Sparse Shrunk Additive Models

Guodong Liu [*1] **Hong Chen** [*2] **Heng Huang** [1 3]

Abstract

Most existing feature selection methods in literature are linear models, so that the nonlinear relations between features and response variables are not considered. Meanwhile, in these feature selection models, the interactions between features are often ignored or just discussed under prior structure information. To address these challenging issues, we consider the problem of sparse additive models for high-dimensional nonparametric regression with the allowance of the flexible interactions between features. A new method, called as *sparse shrunk additive models* (SSAM), is proposed to explore the structure information among features. This method bridges sparse kernel regression and sparse feature selection. Theoretical results on the convergence rate and sparsity characteristics of SSAM are established by the novel analysis techniques with integral operator and concentration estimate. In particular, our algorithm and theoretical analysis only require the component functions to be continuous and bounded, which are not necessary to be in reproducing kernel Hilbert spaces. Experiments on both synthetic and real-world data demonstrate the effectiveness of the proposed approach.

1. Introduction

Sparse feature selection has attracted much attention in machine learning community for learning tasks with high-dimensional data, especially useful in bioinformatics related applications. Linear models with ℓ_1-norm regularization, such as Lasso (Tibshirani, 1996) and Dantzig selector(Candes & Tao, 2007), have been well studied for their theoretical properties and extensively used for feature selection applications. However, in many applications, the linear assumption could be too restricted to select the optimal features, because the relations between features and response variables could be nonlinear. Because of the difficulties in both computational algorithm and learning theory analysis, only few of existing feature selection methods in literature focus on the nonlinear feature selection.

To enhance the ability of feature selection models with considering nonlinear relationship between features and response variables, several sparse learning based additive models were proposed for regression (Ravikumar et al., 2009; Huang et al., 2010; Raskutti et al., 2012; Yuan & Zhou, 2016; Yin et al., 2012; Chen et al., 2020, In press) and classification (Zhao & Liu, 2012; Chen et al., 2017), which are extensions of original additive models (Hastie & Tibshirani, 1990). Note that, in these additive models, each component function is a univariate smooth function (Ravikumar et al., 2009; Huang et al., 2010; Raskutti et al., 2012; Yuan & Zhou, 2016; Zhao & Liu, 2012) or is defined on grouped features with prior structure information (Chen et al., 2017; Yin et al., 2012). Although these sparse additive models can conduct nonlinear feature selection, all of them do not explore the important feature interaction without prior structure information. Recently, the shrunk additive least square approximation (SALSA) (Kandasamy & Yu, 2016) method was introduced to utilizing the feature interactions, but without feature selection mechanism.

On the other hand, the sparse sample selection arises from learning tasks with large-scale data. The generalized Lasso was proposed in (Roth, 2004) to handle the regression problem with addressing sample sparsity, and its learning theory has been studied in (Shi et al., 2011). Recently, Nyström approximation has been used for selecting important samples (landmark points) in kernel methods, which show that the predictor can be derived efficiently from data dependent hypothesis spaces associated with subsamples (Kumar et al., 2012; Alaoui & Mahoney, 2015; Rudi et al., 2015). While some fast algorithms have been developed for sparse kernel regression, none of them is capable of the feature selection and provides the interpretability of prediction.

To address the above challenges, in this paper, we propose

[*]Equal contribution [1]Department of Electrical and Computer Engineering, University of Pittsburgh, PA, USA [2]Department of Mathematics and Statistics, College of Science, Huazhong Agricultural University, Wuhan, China [3]JD Finance America Corporation, Mountain View, CA, USA. Correspondence to: Hong Chen <chenh@mail.hzau.edu.cn>, Guodong Liu <guodong.liu.e@pitt.edu>.

Proceedings of the 37^{th} International Conference on Machine Learning, Vienna, Austria, PMLR 119, 2020.

a novel *sparse shrunk additive model* (SSAM) for jointly selecting features and samples with learning the feature interactions and mining the structure information among features. Different to previous models, our new method will simultaneously conduct sparse feature selection, sparse sample selection, and feature interactions learning. Our SSAM can utilize the component functions from general continuous and bounded function space (Sun & Wu, 2011; Chen et al., 2016) and can be implemented efficiently via the optimization technique in (Nesterov, 2013).

More important, to better understand the learning theory properties of SSAM, we investigate its convergence rate and sparsity. The proposed SSAM involves the shrunk structure on features and the ℓ_1-norm regularization on data dependent hypothesis spaces. While these features provide the superior flexibility and adaptivity of SSAM, there are new technical difficulties to characterize its theory properties. To address the new difficulties, we introduce a novel decomposition on the excess generalization error, and develop the recent approximation techniques with integral operator and concentration estimates with empirical covering numbers. Our main contributions in this paper include:

- A sparse shrunk additive algorithm is proposed to improve the feature selection ability of nonlinear models. It is a uniform framework to bridge sparse feature selection, sparse sample selection, and feature interaction structure learning tasks. SSAM can be implemented efficiently and its effectiveness is supported by the empirical studies.
- Generalization bound on the excess risk is provided for SSAM under mild conditions, which implies the fast convergence rate can be achieved. Additionally, the necessary and sufficient condition is derived to characterize the sparsity of SSAM.

2. Sparse Shrunk Additive Models

Let $\mathcal{X} \subset \mathbb{R}^n$ be an explanatory feature space and let $\mathcal{Y} \subset [-1,1]$ be the response set. Let $\mathbf{z} := \{z_i\}_{i=1}^m = \{(x_i, y_i)\}_{i=1}^m$ be independent copies of a random sample (x, y) following an unknown intrinsic distribution ρ on $\mathcal{Z} := \mathcal{X} \times \mathcal{Y}$. Denote the marginal distribution of ρ on $\mathcal{X}$ as $\rho_{\mathcal{X}}$ and denote the conditional distribution for given $x \in \mathcal{X}$ as $\rho(\cdot|x)$. Given $\mathbf{z}$, the main goal of regression learning is to infer a functional relation between the input $x \in \mathcal{X}$ and the corresponding output $y \in \mathcal{Y}$. Usually, the expected risk associated with least squares loss is used to evaluate the prediction performance, which is denoted by

$$\mathcal{E}(f) = \int_{\mathcal{Z}} (y - f(x))^2 d\rho(x, y).$$

In theory, the minimizer of $\mathcal{E}(f)$ over all measurable functions is the regression function

$$f_\rho(x) = \int_{\mathcal{Y}} y d\rho(y|x).$$

2.1. Sparse additive models

Additive models (Hastie & Tibshirani, 1990) aim to find the predictor in the special hypothesis space $\mathcal{F} = \{f : f(X) = \sum_{j=1}^n f_j(X_j), X = (X_1, ..., X_n) \in \mathcal{X}\}$. Here, each $f_j \in \mathcal{F}_j$ is one-dimensional smooth function, and its typical examples include the spline function and the Gaussian function. The optimization framework of standard additive model is

$$\min_{f_j} \frac{1}{m} \sum_{i=1}^m (y_i - \sum_{j=1}^n f_j(x_{ij}))^2. \tag{1}$$

Theoretical analysis on (1) shows the good performance of additive model relies on the condition that the number of features n is not large relative to the sample size m.

The algorithm of sparse additive models (SpAM) (Ravikumar et al., 2009) is proposed to address the feature selection in the high dimensional setting, which can be formulated as the following regularized framework

$$\min_{f_j} \Big\{ \frac{1}{m} \sum_{i=1}^m (y_i - \sum_{j=1}^n f_j(x_{ij}))^2 + \lambda \sum_{j=1}^n \|f_j\| \Big\}, \tag{2}$$

where $\lambda > 0$ is a regularization parameter and $\sum_{j=1}^n \|f_j\|$ behaves liken an ℓ_1 ball across different components to encourage functional sparsity (Ravikumar et al., 2009; Yin et al., 2012). The SpAM (2) can be solved efficiently in terms of the back-fitting algorithm (Hastie et al., 2001), and has been extended to the group sparse additive regression (Huang et al., 2010; Raskutti et al., 2012; Yin et al., 2012).

2.2. Shrunk additive models

Although SpAM (2) has nice properties, it ignores the interactions between features. Recently, a novel method, called *shrunk additive least squares approximation* (SALSA), is proposed in (Kandasamy & Yu, 2016) and has shown satisfactory prediction performance.

For any given $1 \le k \le n$ and $\{1, 2, ..., n\}$, we denote $d = \binom{n}{k}$ as the number of index subsets with k elements . It is easy to see that $d = n$ as $k = 1$ and $d = \frac{n(n-1)}{2}$ as $k = 2$. Let $x^{(j)} \in \mathbb{R}^k$ be a subset of x with k features and denote its corresponding space as $\mathcal{X}^{(j)}$.

Denote $\mathcal{H}_{K^{(j)}}$ as a reproducing kernel Hilbert space (RKHS) (Aronszajn, 1950; Scholköpf & Smola, 2001; Shawe-Taylor & Cristianini, 2004) associated with a symmetric and positive definite kernel $K^{(j)} : \mathcal{X}^{(j)} \times \mathcal{X}^{(j)} \to \mathbb{R}, j \in \{1, ..., d\}$.

The SALSA is dependent on the hypothesis space with additive kernels, which is defined by:

$$\mathcal{H} = \Big\{ \sum_{j=1}^{d} f^{(j)} : f^{(j)} \in \mathcal{H}_{K^{(j)}}, j = 1, 2, ..., d \Big\}.$$

Indeed, $(\mathcal{H}, \|\cdot\|_K)$ also is an RKHS for $K = \sum_{j=1}^{d} K^{(j)}$, where $\|f\|_K^2 = \inf\{\sum_{j=1}^{d} \|f^{(j)}\|_{K^{(j)}}^2 : f = \sum_{j=1}^{d} f^{(j)}\}$ (Raskutti et al., 2012; Christmann & Zhou, 2016; Yuan & Zhou, 2016).

Given training samples $\mathbf{z} = \{(x_i, y_i)\}_{i=1}^m$, the SALSA in (Kandasamy & Yu, 2016) can be formulated as the following optimization problem:

$$\tilde{f}_{\mathbf{z}} = \mathop{\arg\min}_{f=\sum_{j=1}^{d} f^{(j)} \in \mathcal{H}} \Big\{ \frac{1}{n} \sum_{i=1}^{m} \Big(y_i - \sum_{j=1}^{d} f^{(j)}(x_i^{(j)}) \Big)^2 + \eta \sum_{j=1}^{d} \|f^{(j)}\|_{K^{(j)}}^2 \Big\}, \tag{3}$$

where $\eta > 0$ is a regularization parameter.

Remark 6 in (Kandasamy & Yu, 2016) tells us that the predictor of SALSA can be expressed as:

$$\tilde{f}_{\mathbf{z}} = \sum_{j=1}^{d} \tilde{f}_{\mathbf{z}}^{(j)} = \sum_{j=1}^{d} \sum_{i=1}^{m} w_i K^{(j)}(x_i^{(j)}, \cdot), w_i \in \mathbb{R}\,.$$

It also has been demonstrated that SALSA in (3) can be considered as kernel ridge regression with shrunk features and additive kernels (Kandasamy & Yu, 2016). Despite nice theoretical and empirical analysis, SALSA does not address the sparsity of shrunk features. For high dimensional data, the sparsity on shrunk features usually is benefit to explore the structure information among features, which will improve the interpretability of learning model.

2.3. New sparse shrunk additive models

To improve the sparsity of SALSA, we propose a new algorithm, named as *sparse shrunk additve models* (SSAM). Some sparse methods (*e.g.*, Lasso (Tibshirani, 1996) and kernelized Lasso (Roth, 2004)) can be considered as the special cases of our new model. It is interesting that SSAM also is a natural but nontrivial extension of sparse regularized regression in data dependent hypothesis spaces (Shi et al., 2011; Sun & Wu, 2011; Feng et al., 2016).

For any given training samples $\mathbf{z}$, we introduce the following data dependent hypothesis space:

$$\mathcal{H}_{\mathbf{z}} = \{f : f(x) = \sum_{j=1}^{d} f^{(j)}(x^{(j)}), f^{(j)} \in \mathcal{H}_{\mathbf{z}}^{(j)}\}, \tag{4}$$

where $\mathcal{H}_{\mathbf{z}}^{(j)} = \{f^{(j)} = \sum_{i=1}^{m} \alpha_i^{(j)} K^{(j)}(x_i^{(j)}, \cdot) : \alpha_i^{(j)} \in \mathbb{R}\}$ and $K^{(j)} : \mathcal{X}^{(j)} \times \mathcal{X}^{(j)} \to \mathbb{R}$ be a continuous function satisfying $\|K^{(j)}\|_\infty < +\infty$. Without loss of generality, this paper assumes $\|K^{(j)}\|_\infty \le 1$ for each $1 \le j \le d$.

The predictor of SSAM can be expressed as

$$f_{\mathbf{z}} = \sum_{j=1}^{d} f_{\mathbf{z}}^{(j)} = \sum_{j=1}^{d} \sum_{t=1}^{m} \hat{\alpha}_t^{(j)} K^{(j)}(x_t^{(j)}, \cdot),$$

where, for $1 \le t \le m$ and $1 \le j \le d$,

$$\{\hat{\alpha}_t^{(j)}\} = \mathop{\arg\min}_{\alpha_t^{(j)} \in \mathbb{R}, t, j} \Big\{ \lambda \sum_{j=1}^{d} \sum_{t=1}^{m} |\alpha_t^{(j)}| + \frac{1}{m} \sum_{i=1}^{m} \Big(y_i - \sum_{j=1}^{d} \sum_{t=1}^{m} \alpha_t^{(j)} K^{(j)}(x_t^{(j)}, x_i^{(j)}) \Big)^2 \Big\}. \tag{5}$$

Let $\alpha^{(j)} = (\alpha_1^{(j)}, ..., \alpha_m^{(j)})^T \in \mathbb{R}^m$ and $\mathbf{K}_i^{(j)} = (K^{(j)}(x_1^{(j)}, x_i^{(j)}), ..., K^{(j)}(x_m^{(j)}, x_i^{(j)}))^T \in \mathbb{R}^m$. Denote $\mathbf{K}_i = ((\mathbf{K}_i^{(1)})^T, ..., (\mathbf{K}_i^{(d)})^T)^T \in \mathbb{R}^{md}$ and $\alpha = ((\alpha^{(1)})^T, ..., (\alpha^{(d)})^T)^T \in \mathbb{R}^{md}$, we can see $\sum_{j=1}^{d} (\mathbf{K}_i^{(j)})^T \alpha^{(j)} = \mathbf{K}_i^T \alpha$. Moreover, by denoting $\mathbf{Y} = (y_1, y_2, ..., y_m)^T \in \mathbb{R}^m$ and $\mathbf{K} = (\mathbf{K}_1, ..., \mathbf{K}_m)^T \in \mathbb{R}^{m \times md}$, we have

$$\hat{\alpha} = \mathop{\arg\min}_{\alpha \in \mathbb{R}^{md}} \Big\{ \frac{1}{m} \|\mathbf{Y} - \mathbf{K}\alpha\|_2^2 + \lambda \|\alpha\|_1 \Big\}. \tag{6}$$

Moreover, for $j \in \{1, ..., d\}$ and $q \in \{1, 2\}$, define

$$\|f^{(j)}\|_{\ell_q}^q = \inf \Big\{ m^{q-1} \sum_{t=1}^{m} |\alpha_t^{(j)}|^q : f^{(j)} = \sum_{t=1}^{m} \alpha_t^{(j)} K^{(j)}(x_t^{(j)}, \cdot) \Big\}$$

and $\|f\|_{\ell_q}^q := \sum_{j=1}^{d} \|f^{(j)}\|_{\ell_q}^q$ for $f = \sum_{j=1}^{d} f^{(j)}$. Then, we can formulate SSAM from the viewpoint of function approximation as below

$$f_{\mathbf{z}} = \mathop{\arg\min}_{f \in \mathcal{H}_{\mathbf{z}}} \Big\{ \frac{1}{m} \sum_{i=1}^{m} (y_i - f(x_i))^2 + \lambda \|f\|_{\ell_1} \Big\}. \tag{7}$$

Except the additive structure on $\mathcal{H}_{\mathbf{z}}$, (7) is consistent with the sparse kernel machine in data dependent hypothesis spaces (Roth, 2004; Shi et al., 2011).

SSAM can be transformed to other methods by explicit selections on $k, K^{(j)}$. When $k = 1$ and $K^{(j)}(x^{(j)}, \tilde{x}^{(j)}) = x^{(j)}$, our model is equivalent to Lasso (Tibshirani, 1996). When $k = n$ and $K^{(j)}(x^{(j)}, \tilde{x}^{(j)}) = K(x, \tilde{x})$, SSAM can be considered the kernelized Lasso (Roth, 2004).

Different from SALSA (Kandasamy & Yu, 2016), our SSAM is based on general kernel, which is not necessary to be a Mercer kernel. Moreover, our SSAM not only can handle regression prediction by using the interactions between features, but also can explore the structure of shrunk features for model selection. The previous SALSA only works for prediction task.

2.4. Comparisons with the related methods

Now we provide some comparisons for SSAM in (5) with the related regularized methods, including *Kernel ridge regression* (KRR), *Least absolute shrinkage and selection operator* (Lasso) (Tibshirani, 1996), *Kernelized Lasso* (KLasso) (Roth, 2004; Sun & Wu, 2015), *Additive model with kernel regularization* (KAM) (Christmann & Zhou, 2016), *Sparse additive models* (SpAM) (Ravikumar et al., 2009), *Component selection and smoothing operator* (COSSO) (Lin & Zhang, 2006), and *Shrunk additive least squares approximation* (SALSA) (Kandasamy & Yu, 2016). A brief summary is presented in Table 1 to show the algorithmic properties including the component function, the regularizer on each component, sample/feature sparsity, feature interaction, and the number of additive components.

From Table 1, we know that SSAM bridges sparse kernel regression and sparse additive models. In theory, SSAM not only can exploit the interactions among features for prediction, but also handle the selections on features and samples simultaneously. In particular, the selection of shrunk features can be used to characterize the structure among features, which is essentially different from the grouped features under prior knowledge (Yin et al., 2012). By introducing the shrunk features, the proposed SSAM encourages the group features to be selected simultaneously, while the previous sparse additive models (Meier et al., 2009; Huang et al., 2010) usually select feature individually.

Indeed, as shown in (Bach, 2008) ,the nonparametric group Lasso can be seen as a variable selection method in a generalized additive model, and can also be seen as equivalent to learning a convex combination of kernel, a framework referred to multiple kernel learning (MKL). The link between the group Lasso and MKL is established in (Bach, 2008) based on the works in (Bach et al., 2004; G.R.G.Lanckrit et al., 2004). However, there are key deferences between our SSAM and MKL (or group Lasso in (Bach, 2008)):
1) Hypothesis space (continuous and bounded function space VS RKHS). The proposed SSAM only requires the component functions to be continuous and bounded, which are not necessary to be in reproducing kernel Hilbert spaces (RKHS). That is to say, we consider the generalized kernel-based hypothesis space (Shi et al., 2011; Sun & Wu, 2011; Chen et al., 2016), which is not necessary to be associated with positive definite kernel used in (Bach, 2008).
2) Regularization (1-norm with data-dependent hypothesis space VS Hilbert norm with data-independent RKHS). We use the 1-norm on coefficients, which is different from the Hilbert norm used in the nonparametric group Lasso (Bach, 2008). From the function approximation point of view, we find the prediction function from data dependent hypothesis spaces (Shi et al., 2011; Sun & Wu, 2011; Chen et al., 2016; Feng et al., 2016) with sparsity restriction on samples and features simultaneously (via 1-norm). However, the nonparametric group Lasso (Bach, 2008) is associated with data independent RKHS and only addresses the feature sparsity. In addition, the kernel Lasso (Roth, 2004) only focuses on the sample sparsity since it does not consider the input variable decomposition.
3) Learning theory (Error bound based on integral operator approximation and concentration estimate with empirical covering numbers VS Consistency based on covariance operator analysis). According to 1) and 2), the theory analysis for MKL (e.g. (Bach et al., 2004; G.R.G.Lanckrit et al., 2004)) or group Lasso (Bach, 2008) doesn't hold true for our approach under mild restriction on component function. As studied in (Shi et al., 2011; Sun & Wu, 2011; Chen et al., 2016; Feng et al., 2016), the learning theory analysis is much more difficult for data-dependent hypothesis space with generalized kernel. In this paper, we overcame the difficulty of theoretical analysis by developing and integrating the integral operator approximation (Smale & Zhou, 2007; Sun & Wu, 2011; Shi, 2013) and the concentration estimation with empirical covering numbers (Wu et al., 2007; Shi et al., 2011).

3. Theoretical Analysis

We begin this section with some necessary definitions and assumptions used in our analysis. Let $L^2_{\rho_{\mathcal{X}^{(j)}}}$ be a square-integrable function space on $\mathcal{X}^{(j)}$ with distribution $\rho_{\mathcal{X}^{(j)}}$. For each $j \in \{1, 2, ..., d\}$ and $f \in L^2_{\rho_{\mathcal{X}^{(j)}}}$, define the integral operator $L_{K^{(j)}} : L^2_{\rho_{\mathcal{X}^{(j)}}} \to L^2_{\rho_{\mathcal{X}^{(j)}}}$ as

$$L_{K^{(j)}}(f)(x^{(j)}) = \int_{\mathcal{X}^{(j)}} K^{(j)}(x^{(j)}, u) f(u) d\rho_{\mathcal{X}^{(j)}}(u).$$

Define $\tilde{K}^{(j)}(x^{(j)}, \tilde{x}^{(j)}) = \int K^{(j)}(x^{(j)}, u) K^{(j)}(\tilde{x}^{(j)}, u) d\rho_{\mathcal{X}^{(j)}}(u)$. It has been verified in (Sun & Wu, 2011) that $\tilde{K}^{(j)}$ is a Mercer kernel and $L_{\tilde{K}^{(j)}} = L_{K^{(j)}} L^T_{K^{(j)}} : L^2_{\rho_{\mathcal{X}^{(j)}}} \to L^2_{\rho_{\mathcal{X}^{(j)}}}$ is a self-adjoint positive operator with decreasing eigenvalues $\{\lambda^{(j)}_t\}_{t=1}^{\infty}$ and eigenfunctions $\{\psi^{(j)}_t\}_{t=1}^{\infty}$, where $\{\psi^{(j)}_t\}_{t=1}^{\infty}$ form an orthonormal basis of $L^2_{\rho_{\mathcal{X}^{(j)}}}$. For given $r > 0$, define the r-th power $L^r_{\tilde{K}^{(j)}}$ of $L_{\tilde{K}^{(j)}}$ by

$$L^r_{\tilde{K}^{(j)}}(\sum_t c_{j,t}\psi^{(j)}_t) = \sum_t c_{j,t}(\lambda^{(j)}_t)^r \psi^{(j)}_t, \forall (c_{j,t})_{t\in\mathbb{N}} \in \ell_2.$$

Table 1. Properties of kernel methods and additive models ($\surd$ means using the given formulation or information and $\times$ means not available for the information)

property	KRR	KLasso	Lasso	KAM	SpAM	COSSO	SALSA	SSAM
Component function	RKHS	continuous	linear	RKHS	Hilbert	Spline	RKHS	continuous
Regularization	K-norm	1-norm	1-norm	K-norm	2,1-norm	2,1-norm	K-norm	1-norm
Sparsity (sample)	$\times$	$\surd$	$\times$	$\times$	$\times$	$\times$	$\times$	$\surd$
Sparsity (feature)	$\times$	$\times$	$\surd$	$\times$	$\surd$	$\surd$	$\times$	$\surd$
Feature Interaction	$--$	$--$	$\times$	$\times$	$\times$	$\surd$	$\surd$	$\surd$
Component number	1	1	n	n	n	$\sum_{k=1}^{d}\binom{n}{k}$	$\binom{n}{k}*$	$\binom{n}{k}*$

K-norm:=Kernel norm. *The number can be reduced largely by incorporating prior information of features.

Assumption 1. *Assume that $f_\rho = \sum_{j=1}^{d} f_\rho^{(j)}$, where for each $j \in \{1,2,...,d\}$, $f_\rho^{(j)} : \mathcal{X}^{(j)} \to \mathbb{R}$ is a function of the form $f_\rho^{(j)} = L^r_{K^{(j)}}(g_\rho^{(j)})$ with some $r > 0$ and $g_\rho^{(j)} \in L^2_{\rho_{\mathcal{X}^{(j)}}}$.*

This regularity condition on f_ρ has been studied for coefficient-based regularized regression with general kernel (Sun & Wu, 2011; Shi, 2013). For the additive model with Mercer kernel, similar assumption has been introduced in (Christmann & Zhou, 2016).

We also need the Lipschitz continuous condition on each kernel $K^{(j)}$. The restrictive condition has been studied extensively in learning theory of kernel methods, *e.g.*, (Shi et al., 2011; Shi, 2013). In particular, the Gaussian kernel satisfies this condition.

Assumption 2. *For each $j \in \{1,2,...,d\}$, the kernel function $K^{(j)} : \mathcal{X}^{(j)} \times \mathcal{X}^{(j)} \to \mathbb{R}$ is $\mathcal{C}^s$ with some $s > 0$ satisfying*

$$\|K^{(j)}(u,v) - K^{(j)}(u,v')\| \le c_s \|v - v'\|_2^s, \forall u, v, v' \in \mathcal{X}^{(j)}$$

for some positive constant c_s.

From the definition of f_ρ and $\mathcal{Y} \in [-1,1]$, we know that $|f_\rho(x)| \le 1$ for any $x \in \mathcal{X}$. Thus, we can utilize the following projection operator to get tight error estimate which is a standard technique in error analysis (Cucker & Zhou, 2007; Steinwart et al., 2009).

Definition 1. *The projection operator π is defined on the space of measurable functions $f : \mathcal{X} \to \mathbb{R}$ as $\pi(f)(x) = \max\{-1, \min\{f(x), 1\}\}$.*

Denote

$$p = \begin{cases} 2k/(k+2s), & \text{if } s \in (0,1]; \\ 2k/(k+2), & \text{if } s \in (1, 1+k/2]; \\ k/s, & \text{if } s \in (1+k/2, \infty). \end{cases} \tag{8}$$

Our first theoretical result is the upper bound of $\mathcal{E}(\pi(f_{\mathbf{z}})) - \mathcal{E}(f_\rho)$.

Theorem 1. *Let Assumptions 1 and 2 be true. For any $0 < \delta < 1$, with confidence $1-\delta$, there exists positive constant $\tilde{c}_1$ independent of m, δ such that:*

(1) If $r \in (0, \frac{1}{2})$ in Assumption 1, setting $\lambda = m^{-\theta_1}$ with $\theta_1 \in (0, \frac{2}{2+p})$,

$$\mathcal{E}(\pi(f_{\mathbf{z}})) - \mathcal{E}(f_\rho) \le \tilde{c}_1 \log(8/\delta) m^{-\gamma_1},$$

where $\gamma_1 = \min\{2r\theta_1, \frac{1-\theta_1+2r\theta_1}{2}, \frac{2}{2+p} - (2-2r)\theta_1, \frac{2(1-p\theta_1)}{2+p}\}$.

(2) If $r \ge \frac{1}{2}$ in Assumption 1, taking $\lambda = m^{-\theta_2}$ with some $\theta_2 \in (0, \frac{2}{2+p})$,

$$\mathcal{E}(\pi(f_{\mathbf{z}})) - \mathcal{E}(f_\rho) \le \tilde{c}_1 \log(8/\delta) m^{-\gamma_2},$$

where $\gamma_2 = \min\left\{\theta_2, \frac{1}{2}, \frac{2}{2+p} - \theta_2\right\}$.

Theorem 1 provides the upper bound of generalization error to SSAM with Lipshitz continuous kernel. For $r \in (0, \frac{1}{2})$, as $s \to \infty$, we have $\gamma_1 \to \min\{2r\theta_1, \frac{1}{2} + (r - \frac{1}{2})\theta, 1 - 2\theta_1 + 2r\theta_1\}$. Moreover, when $r \to \frac{1}{2}$ and $\theta_1 \to \frac{1}{2}$, the convergence rate $O(m^{-\frac{1}{2}})$ can be reached.

For $r \ge \frac{1}{2}$, taking $\theta_2 = \frac{1}{2+p}$, we get the convergence rate $O(m^{-\frac{1}{2+p}})$.

The following result is about a special case when $f_\rho^{(j)} \in \mathcal{H}^{(j)}$.

Theorem 2. *Assume that $f_\rho^{(j)} \in \mathcal{H}^{(j)}$ for each $1 \le j \le d$. Take $\lambda = m^{-\frac{2}{2+3p}}$ in (5). For any $0 < \delta < 1$, with confidence $1-\delta$ we have*

$$\mathcal{E}(\pi(f_{\mathbf{z}})) - \mathcal{E}(f_\rho) \le \tilde{c}_2 \log(1/\delta) m^{-\frac{2}{2+3p}},$$

where $\tilde{c}_2$ is a positive constant independent of m, δ, and p is defined in (8).

Under the strong condition on f_ρ, the convergence rate can be arbitrary close to $O(m^{-1})$ as $s \to \infty$.

Now we summarize the comparisons on the related convergence analysis of additive models with feature interactions.

- For SALSA in (Kandasamy & Yu, 2016), the convergence rate with polynomial decay is also obtained under mild condition on f_ρ. Different from our work, the previous analysis is limited to the Mercer kernel and the error is expressed with the expectation version.
- For the generalized SpAM in (Tyagi et al., 2016), theoretical analysis demonstrates its effectiveness to estimate the underlying component functions, which provides stronger guarantees than generalization bound. However, the condition on $\mathcal{H}^{(j)}$ is much restrictive than SSAM.
- For the fixed design setting, the COSSO estimator in (Lin & Zhang, 2006) has a convergence rate $O(m^{-\frac{\tilde{s}}{2\tilde{s}+1}})$, where $\tilde{s}$ is the order of smoothness of the components in Sobolev space. It can be seen from Theorem 2 that the faster learning rate of SSAM can be reached as $K \in \mathcal{C}^\infty$.

In the future, it is natural to extend the current result from uniform boundedness to unbounded sampling by the analysis techniques in (Steinwart et al., 2009; Wang & Zhou, 2011; Guo & Zhou, 2013).

Besides the generalization ability, SSAM also advocates the sparsity on features and samples by employing the ℓ_1 regularization. The sparsity of SSAM can be characterized as below.

Theorem 3. *For $t \in \{1, 2, ..., m\}$ and $j \in \{1, 2, ..., d\}$, $\hat{\alpha}_t^{(j)} = 0$ if and only if*

$$\Big|\frac{1}{m}\sum_{i=1}^{m}(y_i - f_{\mathbf{z}}(x_i))K^{(j)}(x_t^{(j)}, x_i^{(j)})\Big| < \frac{\lambda}{2}.$$

Theorem 3 provides a necessary and sufficient condition for the zero pattern of $\hat{\alpha}$. In terms of the discussions in (Shi et al., 2011), Theorem 3 also implies the probabilistic confidence bound to ensure the sparsity of $\hat{\alpha}_t^{(j)}$ in (5). In particular, for the fixed design setting, the sparsity recovery may be achieved by adding some conditions (Feng et al., 2016; Yang et al., 2016). We leave it for future study.

4. Proof Sketches

The proofs of Theorems 1 and 2 involve a integration of techniques for error analysis with integral operator approximation (Smale & Zhou, 2007; Sun & Wu, 2011; Shi, 2013; Nie & Wang, 2015) and the empirical process theory for analyzing kernel methods (Pinelis, 1994; Wu et al., 2007; Christmann & Zhou, 2016). The proof of Theorem 3 follows the analysis technique for sparse characterization (Shi et al., 2011; Sun & Wu, 2015).

The key to bound $\mathcal{E}(\pi(f_{\mathbf{z}})) - \mathcal{E}(f_\rho)$ is a novel error decomposition, where some intermediate functions are constructed as the stepping stone functions. Then, we bound the decomposed terms respectively in terms of operator approximation and concentration equalities for empirical processes.

From Proposition 1 in (Shi, 2013), we know that $L_{K^{(j)}}^T = UL_{\tilde{K}^{(j)}}^{\frac{1}{2}}$ and $L_{K^{(j)}} = L_{\tilde{K}^{(j)}}^{\frac{1}{2}}U^T$ for each $j \in \{1, 2, ..., d\}$, where U is a partial isometry on $L^2_{\rho_{\mathcal{X}^{(j)}}}$ with U^TU being the orthogonal prediction onto the RKHS $\mathcal{H}_{\tilde{K}^{(j)}}$.

For any $j \in \{1, 2, ..., d\}$, define the intermediate function $f_\lambda^{(j)}$ by

$$f_\lambda^{(j)} = \arg\min_{f \in L^2_{\rho_{\mathcal{X}^{(j)}}}} \Big\{\|L_{K^{(j)}}f^{(j)} - f_\rho^{(j)}\|^2_{L^2_{\rho_{\mathcal{X}^{(j)}}}} + \lambda\|U^Tf^{(j)}\|^2_{L^2_{\rho_{\mathcal{X}^{(j)}}}}\Big\}. \quad (9)$$

Denote $f_\lambda = \sum_{j=1}^d f_\lambda^{(j)}$ and $g_\lambda = \sum_{j=1}^d g_\lambda^{(j)}$ with $g_\lambda^{(j)} = L_{K^{(j)}}f_\lambda^{(j)}$.

Define the empirical version of g_λ as

$$\hat{g}_\lambda(x) = \frac{1}{m}\sum_{i=1}^{m}\sum_{j=1}^{d} f_\lambda^{(j)}(x_i^{(j)})K^{(j)}(x_i^{(j)}, x^{(j)}), x \in \mathcal{X}. \quad (10)$$

Now we give the following error decomposition.

Proposition 1. *For $f_{\mathbf{z}}, \hat{g}_\lambda$ defined in (5) and (10), respectively, there holds*

$$\mathcal{E}(\pi(f_{\mathbf{z}})) - \mathcal{E}(f_\rho) \le E_1 + E_2 + E_3,$$

where

$$\begin{aligned} E_1 &= \mathcal{E}(\pi(f_{\mathbf{z}})) - \mathcal{E}_{\mathbf{z}}(\pi(f_{\mathbf{z}})) + \mathcal{E}_{\mathbf{z}}(\hat{g}_\lambda) - \mathcal{E}(\hat{g}_\lambda), \\ E_2 &= \mathcal{E}(\hat{g}_\lambda) - \mathcal{E}(g_\lambda) + \lambda\|\hat{g}_\lambda\|_{\ell_1} \end{aligned}$$

and

$$E_3 = \mathcal{E}(g_\lambda) - \mathcal{E}(f_\rho).$$

Proof. According the definition of $f_{\mathbf{z}}$, we have

$$\begin{aligned} & \mathcal{E}(\pi(f_{\mathbf{z}})) - \mathcal{E}(f_\rho) \\ \le\ & \mathcal{E}(\pi(f_{\mathbf{z}})) - \mathcal{E}_{\mathbf{z}}(\pi(f_{\mathbf{z}})) + \mathcal{E}_{\mathbf{z}}(\hat{g}_\lambda) - \mathcal{E}(f_\rho) + \lambda\|\hat{g}_\lambda\|_{\ell_1} \\ & + \Big\{\mathcal{E}_{\mathbf{z}}(f_{\mathbf{z}}) + \lambda\|f_{\mathbf{z}}\|_{\ell_1} - (\mathcal{E}_{\mathbf{z}}(\hat{g}_\lambda) + \lambda\|\hat{g}_\lambda\|_{\ell_1})\Big\} \\ \le\ & \mathcal{E}(\pi(f_{\mathbf{z}})) - \mathcal{E}_{\mathbf{z}}(\pi(f_{\mathbf{z}})) + \mathcal{E}_{\mathbf{z}}(\hat{g}_\lambda) - \mathcal{E}(f_\rho) \\ & + \lambda\|\hat{g}_\lambda\|_{\ell_1}. \end{aligned} \quad (11)$$

Note that

$$\mathcal{E}_{\mathbf{z}}(\hat{g}_\lambda) - \mathcal{E}(f_\rho) = (\mathcal{E}(\hat{g}_\lambda) - \mathcal{E}(g_\lambda)) + \mathcal{E}(g_\lambda) - \mathcal{E}(f_\rho) + \mathcal{E}_{\mathbf{z}}(\hat{g}_\lambda) - \mathcal{E}(\hat{g}_\lambda). \quad (12)$$

Combining both (11) and (12), we get the desired decomposition.

□

The error term E_1 measures the divergence between the empirical risk and the corresponding expected risk, which usually is called sample error in learning theory. In terms of recent theoretical progress for learning with data dependent hypothesis spaces (Shi et al., 2011; Shi, 2013; Feng et al., 2016), we can bound sample error E_1 via concentration inequality associated with empirical covering numbers (Wu et al., 2007; Christmann & Zhou, 2016).

The error term E_2 reflects the drift risk for learning with hypothesis spaces $\mathcal{H}_{\mathbf{z}}$ and $\mathcal{H}$, and hence is called as the hypothesis error. By relating $\mathcal{E}(\hat{g}_\lambda) - \mathcal{E}(g_\lambda)$ with $\sum_{j=1}^d \|\hat{g}_\lambda^{(j)} - g_\lambda^{(j)}\|_{L^2_{\rho_{\mathcal{X}^{(j)}}}}$, we can estimate this hypothesis error through the inequality in Hilbert space (Pinelis, 1994; Smale & Zhou, 2007).

The error term E_3 is called the approximation error, which describes the approximation ability of regularized scheme. Following the approximation analysis with integral operator in (Smale & Zhou, 2007; Shi, 2013; Nie & Wang, 2015), we derive the upper bound of E_2 based on the properties of $L_{\tilde{K}^{(j)}}, 1 \le j \le d$.

Detail technical proofs are provided in the supplementary materials.

5. Experimental Results

This section shows the empirical evaluation of SSAM. We first introduce the experimental setups following (Kandasamy & Yu, 2016), and then validate SSAM's ability for feature selection and regression prediction.

We consider SSAM for pairwise interaction setting, and set $k = 2, d = \binom{n}{2}$. Similar with (Kandasamy & Yu, 2016), each kernel on $\mathcal{X}^{(j)}$ is generated from Gaussian kernel. For example, when $x_s^{(j)} = (x_{s1}, x_{s2})$ and $x_t^{(j)} = (x_{t1}, x_{t2})$, the shrunk kernel $K^{(j)}(x_s^{(j)}, x_t^{(j)}) = \exp\{-\frac{(x_{s1}-x_{t1})^2}{2\mu_1^2}\} \cdot \exp\{-\frac{(x_{s2}-x_{t21})^2}{2\mu_2^2}\}$, where $\mu_i = 4.5\sigma_i m^{-\frac{1}{10}}$ and σ_i is the standard deviation on i-th coordination. The regularization parameter λ is chosen via five-fold cross validation with respect to the mean square error (MSE).

We implement our SSAM method via accelerated proximal gradient methods (Nesterov, 2013) to get the coefficient vector $\hat{\alpha}$. For sparse representation and feature selection, we compute $\sum_{t=1}^m \hat{\alpha}_t^{(j)}$ on the j-th pairwise features, and then select the informative shrunk features. For synthetic data, we compare our model with COSSO (Lin & Zhang, 2006) to validate our motivation for feature selection. For real-word benchmark data, we compare MSE of SSAM with SALSA (Kandasamy & Yu, 2016), COSSO (Lin & Zhang, 2006), SpAM (Ravikumar et al., 2009), and Lasso (Tibshirani, 1996).

5.1. Experiments with synthetic data

Following the ideas in (Lin & Zhang, 2006; Yin et al., 2012), we use two different types of data to evaluate the model selection ability of SSAM. The first type of synthetic data has at most one informative pairwise features and the second one has at least two pairwise features. Since SALSA does not concern the selection of shrunk features, we compare the performance SSAM with COSSO (Lin & Zhang, 2006). As shown in Table 1, COSSO is based on component functions on both single and pairwise input features.

Generate synthetic data: We generate the n-dimensional input $x_i = (x_{i1}, x_{i2}, ..., x_{in})^T$ with $x_{ij} = \frac{W_{ij}+\eta U_i}{1+\eta}$ and $n = 10$, where W and U are sampled from independent uniform distributions defined in $[-0.5, 0.5]$. Parameter η controls the magnitude of correlation. Inputs are independent if $\eta = 0$ and correlated if $\eta = 1$.

Example set I: We apply SSAM with 100 training samples on three underlying functions (a. simple additive model, b. simple pairwise interaction model, c. multi-ways interaction model). For $x_t = (x_{t1}, x_{t2}, ..., x_{tn})^T$,

$$\begin{aligned} a.\ f^*(x_t) &= x_{t1} + x_{t2} + x_{t3} + exp(-x_{t4}) \\ b.\ f^*(x_t) &= (2x_{t1} - 1)(2x_{t2} - 1) \\ c.\ f^*(x_t) &= (2sin(x_{t1}) - 1)(2sin(x_{t2}) - 1) \\ &\quad \cdot(2sin(x_{t3}) - 1)(2sin(x_{t4}) - 1) \end{aligned}$$

Example set II: We also apply SSAM with 100 training samples on much complicated interaction models (e. overlapped pairwise interaction, f. independent pairwise interaction, g. circle related pairwise interaction):

$$\begin{aligned} e.\ f^*(x_t) &= (2sin(x_{t1}) - 1)(2sin(x_{t2}) - 1) \\ &\quad + sin(x_{t1})sin(x_{t3}), \\ f.\ f^*(x_t) &= 2exp(x_{t1} + x_{t2} + 0.2) + 2exp^{-1}(x_{t3} + x_{t4}), \\ g.\ f^*(x_t) &= (2x_{t1} - 1)(2x_{t2} - 1) + (2x_{t2} - 1)(2x_{t3} - 1) \\ &\quad + (2x_{t1} - 1)(2x_{t3} - 1). \end{aligned}$$

The final output is $y = f^*(x) + \epsilon$, where $\epsilon \sim \mathcal{N}(0, 0.25)$. For each example, we make feature selection according to the values of $\sum_{t=1}^{100} \hat{\alpha}_t^{(j)}$ for $j \in \{1, ..., 45\}$. The Precision@τ

Table 2. Precision@τ for feature selection
(a) Synthetic data I

f^*	(m,n,η)	τ	SSAM	COSSO
a	(100,10,0)	4	3.88	3.69
		5	3.92	3.81
		6	3.93	3.85
	(100,10,1)	4	3.37	2.58
		5	3.68	2.80
		6	3.82	2.91
b	(100,10,0)	1	0.97	1
		2	0.97	1
		3	0.97	1
	(100,10,1)	1	0.95	0.62
		2	0.95	0.65
		3	0.98	0.68
c	(100,10,0)	4	3.94	0.63
		5	3.97	0.68
		6	3.97	0.75
	(100,10,1)	4	3.69	0.84
		5	3.87	0.91
		6	3.92	0.94

(b) Synthetic data II

f^*	(m,n,η)	τ	SSAM	COSSO
e	(100,10,0)	2	1.05	0.73
		3	1.13	0.90
		4	1.20	0.90
	(100,10,1)	2	1.04	0.13
		3	1.10	0.16
		4	1.12	0.20
f	(100,10,0)	2	0.72	0.88
		3	0.93	1
		4	1.23	1
	(100,10,1)	2	1.90	0.94
		3	1.94	0.94
		4	1.95	0.97
g	(100,10,0)	3	2.94	2.98
		4	2.94	2.98
		5	2.94	3
	(100,10,1)	3	2.85	2.14
		4	2.85	2.40
		5	2.85	2.49

is used to measure the performance of feature selection, which describes the number of truly informative features in the top-τ selected results. Tables 2(a) and 2(b) provide the average results on Precision@τ after repeating 100 times. In most cases, SSAM performs better than COSSO for feature selection. Especially, SSAM behaves more stable than COSSO in complicated models (*e.g.* c, g) and dependent features.

Table 3. Average MSE on real data.

	SSAM	SALSA	COSSO	SpAM	Lasso
Insulin	**1.0146**	1.0206	1.1379	1.2035	1.1103
Skillcraft	**0.5432**	0.5470	0.5551	0.90545	0.6650
Airfoil	**0.4866**	0.5176	0.5178	0.9623	0.5199
Forestfire	**0.3477**	0.3530	0.3753	0.9694	0.5193
Housing	0.3787	**0.2642**	1.3097	0.8165	0.4452
CCPP	0.0694	0.0678	0.9684	**0.0647**	0.0740
Music	0.6295	**0.6251**	0.7982	0.7683	0.6349
Telemonit	0.0689	**0.0347**	5.7192	0.8643	0.0863

5.2. Experiments with real-world benchmark data

We compare the prediction performance of SSAM with the most related additive models, where eight data sets are used under the same experimental setups in (Kandasamy & Yu, 2016). The data sets from UCI repository (*http://archive.ics.uci.edu/ml*) and (Tu et al., 2012), which include *Insulin* ($n = 50, m = 256$), *Skillcraft* ($n = 18, m = 1700$), *Airfoil* ($n = 40, m = 750$), *Forestfire* ($n = 10, m = 211$), *Housing* ($n = 12, m = 256$), *CCPP* ($n = 59, m = 2000$), *Music* ($n = 90, m = 1000$), *Telemonit* ($n = 19, m = 1000$). As shown in Table 3, on all eight benchmark datasets, our SSAM has best results on four of them, second best results on three of the rest, and third best result on the rest one. Experimental results show that our SSAM has comparable performance with SALSA, even if only pairwise interaction features are used.As shown in (Kandasamy & Yu, 2016), SALSA has shown competitive performance with many nonparametric models and parametric models (but SALSA cannot do feature selection). Therefore, SSAM is effective for regression prediction besides its capacity for sparse feature selection.

5.3. More experimental results

According to the reviewer comments of scalability, we did new experiments on simulated data for the high dimensional setting (20,000 samples and other settings remain the same). The average results (with 20 repeats) in Table 4 demonstrate that SSAM scales well in high dimensional setting.

One reviewer suggested us to compare with more methods beside COSSO. We added new comparison results on simulated data with SpAM and the other new method RMR (Wang et al., 2017) in Table 6. The new results also verify the effectiveness of the proposed method.

In addition, we added new experimental results on real data with RMR in Table 5, This results also show the proposed method is better.

Table 4. Precision@τ for feature selection

f^*	(m,n,η)	τ	SSAM
a	(20000,10,0)	4	3.95
		5	4.00
		6	4.00
	(20000,10,1)	4	3.90
		5	3.90
		6	4.00
c	(20000,10,0)	4	3.90
		5	4.00
		6	4.00
	(20000,10,1)	4	3.70
		5	4.00
		6	4.00
g	(20000,10,0)	3	2.90
		4	2.95
		5	3.00
	(20000,10,1)	3	2.85
		4	2.85
		5	3.00

Table 5. Average MSE on real data.

	SSAM	RMR
Insulin	**1.0146**	1.0198
Skillcraft	**0.5432**	0.6486
Airfoil	**0.4866**	0.5314
Forestfire	**0.3477**	0.3765
Housing	**0.3787**	0.4375
CCPP	0.0694	**0.0667**
Music	0.6295	**0.6210**
Telemonit	**0.0689**	0.0824

Table 6. Precision@τ for feature selection

(a) Synthetic data I

f^*	(m,n,η)	τ	SSAM	SpAM	RMR
a	(100,10,0)	4	3.88	3.85	3.81
		5	3.92	3.90	3.95
		6	3.93	3.97	3.97
	(100,10,1)	4	3.37	3.50	2.64
		5	3.68	3.61	3.02
		6	3.82	4.00	3.06
b	(100,10,0)	1	0.97	0.94	1.00
		2	0.97	1.00	2.00
		3	0.97	1.00	2.00
	(100,10,1)	1	0.95	0.94	0.91
		2	0.95	1.00	0.91
		3	0.98	1.00	0.98
c	(100,10,0)	4	3.94	3.93	3.55
		5	3.97	3.96	3.71
		6	3.97	3.98	3.81
	(100,10,1)	4	3.69	3.54	2.82
		5	3.87	3.83	3.20
		6	3.92	3.40	3.46

(b) Synthetic data II

f^*	(m,n,η)	τ	SSAM	SpAM	RMR
e	(100,10,0)	2	1.05	1.00	1.67
		3	1.13	1.17	1.96
		4	1.20	1.19	2.13
	(100,10,1)	2	1.04	1.00	1.26
		3	1.10	1.13	1.64
		4	1.12	1.30	2.97
f	(100,10,0)	2	0.72	0.89	1.83
		3	0.93	1.00	2.37
		4	1.23	1.00	2.97
	(100,10,1)	2	1.90	2.00	1.13
		3	1.94	2.00	1.66
		4	1.95	2.00	1.93
g	(100,10,0)	3	2.94	2.90	3.00
		4	2.94	2.98	3.00
		5	2.94	3.00	3.00
	(100,10,1)	3	2.85	2.80	2.50
		4	2.85	2.82	2.72
		5	2.85	3.00	2.84

6. Conclusion

In this paper, we proposed a uniform scheme for nonlinear feature and sample selections under additive models. Learning theory analysis has been provided to demonstrate the convergence and sparsity properties of SSAM, where involves novel analysis technique with integral operator and concentration estimation. Experiments on both synthetic and real-world datasets support the effectiveness of our new model.

Acknowledgements

We thank the anonymous reviewers for their helpful comments. G.L. and H.H. were partially supported by U.S. NSF IIS 1836945, IIS 1836938, IIS 1845666, IIS 1852606, IIS 1838627, IIS 1837956. H.C. was partially supported by the National Natural Science Foundation of China (NSFC) under Grants 11671161.

References

Alaoui, A. and Mahoney, M. W. Fast randomized kernel ridge regression with statistical guarantees. In *NIPS*, 2015.

Aronszajn, N. Theory of reproducing kernels. *Trans. Amer.*

Math. Soc., 68:337–404, 1950.

Bach, F. Consistency of the group lasso and multiple kernel learning. *J. Mach. Learn. Res.*, 9:1179–1225, 2008.

Bach, F., Lanckrit, G., and Jordan, M. Multiple kernel learning, conic duality and the smo agorithm. In *ICML*, 2004.

Candes, E. and Tao, T. The dantzig selector: Statistical estimation when p is much larger than n. *Ann. Statist.*, 35(6):2313–2351, 2007.

Chen, H., Xia, H., Cai, W., and Huang, H. Error analysis of generalized nyström kernel regression. In *NIPS*, 2016.

Chen, H., Wang, X., Deng, C., and Huang, H. Group sparse additive machine. In *NIPS*, 2017.

Chen, H., Wang, Y., Zheng, F., Deng, C., and Huang, H. Sparse modal additive model. *IEEE Transactions on Neural Networks and Learning Systems*, 2020, In press.

Christmann, A. and Zhou, D. X. Learning rates for the risk of kernel-based quantile regression estimators in additive models. *Analysis and Applications*, 14(3):449–477, 2016.

Cucker, F. and Zhou, D. X. *Learning Theory: An Approximation Theory Viewpoint*. Cambridge University Press, 2007.

Feng, Y., Lv, S., Hang, H., and Suykens, J. A. Kernelized elastic net regularization: Generalization bounds, and sparse recovery. *Neural Comput.*, 28(3):525–562, 2016.

G.R.G.Lanckrit, N.Cristianini, L. G., P.Barlett, and Jordan, M. Learning the kernel matrix with semidefinite programming. *J. Mach. Learn. Res.*, 5:27–72, 2004.

Guo, Z. and Zhou, D.-X. Concentration estimates for learning with unbounded sampling. *Adv. Comput. Math.*, 38 (1):207–223, 2013.

Hastie, T. and Tibshirani, R. *Generalized Additive Models*, volume 43. Chapman and Hall press, London, 1990.

Hastie, T., Tibshirani, R., and Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer-Verlag, New York, 2001.

Huang, J., Horowitz, J. L., and Wei, F. Variable selection in nonparametric additive models. *Ann. Statist.*, 38(4): 2282–2313, 2010.

Kandasamy, K. and Yu, Y. Additive approximation in high dimensional nonparametric regression via the salsa. In *ICML*, 2016.

Kumar, S., Mohri, M., and Talwalkar, A. Sampling methods for the nyström method. *J. Mach. Learn. Res.*, 13:981–1006, 2012.

Lin, Y. and Zhang, H. H. Component selection and smoothing in smoothing spline analysis of variance models. *Ann. Statist.*, 34(5):2272–2297, 2006.

Meier, L., Van de Geer, S., and Bühlmann, P. High-dimensional additive modeling. *Ann. Statist.*, 37(6B): 3779–3821, 2009.

Nesterov, Y. Gradient methods for minimizing composite functions. *Mathematical Programming*, 140(1):125–161, 2013.

Nie, W. and Wang, C. Constructive analysis for coefficient regularization regression algorithms. *Journal of Mathematical Analysis and Applications*, 431(2):1153–1171, 2015.

Pinelis, I. Optimum bounds for the distributions of martingales in banach spaces. *The Annals of Probability*, pp. 1679–1706, 1994.

Raskutti, G., Wainwright, M. J., and Yu, B. Minimax-optimal rates for sparse additive models over kernel classes via convex programming. *J. Mach. Learn. Res.*, 13:389–427, 2012.

Ravikumar, P., Lafferty, J., Liu, H., and Wasserman, L. Sparse additive models. *J. Royal. Statist. Soc B.*, 71(5): 1009–1030, 2009.

Roth, V. The generalized lasso. *IEEE Trans. Neural Networks*, 15(1):16–28, 2004.

Rudi, A., Camoriano, R., and Rosasco, L. Less is more: Nyström computational regularization. In *NIPS*, pp. 1657–1665, 2015.

Scholköpf, B. and Smola, A. J. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT press, 2001.

Shawe-Taylor, J. and Cristianini, N. *Kernel Methods for Pattern Analysis*. Cambridge university press, 2004.

Shi, L. Learning theory estimates for coefficient-based regularized regression. *Appl. Comput. Harmon. Anal.*, 34 (2):252–265, 2013.

Shi, L., Feng, Y., and Zhou, D. X. Concentration estimates for learning with ℓ_1-regularizer and data dependent hypothesis spaces. *Appl. Comput. Harmon. Anal.*, 31(2): 286–302, 2011.

Smale, S. and Zhou, D. X. Learning theory estimates via integral operators and their approximations. *Constr. Approx.*, 26(2):153–172, 2007.

Steinwart, I., Hush, D., and Scovel, C. Optimal rates for regularized least squares regression. In *COLT*, 2009.

Sun, H. and Wu, Q. Least square regression with indefinite kernels and coefficient regularization. *Appl. Comput. Harmon. Anal.*, 30(1):96–109, 2011.

Sun, H. and Wu, Q. Sparse representation in kernel machines. *IEEE Transactions on Neural Networks and Learning Systems*, 26(10):2576–2582, 2015.

Tibshirani, R. Regression shrinkage and selection via the lasso. *J. Royal. Statist. Soc B.*, 58(1):267–288, 1996.

Tu, Z. et al. Integrative analysis of a cross-loci regulation network identifies app as a gene regulating insulin secretion from pancreatic islets. *PLoS Genet*, 8(12):e1003107, 2012.

Tyagi, H., Kyrillidis, A., Gärtner, B., and Krause, A. Learning sparse additive models with interactions in high dimensions. In *AISTATS*, 2016.

Wang, C. and Zhou, D.-X. Optimal learning rates for least squares regularized regression with unbounded sampling. *J. Complexity*, 27(1):55–67, 2011.

Wang, X., Chen, H., Cai, W., Shen, D., and Huang, H. Regularized modal regression with applications in cognitive impairment prediction. In *Advances in neural information processing systems*, pp. 1448–1458, 2017.

Wu, Q., Ying, Y., and Zhou, D.-X. Multi-kernel regularized classifiers. *J. Complexity*, 23(1):108–134, 2007.

Yang, L., Lv, S., and Wang, J. Model-free variable selection in reproducing kernel hilbert space. *J. Mach. Learn. Res.*, 17:1–24, 2016.

Yin, J., Chen, X., and Xing, E. P. Group sparse additive models. In *ICML*, 2012.

Yuan, M. and Zhou, D. X. Minimax optimal rates of estimation in high dimensional additive models. *Ann. Statist.*, 44(6):2564–2593, 2016.

Zhao, T. and Liu, H. Sparse additive machine. In *AISTATS*, pp. 1435–1443, 2012.