

SCALABLE KERNEL-BASED VARIABLE SELECTION WITH SPARSISTENCY*

BY XIN HE[†], JUNHUI WANG[†] AND SHAOGAO LV[‡]

Department of Mathematics, City University of Hong Kong[†] and Center of Statistics, Southwestern University of Finance and Economics[‡]

Variable selection is central to high-dimensional data analysis, and various algorithms have been developed. Ideally, a variable selection algorithm shall be flexible, scalable, and with theoretical guarantee, yet most existing algorithms cannot attain these properties at the same time. In this article, a three-step variable selection algorithm is developed, involving kernel-based estimation of the regression function and its gradient functions as well as a hard thresholding. Its key advantage is that it assumes no explicit model assumption, admits general predictor effects, allows for scalable computation, and attains desirable asymptotic sparsistency. The proposed algorithm can be adapted to any reproducing kernel Hilbert space (RKHS) with different kernel functions, and can be extended to interaction selection with slight modification. **Its computational cost is only linear in the data dimension, and can be further improved through parallel computing.** The sparsistency of the proposed algorithm is established for general RKHS under mild conditions, including linear and Gaussian kernels as special cases. Its effectiveness is also supported by a variety of simulated and real examples.

1. Introduction. Variable selection has attracted tremendous interests from both researchers and practitioners, due to the availability of large number of variables in many real applications. In such scenarios, identifying the truly informative variables for the objective of analysis has become a key factor to facilitate statistical modeling and analysis. Ideally, a variable selection algorithm shall be flexible, scalable, and with theoretical guarantee. To be more specific, the algorithm shall not assume restrictive model assumptions, so that it is applicable to data with complex structures; its implementation shall be computationally efficient and able to take advantage of high performance computing platform; it shall have theoretical guarantee on its asymptotic sparsistency in identifying the truly informative variables.

In literature, many variable selection algorithms have been developed in the regularization framework assuming certain working model set. The most popular working model set is to assume a linear model, where the variable selection task

*Research is supported in part by HK GRF-11302615 and HK GRF-11331016.

MSC 2010 subject classifications: Primary 68Q32, 62G99; secondary 62J07

Keywords and phrases: Interaction selection, ridge regression, RKHS, scalability, sparsistency, variable selection

simplifies to identifying nonzero coefficients. Under the linear model assumption, the regularization framework consists of a least square loss function for the linear model as well as a sparsity-inducing regularization term. Various regularization terms have been considered, including the least absolute shrinkage and selection operator (Lasso; Tibshirani, 1996, [30]), the smoothly clipped absolute deviation (SCAD; Fan and Li, 2001, [3]), the adaptive Lasso (Zou, 2006, [40]), the minimax concave penalty (MCP; Zhang, 2010, [38]), the truncated l_1 -penalty (TLP; Shen et al., 2012, [23]), and so on. These algorithms have also been extended to the nonparametric models to relax the linear model assumption. For example, under the additive model assumption, a number of variable selection algorithms have been developed (Shively et al., 1999, [24]; Huang et al., 2010, [12]), where each component function depends on one variable only. Further, a component selection and smoothing operator algorithm is proposed in Lin and Zhang (2006, [17]) to allow higher-order interaction components in the additive model, at the cost of increasing computational cost. These nonparametric variable selection algorithms, although more flexible than the linear model, still require some explicit working model sets.

More recently, attempts have been made to develop model-free variable selection algorithms to circumvent the dependency on restrictive model assumptions. Particularly, variable selection is formulated in a dimension reduction framework in Li et al. (2005, [14]) and Bondell and Li (2009, [2]) via searching for the sparse basis of the central dimension reduction space. Fukumizu and Leng (2014, [8]) developed a gradient-based dimension reduction algorithm that can be extended to model-free variable selection. A novel measurement-error-model-based variable selection algorithm is developed in Stefanski et al. (2014, [27]) and Wu and Stefanski (2015, [36]) for nonparametric kernel regression models, and a gradient learning framework is proposed in Yang et al. (2016, [37]) to conduct variable selection in a flexible RKHS (Wahba, 1998, [32]). These algorithms are attractive due to its flexibility and asymptotic sparsistency, yet their high computational cost remains as one of the main obstacles.

Another popular line of research on high-dimensional data is variable screening, which screens out uninformative variables by examining the marginal relationship between the response and each variable. The marginal relationship can be measured by various criteria, including the Pearson's correlation (Fan and Lv, 2008, [4]), the empirical functional norm (Fan et al., 2011, [5]), the distance correlation (Li et al., 2012, [15]), and a quantile-adaptive procedure (He et al., 2013, [11]). All these methods are computationally very efficient, and attain the sure screening property, meaning that all the truly informative variables are retained after screening with probability tending to one. This is a desirable property, yet slightly weaker than the asymptotic sparsistency in variable selection. Another potential weakness of the

marginal screening methods is that they may ignore those marginally unimportant but jointly important variables (He et al., 2013, [11]). To remedy this limitation, some recent work (Hao et al., 2017, [10]; Kong et al., 2017, [13]; She et al., 2017, [22]) has been done to conduct sure screening for variables with interaction effects.

In this article, we propose a scalable kernel-based variable selection algorithm, which is methodologically flexible, computationally efficient, and able to achieve the asymptotic sparsistency without requiring any explicit model assumption. The algorithm consists of three simple steps, involving kernel-based estimation of the regression function and its gradient functions as well as a hard thresholding. It first fits a kernel ridge regression model in a flexible RKHS to obtain an estimated regression function, then estimates its gradient functions along each variable by taking advantage of the derivative reproducing property (Zhou, 2007, [39]), and finally hard-thresholds the empirical norm of each gradient function to identify the truly informative variables. This algorithm is flexible in that it can be adapted to any RKHS with different kernel functions, to accommodate prior information about the true regression function. Its computational cost is only linear in the data dimension, and thus scalable to analyze dataset with large dimensions. For example, the simulated examples with $O(10^5)$ variables can be efficiently analyzed on a standard multi-core PC. More importantly, asymptotic sparsistency can be established for the proposed algorithm without requiring any explicit model assumptions. It is clear that the proposed algorithm is advantageous than the existing algorithms, as it achieves methodological flexibility, numerical efficiency and asymptotic sparsistency. To our knowledge, this algorithm is the first one that can achieve these three desirable properties at the same time.

The rest of the article is organized as follows. In Section 2, we present the proposed general kernel-based variable selection algorithm as well as its computational scheme. In Section 3, the asymptotic sparsistency of the proposed algorithm is established. In Section 4, we give a detailed analysis of the asymptotic behavior of the proposed algorithm under linear model, and compare it against some existing theoretical results. In Section 5, the proposed algorithm is extended to select truly informative interaction terms. The numerical experiments on the simulated and real examples are contained in Section 6, and all the technical proofs are given in Section 7.

2. Proposed algorithm. Suppose a random sample $\mathcal{Z}^n = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$ are independent copies of $\mathcal{Z} = (\mathbf{x}, y)$, drawn from some unknown distribution $\rho_{\mathbf{x}, y}$ with $\mathbf{x} = (x^1, \dots, x^p)^T \in \mathcal{X}$ supported on a compact metric space and $y \in \mathcal{R}$. Consider a general regression setting,

$$y = f^*(\mathbf{x}) + \epsilon,$$

where ϵ is a random error with $\mathbb{E}(\epsilon|\mathbf{x}) = 0$ and $\text{Var}(\epsilon|\mathbf{x}) = \sigma^2$, and thus $f^*(\mathbf{x}) = \int y d\rho_{y|\mathbf{x}}$ with $\rho_{y|\mathbf{x}}$ denoting the conditional distribution of y given \mathbf{x} . It is also assumed that $f^* \in \mathcal{H}_K$, where \mathcal{H}_K is a RKHS induced by some pre-specified kernel function $K(\cdot, \cdot)$. For each $\mathbf{x} \in \mathcal{X}$, denote $K_{\mathbf{x}} = K(\mathbf{x}, \cdot) \in \mathcal{H}_K$, and the reproducing property of RKHS implies that $\langle f, K_{\mathbf{x}} \rangle_K = f(\mathbf{x})$ for any $f \in \mathcal{H}_K$, where $\langle \cdot, \cdot \rangle_K$ is the inner product in \mathcal{H}_K . Examples of RKHS include the standard Euclidean space, the Sobolev space, the Besov space, as well as some domain-specific spaces for various applications (Genton, 2001, [9]).

In sparse modeling, it is generally believed that $f^*(\mathbf{x})$ only depends on a small number of variables, while others are uninformative. Unlike model-based settings, variable selection for a general regression model is challenging due to the lack of explicit regression parameters. Here we measure the importance of variables in a regression function by examining the corresponding gradient functions. It is crucial to observe that if a variable x^l is deemed uninformative, the corresponding gradient function $g_l^*(\mathbf{x}) = \partial f^*(\mathbf{x})/\partial x^l$ should be exactly zero almost surely. Thus the true active set can be defined as

$$\mathcal{A}^* = \{l : \|g_l^*\|_2^2 > 0\},$$

where $\|g_l^*\|_2^2 = \int (g_l^*(\mathbf{x}))^2 d\rho_{\mathbf{x}}$ with the marginal distribution $\rho_{\mathbf{x}}$.

The proposed general variable selection algorithm is presented in Algorithm 1.

Algorithm 1: General variable selection algorithm

- Step 1: Obtain an estimate \hat{f} in a smooth RKHS based on the given sample \mathcal{Z}^n ;
 Step 2: Compute $\hat{g}_l(\mathbf{x}) = \partial \hat{f}(\mathbf{x})/\partial x^l$ for $l = 1, \dots, p$;
 Step 3: Identify the informative variables by checking the norm of each \hat{g}_l .
-

We now give details of each step in Algorithm 1. To obtain \hat{f} in Step 1, we employ the kernel ridge regression model,

$$(2.1) \quad \hat{f}(\mathbf{x}) = \operatorname{argmin}_{f \in \mathcal{H}_K} \frac{1}{n} \sum_{i=1}^n (y_i - f(\mathbf{x}_i))^2 + \lambda \|f\|_K^2,$$

where the first term, denoted as $\mathcal{E}_n(f)$, is an empirical version of $\mathcal{E}(f) = \mathbb{E}(y - f(\mathbf{x}))^2$, and $\|f\|_K = \langle f, f \rangle_K^{1/2}$ is the associated RKHS-norm of $f \in \mathcal{H}_K$. By the representer theorem (Wahba, 1998, [32]), the minimizer of (2.1) must have the form

$$\hat{f}(\mathbf{x}) = \sum_{i=1}^n \hat{\alpha}_i K(\mathbf{x}_i, \mathbf{x}) = \hat{\boldsymbol{\alpha}}^T \mathbf{K}_n(\mathbf{x}),$$

where $\hat{\alpha} = (\hat{\alpha}_1, \dots, \hat{\alpha}_n)^T$ and $\mathbf{K}_n(\mathbf{x}) = (K(\mathbf{x}_1, \mathbf{x}), \dots, K(\mathbf{x}_n, \mathbf{x}))^T$. Then the optimization task in (2.1) can be solved analytically, with

$$(2.2) \quad \hat{\alpha} = (\mathbf{K}^2 + n\lambda \mathbf{K})^+ \mathbf{K} \mathbf{y},$$

where $\mathbf{K} = (K(\mathbf{x}_i, \mathbf{x}_j))_{i,j=1}^n$, and $^+$ denotes the Moore-Penrose generalized inverse of a matrix. When \mathbf{K} is invertible, (2.2) simplifies to $\hat{\alpha} = (\mathbf{K} + n\lambda \mathbf{I})^{-1} \mathbf{y}$.

Next, to obtain \hat{g}_l in Step 2, it follows from Lemma 1 in the Appendix that for any $f \in \mathcal{H}_K$,

$$g_l(\mathbf{x}) = \frac{\partial f(\mathbf{x})}{\partial x^l} = \langle f, \partial_l K_{\mathbf{x}} \rangle_K \leq \|\partial_l K_{\mathbf{x}}\|_K \|f\|_K,$$

where $\partial_l K_{\mathbf{x}} = \frac{\partial K(\mathbf{x}, \cdot)}{\partial x^l}$. This implies that the gradient function of any $f \in \mathcal{H}_K$ can be bounded by its K -norm up to some constant. In other words, if we want to estimate $g_l^*(\mathbf{x})$ within the smooth RKHS, it suffices to estimate f^* itself without loss of information. Consequently, if \hat{f} is obtained in Step 1, $g_l^*(\mathbf{x})$ can be estimated as $\hat{g}_l(\mathbf{x}) = \hat{\alpha}^T \partial_l \mathbf{K}_n(\mathbf{x})$ for each l .

In Step 3, it is difficult to evaluate $\|\hat{g}_l\|_2^2$ directly, as $\rho_{\mathbf{x}}$ is usually unknown in practice. We then adopt the empirical norm of \hat{g}_l as a practical measure,

$$\|\hat{g}_l\|_n^2 = \frac{1}{n} \sum_{i=1}^n (\hat{g}_l(\mathbf{x}_i))^2 = \frac{1}{n} \sum_{i=1}^n (\hat{\alpha}^T \partial_l \mathbf{K}_n(\mathbf{x}_i))^2.$$

The estimated active set can be set as $\hat{\mathcal{A}}_{v_n} = \{l : \|\hat{g}_l\|_n^2 > v_n\}$ for some pre-specified v_n . It is crucial to choose a proper v_n for optimal selection performance, which shall be determined by some data-adaptive tuning schemes such as the stability-based criterion in Sun et al. (2013, [29]).

The proposed Algorithm 1 is general in that it can be adapted to any smooth RKHS with different kernel functions, where the choice of kernel function depends on prior knowledge about f^* . A direct application of the proposed algorithm under linear model will be discussed in Sections 4. The proposed algorithm is also computationally efficient, whose computational cost is about $O(n^3 + n^2p)$. The complexity $O(n^3)$ comes from inverting an $n \times n$ matrix in (2.2), and the complexity $O(n^2p)$ comes from calculating $\|\hat{g}_l\|_n^2$ for $l = 1, \dots, p$. Furthermore, if low-rank approximation of the kernel matrix is employed, such as Nystrom sampling (Williams and Seeger, 2001, [35]) or incomplete Cholesky decomposition (Fine and Scheinberg, 2002, [6]), the complexity can be reduced to $O(n^2d + ndp)$, where d is the preserved rank. The algorithm can also be parallelized to further speed up the computation, which is particularly attractive in the large- p -small- n scenarios.

3. Asymptotic sparsistency. Now we establish the sparsistency of the proposed algorithm. First, we introduce an integral operator $L_K : \mathcal{L}^2(\mathcal{X}, \rho_{\mathbf{x}}) \rightarrow \mathcal{L}^2(\mathcal{X}, \rho_{\mathbf{x}})$, given by

$$L_K(f)(\mathbf{x}) = \int K(\mathbf{x}, \mathbf{u})f(\mathbf{u})d\rho_{\mathbf{x}}(\mathbf{u}),$$

for any $f \in \mathcal{L}^2(\mathcal{X}, \rho_{\mathbf{x}}) = \{f : \int f^2(\mathbf{x})d\rho_{\mathbf{x}} < \infty\}$. Note that if the corresponding RKHS is separable, by the spectral theorem we have

$$L_K f = \sum_j \mu_j \langle f, e_j \rangle_2 e_j,$$

where $\{e_j\}$ is an orthonormal basis of $\mathcal{L}^2(\mathcal{X}, \rho_{\mathbf{x}})$, μ_j is the eigenvalue of the integral operator L_K , and $\langle \cdot, \cdot \rangle_2$ is the inner product in $\mathcal{L}^2(\mathcal{X}, \rho_{\mathbf{x}})$.

We rewrite λ and p as λ_n and p_n to emphasize their dependency on n , and thus p_n is allowed to diverge with n . The cardinality of the true active set \mathcal{A}^* is denoted as $|\mathcal{A}^*| = p_0 \ll p_n$. The following technical assumptions are made.

Assumption 1: Suppose that f^* is in the range of the r -th power of L_K , denoted as $L_K^r f^*$, for some positive constant $r \in (1/2, 1]$.

Note that the operator L_K on $\mathcal{L}^2(\mathcal{X}, \rho_{\mathbf{x}})$ is self-adjoint and semi-positive definite, and thus its fractional operator L_K^r is well defined. Furthermore, the range of L_K^r is contained in \mathcal{H}_K if $r \geq 1/2$ (Smale and Zhou, 2007, [26]), and thus Assumption 1 implies that there exists some function $h \in \mathcal{L}^2(\mathcal{X}, \rho_{\mathbf{x}})$ such that $f^* = L_K^r h = \sum_j \mu_j^r \langle h, e_j \rangle_2 e_j \in \mathcal{H}_K$, ensuring strong estimation consistency under the RKHS-norm. Similar assumptions are also imposed in Mendelson and Neeman (2010, [18]).

Assumption 2: There exist some constants κ_1 and κ_2 such that $\sup_{\mathbf{x} \in \mathcal{X}} \|K_{\mathbf{x}}\|_K \leq \kappa_1$, and $\sup_{\mathbf{x} \in \mathcal{X}} \|\partial_l K_{\mathbf{x}}\|_K \leq \kappa_2$, for any $l = 1, \dots, p_n$.

Assumption 2 assumes the boundedness of the kernel function and its gradient functions, which is satisfied by many popular kernels, including the Gaussian kernel and the Sobolev kernel (Smale and Zhou, 2007, [26]; Rosasco et al., 2013, [20]; Yang et al., 2016, [37]).

THEOREM 1. Suppose Assumptions 1 and 2 are satisfied. For any $\delta_n \geq 4(\sigma^2 + \|f^*\|_2^2)(\log n)^{-2}$, with probability at least $1 - \delta_n/2$, there holds

$$(3.1) \quad \|\hat{f} - f^*\|_K \leq 2 \log \frac{8}{\delta_n} \left(3\kappa_1 \lambda_n^{-1} n^{-1/2} \log n + \lambda_n^{r-1/2} \|L_K^{-r} f^*\|_2 \right).$$

为什么速度和变量维数 p_n 无关呢？这个结果很奇怪

Additionally, let $\lambda_n = n^{-\frac{1}{2r+1}} (\log n)^{\frac{2}{2r+1}}$, then with probability at least $1 - \delta_n$, there holds

$$(3.2) \quad \max_{1 \leq l \leq p_n} \|\hat{g}_l\|_n^2 - \|g_l^*\|_2^2 \leq c_1 \log \left(\frac{8p_n}{\delta_n} \right) n^{-\frac{2r-1}{2(2r+1)}} (\log n)^{\frac{2r-1}{2r+1}},$$

where $c_1 = 4 \max\{\kappa_2^2, \kappa_2^2 \|f^*\|_K, \|f^*\|_K^2\} \max\{3\kappa_1, 2\sqrt{2}\kappa_2^2, \|L_K^{-r} f^*\|_2\}$.

Theorem 1 establishes the convergence rate of the difference between the estimated regression function and the true regression function in terms of the RKHS-norm. In the upper bound in (3.1), the quantity $\|L_K^{-r} f^*\|_2$ may depend on p_0 through f^* , yet such dependency is generally difficult to quantify explicitly (Fukumizu and Leng, 2014, [8]). Theorem 1 also shows that $\|\hat{g}_l\|_n^2$ converges to $\|g_l^*\|_2^2$ with high probability, which is crucial to establish the asymptotic sparsistency. Note that the constant c_1 is spelled out precisely for the subsequent analysis of the asymptotic sparsistency and its dependency on f^* .

The convergence rate in Theorem 1 can be strengthened to obtain an optimal strong convergence rate in a minimax sense as in Fischer and Steinwart (2018, [7]). Yet it further requires the random error ϵ follows a sub-Gaussian distribution and the decay rate of L_K 's eigenvalues fulfills a polynomial upper bound of order $\tau \in (0, 1)$; that is, $\mu_j \leq Cj^{-1/\tau}$ for some positive constant C . Then the convergence rate can be obtained as $O_p(n^{-\frac{2r-1}{2r+\tau}} \log p_n)$, which has a weak and implicit dependency on the ambient dimension. In other words, even though most of the RKHSs are infinite-dimensional spaces, their functional complexity can be well controlled and the curse of dimensionality can be largely avoided without specifying model structures.

Assumption 3: There exist constants $c_2 > 0$ and $\xi_1 > \frac{2r-1}{2r+1}$ such that $\min_{l \in \mathcal{A}^*} \|g_l^*\|_2^2 > c_2 n^{-\frac{2r-1}{2(2r+1)}} (\log n)^{\xi_1} \log p_n$.

Assumption 3 requires the true gradient function contains sufficient information about the truly informative variables. Unlike most nonparametric models, we measure the significance of each gradient function to distinguish the informative and uninformative variables without any explicit model specification. Now we establish the asymptotic sparsistency of the proposed variable selection algorithm.

THEOREM 2. Suppose the assumptions of Theorem 1 and Assumption 3 are satisfied. Let $v_n = \frac{c_2}{2} n^{-\frac{2r-1}{2(2r+1)}} (\log n)^{\xi_1} \log p_n$, then we have

$$P(\hat{\mathcal{A}}_{v_n} = \mathcal{A}^*) \rightarrow 1, \text{ as } n \rightarrow \infty.$$

Theorem 2 shows that the selected active set can exactly recover the true active set with probability tending to 1. This result is particularly interesting given the fact that it is established for any RKHS with different kernel functions. A direct application of the proposed algorithm and Theorem 2 is to conduct nonparametric variable selection with sparsistency (Li et al., 2012, [15]; He et al., 2013, [11]; Yang et al., 2016, [37]). If no prior knowledge about the true regression function is available, the proposed algorithm can be applied with a RKHS associated with the

Gaussian kernel. Asymptotic sparsistency can be established following Theorem 2 provided that f^* is contained in the RKHS associated with the Gaussian kernel. This RKHS is fairly large as the Gaussian kernel is known to be universal in the sense that any continuous function can be well approximated by some function in the induced RKHS under the infinity norm (Steinwart, 2005, [28]).

4. A special case with linear kernel. Variable selection for linear model is of great interest in statistical literature due to its simplicity and interpretability. Particularly, the true regression function is assumed to be a linear function, $f^*(\mathbf{x}) = \beta_0 + \mathbf{x}^T \boldsymbol{\beta}^*$, and the true active set is defined as $\mathcal{A}^* = \{l : \beta_l^* \neq 0\}$. We also centralize the response and each variable, so that β_0 can be discarded from the linear model for simplicity.

With the linear kernel $K(\mathbf{x}, \mathbf{u}) = \mathbf{x}^T \mathbf{u}$, the induced RKHS becomes the linear functional space and contains the true regression function f^* . The proposed Algorithm 1 simplifies to first fit a ridge regression model,

$$(4.1) \quad \hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^n (y_i - \boldsymbol{\beta}^T \mathbf{x}_i)^2 + \lambda_n \|\boldsymbol{\beta}\|^2,$$

where the RKHS-norm induced by the linear kernel reduces to $\|\boldsymbol{\beta}\|^2$. Then the estimated active set is defined as $\hat{\mathcal{A}}_{v_n} = \{l : |\hat{\beta}_l| > v_n\}$ for some pre-specified thresholding value v_n . This algorithm simply truncates the ridge regression coefficients for variable selection, and its asymptotic properties have been considered in Shao and Deng (2012, [21]) and Wang and Leng (2016, [34]).

We now apply the general results in Section 3 to establish the sparsistency of the proposed algorithm under the linear model. We first scale the original data as $\tilde{\mathbf{y}} = p_n^{-1} \mathbf{y}$ and $\tilde{\mathbf{x}} = p_n^{-1/2} \mathbf{x}$, and let \mathcal{H}_K be the RKHS induced by a scaled linear kernel $K(\tilde{\mathbf{x}}, \tilde{\mathbf{u}}) = \tilde{\mathbf{x}}^T \tilde{\mathbf{u}} = p_n^{-1} \mathbf{x}^T \mathbf{u}$. The true regression model can be rewritten as $f^*(\tilde{\mathbf{x}}) = \tilde{\mathbf{x}}^T \tilde{\boldsymbol{\beta}}^*$ with $\tilde{\boldsymbol{\beta}}^* = p_n^{-1/2} \boldsymbol{\beta}^*$. With the scaled data, the ridge regression formula in (4.1) becomes

$$(4.2) \quad \check{\boldsymbol{\beta}} = \underset{\tilde{\boldsymbol{\beta}}}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^n (\tilde{y}_i - \tilde{\boldsymbol{\beta}}^T \tilde{\mathbf{x}}_i)^2 + p_n^{-1} \lambda_n \|\tilde{\boldsymbol{\beta}}\|^2,$$

where $\check{\boldsymbol{\beta}} = p_n^{-1/2} \hat{\boldsymbol{\beta}}$. By the representer theorem, the solution of (4.2) is

$$\check{\boldsymbol{\beta}} = \tilde{\mathbf{X}}^T (\tilde{\mathbf{X}} \tilde{\mathbf{X}}^T + n p_n^{-1} \lambda_n \mathbf{I}_n)^{-1} \tilde{\mathbf{y}},$$

where $\tilde{\mathbf{X}} = (\tilde{\mathbf{x}}_1, \dots, \tilde{\mathbf{x}}_n)^T$ and $\tilde{\mathbf{y}} = (\tilde{y}_1, \dots, \tilde{y}_n)^T$. This is equivalent to the standard formula for the ridge regression $\check{\boldsymbol{\beta}} = (\tilde{\mathbf{X}}^T \tilde{\mathbf{X}} + n p_n^{-1} \lambda_n \mathbf{I})^{-1} \tilde{\mathbf{X}}^T \tilde{\mathbf{y}}$ according to the Sherman-Morrison-Woodbury formula (Wang and Leng, 2015).

Assumption 4: There exist some positive constants τ_1 and τ_2 such that the smallest eigenvalue of $E(\mathbf{x} \mathbf{x}^T)$, $\lambda_{\min}(E(\mathbf{x} \mathbf{x}^T)) = O(p_n n^{-\tau_1})$, and $\|\beta^*\|^2 = O(n^{\tau_2})$.

Assumption 4 implies that $E(\mathbf{x} \mathbf{x}^T)$ is invertible, and that Assumption 1 is satisfied for the scaled linear kernel with $r = 1$. Assumption 2 is also satisfied due to the fact that $\|\tilde{\mathbf{x}}\|^2 = p_n^{-1} \mathbf{x}^T \mathbf{x}$ is bounded on a compact support \mathcal{X} . A similar assumption is made in Shao and Deng (2012, [21]), assuming the decay order of the smallest eigenvalue of $n^{-1} \mathbf{X}^T \mathbf{X}$. We now establish the estimation consistency for $\tilde{\beta}$.

COROLLARY 1. *Suppose Assumption 4 is met. For any $\delta_n \geq 4(\sigma^2 p_n^{-1} + \|\tilde{\beta}^*\|^2)(\log n)^{-2}$, there exists some positive constant c_3 such that, with probability at least $1 - \delta_n$, there holds*

$$\|\tilde{\beta} - \tilde{\beta}^*\| \leq c_3 \log \frac{4}{\delta_n} (p_n \lambda_n^{-1} n^{-1/2} \log n + p_n^{-1} \lambda_n^{1/2} n^{\frac{\tau_1 + \tau_2}{2}}).$$

Specifically, let $\lambda_n = O(p_n^{4/3} n^{-(1+\tau_1+\tau_2)/3} (\log n)^{2/3})$, then with probability at least $1 - \delta_n$, there holds

$$(4.3) \quad \|\tilde{\beta} - \tilde{\beta}^*\| \leq 2c_3 \log \left(\frac{4}{\delta_n} \right) p_n^{-1/3} n^{-\frac{1-2(\tau_1+\tau_2)}{6}} (\log n)^{1/3},$$

p_n 越大, 收敛速度越大, 这和常识违背了

Corollary 1 is a directly application of Theorem 1 under the linear kernel. An additional assumption is made to establish the sparsistency.

Assumption 5: There exist some positive constants c_4 and $\xi_2 > 1/3$ such that $\min_{l \in \mathcal{A}^*} |\beta_l^*| > c_4 p_n^{1/6} n^{-\frac{1-2(\tau_1+\tau_2)}{6}} (\log n)^{\xi_2}$.

Assumption 5 is similar to Assumption 3, and requires the true regression coefficients contains sufficient information about the truly informative variables in the linear model. Similar assumptions are also assumed in Shao and Deng (2012, [21]) and Wang and Leng (2016, [34]).

COROLLARY 2. *Suppose the assumptions of Corollary 1 and Assumption 5 are met. Let $v_n = \frac{c_4}{2} p_n^{1/6} n^{-\frac{1-2(\tau_1+\tau_2)}{6}} (\log n)^{\xi_2}$, then we have*

$$P(\hat{\mathcal{A}}_{v_n} = \mathcal{A}^*) \rightarrow 1, \text{ as } n \rightarrow \infty.$$

It is interesting to point out that Corollary 2 holds when p_n diverges at order $o(\min\{n^{\tau_1}, n^{1-2(\tau_1+\tau_2)} (\log n)^{-6\xi_2}, n^{\frac{1+\tau_1+\tau_2}{4}} (\log n)^{-1/2}\})$. Particularly, when $\tau_1 = 1/3$ and $\tau_2 = 0$, p_n can diverge at a polynomial rate $o(n^{1/3})$. This result is similar as that established in Shao and Deng (2012) under the assumption of a finite second moment of ϵ . In literature, it is possible to allow p_n to diverge at an

exponential rate of n under certain distributional assumptions on ϵ , such as a sub-Gaussian distribution in Shao and Deng (2012, [21]) or a spherically symmetric distribution with q -exponential tails in Wang and Leng (2016, [34]). Such distributional assumptions are not necessary in establishing the sparsistency in Corollary 2.

5. Extension to interaction selection. Interaction selection has become popular in recent literature, including Bien et al. (2013, [1]), Lim and Hastie (2015, [16]), Hao et al. (2018, [10]), She et al. (2018, [22]) and Kong et al. (2017, [13]). In these work, the true regression function is assumed to be a quadratic model $f^*(\mathbf{x}) = \sum_{l=1}^{p_0} \beta_l^* x^l + \sum_{l,k=1}^{p_0} \gamma_{lk}^* x^l x^k$. A strong heredity is often assumed, requiring that if an interaction effect $x^l x^k$ is included in the model, then both main effects x^l and x^k must be included as well.

We now extend the proposed algorithm to identify the truly informative interaction terms without assuming the quadratic model. Following the idea in Section 2, the true interaction effects can be defined as those with nonzero second-order gradient function $g_{lk}^*(\mathbf{x}) = \partial^2 f^*(\mathbf{x}) / \partial x^l \partial x^k$. Note that a nonzero second-order gradient function $g_{lk}^*(\mathbf{x})$ implies that both $g_l^*(\mathbf{x})$ and $g_k^*(\mathbf{x})$ are also nonzero, and thus the strong heredity is automatically satisfied. Specifically, given the true active set \mathcal{A}^* , we denote

$$\mathcal{A}_2^* = \{l \in \mathcal{A}^* : \|g_{lk}^*\|_2 > 0, \text{ for some } k \in \mathcal{A}^*\},$$

which contains the variables that contribute to the interaction effects in f^* . Further, let $\mathcal{A}_1^* = \mathcal{A}^* \setminus \mathcal{A}_2^*$, which contains the variables that contribute to the main effects of f^* only. Therefore, the main goal of interaction selection is to estimate both \mathcal{A}_1^* and \mathcal{A}_2^* .

First, let $K(\cdot, \cdot)$ be a forth-order differentiable kernel function, then it follows from Lemma 1 in the Appendix that for any $f \in \mathcal{H}_K$,

$$g_{lk}(\mathbf{x}) = \frac{\partial^2 f(\mathbf{x})}{\partial x^l \partial x^k} = \langle f, \partial_{lk} K_{\mathbf{x}} \rangle_K \leq \|\partial_{lk} K_{\mathbf{x}}\|_K \|f\|_K,$$

where $\partial_{lk} K_{\mathbf{x}} = \frac{\partial^2 K(\mathbf{x}, \cdot)}{\partial x^l \partial x^k}$. Then, given \hat{f} from (2.1), its second-order gradient function is

$$\hat{g}_{lk}(\mathbf{x}) = \frac{\partial^2 \hat{f}(\mathbf{x})}{\partial x^l \partial x^k} = \hat{\boldsymbol{\alpha}}^T \partial_{lk} \mathbf{K}_n(\mathbf{x}),$$

where $\partial_{lk} \mathbf{K}_n(\mathbf{x}) = \frac{\partial \mathbf{K}_n(\mathbf{x})}{\partial x^l \partial x^k}$. Its empirical norm is $\|\hat{g}_{lk}\|_n^2 = \frac{1}{n} \sum_{i=1}^n (\hat{g}_{lk}(\mathbf{x}_i))^2$. With some pre-defined thresholding value v_n^{int} , the estimated \mathcal{A}_1 and \mathcal{A}_2 are set as

$$\hat{\mathcal{A}}_2 = \{l \in \hat{\mathcal{A}} : \|\hat{g}_{lk}\|_n^2 > v_n^{int}, \text{ for some } k \in \hat{\mathcal{A}}\} \text{ and } \hat{\mathcal{A}}_1 = \hat{\mathcal{A}} \setminus \hat{\mathcal{A}}_2,$$

respectively. The following technical assumption is made to establish the interaction selection consistency for the proposed algorithm.

Assumption 6: There exists some constant κ_3 such that $\sup_{\mathbf{x} \in \mathcal{X}} \|\partial_{lk} K_{\mathbf{x}}\|_K \leq \kappa_3$ for any l and k . Also, there exist some positive constants c_5 and $\xi_3 > \frac{2r-1}{2r+1}$ such that $\min_{l,k \in \mathcal{A}_2^*} \|g_{lk}^*\|_2^2 > c_5 n^{-\frac{2r-1}{2(2r+1)}} (\log n)^{\xi_3} \log p_0$.

Assumption 6 can be regarded as an extension of Assumptions 2 and 3 by requiring the boundedness of the second-order gradients of $K_{\mathbf{x}}$, and that the true second-order gradient functions have sufficient information about the interaction effects.

THEOREM 3. Suppose the assumptions of Theorem 2 and Assumption 6 are met. Let $P(\hat{\mathcal{A}} \neq \mathcal{A}^*) = \Delta_n$. For any $\delta_n \geq 4(\sigma^2 + \|f^*\|_2^2)(\log n)^{-2}$, with probability at least $1 - \delta_n - \Delta_n$, there holds

$$\max_{l,k \in \hat{\mathcal{A}}} \left| \|\hat{g}_{lk}\|_n^2 - \|g_{lk}^*\|_2^2 \right| \leq c_6 \log \left(\frac{8p_0^2}{\delta_n} \right) n^{-\frac{2r-1}{2(2r+1)}} (\log n)^{\frac{2r-1}{2r+1}},$$

where $c_6 = 4 \max\{\kappa_3^2, \|f^*\|_K^2, \kappa_3^2 \|f^*\|_K\} \max\{3\kappa_1, 2\sqrt{2}\kappa_3^2, \|L_K^{-r} f^*\|_2\}$

Theorem 3 shows that $\|\hat{g}_{lk}\|_n^2$ converges to $\|g_{lk}^*\|_2^2$ with high probability, which is crucial to establish the interaction selection consistency.

THEOREM 4. Suppose the assumptions of Theorem 3 are met. Let $v_n^{int} = \frac{c_5}{2} n^{-\frac{2r-1}{2(2r+1)}} (\log n)^{\xi_3} \log p_0$, there holds

$$P(\hat{\mathcal{A}}_2 = \mathcal{A}_2^*, \hat{\mathcal{A}}_1 = \mathcal{A}_1^*) \rightarrow 1, \text{ as } n \rightarrow \infty.$$

Theorem 4 shows that the proposed interaction selection algorithm can exactly detect the interaction structure with probability tending to 1. It is clear that the algorithm can be extended to detect higher-order interaction effects, which is of particular interest in some real applications (Ritchie et al., 2001, [19]).

6. Numerical experiments. In this section, the numerical performance of the proposed algorithm is examined, and compared against some existing methods, including the **distance correlation screening** (Li et al., 2012, [15]) and the **quantile-adaptive screening** (He et al, 2013, [11]). As these two methods are designed for screening only, they are also truncated by some thresholding values to conduct **variable selection**. For simplicity, we denote these three methods as GM, DC-t and QaSIS-t, respectively.

In all the simulation examples, no prior knowledge about the true regression function is assumed, and the Gaussian kernel $K(\mathbf{u}, \mathbf{v}) = \exp\{-\|\mathbf{u} - \mathbf{v}\|^2/2\sigma_n^2\}$

is used to induce the RKHS, where σ_n is set as the median of all the pairwise distances among the training sample. The thresholding values for all the methods are selected by the stability-based selection criterion (Sun et al., 2013, [29]). Its key idea is to measure the stability of variable selection by randomly splitting the training sample into two parts and comparing the disagreement between the two estimated active sets. The maximization of the stability criterion is conducted via a grid search, where the grid is set as $\{10^{-3+0.1s} : s = 0, \dots, 60\}$.

6.1. Simulated examples. Two simulated examples are examined under various scenarios.

Example 1: We first generate $x_i = (x_{i1}, \dots, x_{ip})^T$ with $x_{ij} = \frac{W_{ij} + \eta U_i}{1 + \eta}$, where W_{ij} and U_i are independently drawn from $U(-0.5, 0.5)$. The response y_i is generated as $y_i = f(\mathbf{x}_i) + \epsilon_i$, where $f^*(\mathbf{x}) = 6f_1(x_{i1}) + 4f_2(x_{i2})f_3(x_{i3}) + 6f_4(x_{i4}) + 5f_5(x_{i5})$, with $f_1(u) = u$, $f_2(u) = 2u + 1$, $f_3(u) = 2u - 1$, $f_4(u) = 0.1 \sin(\pi u) + 0.2 \cos(\pi u) + 0.3(\sin(\pi u))^2 + 0.4(\cos(\pi u))^3 + 0.5(\sin(\pi u))^3$, $f_5(u) = \sin(\pi u)/(2 - \sin(\pi u))$, and ϵ_i 's are independently drawn from $N(0, 1)$. Clearly, the first 5 variables are truly informative.

Example 2: The generating scheme is similar as Example 1, except that W_{ij} and U_i are independently drawn from $U(0, 1)$ and $f^*(\mathbf{x}) = 20x_1x_2x_3 + 5x_4^2 + 5x_5$. The first 5 variables are truly informative.

For each example, we consider scenarios with $(n, p) = (400, 500), (400, 1000), (500, 10000)$ and $(500, 50000)$. For each scenario, $\eta = 0$ and $\eta = 0.2$ are examined. When $\eta = 0$, the variables are completely independent, whereas when $\eta = 0.2$, correlation structure are added among the variables. Each scenario is replicated 50 times. The averaged performance measures are summarized in Tables 1 and 2, where Size is the averaged number of selected informative variables, TP is the number of truly informative variables selected, FP is the number of truly uninformative variables selected, and C, U, O are the times of correct-fitting, under-fitting, and over-fitting, respectively.

It is evident that GM outperforms the other methods in both examples. In Example 1, GM is able to identify all the truly informative variables in most replications. However, the other two methods tend to miss some truly informative variables, probably due to the interaction effect between x^2 and x^3 . In Example 2, with a three-way interaction term involved in $f^*(\mathbf{x})$, GM is still able to identify all the truly informative variables with high accuracy, but the other two methods tend to underfit by missing some truly informative variables in the interaction term. Note that if we do not threshold DC and QaSIS, they tend to overfit almost in every replication as both screening methods tend to keep a substantial amount of uninformative variables to attain the sure screening property. Furthermore, when the correlation structure with $\eta = 0.2$ is considered, identifying the truly informative

TABLE 1
The averaged performance measures of various methods in Example 1.

(n, p, η)	Method	Size	TP	FP	C	U	O
(400,500,0)	GM	5.10	5.00	0.10	45	0	5
	QaSIS-t	4.48	4.48	0.00	28	22	0
	DC-t	4.82	4.82	0.00	42	8	0
(400,1000,0)	GM	5.16	5.00	0.16	42	0	8
	QaSIS-t	4.24	4.24	0.00	17	33	0
	DC-t	4.68	4.68	0.00	35	15	0
(500,10000,0)	GM	5.00	5.00	0.00	50	0	0
	QaSIS-t	4.28	4.28	0.00	24	26	0
	DC-t	4.68	4.68	0.00	36	14	0
(500,50000,0)	GM	5.08	4.98	0.10	44	1	5
	QaSIS-t	4.08	4.08	0.00	18	32	0
	DC-t	4.48	4.48	0.00	28	22	0
(400,500,0.2)	GM	5.50	5.00	0.50	32	0	18
	QaSIS-t	4.06	4.04	0.02	18	32	0
	DC-t	4.72	4.70	0.02	40	9	1
(400,1000,0.2)	GM	5.44	4.98	0.46	33	1	16
	QaSIS-t	4.00	4.00	0.00	17	33	0
	DC-t	4.64	4.64	0.00	36	14	0
(500,10000,0.2)	GM	5.50	5.00	0.50	34	0	16
	QaSIS-t	4.10	4.10	0.00	22	28	0
	DC-t	4.56	4.56	0.00	35	15	0
(500,50000,0.2)	GM	5.80	4.98	0.82	34	1	15
	QaSIS-t	4.10	4.10	0.00	22	28	0
	DC-t	4.66	4.66	0.00	36	14	0

variables becomes more difficult, yet GM still outperforms the other two competitors in most scenarios.

6.2. Supermarket dataset. We now apply the proposed algorithm to a supermarket dataset in Wang (2009, [33]). The dataset is collected from a major supermarket located in northern China, consisting of daily sale records of $p = 6398$ products on $n = 464$ days. In this dataset, the response is the number of customers on each day, and the variables are the daily sale volumes of each product. The primary interest is to identify the products whose sale volumes are related with the number of customers, and then to design sale strategies based on those products. The dataset is pre-processed so that both the response and predictors have zero mean and unit variance.

In addition to GM, DC-t and QaSIS-t, we also include the original DC and QaSIS without thresholding, which keep the first $\lceil n / \log n \rceil$ variables to assure the sure screening property. As the truly informative variables are unknown for the supermarket dataset, we report the prediction performance of each method. Specifically, the supermarket dataset is randomly split into two parts, with 164 observations for testing and the remaining for training. We first apply each method to the full dataset to select the informative variables, and then refit a kernel ridge regression

TABLE 2
The averaged performance measures of various methods in Example 2.

(n, p, η)	Method	Size	TP	FP	C	U	O
(400,500,0)	GM	4.98	4.98	0.00	49	1	0
	QaSIS-t	4.26	4.26	0.00	22	28	0
	DC-t	4.92	4.92	0.00	48	2	0
(400,1000,0)	GM	5.00	5.00	0.00	50	0	0
	QaSIS-t	4.04	4.04	0.0	20	30	0
	DC-t	4.96	4.96	0.00	48	2	0
(500,10000,0)	GM	4.96	4.96	0.00	48	2	0
	QaSIS-t	3.82	3.82	0.00	13	37	0
	DC-t	4.92	4.92	0.00	46	4	0
(500,50000,0)	GM	4.94	4.94	0.00	47	3	0
	QaSIS-t	3.04	3.04	0.00	8	42	0
	DC-t	4.66	4.66	0.00	38	12	0
(400,500,0.2)	GM	5.00	5.00	0.00	50	0	0
	QaSIS-t	4.16	4.14	0.02	20	30	0
	DC-t	4.92	4.92	0.00	47	3	0
(400,1000,0.2)	GM	5.00	5.00	0.00	50	0	0
	QaSIS-t	4.24	4.24	0.00	25	25	0
	DC-t	4.98	4.98	0.00	49	1	0
(500,10000,0.2)	GM	4.98	4.98	0.00	49	1	0
	QaSIS-t	3.86	3.86	0.00	17	33	0
	DC-t	4.90	4.90	0.00	46	0	0
(500,50000,0.2)	GM	4.94	4.94	0.00	47	3	0
	QaSIS-t	3.16	3.16	0.00	7	43	0
	DC-t	4.92	4.92	0.00	46	4	0

model with the selected variables for each method on the training set. The prediction performance of each ridge regression model is measured on the testing set. The procedure is replicated 1000 times, and the number of selected variables and the averaged prediction errors are summarized in Table 3.

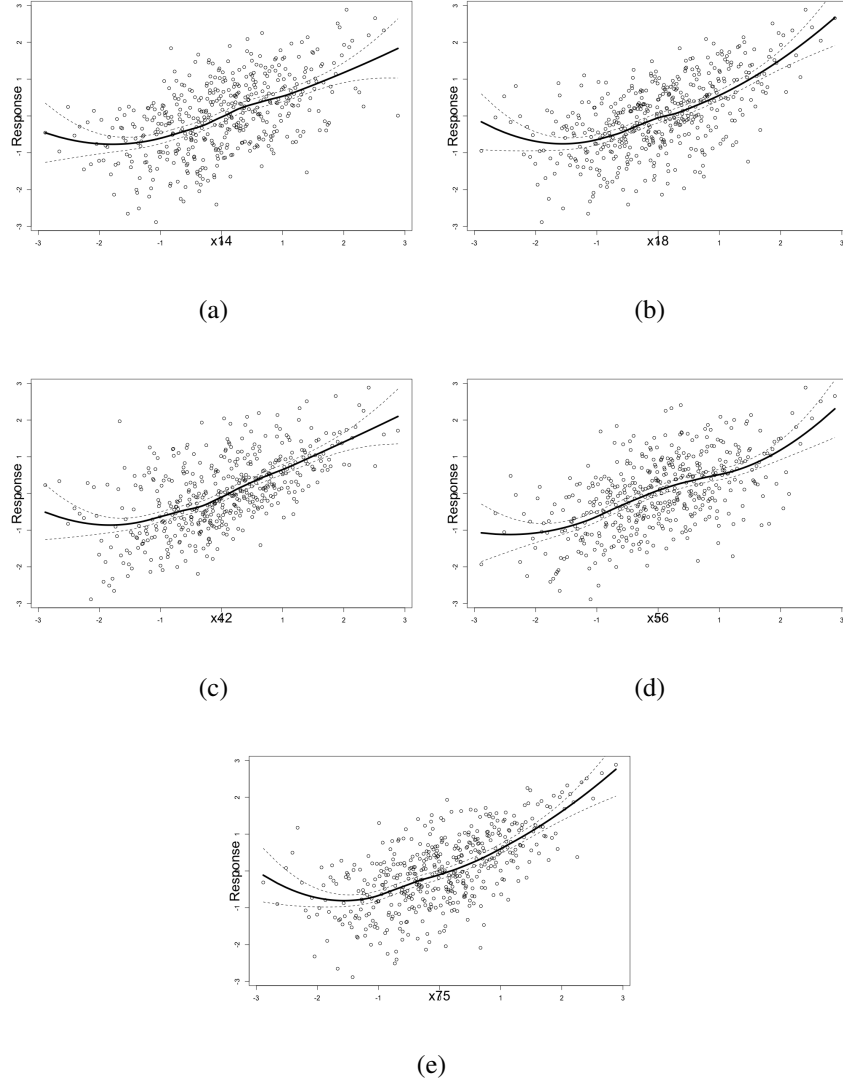
TABLE 3
The number of selected variables as well as the corresponding averaged prediction errors by various methods in the supermarket dataset.

Dataset	Method	Size	Testing error (Std)
	GM	10	0.3688 (0.0016)
	QaSIS-t	7	0.3740 (0.0016)
	DC-t	7	0.3882 (0.0016)
	QaSIS	75	0.3819 (0.0016)
	DC	75	0.3982 (0.0017)

As Table 3 shows, GM selects 10 variables, whereas DC-t and QaSIS-t select 7 variables and DC and QaSIS select 75 variables. The average prediction error of GM is smaller than that of the other four methods, implying that DC-t and QaSIS-t may miss some truly informative variables that deteriorate their prediction accuracy, and DC and QaSIS may include too many noise variables. Precisely, among the 10 selected variables by GM, X^{14} , X^{18} , X^{42} , X^{56} and X^{75} are missed by both

DC-t and QaSIS-t. The scatter plots of the response against these five variables are presented in Figure 1.

Fig 1: The scatter plots of the response against a number of selected variables by GM in the supermarket dataset. The solid lines are the fitted curve by local smoothing, and the dashed lines are the fitted mean plus or minus one standard deviation.



It is evident that the response and these variables have showed some clear relationship, which supports the advantage of GM in identifying the truly informative variables.

7. Lemmas and technical proofs. To be self-contained, we first give a special case of Theorem 1 in Zhou (2007, [39]) as a lemma on the smooth RKHS below, which plays an important role for the subsequent analysis. Its proof follows directly from that of Theorem 1 in Zhou (2007, [39]) and thus omitted here.

LEMMA 1. *Let $K : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ be a Mercer kernel such that $K \in C^4(\mathcal{X} \times \mathcal{X})$, where C^4 is a class of functions whose fourth derivative is continuous. Then the following statements hold:*

- (a) *For any $\mathbf{x} \in \mathcal{X}$, $\partial_l K_{\mathbf{x}}, \partial_{lk} K_{\mathbf{x}} \in \mathcal{H}_K$, for any $l, k = 1, \dots, p_n$.*
- (b) *A derivative reproducing property holds true; that is, for any $f \in \mathcal{H}_K$,*

$$\partial_l f(\mathbf{x}) = \langle f, \partial_l K_{\mathbf{x}} \rangle_K, \quad \text{and } \partial_{lk} f(\mathbf{x}) = \langle f, \partial_{lk} K_{\mathbf{x}} \rangle_K.$$

We now start the proof of Theorem 1 with a proposition, which is crucial to establish the estimation consistency of the proposed algorithm.

PROPOSITION 1. *Suppose Assumption 2 is met and $|y| \leq M_n$. Let \tilde{f} be the minimizer of $\mathcal{E}_{\lambda_n}(f) = E(y - f(\mathbf{x}))^2 + \lambda_n \|f\|_K^2$ in \mathcal{H}_K . Then for any $\delta_n \in (0, 1)$, with probability at least $1 - \delta_n$, there holds*

$$\|\hat{f} - \tilde{f}\|_K \leq \frac{6\kappa_1 M_n}{\lambda_n n^{1/2}} \log \frac{2}{\delta_n}.$$

Proof of Proposition 1: Define the sample operators $S_{\mathbf{x}} : \mathcal{H}_K \rightarrow \mathcal{R}^n$ and $S_{\mathbf{x}}^T : \mathcal{R}^n \rightarrow \mathcal{H}_K$ as

$$S_{\mathbf{x}}(f) = (f(\mathbf{x}_1), \dots, f(\mathbf{x}_n))^T \quad \text{and} \quad S_{\mathbf{x}}^T \mathbf{c} = \sum_{i=1}^n c_i K_{\mathbf{x}_i}.$$

Then solving (2.1) is equivalent to solve

$$\hat{f} = \operatorname{argmin}_{f \in \mathcal{H}_K} \frac{1}{n} \mathbf{y}^T \mathbf{y} - \frac{2}{n} \langle f, S_{\mathbf{x}}^T \mathbf{y} \rangle_K + \frac{1}{n} \langle f, S_{\mathbf{x}}^T S_{\mathbf{x}} f \rangle_K + \lambda \langle f, f \rangle_K,$$

where $\mathbf{y} = (y_1, \dots, y_n)^T$, and hence that

$$\hat{f} = \left(\frac{1}{n} S_{\mathbf{x}}^T S_{\mathbf{x}} + \lambda_n I \right)^{-1} \frac{1}{n} S_{\mathbf{x}}^T \mathbf{y}.$$

Similarly, the minimizer of $\mathcal{E}_{\lambda_n}(f)$ in \mathcal{H}_K must have the form

$$\tilde{f} = (L_K + \lambda_n I)^{-1} L_K f^*.$$

Therefore, we have

$$\begin{aligned} \hat{f} - \tilde{f} &= \left(\frac{1}{n} \mathcal{S}_{\mathbf{x}}^T \mathcal{S}_{\mathbf{x}} + \lambda_n \mathbf{I} \right)^{-1} \left(\frac{1}{n} \mathcal{S}_{\mathbf{x}}^T \mathbf{y} - \frac{1}{n} \mathcal{S}_{\mathbf{x}}^T \mathcal{S}_{\mathbf{x}} \tilde{f} - \lambda_n \tilde{f} \right) \\ &= \left(\frac{1}{n} \mathcal{S}_{\mathbf{x}}^T \mathcal{S}_{\mathbf{x}} + \lambda_n \mathbf{I} \right)^{-1} \left(\frac{1}{n} \sum_{i=1}^n (y_i - \tilde{f}(\mathbf{x}_i)) K_{\mathbf{x}_i} - L_K(f^* - \tilde{f}) \right), \end{aligned}$$

and its RKHS-norm can be upper bounded as

$$\|\hat{f} - \tilde{f}\|_K \leq \lambda_n^{-1} \left\| \frac{1}{n} \sum_{i=1}^n (y_i - \tilde{f}(\mathbf{x}_i)) K_{\mathbf{x}_i} - L_K(f^* - \tilde{f}) \right\|_K = \Delta_1.$$

To bound Δ_1 , denote $\xi = (y - \tilde{f}(\mathbf{x})) K_{\mathbf{x}}$, and it follows from Assumption 2 and direct calculation that

$$\begin{aligned} E\xi &= L_K(f^* - \tilde{f}), \quad \|\xi\|_K \leq \kappa_1(M_n + \|\tilde{f}\|_{\infty}), \\ E(\|\xi\|_K^2) &\leq \kappa_1^2 \int (y - \tilde{f}(x))^2 d\rho_{\mathbf{x},y}. \end{aligned}$$

By Lemma 2 of Smale and Zhou (2005, [25]) and Assumption 2, with probability at least $1 - \delta_n$, there holds

$$\begin{aligned} \Delta_1 &\leq 2n^{-1} \kappa_1 \log \frac{2}{\delta_n} (M_n + \|\tilde{f}\|_{\infty}) + \\ &\quad n^{-1/2} \kappa_1 \left(2 \log \frac{2}{\delta_n} \right)^{1/2} \left(\int (y - \tilde{f}(\mathbf{x}))^2 d\rho_{\mathbf{x},y} \right)^{1/2}. \end{aligned}$$

For $\|\tilde{f}\|_{\infty}$, by the definition of \tilde{f} , we have

$$(7.1) \quad \|\tilde{f} - f^*\|_2^2 + \lambda_n \|\tilde{f}\|_K^2 \leq \|0 - f^*\|_2^2 + \lambda_n \|0\|_K^2 \leq \|f^*\|_2^2,$$

where $\|f^*\|_2^2$ is a bounded quantity. Hence, there holds

$$(7.2) \quad \|\tilde{f}\|_{\infty} \leq \kappa_1 \|\tilde{f}\|_K \leq \kappa_1 \lambda_n^{-1/2} \|f^*\|_2.$$

For $\int (y - \tilde{f}(\mathbf{x}))^2 d\rho_{\mathbf{x},y}$, note that

$$\int (y - f(\mathbf{x}))^2 d\rho_{\mathbf{x},y} - \int (y - f^*(\mathbf{x}))^2 d\rho_{\mathbf{x},y} = \|f - f^*\|_2^2,$$

for any f . Substituting $f = 0$ and $f = \tilde{f}$ yield that

$$(7.3) \quad \int (y - f^*(\mathbf{x}))^2 d\rho_{\mathbf{x},y} + \|f^*\|_2^2 = \int y^2 d\rho_{\mathbf{x},y} \leq M_n^2,$$

$$(7.4) \quad \int (y - \tilde{f}(\mathbf{x}))^2 d\rho_{\mathbf{x},y} = \|\tilde{f} - f^*\|_2^2 + \int (y - f^*(\mathbf{x}))^2 d\rho_{\mathbf{x},y} \leq 2M_n^2,$$

where the last inequality follows from (7.1) and (7.3).

Combing (7.2) and (7.4), we have with probability at least $1 - \delta_n$ that

$$\begin{aligned} \Delta_1 &\leq 2n^{-1}\kappa_1 \log \frac{2}{\delta_n} M_n (1 + \kappa_1 \lambda_n^{-1/2}) + 2n^{-1/2}\kappa_1 \left(\log \frac{2}{\delta_n}\right)^{1/2} M_n \\ &\leq \frac{2\kappa_1 M_n}{n} \log \frac{2}{\delta_n} + \frac{2\kappa_1 M_n}{n^{1/2}} \log \frac{2}{\delta_n} \frac{\kappa_1}{\lambda_n^{1/2} n^{1/2}} + \frac{2\kappa_1 M_n}{n^{1/2}} \left(\log \frac{2}{\delta_n}\right)^{1/2}. \end{aligned}$$

Note that when $\frac{\kappa_1}{\lambda_n^{1/2} n^{1/2}} \leq \left(3 \log \left(\frac{2}{\delta_n}\right)\right)^{-1}$, the above upper bound simplifies to

$$\|\hat{f} - \tilde{f}\|_K \leq \lambda_n^{-1} \Delta_1 \leq \frac{6\kappa_1 M_n}{\lambda_n n^{1/2}} \log \frac{2}{\delta_n}.$$

When $\frac{\kappa_1}{\lambda_n^{1/2} n^{1/2}} > \frac{1}{3} \left(\log \frac{2}{\delta_n}\right)^{-1}$, we have

$$\|\hat{f} - \tilde{f}\|_K \leq \|\hat{f}\|_K + \|\tilde{f}\|_K \leq \frac{2M_n}{\lambda_n^{1/2}} \leq \frac{6\kappa_1 M_n}{\lambda_n n^{1/2}} \log \frac{2}{\delta_n},$$

where the second inequality follows from (7.2), (7.3) and the definition of \hat{f} that $\frac{1}{n} \sum_{i=1}^n (y_i - \hat{f}(\mathbf{x}_i))^2 + \lambda_n \|\hat{f}\|_K^2 \leq \frac{1}{n} \sum_{i=1}^n y_i^2 \leq M_n^2$. The desired inequality then follows immediately. \blacksquare

Proof of Theorem 1. For simplicity, we denote

$$\mathcal{C}_1 = \left\{ \|\hat{f} - f^*\|_K \geq 2 \log \frac{8}{\delta_n} \left(\frac{3\kappa_1 \log n}{n^{1/2} \lambda_n} + \lambda_n^{r-1/2} \|L_K^{-r} f^*\|_2 \right) \right\}.$$

Then $P(\mathcal{C}_1)$ can be decomposed as

$$\begin{aligned} P(\mathcal{C}_1) &= P(\mathcal{C}_1 \cap \{|y| > \log n\}) + P(\mathcal{C}_1 \cap \{|y| \leq \log n\}) \\ &\leq P(|y| > \log n) + P(\mathcal{C}_1 \mid \{|y| \leq \log n\}) = P_1 + P_2. \end{aligned}$$

By Markov's inequality, $P_1 = P(|y| > \log n) \leq (\log n)^{-2} E(|y|^2)$, where $E(y^2) \leq E\epsilon^2 + E|f^*(\mathbf{x})|^2 \leq \sigma^2 + \|f^*\|_2^2$.

For P_2 , note that

$$\|\hat{f} - f^*\|_K \leq \|\hat{f} - \tilde{f}\|_K + \|\tilde{f} - f^*\|_K.$$

We first bound $\|\tilde{f} - f^*\|_K$ following the similar treatment as in Smale and Zhou (2005, [25]). Suppose $\{\mu_i, e_i\}_{i \geq 1}$ are the normalized eigenpairs of the integral operator $L_K : \mathcal{L}^2(\mathcal{X}, \rho_{\mathbf{x}}) \rightarrow \mathcal{L}^2(\mathcal{X}, \rho_{\mathbf{x}})$, we have $L_K^{1/2} e_i = \mu_i^{1/2} e_i \in \mathcal{H}_K$ and $\|\mu_i^{1/2} e_i\|_K = 1$ when $\mu_i > 0$. Thus by Assumption 1, there exists some function $h = \sum_{i \geq 1} \langle h, e_i \rangle_2 e_i \in \mathcal{L}^2(\mathcal{X}, \rho_{\mathbf{x}})$ such that $f^* = L_K^r h = \sum_{i \geq 1} \mu_i^r \langle h, e_i \rangle_2 e_i \in \mathcal{H}_K$. Directly calculation yields to

$$\begin{aligned} \tilde{f} - f^* &= (L_K + \lambda_n \mathbf{I})^{-1} L_K f^* - f^* = (L_K + \lambda_n \mathbf{I})^{-1} (-\lambda_n f^*) \\ &= -\sum_{i \geq 1} \frac{\lambda_n}{\lambda_n + \mu_i} \mu_i^r \langle h, e_i \rangle_2 e_i. \end{aligned}$$

Therefore, the RKHS-norm of $\tilde{f} - f^*$ can be bounded as

$$\begin{aligned} \|\tilde{f} - f^*\|_K^2 &= \sum_{i \geq 1} \left(\frac{\lambda_n}{\lambda_n + \mu_i} \mu_i^{r-1/2} \langle h, e_i \rangle_2 \right)^2 \|\mu_i^{1/2} e_i\|_K^2 \\ &= \sum_{i \geq 1} \left(\frac{\lambda_n}{\lambda_n + \mu_i} \mu_i^{r-1/2} \langle h, e_i \rangle_2 \right)^2 \\ &= \lambda_n^{2r-1} \sum_{i \geq 1} \left(\frac{\lambda_n}{\lambda_n + \mu_i} \right)^{3-2r} \left(\frac{\mu_i}{\lambda_n + \mu_i} \right)^{2r-1} \langle h, e_i \rangle_2^2 \\ (7.5) \quad &\leq \lambda_n^{2r-1} \sum_{i \geq 1} \langle h, e_i \rangle_2^2 = \lambda_n^{2r-1} \|h\|_2^2 = \lambda_n^{2r-1} \|L_K^{-r} f^*\|_2^2. \end{aligned}$$

It then follows from Proposition 1 that

$$P_2 \leq P\left(\|\hat{f} - \tilde{f}\|_K \geq \frac{6\kappa_1 \log n}{\lambda_n n^{1/2}} \log \frac{8}{\delta_n} \mid \{|y| \leq \log n\}\right) \leq \delta_n/4.$$

Combining the upper bounds of P_1 and P_2 yields that $P(\mathcal{C}_1) \leq \delta_n/4 + (\sigma^2 + \|f^*\|_2^2)(\log n)^{-2}$. Since $\delta_n \geq 4(\sigma^2 + \|f^*\|_2^2)(\log n)^{-2}$, we have $P(\mathcal{C}_1) \leq \delta_n/2$. Thus, with probability at least $1 - \delta_n/2$, there holds

$$\|\hat{f} - f^*\|_K \leq 2 \log \frac{8}{\delta_n} \left(\frac{3\kappa_1 \log n}{n^{1/2} \lambda_n} + \lambda_n^{r-1/2} \|L_K^{-r} f^*\|_2 \right).$$

Now we turn to establish the weak convergence rate of \hat{g}_l in estimating g_l^* . We first introduce some notations. Define the sample operators for gradients $\hat{D}_l : \mathcal{H}_K \rightarrow \mathcal{R}^n$ and their adjoint operators $\hat{D}_l^* : \mathcal{R}^n \rightarrow \mathcal{H}_K$ as

$$(\hat{D}_l f)_i = \langle f, \partial_l K_{\mathbf{x}_i} \rangle_K \quad \text{and} \quad \hat{D}_l^* c = \frac{1}{n} \sum_{i=1}^n \partial_l K_{\mathbf{x}_i} c_i,$$

respectively. And the integral operators for gradients $D_l : \mathcal{H}_K \rightarrow \mathcal{L}^2(\rho_{\mathbf{x}}, \mathcal{X})$ and $D_l^* : \mathcal{L}^2(\rho_{\mathbf{x}}, \mathcal{X}) \rightarrow \mathcal{H}_K$ are defined as

$$D_l f = \langle f, \partial_l K_{\mathbf{x}} \rangle_K \text{ and } D_l^* f = \int \partial_l K_{\mathbf{x}} f(\mathbf{x}) d\rho_{\mathbf{x}}.$$

Note that D_l and \widehat{D}_l are the Hilbert-Schmidt operators by Propositions 12 and 13 of Rosasco et al. (2013, [20]), then we have

$$D_l^* D_l f = \int \partial_l K_{\mathbf{x}} g_l(\mathbf{x}) d\rho_{\mathbf{x}} \text{ and } \widehat{D}_l^* \widehat{D}_l f = \frac{1}{n} \sum_{i=1}^n \partial_l K_{\mathbf{x}_i} g_l(\mathbf{x}_i).$$

Furthermore, we denote $HS(K)$ as a Hilbert space with all the Hilbert-Schmidt operators on \mathcal{H}_K , which endows with a norm $\|\cdot\|_{HS}$ such that $\|T\|_K \leq \|T\|_{HS}$ for any $T \in HS(K)$.

With these operators, simple algebra yields that

$$\begin{aligned} & \left| \|\widehat{g}_l\|_n^2 - \|g_l^*\|_2^2 \right| \\ &= \left| \frac{1}{n} \sum_{i=1}^n (\widehat{g}_l(\mathbf{x}_i))^2 - \int (g_l^*(\mathbf{x}))^2 d\rho_{\mathbf{x}} \right| \\ &= \left| \frac{1}{n} \sum_{i=1}^n \widehat{g}_l(\mathbf{x}_i) \langle \widehat{f}, \partial_l K_{\mathbf{x}_i} \rangle_K - \int g_l^*(\mathbf{x}) \langle f^*, \partial_l K_{\mathbf{x}} \rangle_K d\rho_{\mathbf{x}} \right| \\ &= \left| \langle \widehat{f}, \frac{1}{n} \sum_{i=1}^n \widehat{g}_l(\mathbf{x}_i) \partial_l K_{\mathbf{x}_i} \rangle_K - \langle f^*, \int g_l^*(\mathbf{x}) \partial_l K_{\mathbf{x}} d\rho_{\mathbf{x}} \rangle_K \right| \\ &= \left| \langle \widehat{f} - f^*, \widehat{D}_l^* \widehat{D}_l \widehat{f} \rangle_K + \langle f^*, \widehat{D}_l^* \widehat{D}_l (\widehat{f} - f^*) \rangle_K + \langle f^*, (\widehat{D}_l^* \widehat{D}_l - D_l^* D_l) f^* \rangle_K \right| \\ &= \left| \langle \widehat{f} - f^*, \widehat{D}_l^* \widehat{D}_l (\widehat{f} - f^*) \rangle_K + \langle \widehat{D}_l^* \widehat{D}_l f^*, \widehat{f} - f^* \rangle_K + \right. \\ &\quad \left. \langle f^*, \widehat{D}_l^* \widehat{D}_l (\widehat{f} - f^*) \rangle_K + \langle f^*, (\widehat{D}_l^* \widehat{D}_l - D_l^* D_l) f^* \rangle_K \right| \\ &\leq \|\widehat{f} - f^*\|_K^2 \|\widehat{D}_l^* \widehat{D}_l\|_{HS} + 2\|\widehat{f} - f^*\|_K \|f^*\|_K \|\widehat{D}_l^* \widehat{D}_l\|_{HS} + \\ &\quad \|\widehat{D}_l^* \widehat{D}_l - D_l^* D_l\|_{HS} \|f^*\|_K^2, \end{aligned}$$

where the last inequality follows from the Cauchy-Schwartz inequality. It then suffices to bound the terms in the upper bound of $\left| \|\widehat{g}_l\|_n^2 - \|g_l^*\|_2^2 \right|$ separately. Note that $\|f^*\|_K$ is a bounded quantity, and it follows from Assumption 2 and Rosasco et al. (2013, [20]) that $\max_l \|\widehat{D}_l^* \widehat{D}_l\|_{HS} = \max_l \|\partial_l K_{\mathbf{x}}\|_K^2 \leq \kappa_2^2$. Hence, we have

$$\begin{aligned} & \max_{1 \leq l \leq p_n} \left| \|\widehat{g}_l\|_n^2 - \|g_l^*\|_2^2 \right| \\ & \leq a_1 \left(\|\widehat{f} - f^*\|_K^2 + 2\|\widehat{f} - f^*\|_K + \max_{1 \leq l \leq p_n} \|\widehat{D}_l^* \widehat{D}_l - D_l^* D_l\|_{HS} \right), \end{aligned}$$

where $a_1 = \max\{\kappa_2^2, \kappa_2^2\|f^*\|_K, \|f^*\|_K^2\}$. When $\|\hat{f} - f^*\|_K$ is sufficiently small, the upper bound can be simplified to

$$\max_{1 \leq l \leq p_n} \left| \|\hat{g}_l\|_n^2 - \|g_l^*\|_2^2 \right| \leq a_1 \left(3\|\hat{f} - f^*\|_K + \max_{1 \leq l \leq p_n} \|\hat{D}_l^* \hat{D}_l - D_l^* D_l\|_{HS} \right),$$

where $\|\hat{f} - f^*\|_K$ is bounded in the first half of the proof. Furthermore, for any $\epsilon_n \in (0, 1)$, by the concentration inequalities for $HS(K)$ (Rosasco et al., 2013, [20]), we have

$$P\left(\|\hat{D}_l^* \hat{D}_l - D_l^* D_l\|_{HS} \geq \epsilon_n\right) \leq 2p_n \exp\left(-\frac{n\epsilon_n^2}{8\kappa_2^4}\right),$$

for any $l = 1, \dots, p_n$. Therefore, with probability at least $1 - \delta_n/2$, there holds

$$\max_{1 \leq l \leq p_n} \|\hat{D}_l^* \hat{D}_l - D_l^* D_l\|_{HS} \leq \left(\frac{8\kappa_2^4}{n} \log \frac{4p_n}{\delta_n}\right)^{1/2}.$$

Combining all the upper bounds above, we have with probability at least $1 - \delta_n$, there holds

$$\begin{aligned} & \max_{1 \leq l \leq p_n} \left| \|\hat{g}_l\|_n^2 - \|g_l^*\|_2^2 \right| \\ & \leq 2a_1 \left(3 \log \frac{8}{\delta_n} \left(\frac{3\kappa_1 \log n}{n^{1/2} \lambda_n} + \lambda_n^{r-1/2} \|L_K^{-r} f^*\|_2 \right) + \left(\frac{2\kappa_2^4}{n} \log \frac{4p_n}{\delta_n} \right)^{1/2} \right). \end{aligned}$$

This implies the desired results immediately with $\lambda_n = n^{-\frac{1}{2r+1}} (\log n)^{\frac{2}{2r+1}}$. ■

Proof of Theorem 2: We first show that $\mathcal{A}^* \subset \hat{\mathcal{A}}$ in probability. If not, suppose there exists some $l' \in \mathcal{A}^*$ but $l' \notin \hat{\mathcal{A}}$, and thus $\|\hat{g}_{l'}\|_n^2 \leq v_n$. By Assumption 3, we have with probability $1 - \delta_n$ that

$$\begin{aligned} \left| \|\hat{g}_{l'}\|_n^2 - \|g_{l'}^*\|_2^2 \right| & \geq \|g_{l'}^*\|_2^2 - \|\hat{g}_{l'}\|_n^2 > c_2 n^{-\frac{(2r-1)}{2(2r+1)}} \log p_n (\log n)^{\xi_1} - v_n \\ & = \frac{c_2}{2} n^{-\frac{(2r-1)}{2(2r+1)}} \log p_n (\log n)^{\xi_1}, \end{aligned}$$

which contradicts with Theorem 1. This implies that $\mathcal{A}^* \subset \hat{\mathcal{A}}$ with probability at least $1 - \delta_n$.

Next, we show that $\hat{\mathcal{A}} \subset \mathcal{A}^*$ in probability. If not, suppose there exists some $l' \in \hat{\mathcal{A}}$ but $l' \notin \mathcal{A}^*$, which implies $\|\hat{g}_{l'}\|_n^2 > v_n$ but $\|g_{l'}^*\|_2^2 = 0$, and then with probability at least $1 - \delta_n$, there holds

$$\left| \|\hat{g}_{l'}\|_n^2 - \|g_{l'}^*\|_2^2 \right| > v_n = \frac{c_2}{2} n^{-\frac{(2r-1)}{2(2r+1)}} \log p_n (\log n)^{\xi_1}.$$

This contradicts with Theorem 1 again, and thus $\hat{\mathcal{A}} \subset \mathcal{A}^*$ with probability at least $1 - \delta_n$. Combining these two results yields the desired sparsistency. ■

Proof of Corollary 1: Corollary 1 is a direct application of Theorem 1 for the scaled linear kernel $K(\mathbf{x}, \mathbf{u}) = \mathbf{x}^T \mathbf{u} / p_n$, and we just need to verify the assumptions of Theorem 1. In fact, Assumption 4 implies that $E(\mathbf{x} \mathbf{x}^T)$ is invertible, and thus Assumption 1 is satisfied for the scaled linear kernel with $r = 1$. Assumption 2 is also satisfied due to the fact that $\sup_{\mathbf{x} \in \mathcal{X}} \|K_{\mathbf{x}}\|_K = p_n^{-1/2} \|\mathbf{x}\|$ is bounded on a compact support $\mathcal{X} \subset \mathcal{R}^{p_n}$. Furthermore,

$$\begin{aligned} \|L_K^{-1} f^*\|_2 &= \|(E \tilde{\mathbf{x}} \tilde{\mathbf{x}}^T)^{-1} \tilde{\boldsymbol{\beta}}^*\|_2 = (\tilde{\boldsymbol{\beta}}^{*T} (E \mathbf{x} \mathbf{x}^T / p_n)^{-1} \tilde{\boldsymbol{\beta}}^*)^{1/2} \\ &\leq p_n^{-1/2} \lambda_{\min}(E(\mathbf{x} \mathbf{x}^T) / p_n)^{-1/2} \|\boldsymbol{\beta}^*\| = O(p_n^{-1/2} n^{(\tau_1 + \tau_2)/2}). \end{aligned}$$

The desired result then follows from Theorem 1 immediately. ■

Proof of Corollary 2: First, note that $\mathcal{A}^* = \{l : \tilde{\beta}_l^* \neq 0\}$ and $\hat{\mathcal{A}}_{v_n} = \{l : |\check{\beta}_l| > p_n^{-1/2} v_n\}$ with $\check{\beta}_l = p_n^{-1/2} \hat{\beta}_l$. Clearly, (4.3) directly implies that for any $l = 1, \dots, p_n$, with probability at least $1 - \delta_n$, there holds

$$|\check{\beta}_l - \tilde{\beta}_l^*| \leq 2c_3 \log \frac{4}{\delta_n} p_n^{-1/3} n^{-\frac{1-2(\tau_1+\tau_2)}{6}} (\log n)^{1/3}.$$

Therefore, following the proof of Theorem 2 and $v_n = \frac{c_4}{2} p_n^{1/6} n^{-\frac{1-2(\tau_1+\tau_2)}{6}} (\log n)^{\xi_2}$, we have $P(\hat{\mathcal{A}}_{v_n} = \mathcal{A}^*) \rightarrow 1$. ■

Proof of Theorem 3: For simplicity, denote

$$\mathcal{C}_2 = \left\{ \max_{l, k \in \hat{\mathcal{A}}} \left| \|\hat{g}_{lk}\|_n^2 - \|g_{lk}^*\|_2^2 \right| > c_6 \left(\log \frac{8p_0^2}{\delta_n} \right) n^{-\frac{(2r-1)}{2(2r+1)}} (\log n)^{\frac{2r-1}{2r+1}} \right\}.$$

Note that $P(\mathcal{C}_2)$ can be decomposed as

$$\begin{aligned} P(\mathcal{C}_2) &= P(\mathcal{C}_2 \cap \{\hat{\mathcal{A}} = \mathcal{A}^*\}) + P(\mathcal{C}_2 \cap \{\hat{\mathcal{A}} \neq \mathcal{A}^*\}) \\ &\leq P(\hat{\mathcal{A}} \neq \mathcal{A}^*) + P(\mathcal{C}_2 | \hat{\mathcal{A}} = \mathcal{A}^*) P(\hat{\mathcal{A}} = \mathcal{A}^*) = \Delta_n + P_3(1 - \Delta_n), \end{aligned}$$

where $\Delta_n \rightarrow 0$ according to Theorem 2, and P_3 can be bounded as follows.

To bound P_3 , we first introduce some additional notations. Denote the operators for the second-order gradients as

$$D_{lk}^* D_{lk} f = \int \partial_{lk}^2 K_{\mathbf{x}} g_{lk}(\mathbf{x}) d\rho_{\mathbf{x}} \quad \text{and} \quad \hat{D}_{lk}^* \hat{D}_{lk} f = \frac{1}{n} \sum_{i=1}^n \partial_{lk}^2 K_{\mathbf{x}_i} \hat{g}_{lk}(\mathbf{x}_i),$$

where $\partial_{lk}^2 K_{\mathbf{x}} = \frac{\partial^2 K(\mathbf{x}, \cdot)}{\partial x^l \partial x^k}$. Hence, for any $l, k \in \mathcal{A}^*$, we have

$$\begin{aligned}
& \left| \|\widehat{g}_{lk}\|_n^2 - \|g_{lk}^*\|_2^2 \right| \\
&= \left| \frac{1}{n} \sum_{i=1}^n (\widehat{g}_{lk}(\mathbf{x}_i))^2 - \int (g_{lk}^*(\mathbf{x}))^2 d\rho_{\mathbf{x}} \right| \\
&= \left| \frac{1}{n} \sum_{i=1}^n \widehat{g}_{lk}(\mathbf{x}_i) \langle \widehat{f}, \partial_{lk}^2 K_{\mathbf{x}_i} \rangle_K - \int g_{lk}^*(\mathbf{x}) \langle f^*, \partial_{lk}^2 K_{\mathbf{x}} \rangle_K d\rho_{\mathbf{x}} \right| \\
&= \left| \langle \widehat{f}, \widehat{D}_{lk}^* \widehat{D}_{lk} \widehat{f} \rangle_K + \langle f^*, D_{lk}^* D_{lk} f^* \rangle_K \right| \\
&= \left| \langle \widehat{f} - f^*, \widehat{D}_{lk}^* \widehat{D}_{lk} (\widehat{f} - f^*) \rangle_K + 2 \langle f^*, \widehat{D}_{lk}^* \widehat{D}_{lk} (\widehat{f} - f^*) \rangle_K + \right. \\
&\quad \left. \langle f^*, (\widehat{D}_{lk}^* \widehat{D}_{lk} - D_{lk}^* D_{lk}) f^* \rangle_K \right| \\
&\leq \kappa_3^2 \|\widehat{f} - f^*\|_K^2 + 2\kappa_3 \|f^*\|_K \|\widehat{f} - f^*\|_K + \|f^*\|_K^2 \|\widehat{D}_{lk}^* \widehat{D}_{lk} - D_{lk}^* D_{lk}\|_{HS},
\end{aligned}$$

where the last inequality follows from the Cauchy-Schwartz inequality.

Note that $\|f^*\|_K$ is bounded, and D_{lk} and \widehat{D}_{lk} are Hilbert-Schmidt operators on \mathcal{H}_K by Assumption 6 and a slightly modified proof of Proposition 6 in Vito et al. (2005, [31]). It then follows from Rosasco et al. (2013, [20]) that $\max_{l,k \in \mathcal{A}^*} \|\widehat{D}_{lk}^* \widehat{D}_{lk}\|_{HS} \leq \kappa_3^2$. Hence, conditional on $\widehat{\mathcal{A}} = \mathcal{A}^*$, we have

$$\begin{aligned}
& \max_{l,k \in \mathcal{A}^*} \left| \|\widehat{g}_{lk}\|_n^2 - \|g_{lk}^*\|_2^2 \right| \\
&\leq a_2 \left(\|\widehat{f} - f^*\|_K^2 + 2\|\widehat{f} - f^*\|_K + \max_{l,k \in \mathcal{A}^*} \|\widehat{D}_{lk}^* \widehat{D}_{lk} - D_{lk}^* D_{lk}\|_{HS} \right) \\
&\leq a_2 \left(3\|\widehat{f} - f^*\|_K + \max_{l,k \in \mathcal{A}^*} \|\widehat{D}_{lk}^* \widehat{D}_{lk} - D_{lk}^* D_{lk}\|_{HS} \right),
\end{aligned}$$

where $a_2 = \max\{\kappa_3^2, \|f^*\|_K^2, \kappa_3^2 \|f^*\|_K\}$, and the second inequality holds when $\|\widehat{f} - f^*\|_K^2$ is sufficiently small. Here $\|\widehat{f} - f^*\|_K$ is bounded in Theorem 1. Moreover, for any $\epsilon_n \in (0, 1)$ and $l, k \in \mathcal{A}^*$, by the concentration inequalities in $HS(K)$ on \mathcal{H}_K (Rosasco et al., 2013, [20]), we have

$$P\left(\|\widehat{D}_{lk}^* \widehat{D}_{lk} - D_{lk}^* D_{lk}\|_{HS} \geq \epsilon_n\right) \leq 2 \exp\left(-\frac{n\epsilon_n^2}{8\kappa_3^4}\right).$$

Let $\epsilon_n = \left(\frac{8\kappa_3^4}{n} \log \frac{4}{\delta_n}\right)^{1/2}$, then with probability at least $1 - \delta_n/2$, there holds

$$\max_{l,k \in \mathcal{A}^*} \|\widehat{D}_{lk}^* \widehat{D}_{lk} - D_{lk}^* D_{lk}\|_{HS} \leq \left(\frac{8\kappa_3^4}{n} \log \frac{4p_0^2}{\delta_n}\right)^{1/2}.$$

Therefore, conditional on $\hat{\mathcal{A}} = \mathcal{A}^*$, we have with probability at least $1 - \delta_n$, there holds

$$\begin{aligned} & \max_{l,k \in \hat{\mathcal{A}}} \left| \|\hat{g}_{lk}\|_n^2 - \|g_{lk}^*\|_2^2 \right| \\ & \leq a_2 \left(3 \log \frac{8}{\delta_n} \left(\frac{3\kappa_1 \log n}{n^{1/2} \lambda_n} + \lambda_n^{r-1/2} \|L_K^{-r} f^*\|_2 \right) + \left(\frac{8\kappa_3^4}{n} \log \frac{4p_0^2}{\delta_n} \right)^{1/2} \right). \end{aligned}$$

Furthermore, with $\lambda_n = n^{-\frac{1}{2r+1}} (\log n)^{\frac{2}{2r+1}}$, the upper bound reduces to

$$\max_{l,k \in \hat{\mathcal{A}}} \left| \|\hat{g}_{lk}\|_n^2 - \|g_{lk}^*\|_2^2 \right| \leq c_6 \left(\log \frac{8p_0^2}{\delta_n} \right) n^{-\frac{(2r-1)}{2(2r+1)}} (\log n)^{\frac{2r-1}{2r+1}},$$

where c_6 is given in Theorem 3, and hence that $P_3 \leq \delta_n$. Therefore, $P(\mathcal{C}_2) \leq \Delta_n + \delta_n(1 - \Delta_n) \leq \Delta_n + \delta_n$, and the desired result follows immediately. \blacksquare

Proof of Theorem 4: Note that

$$\begin{aligned} & P(\hat{\mathcal{A}}_2 = \mathcal{A}_2^*, \hat{\mathcal{A}}_1 = \mathcal{A}_1^*) \\ & = P(\hat{\mathcal{A}}_2 = \mathcal{A}_2^*, \hat{\mathcal{A}}_1 = \mathcal{A}_1^*, \hat{\mathcal{A}} = \mathcal{A}^*) \\ & = P(\hat{\mathcal{A}}_2 = \mathcal{A}_2^*, \hat{\mathcal{A}}_1 = \mathcal{A}_1^* | \hat{\mathcal{A}} = \mathcal{A}^*) P(\hat{\mathcal{A}} = \mathcal{A}^*) \\ & \geq \left(1 - P(\hat{\mathcal{A}}_2 \neq \mathcal{A}_2^* | \hat{\mathcal{A}} = \mathcal{A}^*) - P(\hat{\mathcal{A}}_1 \neq \mathcal{A}_1^* | \hat{\mathcal{A}} = \mathcal{A}^*) \right) P(\hat{\mathcal{A}} = \mathcal{A}^*) \\ & = \left(1 - 2P(\hat{\mathcal{A}}_2 \neq \mathcal{A}_2^* | \hat{\mathcal{A}} = \mathcal{A}^*) \right) (1 - \Delta_n), \end{aligned}$$

where the last equality follows from the fact that $\hat{\mathcal{A}}_1 \cap \hat{\mathcal{A}}_2 = \mathcal{A}_1^* \cap \mathcal{A}_2^* = \emptyset$, and then $\{\hat{\mathcal{A}}_1 \neq \mathcal{A}_1^*\} = \{\hat{\mathcal{A}}_2 \neq \mathcal{A}_2^*\}$ given $\hat{\mathcal{A}} = \mathcal{A}^*$. By Theorem 2, $\Delta_n \rightarrow 0$ as n diverges. Therefore, it suffices to show $P(\hat{\mathcal{A}}_2 \neq \mathcal{A}_2^* | \hat{\mathcal{A}} = \mathcal{A}^*) \rightarrow 0$ as n diverges.

We first show that $\mathcal{A}_2^* \subset \hat{\mathcal{A}}_2$ in probability conditional on $\hat{\mathcal{A}} = \mathcal{A}^*$. If not, suppose that there exists some $l' \in \mathcal{A}_2^*$, which directly implies that $\|g_{l'k}^*\|_2^2 > c_5 n^{-\frac{(2r-1)}{2(2r+1)}} \log p_0 (\log n)^{\xi_3}$, for some $k \in \mathcal{A}^*$ but $l' \notin \hat{\mathcal{A}}_2$, and thus $\|\hat{g}_{l'k}\|_n^2 \leq v_n^{int}$. By Assumption 6, we have with probability at least $1 - \Delta_n - \delta_n$ that

$$\left| \|\hat{g}_{l'k}\|_n^2 - \|g_{l'k}^*\|_2^2 \right| \geq \|g_{l'k}^*\|_2^2 - \|\hat{g}_{l'k}\|_n^2 > \frac{c_5}{2} n^{-\frac{(2r-1)}{2(2r+1)}} \log p_0 (\log n)^{\xi_3},$$

which contradicts with Theorem 3. This implies that conditional on $\hat{\mathcal{A}} = \mathcal{A}^*$, $\mathcal{A}_2^* \subset \hat{\mathcal{A}}_2$ with probability at least $1 - \Delta_n - \delta_n$.

Next, we show that $\hat{\mathcal{A}}_2 \subset \mathcal{A}_2^*$ in probability conditional on $\hat{\mathcal{A}} = \mathcal{A}^*$. If not, suppose there exists some $l' \in \hat{\mathcal{A}}_2$ but $l' \notin \mathcal{A}_2^*$, which implies $\|\hat{g}_{l'k}\|_n^2 > v_n^{int}$ for

some $k \in \hat{A}_Q$ but $\|g_{l'k}^*\|_2^2 = 0$. Then with probability at least $1 - \Delta_n - \delta_n$, there holds

$$\left| \|\hat{g}_{l'k}\|_n^2 - \|g_{l'k}^*\|_2^2 \right| = \|\hat{g}_{l'k}\|_n^2 > \frac{c_5}{2} n^{-\frac{(2r-1)}{2(2r+1)}} \log p_0 (\log n)^{\xi_3},$$

which contradicts with Theorem 3 again. Therefore, conditional on $\hat{\mathcal{A}} = \mathcal{A}^*$, $\hat{\mathcal{A}}_2 \subset \mathcal{A}_2^*$ with probability at least $1 - \Delta_n - \delta_n$.

Combining these two results yields that $P(\hat{\mathcal{A}}_2 = \mathcal{A}_2^* | \hat{\mathcal{A}} = \mathcal{A}^*) \geq 1 - 2\Delta_n - 2\delta_n$, or equivalently, $P(\hat{\mathcal{A}}_2 \neq \mathcal{A}_2^* | \hat{\mathcal{A}} = \mathcal{A}^*) \leq 2\Delta_n + 2\delta_n \rightarrow 0$. The desired sparsistency then follows immediately. ■

References.

- [1] J. Bien, J. Taylor, and R. Tibshirani. A lasso for hierarchical interactions. *Annals of Statistics*, **41**:1111–1141, 2013.
- [2] H. Bondell and L. Li. Shrinkage inverse regression estimation for model free variable selection. *Journal of the Royal Statistical Society, Series B*, **71**:287–299, 2009.
- [3] J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, **96**:1348–1360, 2001.
- [4] J. Fan and J. Lv. Sure independence screening for ultrahigh dimensional feature space (with discussion). *Journal of the Royal Statistical Society, Series B*, **70**:849–911, 2008.
- [5] J. Fan, F. Yang, and R. Song. Nonparametric independence screening in sparse ultrahigh dimensional additive models. *Journal of the American Statistical Association*, **106**:544–557, 2011.
- [6] S. Fine and K. Scheinberg. Efficient svm training using low-rank kernel representations. *Journal of Machine Learning Research*, **2**:243–264, 2002.
- [7] S. Fischer and I. Steinwart. Sobolev norm learning rates for regularized least square algorithm. *Manuscript*, 2018.
- [8] K. Fukumizu and C. Leng. Gradient-based kernel dimension reduction for regression. *Journal of the American Statistical Association*, **109**:359–370, 2014.
- [9] M. Genton. Classes of kernels for machine learning: a statistical perspective. *Journal of Machine Learning Research*, **2**:299–312, 2001.
- [10] N. Hao, Y. Feng, and H. Zhang. Model selection for high dimensional quadratic regression via regularization. *Journal of the American Statistical Association*, In press, 2018.
- [11] X. He, L. Wang, and H. Hong. Quantile-adaptive model-free variable screening for high-dimensional heterogeneous data. *Annals of Statistics*, **41**:342–369, 2013.
- [12] J. Huang, J. Horowitz, and F. Wei. Variable selection in nonparametric additive models. *Annals of Statistics*, **38**:2282–2313, 2010.
- [13] Y. Kong, D. Li, Y. Fan, and J. Lv. Interaction pursuit in high-dimensional multi-response regression via distance correlation. *Annals of Statistics*, **45**:897–922, 2017.
- [14] B. Li, H. Zha, and F. Chiaromonte. Contour regression: a general approach to dimension reduction. *Annals of Statistics*, **33**:1580–1616, 2005.
- [15] R. Li, W. Zhong, and L. Zhu. Feature screening via distance correlation learning. *Journal of the American Statistical Association*, **107**:1129–1139, 2012.
- [16] M. Lim and T. Hastie. Learning interactions via hierarchical group-lasso regularization. *Journal of Computational and Graphical Statistics*, **24**:627–654, 2015.
- [17] Y. Lin and H. Zhang. Component selection and smoothing in multivariate nonparametric regression. *Annals of Statistics*, **34**:2272–2297, 2006.

- [18] S. Mendelson and J. Neeman. Regularization in kernel learning. *Annal of Statistics*, **38**:526–565, 2010.
- [19] M. Ritchie, L. Hahn, N. Roodi, L. Bailey, W. Dupont, F. Parl, and J. Moore. Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. *The American Journal of Human Genetics*, **69**:138–147, 2001.
- [20] L. Rosasco, S. Villa, S. Mosci, M. Santoro, and A. Verri. Nonparametric sparsity and regularization. *Journal of Machine Learning Research*, **14**:1665–1714, 2013.
- [21] J. Shao and X. Deng. Estimation in high-dimensional linear models with deterministic design matrices. *Annals of Statistics*, **40**:1821–1831, 2012.
- [22] Y. She, Z. Wang, and H. Jiang. Group regularized estimation under structural hierarchy. *Journal of the American Statistical Association*, In press, 2017.
- [23] X. Shen, W. Pan, and Y. Zhu. Likelihood-based selection and sharp parameter estimation. *Journal of the American Statistical Association*, **107**:223–232, 2012.
- [24] T. Shively, R. Kohn, and S. Wood. Variable selection and function estimation in additive non-parametric regression using a data-based prior. *Journal of the American Statistical Association*, **94**:777–794, 1999.
- [25] S. Smale and D. Zhou. Shannon sampling ii: connections to learning theory. *Applied and Computational Harmonic Analysis*, **19**:285–302, 2005.
- [26] S. Smale and D. Zhou. Learning theory estimates via integral operators and their approximations. *Constructive Approximation*, **26**:153–172, 2007.
- [27] L. Stefanski, Y. Wu, and K. White. Variable selection in nonparametric classification via measurement error model selection likelihoods. *Journal of the American Statistical Association*, **109**:574–589, 2014.
- [28] I. Steinwart. Consistency of support vector machines and other regularized kernel classifiers. *IEEE Transactions on Information Theory*, **51**:128–142, 2005.
- [29] W. Sun, J. Wang, and Y. Fang. Consistent selection of tuning parameters via variable selection stability. *Journal of Machine Learning Research*, **14**:3419–3440, 2013.
- [30] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, **58**:267–288, 1996.
- [31] E. Vito, L. Rosasco, A. Caponnetto, U. Giovannini, and F. Odone. Learning from examples as an inverse problem. *Journal of Machine Learning Research*, **6**:883–904, 2005.
- [32] G. Wahba. Support vector machines, reproducing kernel hilbert spaces, and randomized gacv. *Advances in kernel methods: support vector learning*, pages 69–88, 1998.
- [33] H. Wang. Forward regression for ultra-high dimensional variable screening. *Journal of the American Statistical Association*, **104**:1512–1524, 2009.
- [34] X. Wang and C. Leng. High dimensional ordinary least squares projection for screening variables. *Journal of the Royal Statistical Society, Series B*, **78**:589–611, 2016.
- [35] C. Williams and M. Seeger. Using the nystrom method to speed up kernel machines. *Advances in Neural Information Processing Systems*, **13**:682–688, 2001.
- [36] Y. Wu and L. Stefanski. Automatic structure recovery for additive models. *Biometrika*, **102**:381–395, 2015.
- [37] L. Yang, S. Lv, and J. Wang. Model-free variable selection in reproducing kernel hilbert space. *Journal of Machine Learning Research*, **17**:1–24, 2016.
- [38] C. Zhang. Nearly unbiased variable selection under minimax concave penalty. *Annals of Statistics*, **38**:894–942, 2010.
- [39] D. Zhou. Derivative reproducing properties for kernel methods in learning theory. *Journal of Computational and Applied Mathematics*, **220**:456–463, 2007.
- [40] H. Zou. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, **101**:1418–1429, 2006.

XIN HE AND JUNHUI WANG
DEPARTMENT OF MATHEMATICS
CITY UNIVERSITY OF HONG KONG
E-MAIL: xinhe6-c@my.cityu.edu.hk
j.h.wang@cityu.edu.hk

SHAOGAO LV
CENTER OF STATISTICS
SOUTHWESTERN UNIVERSITY OF FINANCE AND ECONOMICS
E-MAIL: lvsg716@swufe.edu.cn