Mathematical and Computer Modelling 57 (2013) 1395–1407

Contents lists available at SciVerse ScienceDirect

Mathematical and Computer Modelling

journal homepage: www.elsevier.com/locate/mcm

# Learning performance of elastic-net regularization

Yu-long Zhao *, Yun-long Feng

*Department of Mathematics, City University of Hong Kong, Kowloon, Hong Kong, China*

ARTICLE INFO

*Article history:*
Received 13 June 2011
Received in revised form 22 November 2012
Accepted 25 November 2012

*Keywords:*
Learning theory
Elastic-net regularization
$\ell^2$-empirical covering number
Learning rate

ABSTRACT

In this paper, within the framework of statistical learning theory we address the elastic-net regularization problem. Based on the capacity assumption of hypothesis space composed by infinite features, significant contributions are made in several aspects. First, concentration estimates for sample error are presented by introducing $\ell^2$-empirical covering number and utilizing an iteration process. Second, a constructive approximation approach for estimating approximation error is presented. Third, the elastic-net learning with infinite features is studied and the role that the tuning parameter $\zeta$ plays is also discussed. Finally, our learning rate is shown to be faster compared with existing results.



## 1. Introduction and main results

During the last few decades, several regularized methods for linear regression have been adopted to overcome deficiencies of ordinary least square regression on prediction and interpretation. Shrinking coefficients toward zero, ridge regression [1] achieves better prediction performance through a bias-variance trade-off. However, ridge regression is not able to provide a sparse model which can be interpreted better since the coefficients are shrunken toward zero but never become zero exactly. Aiming at continuous shrinkage and automatic variable selection simultaneously, a penalized least squares method called LASSO is proposed [2] by imposing an $\ell^1$-regularizer on regression coefficients. Different from ridge regression, coefficients in LASSO can be shrunken toward zero exactly, which leads to much better interpretability. However, in some special cases the LASSO also shows its deficiency, for example when the variables have group effects or the number of predictors is much larger than the number of observations [3]. Mindful of these flaws, a regularized regression scheme generated by a combination of the LASSO and ridge penalty is proposed. It was first introduced in [3] and then analyzed in [4]. It is demonstrated that the elastic net often outperforms LASSO and simultaneously preserves the sparse property [4,3]. The advantages of this regularization scheme have been also confirmed by various applications [4,3,5,6].

In this paper, we focus on the statistical properties of this scheme and in particular its consistency property, which is studied within the framework of statistical learning theory. To address this problem, we first present a mathematical setup, which follows the setting in [4].

The regression problem aims at learning a regression function on a separable metric space $X$ (called the input space) with values in $Y = \mathbb{R}$. The elastic net algorithm is given in terms of finite set $\{\varphi_k\}_{k=1}^N$ of continuous functions on $X$ with sufficiently large $N$, which is a subset of a dictionary $\{\varphi_k\}_{k\in\Gamma}$ with cardinality $|\Gamma|$ countable, where $|\Gamma| \geq N$. Its regularizer is an elastic net penalty on $\mathbb{R}^N$. In fact the learning algorithm can be extended to the infinite case, as we will explain later. We first present the definition of elastic net penalty.

* Corresponding author.
*E-mail addresses:* zhaoyulong@gmail.com, zyulong2@student.cityu.edu.hk (Y.-l. Zhao), yunlfeng@cityu.edu.hk (Y.-l. Feng).

0895-7177/$ – see front matter © 2012 Elsevier Ltd. All rights reserved.
doi:10.1016/j.mcm.2012.11.028

**Definition 1.** Let $\zeta > 0$, the *elastic net penalty* $p_\zeta : \mathbb{R}^N \to [0, \infty)$ is defined as

$$p_\zeta(\beta) = \sum_{k=1}^{N} \{|\beta_k| + \zeta \beta_k^2\}$$

where $N$ is a positive integer.

The hypothesis space $\mathcal{H}_N$ for the regularization scheme consists of linear combinations of $N$ features: $f_\beta = \sum_{k=1}^N \beta_k \varphi_k$. We adopt the setting introduced in [4], which also can be found in [7]. Explicitly, $\mathcal{H}_N = \left\{ f : f = \sum_{k=1}^N \beta_k \varphi_k \right\}$ is a subset of $\mathcal{H}_\Gamma$, which is defined as

$$\mathcal{H}_\Gamma = \left\{ f : f = \sum_{k \in \Gamma} \beta_k \varphi_k, \beta_k \in \mathbb{R} \right\}. \tag{1.1}$$

Then elastic net algorithm is now defined for a given sample $\mathbf{z} = \{(x_i, y_i)\}_{i=1}^m \in (X \times Y)^m$ by

$$f_{\mathbf{z}} = \arg\min_{f_\beta \in \mathcal{H}_N} \left( \frac{1}{m} \sum_{i=1}^{m} (f_\beta(x_i) - y_i)^2 + \lambda \mathcal{P}_\zeta(f_\beta) \right), \tag{1.2}$$

where $\lambda = \lambda(m) \geq 0$ is a regularization parameter and $\mathcal{P}_\zeta(f_\beta) := p_\zeta(\beta)$.

As mentioned above, in this paper we are interested in the learning ability of the algorithm (1.2). To this end, we take a common model in learning theory and assume that $\rho$ is a Borel probability measure on $Z := X \times Y$ and the *regression function* is defined by

$$f_\rho(x) = \int_Y y d\rho(y|x), \quad x \in X, \tag{1.3}$$

where $\rho(\cdot|x)$ is the conditional probability measure induced by $\rho$ at $x \in X$.

In the supervised learning framework, $\rho$ is unknown and one cannot obtain the regression function $f_\rho$ directly. Indeed, we learn the regression function from the sample $\mathbf{z} = \{(x_i, y_i)\}_{i=1}^m \in Z^m$, which is assumed to be drawn independently according to the measure $\rho$. Throughout this paper, we assume that $\rho(\cdot|x)$ is supported on $[-M, M]$, for some $M > 0$. The learning ability of the algorithm (1.2) is measured by the error $\|f_{\mathbf{z}} - f_\rho\|_{L^2_{\rho_X}}$ of the difference function $f_{\mathbf{z}} - f_\rho$ in the space $L^2_{\rho_X}$ where $\rho_X$ is the marginal distribution of $\rho$ on $X$.

Considering that the analysis in this paper is based on the complexity assumption of the hypothesis space, we need the following capacity condition for $\mathcal{H}_\Gamma$ in terms of $\ell^2$-empirical covering numbers.

**Definition 2.** Let $(\mathcal{M}, d)$ be a pseudo-metric space and $S \subset \mathcal{M}$ a subset. For every $\epsilon > 0$, the covering number $\mathcal{N}(S, \epsilon, d)$ is defined as the minimal number of balls of radius $\epsilon$ whose union covers $S$, that is,

$$\mathcal{N}(S, \epsilon, d) = \min \left\{ \ell \in \mathbb{N} : S \subset \bigcup_{j=1}^{\ell} B(s_j, \epsilon) \text{ for some } \{s_j\}_{j=1}^{\ell} \subset \mathcal{M} \right\},$$

where $B(s_j, \epsilon) = \{s \in \mathcal{M} : d(s, s_j) \leq \epsilon\}$.

Let $d_2$ be the normalized metric on the Euclidian space $\mathbb{R}^n$ given by

$$d_2(\mathbf{a}, \mathbf{b}) = \left( \frac{1}{n} \sum_{i=1}^{n} |a_i - b_i|^2 \right)^{1/2} \quad \text{for } \mathbf{a} = \{a_i\}_{i=1}^n, \ \mathbf{b} = \{b_i\}_{i=1}^n \in \mathbb{R}^n.$$

**Definition 3.** Let $\mathcal{F}$ be a set of functions on $X$, $\mathbf{x} = (x_i)_{i=1}^n \subset X^n$ and $\mathcal{F}|_{\mathbf{x}} = \{(f(x_i))_{i=1}^n : f \in \mathcal{F}\} \subset \mathbb{R}^n$. Set $\mathcal{N}_{2,\mathbf{x}}(\mathcal{F}, \epsilon) = \mathcal{N}(\mathcal{F}|_{\mathbf{x}}, \epsilon, d_2)$. The $\ell^2$-empirical covering number of $\mathcal{F}$ is defined by

$$\mathcal{N}_2(\mathcal{F}, \epsilon) = \sup_{n \in \mathbb{N}} \sup_{\mathbf{x} \in X^n} \mathcal{N}_{2,\mathbf{x}}(\mathcal{F}, \epsilon), \quad \epsilon > 0.$$

**Assumption 1.** The space $\mathcal{H}_\Gamma$ has empirical polynomial complexity with exponent $p$, where $0 < p < 2$. That is, there exists a constant $c_{p,\mathcal{H}_\Gamma} > 0$ such that

$$\log \mathcal{N}_2(B_1, \epsilon) \leq c_{p,\mathcal{H}_\Gamma} \left( \frac{1}{\epsilon} \right)^p, \quad \forall \epsilon > 0, \tag{1.4}$$

where $B_1$ is the subset of $\mathcal{H}_\Gamma$ defined by $B_R = \{f_\beta \in \mathcal{H}_\Gamma : \|\beta\|_{\ell^1} \leq R\} \cap \{f_\beta \in \mathcal{H}_\Gamma : \|\beta\|_{\ell^2} \leq \sqrt{\frac{R}{\zeta}}\}$ with $R = 1$.

For the dictionary, we assume that

$$\sum_{k\in\Gamma}|\varphi_k(x)|^2 \le 1, \quad \forall x \in X. \tag{1.5}$$

Assumption 1 could be considered as a condition on the complexity of $\ell^1$-ball and $\ell^2$-ball in $\mathcal{H}_N$ with fixed $\zeta$, since $\mathcal{H}_N$ is a subspace of $\mathcal{H}_\Gamma$. Concerning the $\ell^2$-empirical covering number, some references can be found in [8–10]. As pointed out there, compared with uniform covering number [11], empirical covering number often leads to sharper learning rates [9].

Besides the assumptions on the dictionary $\{\varphi_k\}_{k\in\Gamma}$ and the capacity of the space $\mathcal{H}_\Gamma$, our estimates also depend on the regularity of the regression function $f_\rho$, which is characterized in terms of decay of its coefficients.

**Assumption 2.** We assume that for some $1 < q \le 2$, $0 < s \le \frac{2(q-1)}{q}$ and a non-increasing sequence $\lambda_1 \ge \lambda_2 \ge \cdots \ge 0$, with $\Lambda := \sum_{k\in\Gamma}\lambda_k^2 < +\infty$ and $\lim_{k\to\infty}\lambda_k = 0$, the regression function is given by $f_\rho = f_{\beta^*}$ where $\beta^* \in \ell^2$ is a sequence given by $\beta_k^* = \lambda_k^s\alpha_k (k \in \Gamma)$ with $C_\alpha := \left(\sum_{k\in\Gamma}|\alpha_k|^q\right)^{1/q} < \infty$.

With these preparations, our main result on the learning rate of algorithm (1.2) can be stated as follows.

**Theorem 1.** *Assume that Assumptions 1 and 2 hold, and $f_{\mathbf{z}}$ is given by (1.2). Let $\tau = \frac{2qs}{qs+2(q-1)}$, for any $0 < \delta < 1$ and $\theta \in [\tau-1, +\infty)$, take $\zeta = \lambda^\theta$, $\lambda = m^{-\gamma}$, where*

$$\begin{cases} 0 < \gamma < \dfrac{1}{2+p}, & \text{if } \theta \in (1-\tau, +\infty), \\ 0 < \gamma < \dfrac{2}{(2+p)(1+\theta)}, & \text{if } \theta \in [\tau-1, 1-\tau]. \end{cases}$$

*Then with confidence $1-\delta$ there holds*

$$\|f_{\mathbf{z}} - f_\rho\|_{L^2_{\rho_X}}^2 \le c_0\left(\log\left(\frac{2J_0}{\delta}\right)\right)^{2J_0+1}\left(\frac{1}{m}\right)^{\Theta},$$

*where $c_0$ is a positive constant independent of $m$ and $\delta$, $J_0$ is explicitly given by (3.9) and*

$$\Theta = \begin{cases} \min\left\{\gamma\tau, \dfrac{2}{2+p} - (2-2\tau)\gamma\right\}, & \text{if } \theta \in (1-\tau, +\infty), \\ \min\left\{\gamma\tau, \gamma(\tau-1-\theta) + \dfrac{2}{2+p}\right\}, & \text{if } \theta \in [\tau-1, 1-\tau]. \end{cases}$$

To illustrate the learning rate in Theorem 1, let us reformulate Theorem 1 for the specific case that the involved covering number and the regularity of the regression function have certain behaviors.

**Corollary 1.** *Under assumptions of Theorem 1, suppose that $\tau$ is sufficiently close to 1 and $p$ is small enough, let $0 < \varepsilon < 1$ and choose $\zeta = \lambda^\theta$, $\theta \in [-\epsilon, +\infty)$, $\lambda = m^{-\gamma}$ where*

$$\begin{cases} \gamma = \dfrac{1}{2} - \dfrac{1}{2}\varepsilon, & \text{if } \theta \in (\varepsilon, +\infty), \\ \gamma = \dfrac{1-\varepsilon}{1+\theta}, & \text{if } \theta \in [-\varepsilon, \varepsilon]. \end{cases}$$

*For any $0 < \delta < 1$, with confidence $1-\delta$ we have*

$$\|f_{\mathbf{z}} - f_\rho\|_{L^2_{\rho_X}}^2 \le c_0\left(\log\left(\frac{2J_0}{\delta}\right)\right)^{2J_0+1}\left(\frac{1}{m}\right)^{\Theta},$$

*where $c_0$ is a positive constant that independent of $m$ or $\delta$, $J_0$ is explicitly given in (3.9) and*

$$\Theta = \begin{cases} \dfrac{1}{2} - \varepsilon + \dfrac{1}{2}\varepsilon^2, & \text{if } \theta \in (\varepsilon, +\infty), \\ \dfrac{(1-\varepsilon)^2}{1+\theta}, & \text{if } \theta \in [-\varepsilon, \varepsilon]. \end{cases}$$

**Remark 1.** As will be shown later, the parameter $\tau$ specifies the approximation ability of functions in the hypothesis space $\mathcal{H}_N$ to the regression function $f_\rho$ and further characterizes the regularity of $f_\rho$. Specifically, $\tau = 1$ indicates that $f_\rho \in \mathcal{H}_N$. Moreover, it is easy to see that the index $p$ reflects the regularity of functions in $\mathcal{H}_N$ and hence characterizes the complexity of $\mathcal{H}_N$. The assumption that $p$ is sufficiently small indicates that the complexity of the hypothesis space is small enough.

Apparently, Corollary 1 shows the influence of the tuning parameter $\zeta$ on the learning rate. A closer look reveals that with the to be chosen $\zeta$, the elastic net regularization can be interpreted as an interpolation of $\ell^1$ regularization and $\ell^2$ regularization. It coincides with the motivation of introducing the elastic net regularizer [3,4,7].

Our contributions are summarized as follows.

- Approximation error is explicitly derived. Our work follows the setting of [4], while not much attention is paid to the approximation error in [4]. The approximation error further indicates the influence of the $\ell^2$ term to the approximation ability of $f_{\mathbf{z}}$.
- Sharper learning rates are obtained compared with existing learning literatures.
- We show that the elastic-net learning algorithm can be extended to the infinite features. Besides the results on consistency and learning rates, the support of the empirical target function $f_{\mathbf{z}}$ is also explicitly given in this case. Under proper conditions, this support set can be finite, which immediately gives computational implications.

Proofs of Theorem 1 and Corollary 1 are provided in Section 5. The rest of the paper is organized as follows. In Section 2, error analysis and estimates for the sample error and approximation error are presented respectively. In Section 3, an iteration technique is applied to improve excess generalization error. Besides the proofs, further analysis on the learning rate and the tuning parameter $\zeta$ is also given in Section 5.

## 2. Error analysis and estimation

In this section, we apply an error decomposition procedure to conduct error analysis [21]. Some notations are defined as follows.

Define the *generalization error* for a function $f : X \to Y$ as

$$\mathcal{E}(f) = \int_Z (f(x) - y)^2 d\rho. \tag{2.1}$$

Then $f_\rho$ is the minimizer of the generalization error since $\mathcal{E}(f) = \|f - f_\rho\|^2_{L^2_{\rho_X}} + \mathcal{E}(f_\rho)$. Moreover, the error $\|f_{\mathbf{z}} - f_\rho\|_{L^2_{\rho_X}}$ can be written in terms of the *excess generalization error* as

$$\|f_{\mathbf{z}} - f_\rho\|^2_{L^2_{\rho_X}} = \mathcal{E}(f_{\mathbf{z}}) - \mathcal{E}(f_\rho),$$

which can be used to measure the learning ability or statistical performance of the learning algorithm. Replacing the integral by the sample mean, for a fixed function $f$, the generalization error $\mathcal{E}(f)$ can be approximated by the *empirical error*

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

Accordingly, algorithm (1.2) can be rewritten as

$$f_{\mathbf{z}} = \arg\min_{f\in\mathcal{H}_N} \left\{\mathcal{E}_{\mathbf{z}}(f) + \lambda\mathcal{P}_\zeta(f)\right\}.$$

To derive error bounds, we introduce a regularizing function $f_\lambda$ which is given by

$$f_\lambda = \arg\min_{f\in\mathcal{H}_N} \left\{\mathcal{E}(f) - \mathcal{E}(f_\rho) + \lambda\mathcal{P}_\zeta(f)\right\}.$$

The following proposition divides the excess generalization error into three parts, which we will estimate separately in the sequel.

**Proposition 1.** *Let $\lambda > 0$, there holds*

$$\mathcal{E}(f_{\mathbf{z}}) - \mathcal{E}(f_\rho) \le \{\mathcal{E}(f_{\mathbf{z}}) - \mathcal{E}_{\mathbf{z}}(f_{\mathbf{z}})\} + \{\mathcal{E}_{\mathbf{z}}(f_\lambda) - \mathcal{E}(f_\lambda)\} + \{\mathcal{E}(f_\lambda) - \mathcal{E}(f_\rho) + \lambda\mathcal{P}_\zeta(f_\lambda)\}.$$

**Proof.** When $\lambda > 0$, it is easy to see that

$$\begin{aligned}\mathcal{E}(f_{\mathbf{z}}) - \mathcal{E}(f_\rho) &\le \mathcal{E}(f_{\mathbf{z}}) - \mathcal{E}(f_\rho) + \lambda\mathcal{P}_\zeta(f_{\mathbf{z}})\\ &\le \{\mathcal{E}(f_{\mathbf{z}}) - \mathcal{E}_{\mathbf{z}}(f_{\mathbf{z}})\} + \{\mathcal{E}_{\mathbf{z}}(f_{\mathbf{z}}) + \lambda\mathcal{P}_\zeta(f_{\mathbf{z}}) - \mathcal{E}_{\mathbf{z}}(f_\lambda) - \lambda\mathcal{P}_\zeta(f_\lambda)\}\\ &\quad + \{\mathcal{E}_{\mathbf{z}}(f_\lambda) - \mathcal{E}(f_\lambda)\} + \{\mathcal{E}(f_\lambda) - \mathcal{E}(f_\rho) + \lambda\mathcal{P}_\zeta(f_\lambda)\}.\end{aligned}$$

Following the definition of $f_{\mathbf{z}}$ and $f_\lambda$, one has $\mathcal{E}_{\mathbf{z}}(f_{\mathbf{z}}) + \lambda\mathcal{P}_\zeta(f_{\mathbf{z}}) - \mathcal{E}_{\mathbf{z}}(f_\lambda) - \lambda\mathcal{P}_\zeta(f_\lambda) \le 0$. This completes the proof. □

**Remark 2.** The term $\{\mathcal{E}(f_{\mathbf{z}}) - \mathcal{E}_{\mathbf{z}}(f_{\mathbf{z}})\} + \{\mathcal{E}_{\mathbf{z}}(f_\lambda) - \mathcal{E}(f_\lambda)\}$ is called sample error, while the term $\mathcal{D}(\lambda) = \mathcal{E}(f_\lambda) - \mathcal{E}(f_\rho) + \lambda\mathcal{P}_\zeta(f_\lambda)$ is called approximation error.

### 2.1. Bounding approximation error

In this subsection, we present a constructive approximation approach to estimate approximation error [22], which is one of our main technical contributions in the analysis of elastic net regularization scheme in this paper.

**Theorem 2.** *Under the Assumption 2 for $f_\rho$, if $\zeta \le \lambda^{\frac{qs-2(q-1)}{qs+2(q-1)}}$, then for any $\lambda > 0$, there exists some $f_\lambda \in \mathcal{H}_N$ such that*

$$\mathcal{D}(\lambda) = \|f_\lambda - f_\rho\|^2_{L^2_{\rho_X}} + \lambda \mathcal{P}_\zeta(f_\lambda) \le C_1 \lambda^{\frac{2qs}{qs+2(q-1)}}, \tag{2.2}$$

*where $C_1 = C_\alpha \Lambda^{\frac{q-1}{q}} + C_\alpha^2 \lambda_1^{2s} + C_\alpha^2$, $\Lambda := \sum_{k\in\Gamma} \lambda_k^2$ and $C_\alpha := \left(\sum_{k\in\Gamma} |\alpha_k|^q\right)^{1/q}$.*

**Proof.** According to Assumption 2, we know that $f_\rho = \sum_{k\in\Gamma} \lambda_k^s \alpha_k \varphi_k$. Taking into account the value of $\lambda$, we can bound the approximation error as follows.

If $0 < \lambda \le \lambda_1^{s+\frac{2(q-1)}{q}}$, then there exists some $N_1 \in \mathbb{N}$ such that $\lambda_{N_1+1} < \lambda^{\frac{q}{qs+2(q-1)}} \le \lambda_{N_1}$. Choose $f_\lambda = \sum_{k=1}^{N_1} \lambda_k^s \alpha_k \varphi_k$. For $1 \le k \le N_1$, there holds $\lambda^{\frac{q}{qs+2(q-1)}} \le \lambda_{N_1} \le \lambda_k$. Under regularity assumption this yields

$$\sum_{k=1}^{N_1} \lambda_k^{\frac{qs}{q-1}} = \sum_{k=1}^{N_1} \lambda_k^2 \lambda_k^{\frac{sq}{q-1}-2} \le \lambda^{\frac{\frac{sq}{q-1}-2}{s+\frac{2(q-1)}{q}}} \Lambda.$$

Together with the Hölder's inequality, we have

$$\begin{aligned}\mathcal{P}_\zeta(f_\lambda) &= \sum_{k=1}^{N_1} |\alpha_k| \lambda_k^s + \zeta \sum_{k=1}^{N_1} |\alpha_k|^2 \lambda_k^{2s} \le \left(\sum_{k=1}^{N_1} |\alpha_k|^q\right)^{\frac{1}{q}} \left(\sum_{k=1}^{N_1} \lambda_k^{s\frac{q}{q-1}}\right)^{\frac{q-1}{q}} + \zeta \sum_{k=1}^{N_1} |\alpha_k|^2 \lambda_k^{2s} \\ &\le (\Lambda^{\frac{q-1}{q}} C_\alpha + C_\alpha^2 \lambda_1^{2s}) \left(\lambda^{\frac{sq-2(q-1)}{sq+2(q-1)}} + \zeta\right),\end{aligned}$$

where the second inequality follows from the fact that $s \le \frac{2(q-1)}{q}$. Meanwhile, by the Schwartz inequality one also gets

$$\begin{aligned}\|f_\rho - f_\lambda\|^2_{L^2_{\rho_X}} &= \left\|\sum_{k\ge N_1+1} \alpha_k \lambda_k^s \varphi_k\right\|^2_{L^2_{\rho_X}} = \int_X \left(\sum_{k\ge N_1+1} \alpha_k \lambda_k^s \varphi_k\right)^2 d\rho_X \\ &\le \int_X \left(\sum_{k\ge N_1+1} \alpha_k^2 \lambda_k^{2s}\right) \sum_{k\ge N_1+1} \varphi_k^2 d\rho_X \le C_\alpha^2 \lambda^{\frac{2sq}{sq+2(q-1)}}.\end{aligned}$$

Thus, if $\zeta \le \lambda^{\frac{sq-2(q-1)}{sq+2(q-1)}}$, there holds

$$\mathcal{D}(\lambda) \le \lambda^{\frac{2sq}{sq+2(q-1)}} \left(C_\alpha \Lambda^{\frac{q-1}{q}} + C_\alpha^2 \lambda_1^{2s} + C_\alpha^2\right) = C_1 \lambda^{\frac{2sq}{sq+2(q-1)}}.$$

If $\lambda > \lambda_1^{s+2\frac{q-1}{q}}$, by taking $f_\lambda = 0$ there still holds

$$\mathcal{D}(\lambda) \le \|f_\rho\|^2_{L^2_{\rho_X}} = \left\|\sum_{k\ge 1} \alpha_k \lambda_k^s \varphi_k\right\|^2_{L^2_{\rho_X}} \le C_\alpha^2 \lambda_1^{2s} \le C_\alpha^2 \lambda^{\frac{2sq}{sq+2(q-1)}}. \tag{2.3}$$

Hence we obtain desired estimation by combining two cases above. □

To simplify analysis, in what follows, we always assume that $\mathcal{D}(\lambda) \le C_1 \lambda^\tau$ under the condition $\zeta \le \lambda^{\tau-1}$ with $C_1$ given above and $\tau \in (0, 1]$. According to Theorem 2, we know that $\tau = \frac{2sq}{sq+2(q-1)}$.

**Remark 3.** According to the proof of Theorem 2, it is necessary to assume that $N$ is large enough to ensure $f_\lambda \in \mathcal{H}_N$.

### 2.2. Concentration estimates for sample error involving $f_\lambda$

In this subsection, by one-side Bernstein-type inequality the first term $S_1(\mathbf{z}, \lambda)$ is estimated as follows [8].

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

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

Denote $\xi(z) = (y - f_\lambda(x))^2 - (y - f_\rho(x))^2$ as a random variable on $z = (x, y) \in Z$. Applying Lemma 1, we come to the following proposition.

**Proposition 2.** *For any $0 < \delta < 1$, with confidence $1 - \frac{\delta}{2}$ there holds,*

$$S_1(\mathbf{z}, \lambda) \le \frac{1}{2}\mathcal{D}(\lambda) + \frac{7(3M + r_\zeta)^2 \log\frac{2}{\delta}}{3m},$$

*where*

$$r_\zeta := \min\left\{\frac{\mathcal{D}(\lambda)}{\lambda}, \sqrt{\frac{\mathcal{D}(\lambda)}{\zeta\lambda}}\right\}. \tag{2.4}$$

**Proof.** According to the definition of $\mathcal{D}(\lambda)$, we have

$$\|f_\lambda\|_\infty = \left\|\sum_{k=1}^{N}\varphi_k\beta_k^\lambda\right\|_\infty \le \min\left\{\left(\sum_{k=1}^{N}(\beta_k^\lambda)^2\right)^{\frac{1}{2}}, \sum_{k=1}^{N}|\beta_k^\lambda|\right\} \le r_\zeta.$$

Recalling that the random variable $\xi$ is defined by $\xi(z) = (y - f_\lambda(x))^2 - (y - f_\rho(x))^2$ with $z = (x, y) \in Z$ and $|f_\rho(x)| \le M$ almost surely, we get

$$|\xi(z)| \le (3M + \|f_\lambda\|_\infty)(M + \|f_\lambda\|_\infty) \le c := (3M + r_\zeta)^2.$$

Hence one obtains $|\xi - \mathbb{E}\xi| \le 2c$. Moreover

$$\sigma^2 \le \mathbb{E}\xi^2 = \int_Z (f_\lambda - f_\rho)^2(2y - f_\lambda - f_\rho)^2 d\rho \le (3M + \|f_\lambda\|_\infty)^2\|f_\lambda - f_\rho\|_{L^2_{\rho X}}^2 \le c\mathcal{D}(\lambda).$$

Applying one-side Bernstein inequality in Lemma 1, with confidence $1 - \delta/2$ we have

$$S_1(z, \lambda) \le \frac{1}{2}\mathcal{D}(\lambda) + \frac{7(3M + r_\zeta)^2 \log\frac{2}{\delta}}{3m}.$$

Hence we obtain our desired estimate on $\mathcal{S}_1(\mathbf{z}, \lambda)$. □

### 2.3. Concentration estimates for sample error involving $f_\mathbf{z}$

In this subsection, we provide a concentration estimate for $f_\mathbf{z}$. Since the empirical target function $f_\mathbf{z}$ varies with sample $\mathbf{z}$, the estimate for $f_\mathbf{z}$ involves the complexity of the hypothesis space $\mathcal{H}_N$. This complicates our analysis. To overcome this difficulty, we introduce the following lemma, which was proposed in [9] and also employed in [10,8]. Based on the assumption of the hypothesis space, it provides a concentration inequality, which plays a key role in our analysis for $\mathcal{S}_2(\mathbf{z}, \lambda)$.

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

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

*then there exists a constant $c_p$ depending only on $p$ such that for any $t > 0$, with probability at least $1 - e^{-t}$, there holds*

$$\mathbb{E}f - \frac{1}{m}\sum_{i=1}^{m}f(z_i) \le \frac{1}{2}\eta^{1-\theta}(\mathbb{E}f)^\theta + c_p\eta + 2\left(\frac{ct}{m}\right)^{\frac{1}{2-\theta}} + \frac{18Bt}{m}, \quad \forall f \in \mathcal{F},$$

*where*

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

Denote the function set $\mathcal{F}_R = \left\{(y-f(x))^2-(y-f_\rho(x))^2 : f \in B_R\right\}$ with $R>0$. Applying Lemma 2, we get the following proposition which gives an upper bound of $\mathcal{S}_2(\mathbf{z},\lambda)$.

**Proposition 3.** *Suppose that the Assumption 1 holds and $R_\zeta \geq M$. Then for any $0<\delta<1$, with confidence $1-\frac{\delta}{2}$, there holds*

$$\left\{\mathcal{E}(f)-\mathcal{E}(f_\rho)\right\}-\left\{\mathcal{E}_{\mathbf{z}}(f)-\mathcal{E}_{\mathbf{z}}(f_\rho)\right\} \leq \frac{1}{2}\left\{\mathcal{E}(f)-\mathcal{E}(f_\rho)\right\}+c_1\log\left(\frac{2}{\delta}\right)\frac{R_\zeta^2}{m^{2/(2+p)}}, \quad \forall f \in B_R,$$

*where $c_1$ is a positive constant independent of $m$, $\delta$ and $R_\zeta$, which will be given explicitly in the proof and*

$$R_\zeta = \min\left\{R, \sqrt{\frac{R}{\zeta}}\right\}. \tag{2.5}$$

**Proof.** Consider the function set $\mathcal{F}_R$ on $Z$ and let $g \in \mathcal{F}_R$. Following the definition of $\mathcal{F}_R$ we have

$$g(z) = (f(x)-f_\rho(x))\left\{(f(x)-y)+(f_\rho(x)+y)\right\}.$$

Recalling that $\|f\|_\infty \leq \min\left\{\|\beta\|_{\ell^1}, \|\beta\|_{\ell^2}\right\} \leq R_\zeta$ and $|f_\rho(x)| \leq M$ almost surely, we get

$$|g(z)| \leq (R_\zeta+M)(R_\zeta+3M) \leq (3M+R_\zeta)^2$$

and

$$Eg^2 \leq \int_Z (2y-f(x)-f_\rho(x))^2(f(x)-f_\rho(x))^2 d\rho \leq (3M+R_\zeta)^2 Eg.$$

It follows that given $g_1,\ g_2 \in \mathcal{F}_R$, one gets

$$|g_1(z)-g_2(z)| = |(y-f_1(x))^2-(y-f_2(x))^2| \leq (2M+2R_\zeta)|f_1(x)-f_2(x)|.$$

This in connection with the definition of $\ell^2$-empirical covering number implies that

$$\mathcal{N}_{2,\mathbf{z}}(\mathcal{F}_R,\epsilon) \leq \mathcal{N}_{2,\mathbf{z}}\left(B_R, \frac{\epsilon}{2M+2R_\zeta}\right) \leq \mathcal{N}_{2,\mathbf{z}}\left(B_1, \frac{\epsilon}{R_\zeta(2M+2R_\zeta)}\right)$$

which further implies

$$\log \mathcal{N}_{2,\mathbf{z}}(\mathcal{F}_R,\epsilon) \leq c_{p,\mathcal{H}_\Gamma} R_\zeta^p (2M+2R_\zeta)^p \epsilon^{-p}.$$

Applying Lemma 2 with $B=c=(3M+R_\zeta)^2$, $\alpha=1$ and $a=c_{p,\mathcal{H}_\Gamma}R_\zeta^p(2M+2R_\zeta)^p$, for any $0<\delta<1$, with confidence $1-\delta$ there holds

$$\left\{\mathcal{E}(f)-\mathcal{E}(f_\rho)\right\}-\left\{\mathcal{E}_{\mathbf{z}}(f)-\mathcal{E}_{\mathbf{z}}(f_\rho)\right\} \leq \frac{1}{2}\left\{\mathcal{E}(f)-\mathcal{E}(f_\rho)\right\}+c_1\log\left(\frac{2}{\delta}\right)\frac{R_\zeta^2}{m^{2/(2+p)}},$$

where $c_1 = 320+512c_p c_{p,\mathcal{H}_\Gamma}^{2/(2+p)}$. □

Based on the evaluation of approximation error and sample error, the following upper bound on excess generalization error is a direct result after simple computations, hence we omit its proof here.

**Proposition 4.** *Let $0<\lambda\leq 1$, then for any $0<\delta<1$, with confidence $1-\delta$ there holds,*

$$\mathcal{E}(f_{\mathbf{z}})-\mathcal{E}(f_\rho)+\lambda\mathcal{P}_\zeta(f_{\mathbf{z}}) \leq C\log\left(\frac{2}{\delta}\right)\left(\mathcal{D}(\lambda)+\frac{1}{m}+\frac{R_\zeta^2}{m^{2/(2+p)}}+\frac{r_\zeta^2}{m}\right), \tag{2.6}$$

*where $R_\zeta$ and $r_\zeta$ are defined as above, and $C=2c_1+11+84M^2$.*

## 3. Improved estimates via iteration process

To get an explicit upper bound of excess generalization error, we need to evaluate $R$. The following lemma presents a rough upper bound for $R$.

**Lemma 3.** *Let $\lambda > 0$, for almost all $\mathbf{z} \in Z^m$, there holds*

$$\|\beta_{\mathbf{z}}\|_{\ell^1} \leq \frac{M^2}{\lambda} \quad \text{and} \quad \zeta \|\beta_{\mathbf{z}}\|_{\ell^2}^2 \leq \frac{M^2}{\lambda}.$$

**Proof.** Following the definition of the algorithm, we have

$$\lambda \|\beta_{\mathbf{z}}\|_{\ell^1} \leq \lambda \mathcal{P}_\zeta(f_{\mathbf{z}}) \leq \mathcal{E}_{\mathbf{z}}(f_{\mathbf{z}}) + \lambda \mathcal{P}_\zeta(f_{\mathbf{z}}) \leq \mathcal{E}_{\mathbf{z}}(0) + 0 \leq M^2.$$

Moreover, we have

$$\lambda\, \zeta \|f_{\mathbf{z}}\|_{\ell^2}^2 \leq \lambda \mathcal{P}_\zeta(f_{\mathbf{z}}) \leq \mathcal{E}_{\mathbf{z}}(f_{\mathbf{z}}) + \lambda \mathcal{P}_\zeta(f_{\mathbf{z}}) \leq \mathcal{E}_{\mathbf{z}}(0) + 0 \leq M^2.$$

Hence we obtain the desired upper bounds. □

For $R > 0$, define $\mathcal{W}_1(R) = \left\{\mathbf{z} \in Z^m : \|\beta_{\mathbf{z}}\|_{\ell^1} \leq R\right\}$ and $\mathcal{W}_2(R) = \left\{\mathbf{z} \in Z^m : \zeta \|\beta_{\mathbf{z}}\|_{\ell^2}^2 \leq R\right\}$. Apparently, Lemma 3 asserts that $\mathcal{W}_1(\frac{M^2}{\lambda}) = \mathcal{W}_2(\frac{M^2}{\lambda}) = Z^m$. Following the definition of $\mathcal{D}(\lambda)$, we know that

$$\lambda \mathcal{P}_\zeta(f_\lambda) \leq \mathcal{E}(f_\lambda) - \mathcal{E}(f_\rho) + \lambda \mathcal{P}_\zeta(f_\lambda) \leq \mathcal{D}(\lambda).$$

Hence, $\|\beta\|_{\ell^1} \leq \frac{\mathcal{D}(\lambda)}{\lambda}$ and $\|\beta\|_{\ell^2} \leq \sqrt{\frac{\mathcal{D}(\lambda)}{\lambda\zeta}}$. Due to the definition of $f_\lambda$, it is reasonable to expect that $\|\beta_{\mathbf{z}}\|_{\ell^1}$ and $\|\beta_{\mathbf{z}}\|_{\ell^2}$ can be bounded by $\frac{\mathcal{D}(\lambda)}{\lambda}$ and $\sqrt{\frac{\mathcal{D}(\lambda)}{\lambda\zeta}}$, which is sharper compared with $\frac{M^2}{\lambda}$. This is realized via an iteration technique presented in the following proposition.

**Proposition 5.** *Assume that Assumption 1 and Regularity Assumption hold, let $f_{\mathbf{z}}$ given by (1.2). For any $0 < \delta < 1$, take $\zeta = \lambda^\theta$, $\lambda = m^{-\gamma}$ with*

$$\begin{cases} 0 < \gamma < \dfrac{1}{2+p}, & \text{if } \theta \in (1-\tau, +\infty), \\ 0 < \gamma < \dfrac{2}{(2+p)(1+\theta)}, & \text{if } \theta \in [\tau - 1, 1-\tau]. \end{cases}$$

*Then with confidence $1 - \delta$ there holds*

$$\max\left\{\|\beta_{\mathbf{z}}\|_{\ell^1}, \zeta \|\beta_{\mathbf{z}}\|_{\ell^2}^2\right\} \leq 2^{2J_0} C_0 \left(\log\left(2J_0/\delta\right)\right)^{2J_0} m^{\gamma(1-\tau)},$$

*where $J_0$ and $C_0$ are positive constants that independent of $m$ or $\delta$, which are explicitly given in (3.8) and (3.9).*

**Proof.** Following the excess generalization bound in (2.6), there is a set $V_R \subset Z^m$ with $\rho(V_R) \leq \delta$ such that

$$\lambda \|\beta_{\mathbf{z}}\|_{\ell^1} + \lambda\zeta \|\beta_{\mathbf{z}}\|_{\ell^2}^2 = \lambda \mathcal{P}_\zeta(f_{\mathbf{z}}) \leq C \log\left(\frac{2}{\delta}\right)\left(\mathcal{D}(\lambda) + \frac{1}{m} + \frac{R_\zeta^2}{m^{2/(2+p)}} + \frac{r_\zeta^2}{m}\right).$$

Hence it is easy to see that

$$\begin{aligned} \max\left\{\|\beta_{\mathbf{z}}\|_{\ell^1}, \zeta \|\beta_{\mathbf{z}}\|_{\ell^2}^2\right\} &\leq \frac{C \log\left(\frac{2}{\delta}\right)}{m^{2/(2+p)}\lambda} R_\zeta^2 + C \log\left(\frac{2}{\delta}\right)\left(\frac{\mathcal{D}(\lambda)}{\lambda} + \frac{1}{m\lambda} + \frac{r_\zeta^2}{m\lambda}\right) \\ &= a_{m,\lambda} R_\zeta^2 + b_{m,\lambda}, \end{aligned} \tag{3.1}$$

where

$$a_{m,\lambda} = \frac{C \log\left(\frac{2}{\delta}\right)}{m^{2/(2+p)}\lambda} \tag{3.2}$$

and

$$b_{m,\lambda} = C \log\left(\frac{2}{\delta}\right)\left(\frac{\mathcal{D}(\lambda)}{\lambda} + \frac{1}{m\lambda} + \frac{r_\zeta^2}{m\lambda}\right). \tag{3.3}$$

Recall that $R_\zeta = \min\left\{R, \sqrt{\frac{R}{\zeta}}\right\}$. We derive our estimates in the following two cases.

We first consider the case $R_\zeta \le R$. Then following (3.1) we know that

$$\max\left\{\|\beta_{\mathbf{z}}\|_{\ell^1}, \zeta\|\beta_{\mathbf{z}}\|_{\ell^2}^2\right\} \le a_{m,\lambda}R^2 + b_{m,\lambda},$$

which yields

$$\mathcal{W}_1(R) \subset \mathcal{W}_1\left(\max\left\{2a_{m,\lambda}R^2, 2b_{m,\lambda}\right\}\right) \cup V_R. \tag{3.4}$$

Defining a sequence of radii $\{R^{(j)}\}_{j\in\mathbb{N}}$ with $R^{(0)} = M^2/\lambda$ and applying the inclusion (3.4) to the sequence, we have

$$R^{(j)} = \max\left\{2a_{m,\lambda}(R^{(j-1)})^2, 2b_{m,\lambda}\right\}, \quad j \in \mathbb{N}. \tag{3.5}$$

According to (3.4), for each $R^{(j)}$ there holds $\mathcal{W}(R^{(j-1)}) \subset \mathcal{W}(R^{(j)}) \cup V_{R^{(j-1)}}$ with $\rho(V_R^{(j-1)}) \le \delta$. Applying this inclusion for $j = 1, 2, \ldots, J$ with $J$ to be determined later, we get

$$Z^m = \mathcal{W}_1(R^{(0)}) \subset \mathcal{W}_1(R^{(1)}) \cup V_{R^{(0)}} \subset \cdots \subset \mathcal{W}_1(R^{(J)}) \cup \left(\cup_{j=0}^{J-1} V_{R^{(j)}}\right).$$

It follows that the measure of the set $\mathcal{W}_1(R^{(J)})$ is at least $1 - J\delta$.

By the definition of sequence $\{R^{(j)}\}_{j\in\mathbb{N}}$ and (3.5), we see that

$$R^{(J)} = \max\left\{(a_{m,\lambda})^{1+2+2^2+\cdots+2^{J-1}}(R^{(0)})^{2^J}, (a_{m,\lambda})^{1+2+2^2+\cdots+2^{J-2}}(b_{m,\lambda})^{2^{J-1}}, \ldots, a_{m,\lambda}b_{m,\lambda}^2, b_{m,\lambda}\right\}.$$

Following (3.2), the first term on the right-hand side can be bounded by

$$(a_{m,\lambda})^{1+2+2^2+\cdots+2^{J-1}}(R^{(0)})^{2^J} \le \left(C\log\left(\frac{2}{\delta}\right)\right)^{2^J-1} M^{2^J+1}m^{-\frac{2}{2+p}(2^J-1)}\lambda^{1-2^{J+1}}.$$

For remainder terms, we have

$$\begin{aligned}\max\left\{(a_{m,\lambda})^{1+2+2^2+\cdots+2^{J-2}}(b_{m,\lambda})^{2^{J-1}}, \ldots, a_{m,\lambda}b_{m,\lambda}^2, b_{m,\lambda}\right\} &= b_{m,\lambda}\max\left\{(a_{m,\lambda}b_{m,\lambda})^{1+2+2^2+\cdots+2^{J-2}}, \ldots, a_{m,\lambda}b_{m,\lambda}, 1\right\}\\ &= b_{m,\lambda}\max\left\{1, (a_{m,\lambda}b_{m,\lambda})^{2^{J-1}-1}\right\}.\end{aligned}$$

Following (3.2) and (3.3), the following equality holds

$$(a_{m,\lambda}b_{m,\lambda})^{2^{J-1}-1} = \left(C(2C_1+2)\log\left(\frac{2}{\delta}\right)\right)^{2^J-2}\left(\lambda^{-1}m^{-\frac{2}{2+p}}\lambda^{\tau-1}\right)^{2^{J-1}-1}.$$

Next, we choose $\lambda = m^{-\gamma}$ with $\gamma > 0$ by restricting

$$2\gamma - \frac{2}{2+p} < 0 \quad \text{and} \quad (2-\tau)\gamma - \frac{2}{2+p} \le 0$$

and we choose $J$ as the smallest integer satisfying

$$J = J_1 \ge \max\left\{\frac{\log(2-\gamma(2+p)) - \log(2-2\gamma(2+p))}{\log 2}, 1\right\}.$$

With choices of $\gamma$ and $J = J_1$, and the fact that the measure of the set $\mathcal{W}_1(R^{(J)})$ is at least $1 - J_1\delta$, we see that with confidence at least $1 - J_1\delta$, there holds

$$R^{(J)} \le (C(1+C_1))^{2^{J_1}-1}(2\log(2/\delta))^{2^{J_1}-1}m^{\gamma(1-\tau)}. \tag{3.6}$$

Next, we consider the case when $R_\zeta \le \sqrt{\frac{R}{\zeta}}$. According to (3.1), we have

$$\max\left\{\|\beta_{\mathbf{z}}\|_{\ell^1}, \zeta\|\beta_{\mathbf{z}}\|_{\ell^2}^2\right\} \le \frac{a_{m,\lambda}}{\zeta}R + b_{m,\lambda}.$$

Applying similar iteration process and taking $\zeta = \lambda^\theta$, $\lambda = m^{-\gamma}$ where $\gamma$ satisfying the restriction

$$\gamma < \frac{2}{(2+p)(1+\theta)} \quad \text{and} \quad J_2 = \left\lceil \frac{\gamma(1+\theta)}{\frac{2}{2+p} - \gamma(1+\theta)} \right\rceil,$$

it follows that with confidence at least $1 - J_2\delta$, there holds

$$R^{(J)} \leq (C_1 + 1 + M^2)C^{J_2}\,(2\log(2/\delta))^{J_2}\, m^{\gamma(1-\tau)}. \tag{3.7}$$

Choose

$$C_0 = \max\left\{(C_1 + 1 + M^2)C^{J_2}, (C(1 + C_1))^{2^{J_1}-1}\right\} \tag{3.8}$$

and $J_0 = 2\max\{J_1, J_2\}$, that is, the smallest integer satisfying

$$J_0 \geq \max\left\{\frac{\log(2 - \gamma(2+p)) - \log(2 - 2\gamma(2+p))}{\log 2}, \frac{\gamma(1+\theta)}{\frac{2}{2+p} - \gamma(1+\theta)}, 1\right\}. \tag{3.9}$$

Combining (3.6) and (3.7), we complete our proof by scaling $J_0\delta$ to $\delta$. □

## 4. Elastic-net learning with infinite features

In this section, we show that algorithm (1.2) can also be extended to the infinite-feature case, where the hypothesis space is composed by infinite features. For this infinite-feature case, we assume that the dictionary $\Gamma = \mathbb{N}$ and for a constant $\kappa > 0$, there holds

$$\sum_{k=1}^{+\infty} |\varphi_k(x)|^2 \leq \kappa, \quad \forall x \in X. \tag{4.1}$$

To better understand this assumption, we provide an example in the following which can also be found in [4].

**Example.** Take $X = [0, 1]$. Let $\{\psi_{jk} | j = 0, 1\cdots; k \in \Delta_j\}$ be an orthonormal wavelet basis in $L^2([0, 1])$ with regularity $C^r$, $r > \frac{1}{2}$, where for $j \geq 1$, $\{\psi_{jk} | k \in \Delta_j\}$ is the orthonormal wavelet basis(with suitable boundary conditions) spanning the detail space at level $j$. It is assumed that the set $\{\psi_{0k} : k \in \Delta_0\}$ contains both the wavelets and the scaling functions at level $j = 0$. Fix $s$ such that $\frac{1}{2} < s < r$ and let $\phi_{jk} = 2^{-js}\psi_{j,k}$. Then

$$\sum_{j=0}^{\infty}\sum_{k\in\Delta_j} |\phi_{jk}|^2 = \sum_{j=0}^{\infty}\sum_{k\in\Delta_j} 2^{-2js}|\psi_{jk}|^2 \leq C\sum_{j=0}^{\infty} 2^{-2js}2^j = C\frac{1}{1 - 2^{1-2s}} = \kappa$$

where $C$ is a suitable constant depending on the number of wavelets that are non-zero at a point $x \in [0, 1]$ for a given level $j$, and on the maximum values of the scaling function and of the mother wavelet.

The hypothesis space constituted by infinite features, where we denote as $\mathcal{H}_0$, is given by

$$\mathcal{H}_0 = \left\{f : f = \sum_{\gamma=1}^{+\infty} \beta_\gamma\varphi_\gamma, \{\beta_\gamma\}_{\gamma=1}^{+\infty} \in \ell_2\right\}.$$

Under assumption (4.1), the construction of $\mathcal{H}_0$ ensures the boundedness of $f_\beta(x)$, for every $f_\beta \in \mathcal{H}_0$. To control both the sparsity and regularity of the regression function $f_\rho$, the following assumption is further introduced.

**Assumption 3.** We assume that the regression function $f_\rho = f_\beta$ for some $\beta \in \ell_2$ with $\sum_{\gamma=1}^{+\infty} |\beta_\gamma| < +\infty$.

Now our elastic net regularization learning scheme with infinite features turns to the following form.

$$f_{\mathbf{z}} = \arg\min_{f_\beta\in\mathcal{H}_0}\left(\frac{1}{m}\sum_{i=1}^{m}(f_\beta(x_i) - y_i)^2 + \lambda\mathcal{P}_\zeta(f_\beta)\right). \tag{4.2}$$

At a first glance, one has to solve optimization problem (4.2) in an infinite dimensional space $\mathcal{H}_0$, which is computationally infeasible. In fact, this is trackable in the sense that the support of $f_{\mathbf{z}}$ is included in a set $\Gamma_\lambda$ and can be finite under proper conditions. To explain this, we denote $\mathrm{supp}(f_\beta) = \{\gamma \in \Gamma : \beta_\gamma \neq 0, \text{ where } f_\beta = \sum_{\gamma=1}^{+\infty}\beta_\gamma\varphi_\gamma\}$, for $f_\beta \in \mathcal{H}_0$.

**Theorem 3.** *Given $\zeta > 0$ and $\lambda > 0$, let $f_{\mathbf{z}}$ produced by (4.2), denote*

$$\Gamma_\lambda = \left\{\gamma \in \Gamma : \frac{1}{m}\sum_{i=1}^{m}|\varphi_\gamma(x_i)|^2 \neq 0, \sqrt{\frac{1}{m}\sum_{i=1}^{m}|\varphi_\gamma(x_i)|^2} \geq \frac{\lambda}{2M} - \sqrt{\zeta\lambda}\right\},$$

*then* $\mathrm{supp}(f_{\mathbf{z}}) \subset \Gamma_\lambda$. *Moreover, $\Gamma_\lambda$ is finite when $\lambda > 4M^2\zeta$ and* $\lim_{\gamma\to+\infty}\|\varphi_\gamma\|_\infty = 0$.

**Proof.** Recalling the definition of $\mathcal{E}_{\mathbf{z}}(\cdot)$, we know that $f_{\mathbf{z}}$ is a minimizer if and only if $\mathbf{0} \in \partial\left(\mathcal{E}_{\mathbf{z}}(f_{\mathbf{z}}) + \lambda\mathcal{P}_{\zeta}(f_{\mathbf{z}})\right)$. Denote $f_{\mathbf{z}} = \sum_{\gamma=1}^{+\infty} \beta_{\gamma}^{\mathbf{z}}\varphi_{\gamma}$ and rewrite algorithm (4.2) into its coefficient-based form, one gets

$$(\beta_{\gamma}^{\mathbf{z}})_{\gamma\in\Gamma} = \arg\min_{(\beta_{\gamma})_{\gamma\in\Gamma}\in\ell^2}\left(\frac{1}{m}\sum_{i=1}^{m}\left(\sum_{\gamma=1}^{+\infty}\beta_{\gamma}\varphi_{\gamma}(x_i) - y_i\right)^2 + \lambda\sum_{\gamma=1}^{+\infty}(|\beta_{\gamma}| + \zeta\beta_{\gamma}^2)\right),$$

this in connection with the condition $\mathbf{0} \in \partial\left(\mathcal{E}_{\mathbf{z}}(f_{\mathbf{z}}) + \lambda\mathcal{P}_{\zeta}(f_{\mathbf{z}})\right)$ implies that

$$\lambda\,\mathrm{sgn}(\beta_{\gamma}^{\mathbf{z}}) = \frac{2}{m}\sum_{i=1}^{m}\varphi_{\gamma}(x_i)\left(y_i - \sum_{\gamma'=1}^{+\infty}\beta_{\gamma'}^{\mathbf{z}}\varphi_{\gamma'}(x_i)\right) - 2\lambda\zeta\beta_{\gamma}^{\mathbf{z}},$$

where the definition of $\mathrm{sgn}(t)$ is given by

$$\begin{cases}\mathrm{sgn}(t) = 1, & \text{if } t > 0;\\ \mathrm{sgn}(t) = 0, & \text{if } t = 0;\\ \mathrm{sgn}(t) = -1, & \text{if } t < 0.\end{cases}$$

Following the definition of $f_{\mathbf{z}}$, one has

$$\mathcal{E}_{\mathbf{z}}(f_{\mathbf{z}}) \le \mathcal{E}_{\mathbf{z}}(f_{\mathbf{z}}) + \lambda\mathcal{P}_{\zeta}(f_{\mathbf{z}}) \le \mathcal{E}_{\mathbf{z}}(0) + 0 \le M^2,$$

and

$$\lambda\mathcal{P}_{\zeta}(f_{\mathbf{z}}) \le \mathcal{E}_{\mathbf{z}}(f_{\mathbf{z}}) + \lambda\mathcal{P}_{\zeta}(f_{\mathbf{z}}) \le \mathcal{E}_{\mathbf{z}}(0) + 0 \le M^2.$$

Using Hölder's inequality and combining two upper bounds above, one gets

$$\begin{aligned}|\lambda\,\mathrm{sgn}(\beta_{\gamma}^{\mathbf{z}})| &= \left|\frac{2}{m}\sum_{i=1}^{m}\varphi_{\gamma}(x_i)\left(y_i - \sum_{\gamma'=1}^{+\infty}\beta_{\gamma'}^{\mathbf{z}}\varphi_{\gamma'}(x_i)\right) - 2\lambda\zeta\beta_{\gamma}^{\mathbf{z}}\right|\\ &\le 2\sqrt{\frac{1}{m}\sum_{i=1}^{m}\varphi_{\gamma}(x_i)^2}\sqrt{\frac{1}{m}\sum_{i=1}^{m}(y_i - f_{\mathbf{z}}(x_i))^2} + 2\lambda\zeta|\beta_{\gamma}^{\mathbf{z}}|\\ &\le 2M\left(\sqrt{\frac{1}{m}\sum_{i=1}^{m}\varphi_{\gamma}(x_i)^2} + \sqrt{\lambda\zeta}\right),\end{aligned}$$

therefore, we have

$$|\mathrm{sgn}(\beta_{\gamma}^{\mathbf{z}})| \le \frac{2M}{\lambda}\left(\sqrt{\frac{1}{m}\sum_{i=1}^{m}\varphi_{\gamma}(x_i)^2} + \sqrt{\lambda\zeta}\right).$$

Hence following the definition of the sign function, we have

$$|\mathrm{sgn}(\beta_{\gamma}^{\mathbf{z}})| = 1, \quad \text{if } \beta_{\gamma}^{\mathbf{z}} \ne 0,$$

and

$$\beta_{\gamma}^{\mathbf{z}} = 0, \quad \text{if } \frac{2M}{\lambda}\left(\sqrt{\frac{1}{m}\sum_{i=1}^{m}\varphi_{\gamma}(x_i)^2} + \sqrt{\lambda\zeta}\right) < 1.$$

When $\lambda > 4M^2\zeta$ and $\lim_{\gamma\to+\infty}\|\varphi_{\gamma}\|_{\infty} = 0$, it is easy to see that $\Gamma_{\lambda}$ is finite. This completes the proof. □

As illustrated in Theorem 3, the support of $f_{\mathbf{z}}$ is included in $\Gamma_{\lambda}$ and under proper conditions $\Gamma_{\lambda}$ can be a finite set, which has immediate computational implications. To illustrate this, denote the cardinality of $\Gamma_{\lambda}$ as $N_0$ and it is easy to see that in this case algorithm (4.2) turns to the following form

$$f_{\mathbf{z}} = \arg\min_{f_{\beta}\in\mathcal{H}_{N_0}}\left(\frac{1}{m}\sum_{i=1}^{m}(f_{\beta}(x_i) - y_i)^2 + \lambda\mathcal{P}_{\zeta}(f_{\beta})\right), \tag{4.3}$$

where $\mathcal{H}_{N_0}$ is a subset of $\mathcal{H}_0$ with finite dimension. Hence, algorithm (4.2) is trackable in the sense that it is equivalent to (4.3) under proper conditions. Following analogous analyzing process as we did in previous sections, one can easily show the consistency of algorithm (4.2). Moreover, under Assumptions 1 and 3, with properly chosen parameters, it can be proved that learning rates in Theorem 1 can also be achieved for algorithm (4.2).

## 5. Proofs of main results and discussion

### 5.1. Proofs of main results

**Proof of Theorem 1.** As claimed in Proposition 5, by taking $\zeta = \lambda^\theta, \lambda = m^{-\gamma}$ with constraints on $\gamma$ satisfied, for any $0 < \delta < 1$, with confidence $1-\delta$ there holds

$$\max\left\{\|\beta_{\mathbf{z}}\|_{\ell^1}, \zeta\|\beta_{\mathbf{z}}\|_{\ell^2}^2\right\} \le 2^{2J_0} C_0 \left(\log\left(2J_0/\delta\right)\right)^{2J_0} m^{\gamma(1-\tau)}.$$

This combined with the definition of $R_\zeta$ yields

$$R_\zeta = 2^{2J_0} C_0 (\log(2J_0/\delta))^{2J_0} \min\left\{\lambda^{\tau-1}, \sqrt{\frac{\lambda^{\tau-1}}{\zeta}}\right\}.$$

Together with (2.6), $\|f_{\mathbf{z}} - f_\rho\|_{L^2_{\rho_X}}^2$ can be bounded by

$$\left(C + C_0^2 2^{2J_0+1}\right)\left(\log\left(\frac{2J_0}{\delta}\right)\right)^{2J_0+1}\left(\lambda^\tau + \frac{1}{m} + \min\left\{\lambda^{2\tau-2}, \frac{\lambda^{\tau-1}}{\zeta}\right\}\left(\frac{1}{m}\right)^{2/(2+p)}\right).$$

With simple computations when $\theta$ belongs to corresponding interval and recalling the constraints on $\gamma$, we complete the proof of Theorem 1. □

**Proof of Corollary 1.** Considering that $p$ is sufficiently small and $\tau$ is sufficiently close to 1, when $\theta \in (\varepsilon, +\infty)$, we choose $p = \frac{2\varepsilon}{1-\varepsilon}$, $\tau = 1-\varepsilon$. Together with the choice of $\gamma$, Theorem 1 implies the desired learning rate. Moreover, when $\theta \in [-\varepsilon, \varepsilon]$, we choose the same $p$ and $\tau$. Analogously, following the choice of $\gamma$ and applying Theorem 1, we obtain the desired learning rate. Hence this completes the proof. □

### 5.2. Toward learning rates and the role that $\zeta$ plays

In this subsection, we consider the learning rate and the role that $\zeta$ plays. As shown in Corollary 1, when $\zeta$ is small enough, the learning rate is of type $\mathcal{O}(m^{\varepsilon-\frac{1}{2}})$ with arbitrary small $\varepsilon$. By choosing appropriate $\zeta$, the learning rate we obtained is of type $\mathcal{O}(m^{\varepsilon-1})$ with arbitrary small $\varepsilon$. That is, with a proper chosen $\zeta$, the convexity is enhanced and the convergence rate is also promoted, compared with lasso-type algorithms.

We can further explain this by revisiting the evaluation process of the sample error. Note that we use $r_\zeta$ and $R_\zeta$ to bound $\|f_\lambda\|_\infty$ and $\|f_{\mathbf{z}}\|_\infty$ separately. Choose $\zeta = \zeta_0$ such that $\frac{\mathcal{D}(\lambda)}{\lambda} = \sqrt{\frac{\mathcal{D}(\lambda)}{\lambda\zeta_0}}$. Then Corollary 1 reveals that when $\zeta < \zeta_0$, we use $\frac{\mathcal{D}(\lambda)}{\lambda}$ to bound $\|f_{\mathbf{z}}\|_\infty$ and $\|f_\lambda\|_\infty$ while the upper bound $\sqrt{\frac{\mathcal{D}(\lambda)}{\lambda\zeta}}$ is utilized when $\zeta \ge \zeta_0$. It means that algorithm (1.2) searches for empirical target function $f_{\mathbf{z}}$ in different balls when $\zeta$ is chosen with different values. The tuning parameter $\zeta$ in fact controls the switch between the $\ell^1$-ball and $\ell^2$-ball in $\mathcal{H}_\Gamma$, with critical point $\zeta = \zeta_0$. This provides a theoretical characterization on the role that $\zeta$ plays and hence guides the choice of tuning parameter in practice.

When considering the learning rate, some theoretical analysis on elastic-net regularization has been presented in [4]. Based on a nonlinear operator approach, they conduct their analysis and present the learning rates. As claimed in [4], under mild conditions one of the optimal learning rates they obtain is of type $\mathcal{O}(m^{-\frac{1}{2}})$. Apparently, our learning rate is always faster. In fact, this is a natural result of our approach since we conduct error analysis by taking into account the complexity of the hypothesis space. To understand this, it is necessary to note that two approaches, capacity independent approach and capacity dependent approach, are frequently used to derive learning rates in learning theory literature. For the capacity independent approach, one may refer to [12,13]. For the capacity dependent approach, there are a vast of learning theory literatures [14,15,10,16,17,8,18]. As we know, the capacity dependent approach reflects the regularity of functions in hypothesis space while the capacity independent approach neglects such information. Hence, it is reasonable that our learning rates are faster.

Recently, [10] investigates the kernel-based lasso-type algorithms within a statistical learning framework, which is mainly concerned with generalization bounds. Learning rates of type $\mathcal{O}(m^{\varepsilon-\frac{1}{2}})$ are obtained in terms of the regularity of the kernel and the regression function $f_\rho$ and the capacity of the hypothesis space. Comparing with the lasso-type algorithms in [10], the elastic net regularization algorithm we study in this paper, which is initially aiming at producing group effect, is essentially different as explained in the Introduction and also in [3,19]. As shown in [3], the strongly convex regularizer has a group effect while the lasso-type $\ell_1$ regularizer does not. It is easy to see that the algorithm we study in this paper is also a lasso-type algorithm by removing the term including $\zeta$ in the regularizer. In fact, according to Corollary 1, the learning rate we obtain is of type $\mathcal{O}(m^{\varepsilon-\frac{1}{2}})$ when $\zeta$ is small enough, which coincides with that in [10] while analysis in this paper is conducted based on a more general setting [4]. Besides, we also study the role that the parameter $\zeta$ plays by analyzing the

generalization bounds when $\zeta$ is not too small. It is worth mentioning that error analysis both in [10] and in our paper are conducted in a standard framework proposed in [15,20].

We end this paper with two remarks on the sparse and group sparse property of elastic net.

**Remark 4.** We say that the empirical target function $f_{\mathbf{z}}$ produced by elastic-net algorithm is sparse in the sense that the support of $f_{\mathbf{z}}$ is explicitly given and under proper conditions the support set is finite, as illustrated in Theorem 3. The elastic-net regularizer implies the sparse property. It is an interesting problem to explore some other criteria to characterize the sparseness of elastic-net algorithms, which is beyond the scope of this paper.

**Remark 5.** As illustrated in [3], when a group of highly correlated variables are selected, elastic net regularization performs well in the sense that the regression coefficients of such variables tend to be equal. One may also expect that algorithm (1.2) preserves the group sparse property with respect to the features $\{\varphi_k\}_{k\in\Gamma}$, since the group effect benefits from the strictly convexity of the regularizer. Further explanation will be investigated in our future work.

## Acknowledgments

The authors would like to thank Dr. Quan-Wu Xiao and Mr. Jun Fan for helpful discussions and anonymous referees for their valuable suggestions which helped improve this paper.

## References

[1] A. Hoerl, R. Kennard, Ridge regression: biased estimation for nonorthogonal problems, Technometrics 12 (1) (1970) 55–67.
[2] R. Tibshirani, Regression shrinkage and selection via the lasso, J. R. Stat. Soc. Ser. B 58 (1996) 267–288.
[3] H. Zou, T. Hastie, Regularization and variable selection via the elastic net, J. R. Stat. Soc. Ser. B 67 (2005) 301–320.
[4] C. De Mol, E. De Vito, L. Rosasco, Elastic-net regularization in learning theory, J. Complexity 25 (2) (2009) 201–230.
[5] C. De Mol, S. Mosci, M. Traskine, A. Verri, A regularized method for selecting nested groups of relevant genes from microarray data, J. Comput. Biol. 16 (2009) 677–690.
[6] A. Destrero, S. Mosci, C. De Mol, A. Verri, F. Odone, Feature selection for high dimensional data, Comput. Manag. Sci. 6 (2009) 25–40.
[7] A. Wibisono, L. Rosasco, T. Poggio, Sufficient conditions for uniform stability of regularized algorithms, Technical Report, 2009.
[8] Y.L. Feng, S.G. Lv, Unified approach to coefficient-based regularized regression, Comput. Math. Appl. 62 (2011) 506–515.
[9] Q. Wu, Y.M. Ying, D.X. Zhou, Multi-kernel regularized classifier, J. Complexity 23 (2007) 108–134.
[10] L. Shi, Y.L. Feng, D.X. Zhou, Concentration estimates for learning with $\ell^1$-regularizer and data dependent hypothesis spaces, Appl. Comput. Harmon. Anal. 31 (2011) 286–302.
[11] D.X. Zhou, The covering number in learning theory, J. Complexity 18 (2002) 739–767.
[12] L. Rasasco, M. Belkin, On learning with integral operators, J. Mach. Learn. Res. (2010) 905–934.
[13] S. Smale, D.X. Zhou, Learning theory estimates via integral operator and their approximations, Constr. Approx. 26 (2007) 152–172.
[14] A. Caponnetto, E. De Vito, Optimal rates for regularized least squares algorithm, Found. Comput. Math. 7 (2007) 331–368.
[15] F. Cucker, S. Smale, On the mathematical fundations of learning, Bull. Amer. Math. Soc. 39 (2001) 1–49.
[16] I. Steinwart, D. Hush, C. Scovel, Optimal rates for regularized least square regression, in: Proceedings of 22nd Annual Conference on Learning Theory, 2009.
[17] I. Steinwart, A. Christmann, Support Vector Machines, Springer-Verlag, New York, 2008.
[18] P. Bartlett, S. Mendelson, Rademacher and Gaussian complexities: risk bounds and structural results, J. Mach. Learn. Res. 3 (2002) 463–482.
[19] T.C. Hesterberg, N.H. Choi, L. Meier, C. Fraley, Least angle and $\ell_1$ penalized regression: a review, Statist. Surv. 2 (2008) 61–93.
[20] F. Cucker, D.X. Zhou, Learning Theory: An Approximation Theory Viewpoint, Cambridge University Press, 2004.
[21] S. Smale, D.X. Zhou, Estimating the approximation error in learning theory, Anal. Appl. 1 (2003) 17–41.
[22] Q. Wu, Y.M. Ying, D.X. Zhou, Learning rates of least-square regularized regression, Found. Comput. Math. 6 (2) (2006) 171–192.