Learning with Kernelized Elastic Net Regularization

Yunlong Feng YUNLONG.FENG@ESAT.KULEUVEN.BE
Yuning Yang YUNING.YANG@ESAT.KULEUVEN.BE
Department of Electrical Engineering, ESAT-STADIUS, KU Leuven
Kasteelpark Arenberg 10, Leuven, B-3001, Belgium
Shao-Gao Lv LVSG716@SWUFE.EDU.CN
Center of Statistics, Statistics School
Southwestern University of Finance and Economics, Chengdu, China
Yulong Zhao YZ367@MSSTATE.EDU
Department of Basic Sciences
Mississippi State University, Mississippi State, MS 39762, USA
Johan A.K. Suykens JOHAN.SUYKENS@ESAT.KULEUVEN.BE
Department of Electrical Engineering, ESAT-STADIUS, KU Leuven
Kasteelpark Arenberg 10, Leuven, B-3001, Belgium

Abstract

Within the statistical learning framework, this paper proposes a new kernel-based sparse regularization model induced by an elastic net type regularizer. In our study, the purpose of enforcing the elastic net regularizer is to pursue a sparse and stable approximation to the regression function. This is different from the original motivation of introducing the elastic net regularizer in the statistics community, which aims at selecting group and correlated features. The term "kernel-based" refers to the fact that the proposed kernelized elastic net regularization model learns with a kernelized dictionary. We show that the proposed model outperforms the kernelized Lasso in terms of characterizable sparsity, computational efficiency, and algorithmic stability. Generalization bounds of the proposed model are derived for both cases when the kernels are positive definite and indefinite, respectively.

Keywords: Elastic net regularization, kernelized dictionary, kernelized Lasso, stability, covering number, generalization bounds, regression

1. Introduction and Motivation

Nowadays, in many real-world data-driven applications, there exists a great demand of parsimonious and accurate models, and as such, this gives birth to various sparsity-producing algorithms and techniques. Among them, the Lasso (Tibshirani, 1996) is a very popular tool for researchers and practitioners in various areas. However, the Lasso suffers from the deficiency of choosing correlated features, as explained in Zou and Hastie (2005). As a remedy, Zou and Hastie (2005) proposed the well-known elastic net regularization model, which can successfully select the correlated features while preserving the sparseness of the model. This purpose is achieved by introducing an additional ℓ^2-regularization term (w.r.t the covariate coefficients). Since then, the elastic net regularizers have been imposed into various data-fitting models and earned significant successes both empirically and theoretically. In addition, as a parametric linear model, the Lasso cannot capture the nonlinearity

© 2016 Yunlong Feng, Yuning Yang, Shao-Gao Lv, Yulong Zhao, and Johan A.K. Suykens.

of the raw data. This is a limitation of the Lasso, especially when the prediction ability of the model is the main concern. To handle nonlinearity, various nonlinear regression models have been proposed. A breakthrough is put forward since the introduction of the kernel-based models into the nonlinear regression problems, which is motivated by the success of support vector machines (Vapnik, 1995). Typical kernel-based regression models include SVR (Vapnik, 1995), LS-SVM (Suykens et al., 2002), and ν-SVR (Schölkopf et al., 2000).

Before delving into technical details, we first give a formal setting and some notations on the nonlinear regression problem. Assume that $\mathcal{X}$ is a compact subset of $\mathbb{R}^d$, which stands for the instance space and $\mathcal{Y} \subset \mathbb{R}$ is referred as the output space. The underlying nonlinear regression model is assumed as

$$Y = f^\star(X) + \epsilon, \tag{1}$$

where ϵ is the zero-mean additive noise, X and Y are the explanatory variable and the response variable taking values in $\mathcal{X}$, and $\mathcal{Y}$, respectively. Suppose that we are given a set of i.i.d. observations $\mathbf{z} = \{(x_i, y_i)\}_{i=1}^m$ generated by model (1) from an unknown distribution ρ over $\mathcal{X} \times \mathcal{Y}$. The regression problem aims at approximating the conditional mean, $f^\star$, from the given observations $\mathbf{z}$.

In the kernel-based regression settings, there exist several different approaches to seek a sparse output function. One approach is to use special loss functions. For instance, in SVR and ν-SVR, an ε-insensitive zone is placed to the loss function. Some approaches seek sparseness by using various pruning techniques (Suykens et al., 2000; Hoegaerts et al., 2004) and considering reduced-order models (De Brabanter et al., 2010; Karsmakers et al., 2011). Recently, another "built-in" sparse regression model, the kernelized Lasso, was proposed in Schölkopf and Smola (2001) and investigated in Roth (2004); Wang et al. (2007); Wu and Zhou (2008); Xiao and Zhou (2010); Shi et al. (2011); Zhang et al. (2016). It advocates a sparse regression model by penalizing the ℓ^1-norm over the coefficients of the kernelized dictionary, $\mathcal{D}_m := \{\mathcal{K}(x, x_i)\}_{i=1}^m$, where $\mathcal{K} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a kernel[1] which is allowed to be indefinite (Haasdonk, 2005). The idea is to interpret $(\mathcal{K}(x, x_1), \ldots, \mathcal{K}(x, x_m))^\top$ as the feature vector and then the Lasso procedure can be applied. Consequently, the sparseness is delivered by the ℓ^1 penalty term, which is indeed instance-wise sparse. To be more precise, let us denote $\mathbf{y} = (y_1, y_2, \ldots, y_m)^\top$ and $\boldsymbol{\mathcal{K}}$ as the $m \times m$ matrix with entries $\boldsymbol{\mathcal{K}}_{i,j} = \mathcal{K}(x_i, x_j)$ for $i, j = 1, \ldots, m$. Then the kernelized Lasso can be formulated as

$$\boldsymbol{\alpha}_{\mathbf{z},\eta} = \arg\min_{\boldsymbol{\alpha} \in \mathbb{R}^m} \|\mathbf{y} - \boldsymbol{\mathcal{K}}\boldsymbol{\alpha}\|_2^2 + \eta \|\boldsymbol{\alpha}\|_1, \tag{2}$$

where $\eta > 0$ is the regularization parameter, $\boldsymbol{\alpha}_{\mathbf{z},\eta} = (\alpha_{\mathbf{z},\eta,1}, \alpha_{\mathbf{z},\eta,2}, \ldots, \alpha_{\mathbf{z},\eta,m})^\top$ and $\|\boldsymbol{\alpha}\|_1$, $\|\boldsymbol{\alpha}\|_2$ are the ℓ^1-norm and ℓ^2-norm of the vector $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_m)^\top$ respectively. Following (2), the learned output function can be modeled as

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

Concerning the approximation ability of $f_{\mathbf{z},\eta}$ to $f^\star$, learning theory estimates were derived in Xiao and Zhou (2010); Shi et al. (2011); Feng and Lv (2011) and Guo and Shi (2013)

1. Throughout this paper, we assume that $\sup_{x,x' \in \mathcal{X}} |\mathcal{K}(x, x')| \leq 1$.

while some computational and model selection considerations can be found in Wang et al. (2007); Gao et al. (2010).

However, despite its capabilities of capturing the nonlinearity and producing sparsity, the kernelized Lasso still suffers from the stability problem. That is, the ℓ^1-norm regularizer based kernelized Lasso is not stable. In this context, the implications of "stable" are two-fold: the algorithmic stability (Bousquet and Elisseeff, 2002; De Mol et al., 2009; Xu et al., 2012) and the computational stability (Jin et al., 2009), as we shall discuss in detail later.

To alleviate the above-mentioned problems, in this paper we study the following Kernelized Elastic Net Regularization (KENReg) model:

$$\text{(KENReg)} \qquad \boldsymbol{\alpha}_{\mathbf{z},\lambda_1,\lambda_2} = \arg\min_{\boldsymbol{\alpha}\in\mathbb{R}^m} \|\mathbf{y} - \boldsymbol{\mathcal{K}}\boldsymbol{\alpha}\|_2^2 + \lambda_1\|\boldsymbol{\alpha}\|_1 + \lambda_2\|\boldsymbol{\alpha}\|_2^2, \tag{3}$$

where $\lambda_1, \lambda_2 > 0$ are regularization parameters. To simplify notations, in what follows we suppress the subscripts λ_1, λ_2 and denote $\boldsymbol{\alpha}_{\mathbf{z}} := \boldsymbol{\alpha}_{\mathbf{z},\lambda_1,\lambda_2}$, where $\boldsymbol{\alpha}_{\mathbf{z}} = (\alpha_{\mathbf{z},1}, \alpha_{\mathbf{z},2}, \ldots, \alpha_{\mathbf{z},m})^\top$. Following (3), the learned empirical target function is given by

$$f_{\mathbf{z}}(x) = \sum_{i=1}^{m} \alpha_{\mathbf{z},i}\, \mathcal{K}(x_i, x), \quad x \in \mathcal{X}. \tag{4}$$

The rest of this paper is organized as follows. In Section 2, we discuss the advantages of learning schemes induced by the kernelized dictionary and the sparseness of KENReg. Section 3 explores the stability of KENReg, which is the main motivation of introducing KENReg. We provide generalization bounds for KENReg in Section 4 by considering the case when the kernel function is positive semidefinite and the case when the kernel is indefinite. Barriers in deriving these generalization bounds will be also commented on in this section. Comparisons between KENReg and related studies are presented in Section 5. We conclude the paper in Section 6. All proofs of Section 4 are provided in the appendix.

2. KENReg: Kernelized Dictionary, and Sparseness

2.1 Learning with the Kernelized Dictionary

KENReg learns over the hypothesis space spanned by the kernelized dictionary $\mathcal{D}_m$. Learning models induced by $\mathcal{D}_m$ are also usually termed as coefficient-based learning models, see e.g., Wu and Zhou (2005); Xiao and Zhou (2010); Sun and Wu (2011); Tong et al. (2010); Shi et al. (2011); Shi (2013); Chen et al. (2013); Guo and Shi (2013); Lin et al. (2014) and many others.

The kernelized dictionary together with its induced learning models has been previously considered in literature. In a supervised regression setting, to pursue a sparse nonlinear regression machine, Schölkopf and Smola (2001) proposed the ℓ^1-norm regularized learning model induced by the kernelized dictionary, namely, the kernelized Lasso. It is indeed a basis pursuit method, the idea of which can be dated back to Chen and Donoho (1994); Girosi (1998). To seek a sparse representation, Vincent and Bengio (2002) proposed a matching pursuit algorithm based on the kernelized dictionary. It was in Wu and Zhou (2008) that a framework of analyzing the generalization bounds for learning schemes induced by the kernelized dictionary was proposed. Since then, a series of interesting following-up studies

have been expanded for various learning models induced by the kernelized dictionary. For example, probabilistic generalization bounds for different learning models were derived in Xiao and Zhou (2010); Shi et al. (2011); Tong et al. (2010, 2014); Shi (2013); Chen et al. (2013); Guo and Shi (2013) and many others. As noticed, learning with the kernelized dictionary provides us more flexibility. For instance:

- "*Testing Mercer's condition for a given kernel can be a challenging task which may well lie beyond the abilities of a practitioner*" (Ong et al., 2004). By learning with the kernelized dictionary, the positive semi-definite constraint on the kernel function is removed and this enables us to utilize some specific indefinite kernels to cope with real-world applications, see e.g., Ong et al. (2004); Roth (2004); Haasdonk (2005); Luss and d'Aspremont (2009); Balcan et al. (2008); Ying et al. (2009); Kar and Jain (2012);

- When being restricted to Mercer kernels, models induced by the kernelized dictionary usually provide comparable learnabilities as regularization schemes in the reproducing kernel Hilbert spaces, see e.g., Wu and Zhou (2005, 2008); Sun and Wu (2011); Zhang et al. (2016);

- Learning models induced by the kernelized dictionary with specific choices of kernel functions possess the rooted robustness property, where the "robustness property" refers to the ability of being resistant to the leverage points (outliers in the explanatory variables). In this sense, they inherit the robustness properties of support vector machines as revealed in Steinwart and Christmann (2008);

- By learning with the kernelized dictionary, various regularization models can be encompassed by introducing different penalty terms and loss functions, e.g., ℓ^1-norm regularized regression (Xiao and Zhou, 2010; Zhang and Zhou, 2010; Shi et al., 2011; Zhang et al., 2016), ℓ^2-norm regularized regression (Sun and Wu, 2011; Shi, 2013; Wu, 2013), ℓ^p-regularization (Tong et al., 2010), ℓ^1-norm novelty detection (Zhang and Zhou, 2013; Chen et al., 2013), kernelized elastic-net regularization in this paper and many others;

- Learning with the kernelized dictionary enables us to consider a reduced order model to speed up the training process in large scale applications. The reduced order model can be selected via a kernel matching pursuit method (Vincent and Bengio, 2002; Nair et al., 2003), a kernel basis pursuit strategy (Guigue et al., 2005), an online manner (Honeine, 2012) and an information-theoretic criterion (Richard et al., 2009), as well as many other heuristics.

2.2 On the Sparseness of KENReg

We now discuss the sparseness of the empirical target function $f_{\mathbf{z}}$ trained by KENReg. Specifically, we will show that the zero components of $\boldsymbol{\alpha}_{\mathbf{z}}$ can be characterized, which is not the case for the kernelized Lasso.

Denote $\mathcal{R}_{\mathbf{z}}(\boldsymbol{\alpha})$ as the total data-fitting risk of the coefficient vector $\boldsymbol{\alpha}$, that is

$$\mathcal{R}_{\mathbf{z}}(\boldsymbol{\alpha}) := \|\mathbf{y} - \boldsymbol{\mathcal{K}}\boldsymbol{\alpha}\|_2^2 + \lambda_1\|\boldsymbol{\alpha}\|_1 + \lambda_2\|\boldsymbol{\alpha}\|_2^2. \tag{5}$$

Recall that $\boldsymbol{\alpha}_{\mathbf{z}}$ is the unique minimizer of the above risk functional. The uniqueness of $\boldsymbol{\alpha}_{\mathbf{z}}$ is due to the strong convexity of $\mathcal{R}_{\mathbf{z}}(\boldsymbol{\alpha})$ with respect to $\boldsymbol{\alpha}$. Based on the above notations, the following conclusion that characterizes the non-zero components of $\boldsymbol{\alpha}$, i.e., $\{\alpha_{\mathbf{z},i} \,|\, \alpha_{\mathbf{z},i} \neq 0,\, i = 1, \ldots, m\}$ can be deduced.

Theorem 1 *For $i = 1, \ldots, m$, $\alpha_{\mathbf{z},i} \neq 0$ if and only if*

$$\left|(\mathbf{y} - \boldsymbol{\mathcal{K}}\boldsymbol{\alpha}_{\mathbf{z}})^{\top}\boldsymbol{\mathcal{K}}_{x_i}\right| > \frac{\lambda_1}{2},$$

where $\boldsymbol{\mathcal{K}}_{x_i} = (\mathcal{K}(x_i, x_1), \mathcal{K}(x_i, x_2), \ldots, \mathcal{K}(x_i, x_m))^{\top}$. Equivalently we have, for $i = 1, \ldots, m$, $\alpha_{\mathbf{z},i} = 0$ if and only if

$$\left|(\mathbf{y} - \boldsymbol{\mathcal{K}}\boldsymbol{\alpha}_{\mathbf{z}})^{\top}\boldsymbol{\mathcal{K}}_{x_i}\right| \leq \frac{\lambda_1}{2}.$$

Proof Recalling the fact that $\boldsymbol{\alpha}_{\mathbf{z}}$ is the unique minimizer of the risk functional (5) and the strong convexity property of $\mathcal{R}_{\mathbf{z}}(\boldsymbol{\alpha})$ with respect to $\boldsymbol{\alpha}$, we have

$$\mathbf{0} \in \partial\mathcal{R}_{\mathbf{z}}(\boldsymbol{\alpha}_{\mathbf{z}}) = 2(\mathbf{y} - \boldsymbol{\mathcal{K}}\boldsymbol{\alpha}_{\mathbf{z}})^{\top}\boldsymbol{\mathcal{K}} + \lambda_1\partial\|\boldsymbol{\alpha}_{\mathbf{z}}\|_1 + 2\lambda_2\boldsymbol{\alpha}_{\mathbf{z}}.$$

Therefore, for $i = 1, 2, \ldots, m$, there holds

$$0 \in \partial\mathcal{R}_{\mathbf{z}}(\alpha_{\mathbf{z},i}) = 2(\mathbf{y} - \boldsymbol{\mathcal{K}}\boldsymbol{\alpha}_{\mathbf{z}})^{\top}\boldsymbol{\mathcal{K}}_{x_i} + \lambda_1\partial|\alpha_{\mathbf{z},i}| + 2\lambda_2\alpha_{\mathbf{z},i}.$$

In the following, we discuss the three different cases when $\operatorname{sgn}(\alpha_{\mathbf{z},i})$ takes different values, i.e., $\alpha_{\mathbf{z},i} > 0$, $\alpha_{\mathbf{z},i} < 0$, and $\alpha_{\mathbf{z},i} = 0$. After simple computations, we come to the following conclusions:

(i) if $\alpha_{\mathbf{z},i} > 0$, then $\left|(\mathbf{y} - \boldsymbol{\mathcal{K}}\boldsymbol{\alpha}_{\mathbf{z}})^{\top}\boldsymbol{\mathcal{K}}_{x_i}\right| > \frac{\lambda_1}{2}$,

(ii) if $\alpha_{\mathbf{z},i} < 0$, then $\left|(\mathbf{y} - \boldsymbol{\mathcal{K}}\boldsymbol{\alpha}_{\mathbf{z}})^{\top}\boldsymbol{\mathcal{K}}_{x_i}\right| > \frac{\lambda_1}{2}$,

(iii) if $\alpha_{\mathbf{z},i} = 0$, then $\left|(\mathbf{y} - \boldsymbol{\mathcal{K}}\boldsymbol{\alpha}_{\mathbf{z}})^{\top}\boldsymbol{\mathcal{K}}_{x_i}\right| \leq \frac{\lambda_1}{2}$.

Summarizing the above three conclusions lead to the results in Theorem 1. This completes the proof of Theorem 1. ■

In light of Theorem 1, KENReg favors the characterizable sparsity property in the sense that it provides a necessary and sufficient condition to characterize the non-zero coefficients of $f_{\mathbf{z}}$. It follows that these non-zero coefficients of $f_{\mathbf{z}}$ are determined by the data-fitting risk, the regularity of the kernel function $\mathcal{K}$, and the regularization parameter λ_1. As a consequence of Theorem 1, confidence bounds on the sparseness of $f_{\mathbf{z}}$ can be established as done in Shi et al. (2011). According to Theorem 1, it seems that the regularization parameter λ_2 does not contribute to the sparseness of $f_{\mathbf{z}}$. In fact, this is not the case. For instance, by setting $\lambda_2 = 0$, KENReg can be reduced to the kernelized Lasso. However, due to the non-uniqueness of $f_{\mathbf{z},\eta}$, a necessary and sufficient condition that characterizes the support vectors of $f_{\mathbf{z},\eta}$ is not obtainable thus far. For the classical Lasso procedure, the characterization of the zero or nonzero coordinates of its output has been extensively

studied, the problem of which is usually termed as the oracle property, the variable selection property, the model selection, or the sparse recovery of the Lasso, see e.g., Zou (2006); Zhao and Yu (2006); van de Geer and Bühlmann (2009); Meinshausen and Yu (2009); Tibshirani (2013), among others. The significance of Theorem 1 lies in that it entails us to study similar inference problem with respect to KENReg. For example, by considering a fixed design setup and imposing certain prior information on $f^\star$, some initial studies on the sparse recovery of KENReg are conducted in Feng et al. (2016) recently.

3. Stability of Kernelized Elastic Net Regularization

KENReg is proposed in this study as a stabilized alternative to the kernelized Lasso. As mentioned in the introduction, the implications of stability here refers to the computational stability and the algorithmic stability. The algorithmic stability is in the sense of Bousquet and Elisseeff (2002). It is generally accepted that sparse learning algorithms cannot be stable (Xu et al., 2012). In this section, we will show that this is not the case for KENReg.

To discuss the computational stability, let us consider a deterministic sampling. In this case, discussions on the instability of the ℓ^1-regularization model as well as the computational efficiency of the elastic net regularization scheme can be found in Jin et al. (2009); Mosci et al. (2010). The ℓ^2-regularizer "*has a preconditioning effect on the iterative procedure and can substantially reduce the number of required computations without affecting the prediction property of the obtained solution*" (Mosci et al., 2010). To explain more on the computational efficiency of KENReg, let us consider the following more general optimization problem: $\min_{\alpha\in\mathbb{R}^n} f(\alpha) + g(\alpha)$, where f, g are both convex and g is non-smooth. To solve the above optimization problem, when f is strongly convex, some first-order methods, such as the proximal gradient methods and the accelerated proximal gradient methods (Beck and Teboulle, 2009), can be applied and linear convergence rates can be ensured Schmidt et al. (2011). However, when f is not strongly convex, only sublinear convergence rates can be obtained by the above methods (Beck and Teboulle, 2009). Additionally, comparing with the kernelized Lasso, KENReg has an additional ℓ^2-penalty term, which decreases the condition number of the Hessian of the smoothed term (e.g., from $\boldsymbol{\mathcal{K}}^\top\boldsymbol{\mathcal{K}}$ to $\boldsymbol{\mathcal{K}}^\top\boldsymbol{\mathcal{K}} + \lambda_2\mathbf{I}$). Therefore, theoretically, one may expect fewer iterations when solving the problem. Therefore, in the above sense, KENReg possesses the computational efficiency due to its strong convexity.

Coming from the sensitivity analysis, the algorithmic stability refers to the property that the learned function does not change much with small changes in the given sample $\mathbf{z}$. In the statistical machine learning literature, various notions of stability have been proposed to study supervised and unsupervised learning algorithms (Rogers and Wagner, 1978; Devroye and Wagner, 1979; Kearns and Ron, 1999; Bousquet and Elisseeff, 2002; Kutin and Niyogi, 2002; Poggio et al., 2004).

To continue our discussion on the stability property of KENReg, we follow notions of uniform stability in Bousquet and Elisseeff (2002). 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 any $1 \le i \le m$, $(x_i', y_i') \in \mathcal{X} \times \mathcal{Y}$ and $m \in \mathbb{N}$. $\mathcal{A}$ is denoted as a learning model with $\mathbf{z} \in (\mathcal{X} \times \mathcal{Y})^m$ as the input and $\mathcal{A}_\mathbf{z}$ as the output. Denote $a \wedge b := \min\{a, b\}$, and $a \vee b := \max\{a, b\}$ for $a, b \in \mathbb{R}$. In the regression setup, the uniform β-stable is defined as follows:

Definition 2 (Uniform β-stable (Bousquet and Elisseeff, 2002)) *Let $\beta : \mathbb{N} \to \mathbb{R}_+$. An algorithm $\mathcal{A}$ 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 $1 \leq i \leq m$, for all $(x, y) \in (\mathcal{X}, \mathcal{Y})$, there holds*

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

Theorem 3 KENReg is uniform β-stable with $\beta = \frac{8}{\lambda_2}\left(1 + \frac{2}{\lambda_1} \bigwedge \frac{2\sqrt{m}}{\sqrt{\lambda_2}}\right)^2$.

To prove Theorem 3, we need to introduce some notations. Let $f_{\boldsymbol{\alpha}}(x) = \sum_{i=1}^m \alpha_i \mathcal{K}(x_i, x)$ for any $x \in \mathcal{X}$ with $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_m)^\top$. The empirical risk $\mathcal{E}_{\mathbf{z}}(f_{\boldsymbol{\alpha}})$ is denoted as

$$\mathcal{E}_{\mathbf{z}}(f_{\boldsymbol{\alpha}}) = \frac{1}{m}\sum_{k=1}^m \left(y_k - f_{\boldsymbol{\alpha}}(x_k)\right)^2, \text{ and } \Omega(f_{\boldsymbol{\alpha}}) = \lambda_1 \|\boldsymbol{\alpha}\|_1 + \lambda_2 \|\boldsymbol{\alpha}\|_2^2.$$

When no ambiguity arises we also denote $\Omega(\boldsymbol{\alpha}) := \Omega(f_{\boldsymbol{\alpha}})$ and $\mathcal{E}_{\mathbf{z}}(\boldsymbol{\alpha}) := \mathcal{E}_{\mathbf{z}}(f_{\boldsymbol{\alpha}})$. The same notations of $\mathcal{E}_{\mathbf{z}}(\cdot)$ and $\mathcal{E}(\cdot)$ are also applied to f for any $f \in \mathcal{H}_{\mathcal{K}}$. Due to the dependency of λ_1 and λ_2 on m, we can slightly modify KENReg as follows:

$$\boldsymbol{\alpha}_{\mathbf{z}} = \arg\min_{\boldsymbol{\alpha} \in \mathbb{R}^m} \mathcal{R}_{\mathbf{z}}(\boldsymbol{\alpha}), \quad \text{where} \quad \mathcal{R}_{\mathbf{z}}(\boldsymbol{\alpha}) = \mathcal{E}_{\mathbf{z}}(\boldsymbol{\alpha}) + \Omega(\boldsymbol{\alpha}).$$

$\mathcal{E}_{\mathbf{z}^i}(\boldsymbol{\alpha})$ and $\mathcal{R}_{\mathbf{z}^i}(\boldsymbol{\alpha})$ are denoted as the perturbed versions of $\mathcal{E}_{\mathbf{z}}(\boldsymbol{\alpha})$ and $\mathcal{R}_{\mathbf{z}}(\boldsymbol{\alpha})$ as follows:

$$\mathcal{E}_{\mathbf{z}^i}(\boldsymbol{\alpha}) = \frac{1}{m}\left\{\sum_{k=1:\, k\neq i}^m \left(y_k - \sum_{l=1}^m \alpha_l \mathcal{K}(x_l, x_k)\right)^2 + \left(y_i' - \sum_{l=1}^m \alpha_l \mathcal{K}(x_l, x_i')\right)^2\right\},$$

and

$$\mathcal{R}_{\mathbf{z}^i}(\boldsymbol{\alpha}) = \mathcal{E}_{\mathbf{z}^i}(\boldsymbol{\alpha}) + \Omega(\boldsymbol{\alpha}).$$

We also denote $\boldsymbol{\alpha}_{\mathbf{z}^i}$ as the output coefficient learned from KENReg (3) with the input $\mathbf{z}^i$.

Proposition 4 *For any $0 \leq t \leq 1$, we have*

$$\Omega(\boldsymbol{\alpha}_{\mathbf{z}}) - \Omega(\boldsymbol{\alpha}_{\mathbf{z}} + t\triangle\boldsymbol{\alpha}) + \Omega(\boldsymbol{\alpha}_{\mathbf{z}^i}) - \Omega(\boldsymbol{\alpha}_{\mathbf{z}^i} - t\triangle\boldsymbol{\alpha}) \lesssim \frac{4t}{m}\left(1 + \frac{2}{\lambda_1} \bigwedge \frac{2\sqrt{m}}{\sqrt{\lambda_2}}\right) \|f_{\mathbf{z}} - f_{\mathbf{z}^i}\|_\infty,$$

where $\triangle\boldsymbol{\alpha} := \boldsymbol{\alpha}_{\mathbf{z}^i} - \boldsymbol{\alpha}_{\mathbf{z}}$.

Proof [Proof of Proposition 4] Recall that a convex function $g : \mathcal{S} \to \mathbb{R}$ satisfies that for all $\alpha, \beta \in \mathcal{S}$ and $t \in [0, 1]$, there holds

$$g(\alpha + t(\beta - \alpha)) - g(\alpha) \leq t(g(\beta) - g(\alpha)). \tag{6}$$

The convexity property of $\mathcal{E}_{\mathbf{z}}(\boldsymbol{\alpha})$ and $\mathcal{E}_{\mathbf{z}^i}(\boldsymbol{\alpha})$ with respect to $\boldsymbol{\alpha}$ is obvious due to the convexity of the least squares loss. Following from (6), we know that

$$\mathcal{E}_{\mathbf{z}^i}(\boldsymbol{\alpha}_{\mathbf{z}} + t\triangle\boldsymbol{\alpha}) - \mathcal{E}_{\mathbf{z}^i}(\boldsymbol{\alpha}_{\mathbf{z}}) \leq t(\mathcal{E}_{\mathbf{z}^i}(\boldsymbol{\alpha}_{\mathbf{z}^i}) - \mathcal{E}_{\mathbf{z}^i}(\boldsymbol{\alpha}_{\mathbf{z}})). \tag{7}$$

Similarly, (6) also implies

$$\mathcal{E}_{\mathbf{z}^i}(\boldsymbol{\alpha}_{\mathbf{z}^i} - t\triangle\boldsymbol{\alpha}) - \mathcal{E}_{\mathbf{z}^i}(\boldsymbol{\alpha}_{\mathbf{z}^i}) \leq t(\mathcal{E}_{\mathbf{z}^i}(\boldsymbol{\alpha}_{\mathbf{z}}) - \mathcal{E}_{\mathbf{z}^i}(\boldsymbol{\alpha}_{\mathbf{z}^i})). \tag{8}$$

Combining inequalities (7) and (8), we obtain

$$\mathcal{E}_{\mathbf{z}^i}(\boldsymbol{\alpha}_{\mathbf{z}} + t\triangle\boldsymbol{\alpha}) - \mathcal{E}_{\mathbf{z}^i}(\boldsymbol{\alpha}_{\mathbf{z}}) + \mathcal{E}_{\mathbf{z}^i}(\boldsymbol{\alpha}_{\mathbf{z}^i} - t\triangle\boldsymbol{\alpha}) - \mathcal{E}_{\mathbf{z}^i}(\boldsymbol{\alpha}_{\mathbf{z}^i}) \leq 0. \tag{9}$$

On the other hand, considering that $\boldsymbol{\alpha}_{\mathbf{z}}$ and $\boldsymbol{\alpha}_{\mathbf{z}^i}$ are minimizers of the risk functionals $\mathcal{R}_{\mathbf{z}}(\cdot)$ and $\mathcal{R}_{\mathbf{z}^i}(\cdot)$ respectively, there holds

$$\mathcal{R}_{\mathbf{z}}(\boldsymbol{\alpha}_{\mathbf{z}}) - \mathcal{R}_{\mathbf{z}}(\boldsymbol{\alpha}_{\mathbf{z}} + t\triangle\boldsymbol{\alpha}) \leq 0, \tag{10}$$

and

$$\mathcal{R}_{\mathbf{z}^i}(\boldsymbol{\alpha}_{\mathbf{z}^i}) - \mathcal{R}_{\mathbf{z}^i}(\boldsymbol{\alpha}_{\mathbf{z}^i} - t\triangle\boldsymbol{\alpha}) \leq 0. \tag{11}$$

Combining inequalities (9), (10), and (11), we obtain

$$\begin{aligned}\mathcal{E}_{\mathbf{z}^i}(\boldsymbol{\alpha}_{\mathbf{z}} + t\triangle\boldsymbol{\alpha}) - \mathcal{E}_{\mathbf{z}^i}(\boldsymbol{\alpha}_{\mathbf{z}}) + \mathcal{E}_{\mathbf{z}^i}(\boldsymbol{\alpha}_{\mathbf{z}^i} - t\triangle\boldsymbol{\alpha}) - \mathcal{E}_{\mathbf{z}^i}(\boldsymbol{\alpha}_{\mathbf{z}^i}) + \mathcal{R}_{\mathbf{z}}(\boldsymbol{\alpha}_{\mathbf{z}}) \\ -\mathcal{R}_{\mathbf{z}}(\boldsymbol{\alpha}_{\mathbf{z}} + t\triangle\boldsymbol{\alpha}) + \mathcal{R}_{\mathbf{z}^i}(\boldsymbol{\alpha}_{\mathbf{z}^i}) - \mathcal{R}_{\mathbf{z}^i}(\boldsymbol{\alpha}_{\mathbf{z}^i} - t\triangle\boldsymbol{\alpha}) \leq 0.\end{aligned}$$

Recalling the definitions of the functionals $\mathcal{E}_{\mathbf{z}}(\cdot)$, $\mathcal{E}_{\mathbf{z}^i}(\cdot)$, $\mathcal{R}_{\mathbf{z}}(\cdot)$, and $\mathcal{R}_{\mathbf{z}^i}(\cdot)$, we see that

$$\begin{aligned}&m(\Omega(\boldsymbol{\alpha}_{\mathbf{z}}) - \Omega(\boldsymbol{\alpha}_{\mathbf{z}} + t\triangle\boldsymbol{\alpha}) + \Omega(\boldsymbol{\alpha}_{\mathbf{z}^i}) - \Omega(\boldsymbol{\alpha}_{\mathbf{z}^i} - t\triangle\boldsymbol{\alpha}) \\ &\leq (y_i' - f_{\mathbf{z}}(x_i'))^2 - (y_i - f_{\mathbf{z}}(x_i))^2 + (y_i - (f_{\mathbf{z}} + t\triangle f_{\mathbf{z}})(x_i))^2 - (y_i' - (f_{\mathbf{z}} + t\triangle f_{\mathbf{z}})(x_i'))^2) \\ &\leq |t\triangle f_{\mathbf{z}}(x_i')|(|2y_i'| + 2|f_{\mathbf{z}}(x_i')| + t|\triangle f_{\mathbf{z}}(x_i')|) + |t\triangle f_{\mathbf{z}}(x_i)|(|2y_i| + 2|f_{\mathbf{z}}(x_i)| + t|\triangle f_{\mathbf{z}}(x_i)|) \\ &\leq 2t\|f_{\mathbf{z}} - f_{\mathbf{z}^i}\|_\infty(2 + 3\|f_{\mathbf{z}}\|_\infty + \|f_{\mathbf{z}^i}\|_\infty).\end{aligned}$$

On the other hand, according to the definition of $\boldsymbol{\alpha}_{\mathbf{z}}$ in (3), we know that

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

Applying the Hölder's inequality, we see that

$$\|f_{\mathbf{z}}\|_\infty = \left\|\sum_{k=1}^m \alpha_k\mathcal{K}(x_k, \cdot)\right\|_\infty \lesssim \|\boldsymbol{\alpha}_{\mathbf{z}}\|_1 \leq \sqrt{m}\|\boldsymbol{\alpha}_{\mathbf{z}}\|_2.$$

Consequently, we have $\|f_{\mathbf{z}}\|_\infty \lesssim \lambda_1^{-1} \wedge \sqrt{m\lambda_2^{-1}}$. Analogously, $\|f_{\mathbf{z}^i}\|_\infty \lesssim \lambda_1^{-1} \wedge \sqrt{m\lambda_2^{-1}}$. Combining the above estimates, we complete the proof of Proposition 4. ■

Proof [Proof of Theorem 3] Applying Proposition 4 with $t = \frac{1}{2}$, we know that

$$\Omega(\boldsymbol{\alpha}_{\mathbf{z}}) - \Omega(\boldsymbol{\alpha}_{\mathbf{z}} + \frac{1}{2}\triangle\boldsymbol{\alpha}) + \Omega(\boldsymbol{\alpha}_{\mathbf{z}^i}) - \Omega(\boldsymbol{\alpha}_{\mathbf{z}^i} - \frac{1}{2}\triangle\boldsymbol{\alpha}) \lesssim \frac{2}{m}\left(1 + \frac{2}{\lambda_1} \bigwedge \frac{2\sqrt{m}}{\sqrt{\lambda_2}}\right)\|f_{\mathbf{z}} - f_{\mathbf{z}^i}\|_\infty.$$

On the other hand, due to the strict convexity of $\Omega(\boldsymbol{\alpha})$ with respect to $\boldsymbol{\alpha}$, it is easy to deduce that

$$\frac{\lambda_2}{2}\|\boldsymbol{\alpha}_{\mathbf{z}}-\boldsymbol{\alpha}_{\mathbf{z}^i}\|_2^2 \leq \Omega(\boldsymbol{\alpha}_{\mathbf{z}})-\Omega(\boldsymbol{\alpha}_{\mathbf{z}}+\frac{1}{2}\triangle\boldsymbol{\alpha})+\Omega(\boldsymbol{\alpha}_{\mathbf{z}^i})-\Omega(\boldsymbol{\alpha}_{\mathbf{z}^i}-\frac{1}{2}\triangle\boldsymbol{\alpha}).$$

Combining the above two inequalities, we get

$$\frac{\lambda_2}{2}\|\boldsymbol{\alpha}_{\mathbf{z}}-\boldsymbol{\alpha}_{\mathbf{z}^i}\|_2^2 \lesssim \frac{2}{m}\left(1+\frac{2}{\lambda_1}\bigwedge\frac{2\sqrt{m}}{\sqrt{\lambda_2}}\right)\|f_{\mathbf{z}}-f_{\mathbf{z}^i}\|_\infty \lesssim \frac{2}{\sqrt{m}}\left(1+\frac{2}{\lambda_1}\bigwedge\frac{2\sqrt{m}}{\sqrt{\lambda_2}}\right)\|\boldsymbol{\alpha}_{\mathbf{z}}-\boldsymbol{\alpha}_{\mathbf{z}^i}\|_2.$$

As a result, the following estimate holds

$$\|\boldsymbol{\alpha}_{\mathbf{z}}-\boldsymbol{\alpha}_{\mathbf{z}^i}\|_2 \lesssim \frac{4}{\lambda_2\sqrt{m}}\left(1+\frac{2}{\lambda_1}\bigwedge\frac{2\sqrt{m}}{\sqrt{\lambda_2}}\right).$$

Recalling the definition of uniform β-stable, for any $(x,y)\in\mathcal{X}\times\mathcal{Y}$, there holds

$$|(y-f_{\mathbf{z}}(x))^2-(y-f_{\mathbf{z}^i}(x))^2| \lesssim (2+\|f_{\mathbf{z}}\|_\infty+\|f_{\mathbf{z}^i}\|_\infty)\|f_{\mathbf{z}}-f_{\mathbf{z}^i}\|_\infty \lesssim \frac{8}{\lambda_2}\left(1+\frac{2}{\lambda_1}\bigwedge\frac{2\sqrt{m}}{\sqrt{\lambda_2}}\right)^2.$$

That is, KENReg is uniform β-stable with $\beta=\frac{8}{\lambda_2}\left(1+\frac{2}{\lambda_1}\bigwedge\frac{2\sqrt{m}}{\sqrt{\lambda_2}}\right)^2$. This completes the proof of Theorem 3. ■

According to Theorem 3, KENReg is uniform β-stable. This is, in fact, the most prominent advantage of KENReg over the kernelized Lasso that motivates this study. Several remarks towards this can be made here:

- As mentioned earlier, sparse learning algorithms are generally thought of being unstable (Xu et al., 2012). However, from the above discussions, we know that KENReg advocates a sparse empirical target function and possesses the stability property simultaneously. That is to say, in the regime of kernel dictionary based learning, sparseness and stability are not necessarily to be mutually exclusive, as argued in Feng et al. (2016). However, this property does not apply to the kernelized Lasso.

- Stability is frequently pursued when designing machine learning algorithms since, as we shall discuss later, it is closely related to their generalization ability, see e.g., Kearns and Ron (1999); Kutin and Niyogi (2002); Bousquet and Elisseeff (2002); Poggio et al. (2004); Mukherjee et al. (2006); Shalev-Shwartz et al. (2010). On the other hand, from a robust statistical point of view, the stability of an algorithm can be considered as a weak notion of robustness (Breiman, 1996; Yu, 2013), which plays an important role in statistical estimation.

- It should be noted from the above discussion that the motivation of imposing the elastic net type penalty term in KENReg is essentially different from its original motivation in Zou and Hastie (2005). Being confined to a linear model, the elastic net regularization in Zou and Hastie (2005) aims primarily at selecting correlated covariates. While within a kernel-based learning framework, KENReg in this study serves as a stabilized sparsity-producing regression estimator that outperforms the kernelized Lasso.

- We also remark that in the regime of kernel dictionary based learning, we give some first discussions towards the stability issue. Some related properties may be pursued under relaxed notions of stability, see e.g., Bousquet and Elisseeff (2002); Mukherjee et al. (2006).

In the statistical learning literature, it has been well understood that the algorithmic stability of an algorithm is closely connected to its generalization ability. Here the generalization ability is referred to as the ability that a learning machine can generalize on unseen data. More specifically, the stability property of an algorithm is usually employed to study its generalization ability when it cannot be easily analyzed via capacity-dependent approaches (Maurer, 2005; Agarwal and Niyogi, 2009). However, in the context of KENReg, one cannot count on the uniform stability constant β in Theorem 3 to get meaningful generalization bounds, due to its lack of dependence on m^{-1}. In addition, the generalization bounds obtained from the stability properties of the model do not make use of the variance information of the function-valued random variables, and consequently usually are loose. In view of these, in the following section, we will pursue capacity-dependent generalization bounds by applying and extending the stepping-stone technique introduced in Wu and Zhou (2005).

Before proceeding to derive the generalization bounds for KENReg, let us summarize the advantages of KENReg over the kernelized Lasso in the following table.

Table 1: Comparisons on properties of the kernelized Lasso and KENReg

	strong convexity	characterizable sparsity	computational efficiency	algorithmic stability
Kernelized Lasso	×	×	×	×
KENReg	✓	✓	✓	✓

4. Generalization Bounds of Kernelized Elastic Net Regularization

Analyzing the generalization ability of a learning machine is one of the focal research problems in statistical learning theory. Denote $\rho_{\mathcal{X}}$ as the marginal distribution of ρ on $\mathcal{X}$ and $f_\rho(x) := \mathbb{E}(y \,|\, \mathcal{X} = x)$ for any $x \in \mathcal{X}$ as the regression function. Due to the zero-mean noise assumption, almost surely we have $f_\rho = f^\star$. The generalization bounds of KENReg refer to the convergence rates of $\|f_{\mathbf{z}} - f^\star\|^2_{L^2_{\rho_{\mathcal{X}}}}$.

4.1 Generalization Bounds with Positive Semidefinite Kernels

We first consider the case when the kernel $\mathcal{K}$ is restricted to a positive semidefinite kernel. In this case, we denote $\mathcal{H}_{\mathcal{K}}$ as the reproducing kernel Hilbert space (RKHS) induced by $\mathcal{K}$. Denote $\|\cdot\|_{\mathcal{K}}$ as the norm induced by the inner product $\langle\cdot,\cdot\rangle = \langle\cdot,\cdot\rangle_{\mathcal{K}}$, which satisfies $\langle \mathcal{K}_u, \mathcal{K}_v\rangle = \mathcal{K}(u,v)$ for $u, v \in \mathcal{X}$, with $\mathcal{K}_u := \mathcal{K}(u,\cdot) : \mathcal{X} \to \mathbb{R}$ for any fixed $u \in \mathcal{X}$. In what follows, the notation $f \lesssim g$ means that there exists a universal positive constant c such that $f \le c\, g$ for any functions f and g on $\mathbb{R}$.

Assumption 5 *We make the following assumptions:*

(i) $|y| \leq M$, *for some* $M > 0$ *and without loss of generality we set* $M = 1$.

(ii) There exists a positive constant p *with* $0 < p < 2$, *such that*

$$\log \mathcal{N}_2(B_1, \epsilon) \lesssim \epsilon^{-p}, \text{ for all } \epsilon > 0,$$

where B_1 *is the unit ball of* $\mathcal{H}_K$ *defined by* $B_R = \{f \in \mathcal{H}_K : \|f\|_K \leq R\}$ *with* $R = 1$, *and* $\mathcal{N}_2(B_1, \epsilon)$ *denotes the* ℓ^2*-empirical covering number of* B_1, *the definition of which is given in Appendix A.*

(iii) There exists a constant $0 < \beta \leq 1$, *such that for any* $\lambda > 0$, *there holds*

$$\inf_{f \in \mathcal{H}_K} \left\{ \|f - f_\rho\|^2_{\mathcal{L}^2_{\rho_\mathcal{X}}} + \lambda \|f\|^2_K \right\} \lesssim \lambda^\beta.$$

Condition (i) on the boundedness of Y in Assumption 5 is frequently applied in the statistical learning theory literature (Vapnik, 1998; Cucker and Zhou, 2007; Steinwart and Christmann, 2008; Mendelson and Neeman, 2010). From Shi et al. (2011) we know that Condition (ii) holds with $0 < p < 2$ when $\mathcal{K} \in \mathcal{C}^s(\mathcal{X} \times \mathcal{X})$ with $s > 0$. Condition (iii) is the model assumption that reflects the approximation ability of the associated hypothesis space, which is typical in learning theory (Smale and Zhou, 2003). Specifically, when $f_\rho \in \mathcal{H}_K$, there holds $\beta = 1$.

Notice that the boundedness assumption implies that $|f_\rho(x)| \leq 1$ for any $x \in \mathcal{X}$. Consequently, it is natural to require that the learned output function can be also upper bounded by 1. In view of this and also for ease of simplifying our analysis, we introduce the projected version, $\bar{f}_\mathbf{z}$, of the empirical target function $f_\mathbf{z}$. For any $x \in \mathcal{X}$, $\bar{f}_\mathbf{z}(x)$ takes the value 1 if $f_\mathbf{z}(x) > 1$ and takes the value -1 if $f_\mathbf{z}(x) < -1$. It equals $f_\mathbf{z}(x)$ if $|f_\mathbf{z}(x)| \leq 1$ for any $x \in \mathcal{X}$. This projection operator was introduced in Bartlett (1998) to analyze classification schemes induced by neural networks. It was then introduced into learning scenarios in the context of support vector machines (Bousquet and Elisseeff, 2002; Chen et al., 2004; Wu et al., 2007; Steinwart et al., 2006). Generalization bounds provided below are derived with respect to the projected empirical target function $\bar{f}_\mathbf{z}$.

Theorem 6 *Suppose that Assumption 5 holds and* $f_\mathbf{z}$ *is given by (4). For any* $0 < \delta < 1$, *with confidence* $1 - \delta$, *there holds*

$$\left\|\bar{f}_\mathbf{z} - f_\rho\right\|^2_{\mathcal{L}^2_{\rho_\mathcal{X}}} \lesssim \log\left(\frac{2}{\delta}\right)\left(\frac{\lambda^{\beta+1} + \lambda_2 + m\lambda_1\lambda + m\lambda^{\beta+2}}{m\lambda^2} + \left(\frac{1}{\lambda_1} \bigwedge \sqrt{\frac{m}{\lambda_2}}\right)^{\frac{2p}{2+p}} \left(\frac{1}{m}\right)^{\frac{2}{2+p}}\right),$$

where $\lambda > 0$ *is a parameter that will be specified in the proof.*

The following corollary is a consequence of Theorem 6.

Corollary 7 *Under the assumptions of Theorem 6, let* $\lambda_1 = m^{-\theta_1}$, *and* $\lambda_2 = m^{-\theta_2}$. *For any* $0 < \delta < 1$, *with confidence* $1 - \delta$ *there holds*

$$\left\|\bar{f}_\mathbf{z} - f_\rho\right\|^2_{\mathcal{L}^2_{\rho_\mathcal{X}}} \lesssim \left(\log\frac{2}{\delta}\right)\left(\frac{1}{m}\right)^{\beta\Theta/(2+p)}$$

with

$$\Theta = \begin{cases} \theta_1(2+3p)-2, & \textit{for case I,} \\ \theta_2(1+p)+p, & \textit{for case II,} \end{cases}$$

where

$$\textit{case I}: \begin{cases} \theta_1 \in \left[\frac{1}{2} \bigwedge \frac{6+p+2\beta}{(\beta+2)(2+3p)}, \frac{4+p}{2+3p} \bigwedge \frac{2(\beta+1)}{(2+3p)(\beta+1)-(2+p)}\right], \\ \theta_2 \in \left[(2\theta_1-1) \bigvee \frac{((2+3p)\theta_1-2)(\beta+2)-2-p}{2+p}, +\infty\right], \end{cases}$$

and

$$\textit{case II}: \begin{cases} \theta_1 \in \left[\frac{1+\theta_2}{2} \bigvee \frac{(\beta+1)((1+p)\theta_2+p)}{2+p}, +\infty\right], \\ \theta_2 \in \left(0, \frac{2}{1+p} \bigwedge \frac{2-p\beta-p}{(1+p)(\beta+2)-2-p}\right] \ \textit{and} \ p \in (0, 2/(1+\beta)). \end{cases}$$

The proofs of Theorem 6 and Corollary 7 are provided in Appendix B. From Corollary 7, we see that with properly chosen parameters λ_1 and λ_2, when $\beta = 1$ and p tends to 0, generalization bounds of the type $\mathcal{O}(m^{-1})$ can be achieved. In fact, from the proof of Corollary 7, we see that the derived generalization bounds not only rely on the ℓ^1-regularization term but also on the ℓ^2-regularizer. More specifically, when the ℓ^1-penalty term plays a leading role, the obtained bounds will be controlled by the parameter λ_1. On the other hand, when the ℓ^2-regularizer is dominant in the regularization term, λ_2 will play a leading role in the derived generalization bounds. That is, the ℓ^2 regularization term in KENReg does help to generalize.

In light of Theorem 6 and Corollary 7, λ_2 should tend to zero when the sample size m goes to infinity in order to ensure the generalization consistency of KENReg. However, the stability results in Theorem 3 suggest that $\lambda_2 \to \infty$ when $m \to \infty$ so that KENReg can generalize. This seeming contradiction is due to the fact that the generalization results in Theorem 6 and Corollary 7 are derived with respect to the projected version of $f_{\mathbf{z}}$ instead of $f_{\mathbf{z}}$ itself while the stability analysis is conducted with respect to $f_{\mathbf{z}}$. In fact, some further study conducted in Feng et al. (2016) confirms that $f_{\mathbf{z}}$ is generalization consistent under the condition that λ_2 takes the form $\lambda_2 = m^{\theta}$ for some $\theta > 0$.

4.2 Generalization Bounds with Indefinite Kernels

Indefinite kernels refer to kernels that are not positive semidefinite. KENReg learns from the kernelized dictionary $\mathcal{D}_m$, which does not limit us to the representer theorem and consequently allows us to use various indefinite kernels. However, as explained later, this also brings inherent barriers in deriving generalization bounds. In fact, we will show that when the kernels are indefinite, generalization bounds for KENReg can be also derived by using the stepping-stone technique.

The basic idea is to as follows: first, one constructs a proxy positive semidefinite kernel from the indefinite kernel, and then, one introduces an instrumental empirical target function which is generated by the kernel ridge regression scheme induced by the proxy kernel and the same given observations. With the proxy definite kernel and the instrumental empirical target function at hand, one can again apply the stepping-stone technique. By introducing a novel error decomposition method, we can bound the generalization error. Detailed steps in deriving the generalization bounds for this case are listed in Appendix C.

4.3 Comments on the Generalization Analysis for KENReg

We now give some comments on the barriers and ideas in the generalization bounds analysis for KENReg. To this end, we remind that within the statistical learning framework it is crucial to control the complexity of the hypothesis space when deriving capacity-dependent generalization bounds. The hypothesis space in which KENReg works can be denoted as $\mathcal{H}_{\mathcal{K},\mathbf{z}} = \{f_{\boldsymbol{\alpha}} \mid f_{\boldsymbol{\alpha}}(x) := \sum_{i=1}^{m} \alpha_i \mathcal{K}(x_i, x) : \alpha_i \in \mathbb{R}, x \in \mathcal{X}\}$. As noticed in Wu and Zhou (2008), the hypothesis spaces $\mathcal{H}_{\mathcal{K},\mathbf{z}}$ are instance-dependent and consequently lead to instance-dependent biases (i.e., instance-dependent approximation errors). This gives rise to the first barrier encountered in deriving the generalization bounds for KENReg since the instance-dependent biases disable usual techniques of analyzing the learning schemes (Cucker and Zhou, 2007; Steinwart and Christmann, 2008). A strategy provided in Wu and Zhou (2008) is to consider a universal hypothesis $\mathcal{H}_0$, e.g., unifying $\mathcal{H}_{\mathcal{K},\mathbf{z}}$ with respect to $\mathbf{z}$. Basically, one needs to ensure that $\mathcal{H}_0 \supseteq \cup_{m\in\mathbb{N}} \cup_{\mathbf{z}\in(\mathcal{X}\times\mathcal{Y})^m} \mathcal{H}_{\mathcal{K},\mathbf{z}}$. This was adopted in analyzing the generalization property of the kernelized Lasso in Xiao and Zhou (2010); Shi et al. (2011) by taking the approximation ability of the universal hypothesis space $\mathcal{H}_0$ into account and using capacity-involved arguments, where $\mathcal{H}_0$ is defined as $\mathcal{H}_0 := \{f : f = \sum_{i=1}^{\infty} \alpha_i \mathcal{K}_{u_i}, \{\alpha_i\} \subset \ell^1, \{u_i\} \subset \mathcal{X}\}$ with the norm $\|f\| := \inf\big\{\sum_{i=1}^{\infty} |\alpha_i| : f = \sum_{i=1}^{\infty} \alpha_i \mathcal{K}_{u_i}\big\}$.

For regularized learning models, their complexities are controlled by the regularization parameters. However, in the context of KENReg, another technical difficulty in deriving its generalization bounds comes up since the radius of the hypothesis space not only depends on the parameter λ_1 but also on λ_2. In fact, a closer look reveals that one needs to measure the capacity of the function set $\mathcal{F}_{\lambda_1,\lambda_2} := \{f \mid \|f\|_\infty \le \lambda_1^{-1} \wedge \sqrt{m\lambda_2^{-1}},\ f \in \mathcal{H}_{\mathcal{K},\mathbf{z}},\ \mathbf{z} \in (\mathcal{X}\times\mathcal{Y})^m\}$. It is obvious that the upper bounds of the supremum norm of functions in $\mathcal{F}_{\lambda_1,\lambda_2}$ are specified by the two parameters λ_1 and λ_2. This together with the instance-dependent fact of $\mathcal{H}_{\mathcal{K},\mathbf{z}}$ again makes it non-trivial. The third difficulty in analyzing KENReg is due to the lack of understanding on the contribution of the ℓ^2-regularizer to the approximation error, i.e., $\mathbb{E}f_{\mathbf{z}} - f_\rho$. For a fixed hypothesis space, De Mol et al. (2009) showed that under some regularity conditions on f_ρ, the approximation error goes to zero, provided that λ_1, λ_2 tend to zero and λ_1/λ_2 is equal to a certain positive constant. However, this does not illustrate the role that λ_2 plays in the approximation error since the same observation can be obtained even if one simply sets $\lambda_2 = 0$, i.e., the case of the kernelized Lasso (Xiao and Zhou, 2010).

In this study, to overcome these difficulties, the stepping-stone technique introduced in Wu and Zhou (2005) is employed and extended to derive probabilistic generalization bounds. The idea of the stepping-stone technique in the generalization analysis is that instead of estimating the generalization bounds of the empirical target function $f_{\mathbf{z}}$ directly, one estimates the generalization bounds of an instrumental empirical target function and then evaluates the distance between the two empirical estimators. Constructing an instrumental empirical target function and choosing the hypothesis space that the instrumental empirical target function works in are crucial when applying the stepping-stone technique. When the kernel $\mathcal{K}$ is positive semidefinite, this analysis can be easily conducted since the unification $\mathcal{H}_0$ lies in the RKHS $\mathcal{H}_{\mathcal{K}}$. As we shall see later, in this case, the instrumental empirical target function could be chosen as the kernel ridge regression estimator obtained from the same kernel $\mathcal{K}$ and observations $\mathbf{z}$. The generalization ability of the kernel ridge regression estimator has been well understood. However, when the kernel $\mathcal{K}$ is indefinite, the analysis

could be more complicated since in general there does not exist a RKHS that contains the unification $\mathcal{H}_0$. In this case, we first construct a proxy positive semidefinite kernel based on the indefinite kernel and then apply the stepping-stone technique, as mentioned in the preceding subsection. The generalization analysis procedures for KENReg mentioned above are illustrated in Figs. 1 and 2. Detailed analysis and proofs for the two different situations are provided in Appendices B and C.

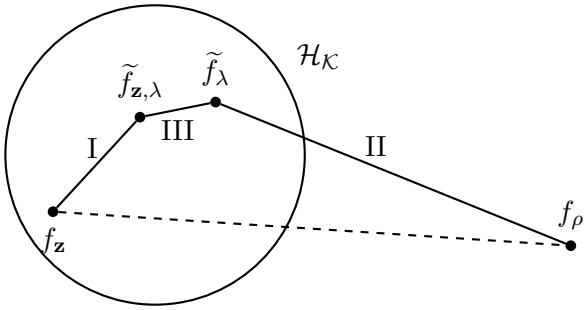

Figure 1: An illustration of the idea in analyzing the generalization bounds for KENReg when the kernel is positive semidefinite. Instead of estimating the $\mathcal{L}^2_{\rho_{\mathcal{X}}}$-distance between $f_{\mathbf{z}}$ and f_ρ directly, one introduces an instrumental empirical target function $\widetilde{f}_{\mathbf{z},\lambda}$ and decomposes the distance into I, II, and III. $\widetilde{f}_\lambda$ is the population version of $\widetilde{f}_{\mathbf{z},\lambda}$.

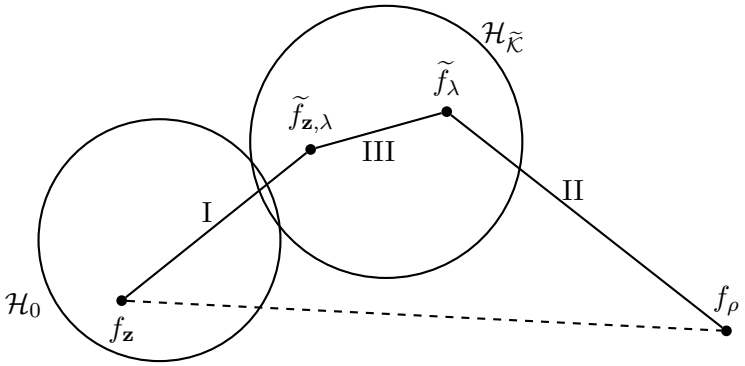

Figure 2: An illustration of the idea in analyzing the generalization bounds for KENReg when the kernel is indefinite. Similarly, instead of estimating the $\mathcal{L}^2_{\rho_{\mathcal{X}}}$-distance between $f_{\mathbf{z}}$ and f_ρ directly, one introduces an instrumental empirical target function $\widetilde{f}_{\mathbf{z},\lambda}$ and decomposes the distance into I, II, and III. $\widetilde{f}_\lambda$ is the population version of $\widetilde{f}_{\mathbf{z},\lambda}$, $\mathcal{H}_0$ is the unification of $\mathcal{H}_{\mathcal{K},\mathbf{z}}$ with respect to $\mathbf{z}$, and $\mathcal{H}_{\widetilde{\mathcal{K}}}$ is the RKHS induced by the proxy Mercer kernel $\widetilde{\mathcal{K}}$.

5. Comparisons with Related Studies

In the statistical learning literature, as far as we know, only few existing work towards theoretical investigations on the elastic net regularization model can be found. However, it should be mentioned that De Mol et al. (2009) presented interesting investigations into the elastic net regularization problems, which have some connections with our study.

Suppose that Γ is an index set, $\{\phi_\gamma\}_{\gamma\in\Gamma}$ represents the set of dictionaries, and $\{\omega_\gamma\}_{\gamma\in\Gamma}$ with $\omega_\gamma \geq 0$ stands for a set of weights. Concerning the data-generating model (1), it is

assumed in De Mol et al. (2009) that the regression function $f^\star$ admits the following sparse representation on the dictionary:

$$f^\star := f_{\boldsymbol{\beta}^\star} = \sum_{\gamma\in\Gamma} \beta^\star_\gamma \phi_\gamma, \text{ with } \boldsymbol{\beta}^\star = (\beta^\star_\gamma)_{\gamma\in\Gamma} \in \ell^2, \text{ and } \sum_{\gamma\in\Gamma} \omega_\gamma |\beta^\star_\gamma| < \infty.$$

Given the i.i.d observations $\mathbf{z}$, under a uniform boundedness assumption on the dictionary, De Mol et al. (2009) learns the coefficient vector $\boldsymbol{\beta}_{\mathbf{z}} = (\beta_{\mathbf{z},\gamma})_{\gamma\in\Gamma}$ from $\ell^2(\Gamma)$ via the following elastic net regularization model

$$\boldsymbol{\beta}_{\mathbf{z}} := \arg\min_{\beta\in\ell^2} \frac{1}{m}\sum_{i=1}^m (y_i - \sum_{\gamma\in\Gamma} \beta_\gamma \phi_\gamma(x_i))^2 + \sum_{\gamma\in\Gamma} (\lambda_1 \omega_\gamma |\beta_\gamma| + \lambda_2 \beta_\gamma^2), \tag{12}$$

where λ_1, $\lambda_2 > 0$ are regularization parameters. The resulting empirical target function of (12) is given as $f_{\boldsymbol{\beta}_{\mathbf{z}}} = \sum_{\gamma\in\Gamma} \beta_{\mathbf{z},\gamma}\phi_\gamma$. The variable selection consistency, as well as the regression consistency, is reported in De Mol et al. (2009) which is derived via a capacity-independent approach. There are several differences between De Mol et al. (2009) and our work:

- The first and also the most significant difference lies in the motivations of enforcing the elastic net regularization term. In De Mol et al. (2009), by making use of the elastic net regularizer, they aim at seeking the underlying true model and their main concerns are the variable identification abilities of the model. However, in our setup, by introducing the elastic net regularizer, our purpose is to seek a stable and sparse approximation to the regression function. In fact, the elastic net regularization model in De Mol et al. (2009) is feature-wise while KENReg in our work is instance-wise;

- The second important difference concerns theoretical results on convergence rates. Aiming at selecting variables, De Mol et al. (2009) showed the variable selection consistency property of (12) and derived probabilistic convergence rates for $\|\boldsymbol{\beta}_{\mathbf{z}} - \boldsymbol{\beta}^\star\|_2$. As commented there, the variable selection consistency implies the regression consistency. For instance, under certain conditions, they reported the following generalization bounds: for any $0 < \delta < 1$, with confidence $1 - \delta$, there holds

$$\left\| f_{\boldsymbol{\beta}_{\mathbf{z}}} - f_{\boldsymbol{\beta}^\star} \right\|^2_{\mathcal{L}^2_{\rho_\mathcal{X}}} \lesssim \log(4/\delta)\lambda_2^{-2}(m^{-1} + m^{-2}\lambda_1^{-1}). \tag{13}$$

  It is obvious that the generalization bounds reported in our study can be sharper with a nontrivial choice of the regularization parameter λ_2, e.g, $\lambda_2 = \mathcal{O}(m^{-\tau})$ with certain $\tau > 0$;

- The third important difference concerns the interpretation of the regularization parameter λ_2. In our study, Theorem 6 together with Corollary 7 shows that the generalization ability of KENReg depends on the leading term between the ℓ^1 and ℓ^2 regularizers. However, similar interpretations cannot be found explicitly in De Mol et al. (2009). For instance, one can not expect any non-trivial convergence rates from (13) by letting $\lambda_2 = 0$.

6. Conclusion

The kernelized Lasso is a sparse kernel-based learning algorithm and has been well studied in the literature. However, it is noticed that the kernelized Lasso is not stable. In this paper, another sparse kernel-based learning algorithm, namely, the kernelized elastic net regularization scheme, was proposed as an stabilized alternative to the kernelized Lasso. Compared with the kernelized Lasso, an additional ℓ^2-regularizer was employed in the proposed model. More specifically, several nice properties were delivered to the proposed model with the introduction of the ℓ^2-regularizer, which included the characterizable sparseness, computational efficiency, and algorithmic stability. We then derived its generalization bounds for both cases when the kernel is positive semidefinite and indefinite.

Acknowledgments

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 paper reflects only the authors' views, the Union is not liable for any use that may be made of the contained information. Research Council KUL: GOA/10/09 MaNet, CoE PFV/10/002 (OPTEC), BIL12/11T; PhD/Postdoc grants. Flemish Government: FWO: projects: G.0377.12 (Structured systems), G.088114N (Tensor based data similarity); PhD/Postdoc grants. IWT: projects: SBO POM (100031); PhD/Postdoc grants. iMinds Medical Information Technologies SBO 2014. Belgian Federal Science Policy Office: IUAP P7/19 (DYSCO, Dynamical systems, control and optimization, 2012-2017).

Appendix A. Definitions and Concentration Inequalities

Definition 8 (ℓ^2-Empirical Covering Number) *Let $\mathbf{x} = \{x_1, x_2, \ldots, x_n\} \subset \mathcal{X}^n$. The ℓ^2-empirical covering number of the hypothesis space $\mathcal{H}$, which is denoted as $\mathcal{N}_2(\mathcal{H}, \eta)$ with radius $\eta > 0$, is defined by*

$$\mathcal{N}_2(\mathcal{H}, \eta) := \sup_{n \in \mathbb{N}} \sup_{\mathbf{x} \in \mathcal{X}^n} \inf \Big\{ \ell \in \mathbb{N} : \exists \{f_i\}_{i=1}^{\ell} \subset \mathcal{H} \text{ such that for each } f \in \mathcal{H}, \text{there is some } i \in \{1, 2, \ldots, \ell\} \text{ with } \frac{1}{n} \sum_{j=1}^{n} |f(x_j) - f_i(x_j)|^2 \leq \eta^2 \Big\}.$$

Lemma 9 (One-sided Bernstein's Inequality, see Bernstein (1946)) *Let ξ be a random variable on a probability space Z with variance σ^2 satisfying $|\xi - \mathbb{E}\xi| \leq M_\xi$ for some constant M_ξ. Then for any $0 < \delta < 1$, we have*

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

Lemma 10 (Wu et al. (2007)) *Let $\mathcal{F}$ be a class of measurable functions on Z. Assume that there are constants $B, c > 0$ and $\theta \in [0, 1]$ such that $\|f\|_\infty \leq B$ and $\mathbb{E}f^2 \leq c(\mathbb{E}f)^\theta$ for*

every $f \in \mathcal{F}$. *If for some* $a > 0$ *and* $p \in (0, 2)$,

$$\log \mathcal{N}_2(\mathcal{F}, \epsilon) \leq a\epsilon^{-p}, \qquad \text{for all } \epsilon > 0,$$

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

$$\mathbb{E}f - \frac{1}{m}\sum_{i=1}^{m} f(z_i) \leq \frac{1}{2}\zeta^{1-\theta}(\mathbb{E}f)^{\theta} + c_p\zeta + 2\left(\frac{ct}{m}\right)^{\frac{1}{2-\theta}} + \frac{18Bt}{m}, \qquad \text{for all } f \in \mathcal{F},$$

where

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

Appendix B. Error Analysis with Positive Semidefinite Kernels

We now derive the probabilistic generalization bounds stated in Theorem 6 by following steps stated below and employing the stepping-stone technique introduced in Wu and Zhou (2005). To this end, we first introduce the following intermediate regularization scheme

$$\widetilde{f}_{\mathbf{z},\lambda} = \arg\min_{f\in\mathcal{H}_\mathcal{K}} \frac{1}{m}\sum_{k=1}^{m}(y_k - f(x_k))^2 + \lambda\|f\|_\mathcal{K}^2, \tag{14}$$

and its data-free counterpart

$$\widetilde{f}_\lambda = \arg\min_{f\in\mathcal{H}_\mathcal{K}} \int_{\mathcal{X}\times\mathcal{Y}}(y - f(x))^2 d\rho + \lambda\|f\|_\mathcal{K}^2, \tag{15}$$

where $\lambda > 0$ is another regularization parameter, which can be chosen differently from λ_1 and λ_2.

Step 1: Decomposing the Total Error

A sketch of the error analysis for KENReg is summarized in the following proposition. It splits the generalization error into several parts, where the method of decomposition follows from Wu and Zhou (2008).

Proposition 11 *For any* $\mathbf{z} \in (\mathcal{X}\times\mathcal{Y})^m$, *the following error decomposition holds*

$$\|\bar{f}_\mathbf{z} - f_\rho\|_{\mathcal{L}^2_{\rho_\mathcal{X}}}^2 \leq \mathbf{T}_1 + \mathbf{T}_2 + \mathbf{T}_3 + \mathbf{T}_4$$

where

$$\begin{aligned}
\mathbf{T}_1 &:= \mathcal{E}(\bar{f}_\mathbf{z}) - \mathcal{E}_\mathbf{z}(\bar{f}_\mathbf{z}),\\
\mathbf{T}_2 &:= \left\{\mathcal{E}_\mathbf{z}(\bar{f}_\mathbf{z}) + \Omega(f_\mathbf{z})\right\} - \left\{\mathcal{E}_\mathbf{z}(\widetilde{f}_{\mathbf{z},\lambda}) + \lambda\|\widetilde{f}_{\mathbf{z},\lambda}\|_\mathcal{K}^2\right\},\\
\mathbf{T}_3 &:= \mathcal{E}_\mathbf{z}(\widetilde{f}_\lambda) - \mathcal{E}(\widetilde{f}_\lambda),\\
\mathbf{T}_4 &:= \mathcal{E}(\widetilde{f}_\lambda) - \mathcal{E}(f_\rho) + \lambda\|\widetilde{f}_\lambda\|_\mathcal{K}^2.
\end{aligned}$$

Proof For any $\mathbf{z} \in (\mathcal{X} \times \mathcal{Y})^m$, it is obvious that

$$\|\bar{f}_{\mathbf{z}} - f_\rho\|^2_{\mathcal{L}^2_{\rho_\mathcal{X}}} = \mathcal{E}(\bar{f}_{\mathbf{z}}) - \mathcal{E}(f_\rho) \leq \mathcal{E}(\bar{f}_{\mathbf{z}}) - \mathcal{E}(f_\rho) + \Omega(f_{\mathbf{z}}).$$

By introducing intermediate terms $\mathcal{E}_{\mathbf{z}}(\bar{f}_{\mathbf{z}})$, $\mathcal{E}_{\mathbf{z}}(\widetilde{f}_{\mathbf{z},\lambda})$, $\mathcal{E}(\widetilde{f}_\lambda)$, $\mathcal{E}_{\mathbf{z}}(\widetilde{f}_\lambda)$, $\lambda\|\widetilde{f}_{\mathbf{z},\lambda}\|^2_\mathcal{K}$, and $\lambda\|\widetilde{f}_\lambda\|^2_\mathcal{K}$, we come to the following inequality

$$\mathcal{E}(\bar{f}_{\mathbf{z}}) - \mathcal{E}(f_\rho) + \Omega(f_{\mathbf{z}}) \leq \mathbf{T}_1 + \mathbf{T}_2 + \mathbf{T}_3 + \mathbf{T}_4 + \mathbf{T}_5,$$

where

$$\mathbf{T}_5 := \mathcal{E}_{\mathbf{z}}(\widetilde{f}_{\mathbf{z},\lambda}) + \lambda\|\widetilde{f}_{\mathbf{z},\lambda}\|^2_\mathcal{K} - \mathcal{E}_{\mathbf{z}}(\widetilde{f}_\lambda) - \lambda\|\widetilde{f}_\lambda\|^2_\mathcal{K}.$$

Recalling that $\widetilde{f}_{\mathbf{z},\lambda}$ is the minimizer of the risk functional (14), we see that $\mathbf{T}_5 \leq 0$. Therefore, the desired estimate follows. ■

$\mathbf{T}_1$ and $\mathbf{T}_3$ are sample error terms caused by random sampling. The definition of $\widetilde{f}_\lambda$ implies that

$$\mathbf{T}_4 = \mathcal{E}(\widetilde{f}_\lambda) - \mathcal{E}(f_\rho) + \lambda\|\widetilde{f}_\lambda\|^2_\mathcal{K} = \inf_{f \in \mathcal{H}_\mathcal{K}} \left\{ \|f - f_\rho\|^2_{\mathcal{L}^2_{\rho_\mathcal{X}}} + \lambda\|f\|^2_\mathcal{K} \right\}.$$

As aforementioned, $\mathbf{T}_4$ is usually called the model error or the approximation error. In view of the Model Assumption, $\mathbf{T}_4$ is upper bounded by $\mathcal{O}(\lambda^\beta)$ with $0 < \beta \leq 1$. The deviation error term $\mathbf{T}_2$ measures the difference of the empirical risks incurred by using the two different empirical target functions, $f_{\mathbf{z}}$ and $\widetilde{f}_{\mathbf{z},\lambda}$. In the literature of learning with the kernelized dictionary induced models, the deviation error term $\mathbf{T}_2$ is also referred as the hypothesis error (Wu and Zhou, 2008; Xiao and Zhou, 2010; Tong et al., 2010; Shi et al., 2011; Song and Zhang, 2011).

As a consequence of the above discussions, to obtain the generalization bounds it remains to bound the sample error terms $\mathbf{T}_1$ and $\mathbf{T}_3$, and the deviation error term $\mathbf{T}_2$.

Step 2: Bounding the Deviation Error Term $\mathbf{T}_2$

Inspired by techniques employed in Wu and Zhou (2008); Tong et al. (2010), we can bound the deviation error term $\mathbf{T}_2$ as follows:

Proposition 12 *Suppose that Assumption 5 holds. For all $\mathbf{z} \in (\mathcal{X} \times \mathcal{Y})^m$, the deviation error term $\mathbf{T}_2$ can be bounded by*

$$\mathbf{T}_2 \leq \frac{\lambda_1}{\lambda} + \frac{\lambda_2}{m\lambda^2}.$$

Proof Applying the representer theorem to (14), we see that $\widetilde{f}_{\mathbf{z},\lambda} = \sum_{i=1}^m \alpha_{\lambda,\mathbf{z},i}\mathcal{K}_{x_i}$. Denote $\alpha_{\lambda,\mathbf{z}} = (\alpha_{\lambda,\mathbf{z},1}, \ldots, \alpha_{\lambda,\mathbf{z},i})$. By setting the derivative of the risk functional (14) with respect to $\alpha_{\lambda,\mathbf{z}}$ to be zero, we have

$$\alpha_{\lambda,\mathbf{z},i} = \frac{1}{\lambda m}(y_i - \widetilde{f}_{\mathbf{z},\lambda}(x_i)), \qquad i = 1, ..., m.$$

Applying Jensen's inequality, we get

$$\lambda_1 \sum_{i=1}^{m} |\alpha_{\lambda,\mathbf{z},i}| = \frac{\lambda_1}{m\lambda} \sum_{i=1}^{m} |y_i - \widetilde{f}_{\mathbf{z},\lambda}(x_i)| \leq \frac{\lambda_1}{\lambda} \left(\mathcal{E}_{\mathbf{z}}(\widetilde{f}_{\mathbf{z},\lambda})\right)^{\frac{1}{2}},$$

and similarly we have

$$\lambda_2 \sum_{i=1}^{m} |\alpha_{\lambda,\mathbf{z},i}|^2 = \frac{\lambda_2}{m^2\lambda^2} \sum_{i=1}^{m} (y_i - \widetilde{f}_{\mathbf{z},\lambda}(x_i))^2 = \frac{\lambda_2}{m\lambda^2} \mathcal{E}_{\mathbf{z}}(\widetilde{f}_{\mathbf{z},\lambda}).$$

The above two inequalities in connection with the definition of $f_{\mathbf{z}}$ tell us that

$$\begin{aligned}
\mathcal{E}_{\mathbf{z}}(\bar{f}_{\mathbf{z}}) + \Omega(f_{\mathbf{z}}) &\leq \mathcal{E}_{\mathbf{z}}(f_{\mathbf{z}}) + \lambda_1 \sum_{i=1}^{m} |\alpha_{\mathbf{z},i}| + \lambda_2 \sum_{i=1}^{m} |\alpha_{\mathbf{z},i}|^2 \\
&\leq \mathcal{E}_{\mathbf{z}}(\widetilde{f}_{\mathbf{z},\lambda}) + \lambda_1 \sum_{i=1}^{m} |\alpha_{\lambda,\mathbf{z},i}| + \lambda_2 \sum_{i=1}^{m} |\alpha_{\lambda,\mathbf{z},i}|^2 \\
&\leq \mathcal{E}_{\mathbf{z}}(\widetilde{f}_{\mathbf{z},\lambda}) + \frac{\lambda_1}{\lambda} \left(\mathcal{E}_{\mathbf{z}}(\widetilde{f}_{\mathbf{z},\lambda})\right)^{\frac{1}{2}} + \frac{\lambda_2}{m\lambda^2} \mathcal{E}_{\mathbf{z}}(\widetilde{f}_{\mathbf{z},\lambda}).
\end{aligned}$$

The fact that $\widetilde{f}_{\mathbf{z},\lambda}$ is the minimizer of the risk functional in (14) implies

$$\mathcal{E}_{\mathbf{z}}(\widetilde{f}_{\mathbf{z},\lambda}) \leq \mathcal{E}_{\mathbf{z}}(\widetilde{f}_{\mathbf{z},\lambda}) + \lambda \|\widetilde{f}_{\mathbf{z},\lambda}\|_{\mathcal{K}}^2 \leq \mathcal{E}_{\mathbf{z}}(0) \leq 1.$$

As a result, we see that

$$\mathcal{E}_{\mathbf{z}}(\bar{f}_{\mathbf{z}}) + \Omega(f_{\mathbf{z}}) \leq \mathcal{E}_{\mathbf{z}}(\widetilde{f}_{\mathbf{z},\lambda}) + \lambda \|\widetilde{f}_{\mathbf{z},\lambda}\|_{\mathcal{K}}^2 + \frac{\lambda_1}{\lambda} + \frac{\lambda_2}{m\lambda^2}.$$

This gives the desired estimate. ■

Step 3: Bounding the Sample Error Terms $\mathbf{T}_1$ and $\mathbf{T}_3$

In this part, we bound the two sample error terms $\mathbf{T}_1$ and $\mathbf{T}_3$. For convenience of analysis, without affecting the summation of $\mathbf{T}_1$ and $\mathbf{T}_3$, we slightly modify the two error terms $\mathbf{T}_1$ and $\mathbf{T}_3$ by introducing an additional term as follows

$$\mathbf{T}_1 := \left\{\mathcal{E}(\bar{f}_{\mathbf{z}}) - \mathcal{E}_{\mathbf{z}}(\bar{f}_{\mathbf{z}})\right\} - \left\{\mathcal{E}(f_\rho) - \mathcal{E}_{\mathbf{z}}(f_\rho)\right\},$$

and

$$\mathbf{T}_3 := \mathcal{E}_{\mathbf{z}}(\widetilde{f}_\lambda) - \mathcal{E}(\widetilde{f}_\lambda) + \mathcal{E}(f_\rho) - \mathcal{E}_{\mathbf{z}}(f_\rho).$$

The following estimates are obtained by applying concentration inequalities in Appendix A.

Proposition 13 *Suppose that Assumption 5 holds. For any $0 < \delta < 1$, with confidence $1 - \delta/2$, there holds*

$$\mathbf{T}_1 \lesssim \frac{1}{2}\left\|\bar{f}_{\mathbf{z}} - f_\rho\right\|^2_{L^2_{\rho_{\mathcal{X}}}} + \left(\log\frac{2}{\delta}\right)\left(\frac{1}{m} + \left(\lambda_1^{-1}\bigwedge\sqrt{m\lambda_2^{-1}}\right)^{2p/(2+p)}\left(\frac{1}{m}\right)^{2/(2+p)}\right).$$

Proposition 14 *Suppose that Assumption 5 holds. For any $0 < \delta < 1$, with confidence $1 - \delta/2$, there holds*

$$\mathbf{T}_3 \lesssim \log(2/\delta)\left(\lambda^\beta + m^{-1} + m^{-1}\lambda^{\beta-1}\right).$$

Step 4: Deriving Generalization Bounds

This part provides proofs of Propositions 13, 14, and Theorem 6. We first give an upper bound of the $\|\cdot\|_{\mathcal{K}}$ norm of $f_{\mathbf{z}}$.

Lemma 15 *Suppose that Assumption 5 holds. For any $\mathbf{z} \in (\mathcal{X}\times\mathcal{Y})^m$, there holds*

$$\|f_{\mathbf{z}}\|_{\mathcal{K}} \le \lambda_1^{-1}\bigwedge\sqrt{m\lambda_2^{-1}}.$$

Proof According to the definition of $\boldsymbol{\alpha}_{\mathbf{z}}$ in (3), 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, the Hölder's inequality tells us that

$$\|f_{\mathbf{z}}\|_{\mathcal{K}} = \left\|\sum_{k=1}^m \alpha_k\mathcal{K}_{x_k}\right\|_{\mathcal{K}} \le \|\boldsymbol{\alpha}_{\mathbf{z}}\|_1 \le \sqrt{m}\|\boldsymbol{\alpha}_{\mathbf{z}}\|_2.$$

Consequently, we have

$$\|f_{\mathbf{z}}\|_{\mathcal{K}} \le \lambda_1^{-1}\bigwedge\sqrt{m\lambda_2^{-1}}.$$

This completes the proof of Lemma 15. ■

Proof [Proof of Proposition 13] We now present a concentration estimation with respect to the random variable $\xi_2 := (\bar{f}_{\mathbf{z}}(x) - y)^2 - (f_\rho(x) - y)^2$ to bound the sample error term $\mathbf{T}_1$. Obviously, ξ_2 varies among a set of functions in accordance with the varying sample $\mathbf{z}$. In this case, we turn to investigating the function set

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

To apply Lemma 10 to the function set B_R, it suffices to verify conditions in Lemma 10.

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

$$\|g\|_\infty = \sup_{(x,y)\in\mathcal{X}\times\mathcal{Y}}\left\{\left|\bar{f}(x) - f_\rho(x)\right|\left|\bar{f}(x) + f_\rho(x) - 2y\right|\right\} \le 8.$$

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

$$\begin{aligned}\mathbb{E}g^2 &= \mathbb{E}\{(\bar{f}(x) - f_\rho(x))^2(\bar{f}(x) + f_\rho(x) - 2y)^2\} \\ &\leq 16\mathbb{E}(\bar{f}(x) - f_\rho(x))^2 = 16\mathbb{E}g,\end{aligned}$$

where the above inequality again follows from the definition of the projection operator and the boundedness of Y. We now evaluate the capacity of $\mathcal{F}_R$ based on the Complexity Assumption. Observing that for any $g_1, g_2 \in \mathcal{F}_R$, with

$$g_1(z) = (\bar{f}_1(x) - y)^2 - (f_\rho(x) - y)^2 \text{ and } g_2(z) = (\bar{f}_2(x) - y)^2 - (f_\rho(x) - y)^2,$$

the supremum norm of the difference between g_1 and g_2 can be bounded as follows

$$\|g_1 - g_2\|_\infty = \sup_{(x,y)\in\mathcal{X}\times\mathcal{Y}} |(\bar{f}_1(x) - y)^2 - (\bar{f}_2(x) - y)^2| \leq 4\|f_1 - f_2\|_\infty.$$

Consequently, the following relation holds

$$\mathcal{N}_2(F_R, \epsilon) \leq \mathcal{N}_2\left(B_R, \epsilon/4\right) \leq \mathcal{N}_2\left(B_1, \epsilon/(4R)\right).$$

This in connection with the Complexity Assumption tells us that $\log \mathcal{N}_2(F_R, \epsilon) \lesssim R^p \epsilon^{-p}$.

Now applying Lemma 10 and substituting the corresponding coefficients and indices, we see that for any $g \in \mathcal{F}_R$, $0 < \delta < 1$, with confidence $1 - \delta/2$, there holds

$$\mathbb{E}g - \frac{1}{m}\sum_{i=1}^m g(z_i) \lesssim \frac{1}{2}\mathbb{E}g + \frac{C_{\mathcal{K},p}\log(2/\delta)}{m} + \frac{C_{\mathcal{K},p}R^{2p/(2+p)}}{m^{2/(2+p)}}, \tag{16}$$

where $C_{\mathcal{K},p}$ is a positive constant independent of m or δ. To complete the proof of Proposition 13, we consider

$$g(z) = (\bar{f}_{\mathbf{z}}(x) - y)^2 - (f_\rho(x) - y)^2, \ z = (x, y) \in \mathcal{X} \times \mathcal{Y}.$$

According to Lemma 15, for any $\mathbf{z} \in (\mathcal{X} \times \mathcal{Y})^m$, it is clear that $g \in \mathcal{F}_R$ with $R = \lambda_1^{-1} \bigwedge \sqrt{m\lambda_2^{-1}}$. Substituting R with $R = \lambda_1^{-1} \bigwedge \sqrt{m\lambda_2^{-1}}$ into inequality (16), we come to the conclusion that for any $0 < \delta < 1$, with confidence $1 - \delta/2$, there holds

$$\mathbf{T}_1 \lesssim \frac{1}{2}\left\|\bar{f}_{\mathbf{z}} - f_\rho\right\|^2_{\mathcal{L}^2_{\rho_\mathcal{X}}} + \left(\log\frac{2}{\delta}\right)\left(\frac{1}{m} + \left(\lambda_1^{-1} \bigwedge \sqrt{m\lambda_2^{-1}}\right)^{2p/(2+p)}\left(\frac{1}{m}\right)^{2/(2+p)}\right).$$

This completes the proof. ■

Proof [Proof of Proposition 14] To prove Proposition 14, we apply the one-sided Bernstein's inequality in Lemma 9 in Appendix A. According to the definition of $\widetilde{f}_\lambda$ in (15) and the Model Assumption, we know that

$$\lambda\|\widetilde{f}_\lambda\|^2_\mathcal{K} \leq \inf_{f\in\mathcal{H}_\mathcal{K}}\left\{\|f - f_\rho\|^2_{\mathcal{L}^2_{\rho_\mathcal{X}}} + \lambda\|f\|^2_\mathcal{K}\right\} \lesssim \lambda^\beta.$$

The above inequality immediately yields an upper bound of $\|\widetilde{f}_\lambda\|_\infty$, that is, $\left\|\widetilde{f}_\lambda\right\|_\infty \lesssim \sqrt{\lambda^{\beta-1}}$.

To apply the one-sided Bernstein's inequality, we introduce the random variable ξ_1 as follows

$$\xi_1(z) := (\widetilde{f}_\lambda(x) - y)^2 - (f_\rho(x) - y)^2, \ z \in \mathcal{X} \times \mathcal{Y}.$$

From the upper bounds of $\widetilde{f}_\lambda$ and the boundedness of Y, we have

$$|\xi_1 - \mathbb{E}\xi_1| \leq 2|\xi_1| \leq \left(\|\widetilde{f}_\lambda\|_\infty + 3\right)^2.$$

The variance of ξ_1, which is denoted as $\mathrm{var}(\xi_1)$, can be bounded as

$$\mathrm{var}(\xi_1) \leq \mathbb{E}(\xi_1^2) \leq \left(\|\widetilde{f}_\lambda\|_\infty + 3\right)^2 \|\widetilde{f}_\lambda - f_\rho\|^2_{\mathcal{L}^2_{\rho_\mathcal{X}}}.$$

Applying Lemma 9, with simple computations, we come to the conclusion that for any $0 < \delta < 1$, with confidence $1 - \delta/2$, there holds

$$\mathbf{T}_3 \lesssim \log(2/\delta)\left(\lambda^\beta + m^{-1}\lambda^{\beta-1}\right).$$

This completes the proof of Proposition 14. ■

Proof [Proof of Theorem 6] As a consequence of Proposition 11, to bound $\|\bar{f}_\mathbf{z} - f_\rho\|^2_{\mathcal{L}^2_{\rho_\mathcal{X}}}$, it remains to combine the upper bounds of $\mathbf{T}_1$, $\mathbf{T}_2$, $\mathbf{T}_3$, and $\mathbf{T}_4$. As a result, we come to the conclusion that for any $0 < \delta < 1$, with confidence $1 - \delta$, there holds

$$\left\|\bar{f}_\mathbf{z} - f_\rho\right\|^2_{\mathcal{L}^2_{\rho_\mathcal{X}}} \lesssim \log\left(\frac{2}{\delta}\right)\left(\frac{\lambda^{\beta+1} + \lambda_2 + m\lambda_1\lambda + m\lambda^{\beta+2}}{m\lambda^2} + \left(\frac{1}{\lambda_1} \bigwedge \sqrt{\frac{m}{\lambda_2}}\right)^{\frac{2p}{2+p}} \left(\frac{1}{m}\right)^{\frac{2}{2+p}}\right).$$

This completes the proof of Theorem 6.

■

Proof [Proof of Corollary 7] We prove Corollary 7 by discussing the two different cases: $2\theta_1 \leq 1 + \theta_2$ and $2\theta_1 \geq 1 + \theta_2$, respectively.

We first consider the case when $2\theta_1 \leq 1 + \theta_2$. As a consequence of Theorem 6 and recalling that $\lambda_1 = m^{-\theta_1}, \lambda_2 = m^{-\theta_2}$, for any $0 < \delta < 1$, with confidence $1 - \delta$ we have

$$\left\|\bar{f}_\mathbf{z} - f_\rho\right\|^2_{\mathcal{L}^2_{\rho_\mathcal{X}}} \lesssim \left(\log\frac{2}{\delta}\right)\left(\lambda^\beta + \frac{\lambda^{\beta-1}}{m} + \frac{\lambda_1}{\lambda} + \frac{\lambda_2}{m\lambda^2} + \left(\frac{1}{m\lambda_1^p}\right)^{2/(2+p)}\right).$$

In order to obtain nontrivial learning rates, we need to choose proper regularization parameters λ_1, λ_2, and λ, the steps of which are given as follows. First, we choose the regularization parameter λ by letting

$$\frac{\lambda_1}{\lambda} = \left(\frac{1}{\lambda_1}\right)^{\frac{2p}{2+p}} \left(\frac{1}{m}\right)^{\frac{2}{2+p}},$$

which yields, $\lambda = m^{-\frac{2+3p}{2+p}\theta_1+\frac{2}{2+p}}$. Keeping this choice of λ in mind, the index θ_2 can be chosen by forcing $\frac{\lambda_2}{m\lambda^2} \leq \lambda^\beta$ while θ_1 can be decided by letting

$$\frac{\lambda^{\beta-1}}{m} \leq \lambda^\beta, \text{ and } \frac{\lambda_1}{\lambda} \leq \lambda^\beta.$$

A simple computation shows that

$$\begin{cases} \theta_1 \in \left[\frac{1}{2} \bigwedge \frac{6+p+2\beta}{(\beta+2)(2+3p)}, \frac{4+p}{2+3p} \bigwedge \frac{2(\beta+1)}{(2+3p)(\beta+1)-(2+p)}\right], \\ \theta_2 \in \left[(2\theta_1-1) \bigvee \frac{((2+3p)\theta_1-2)(\beta+2)-2-p}{2+p}, +\infty\right], \end{cases}$$

where the lower bound of θ_1 is enforced to ensure the positiveness of the lower bound of θ_2. As a result, we see that for any $0 < \delta < 1$ with confidence $1-\delta$ there holds

$$\left\|\bar{f}_{\mathbf{z}} - f_\rho\right\|^2_{\mathcal{L}^2_{\rho_\mathcal{X}}} \lesssim \left(\log\frac{2}{\delta}\right)\left(\frac{1}{m}\right)^{\frac{\beta(\theta_1(2+3p)-2)}{(2+p)}}.$$

This completes the proof for the first case.

We now proceed with proving the second case when $2\theta_1 \geq 1+\theta_2$. By choosing $\lambda_1 = m^{-\theta_1}, \lambda_2 = m^{-\theta_2}$, Theorem 6 tells us again that, for $0 < \delta < 1$, with confidence $1-\delta$ there holds

$$\left\|\bar{f}_{\mathbf{z}} - f_\rho\right\|^2_{\mathcal{L}^2_{\rho_\mathcal{X}}} \lesssim \left(\log\frac{2}{\delta}\right)\left(\lambda^\beta + \frac{\lambda^{\beta-1}}{m} + \frac{\lambda_1}{\lambda} + \frac{\lambda_2}{m\lambda^2} + \left(\frac{1}{m^{2-p}\lambda_2^p}\right)^{1/(2+p)}\right).$$

Applying similar techniques of choosing regularization parameters λ, λ_1 and λ_2, and by setting

$$\frac{\lambda_2}{m\lambda^2} = \left(\frac{m}{\lambda_2}\right)^{\frac{p}{2+p}}\left(\frac{1}{m}\right)^{\frac{2}{2+p}},$$

we know that $\lambda = m^{-\frac{1+p}{2+p}\theta_2-\frac{p}{2+p}}$. Let $p \in (0, 2/(1+\beta))$. Then θ_1 is decided by forcing $\frac{\lambda_1}{\lambda} \leq \lambda^\beta$ and we choose θ_2 by ensuring

$$\frac{\lambda^{\beta-1}}{m} \leq \lambda^\beta, \text{ and } \frac{\lambda_2}{m\lambda^2} \leq \lambda^\beta.$$

Analogously, a simple computation shows that

$$\begin{cases} \theta_1 \in \left[\frac{1+\theta_2}{2} \bigvee \frac{(\beta+1)((1+p)\theta_2+p)}{2+p}, +\infty\right], \\ \theta_2 \in \left(0, \frac{2}{1+p} \bigwedge \frac{2+p\beta-p}{(1+p)(\beta+2)-2-p}\right]. \end{cases}$$

As a consequence, for any $0 < \delta < 1$, with confidence $1-\delta$ there holds

$$\left\|\bar{f}_{\mathbf{z}} - f_\rho\right\|^2_{\mathcal{L}^2_{\rho_\mathcal{X}}} \lesssim \left(\log\frac{2}{\delta}\right)\left(\frac{1}{m}\right)^{\frac{\beta(\theta_2(1+p)+p)}{(2+p)}}.$$

Introducing the notation Θ and combining the above estimates, we complete the proof. ■

Appendix C. Error Analysis with Indefinite Kernels

Error analysis for KENReg with indefinite kernels can be proceeded by the following steps:

Step 1: Constructing a Proxy Positive Semidefinite Kernel

We first construct a proxy positive semidefinite kernel $\widetilde{\mathcal{K}}$ by using the indefinite kernel $\mathcal{K}$ according to the following fact, which was noticed in Xiao and Zhou (2010).

Fact 1 *Let $\mathcal{K}: \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a continuous kernel. Define $\widetilde{\mathcal{K}}: \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ as*

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

Then, $\widetilde{\mathcal{K}}$ is a Mercer kernel.

Step 2: Constructing an Instrumental Empirical Target Function

The second step is to construct an instrumental empirical target function, which will play a stepping-stone role. In our study, the construction of such an instrumental function follows from that in Guo and Shi (2013). Similar to the positive semidefinite case, we first introduce an intermediate regularization scheme

$$\widetilde{f}_{\lambda} = \arg \min_{f \in \mathcal{H}_{\widetilde{\mathcal{K}}}} \left\{ \mathcal{E}(f) - \mathcal{E}(f_{\rho}) + \lambda \|f\|^2_{\widetilde{\mathcal{K}}} \right\},$$

where $\lambda > 0$ is a regularization parameter and $\mathcal{H}_{\widetilde{\mathcal{K}}}$ is the RKHS induced by the Mercer kernel $\widetilde{\mathcal{K}}$. For any continuous kernel function $\mathcal{K}: \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, we define the integral operator $L_{\mathcal{K}}: \mathcal{L}^2_{\rho_{\mathcal{X}}} \to \mathcal{L}^2_{\rho_{\mathcal{X}}}$ as follows:

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

$L_{\mathcal{K}}$ and its adjoint $L^{\star}_{\mathcal{K}}$ are compact operators due to the compactness of $\mathcal{X}$ and the continuity of $\mathcal{K}$. Moreover, one can easily see that $L_{\widetilde{\mathcal{K}}} = L_{\mathcal{K}} L^{\star}_{\mathcal{K}}$. As shown in Cucker and Smale (2002), the existence and uniqueness of $\widetilde{f}_{\lambda}$ are ensured by the positiveness of λ, which can be explicitly expressed as

$$\widetilde{f}_{\lambda} = (\lambda I + L_{\widetilde{\mathcal{K}}})^{-1} L_{\widetilde{\mathcal{K}}} f_{\rho}.$$

We also define g_{λ} by

$$g_{\lambda} = L^{\star}_{\mathcal{K}} (\lambda I + L_{\widetilde{\mathcal{K}}})^{-1} f_{\rho}.$$

The following conclusion is due to Shi (2013) which is obtained by applying the polar decomposition to the compact operator $L_{\mathcal{K}}$. It will be useful in the construction of the instrumental empirical target function.

Proposition 16 *Let $\overline{\mathcal{H}_{\widetilde{\mathcal{K}}}}$ be the closure of $\mathcal{H}_{\widetilde{\mathcal{K}}}$ in $\mathcal{L}^2_{\rho_{\mathcal{X}}}$. Considering that $\mathcal{H}_{\widetilde{\mathcal{K}}}$ is a subspace of $\mathcal{L}^2_{\rho_{\mathcal{X}}}$, we have $L^{\star}_{\mathcal{K}} = U L^{\frac{1}{2}}_{\widetilde{\mathcal{K}}}$ and $L_{\mathcal{K}} = L^{\frac{1}{2}}_{\widetilde{\mathcal{K}}} U^{\star}$, where U is a partial isometry on $\mathcal{L}^2_{\rho_{\mathcal{X}}}$ with $U^{\star}U$ being the orthogonal projection onto $\overline{\mathcal{H}_{\widetilde{\mathcal{K}}}}$.*

We are now ready to construct the instrumental empirical target function $\widetilde{f}_{\mathbf{z},\lambda}$, which is defined as

$$\widetilde{f}_{\mathbf{z},\lambda}(x) = \frac{1}{m}\sum_{i=1}^{m} g_\lambda(x_i)\mathcal{K}(x_i, x), \quad x \in \mathcal{X}. \tag{17}$$

From Proposition 16, we know that

$$\widetilde{f}_\lambda = L_{\mathcal{K}} g_\lambda = \mathbb{E}\widetilde{f}_{\mathbf{z},\lambda}. \tag{18}$$

Step 3: Decomposing the Total Error

For any $f_{\boldsymbol{\alpha}}(x) = \sum_{l=1}^{m} \alpha_l \mathcal{K}(x_l, x)$, $x \in \mathcal{X}$ and $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_m)^\top$, due to the dependence of λ_2 on m, again we can slightly modify the model by denoting $\Omega(f_{\boldsymbol{\alpha}})$ as

$$\Omega(f_{\boldsymbol{\alpha}}) = \lambda_1 \|\boldsymbol{\alpha}\|_1 + \lambda_2 m \|\boldsymbol{\alpha}\|_2^2.$$

Based on the aforementioned notations, we now state the following novel error decomposition strategy:

Proposition 17 *Given Fact 1, for any $\mathbf{z} \in (\mathcal{X} \times \mathcal{Y})^m$, the following error decomposition holds*

$$\|\bar{f}_{\mathbf{z}} - f_\rho\|_{\mathcal{L}^2_{\rho_{\mathcal{X}}}}^2 \leq \mathbf{T}_1 + \mathbf{T}_2 + \mathbf{T}_3 + \mathbf{T}_4 + \mathbf{T}_5$$

where

$$\begin{aligned}
\mathbf{T}_1 &:= \mathcal{E}(\widetilde{f}_\lambda) - \mathcal{E}(f_\rho), \\
\mathbf{T}_2 &:= \mathcal{E}_{\mathbf{z}}(\widetilde{f}_{\mathbf{z},\lambda}) - \mathcal{E}(\widetilde{f}_{\mathbf{z},\lambda}), \\
\mathbf{T}_3 &:= \Omega(\widetilde{f}_{\mathbf{z},\lambda}), \\
\mathbf{T}_4 &:= \mathcal{E}(\widetilde{f}_{\mathbf{z},\lambda}) - \mathcal{E}(\widetilde{f}_\lambda), \\
\mathbf{T}_5 &:= \mathcal{E}(\bar{f}_{\mathbf{z}}) - \mathcal{E}_{\mathbf{z}}(\bar{f}_{\mathbf{z}}).
\end{aligned}$$

Proof For any $\mathbf{z} \in (\mathcal{X} \times \mathcal{Y})^m$, we can see that

$$\|\bar{f}_{\mathbf{z}} - f_\rho\|_{\mathcal{L}^2_{\rho_{\mathcal{X}}}}^2 = \mathcal{E}(\bar{f}_{\mathbf{z}}) - \mathcal{E}(f_\rho) \leq \mathcal{E}(\bar{f}_{\mathbf{z}}) - \mathcal{E}(f_\rho) + \Omega(f_{\mathbf{z}}).$$

By introducing intermediate terms $\mathcal{E}_{\mathbf{z}}(\bar{f}_{\mathbf{z}})$, $\mathcal{E}_{\mathbf{z}}(\widetilde{f}_{\mathbf{z},\lambda})$, $\mathcal{E}_{\mathbf{z}}(\widetilde{f}_\lambda)$, $\mathcal{E}(\widetilde{f}_\lambda)$, $\mathcal{E}(\widetilde{f}_{\mathbf{z},\lambda})$, and $\Omega(\widetilde{f}_{\mathbf{z},\lambda})$, with simple computations, we see that

$$\mathcal{E}(\bar{f}_{\mathbf{z}}) - \mathcal{E}(f_\rho) + \Omega(f_{\mathbf{z}}) \leq \mathbf{T}_1 + \mathbf{T}_2 + \mathbf{T}_3 + \mathbf{T}_4 + \mathbf{T}_5 + \mathbf{T}_6,$$

where the term

$$\mathbf{T}_6 := \mathcal{E}_{\mathbf{z}}(\bar{f}_{\mathbf{z}}) + \Omega(f_{\mathbf{z}}) - \mathcal{E}_{\mathbf{z}}(\widetilde{f}_{\mathbf{z},\lambda}) - \Omega(\widetilde{f}_{\mathbf{z},\lambda})$$

is at most zero according to the definition of $f_{\mathbf{z}}$. This proves Proposition 17. ■

Step 4: Bounding Error Terms

We now give some discussions on bounding error terms in Proposition 17, which are listed as follows:

- $\mathbf{T}_1$ represents the approximation ability of functions in $\mathcal{H}_{\widetilde{\mathcal{K}}}$ to f_ρ. As discussed previously, one may again apply the assumption on $\mathbf{T}_1$ that $\mathbf{T}_1$ is upper bounded by $\mathcal{O}(\lambda^\beta)$ with $0 < \beta \leq 1$;

- $\mathbf{T}_2$, $\mathbf{T}_4$, and $\mathbf{T}_5$ are sample error terms. $\mathbf{T}_4$ can be bounded directly by using the concentration inequality in Appendix A while bounding $\mathbf{T}_2$ and $\mathbf{T}_5$ needs to introduce concentration inequalities for function space valued random variables and consequently needs to measure the complexities of corresponding hypothesis spaces. The above three sample error terms can be bounded by using similar techniques employed in Shi (2013); Guo and Shi (2013);

- Following the definition of $\widetilde{f}_{\mathbf{z},\lambda}$ and g_λ, we have

$$\mathbf{T}_3 = \frac{\lambda_1}{m}\sum_{i=1}^m |g_\lambda(x_i)| + \frac{\lambda_2}{m}\sum_{i=1}^m |g_\lambda(x_i)|^2 \leq \lambda_1 \|g_\lambda\|_\infty + \lambda_2 \|g_\lambda\|_\infty^2.$$

  As shown in Guo and Shi (2013), $\|g_\lambda\|_\infty \lesssim \lambda^{(\beta-2)/2}$. As a result, we see that

$$\mathbf{T}_3 \lesssim \lambda_1 \lambda^{(\beta-2)/2} + \lambda_2 \lambda^{(\beta-2)}.$$

Following the above steps, generalization bounds can be derived by combining the above estimates and choosing proper regularization parameters λ_1, λ_2, and λ.

References

Shivani Agarwal and Partha Niyogi. Generalization bounds for ranking algorithms via algorithmic stability. *Journal of Machine Learning Research*, 10:441–474, 2009.

Maria-Florina Balcan, Avrim Blum, and Nathan Srebro. A theory of learning with similarity functions. *Machine Learning*, 72(1-2):89–112, 2008.

Peter L. Bartlett. The sample complexity of pattern classification with neural networks: the size of the weights is more important than the size of the network. *Information Theory, IEEE Transactions on*, 44(2):525–536, 1998.

Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.

Sergei N. Bernstein. *The Theory of Probabilities*. Gastehizdat Publishing House, Moscow, 1946.

Olivier Bousquet and André Elisseeff. Stability and generalization. *Journal of Machine Learning Research*, 2:499–526, 2002.

Leo Breiman. Heuristics of instability and stabilization in model selection. *The Annals of Statistics*, 24(6):2350–2383, 1996.

Di-Rong Chen, Qiang Wu, Yiming Ying, and Ding-Xuan Zhou. Support vector machine soft margin classifiers: error analysis. *Journal of Machine Learning Research*, 5:1143–1175, 2004.

Hong Chen, Zhibin Pan, Luoqing Li, and Yuanyan Tang. Error analysis of coefficient-based regularized algorithm for density-level detection. *Neural Computation*, 25(4):1107–1121, 2013.

Shaobing Chen and David Donoho. Basis pursuit. In *Signals, Systems and Computers, 1994. 1994 Conference Record of the Twenty-Eighth Asilomar Conference on*, volume 1, pages 41–44. IEEE, 1994.

Felipe Cucker and Steve Smale. Best choices for regularization parameters in learning theory: on the bias-variance problem. *Foundations of Computational Mathematics*, 2(4):413–428, 2002.

Felipe Cucker and Ding-Xuan Zhou. *Learning Theory: An Approximation Theory Viewpoint*. Cambridge University Press, 2007.

Kris De Brabanter, Jos De Brabanter, Johan A.K. Suykens, and Bart De Moor. Optimized fixed-size kernel models for large data sets. *Computational Statistics & Data Analysis*, 54(6):1484–1504, 2010.

Christine De Mol, Ernesto De Vito, and Lorenzo Rosasco. Elastic-net regularization in learning theory. *Journal of Complexity*, 25(2):201–230, 2009.

Luc P. Devroye and Terry J. Wagner. Distribution-free performance bounds for potential function rules. *Information Theory, IEEE Transactions on*, 25(5):601–604, 1979.

Yunlong Feng and Shao-Gao Lv. Unified approach to coefficient-based regularized regression. *Computers & Mathematics with Applications*, 62(1):506–515, 2011.

Yunlong Feng, Shao-Gao Lv, Hanyuan Hang, and Johan A.K. Suykens. Kernelized elastic net regularization: Generalization bounds, and sparse recovery. *Neural Computation*, 28(3):525–562, 2016.

Junbin Gao, Paul W. Kwan, and Daming Shi. Sparse kernel learning with Lasso and Bayesian inference algorithm. *Neural Networks*, 23(2):257–264, 2010.

Federico Girosi. An equivalence between sparse approximation and support vector machines. *Neural Computation*, 10(6):1455–1480, 1998.

Vincent Guigue, Alain Rakotomamonjy, and Stéphane Canu. Kernel basis pursuit. In *Machine Learning: ECML 2005*, pages 146–157. Springer, 2005.

Zheng-Chu Guo and Lei Shi. Learning with coefficient-based regularization and ℓ^1-penalty. *Advances in Computational Mathematics*, 39(3-4):493–510, 2013.

Bernard Haasdonk. Feature space interpretation of SVMs with indefinite kernels. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 27(4):482–492, 2005.

Luc Hoegaerts, Johan A.K. Suykens, Joos Vandewalle, and Bart De Moor. A comparison of pruning algorithms for sparse least squares support vector machines. In *Neural Information Processing*, pages 1247–1253. Springer, 2004.

Paul Honeine. Online kernel principal component analysis: a reduced-order model. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 34(9):1814–1826, 2012.

Bangti Jin, Dirk A Lorenz, and Stefan Schiffler. Elastic-net regularization: error estimates and active set methods. *Inverse Problems*, 25(11):115022, 2009.

Purushottam Kar and Prateek Jain. Supervised learning with similarity functions. In *Advances in Neural Information Processing Systems*, pages 215–223, 2012.

Peter Karsmakers, Kristiaan Pelckmans, Kris Brabanter, Hugo Hamme, and Johan A.K. Suykens. Sparse conjugate directions pursuit with application to fixed-size kernel models. *Machine Learning*, 85(1-2):109–148, 2011.

Michael Kearns and Dana Ron. Algorithmic stability and sanity-check bounds for leave-one-out cross-validation. *Neural Computation*, 11(6):1427–1453, 1999.

Samuel Kutin and Partha Niyogi. Almost-everywhere algorithmic stability and generalization error. In *Proceedings of the Eighteenth Conference on Uncertainty in Artificial Intelligence*, pages 275–282. Morgan Kaufmann Publishers Inc., 2002.

Shaobo Lin, Jinshan Zeng, Jian Fang, and Zongben Xu. Learning rates of ℓ^q-coefficient regularization learning with Gaussian kernel. *Neural Computation*, 26(10):2350–2378, 2014.

Ronny Luss and Alexandre d'Aspremont. Support vector machine classification with indefinite kernels. *Mathematical Programming Computation*, 1(2-3):97–118, 2009.

Andreas Maurer. Algorithmic stability and meta-learning. *Journal of Machine Learning Research*, 6:967–994, 2005.

Nicolai Meinshausen and Bin Yu. Lasso-type recovery of sparse representations for high-dimensional data. *The Annals of Statistics*, 37(1):246–270, 2009.

Shahar Mendelson and Joseph Neeman. Regularization in kernel learning. *The Annals of Statistics*, 38(1):526–565, 2010.

Sofia Mosci, Lorenzo Rosasco, Matteo Santoro, Alessandro Verri, and Silvia Villa. Solving structured sparsity regularization with proximal methods. In *Machine Learning and Knowledge Discovery in Databases*, pages 418–433. Springer, 2010.

Sayan Mukherjee, Partha Niyogi, Tomaso Poggio, and Ryan Rifkin. Learning theory: stability is sufficient for generalization and necessary and sufficient for consistency of empirical risk minimization. *Advances in Computational Mathematics*, 25(1-3):161–193, 2006.

Prasanth B. Nair, Arindam Choudhury, and Andy J. Keane. Some greedy learning algorithms for sparse regression and classification with mercer kernels. *Journal of Machine Learning Research*, 3:781–801, 2003.

Cheng Soon Ong, Xavier Mary, Stéphane Canu, and Alexander J. Smola. Learning with non-positive kernels. In *Proceedings of the Twenty-First International Conference on Machine Learning*, page 81. ACM, 2004.

Tomaso Poggio, Ryan Rifkin, Sayan Mukherjee, and Partha Niyogi. General conditions for predictivity in learning theory. *Nature*, 428(6981):419–422, 2004.

Cédric Richard, José Carlos M. Bermudez, and Paul Honeine. Online prediction of time series data with kernels. *Signal Processing, IEEE Transactions on*, 57(3):1058–1067, 2009.

William H. Rogers and Terry J. Wagner. A finite sample distribution-free performance bound for local discrimination rules. *The Annals of Statistics*, 6(3):506–514, 1978.

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

Mark Schmidt, Nicolas L. Roux, and Francis R. Bach. Convergence rates of inexact proximal-gradient methods for convex optimization. In *Advances in Neural Information Processing Systems*, pages 1458–1466, 2011.

Bernhard Schölkopf and Alexander J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT press, 2001.

Bernhard Schölkopf, Alex J. Smola, Robert C. Williamson, and Peter L. Bartlett. New support vector algorithms. *Neural Computation*, 12(5):1207–1245, 2000.

Shai Shalev-Shwartz, Ohad Shamir, Nathan Srebro, and Karthik Sridharan. Learnability, stability and uniform convergence. *Journal of Machine Learning Research*, 11:2635–2670, 2010.

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

Lei Shi, Yunlong Feng, and Ding-Xuan Zhou. Concentration estimates for learning with ℓ^1-regularizer and data dependent hypothesis spaces. *Applied and Computational Harmonic Analysis*, 31(2): 286–302, 2011.

Steve Smale and Ding-Xuan Zhou. Estimating the approximation error in learning theory. *Analysis and Applications*, 1(01):17–41, 2003.

Guohui Song and Haizhang Zhang. Reproducing kernel Banach spaces with the ℓ^1-norm II: error analysis for regularized least square regression. *Neural Computation*, 23(10):2713–2729, 2011.

Ingo Steinwart and Andreas Christmann. *Support Vector Machines*. Springer, New York, 2008.

Ingo Steinwart, Don Hush, and Clint Scovel. An oracle inequality for clipped regularized risk minimizers. In *Advances in Neural Information Processing Systems*, pages 1321–1328, 2006.

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

Johan A.K. Suykens, Lukas Lukas, and Joos Vandewalle. Sparse approximation using least squares support vector machines. In *Circuits and Systems, 2000. Proceedings. ISCAS 2000 Geneva. The 2000 IEEE International Symposium on*, volume 2, pages 757–760. IEEE, 2000.

Johan A.K. Suykens, Tony Van Gestel, Jos De Brabanter, Bart De Moor, and Joos Vandewalle. *Least Squares Support Vector Machines*. World Scientific, Singapore, 2002.

Robert Tibshirani. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996.

Ryan J. Tibshirani. The Lasso problem and uniqueness. *Electronic Journal of Statistics*, 7:1456–1490, 2013.

Hongzhi Tong, Di-Rong Chen, and Fenghong Yang. Least square regression with ℓ^p-coefficient regularization. *Neural Computation*, 22(12):3221–3235, 2010.

Hongzhi Tong, Di-Rong Chen, and Fenghong Yang. Learning with convex loss and indefinite kernels. *Neural Computation*, 26(1):158–184, 2014.

Sara A. van de Geer and Peter Bühlmann. On the conditions used to prove oracle results for the lasso. *Electronic Journal of Statistics*, 3:1360–1392, 2009.

Vladimir Vapnik. *The Nature of Statistical Learning Theory*. Springer, 1995.

Vladimir Vapnik. *Statistical Learning Theory*. Wiley-Interscience, 1998.

Pascal Vincent and Yoshua Bengio. Kernel matching pursuit. *Machine Learning*, 48(1-3):165–187, 2002.

Gang Wang, Dit-Yan Yeung, and Frederick H. Lochovsky. The kernel path in kernelized Lasso. In *International Conference on Artificial Intelligence and Statistics*, pages 580–587, 2007.

Qiang Wu. Regularization networks with indefinite kernels. *Journal of Approximation Theory*, 166: 1–18, 2013.

Qiang Wu and Ding-Xuan Zhou. SVM soft margin classifiers: linear programming versus quadratic programming. *Neural Computation*, 17(5):1160–1187, 2005.

Qiang Wu and Ding-Xuan Zhou. Learning with sample dependent hypothesis spaces. *Computers & Mathematics with Applications*, 56(11):2896–2907, 2008.

Qiang Wu, Yiming Ying, and Ding-Xuan Zhou. Multi-kernel regularized classifiers. *Journal of Complexity*, 23(1):108–134, 2007.

Quan-Wu Xiao and Ding-Xuan Zhou. Learning by nonsymmetric kernels with data dependent spaces and ℓ^1-regularizer. *Taiwanese Journal of Mathematics*, 14(5):1821–1836, 2010.

Huan Xu, Constantine Caramanis, and Shie Mannor. Sparse algorithms are not stable: a no-free-lunch theorem. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 34(1):187–193, 2012.

Yiming Ying, Colin Campbell, and Mark Girolami. Analysis of SVM with indefinite kernels. In *Advances in Neural Information Processing Systems*, pages 2205–2213, 2009.

Bin Yu. Stability. *Bernoulli*, 19(4):1484–1500, 2013.

Chong Zhang, Yufeng Liu, and Yichao Wu. On quantile regression in reproducing kernel Hilbert spaces with data sparsity constraint. *Journal of Machine Learning Research*, 17:1–45, 2016.

Li Zhang and Weida Zhou. On the sparseness of 1-norm support vector machines. *Neural Networks*, 23(3):373–385, 2010.

Li Zhang and Weida Zhou. 1-norm support vector novelty detection and its sparseness. *Neural Networks*, 48:125–132, 2013.

Peng Zhao and Bin Yu. On model selection consistency of Lasso. *Journal of Machine Learning Research*, 7:2541–2563, 2006.

Hui Zou. The adaptive Lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429, 2006.

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