5) together with Bernstein inequality (Shi, Feng, and Zhou (2011)) imply that

$$\mathcal{S}_1 \le \frac{4\xi \log \frac{1}{\delta}}{3n} + \sqrt{\frac{2\xi\mathcal{D}\log\frac{1}{\delta}}{n}} \le \frac{7\xi\log\frac{2}{\delta}}{3n} + \frac{\mathcal{D}}{2} \tag{S2.6}$$

with probability at least $1 - \delta/2$ for any $\delta \in (0, 1)$.

We now turn to bound $\mathcal{S}_2$ in (S2.4). Recall that V_k in Algorithm 1 is the active set formed by the k basis functions from a k-step OGA procedure. Let $\mathcal{F}_k = \{T_M[v] : v \in \text{span}\{V_k\}\}$ and g be an arbitrary element from

$$\mathcal{G}_k = \left\{g(X, Y) = \{f(X) - Y\}^2 - \{f^*(X) - Y\}^2,\ f \in \mathcal{F}_k\right\}.$$

Since both $|Y|$ and $|f^*|$ are bounded by M, it is straightforward to show that $|g| \le 8M^2$ and $|g - E(g)| \le 16M^2$. Also, we have

$$\begin{aligned} E(g^2) &= E\left[\{f(X) - f^*(X)\}^2 \{(f(X) - Y) + (f^*(X) - Y)\}^2\right] \\ &\le 16M^2 E(g). \end{aligned}$$

Thus, Lemma 1 becomes applicable to $\mathcal{G}_k$ with $C_1 = C_2 = 16M^2$. Note that

$$E(g) = \mathcal{L}(f) = \mathcal{E}(f) - \mathcal{E}(f^*), \quad \frac{1}{n}\sum_{i=1}^{n} g(y_i, \boldsymbol{x}_i) = \mathcal{E}_n(f) - \mathcal{E}_n(f^*)$$

for some corresponding $f \in \mathcal{F}_k$. This together with Lemma 1 implies that

$$\sup_{f\in\mathcal{F}_k} \left\{\frac{\mathcal{L}(f) - \{\mathcal{E}_n(f) - \mathcal{E}_n(f^*)\}}{\sqrt{\mathcal{L}(f) + \varepsilon}}\right\} \le \sqrt{\varepsilon} \tag{S2.7}$$

with probability at least

$$1 - \mathcal{N}_{\varepsilon/4}\left(\mathcal{G}_k, \|.\|_\infty\right) \exp\left\{-\frac{3n\varepsilon}{128M^2}\right\}.$$

Note that, for any $f_1, f_2 \in \mathcal{F}_k$ and the corresponding $g_1, g_2 \in \mathcal{G}_k$, we have

$$\begin{aligned} \|g_1 - g_2\|_\infty &= \max_{x,y}\left|(f_1(x) - y)^2 - (f_2(x) - y)^2\right| \\ &\le 4M\|f_1 - f_2\|_\infty, \end{aligned}$$

where (x, y) denotes an arbitrary realization from (X, Y). This implies that

$$\begin{aligned}\mathcal{N}_{\varepsilon/4}\left(\mathcal{G}_k, \|.\|_\infty\right) &\leq \mathcal{N}_{\varepsilon/(16M)}\left(\mathcal{F}_k, \|.\|_\infty\right)\\ &\leq \mathcal{N}_{\varepsilon/(16M)}\left(\mathcal{F}_k, \|.\|_2\right)\\ &\leq \exp\left\{ck\log\frac{16M^2}{\varepsilon}\right\}, \qquad \text{(S2.8)}\end{aligned}$$

where the last inequality follows from Lemma 2 with $T = M$. By (S2.7) and (S2.8), we have

$$P\left\{\mathcal{S}_2 \leq \frac{1}{2}\mathcal{L}(\hat{f}_k) + \varepsilon\right\} \geq 1 - \exp\left\{ck\log\frac{16M^2}{\varepsilon} - \frac{3n\varepsilon}{128M^2}\right\}. \qquad \text{(S2.9)}$$

To further specify (S2.9), let

$$h(\varepsilon) = ck\log\frac{16M^2}{\varepsilon} - \frac{3n\varepsilon}{128M^2}$$

and ε_0 be the value of ε such that $h(\varepsilon_0) = \log(\delta/2)$ for the same δ used in (S2.6). It can be shown that, by choosing

$$\varepsilon_1 = \omega\frac{k\log n + \log\frac{2}{\delta}}{n}$$

with some constant $\omega > 0$, we have $h(\varepsilon_1) \leq h(\varepsilon_0)$. Since $h(.)$ is a decreasing function, this implies $\varepsilon_1 \geq \varepsilon_0$, and therefore

$$P\left\{\mathcal{S}_2 \leq \frac{1}{2}\mathcal{L}(\hat{f}_k) + \varepsilon_1\right\} \geq 1 - \delta/2. \qquad \text{(S2.10)}$$

Combining the results from (S2.6) and (S2.10), we have

$$P\left\{\mathcal{S} \leq \frac{\mathcal{D} + \mathcal{L}(\hat{f}_k)}{2} + \frac{7\xi\log\frac{2}{\delta}}{3n} + \varepsilon_1\right\} \geq 1 - \delta. \qquad \text{(S2.11)}$$

Inequality (S2.11) together with (S2.2) and (S2.3) further implies that, with probability at least $1 - \delta$,

$$\begin{aligned}\mathcal{L}(\hat{f}_k) &\leq 3\|f^* - h\|_{\rho_X}^2 + \frac{8\|h\|_{l_1}^2}{k} + \frac{14\xi\log\frac{2}{\delta}}{3n} + 2\varepsilon_1\\ &\leq 3\|f^* - h\|_{\rho_X}^2 + \frac{8\|h\|_{l_1}^2}{k} + \frac{28\log\frac{2}{\delta}\|h\|_\infty^2}{3n} + \frac{2\omega k\log n + 6M^2 + \log\frac{2}{\delta}}{n}.\end{aligned}$$

Noting $2\log(2/\delta) > 1$, we then have, for a sufficiently large n,

$$\begin{aligned}\mathcal{L}(\hat{f}_k) &\leq 3\|f^* - h\|_{\rho_X}^2 + \frac{16\log\frac{2}{\delta}\|h\|_{l_1}^2}{k} + \frac{28\log\frac{2}{\delta}\|h\|_\infty^2}{3n} + \frac{4\omega\log\frac{2}{\delta}k\log n}{n}\\ &\leq C\left[\|f^* - h\|_{\rho_X}^2 + \log\frac{2}{\delta}\left(\frac{\|h\|_{l_1}^2}{k} + \frac{\|h\|_\infty^2}{n} + \frac{k\log n}{n}\right)\right]\end{aligned}$$

with probability at least $1-\delta$, where $C = \max\{16, 4\omega\}$. This completes the proof of Proposition 1.

S3 Proof of Theorem 1

Let $\mathcal{H}_\infty = \lim_{n\to\infty} \operatorname{span}\{D_z^*\}$. For an arbitrary $h \in \mathcal{H}_\infty$, we decompose $\mathcal{L}(\hat{f}_k)$ by

$$\mathcal{L}(\hat{f}_k) = B_1 + B_2 + B_3 + B_4, \tag{S3.1}$$

where

$$B_1 = \|h - \boldsymbol{y}\|_n^2 - \mathcal{E}(h), \quad B_2 = \mathcal{E}(\hat{f}_{k^*}) - \|\hat{f}_k^* - \boldsymbol{y}\|_n^2,$$
$$B_3 = \mathcal{E}(h) - \mathcal{E}(f^*), \quad B_4 = \|\hat{f}_k^* - \boldsymbol{y}\|_n^2 - \|h - \boldsymbol{y}\|_n^2.$$

Since $\mathcal{L}(\hat{f}_k) \geq 0$, the theorem is proved if

$$P\left\{\lim_{n\to\infty} B_j \leq 0\right\} = 1 \tag{S3.2}$$

for $j = 1, 2, 3, 4$. By the strong law of large numbers, (S3.2) readily holds for B_1. Thus, it suffices to show (S3.2) for B_2, B_3, and B_4.

We first show (S3.2) for B_2. Let

$$\mathcal{G}' = \left\{g(X, Y) = [f(X) - Y]^2 : f \in \mathcal{F}_k\right\}$$

with $\mathcal{F}_k$ same defined as in the proof of Proposition 1. Since $|Y| \leq M$, it is straightforward to show that, for any $g \in \mathcal{G}'$,

$$|g| \leq 4M^2, \quad |g - E(g)| \leq 8M^2, \quad E(g^2) \leq 4M^2 E(g).$$

Thus, by applying Lemma 1 to $\mathcal{G}'$ with $C_1 = C_2 = 8M^2$ and some arbitrary $\varepsilon > 0$, we have

$$\sup_{f\in\mathcal{F}_k} \left\{\frac{\mathcal{E}(f) - \|f - \boldsymbol{y}\|_n^2}{\sqrt{\mathcal{E}(f) + \varepsilon}}\right\} > \sqrt{\varepsilon} \tag{S3.3}$$

with probability at most

$$\mathcal{N}_{\varepsilon/4}\left(\mathcal{G}', \|.\|_\infty\right) \exp\left\{-\frac{3n\varepsilon}{64M^2}\right\}.$$

Following the same arguments in (S2.8), we have

$$N_{\varepsilon/4}\left(\mathcal{G}', \|.\|_\infty\right) \leq \exp\left\{ck \log \frac{16M^2}{\varepsilon}\right\}$$

for some positive constant c. This together with (S3.3) implies that

$$\mathcal{E}(\hat{f}_k) - \|\hat{f}_k - \boldsymbol{y}\|_n^2 > \left[\varepsilon(4M^2 + \varepsilon)\right]^{1/2} \tag{S3.4}$$

with probability at most

$$P_k = \exp\left\{ck \log \frac{16M^2}{\varepsilon} - \frac{3n\varepsilon}{64M^2}\right\}. \tag{S3.5}$$

By setting $k = k^* = T\sqrt{n/\log n}$ with some constant $T \geq 0$, we have $\sum_{n=1}^{\infty} P_{k^*} < \infty$. Thus, by Borel-Cantelli lemma, (S3.4) and (S3.5) imply that

$$P\left\{\lim_{n\to\infty} B_2 \leq \left[\varepsilon(4M^2+\varepsilon)\right]^{1/2}\right\} = 1. \tag{S3.6}$$

Since ε is arbitrary, (S3.6) further implies that (S3.2) holds for B_2.

We now proceed to show (S3.2) for B_3 and B_4. Since $|f^*(X)| \leq M$, we have $\|f^*\|_{\rho_X} \leq M$. By Theorem A.1 of Györfy et al. (2002), for any $\varepsilon' > 0$, there exists a $f' \in \mathcal{C}(\mathcal{X})$ such that $\|f' - f^*\|_{\rho_X} \leq \varepsilon'$. Also, Condition C1 implies that $\mathcal{H}_\infty$ is dense in H_K. These results together with Condition C2 imply that, for any $\varepsilon > 0$, there exists a $h_\varepsilon \in \mathcal{H}_\infty$ such that

$$\|h_\varepsilon - f^*\|^2_{\rho_X} \leq \varepsilon. \tag{S3.7}$$

By choosing $h = h_\varepsilon$ in (S3.1), we have (S3.2) holds for B_3 due to the arbitrariness of ε. Meanwhile, by setting $k = k^*$, Lemma 3 implies that

$$B_4 \leq \frac{4\|h_\varepsilon\|^2_{l_1}}{k^*}. \tag{S3.8}$$

Since D_z^* is a normalized dictionary, (S3.7) implies that $\|h_\varepsilon\|_{l_1} < \infty$. Thus, the right hand side of (S3.8) goes to zero as $n \to \infty$, which implies that (S3.2) holds for B_4. The theorem is therefore proved.

S4 Proof of Theorem 2

Proposition 1 implies that, for any $h \in \text{span}\{D_z^*\}$ and n large enough,

$$\mathcal{L}(\hat{f}_k) \leq C\left\{\|f^* - h\|^2_{\rho_X} + \log\frac{2}{\delta}\left(\frac{\|h\|^2_{l_1}}{k} + \frac{\|h\|^2_\infty + k\log n}{n}\right)\right\}$$

with probability at least $1-\delta$ for $\delta \in (0,1)$. When Condition C3 is satisfied with $r > 0.5$, we have $\|h'\|_{l_1} \leq B$ and $\|f^* - h'\|_{\rho_X} \leq \|f^* - h'\|_\infty \leq Bn^{-1/2}$ for some $h' \in \text{span}\{D_z^*\}$. Since $K(.,.)$ is continues and $\mathcal{X}$ is compact, Condition C3 also implies that $\|h'\|^2_\infty$ is bounded by some positive constant B'. Based on these results, we have

$$\mathcal{L}(\hat{f}_k) \leq C\left\{B^2 n^{-1} + \log\frac{2}{\delta}\left(\frac{B^2}{k} + \frac{B' + k\log n}{n}\right)\right\}$$

with probability at least $1-\delta$. By setting $k = k^* = T(n/\log n)^{1/2}$, we have

$$P\left\{\mathcal{L}(\hat{f}_k) > C'\log\frac{2}{\delta}\sqrt{\frac{\log n}{n}}\right\} \leq \delta$$

for some generic positive constant C' with a sufficiently large n. Let $t = C' \log \frac{2}{\delta} (\log n/n)^{1/2}$, we then have

$$\begin{aligned} E[\mathcal{L}(\hat{f}_k)] &= \int_0^\infty P\{\mathcal{L}(\hat{f}_k) > t\} dt \\ &\leq \int_0^\infty 2 \exp\left\{-\frac{t}{C'}\sqrt{\frac{n}{\log n}}\right\} dt \\ &\leq 2C'\sqrt{\frac{\log n}{n}}. \end{aligned}$$

The theorem is therefore proved.