LETTER Communicated by Bin Zou

Kernelized Elastic Net Regularization: Generalization Bounds, and Sparse Recovery

Yunlong Feng
yunlong.feng@esat.kuleuven.be
Department of Electrical Engineering, ESAT-STADIUS, KU Leuven 3000, Belgium

Shao-Gao Lv
lvsg716@swufe.edu.cn
Statistics School, Southwestern University of Finance and Economics, ChengDu 611130, China

Hanyuan Hang
Hanyuan.Hang@esat.kuleuven.be
Johan A. K. Suykens
johan.suykens@esat.kuleuven.be
Department of Electrical Engineering, ESAT-STADIUS, KU Leuven 3000, Belgium

Kernelized elastic net regularization (KENReg) is a kernelization of the well-known elastic net regularization (Zou & Hastie, 2005). The kernel in KENReg is not required to be a Mercer kernel since it learns from a kernelized dictionary in the coefficient space. Feng, Yang, Zhao, Lv, and Suykens (2014) showed that KENReg has some nice properties including stability, sparseness, and generalization. In this letter, we continue our study on KENReg by conducting a refined learning theory analysis. This letter makes the following three main contributions. First, we present refined error analysis on the generalization performance of KENReg. The main difficulty of analyzing the generalization error of KENReg lies in characterizing the population version of its empirical target function. We overcome this by introducing a weighted Banach space associated with the elastic net regularization. We are then able to conduct elaborated learning theory analysis and obtain fast convergence rates under proper complexity and regularity assumptions. Second, we study the sparse recovery problem in KENReg with fixed design and show that the kernelization may improve the sparse recovery ability compared to the classical elastic net regularization. Finally, we discuss the interplay among different properties of KENReg that include sparseness, stability, and generalization. We show that the stability of KENReg leads to generalization, and its sparseness confidence can be derived from generalization. Moreover, KENReg is stable and can be simultaneously sparse, which makes it attractive theoretically and practically.

Neural Computation **28, 525–562** (2016)
doi:10.1162/NECO_a_00812
© 2016 Massachusetts Institute of Technology

1 Introduction

Kernels and related kernel-based learning paradigms have become popular tools in the machine learning community owing to their great successes in theoretical interpretation and computational efficiency. The key idea is to implicitly map observations from the input space to some feature space where simple algorithms can be simultaneously delivered—the so-called kernel trick. Based on the kernels, a variety of learning machines have been developed in both supervised and unsupervised learning. Several canonical examples are the kernel principal component analysis in unsupervised learning, support vector machine for classification (SVC), and support vector regression (SVR) in supervised learning.

In this letter, we focus on the supervised learning problem. Specifically, we are interested in learning for the regression problem, the main goal of which is to infer a functional relation between input and output that gives good prediction performance on future observations. To be more specific, suppose that $\mathcal{X}$ and $\mathcal{Y}$ are the explanatory variable space and response variable space, respectively, with explanatory variable X and response variable Y. We assume that $\mathcal{X} \subset \mathbb{R}^d$ is a compact metric space, and the observations are generated by

$$y = f^\star(x) + \varepsilon,$$

where ε is the additive zero-mean noise. Suppose that we are given a set of observations $\mathbf{z} = \{(x_i, y_i)\}_{i=1}^m$ that are independent and identically distributed (i.i.d.) copies of $X \times Y$ drawn from an unknown probability distribution ρ. In the statistical learning literature, there are two typical settings according to the different forms of the given design points $x_1, \ldots, x_m$. When $x_1, \ldots, x_m$ are randomly drawn from the marginal distribution of ρ, the setting is termed the random design setting, in contrast to the fixed design setting, where $x_1, \ldots, x_m$ are fixed and only the responses $y_1, \ldots, y_m$ are treated as random. The purpose of regression is to approximate the ground truth $f^\star$ in some function space based on given observations $\mathbf{z}$ with random or fixed design.

In the statistical learning literature, kernel-based learning models for regression have been extensively studied. One form can be expressed by the following generic formula,

$$f_{\mathbf{z}} = \arg\min_{f \in \mathcal{H}_{\mathcal{K}}} \mathcal{E}_{\mathbf{z}}^{\ell}(f) + \lambda \|f\|_{\mathcal{K}}^2, \tag{1.1}$$

where $\mathcal{E}_{\mathbf{z}}^{\ell}(\cdot)$ stands for the empirical risk term associated with a loss function $\ell : \mathbb{R} \to \mathbb{R}$, $\mathcal{H}_{\mathcal{K}}$ is a reproducing kernel Hilbert space (RKHS) induced by a Mercer kernel $\mathcal{K} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, and $\lambda > 0$ is a regularization parameter.

Depending on different choices of the loss function ℓ, equation 1.1 has been termed as different learning machines. For example, being equipped with an ϵ-insensitive loss, this equation gives the formula for SVR. The representer theorem ensures that the optimal solution to equation 1.1 admits the presentation

$$f_{\mathbf{z},\lambda}(x) = \sum_{i=1}^{m} \beta_i \mathcal{K}(x_i, x), \quad x \in \mathcal{X}. \tag{1.2}$$

Besides being amenable to analysis and computationally efficient, another property that the learning schemes enjoy is that $f_{\mathbf{z}}$ in equation 1.2 is sparse. As a result, only a fraction of the observations contribute to the final output, which is crucial for pursuing a parsimonious model. In this sense, the above two kernel-based learning models can be cast as an instance-selection learning machine. This also explains the terminology: *support vector*. Here the sparseness is delivered by the special mechanism of the loss function.

However, in some practical regression problems, it might be the case that the output predictor $f_{\mathbf{z}}$ learning from SVR is not sufficiently sparse (Drezet & Harrison, 1998). This might be improved to meet the requirement of pursuing a parsimonious model in SVR by choosing a wider zero zone in the loss function. Nevertheless, with such choice of loss functions, the generalization ability might also be weakened. As an alternative, recently another kernel-based sparsity-producing learning model, kernelized dictionary learning model, has been extensively studied in the statistical learning literature, in which sparseness is delivered via the penalty term instead of the loss function. Mathematically, the kernelized dictionary learning model takes the form

$$\boldsymbol{\alpha}_{\mathbf{z}} = \arg\min_{\boldsymbol{\alpha} \in \mathbb{R}^m} \mathcal{R}_{\mathbf{z}}^{\ell}(\boldsymbol{\alpha}) + \lambda \Omega(\boldsymbol{\alpha}), \tag{1.3}$$

where $\mathcal{R}_{\mathbf{z}}^{\ell}(\cdot)$ denotes the empirical risk term associated with a loss function ℓ and the kernelized dictionary $\{\mathcal{K}(x, x_i)\}_{i=1}^{m}$, $\Omega(\cdot)$ is the penalty term and λ is a positive regularization parameter. Notice that there is no representer theorem for equation 1.3. However, due to the utilization of the kernelized dictionary, it is a finite-dimensional optimization problem, and the output predictor also admits the same form as equation 1.2.

It can be seen from equation 1.3 that the penalization is applied directly to the coefficients. This differentiates the learning scheme, equation 1.3, from previous kernel-based ones and also brings more flexibility. A notable one is that the kernel $\mathcal{K}$ in equation 1.3 is merely a continuous function and does not necessarily need to be a Mercer kernel. This makes the learning

machine, equation 1.3, more applicable considering that the Mercer condition on the kernel in some cases could be cumbersome to verify. For more detailed explanations on the flexibility of learning with kernelized dictionary, see the short review in Feng, Yang, Zhao, Lv, and Suykens (2014). Note that the learning paradigm, equation 1.3, belongs to the category of learning with kernels in the coefficient space that was introduced in Schölkopf and Smola (2001); and Suykens, Van Gestel, De Brabanter, De Moor, and Vandewalle (2002) and recently has been extensively studied empirically and theoretically in Roth (2004), Wu and Zhou (2005, 2008), Wang, Yeung, and Lochovsky (2007), Wu and Zhou (2008), Xiao and Zhou (2010), Sun and Wu (2011), Shi, Feng, and Zhou (2011), Tong, Chen, and Yang (2010), Lin, Zeng, Fang, and Xu (2014), Chen, Pan, Li, and Tang (2013), and Feng et al. (2014). On the other hand, sparseness can be easily delivered in equation 1.3 by choosing corresponding penalty terms. A frequent choice of sparsity-producing penalty term is $\Omega(\boldsymbol{\alpha}) = \|\boldsymbol{\alpha}\|_1$, associated with which equation 1.3 is termed the kernelized Lasso. However, Xu, Caramanis, and Mannor (2012) and Feng et al. (2014) have argued that the kernelized Lasso is not stable from both a computational and an algorithmic viewpoint. Alternatively, Feng et al. (2014) proposed a stabilized regularization scheme by introducing an additional ℓ^2-norm (with respect to the coefficient) to the kernelized Lasso, namely, the kernelized elastic net regularization (KENReg) model, which can be stated as

$$\text{(KENReg)} \quad \boldsymbol{\alpha}_{\mathbf{z}} = \arg\min_{\boldsymbol{\alpha}\in\mathbb{R}^m} \frac{1}{m}\|\mathbf{y} - \mathcal{K}\boldsymbol{\alpha}\|_2^2 + \lambda_1\|\boldsymbol{\alpha}\|_1 + \lambda_2\|\boldsymbol{\alpha}\|_2^2, \tag{1.4}$$

where $\boldsymbol{\alpha}_{\mathbf{z}} = (\alpha_{\mathbf{z},1}, \ldots, \alpha_{\mathbf{z},m})^\top$, $\mathbf{y} = (y_1, \ldots, y_m)^\top$, $\mathcal{K}$ denotes the $m \times m$ matrix with entries $\mathcal{K}_{i,j} = \mathcal{K}(x_i, x_j)$ for $i, j = 1, \ldots, m$, and $\lambda_1, \lambda_2 > 0$ are regularization parameters. Obviously KENReg learns in a finite-dimensional space, and the corresponding empirical target function can be expressed as

$$f_{\mathbf{z}}(x) = \sum_{i=1}^{m} \alpha_{\mathbf{z},i}\,\mathcal{K}(x_i, x), \quad x \in \mathcal{X}. \tag{1.5}$$

Theoretically, Feng et al. (2014) showed that the KENReg model has advantages over kernelized Lasso when taking the stability and characterizable sparsity into account. Meanwhile, the generalization ability that indicates the consistency property as well as the convergence rates of KENReg have been derived.

1.1 Objectives of the Letter. As a continuation of our previous study on KENReg, the main scope of this letter is to study the following aspects:

generalization ability, sparse recovery ability, and the interplay among sparseness, stability, and generalization.

First, generalization ability is one of the major concerns when designing a learning machine and also contributes to the major theme of statistical learning theory. With the random design setting, in the generalization ability analysis, one is concerned with the error term $\mathbb{E}(f_{\mathbf{z}} - f^{\star})^2$, where the expectation is taken with respect to the observations over the marginal distribution of ρ on $\mathcal{X}$. The generalization ability of KENReg has been analyzed in Feng et al. (2014) by introducing an instrumental empirical target function, following the path of Wu and Zhou (2008).

However, as Feng et al. (2014), remarked, the generalization bounds provided there are in fact derived with respect to a truncated version of the empirical target function instead of $f_{\mathbf{z}}$ itself, which is a compromise since the analysis on the truncated version can be significantly simplified. On the other hand, according to the generalization analysis given in Feng et al. (2014), to ensure the convergence of KENReg, the regularization parameter λ_2 is required to go to zero when the sample size m tends to infinity. Nevertheless, the stability arguments in Feng et al. (2014) show that to ensure the convergence of KENReg, λ_2 should go to infinity in accordance with m. This contradiction is again caused by the fact that the analysis on the convergence rates of KENReg in Feng et al. (2014) was carried out with respect to a truncated version of $f_{\mathbf{z}}$.

In fact, the generalization analysis that Feng et al. (2014) presented cannot be easily tailored to $f_{\mathbf{z}}$ in equation 1.5. As Feng et al. (2014) detailed and also as we explain in section 3, from a function approximation point of view, the main difficulties encountered in the analysis lie in defining the hypothesis space and characterizing the approximation error associated with the regularization parameters λ_1 and λ_2. Given that the analysis of the generalization bounds plays an important role in studying a learning machine, the first major concern of this letter is to present a refined analysis on the generalization bounds of KENReg with respect to the empirical target function $f_{\mathbf{z}}$ itself.

Next, we look at sparse recovery ability. KENReg advocates sparse solutions due to the use of a sparsity-producing penalty term. Following the literature on sparse recovery (Candès & Tao, 2005; Candès, Romberg, & Tao, 2006; Zhang, 2011), it is natural to ask whether KENReg can identify the zero pattern of the true solution if it is assumed to be sparse in some sense. This is the second major concern in this work, and a positive answer is presented in section 4 with the fixed design setting.

Finally, we look at the interplay among sparseness, stability, and generalization. Sparseness and stability are two motivating properties for which KENReg is proposed, as Feng et al. (2014) stated. These properties, together with its generalization ability, make KENReg attractive for regression. However, it is generally thought that sparse learning algorithms cannot be stable. Moreover, it is not clear that whether in learning with the kernelized

dictionary setting the algorithmic stability can also lead to generalization. And we are also concerned with the relation between generalization and sparseness. Thus, the third purpose of this letter is to illustrate the interplay among these different and important properties of KENReg.

This letter is organized as follows. In section 2, we present results on generalization bounds. Section 3 is dedicated to outlining the steps and main ideas in doing error analysis. We study the sparse recovery ability of the KENReg model in section 4. Section 5 presents discussion on the interplay among the three different properties of the KENReg model: sparseness, stability, and generalization. We end this letter by summarizing contributions in section 6. Proofs of propositions and theorems are provided in the appendix if they are not presented immediately in the following sections.

2 Preliminaries and Main Results on Generalization Bounds

This section presents our main results on generalization bounds, which refer to the convergence rates of $f_{\mathbf{z}}$ to the regression function f_ρ with $f_\rho(x) = \mathbb{E}(y \mid x)$ for any $x \in \mathcal{X}$ under the $\mathcal{L}^2_{\rho_{\mathcal{X}}}$-metric with random design. Here the regression function f_ρ is in fact $f^\star$ due to the zero-mean noise assumption. To this end, we first discuss difficulties encountered in error analysis and introduce some notation and assumptions.

2.1 Difficulties in Error Analysis and Proposed Method. KENReg learns with the kernelized dictionary $\{\mathcal{K}(x_i, x)\}_{i=1}^m$, which depends on the data $\{x_i\}_{i=1}^m$. Therefore, the hypothesis space of KENReg is drifting with the varying observations $\mathbf{z}$. This leads to the so-called hypothesis error, which also contributes to the variance of the estimator $f_{\mathbf{z}}$. On the other hand, without a fixed hypothesis space, one is not able to characterize the population version of $f_{\mathbf{z}}$, and this makes it difficult to conduct an error analysis via the classical bias-variance trade-off approach. As Feng et al. (2014) commented, besides the varying observations $\mathbf{z}$, such difficulty also stems from the combined penalty term in KENReg and the stepping-stone technique employed there to bypass it.

In this letter, we overcome this difficulty by first constructing an instrumental hypothesis space to help characterize the population version of $f_{\mathbf{z}}$. We then construct an instrumental empirical target function that mimics $f_{\mathbf{z}}$ well and is easier to analyze. The last step is to conduct an error analysis with respect to the newly constructed instrumental empirical target function. This process is illustrated in Figure 1.

In Figure 1, $\mathcal{F}_w$ is the instrumental hypothesis space that is constructed to characterize the population version of $f_{\mathbf{z}}$: f_{w,λ_1}. $g_{\mathbf{x},\lambda_1}$ is an instrumental empirical target function that has a similar generalization performance as $f_{\mathbf{z}}$. g_{w,λ_1} is the population version of $g_{\mathbf{x},\lambda_1}$. $\mathcal{H}_w$ is another space that is closely related to $\mathcal{F}_w$ and is introduced to include both $g_{\mathbf{x},\lambda_1}$ and g_{w,λ_1}. In our analysis,

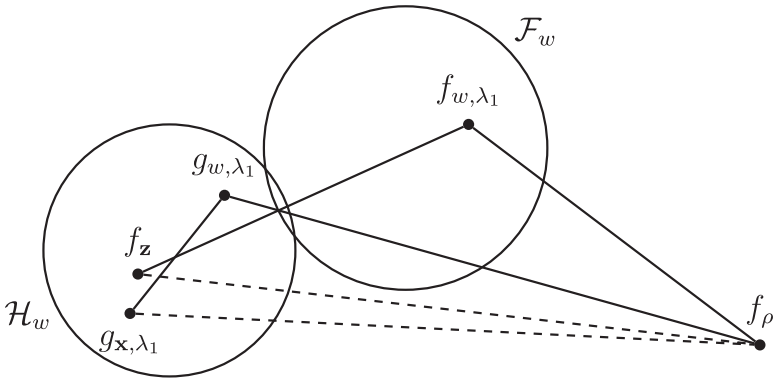

Figure 1: An illustration on the idea of generalization error analysis in our study.

instead of bounding the distance between $f_{\mathbf{z}}$ and f_ρ directly, we bound the distance between $g_{\mathbf{x},\lambda_1}$ and f_ρ and show that this can upper-bound the previous one. Rigorous definitions of these spaces and functions are given below. Here, it should be mentioned that the instrumental empirical target function $g_{\mathbf{x},\lambda_1}$ essentially plays a stepping-stone role; this technique was proposed by Wu and Zhou (2005, 2008) and employed in much follow-up work (see Shi, 2013; Tong et al., 2010; and Chen et al., 2013).

2.2 Construction of the Instrumental Hypothesis Space and the Instrumental Empirical Target Function. For any function f defined on $\mathcal{X}$, the usual $\mathcal{L}^q$-norm is defined as $\|f\|^q_{\mathcal{L}^q_{\rho_\mathcal{X}}} = \int_{\rho_\mathcal{X}} |f(x)|^q d\rho_\mathcal{X}(x)$ for any $1 \le q < \infty$. For notational simplicity, we denote $\|f\|_q = \|f\|_{\mathcal{L}^q_{\rho_\mathcal{X}}}$. The instrumental hypothesis space we construct is the following Banach function space with a weighted norm,

$$\mathcal{F}_w = \left\{ f \in \mathcal{L}^2_{\rho_\mathcal{X}}, \quad \text{with norm } \|f\|_{\mathcal{F}_w} = \sqrt{\|f\|_1^2 + w\|f\|_2^2} \right\},$$

where w is a weight parameter related to the penalty parameter λ_2 that will be specified below. For any bounded kernel function $\mathcal{K} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, we define the integral operator $L_\mathcal{K} : \mathcal{F}_w \to \mathcal{F}_w$ as follows:

$$L_\mathcal{K} f(x) = \int_\mathcal{X} \mathcal{K}(x,t) f(t) d\rho_\mathcal{X}(t), \quad \forall x \in \mathcal{X}.$$

It is easily seen from the boundedness of $\mathcal{K}$ that the operator $L_\mathcal{K}$ is well defined on $\mathcal{F}_w$. With this notation, we define the range of $L_\mathcal{K}$ as another

Banach space,

$$\mathcal{H}_w := \left\{ g : \, g(\cdot) = L_{\mathcal{K}} f(\cdot) \text{ for some } f \in \mathcal{F}_w \right\},$$

where the norm $\|\cdot\|_{\mathcal{H}_w}$ is given by

$$\|g\|_{\mathcal{H}_w} = \inf \left\{ \|f\|_{\mathcal{F}_w}, \text{ where } g(x) = L_{\mathcal{K}} f(x), x \in \mathcal{X} \right\}.$$

Obviously $\mathcal{H}_w$ can be embedded into some subset of $C(\mathcal{X})$ following the continuity of $\mathcal{K}$.

For ease of error analysis, we now introduce the regularization function that is associated with parameter w:

$$f_{w,\lambda_1} = \arg\min_{f \in \mathcal{F}_w} \{ \|L_{\mathcal{K}} f - f_\rho\|_2^2 + \lambda_1 \|f\|_{\mathcal{F}_w}^2 \}, \quad \lambda_1, \, w > 0. \tag{2.1}$$

The instrumental empirical target function $g_{\mathbf{x},\lambda_1}$ is constructed as follows

$$g_{\mathbf{x},\lambda_1}(x) := \frac{1}{m} \sum_{i=1}^{m} f_{w,\lambda_1}(x_i) \mathcal{K}(x_i, x). \tag{2.2}$$

To characterize the approximation error, we also denote $\mathcal{D}_w(\lambda_1)$ as

$$\mathcal{D}_w(\lambda_1) = \inf_{g \in \mathcal{H}_w} \{ \|g - f_\rho\|_2^2 + \lambda_1 \|g\|_{\mathcal{H}_w}^2 \}, \quad \lambda_1 > 0. \tag{2.3}$$

To characterize the population version of $g_{\mathbf{x},\lambda_1}$, we also denote

$$g_{w,\lambda_1} = \arg\min_{g \in \mathcal{H}_w} \|g - f_\rho\|_2^2 + \lambda_1 \|g\|_{\mathcal{H}_w}^2. \tag{2.4}$$

It is easy to see that $\mathcal{D}_w(\lambda_1)$ represents the approximation ability of the function space $\mathcal{H}_w$ to the regression function f_ρ. The functional in equation 2.4 is strictly convex, and $\mathcal{H}_w$ is a reflexive Banach space for $w > 0$. From the classical functional analysis, we know the existence of g_{w,λ_1} in equation 2.4. Moreover, there exists a function $f_{w,\lambda_1} \in \mathcal{F}_w$, such that $g_{w,\lambda_1} = L_{\mathcal{K}} f_{w,\lambda_1}$ satisfying $\|g_{w,\lambda_1}\|_{\mathcal{H}_w} = \|f_{w,\lambda_1}\|_{\mathcal{F}_w}$.

2.3 Assumptions and Generalization Bounds. In order to state the main results on generalization bounds, we introduce several assumptions. The generalization bounds in our study are derived by controlling the capacity of the hypothesis space that KENReg works in. Its hypothesis

space is spanned by the kernelized dictionary $\{\mathcal{K}(\cdot, x_i)\}_{i=1}^m$, and its capacity assumption is given as follows:

Assumption 1 (Capacity Assumption). There exists a positive constant p with $0 < p < 2$ such that

$$\log \mathcal{N}_\infty(B_1, \epsilon) \leq c_p \epsilon^{-p}, \quad \forall \epsilon > 0, \tag{2.5}$$

where c_p is a positive constant independent of ϵ, and

$$B_R = \left\{ f = \sum_{i=1}^m \alpha_i \mathcal{K}(\cdot, u_i), \quad \sum_{i=1}^m |\alpha_i| \leq R,\ u_i \in \mathcal{X} \right\}, \text{ for } R > 0,$$

and $\mathcal{N}_\infty(B_R, \epsilon)$ denotes the covering number of B_R, that is, the smallest integer $\ell \in \mathbb{N}$ such that there exist ℓ disks in $\mathcal{C}(\mathcal{X})$ with radius ϵ and centers in B_R covering B_R.

Shi et al. (2011) showed that the above capacity assumption holds under certain regularity conditions on the kernel $\mathcal{K}$. This indicates that the behavior (see equation 2.5) is typical because the case of the hypothesis space induced by the gaussian kernel can be included. Assumption 2 is on the boundedness of the response variable Y, which is again a canonical assumption in the statistical learning theory literature (Cucker & Zhou, 2007; Steinwart & Christmann, 2008; Hang & Steinwart, 2014; Lv, 2015) and also applied in Feng et al. (2014).

Assumption 2 (Boundedness Assumption). Assume that $|y| \leq M$, $\sup_{x,x' \in \mathcal{X}} \mathcal{K}(x, x') \leq \kappa$, for some $M, \kappa > 0$ and without loss of generality we let $M = 1$ and $\kappa = 1$.

Following the assumption 2, it is easy to see that $|f_\rho(x)| \leq 1$. As Feng et al. (2014) remarked, assumption 2 can be easily relaxed to the assumption that the response variable has a subgaussian tail without leading to essential difficulties in analysis. Moreover, by replacing the least-square loss with some robust loss functions, it can be further relaxed to the assumption that the response variable satisfies certain moment conditions as shown in Huber and Ronchetti (2009), Vapnik (1998), Györfi, Kohler, Krzyżak, and Walk (2002), and Feng, Huang, Shi, Yang, and Suykens (2015).

Assumption 3 (Model Assumption). There exist positive constants τ with $0 < \tau \leq 1$ and c_τ, such that for any $\lambda_1 > 0$, there holds

$$\mathcal{D}_w(\lambda_1) \leq c_\tau \lambda_1^\tau.$$

This model assumption specifies the model approximation ability. In fact, from the definition of $\mathcal{D}_w(\lambda_1)$, it is easy to see that it depends on

the regularity of f_ρ. For instance, when $f_\rho \in \mathcal{H}_w$, there holds $\tau = 1$. More discussions on the model assumption are provided in section 3.2. Now we are ready to state our main results on the generalization bounds.

Theorem 1. *Suppose that $\mathcal{X}$ is a compact subset of $\mathbb{R}^n$. Assume that the boundedness assumption, the capacity assumption, and the model assumption hold. For any $0 < \delta < 1$, with confidence $1 - 4\delta$, there holds*

$$\left\| f_{\mathbf{z}} - f_\rho \right\|_2^2 = C_{\mathcal{K},\rho} \log^2(2/\delta) m^{-\Theta},$$

where

$$\Theta = \begin{cases} \min\left\{ \frac{2\tau}{3p+2-2p\tau}, 1 - \frac{2(\gamma-\tau)}{3p+2-2p\tau} \right\}, & \text{if } 2 - p \geq (3p+2)\tau, \\ \min\left\{ \frac{2\tau}{(4+\tau)p+2\tau}, 1 - \frac{2(\gamma-\tau)}{(4+\tau)p+2\tau} \right\}, & \text{otherwise,} \end{cases}$$

and the above rate is derived by choosing

$$\begin{cases} \lambda_1 = m^{-\frac{2}{3p+2-2p\tau}}, \ \lambda_2 = m\lambda_1^\gamma & \text{if } 2 - p \geq (3p+2)\tau, \\ \lambda_1 = m^{-\frac{2}{(4+\tau)p+2\tau}}, \ \lambda_2 = m\lambda_1^\gamma, & \text{otherwise,} \end{cases}$$

with the parameters $0 < \gamma \leq 2\tau$ and $C_{\mathcal{K},\rho}$ a positive constant independent of m, δ or τ.

As a consequence of theorem 1, we can present the convergence rates of KENReg in the following corollary in a more explicit manner when the regression function f_ρ is smooth enough.

Corollary 1. *Suppose that the assumptions of theorem 1 hold and additionally we assume that $\tau = 1$. For any $0 < \delta < 1$, with confidence $1 - 4\delta$, there holds*

$$\left\| f_{\mathbf{z}} - f_\rho \right\|_2^2 = C_{\mathcal{K},\rho} \log^2(2/\delta) m^{-\frac{2}{5p+2}},$$

where the choice of the parameter pair is $(\lambda_1, \lambda_2) = \left(m^{-2/(5p+2)}, m^{5p/(5p+2)}\right)$.

Proofs of theorem 1 and corollary 1 are provided in the appendix. Here we give several remarks:

- The convergence rates stated above indicate their dependence on the regularization parameter λ_2 and further confirm the involvement of the ℓ^2-term in generalization. They are also optimal in the sense that when p tends to zero, they can be arbitrarily close to $\mathcal{O}(m^{-1})$, which is regarded as the fastest learning rate in the learning theory literature. On the other hand, the convergence rates are conducted with respect to $f_{\mathbf{z}}$ instead of its projected version as presented in Feng et al. (2014).

Noticing that $f_{\mathbf{z}}$ is exactly the empirical target function of interest in our study, in this sense we say that the two types of convergence rates are of a different nature, and refined generalization bounds are presented in this study.

- The convergence rates in corollary 1 are obtained under the condition that λ_2 goes to infinity when the sample size m tends to infinity. As we will detail, this is consistent with the observation on generalization made from the stability arguments in Feng et al. (2014).
- In theorem 1 and corollary 1, the regularization parameters λ_1 and λ_2 are selected to achieve the theoretical convergence rates by balancing between the bias and the variance terms. In practice, they are more frequently chosen by using data-driven techniques (e.g., cross-validation). To reduce the computational burden, a heuristic approach to selecting the two parameters can be conducted as follows: one first chooses λ_2 via cross-validation and sets λ_1 to zero; with fixed λ_2, one can then carry out cross-validation again to determine an appropriate λ_1. (We refer readers to Feng et al., 2014, for more detailed discussion on the model selection problem and numerical studies.)

3 Generalization Error Analysis

This section presents the main analysis in deriving the generalization bounds given in section 2. In the literature of learning with kernelized dictionary, the generalization error consists of the sample error, approximation error, and hypothesis error.

3.1 Decomposing Generalization Error into Bias-Variance Terms. For any function $f : \mathcal{X} \to \mathbb{R}$, denote $\mathcal{E}_{\mathbf{z}}(f)$ and $\mathcal{E}(f)$ as its empirical risk and expected risk under the least squares loss that are given by

$$\mathcal{E}_{\mathbf{z}}(f) = \frac{1}{m}\sum_{i=1}^{m}(y_i - f(x_i))^2, \quad \text{and} \quad \mathcal{E}(f) = \int_{\mathcal{X}\times\mathcal{Y}}(y - f(x))^2 d\rho(x, y).$$

In what follows, we set the weight $w = \lambda_2/(\lambda_1 m)$ and $\lambda_2 = m\lambda_1^{\gamma}$. The following error decomposition splits the generalization error into the above-mentioned three parts: sample error, approximation error, and hypothesis error.

Proposition 1. *Let $f_{\mathbf{z}}$ be produced by equation 1.5, $g_{\mathbf{x},\lambda_1}$ be given by equation 2.2, and g_{w,λ_1} be defined as in equation 2.4. Denote $\Omega(g_{\mathbf{x},\lambda_1}) = \lambda_1\|g_{\mathbf{x},\lambda_1}\|_1 + \lambda_2\|g_{\mathbf{x},\lambda_1}\|_2^2$, where the norms $\|\cdot\|_1$ and $\|\cdot\|_2$ are defined with respect to the coefficients of $g_{\mathbf{x},\lambda_1}$. For any $w > 0$, the following error decomposition holds,*

$$\|f_{\mathbf{z}} - f_\rho\|_2^2 \le \mathbf{T}_1 + \mathbf{T}_2 + \mathbf{T}_3 + \mathbf{T}_4 + \mathbf{T}_5,$$

where

$$\begin{aligned}
\mathbf{T}_1 &:= \mathcal{E}(f_{\mathbf{z}}) - \mathcal{E}(f_\rho) - (\mathcal{E}_{\mathbf{z}}(f_{\mathbf{z}}) - \mathcal{E}_{\mathbf{z}}(f_\rho)),\\
\mathbf{T}_2 &:= \mathcal{E}_{\mathbf{z}}(g_{\mathbf{x},\lambda_1}) - \mathcal{E}_{\mathbf{z}}(f_\rho) - (\mathcal{E}(g_{\mathbf{x},\lambda_1}) - \mathcal{E}(f_\rho)),\\
\mathbf{T}_3 &:= \Omega(g_{\mathbf{x},\lambda_1}) - \lambda_1 \|g_{w,\lambda_1}\|^2_{\mathcal{H}_w},\\
\mathbf{T}_4 &:= \mathcal{E}(g_{\mathbf{x},\lambda_1}) - \mathcal{E}(g_{w,\lambda_1}),\\
\mathbf{T}_5 &:= \mathcal{D}_w(\lambda_1).
\end{aligned}$$

$\mathbf{T}_1$ and $\mathbf{T}_2$ are called the sample error, which are caused by randomly sampling. Notice that the estimation of $\mathbf{T}_1$ involves $f_{\mathbf{z}}$, which varies with respect to the observations $\mathbf{z}$; thus, we need concentration techniques from empirical process theory for bounding it. The same observation can be also applied to $\mathbf{T}_2$ due to the randomness of $g_{\mathbf{x},\lambda_1}$. $\mathbf{T}_3$ and $\mathbf{T}_4$ are called the hypothesis error since $g_{\mathbf{x},\lambda_1}$ and g_{w,λ_1} may lie in different hypothesis spaces with $f_{\mathbf{z}}$ with the varying observations $\mathbf{z}$. $\mathbf{T}_5$ stands for the approximation error that corresponds to the variance term and is independent of randomized sampling. According to the error decomposition in proposition 1, to bound the generalization error of KENReg, it suffices to bound $\mathbf{T}_1$, $\mathbf{T}_2$,$\mathbf{T}_3$, $\mathbf{T}_4$, and $\mathbf{T}_5$, respectively.

3.2 Approximation Error. The approximation error $\mathcal{D}_w(\lambda)$ reflects the approximation ability of the hypothesis space $\mathcal{H}_w$ to the underlying ground truth f_ρ. The model assumption introduced in section 2.3 assumes that for any $\lambda > 0$, $\mathcal{D}_w(\lambda)$ is of polynomial order with respect to λ. In this section, by using techniques introduced in Xiao and Zhou (2010), we will show that the model assumption is typical when a certain regularity assumption on the regression function f_ρ holds.

In the learning theory literature, to characterize the regularity of the regression function f_ρ, it is usually assumed that f_ρ belongs to the range of a compact, symmetric, and positive-definite linear operator on $\mathcal{L}^2_{\rho_{\mathcal{X}}}$ that is associated with the kernel $\mathcal{K}$. Note that the kernel $\mathcal{K}$ in our study is not necessarily positive or symmetric. Xiao and Zhou (2010) noticed that a positive-definite kernel $\widetilde{\mathcal{K}} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ can be constructed from $\mathcal{K}$ as follows:

$$\widetilde{\mathcal{K}}(u, v) = \int_{\mathcal{X}} \mathcal{K}(u, x)\mathcal{K}(v, x) d\rho_{\mathcal{X}}(x), \quad (u, v) \in \mathcal{X} \times \mathcal{X}.$$

Consequently a positive-definite integral operator $L_{\widetilde{\mathcal{K}}} = L_{\mathcal{K}} L^T_{\mathcal{K}}$ can be defined. Due to the compactness of $\mathcal{X}$ and the continuity of $\mathcal{K}$, the integral operator $L_{\widetilde{\mathcal{K}}}$, as well as its fractional power $L^r_{\widetilde{\mathcal{K}}}$, is compact and well defined

on $\mathcal{L}^2_{\rho_{\mathcal{X}}}$ for any $r > 0$. Based on the above notations, we come to the following conclusion:

Proposition 2. *Suppose that there exists $r > 0$ such that $f_\rho = L^r_{\widetilde{\mathcal{K}}} g$ for some $g \in \mathcal{L}^2_{\rho_{\mathcal{X}}}$. If $w \le c_0$ for some $c_0 > 0$, then there holds*

$$\mathcal{D}_w(\lambda) \le (1 + c_0)\|g\|_2^2 \lambda^{2r/3}.$$

Proof. We first denote $\{\sigma_k\}_{k\ge1}$ as eigenvalues of $L_{\widetilde{\mathcal{K}}}$ with $\sigma_1 \ge \sigma_2 \ge \ldots \ge 0$ and $\{\phi_k\}_{k\ge1}$ as the corresponding eigenfunctions. From the spectral theorem for a positive compact operator, we know that $\{\phi_k\}_{k\ge1}$ forms an orthogonal basis of $\mathcal{L}^2_{\rho_{\mathcal{X}}}$. Following the regularity assumption on f_ρ, one has

$$f_\rho = \sum_{k=1}^{\infty} \sigma_k^r \alpha_k \phi_k, \quad \text{and} \quad \|g\|_2^2 = \sum_{k=1}^{\infty} \alpha_k^2 < \infty. \tag{3.1}$$

We now bound $\mathcal{D}(\lambda)$ by considering the cases when λ lies in different intervals.

When $0 < \lambda \le \sigma_1^3$, then there exists $N \in \mathbb{N}$ such that $\sigma_{N+1} < \lambda^{1/3} \le \sigma_N$. Denoting $f_N = \sum_{k=1}^N \sigma_k^r \alpha_k \phi_k$, the fact that $\{\sigma_k\}_{k\ge1}$ are eigenvalues of $L_{\widetilde{\mathcal{K}}}$ implies

$$\begin{aligned} f_N &= \sum_{k=1}^{N} \sigma_k^{r-1} \alpha_k (\sigma_k \phi_k) = \sum_{i=1}^{N} \sigma_k^{r-1} \alpha_k (L_{\mathcal{K}} L_{\mathcal{K}}^T \phi_k) \\ &= L_{\mathcal{K}} \left(\sum_{k=1}^{N} \sigma_k^{r-1} \alpha_k (L_{\mathcal{K}}^T \phi_k) \right). \end{aligned}$$

This in connection with the definition of the norm $\|\cdot\|_{\mathcal{H}_w}$ tells us that

$$\begin{aligned} \|f_N\|_{\mathcal{H}_w}^2 &= \left\| \sum_{i=1}^{N} \sigma_k^{r-1} \alpha_k (L_{\mathcal{K}}^T \phi_k) \right\|_1^2 + w \left\| \sum_{k=1}^{N} \sigma_k^{r-1} \alpha_i (L_{\mathcal{K}}^T \phi_k) \right\|_2^2 \\ &\le \left(\sum_{k=1}^{N} \sigma_k^{r-3/2} \alpha_k \cdot \sqrt{\sigma_k} \right)^2 + w \left(\sum_{k=1}^{N} \sigma_k^{r-3/2} \alpha_k \cdot \sqrt{\sigma_k} \right)^2 \\ &\le (1 + c_0)\|g\|_2^2 \lambda^{2r/3-1}, \end{aligned}$$

where the first inequality is due to the Cauchy-Schwarz inequality and the fact that $\{\phi_k\}_{k\ge1}$ is an orthogonal basis of $\mathcal{L}^2_{\rho_{\mathcal{X}}}$, while the second inequality

is based on the assumption that $w \leq c_0$. On the other hand, for any $k > N$, $\sigma_k < \lambda^{1/3}$ yields

$$\|f_N - f_\rho\|_2^2 = \left\| \sum_{k>N} \sigma_k^r \alpha_k \phi_k \right\|_2^2 = \sum_{k>N} \sigma_k^{2r} \alpha_k^2 \leq \|g\|_2^2 \lambda^{2r/3}.$$

From the above estimates and the definition of $\mathcal{D}(\lambda)$, when $0 < \lambda \leq \sigma_1^3$, there holds

$$\mathcal{D}_w(\lambda) \leq \|f_N - f_\rho\|_2^2 + \lambda \|f_N\|_{\mathcal{H}_w}^2 \leq (1 + c_0)\|g\|_2^2 \lambda^{2r/3}.$$

When $\lambda > \sigma_1^3$, we choose $f = 0 \in \mathcal{H}_w$ and obtain

$$\mathcal{D}_w(\lambda) \leq \|f_\rho\|_2^2 = \sum_{\sigma_k > 0} \alpha_k^2 \sigma_k^{2r} \leq \|g\|_2^2 \lambda^{2r/3}.$$

By combining the above estimates, we accomplish the proof.

3.3 Bounding the Hypothesis Error Terms $\mathbf{T}_3$ and $\mathbf{T}_4$. $\mathbf{T}_3$ can be estimated by applying the classical one-sided Bernstein's inequality. However, the estimation of $\mathbf{T}_4$ involves function-space-valued random variables. Therefore, we need to introduce the following concentration inequality with values in a Hilbert space to estimate $\mathbf{T}_4$, which can be found in Pinelis (1994).

Lemma 1. *Let H be a Hilbert space and ξ be an independent random variable on Z with values in H. Assume that $\|\xi\|_H \leq \widetilde{M} < \infty$ almost surely. Denote $\widetilde{\sigma}^2 = \mathbb{E}(\|\xi\|_H^2)$. Let $\{z_i\}_{i=1}^m$ be an independent random sample from ρ. Then for any $0 < \delta < 1$, with confidence $1 - \delta/2$, there holds*

$$\left\| \frac{1}{m} \sum_{i=1}^m \xi(z_i) - \mathbb{E}(\xi) \right\|_H \leq \frac{2\widetilde{M} \log(2/\delta)}{m} + \sqrt{\frac{2\widetilde{\sigma}^2 \log(2/\delta)}{m}}.$$

The following estimates on $\mathbf{T}_3$ and $\mathbf{T}_4$ can be derived by applying lemma 1.

Proposition 3. *With the choice $w = \lambda_2/(\lambda_1 m)$, for any $0 < \delta < 1$, with confidence at least $1 - 3\delta$, there holds*

$$\mathbf{T}_3 \leq \frac{C_3 \mathcal{D}_w(\lambda_1) \log(2/\delta)}{\lambda_2} + C_3' \mathcal{D}_w(\lambda_1),$$

and

$$\mathbf{T}_4 \leq \frac{2\mathcal{D}_w(\lambda_1)\log(2/\delta)}{\lambda_2}\left(1+\frac{4\log(2/\delta)}{\lambda_2}\right)+\mathcal{D}_w(\lambda_1),$$

where C_3 and C_3' are positive constants independent of m or δ.

3.4 Bounding the Sample Error Terms $\mathbf{T}_1$ and $\mathbf{T}_2$. In this part, we bound the two sample error terms $\mathbf{T}_1$ and $\mathbf{T}_2$, which are more involved due to the dependence of $f_{\mathbf{z}}$ and $g_{\mathbf{x},\lambda_1}$ on the randomized observations $\mathbf{z}$. In the learning theory literature, this is typically done by applying concentration inequalities to empirical processes indexed by a class of functions and also by using the classical tools from empirical processes theory such as peeling and symmetrization techniques. The key idea is to show that the supremum of an empirical process is close enough to its expectation. A crucial step in the estimation is bounding the complexity of the function space that gives rise to empirical processes. To this end, in our study, we introduce the following local Rademacher complexity.

Let $\{\sigma_i\}_{i=1}^m$ be an i.i.d. sequence of Rademacher variables, and let $\{z_i\}_{i=1}^m$ be an i.i.d. sequence of random variables from $\mathcal{Z}$, drawn according to some distribution. Let $\mathcal{F}$ be a class of functions on $\mathcal{Z}$. Let $\mathbb{E}f^2$ be the variance of f^2 with respect to the probability distribution on $\mathcal{Z}$. For each $r>0$, define the Rademacher complexity on the function class $\mathcal{F}$ as

$$R_m(\mathcal{F};r)=\sup_{f\in\mathcal{F},\mathbb{E}f^2\leq r}\frac{1}{m}\left|\sum_{i=1}^m\sigma_i f(z_i)\right|,$$

and call an expression $\mathbb{E}_{z,\sigma}[R_m(\mathcal{F};r)]$ the local Rademacher average of the class $\mathcal{F}$.

It is known that the generalization error bound based on the global Rademacher complexity is of the order $O(\sqrt{1/n})$. In practice, however, the hypothesis selected by a learning algorithm usually performs better than the worst case and belongs to a more favorable subset of all the hypotheses. The advantage of using the local version of Rademacher average is that they can be considerably smaller than the global ones, and the distribution information is also taken into account compared with other complexity measurements. Therefore, we employ the local Rademacher complexity to measure the complexity of smaller subsets, which ultimately leads to sharper learning rates.

In general, a sub-root function is used as an upper bound for the local Rademacher complexity. A function $\psi:[0,\infty)\to[0,\infty)$ is sub-root if it is nonnegative, and nondecreasing and satisfies that $\psi(r)/\sqrt{r}$ is

nonincreasing. The following theorem is due to Bartlett, Bousquet, and Mendelson (2005) with minor changes.

Lemma 2. *Let $\mathcal{F}$ be a class of measurable, square integrable functions such that $\mathbb{E}f - f \leq b$ for all $f \in \mathcal{F}$. Let ψ be a sub-root function, A be some positive constant, and r^* be the unique solution to $\psi(r) = r/A$. Assume that*

$$\mathbb{E}[R_m(\mathcal{F}; r)] \leq \psi(r), \quad r \geq r^*.$$

Then for all $t > 0$ and all $K > A/7$, with probability at least $1 - e^{-t}$ there holds

$$\mathbb{E}f - \frac{1}{m}\sum_{i=1}^{m} f(z_i) \leq \frac{\mathbb{E}f^2}{K} + \frac{50K}{A^2}r^* + \frac{(K+9b)t}{m}, \quad f \in \mathcal{F}.$$

Lemma 2 tells us that to get better bounds for the empirical term, one needs to study properties of the fixed point r^* of a sub-root ψ. Although there does not exist a general method for choosing ψ, tight bounds for local Rademacher complexity have been established in various function spaces such as RKHSs. The following lemma provides a connection between the local Rademacher complexity and entropy integral, which is an immediate result from the proof of theorem A7 in Bartlett et al. (2005).

Lemma 3. *The local Rademacher complexity is upper-bounded as*

$$\mathbb{E}[R_m(f \in \mathcal{F}, \mathbb{E}f^2 \leq r)] \leq \mathbb{E}\left[\frac{A}{\sqrt{m}}\int_0^{\sqrt{2r}} \sqrt{\log \mathcal{N}_2(\mathcal{F}, u, d_m)}\right] du + \frac{1}{m},$$

where A is some constant and $\mathcal{N}_2(\mathcal{F}, u, d_m)$ is the covering number of $\mathcal{F}$ at radius u for the empirical ℓ_2 metric.

By making use of the relation between local Rademacher complexity and the covering number given by lemma 3, we can upper-bound the quantity r^* in lemma 2. Moreover, we come to the following upper bounds for $\mathbf{T}_1$ and $\mathbf{T}_2$ with the notation that $a \wedge b := \min\{a, b\}$ for $a, b \in \mathbb{R}$.

Proposition 4. *Assume that the boundedness assumption and the capacity assumption hold. For any $0 < \delta < 1$, with confidence $1 - \delta/2$, there holds*

$$\mathbf{T}_1 \leq \frac{1}{2}\left\|f_{\mathbf{z}} - f_\rho\right\|_2^2 + \frac{C_1 \log(2/\delta)}{m} + C_1'\left(\lambda_1^{-1} \bigwedge \sqrt{m\lambda_2^{-1}}\right)^{4p/(2+p)}\left(\frac{1}{m}\right)^{2/(2+p)},$$

where C_1 and C_1' are positive constants independent of m or δ.

Proposition 5. *Assume that the boundedness assumption and the capacity assumption hold. For any $0 < \delta < 1$, with confidence $1 - \delta/2$, there holds*

$$\mathbf{T}_2 \leq \frac{1}{2}\mathbf{T}_4 + \frac{1}{2}\mathcal{D}_w(\lambda_1) + C_4 \left\{ \left(\frac{m}{\lambda_2}\right)^{\frac{2(3p+2)}{p+2}} \mathcal{D}_w(\lambda_1)^{\frac{3p+2}{p+2}} \left(\frac{1}{m}\right)^{\frac{2}{p+2}} + \frac{m\mathcal{D}_w(\lambda_1)\log(2/\delta)}{(\lambda_2)^2} \right\},$$

where C_4 is a positive constant independent of m or δ.

3.5 Bounding the Local Rademacher Complexity: A By-Product. In learning theory analysis, to bound the generalization error, it is crucial to take into account the complexity of the hypothesis space. Various notions of complexity measurements have been employed; they include VC-dimension, covering number, Rademacher complexity, and local Rademacher complexity. One advantage of using local Rademacher complexity as the notion of complexity over the others is that it can be computed directly from the data (Bartlett et al. 2005). In our preceding analysis, we bounded the local Rademacher complexity by applying the relation in lemma 3. In fact, as a by-product, in the following proposition we provide another upper bound for the local Rademacher complexity when learning with the kernelized dictionary $\{\mathcal{K}(x_i, \cdot)\}_{i=1}^m$.

To this end, let $\mathbf{u} = \{u_i\}_{i=1}^m$ be any m-size i.i.d. copies of X drawn from $\rho_{\mathcal{X}}$. For any $x \in \mathcal{X}$, denote $K_{\mathbf{u}}(x) = (\mathcal{K}(u_1, x), \ldots, \mathcal{K}(u_m, x))^\top$. Let $\mathcal{F}$ be the function set given by

$$\mathcal{F} = \left\{ f \,\middle|\, f = \sum_{i=1}^m \alpha_i \mathcal{K}(x_i, \cdot), \boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_m)^\top \in \mathbb{R}^m, \|\boldsymbol{\alpha}\|_2 \leq 1 \right\}.$$

Proposition 6. *Let $\nu_{\mathbf{u},l}$ be the l-th largest eigenvalue of $\mathbb{E}_x[K_{\mathbf{u}}(x)K_{\mathbf{u}}(x)^\top]$ and assume that $\nu_{\mathbf{u},l} \leq \mu_l$ for $l = 1, \ldots, m$. Then there holds*

$$\mathbb{E}\left[R_m(f \in \mathcal{F}, \mathbb{E}f^2 \leq r)\right] \leq \frac{2}{\sqrt{m}} \left(\sum_{i=1}^m \min\{r, \mu_i\}\right)^{1/2}.$$

Proof. Let $e_{\mathbf{u},i}$ be the eigenfunction of $\mathbb{E}_x\left[K_{\mathbf{u}}(x)K_{\mathbf{u}}(x)^\top\right]$ that corresponds to $\nu_{\mathbf{u},i}$ and $\{e_{\mathbf{u},i}\}_{i=1}^m$ be an orthogonal basis of $\mathbb{R}^m$. For simplicity, we denote $e_i = e_{\mathbf{u},i}$, $\nu_i = \nu_{\mathbf{u},i}$, $i = 1, \ldots, m$, and further denote $X_\sigma = \frac{1}{m}\sum_{i=1}^m \sigma_i K_{\mathbf{u}}(x_i)$.

For any $h < m$, there holds

$$\begin{aligned}
&\frac{1}{m}\sum_{i=1}^{m}\sigma_i\boldsymbol{\alpha}^\top K_{\mathbf{u}}(x_i) = \left\langle \frac{1}{m}\sum_{i=1}^{m}\sigma_i K_{\mathbf{u}}(x_i), \boldsymbol{\alpha}\right\rangle \\
&= \left\langle \sum_{j=1}^{h}\nu_j^{-1/2}\langle X_\sigma, e_j\rangle e_j, \sum_{j=1}^{h}\sqrt{\nu_j}\langle\boldsymbol{\alpha}, e_j\rangle e_j\right\rangle + \left\langle \sum_{j>h}^{m}\langle X_\sigma, e_j\rangle e_j, \boldsymbol{\alpha}\right\rangle \qquad (3.2)\\
&\le \left\|\sum_{j=1}^{h}\nu_j^{-1/2}\langle X_\sigma, e_j\rangle e_j\right\| \cdot \left\|\sum_{j=1}^{h}\sqrt{\nu_j}\langle\boldsymbol{\alpha}, e_j\rangle e_j\right\| + \left\|\sum_{j>h}^{m}\langle X_\sigma, e_j\rangle e_j\right\| \cdot \|\boldsymbol{\alpha}\|.
\end{aligned}$$

Note that

$$\begin{aligned}
\mathbb{E}\left[\left\|\sum_{j=1}^{h}\nu_j^{-1/2}\langle X_\sigma, e_j\rangle e_j\right\|\right] &= \mathbb{E}\left(\sum_{j=1}^{h}\nu_j^{-1}\langle X_\sigma, e_j\rangle^2\right)^{1/2} \\
&\le \left(\sum_{j=1}^{h}\nu_j^{-1}\mathbb{E}[\langle X_\sigma, e_j\rangle^2]\right)^{1/2}, \qquad (3.3)
\end{aligned}$$

where the inequality follows from Jensen's inequality. Since σ_i's are independent random variables with zero mean and $\mathbb{E}_x[\langle K_{\mathbf{u}}(x), e_j\rangle^2] = \nu_j$, there holds

$$\begin{aligned}
\mathbb{E}[\langle X_\sigma, e_j\rangle^2] &= \frac{1}{m^2}\sum_{i,k=1}^{m}\mathbb{E}[\sigma_i\sigma_k\langle K_{\mathbf{u}}(x_i), e_j\rangle\langle K_{\mathbf{u}}(x_k), e_j\rangle] \\
&= \frac{1}{m}\mathbb{E}_x[\langle K_{\mathbf{u}}(x), e_j\rangle^2] = \frac{\nu_j}{m}.
\end{aligned}$$

This, in connection with equation 3.3, implies that

$$\mathbb{E}\left[\left\|\sum_{j=1}^{h}\nu_j^{-1/2}\langle X_\sigma, e_j\rangle e_j\right\|\right] \le \sqrt{\frac{h}{m}}.$$

On the other hand, we have

$$r \ge \langle\boldsymbol{\alpha}, \mathbb{E}_x[K_{\mathbf{u}}(x)K_{\mathbf{u}}(x)^\top]\boldsymbol{\alpha}\rangle = \sum_{i=1}^{m}\nu_j\langle\boldsymbol{\alpha}, e_i\rangle^2 \ge \left\|\sum_{j=1}^{h}\sqrt{\nu_j}\langle\boldsymbol{\alpha}, e_j\rangle e_j\right\|^2.$$

This in connection with equation 3.2, and Jensen's inequality implies that

$$\mathbb{E}\left(\frac{1}{m}\sum_{i=1}^{m}\sigma_i\boldsymbol{\alpha}^\top K_{\mathbf{u}}(x_i)\right) \le \frac{1}{\sqrt{m}}\min_{0\le h\le m}\left\{\sqrt{hr}+\sqrt{\sum_{j=h+1}^{m}\nu_j}\right\}.$$

Following the subadditivity of $\sqrt{\cdot}$ and taking the supremum on $\mathbf{u}$, we accomplish the proof.

4 Sparse Recovery via Kernelized Elastic Net Regularization

The learning theory estimates presented in section 2 state with overwhelming confidence that $f_{\mathbf{z}}$ can be a good estimator of the regression function. In this section, we focus on the inference aspect of the kernelized elastic net estimator $f_{\mathbf{z}}$ with specific emphasis on its sparse recovery ability.

In recent years, compressed sensing and related sparse recovery schemes have become hot research topics, along with the advent of big data. Essentially the main concern of these sparse recovery schemes is to what extent an algorithm can recover the underlying true signal, which (is assumed) can be sparsely presented in some sense. Given that $f_{\mathbf{z}}$ is also a sparse approximation estimator to f_ρ and being parallel to those sparse recovery schemes (Candès & Tao, 2005; Candès et al., 2006; Zhang, 2011), we now study the sparse recovery property of $f_{\mathbf{z}}$ by assuming that the f_ρ possesses a sparse representation or can be sparsely approximated.

We start with introducing some notations and assumptions. Recall that the regression model we study in this letter is given by

$$y = f_\rho(x) + \varepsilon,$$

where more generally in this section, we assume that $\varepsilon \sim N(0, \sigma^2)$. For the regression function, we assume that the following sparse representation assumption holds, which has also been employed in the machine learning literature (see Xu, Jin, Shen, & Zhu, 2015).

Assumption 4 (Sparse Representation Assumption). Let S (possibly unknown) be a subset of $\{1, \ldots, m\}$ with cardinality $s = |S| \ll m$. We assume that the regression function f_ρ has the following sparse representation:

$$f_\rho(x) = \sum_{i\in S}\alpha_{*,i}\mathcal{K}(x_i, x), \quad x \in \mathcal{X}.$$

In the above sparse representation assumption, it is assumed that f_ρ can be sparsely represented by the kernelized dictionary $\{\mathcal{K}(x, x_i)\}_{i=1}^m$, which depends on the design points $x_1, \ldots, x_m$. Therefore, to study the sparse

recovery ability of the KENReg estimator, in this section we adopt the fixed design setting in order to avoid the drifting ground truth. In fact, the fixed design setting has been commonly adopted in the sparse learning literature (Zhang, 2010; Huang & Zhang, 2010; Huang, Zhang, & Metaxas, 2011; Koltchinskii, 2011; Bach, 2013) and can be more practical in certain real-world applications (Koltchinskii, Sakhanenko, & Cai, 2007). In what follows, we denote $\boldsymbol{\alpha}_\star = (\alpha_{*,i}, \ldots, \alpha_{*,m})^\top$. For any $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_m)^\top \in \mathbb{R}^m$, we denote $\operatorname{supp}(\boldsymbol{\alpha})$ as the support set of $\boldsymbol{\alpha}$ with $\operatorname{supp}(\boldsymbol{\alpha}) = \{i : \alpha_i \neq 0\}$. Under the sparse representation assumption, the main concerns of sparse recovery that could also be tailored to our context are the following two aspects:

- The approximation ability of the solution of KENReg to $\boldsymbol{\alpha}_\star$, that is, the distance between $\boldsymbol{\alpha}_\mathbf{z}$ and $\boldsymbol{\alpha}_\star$ under a certain metric
- The sparse recovery ability of KENReg: under what conditions KENReg can correctly find the zero pattern of $\boldsymbol{\alpha}_\star$, that is, $\operatorname{supp}(\boldsymbol{\alpha}_\mathbf{z}) \subseteq S$

In fact, under assumption 4, if we denote $\mathcal{K}_x = (\mathcal{K}(x_1, x), \ldots, \mathcal{K}(x_m, x))^\top$ and further define the seminorm $\|\cdot\|_\mathcal{K}$ as

$$\|\boldsymbol{\alpha}_\mathbf{z} - \boldsymbol{\alpha}_\star\|_\mathcal{K}^2 := (\boldsymbol{\alpha}_\mathbf{z} - \boldsymbol{\alpha}_\star)^\top \mathcal{K}_x \mathcal{K}_x^\top (\boldsymbol{\alpha}_\mathbf{z} - \boldsymbol{\alpha}_\star),$$

it is easy to see that

$$\left\| f_\mathbf{z} - f_\rho \right\|_2^2 = \mathbb{E}\|\boldsymbol{\alpha}_\mathbf{z} - \boldsymbol{\alpha}_\star\|_\mathcal{K}^2,$$

where the expectation is taking over the randomized sampling from $\rho_\mathcal{X}$. Therefore, for the more general randomized setting, the approximation ability of the solution of KENReg to $\boldsymbol{\alpha}_\star$ has been studied in section 2. With respect to the second aspect, the following results make some effort to address the sparse recovery property of $\boldsymbol{\alpha}_\mathbf{z}$.

Assumption 5 (Kernelized Irrepresentable Assumption). There exists a constant $0 < \eta \leq 1$, such that

$$\left\| \mathcal{K}_{S^c}^\top \mathcal{K}_S \left(\mathcal{K}_S^\top \mathcal{K}_S + \lambda_2 \right)^{-1} \right\|_\infty \leq 1 - \eta,$$

where S^c denotes the complement of S and $\mathcal{K}_S$ is the $m \times s$ matrix acting on the subset S with $(\mathcal{K}_S)_{ij} = \mathcal{K}(x_i, x_j)$, $i \in \{1, \ldots, m\}$, $j \in S$.

In the literature of compressed sensing (Donoho & Huo, 2001; Donoho & Elad, 2003; Tropp & Gilbert, 2007) and sparse linear regression (Zhao & Yu, 2006; Tibshirani, 2013), the coherence criterion and the irrepresentable condition have been introduced to examine correlations between pairs of atoms in a dictionary, which are shown to be closely related. In learning with the kernelized dictionary setting, the kernelized irrepresentable assumption introduced above also measures the correlation between atoms

that belong to different sets of dictionaries. In fact, we notice that the kernelized coherence criterion has been introduced within this setting (Richard, Bermudez, & Honeine, 2009; Honeine, 2014, 2015). In Richard et al. (2009), the kernelized coherence μ is defined as $\mu = \max_{i \neq j} |\mathcal{K}(x_i, x_j)|$, whereas dictionaries are selected to ensure that $\mu \leq \mu_0$ with μ_0 a given threshold. Simple computations show that with properly chosen λ_2 and μ_0, when $\mu \leq \mu_0$, assumption 5 also holds. On the other hand, it is easy to see that the kernel $\mathcal{K}$ and the ℓ^2-regularization term do have an influence in the kernelized irrepresentable assumption.

Theorem 2. *Suppose that assumptions 4 and 5 hold and* (λ_1, λ_2) *is chosen to satisfy*

$$\log(2m) \leq \frac{\eta^4 \lambda_1^2 m}{16\sigma^2}, \text{ and } \left\| \frac{2\lambda_2}{\lambda_1} \mathbf{a}_\star \right\|_\infty \leq \eta. \tag{4.1}$$

Then, with probability at least $1 - \exp\left(-\frac{\eta^4 \lambda_1^2 m}{16\sigma^2}\right)$, *KENReg has a unique solution with its support contained within the true support, that is,* $\operatorname{supp}(\boldsymbol{\alpha}_\mathbf{z}) \subseteq S$.

Theorem 2 states that low correlation between relevant and irrelevant kernel-spanned spaces leads to good model selection ability. Conditions 4.1 require that the sample size be large enough; meanwhile, the ratio between the two regularization parameters λ_2 and λ_1 should be upper-bounded. Under the kernelized irrepresentable assumption, theorem 2 tells us with overwhelming confidence that the zero pattern of $\boldsymbol{\alpha}_\star$ can be correctly identified.

Proof of Theorem 2. We first construct a sparse candidate estimator for a reduced-order form of KENReg; based on this, we construct an augmented candidate estimator. Eventually we show that the constructed candidate estimator is the unique solution to the original optimization problem in KENReg.

We first construct a sparse candidate estimator $\hat{\boldsymbol{\alpha}}_S$ by solving the following reduced-order elastic net regularization model,

$$\hat{\boldsymbol{\alpha}}_S = \arg\min_{\boldsymbol{\alpha}_S \in \mathbb{R}^s} \frac{1}{m} \|\mathbf{y} - \mathcal{K}_S \boldsymbol{\alpha}_S\|_2^2 + \lambda_1 \|\boldsymbol{\alpha}_S\|_1 + \lambda_2 \|\boldsymbol{\alpha}_S\|_2^2, \tag{4.2}$$

where $s \ll m$. The solution to this reduced-order convex problem is guaranteed to be unique due to its strict convexity, regardless of the reduced matrix $\mathcal{K}_S$. From optimality conditions for convex programs (Boyd & Vandenberghe, 2004), we know that the vector $\hat{\boldsymbol{\alpha}}_S$ is the solution to equation 4.2

if and only if there exists a subgradient $\hat{\boldsymbol{\gamma}}_S \in \mathbb{R}^s$ such that

$$\frac{1}{m}\mathcal{K}_S^\top \mathcal{K}_S \hat{\boldsymbol{\alpha}}_S - \frac{1}{m}\mathcal{K}_S^\top \mathbf{y} + \frac{\lambda_1}{2}\hat{\boldsymbol{\gamma}}_S + \lambda_2 \hat{\boldsymbol{\alpha}}_S = 0, \tag{4.3}$$

and moreover, there holds $\|\hat{\boldsymbol{\gamma}}_S\|_\infty \leq 1$.

We then construct $\hat{\boldsymbol{\alpha}}$ by augmenting $\hat{\boldsymbol{\alpha}}_S$ such that $\hat{\boldsymbol{\alpha}}^\top = (\hat{\boldsymbol{\alpha}}_S^\top, \mathbf{0}^\top)$ with $\mathbf{0} \in \mathbb{R}^{m-s}$, which is taken as a candidate solution to KENReg. Then we choose $\hat{\boldsymbol{\gamma}}_{S^c} \in \mathbb{R}^{m-s}$ that satisfies the following zero-subgradient optimality condition of the original KENReg,

$$\frac{1}{m}\mathcal{K}^\top \mathcal{K} \hat{\boldsymbol{\alpha}} - \frac{1}{m}\mathcal{K}^\top \mathbf{y} + \frac{\lambda_1}{2}\hat{\boldsymbol{\gamma}} + \lambda_2 \hat{\boldsymbol{\alpha}} = 0, \tag{4.4}$$

where $\hat{\boldsymbol{\gamma}}^\top = (\hat{\boldsymbol{\gamma}}_S^\top, \hat{\boldsymbol{\gamma}}_{S^c}^\top)$.

We now turn to prove that the estimator $\hat{\boldsymbol{\alpha}}$ constructed above is the unique solution to the original KENReg. In fact, from the construction of $\hat{\boldsymbol{\alpha}}$ and the optimality conditions for strictly convex programs, it suffices to prove that $\|\hat{\boldsymbol{\gamma}}_{S^c}\|_\infty \leq 1$. This can be obtained by first rewriting condition 4.4 in block form, as follows:

$$\begin{bmatrix} \mathcal{K}_S^\top \mathcal{K}_S/m & \mathcal{K}_S^\top \mathcal{K}_{S^c}/m \\ \mathcal{K}_{S^c}^\top \mathcal{K}_S/m & \mathcal{K}_{S^c}^\top \mathcal{K}_{S^c}/m \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\alpha}}_S \\ \mathbf{0} \end{bmatrix} - \begin{bmatrix} \mathcal{K}_S^\top/m \\ \mathcal{K}_{S^c}^\top/m \end{bmatrix} \mathbf{y} + \frac{\lambda_1}{2} \begin{bmatrix} \hat{\boldsymbol{\gamma}}_S \\ \hat{\boldsymbol{\gamma}}_{S^c} \end{bmatrix} + \lambda_2 \begin{bmatrix} \hat{\boldsymbol{\alpha}}_S \\ \mathbf{0} \end{bmatrix} = \mathbf{0}. \tag{4.5}$$

Recalling that the pair $(\hat{\boldsymbol{\alpha}}_S, \hat{\boldsymbol{\gamma}}_S)$ is obtained by solving the convex program, equation 4.3, therefore they must satisfy the top block of these equations. In addition, from the bottom block of equation 4.5, we obtain that

$$\hat{\boldsymbol{\gamma}}_{S^c} = \frac{2}{\lambda_1 m}\left(\mathcal{K}_{S^c}^\top \mathbf{y} - \mathcal{K}_{S^c}^\top \mathcal{K}_S \hat{\boldsymbol{\alpha}}_S\right).$$

The previous formula, together with equation 4.3, implies

$$\begin{aligned} \hat{\boldsymbol{\gamma}}_{S^c} &= \frac{2}{\lambda_1 m}\mathcal{K}_{S^c}^T \left(I_{m\times m} - \mathcal{K}_S \left(\mathcal{K}_S^\top \mathcal{K}_S + \lambda_2\right)^{-1} \mathcal{K}_S^\top\right) \mathbf{y} \\ &\quad + \mathcal{K}_{S^c}^\top \mathcal{K}_S \left(\mathcal{K}_S^\top \mathcal{K}_S + \lambda_2\right)^{-1} \hat{\boldsymbol{\gamma}}_S \\ &:= \Lambda_1 + \Lambda_2, \end{aligned}$$

where $I_{m\times m}$ stands for the identity matrix of size $m \times m$:

$$\Lambda_1 = \mathcal{K}_{S^c}^\top \mathcal{K}_S \left(\mathcal{K}_S^\top \mathcal{K}_S + \lambda_2\right)^{-1} \left(\frac{2\lambda_2}{\lambda_1}\mathbf{a}_\star + \hat{\boldsymbol{\gamma}}_S\right),$$

$$\Lambda_2 = \frac{2}{\lambda_1 m}\mathcal{K}_{S^c}^\top \left(I_{m\times m} - \mathcal{K}_S \left(\mathcal{K}_S^\top \mathcal{K}_S + \lambda_2\right)^{-1} \mathcal{K}_S^\top\right) \boldsymbol{\varepsilon}_m.$$

Therefore, to bound $\|\hat{\boldsymbol{\gamma}}_{S^c}\|_\infty$, it suffices to bound $\|\Lambda_1\|_\infty$ and $\|\Lambda_2\|_\infty$, respectively. Under the kernelized irrepresentable condition and recalling that $\|\hat{\boldsymbol{\gamma}}_S\|_\infty \leq 1$, we have

$$\|\Lambda_1\|_\infty \leq \left\|\mathcal{K}_{S^c}^\top \mathcal{K}_S \left(\mathcal{K}_S^\top \mathcal{K}_S + \lambda_2\right)^{-1}\right\|_\infty \left\|\frac{2\lambda_2}{\lambda_1}\mathbf{a}_\star + \hat{\boldsymbol{\gamma}}_S\right\|_\infty \leq 1 - \eta^2, \quad (4.6)$$

where the second inequality is due to the choice of the regularization parameters in equation 4.1. On the other hand, to bound $\|\Lambda_2\|_\infty$, we define the orthogonal projection matrix $\Pi_{S^\perp}$ by

$$\Pi_{S^\perp} := I_{m\times m} - \mathcal{K}_S\left(\mathcal{K}_S^\top \mathcal{K}_S + \lambda_2\right)^{-1}\mathcal{K}_S^\top.$$

Thus, Λ_2 can be element-wisely rewritten as

$$\Lambda_{2,j} = K_j^\top \Pi_{S^\perp} \frac{2\boldsymbol{\varepsilon}_m}{\lambda_1 m}, \quad j \in S^c,$$

where K_j denotes the jth column of $\mathcal{K}$. According to lemma 1.7 in Buldygin and Kozachenko (2000), we see that each $\Lambda_{2,j}$ is a zero-mean subgaussian noise variable with variance,

$$\mathrm{Var}(\Lambda_{2,j}) \leq \frac{4\sigma^2}{\lambda_1^2 m^2}\left\|\Pi_{S^\perp} K_j\right\|_2^2 \leq \frac{4\sigma^2}{\lambda_1^2 m},$$

where we use the fact that the projection matrix $\Pi_{S^\perp}$ has spectral norm one and the boundedness assumption that $|\mathcal{K}(x, x')| \leq 1$. As a result, by the subgaussian tail bound together with the union bound, we have

$$\mathbb{P}\left[\max_{j\in S^c} |\Lambda_{2,j}| \geq t\right] \leq 2(m-s)\exp\left(-\frac{t^2\lambda_1^2 m}{2\sigma^2}\right).$$

By letting $t = \tau^2/2$, we further obtain

$$\mathbb{P}\left[\max_{j\in S^c} |\Lambda_{2,j}| \geq \frac{\eta^2}{2}\right] \leq 2(m-s)\exp\left(-\frac{\eta^4\lambda_1^2 m}{8\sigma^2}\right) \leq \exp\left(-\frac{\eta^4\lambda_1^2 m}{16\sigma^2}\right), \quad (4.7)$$

where the last inequality follows from the relation $\log(2m) \leq \frac{\tau^4\lambda_1^2 m}{16\sigma^2}$.

Finally, by combining estimates in equations 4.6 and 4.7, we conclude that with probability at least $1 - \exp\left(-\frac{\eta^4\lambda_1^2 m}{16\sigma^2}\right)$, there holds

$$\|\hat{\boldsymbol{\gamma}}_{S^c}\|_\infty \leq \|\Lambda_1\|_\infty + \|\Lambda_2\|_\infty \leq 1 - \eta^2/2.$$

This shows that $\hat{\boldsymbol{\gamma}}$ is one subgradient of $\|\hat{\boldsymbol{\alpha}}\|_1$, and thus equation 4.4 is a zero subgradient equation of KENReg. According to optimality conditions for strictly convex programs (Boyd & Vandenberghe, 2004), we conclude that $\hat{\boldsymbol{\alpha}}$ is the unique solution of KENReg and due to the construction of $\hat{\boldsymbol{\alpha}}$, it is obvious that its support is contained within the true support. This completes the proof of theorem 2.

5 Interplay among Sparseness, Stability, and Generalization

KENReg possesses the algorithmic stability and sparseness property, while simultaneous sparseness and stability indeed give the main motivation of its introduction. Here, sparseness means that some components of the solution to KENReg are zero. This section is dedicated to discussing the sparseness, stability, and generalization properties of KENReg and their interplay. Conclusions drawn in this section are threefold: first, in learning with the kernelized dictionary setting, stability can also lead to the generalization ability as it does for learning machines over reproducing kernel Hilbert spaces; second, the confidence bounds on the sparseness of the output can be derived from the generalization results; third, the two properties of KENReg, sparseness and stability, are not mutually exclusive, which makes KENReg interesting because in general, sparse learning schemes are considered to be not stable (Xu et al., 2012). We next detail these arguments.

5.1 Stability Brings Generalization. Stability and generalization ability are two different but relevant important properties of a learning machine. In connection to the sensitivity analysis, the stability of a learning algorithm means that its output does not change much under small changes in the input observations, while generalization refers to its prediction ability on future unseen observations.

It is now commonly accepted that when learning within reproducing kernel Hilbert spaces, for some kernel-based learning machines including SVC, SVR, and kernel ridge regression, its stability in some sense is equivalent

to its generalization ability (Bousquet & Elisseeff, 2002). When moving attention to KENReg, as mentioned in section 1, it enjoys the property of algorithmic stability. Moreover, under the notion of uniform β-stable, such algorithmic stability property has been theoretically verified in Feng et al. (2014).

In the regression setup, the uniform β-stable of a learning algorithm (Bousquet & Elisseeff, 2002) is defined as follows.

Definition 1. *Let $\beta : \mathbb{N} \to \mathbb{R}$. Denote $\mathbf{z}^i$ as the modified samples by replacing the instance pair (x_i, y_i) with (x_i', y_i') in $\mathbf{z}$, for a fixed $i \in \{1, \ldots, m\}$, $(x_i', y_i') \in \mathcal{X} \times \mathcal{Y}$ and $m \in \mathbb{N}$. Let $\mathcal{A}$ be a learning algorithm with $\mathbf{z} \in (\mathcal{X} \times \mathcal{Y})^m$ as the input and $\mathcal{A}_\mathbf{z}$ as the output. We say that $\mathcal{A}$ has uniform stability β with respect to the loss function $\varphi : \mathbb{R}^\mathcal{X} \times \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^+ \cup \{0\}$ if for any $m \in \mathbb{N}$, $\mathbf{z} \in (\mathcal{X} \times \mathcal{Y})^m$, and $(x_i', y_i') \in \mathcal{X} \times \mathcal{Y}$ with $1 \le i \le m$, for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$, there holds*

$$\left|\varphi(\mathcal{A}_\mathbf{z}, (x, y)) - \varphi(\mathcal{A}_{\mathbf{z}^i}, (x, y))\right| \le \beta(m).$$

The following fact is due to Feng et al. (2014).

Fact 1. KENReg is uniform β-stable with $\beta = 8\big(1 + 2/\lambda_1 \wedge 2\sqrt{m/\lambda_2}\big)^2/\lambda_2$.

Simple calculations show that with the choice $\lambda_1 = m^{-2/(5p+2)}$, and $\lambda_2 = m\lambda_1$ as in corollary 1, the coefficient β in fact 1 is of the type $\mathcal{O}(m^{-(1-5p)/(2+5p)})$. Therefore, according to Bousquet and Elisseeff (2002), starting from the above algorithmic stability results, nontrivial generalization bounds can be derived. This coincides with the information brought by generalization bounds presented in theorem 1, that is, KENReg does generalize when λ_2 tends to infinity in accordance with m. However, this is not the case for the convergence rates reported in Feng et al. (2014), which are conducted with respect to a projected version of $f_\mathbf{z}$ and require that λ_2 goes to zero when m tends to infinity to ensure its convergence. This distinguishes the two types of convergence rates, and this seeming contradiction is in fact caused by the projection operator.

It should be mentioned that the convergence rates of KENReg derived from the algorithmic stability are not as fast as those derived by using concentration arguments stated in theorem 1. This is mainly because the derivation of the convergence rates in theorem 1 takes the second-order information of the noise term into account. On the other hand, in Feng et al. (2014), generalization bounds of KENReg with respect to its projected output, $\bar{f}_\mathbf{z}$, are also derived, which roughly states that to ensure the generalization ability of $\bar{f}_\mathbf{z}$, the regularization parameter λ_2 should go to zero when m tends to infinity. This obviously contradicts the stability results in fact 1 since with such a choice of λ_2, the algorithm stability coefficient β is useless in deriving the convergence rates of KENReg. This makes the

generalization bounds in theorem 1 more attractive as it is compatible with the stability arguments stated above.

In short, KENReg enjoys the algorithmic stability property, which can also bring us nontrivial generalization bounds.

5.2 Sparseness Confidence from Generalization. Besides the property of stability, another important property of KENReg lies in the fact that it advocates sparse output, being attributed to the sparsity-producing penalty it employs. More specifically, as Feng et al. (2014) shows, the sparseness of the solution to KENReg is characterizable. This can be indicated by the following fact:

Fact 2. For $i = 1, \ldots, m$, $\alpha_{\mathbf{z},i} = 0$ if and only if

$$\left| \frac{1}{m} \sum_{j=1}^{m} (y_j - f_{\mathbf{z}}(x_j)) \mathcal{K}(x_i, x_j) \right| < \frac{\lambda_1}{2}, \ \forall i = 1, \ldots, m. \tag{5.1}$$

Obviously, fact 2 provides a necessary and sufficient condition that characterizes the zero pattern of $\boldsymbol{\alpha}_{\mathbf{z}}$. Starting from fact 2, we now show that by using standard learning theory arguments, probabilistic confidence bounds that ensure the sparseness of KENReg can be established, as done in Shi et al. (2011).

The main idea is to derive a probabilistic confidence bound within which equation 5.1 holds. To this purpose, we provide a reminder that the population version of the empirical event

$$\frac{1}{m} \sum_{j=1}^{m} (y_j - f_{\mathbf{z}}(x_j)) \mathcal{K}(x_i, x_j) \tag{5.2}$$

is given by

$$\int_{\mathcal{X}} \mathcal{K}(x, u)(f_\rho(x) - f_{\mathbf{z}}(x)) d\rho_{\mathcal{X}}(x). \tag{5.3}$$

Applying the Cauchy-Schwarz inequality to equality 5.3, one gets

$$\left| \int_{\mathcal{X}} \mathcal{K}(x, u)(f_\rho(x) - f_{\mathbf{z}}(x)) d\rho_{\mathcal{X}}(x) \right| \leq \left\{ \int_{\mathcal{X}} (\mathcal{K}(x, u))^2 dx \right\}^{1/2} \left\| f_{\mathbf{z}} - f_\rho \right\|_{\mathcal{L}^2_{\rho_{\mathcal{X}}}}.$$

The term $\int_{\mathcal{X}} (\mathcal{K}(x, u))^2 dx$ involves only the kernel function $\mathcal{K}$ and can be bounded under restrictions on the regularity of $\mathcal{K}$, while the term $\left\| f_{\mathbf{z}} - f_\rho \right\|_{\mathcal{L}^2_{\rho_{\mathcal{X}}}}$ is the generalization error of $f_{\mathbf{z}}$ that is given in theorem 1.

Therefore, to derive a probabilistic bound that ensures equation 5.1, it remains to bound the difference between equations 5.2 and 5.3. Observe that equation 5.2 is in fact an empirical process that is associated with the randomized sampling $\mathbf{z}$. Standard learning theory arguments that involve empirical processes indexed by a class of functions can be applied, and a concentration estimate on the difference between equations 5.2 and 5.3 can be pursued. Here we omit the details.

To conclude, the sparseness of the solution to KENReg can be theoretically and probabilistically ensured as a consequence of the results on generalization bounds in theorem 1.

5.3 Sparseness and Stability Are Not Mutually Exclusive. As we have shown, KENReg possesses the sparseness and stability properties, both of which are arguably important for a good learning machine. However, Xu et al. (2012) argued that sparse learning algorithms are in some sense not stable, which seems to contradict our conclusion of the coexistence of sparseness and stability. In this section, we clarifying this aspect.

To illustrate, we first remind that KENReg is a kernel-based regression model. Like classic kernel-based regression methods, learning with KENReg is essentially an instance selection procedure. In this context, sparseness refers to the fact that some of the instances do not contribute to the output $f_{\mathbf{z}}$. However, the learning schemes that Xu et al. (2012) discuss give more emphasis on the feature selection problem, whereas an algorithm is said to be sparse if it is IRF (identify redundant features). According to Xu et al. (2012), "Being IRF means that at least one solution of the algorithm does not select both features if they are identical." This obviously excludes the well-known elastic net regularization scheme since it does not encourage the grouping effect property, the main advantage of elastic net regularization. In this sense, we say that the sparseness of a learning algorithm defined in Xu et al. (2012) is somewhat more restrictive and is different from the notion of sparseness in our context. This explains the apparent contradiction between conclusions we draw for KENReg and that in Xu et al. (2012).

From a statistical point of view, the stability of a learning machine is more related to its robustness, since "robustness theories can be viewed as stability theories of statistical inference" (Hampel, Ronchetti, Rousseeuw, & Stahel, 2011). For parametric learning models, their robustness is frequently pursued by applying robust loss functions, which take charge of the residuals (Huber & Ronchetti, 2009; Maronna, Martin, & Yohai, 2006; Hampel et al., 2011); similar results for kernel-based learning models can be also derived (Debruyne et al., 2008; Steinwart & Christmann, 2008; Feng et al., 2015). On the other hand, for a regularization model, more frequently the sparseness is brought by its penalty term, which sets restrictions on the structure of the hypothesis space. Therefore, more emphasis should be placed on designing the penalty term when pursuing a parsimonious model. In the above sense,

the stability and sparseness of a learning machine are of different natures and are not mutually exclusive.

6 Conclusion

In this letter, we studied the KENReg model in which a combined penalty term was employed. A learning theory analysis was presented, where emphasis was placed on its generalization bounds as well as the sparse recovery ability. The following three main results were presented:

- By constructing a new hypothesis space, we presented a concentration estimate to the KENReg model. As a result, refined generalization bounds were obtained, which are optimal in an asymptotic sense. Moreover, the newly obtained bounds were shown to coincide with that obtained via the algorithmic stability analysis.
- The KENReg model is a sparse model, so we studied its sparse recovery ability by assuming that the regression function has a sparse representation with respect to the kernelized dictionary. Theoretically, we showed that under the above assumption the KENReg model can correctly pick up the sparse pattern with overwhelming probability.
- We discussed different properties of the KENReg model, including sparseness, stability, and generalization, with special emphasis on their interplay. Roughly speaking, we showed that for the KENReg model, algorithmic stability also ensured that it can generalize, while its generalization bounds can be used to derive probabilistic bounds on the sparseness. Moreover, we also showed that for the KENReg model, its algorithmic stability and sparseness properties are not mutually exclusive.

Appendix: Technical Proofs

A.1 Proofs of Proposition 1 and Lemma 4

Proof of Proposition 1. According to definitions, it is easy to see that $\|f_{\mathbf{z}} - f_\rho\|_2^2 = \mathcal{E}(f_{\mathbf{z}}) - \mathcal{E}(f_\rho)$. For any $\mathbf{z} \in (\mathcal{X} \times \mathcal{Y})^m$, $\|f_{\mathbf{z}} - f_\rho\|_2^2 + \Omega(f_{\mathbf{z}})$ is equal to

$$\begin{aligned}
&\left\{\mathcal{E}(f_{\mathbf{z}}) - \mathcal{E}(f_\rho) - (\mathcal{E}_{\mathbf{z}}(f_{\mathbf{z}}) - \mathcal{E}_{\mathbf{z}}(f_\rho))\right\} + \left\{\mathcal{E}_{\mathbf{z}}(f_{\mathbf{z}}) + \Omega(f_{\mathbf{z}}) - \mathcal{E}_{\mathbf{z}}(g_{\mathbf{x},\lambda_1}) - \Omega(g_{\mathbf{x},\lambda_1})\right\} \\
&\quad + \left\{\mathcal{E}_{\mathbf{z}}(g_{\mathbf{x},\lambda_1}) - \mathcal{E}_{\mathbf{z}}(f_\rho) - (\mathcal{E}(g_{\mathbf{x},\lambda_1}) - \mathcal{E}(f_\rho))\right\} + \left\{\|g_{w,\lambda_1} - f_\rho\|_2 + \lambda_1 \|g_{w,\lambda_1}\|_{\mathcal{H}_w}^2\right\} \\
&\quad + \left\{\Omega(g_{\mathbf{x},\lambda_1}) - \lambda_1 \|g_{w,\lambda_1}\|_{\mathcal{H}_w}^2\right\} + \left\{\mathcal{E}(g_{\mathbf{x},\lambda_1}) - \mathcal{E}(g_{w,\lambda_1})\right\}.
\end{aligned}$$

From the definition of $f_{\mathbf{z}}$, we know that $\mathcal{E}_{\mathbf{z}}(f_{\mathbf{z}}) + \Omega(f_{\mathbf{z}}) - \mathcal{E}_{\mathbf{z}}(g_{\mathbf{x},\lambda_1}) - \Omega(g_{\mathbf{x},\lambda_1}) \le 0$. Consequently, the desired conclusion follows.

To bound $\mathbf{T}_3$ and $\mathbf{T}_4$ by applying lemma 1, we need to upper-bound f_{w,λ_1} and g_{w,λ_1} in terms of their infinity norm:

Lemma 4 *Let the function f_{w,λ_1} be defined as in equation 2.1; then*

$$\|f_{w,\lambda_1}\|_\infty \leq 2\sqrt{\mathcal{D}_w(\lambda_1)}/(\lambda_1 w),$$

and

$$\|g_{w,\lambda_1}\|_\infty \leq \sqrt{\mathcal{D}_w(\lambda_1)/\lambda_1}.$$

Proof. To proceed with the proof, we first introduce a univariate convex auxiliary function Φ on $\mathbb{R}$, that is, for any fixed function h,

$$\Phi(\theta) = \int_{\mathcal{X}} \left(L_{\mathcal{K}}(f_{w,\lambda_1} + \theta h) - f_\rho\right)^2 d\rho_{\mathcal{X}} + \lambda_1 \left(\int_{\mathcal{X}} \left|f_{w,\lambda_1} + \theta h\right| d\rho_{\mathcal{X}}\right)^2 + \lambda_1 w \int_{\mathcal{X}} \left(f_{w,\lambda_1} + \theta h\right)^2 d\rho_{\mathcal{X}}.$$

According to the definition of f_{w,λ_1} in equation 2.1, we see that $\partial\Phi(0) = 0$ for any fixed h, where $\partial\Phi(\tau)$ denotes the subgradient of Φ at τ. Due to the arbitrariness of h, this further implies that

$$\lambda_1 \|f_{w,\lambda_1}\|_1 \text{sign}(f_{w,\lambda_1}(t)) + \lambda_1 w f_{w,\lambda_1}(t) = -\int_{\mathcal{X}} \left(g_{w,\lambda_1}(x) - f_\rho(x)\right) \mathcal{K}(x,t) d\rho_{\mathcal{X}}(x), \quad \forall t \in \mathcal{X}, \tag{A.1}$$

where we use the fact that the subgradient of $|t|$ is $\text{sign}(t)$, given by

$$\text{sign}(t) = \begin{cases} +1, & t > 0, \\ -1, & t < 0, \\ [-1, 1], & t = 0. \end{cases}$$

If $f_{w,\lambda_1}(t) = 0$ for some t, the result is trivial. Otherwise, there holds $|\text{sign}(f_{w,\lambda_1}(t))| = 1$; by equation A.1, we have

$$w|f_{w,\lambda_1}(t)| \leq \|f_{w,\lambda_1}\|_1 + \sqrt{\mathcal{D}_w(\lambda_1)}/\lambda_1, \quad \forall t \in \mathcal{X}, \tag{A.2}$$

where this inequality follows from the Cauchy-Schwarz inequality and equation 2.3. Besides, it is clear from equation 2.1 that $\|f_{w,\lambda_1}\|_1 \leq$

$\sqrt{\mathcal{D}_w(\lambda_1)/\lambda_1}$. Since $\lambda_1 < 1$ is set in our setting, this, together with equation A.2, yields our first desired result.

To conclude the proof, we need to bound $\|g_{w,\lambda_1}\|_\infty$. Following from the boundedness assumption and the definition of g_{w,λ_1}, we have

$$g_{w,\lambda_1}(t) = \int_{\mathcal{X}} \mathcal{K}(x,t) f_{w,\lambda_1}(x) \rho_{\mathcal{X}}(x) \le \|f_{w,\lambda_1}\|_1 \le \sqrt{\mathcal{D}_w(\lambda_1)/\lambda_1}, \quad \forall t \in \mathcal{X}.$$

This completes the proof of lemma 4.

A.2 Proof of Proposition 3. We first bound the hypothesis error term $\mathbf{T}_3$, which can be split as follows,

$$\begin{aligned}\mathbf{T}_3 = &\frac{\lambda_1}{m}\sum_{i=1}^{m} |f_{w,\lambda_1}(x_i)| + \frac{\lambda_2}{m^2}\sum_{i=1}^{m} (f_{w,\lambda_1}(x_i))^2 - \lambda_1 \|f_{w,\lambda_1}\|_1^2 \\ &- \lambda_1 w \|f_{w,\lambda_1}\|_2^2 := \mathbf{I}_1 + \mathbf{I}_2 + \mathbf{I}_3,\end{aligned}$$

where

$$\begin{aligned}\mathbf{I}_1 &= \frac{\lambda_1}{m}\sum_{i=1}^{m} |f_{w,\lambda_1}(x_i)| - \lambda_1 \|f_{w,\lambda_1}\|_1, \\ \mathbf{I}_2 &= \frac{\lambda_2}{m^2}\sum_{i=1}^{m} (f_{w,\lambda_1}(x_i))^2 - \frac{\lambda_2}{m}\|f_{w,\lambda_1}\|_2^2, \\ \mathbf{I}_3 &= \lambda_1 \|f_{w,\lambda_1}\|_1 - \lambda_1 \|f_{w,\lambda_1}\|_1^2.\end{aligned}$$

Note that in the above expressions, $\mathbf{I}_2$ is denoted under the choice that $w = \lambda_2/(\lambda_1 m)$.

Bounding the quantities $\mathbf{I}_1$ and $\mathbf{I}_2$ involves the concentration arguments with respect to random variables that are associated with the deterministic function f_{w,λ_1}. This can be dealt with by applying lemma 1. To this end, we first introduce the random variable $\xi_1 = f_{w,\lambda_1}(x)$ on $(\mathcal{X}, \rho_{\mathcal{X}})$ with values in $\mathbb{R}$. From lemma 4, we know that $|\xi_1| \le \widetilde{M} := 2\sqrt{\mathcal{D}_w(\lambda_1)}/(\lambda_1 w)$. In addition, we see from equation 2.1 that

$$\widetilde{\sigma}^2 = \mathbb{E}(\xi_1^2) = \int_{\mathcal{X}} [f_{w,\lambda_1}(x)]^2 d\rho_{\mathcal{X}}(x) \le \|f_{w,\lambda_1}\|_\infty \|f_{w,\lambda_1}\|_1.$$

Therefore, lemma 1 tells us that with probability at least $1-\delta$, there holds

$$\frac{\mathbf{I}_1}{\lambda_1} \le \frac{2\widetilde{M}\log(2/\delta)}{m} + \sqrt{\frac{2\widetilde{\sigma}^2\log(2/\delta)}{m}} \le \frac{8\sqrt{\mathcal{D}_w(\lambda_1)}\log(2/\delta)}{m\lambda_1 w} + \frac{\mathcal{D}_w(\lambda_1)}{\lambda_1}.$$

An argument similar to that of $\mathbf{I}_1$ also tells us that with confidence $1-\delta$, there holds

$$\frac{m}{\lambda_2}\mathbf{I}_2 \le \frac{10\mathcal{D}_w(\lambda_1)\log(2/\delta)}{m(\lambda_1 w)^2} + \frac{\mathcal{D}_w(\lambda_1)}{\lambda_1 w}.$$

On the other hand, it is easy to see that

$$\mathbf{I}_3 \le \lambda_1\|f_{w,\lambda_1}\|_1(1-\|f_{w,\lambda_1}\|_1) \le \lambda_1/4 \le \mathcal{D}_w(\lambda_1)/(4c_\tau).$$

Combining the above estimates for $\mathbf{I}_1$, $\mathbf{I}_2$, and $\mathbf{I}_3$, we obtain the desired estimate of $\mathbf{T}_3$.

We now turn to bounding the term $\mathbf{T}_4$. Note that

$$\begin{aligned}\mathbf{T}_4 &= \|g_{\mathbf{x},\lambda_1} - f_\rho\|_2^2 - \|g_{w,\lambda_1} - f_\rho\|_2^2 \\ &\le \|g_{\mathbf{x},\lambda_1} - g_{w,\lambda_1}\|_2\left\{\|g_{\mathbf{x},\lambda_1} - g_{w,\lambda_1}\|_2 + 2\|g_{w,\lambda_1} - f_\rho\|_2\right\} \qquad \text{(A.3)} \\ &\le 2\|g_{\mathbf{x},\lambda_1} - g_{w,\lambda_1}\|_2^2 + \mathcal{D}_w(\lambda_1).\end{aligned}$$

Denoting ξ_2 as the random variable $\xi_2(x_i) = \mathcal{K}(x_i,x)f_{w,\lambda_1}(x_i)$ on $(\mathcal{X},\rho_{\mathcal{X}})$ with values in $\mathcal{L}^2_{\rho_{\mathcal{X}}}$, we then apply lemma 1 again to the random variable ξ_2. In this case, $g_{\mathbf{x},\lambda_1} = \frac{1}{m}\sum_{i=1}^m \xi_2(x_i)$ and $\mathbb{E}(\xi_2(x_i)) = (L_{\mathcal{K}}f_{w,\lambda_1})(x) = g_{\mathbf{x},\lambda_1}(x)$. Besides, it is easy to verify that

$$\widetilde{M} \le \|f_{w,\lambda_1}\|_\infty \quad \text{and} \quad \widetilde{\sigma}^2 = \mathbb{E}(\|\xi_2\|_2^2) \le \|f_{w,\lambda_1}\|_2^2.$$

Therefore, by applying lemma 1, we know that for any $0<\delta<1$, with confidence at least $1-\delta$, there holds

$$\|g_{\mathbf{x},\lambda_1} - g_{w,\lambda_1}\|_2 \le \frac{2\|f_{w,\lambda_1}\|_\infty\log(2/\delta)}{m} + \sqrt{\frac{2\log(2/\delta)}{m}}\|f_{w,\lambda_1}\|_2,$$

and thus the desired upper bound for $\mathbf{T}_4$ follows from equations 2.1 and A.3 and lemma 4.

A.3 Proof of Proposition 4. We now present a concentration estimation with respect to the random variable $\xi := (f_{\mathbf{z}}(x)-y)^2 - (f_\rho(x)-y)^2$

to bound the sample error term $\mathbf{T}_1$. Obviously, ξ varies among a set of functions in accordance with the varying sample $\mathbf{z}$. In this case, we turn to investigating the function set

$$\mathcal{F}_R = \left\{ g \mid g(z) = (f(x) - y)^2 - (f_\rho(x) - y)^2 : \ f \in B_R, \ z = (x, y) \in \mathcal{X} \times \mathcal{Y} \right\}.$$

To apply lemma 2 on the function set B_R, it suffices to verify conditions in lemma 2.

We first check the boundedness of functions in $\mathcal{F}_R$. In fact, for any $g \in \mathcal{F}_R$, the definition of the projecting operation and the boundedness of Y tell us that

$$\|g\|_\infty = \sup_{(x,y)\in\mathcal{X}\times\mathcal{Y}} \left\{ \left| f(x) - f_\rho(x) \right| \left| f(x) + f_\rho(x) - 2y \right| \right\} \leq (R+3)^2.$$

The variance of function-valued variables taking from $\mathcal{F}_R$ can be bounded as follows,

$$\begin{aligned} \mathbb{E}g^2 &= \mathbb{E}\{(f(x) - f_\rho(x))^2 (f(x) + f_\rho(x) - 2y)^2\} \\ &\leq (R+3)^2 \mathbb{E}(f(x) - f_\rho(x))^2 = (R+3)^2 \mathbb{E}g, \end{aligned}$$

where the above inequality again follows from the boundedness of Y. We now evaluate the capacity of $\mathcal{F}_R$ based on the capacity assumption. Observe that for any $g_1, g_2 \in \mathcal{F}_R$, with

$$\begin{aligned} g_1(z) &= (f_1(x) - y)^2 - (f_\rho(x) - y)^2 \ \text{ and } \ g_2(z) \\ &= (f_2(x) - y)^2 - (f_\rho(x) - y)^2, \end{aligned}$$

for any functions $f_1 = \sum_{i=1}^m \alpha_i \mathcal{K}(v_i, \cdot)$ and $f_2 = \sum_{i=1}^m \beta_i \mathcal{K}(u_i, \cdot)$, $u_i, v_i \in \mathcal{X}$, $i = 1, \ldots, m$. The supremum norm of the difference between g_1 and g_2 can be bounded as

$$\|g_1 - g_2\|_\infty = \sup_{(x,y)} \left| (f_1(x) - y)^2 - (f_2(x) - y)^2 \right| \leq 2(R+1) \|f_1 - f_2\|_\infty.$$

Consequently, it follows from the capacity assumption that

$$\begin{aligned} \log \mathcal{N}_2(\mathcal{F}_R, \epsilon) &\leq \log \mathcal{N}_\infty(\mathcal{F}_R, \epsilon) \leq \log \mathcal{N}_\infty(B_R, \epsilon/2(R+1)) \\ &\leq \tilde{c}_p \left(2R(R+1)/\epsilon\right)^p, \end{aligned}$$

which together with lemma 3 yields that the fixed point of lemma 2 will be of the order

$$r^* = c_0 R^{4p/(p+2)} m^{-2/(p+2)},$$

where c_0 is a positive constant independent of m or p.

Now applying lemma 2 and substituting the corresponding coefficients and indices, we see that for any $g \in \mathcal{F}_R$ and $0 < \delta < 1$, with confidence $1 - \delta/2$, there holds

$$\mathbb{E}g - \frac{1}{m}\sum_{i=1}^{m} g(z_i) \le \frac{(R+3)^2 \mathbb{E}g}{K} + 50Kc'_p R^{4p/(p+2)} \left(\frac{1}{m}\right)^{2/(p+2)} + \frac{(K + 18(R+3)^2)\log(2/\delta)}{m}, \tag{A.4}$$

where c'_p is a positive constant depending only on p. To complete the proof of proposition 4, we need the upper bound of R. By the definition of $\boldsymbol{\alpha}_{\mathbf{z}}$ in equation 1.4, we know that

$$\lambda_1 \|\boldsymbol{\alpha}_{\mathbf{z}}\|_1 + \lambda_2 \|\boldsymbol{\alpha}_{\mathbf{z}}\|_2^2 \le \mathcal{E}_{\mathbf{z}}(\boldsymbol{\alpha}_{\mathbf{z}}) + \lambda_1 \|\boldsymbol{\alpha}_{\mathbf{z}}\|_1 + \lambda_2 \|\boldsymbol{\alpha}_{\mathbf{z}}\|_2^2 \le \mathcal{E}_{\mathbf{z}}(0) \le 1.$$

On the other hand, Hölder's inequality tells us that $\|\boldsymbol{\alpha}_{\mathbf{z}}\|_1 \le \sqrt{m}\|\boldsymbol{\alpha}_{\mathbf{z}}\|_2$. Consequently, we can take $R = \lambda_1^{-1} \bigwedge \sqrt{m\lambda_2^{-1}}$. By letting $K = 2(R+3)^2$ and substituting R with $R = \lambda_1^{-1} \bigwedge \sqrt{m\lambda_2^{-1}}$ in inequality A.4, we come to the conclusion that for any $0 < \delta < 1$, with confidence $1 - \delta/2$, there holds

$$\mathbf{T}_1 \le \frac{1}{2}\left\|f_{\mathbf{z}} - f_\rho\right\|_2^2 + \frac{C_1 \log(2/\delta)}{m} + C'_1 \left(\lambda_1^{-1} \bigwedge \sqrt{m\lambda_2^{-1}}\right)^{4p/(2+p)} \left(\frac{1}{m}\right)^{2/(2+p)},$$

where C_1 and C'_1 are positive constants independent of m or δ. This completes the proof of proposition 4.

A.4 Proof of Proposition 5. We apply lemma 2 to prove proposition 5. To this end, we first define the function set

$$H = \left\{ g_{\mathbf{v},\lambda_1} \,\middle|\, g_{\mathbf{v},\lambda_1} = \frac{1}{m}\sum_{i=1}^{m} f_{w,\lambda_1}(v_i)\mathcal{K}(v_i, \cdot),\ v_i \in \mathcal{X} \right\},$$

and then we denote

$$\mathcal{F}_R = \left\{ g \mid g(z) = (g_{\mathbf{v},\lambda_1}(x) - y)^2 - (f_\rho(x) - y)^2,\ g_{\mathbf{v},\lambda_1} \in H \right\}.$$

It is easy to verify that for any $g \in \mathcal{F}_R$, there holds

$$\|g\|_\infty \leq c_4 \|f_{w,\lambda_1}\|_\infty^2 \leq c_4 m^2 \mathcal{D}_w(\lambda_1)/\lambda_2^2,$$

where c_4 is a positive constant. That is, for any $g \in \mathcal{F}_R$, there holds $\mathbb{E}g - g \leq 2c_4 m^2 \mathcal{D}_w(\lambda_1)/\lambda_2^2$. Consequently, it is obvious that

$$\mathbb{E}(g^2) \leq (c_4 m^2 \mathcal{D}_w(\lambda_1)/\lambda_2^2)\mathbb{E}(g).$$

On the other hand, for any $g_{\mathbf{v},\lambda_1}, g_{\mathbf{u},\lambda_1} \in H$ associated with $v_i, u_i \in \mathcal{X}$, we have

$$\begin{aligned}\|g_1 - g_2\|_\infty &= \sup_{(x,y)\in\mathcal{X}\times\mathcal{Y}} |(g_{\mathbf{v},\lambda_1}(x) - y)^2 - (g_{\mathbf{u},\lambda_1}(x) - y)^2| \\ &\leq 2(1 + \|f_{w,\lambda_1}\|_\infty)\|g_{\mathbf{v},\lambda_1} - g_{\mathbf{u},\lambda_1}\|_\infty. \end{aligned} \tag{A.5}$$

Noticing that $H \subseteq B_R$ with $R = \|f_{w,\lambda_1}\|_\infty \leq m\sqrt{\mathcal{D}_w(\lambda_1)}/\lambda_2$, and following from equation A.5 and capacity assumption 1, we know that

$$\log \mathcal{N}_2(\mathcal{F}_R, \epsilon) \leq \log \mathcal{N}_\infty(\mathcal{F}_R, \epsilon) \leq \log \mathcal{N}_\infty(B_R, \epsilon/D) \leq \tilde{c}_p\, (DR/\epsilon)^p\,,$$

where $D = 2(1 + \|f_{w,\lambda_1}\|_\infty)$. By letting $K = 2c_p m^2 \mathcal{D}_w(\lambda_1)/\lambda_2^2$ and applying lemma 2 to the function $\mathcal{F}_R$, we see that for any $0 < \delta < 1$, with confidence at least $1 - \delta/2$, there holds

$$\begin{aligned}\mathbf{T}_2 \leq \frac{1}{2}(\mathcal{E}(g_{\mathbf{x},\lambda_1}) - \mathcal{E}(f_\rho)) + 50Kc_p'(DR)^{2p/(p+2)} \left(\frac{1}{m}\right)^{2/(p+2)} \\ + \frac{(K + 9b)\log(2/\delta)}{m}. \end{aligned} \tag{A.6}$$

Substituting b, R, D, and K with the corresponding estimate given above into equation A.6 and recalling the following relation,

$$\mathcal{E}(g_{\mathbf{x},\lambda_1}) - \mathcal{E}(f_\rho) = \mathbf{T}_4 + \mathcal{E}(g_{w,\lambda_1}) - \mathcal{E}(f_\rho),$$

we obtain the desired estimate in proposition 5 after simple computations. This completes the proof.

A.5 Proofs of Theorem 1 and Corollary 1

Proof of Theorem 1. As a consequence of proposition 1, to bound $\|f_{\mathbf{z}} - f_\rho\|_2^2$, it remains to combine the upper bounds of $\mathbf{T}_1$, $\mathbf{T}_2$, $\mathbf{T}_3$, and $\mathbf{T}_4$. With

simple computations, we obtain that for any $0 < \delta < 1$, with confidence $1 - 4\delta$, $\|f_{\mathbf{z}} - f_\rho\|_2^2$ can be upper-bounded by

$$C_0 \log^2\left(\frac{2}{\delta}\right)\left\{\frac{\mathcal{D}_w(\lambda_1)}{m\lambda_1^\gamma} + \left(\frac{1}{\lambda_1}\right)^{\frac{4p}{p+2}}\left(\frac{1}{m}\right)^{\frac{2}{p+2}} + \left(\frac{1}{\lambda_1}\right)^{\frac{3p+2}{p+2}}\left(\frac{1}{m}\right)^{\frac{2}{p+2}}\mathcal{D}_w(\lambda_1)^{\frac{3p+2}{p+2}} + \mathcal{D}_w(\lambda_1)\right\},$$

where C_0 is a positive constant independent of m or δ.

Recalling the model assumption that $\mathcal{D}_w(\lambda_1) = O(\lambda_1^\tau)$, when $2 - p \geq (3p+2)\tau$, it is easy to verify that in the above estimate, the third term dominates the second one. For this case, simple computations show that with confidence $1 - 4\delta$, there holds

$$\|f_{\mathbf{z}} - f_\rho\|_2^2 \leq C_{\mathcal{K},\rho} \log^2\left(\frac{2}{\delta}\right)\left(m^{-\frac{2\tau}{3p+2-2p\tau}} + m^{-1+\frac{2(\gamma-\tau)}{3p+2-2p\tau}}\right),$$

where $C_{\mathcal{K},\rho}$ is a positive constant independent of m or δ. On the other hand, when $2 - p < (3p+2)\tau$, with confidence $1 - 4\delta$, we have

$$\|f_{\mathbf{z}} - f_\rho\|_2^2 \leq C_{\mathcal{K},\rho} \log^2\left(\frac{2}{\delta}\right)\left(m^{-\frac{2\tau}{(4+\tau)p+2\tau}} + m^{-1+\frac{2(\gamma-\tau)}{(4+\tau)p+2\tau}}\right),$$

with the choice of $\lambda_1 = m^{-\frac{2}{2\tau+p(4+\tau)}}$. To ensure nontrivial convergence rates, we let $0 < \gamma \leq 2\tau$. This completes the proof of theorem 1.

Proof of Corollary 1. When $\tau = 1$, the desired estimate in corollary 1 is a direct result of theorem 1 by taking $\gamma = 1$.

Acknowledgments

We thank the editor and the reviewers for their insightful comments and helpful suggestions that helped to improve the quality of this letter. We would also like to thank Dr. Yunwen Lei for pointing out a mistake in an early version of this letter. The corresponding author is S.-G. L. The research leading to these results has received funding from the European Research Council under the European Union's Seventh Framework Programme (FP7/2007-2013)/ERC AdG A-DATADRIVE-B (290923). This letter reflects our own views; the EU is not liable for any use that may be made of the contained information. Funding has also been received from Research Council KUL: GOA/10/09 MaNet, CoE PFV/10/002 (OPTEC), BIL12/11T;

PhD/postdoc grants; Flemish Government: FWO: PhD/postdoc grants, projects: G.0377.12 (Structured systems), G.088114N (Tensor-Based Data Similarity); IWT: PhD/postdoc grants, projects: SBO POM (100031); iMinds Medical Information Technologies SBO 2014; Belgian Federal Science Policy Office: IUAP P7/19 (DYSCO, Dynamical systems, Control and Optimization, 2012–2017). S.-G.L is supported partially by the National Natural Science Foundation of China (no. 11301421), and Fundamental Research Funds for the Central Universities of China (grants JBK141111, 14TD0046, and JBK151134).

References

Bach, F. (2013). Sharp analysis of low-rank kernel matrix approximations. In *Proceedings of the Twenty-Sixth Annual Conference on Computational Learning Theory*. New York: Springer.

Bartlett, P. L., Bousquet, O., & Mendelson, S. (2005). Local Rademacher complexities. *Annals of Statistics*, *33*(4), 1497–1537.

Bousquet, O., & Elisseeff, A. (2002). Stability and generalization. *Journal of Machine Learning Research*, *2*, 499–526.

Boyd, S., & Vandenberghe, L. (2004). *Convex optimization*. Cambridge: Cambridge University Press.

Buldygin, V. V., & Kozachenko, Y. V. (2000). *Metric characterization of random variables and random processes*. Providence, RI: American Mathematical Society.

Candès, E. J., Romberg, J. K., & Tao, T. (2006). Stable signal recovery from incomplete and inaccurate measurements. *Communications on Pure and Applied Mathematics*, *59*(8), 1207–1223.

Candès, E. J., & Tao, T. (2005). Decoding by linear programming. *IEEE Transactions on Information Theory*, *51*(12), 4203–4215.

Chen, H., Pan, Z., Li, L., & Tang, Y. (2013). Error analysis of coefficient-based regularized algorithm for density-level detection. *Neural Computation*, *25*(4), 1107–1121.

Cucker, F., & Zhou, D.-X. (2007). *Learning theory: An approximation theory viewpoint*. Cambridge: Cambridge University Press.

Debruyne, M., Hubert, M., & Suykens, J. A. K. (2008). Model selection in kernel based regression using the influence function. *Journal of Machine Learning Research*, *9*, 2377–2400.

Donoho, D. L., & Elad, M. (2003). Optimally sparse representation in general (nonorthogonal) dictionaries via ℓ^1 minimization. *Proceedings of the National Academy of Sciences*, *100*(5), 2197–2202.

Donoho, D. L., & Huo, X. (2001). Uncertainty principles and ideal atomic decomposition. *IEEE Transactions on Information Theory*, *47*(7), 2845–2862.

Drezet, P.M.L., & Harrison, R. R. (1998). Support vector machines for system identification. In *Proceedings of UKACC International Conference on Control* (vol. 1, pp. 688–692). London: Institution of Electrical Engineers.

Dudley, R. M. (1999). *Uniform central limit theorems*. Cambridge: Cambridge University Press.

Feng, Y., Huang, X., Shi, L., Yang, Y., & Suykens, J. A. K. (2015). Learning with the maximum correntropy criterion induced losses for regression. *Journal of Machine Learning Research*, *16*, 993–1034.

Feng, Y., Yang, Y., Zhao, Y., Lv, S., & Suykens, J. A. K. (2014). *Learning with kernelized elastic net regularization* (Internal Report, ESAT-STADIUS). Leuven, Belgium: KU Leuven. ftp://ftp.esat.kuleuven.ac.be/stadius/yfeng/KENReg2014.pdf

Györfi, L., Kohler, M., Krzyżak, A., & Walk, H. (2002). *A distribution-free theory of nonparametric regression*. New York: Springer.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., & Stahel, W. A. (2011). *Robust statistics: The approach based on influence functions*. Hoboken, NJ: Wiley.

Hang, H., & Steinwart, I. (2014). Fast learning from α-mixing observations. *Journal of Multivariate Analysis*, *127*, 184–199.

Honeine, P. (2014). *Analyzing sparse dictionaries for online learning with kernels*. arXiv:1409.6045

Honeine, P. (2015). Approximation errors of online sparsification criteria. *IEEE Transactions on Signal Processing*, *63*(37), 4700–4709

Huang, J., & Zhang, T. (2010). The benefit of group sparsity. *Annals of Statistics*, *38*(4), 1978–2004.

Huang, J., Zhang, T., & Metaxas, D. (2011). Learning with structured sparsity. *Journal of Machine Learning Research, 12*, 3371–3412.

Huber, P. J., & Ronchetti, E. (2009). *Robust statistics*. Hoboken, NJ: Wiley.

Koltchinskii, V. (2011). *Oracle inequalities in empirical risk minimization and sparse recovery problems*. New York: Springer.

Koltchinskii, V., Sakhanenko, L., & Cai, S. (2007). Integral curves of noisy vector fields and statistical problems in diffusion tensor imaging: Nonparametric kernel estimation and hypotheses testing. *Annals of Statistics*, *35*(4), 1576–1607.

Lin, S., Zeng, J., Fang, J., & Xu, Z. (2014). Learning rates of ℓ^q-coefficient regularization learning with gaussian kernel. *Neural Computation*, *26*(10), 2350–2378.

Lv, S. (2015). Refined generalization bounds of gradient learning over reproducing kernel Hilbert spaces. *Neural Computation*, *27*(6), 1294–1320.

Maronna, R., Martin, D., & Yohai, V. (2006). *Robust statistics: Theory and methods*. Chichester: Wiley.

Pinelis, I. (1994). Optimum bounds for the distributions of martingales in Banach spaces. *Annals of Probability*, *22*(4), 1679–1706.

Richard, C., Bermudez, J. C. M., & Honeine, P. (2009). Online prediction of time series data with kernels. *IEEE Transactions on Signal Processing*, *57*(3), 1058–1067.

Roth, V. (2004). The generalized Lasso. *IEEE Transactions on Neural Networks*, *15*(1), 16–28.

Schölkopf, B., & Smola, A. J. (2001). *Learning with kernels: Support vector machines, regularization, optimization, and beyond*. Cambridge, MA: MIT Press.

Shi L. (2013). Learning theory estimates for coefficient-based regularized regression. *Applied and Computational Harmonic Analysis*, *34*(2), 252–265.

Shi, L., Feng, Y., & Zhou, D.-X. (2011). Concentration estimates for learning with ℓ^1-regularizer and data dependent hypothesis spaces. *Applied and Computational Harmonic Analysis*, *31*(2), 286–302.

Steinwart, I., & Christmann, A. (2008). *Support vector machines*. New York: Springer.

Sun, H., & Wu, Q. (2011). Least square regression with indefinite kernels and coefficient regularization. *Applied and Computational Harmonic Analysis*, *30*(1), 96–109.

Suykens, J. A. K., Van Gestel, T., De Brabanter, J., De Moor, B., & Vandewalle, J. (2002). *Least squares support vector machines*. Singapore: World Scientific.

Tibshirani, R. J. (2013). The Lasso problem and uniqueness. *Electronic Journal of Statistics*, *7*, 1456–1490.

Tong, H., Chen, D.-R., & Yang, F. (2010). Least square regression with ℓ^p-coefficient regularization. *Neural Computation*, *22*(12), 3221–3235.

Tropp, J., & Gilbert, A.C. (2007). Signal recovery from random measurements via orthogonal matching pursuit. *IEEE Transactions on Information Theory*, *53*(12), 4655–4666.

Vapnik, V. (1998). *Statistical learning theory*. Hoboken, NJ: Wiley-Interscience.

Wang, G., Yeung, D.-Y., & Lochovsky, F. H. (2007). The kernel path in kernelized Lasso. In *Proceedings of the International Conference on Artificial Intelligence and Statistics* (pp. 580–587). JMLR.

Wu, Q., & Zhou, D.-X. (2005). SVM soft margin classifiers: Linear programming versus quadratic programming. *Neural Computation*, *17*(5), 1160–1187.

Wu, Q., & Zhou, D.-X. (2008). Learning with sample dependent hypothesis spaces. *Computers & Mathematics with Applications*, *56*(11), 2896–2907.

Xiao, Q.-W., & Zhou, D.-X. (2010). Learning by nonsymmetric kernels with data dependent spaces and ℓ^1-regularizer. *Taiwanese Journal of Mathematics*, *14*(5), 1821–1836.

Xu, H., Caramanis, C., & Mannor, S. (2012). Sparse algorithms are not stable: A no-free-lunch theorem. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *34*(1), 187–193.

Xu, Z., Jin, R., Shen, B., & Zhu, S. (2015). Nyström approximation for sparse kernel methods: Theoretical analysis and empirical evaluation. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*. Palo Alto, CA: AAAI Press.

Zhang, T. (2010). Analysis of multi-stage convex relaxation for sparse regularization. *Journal of Machine Learning Research*, *11*, 1081–1107.

Zhang, T. (2011). Sparse recovery with orthogonal matching pursuit under RIP. *IEEE Transactions on Information Theory*, *57*(9), 6215–6221.

Zhao, P., & Yu, B. (2006). On model selection consistency of Lasso. *Journal of Machine Learning Research*, *7*, 2541–2563.

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B*, *67*(2), 301–320.

Received August 31, 2015; accepted November 4, 2015.