Journal of Machine Learning Research 17 (2016) 1-24 Submitted 7/15; Revised 2/16; Published 5/16

Model-free Variable Selection in Reproducing Kernel Hilbert Space

Lei Yang LY888@NYU.EDU
Department of Population Health
New York Univeristy
New York, NY, 10016, USA

Shaogao Lv LVSG716@SWUFE.EDU.CN
Center of Statistics
Southwestern University of Finance and Economics
Chengdu, Sichuan, 610074, China

Junhui Wang J.H.WANG@CITYU.EDU.HK
Department of Mathematics
City University of Hong Kong
Kowloon Tong, 999077, HongKong

Editor: Jie Peng

Abstract

Variable selection is popular in high-dimensional data analysis to identify the truly informative variables. Many variable selection methods have been developed under various model assumptions. Whereas success has been widely reported in literature, their performances largely depend on validity of the assumed models, such as the linear or additive models. This article introduces a model-free variable selection method via learning the gradient functions. The idea is based on the equivalence between whether a variable is informative and whether its corresponding gradient function is substantially non-zero. The proposed variable selection method is then formulated in a framework of learning gradients in a flexible reproducing kernel Hilbert space. The key advantage of the proposed method is that it requires no explicit model assumption and allows for general variable effects. Its asymptotic estimation and selection consistencies are studied, which establish the convergence rate of the estimated sparse gradients and assure that the truly informative variables are correctly identified in probability. The effectiveness of the proposed method is also supported by a variety of simulated examples and two real-life examples.

Keywords: group Lasso, high-dimensional data, kernel regression, learning gradients, reproducing kernel Hilbert space (RKHS), variable selection

1. Introduction

The rapid advance of technology has led to an increasing demand for modern statistical techniques, such as high-dimensional data analysis that has attracted tremendous interests in the past two decades. When analyzing high-dimensional data, it is often believed that only a small number of variables are truly informative while others are noise. Therefore, identifying the truly informative variables is regarded as one of the primary goals in high-dimensional data analysis as well as many real applications such as health studies.

In literature, a wide spectrum of variable selection methods have been proposed based on various model assumptions. For example, under the linear model assumption, regularized regression models are popularly used for variable selection, including the nonnegative garrote (Breiman and Friedman, 1985), the least absolute shrinkage and selection operator (Tibshirani, 1996), the smoothly clipped absolute deviation (Fan and Li, 2001), the adaptive Lasso (Zou, 2006), the combined L_0 and L_1 penalty (Liu and Wu, 2007), the truncated L_1 penalty (Shen et al., 2012), and many others. The main strategy is to associate the least square loss function with a sparsity-inducing penalty, leading to sparse representation of the resultant regression function. With the linear regression model, the sparse representation leads to variable selection based on whether the corresponding regression coefficient is zero.

The aforementioned variable selection methods have demonstrated superior performance in many real applications. Yet their success largely relies on the validity of the linear model assumption. To relax the model assumption, attempts have been made to extend the variable selection methods to a nonparametric regression context. For example, under the additive regression model assumption, a number of variable selection methods have been developed (Shively et al., 1999; Huang and Yang, 2004; Xue, 2009; Huang et al., 2010). Furthermore, higher-order additive models can be considered, allowing each functional component contain more than one variables, such as the component selection and smoothing operator (Cosso) method (Lin and Zhang, 2006). While this method provides a more flexible and still interpretable model compared to the classical additive models, the number of functional components increases exponentially with the dimension. Another stream of research on variable selection is to conduct screening (Fan et al., 2011; Zhu et al., 2011; Li et al., 2012), which treats each individual variable separately and assures the sure screening properties. To overcome the issue of ignoring interaction effects, a higher-order interaction screening method is also developed (Hao and Zhang, 2014). Model-free variable selection has also been approached in the context of sufficient dimension reduction (Li et al., 2005; Bondell and Li, 2009). More recently, Stefanski et al. (2014) introduced a novel measurement-error-model-based variable selection method that can be adapted to a nonparametric kernel regression.

In this article, we propose a novel model-free variable selection method, which requires no explicit model assumptions and allows for general variable effects. The method is based on the idea that a variable is truly informative with respect to the regression function if the gradient of the regression function along the corresponding coordinate is substantially different from zero. Thus the proposed variable selection method is formulated in a gradient learning framework equipped with a flexible reproducing kernel Hilbert space (Wahba, 1999). Learning gradients can be traced back to Härdle and Gasser (1985). Some of its recent developments include Jarrow et al. (2004), Mukherjee and Zhou (2006), Ye and Xie (2012), and Brabanter et al. (2013), where the main focus is to estimate the gradient functions.

As opposed to estimating the gradient functions, this article focuses on variable selection whose primary interest is to identify the truly informative variables corresponding to the non-zero gradient functions. To attain the sparsity in the estimated gradients, we consider a learning algorithm generated by a coefficient-based regularization scheme (Scholköpf and Smola, 2002), and a group Lasso penalty (Yuan and Lin, 2006) is enforced on the coefficients so that the proposed method can conduct gradient learning and variable selection

simultaneously. Specifically, the proposed variable selection method via gradient learning is formulated in a regularization form that consists of a pairwise loss function for estimating the gradient functions and a group Lasso penalty.

One of the main features of the proposed variable selection method is that it does not require any explicit model assumption and detect informative variables with various effects on the regression function. This is a major advantage over most existing model-based variable selection methods which need to pre-specify a working model. If higher-order variable effects are considered, the model-based methods need to enumerate the possible components, whose number increases exponentially with the dimension p. In sharp contrast, our proposed method only needs to estimate p components, while allowing for general variable effects.

Another interesting feature of the proposed method is the use of coefficient-based representation in estimating the gradient function. It follows directly from the representor theorem (Wahba, 1999) in a RKHS, and turns out to greatly facilitate variable selection in the gradient learning framework. With the coefficient-based representation, the group Lasso penalty can be naturally enforced on all the coefficients associated with the same variable. This leads to a well-structured optimization task, and can be efficiently solved through a blockwise coordinate descent algorithm (Yang and Zou, 2015). This is contrast to the existing gradient learning methods such as Ye and Xie (2012), where standard RHKS is used and a squared RKHS-norm penalty is enforced to attain the sparsity structure in the estimated gradients, and a forward-backward splitting algorithm is required for computation.

Finally, the effectiveness of the proposed method is supported by a variety of simulated and real examples. More importantly, its asymptotic estimation and selection consistencies are established, showing that the proposed method shall recover the truly informative variables with probability tending to one, and estimate the true gradient function at a fast convergence rate. Note that the variable selection consistency is not established in Ye and Xie (2012), and the estimation consistency of our method is more challenging due to the additional hypothesis error arises in the coefficient-based formulation. Also, as in many nonparametric variable selection methods (Lin and Zhang, 2006; Xue, 2009; Huang et al., 2010), the results are obtained in the sneario of fixed dimension, which are particlarly interesting given the fact that the variable selection consistency is obtained without assuming any explicit model.

The rest of the article is organized as follows. Section 2 presents a general framework of the proposed model-free variable selection method as well as an efficient computing algorithm to tackle the resultant large-scale optimization task. Section 3 establishes the asymptotic results of the proposed method in terms of both estimation and variable selection. The numerical experiments on the simulated examples and real applications are contained in Section 4. A brief discussion is provided in Section 5, and the Appendix is devoted to the technical proofs.

2. Model-free variable selection

2.1 Preambles

Suppose that a training set consists of $(\mathbf{x}_i, y_i)$; $i = 1, \ldots, n$, where $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^T \in \mathcal{R}^p$ and $y_i \in \mathcal{R}$ are independently sampled from some unknown joint distribution. We consider

the following regression model,

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

where $\mathrm{E}(\epsilon|\,\mathbf{x}) = 0$, $\mathrm{Var}(\epsilon|\,\mathbf{x}) = \sigma^2$, $\mathbf{x} = (x^{(1)}, \ldots, x^{(p)})^T$ is supported on a compact metric space $\mathcal{X}$, and f^* is the true regression function that is assumed to be twice differentiable everywhere.

When p is large, it is generally believed that only a small number of variables are truly informative. In literature, to define the truly informative variables, f^* is often assumed to be of certain form. For instance, if $f^*(\mathbf{x}) = \mathbf{x}^T\boldsymbol{\beta}^*$ with $\boldsymbol{\beta}^* = (\beta_1^*, \ldots, \beta_p^*)^T$, then $x^{(j)}$ is regarded as truly informative if $\beta_j^* \neq 0$. However, this linear model assumption on f^* can be too restrictive in practice, and whether a variable is informative shall not depend on the assumed model. In this article, a model-free variable selection method is developed without assuming any explicit form for f^*.

Since no explicit form is assumed for f^*, we note that if $x^{(l)}$ is non-informative in f^*, the corresponding gradient function $\nabla f_l^*(\mathbf{x}) = \partial f^*(\mathbf{x})/\partial x^{(l)} \equiv 0$ for any $\mathbf{x}$. This fact motivates the proposed model-free variable selection method in a gradient learning framework. Denote $\mathbf{g}^*(\mathbf{x}) = \nabla f^*(\mathbf{x}) = (\nabla f_1^*(\mathbf{x}), \ldots, \nabla f_p^*(\mathbf{x}))^T$ the true gradient function, and the estimation error as

$$\begin{aligned} \mathcal{E}(\mathbf{g}) &= E_{(\mathbf{x},y),(\mathbf{u},v)} w(\mathbf{x},\mathbf{u})(y - v - \mathbf{g}(\mathbf{x})^T(\mathbf{x}-\mathbf{u}))^2 \\ &= 2\sigma_s^2 + E_{\mathbf{x},\mathbf{u}} w(\mathbf{x},\mathbf{u})(f^*(\mathbf{x}) - f^*(\mathbf{u}) - \mathbf{g}(\mathbf{x})^T(\mathbf{x}-\mathbf{u}))^2, \end{aligned} \tag{1}$$

where $\sigma_s^2 = E_{(\mathbf{x},y),(\mathbf{u},v)}[w(\mathbf{x},\mathbf{u})(y - f^*(\mathbf{x}))^2]$ is independent of $\mathbf{g}$, and $w(\mathbf{x},\mathbf{u})$ is a weight function that decreases as $\|\,\mathbf{x}-\mathbf{u}\,\|$ increases and ensures the local neighborhood of $\mathbf{x}$ contributes more to estimating $\mathbf{g}^*(\mathbf{x})$. Typically, $w(\mathbf{x},\mathbf{u}) = e^{-\|\,\mathbf{x}-\mathbf{u}\,\|^2/\tau_n^2}$ with a pre-specified positive parameter τ_n^2, which plays a key role in the asymptotic estimation consistency and is to be elaborated.

2.2 Coefficient-based formulation

Given the training set $(\mathbf{x}_i, y_i);\ i = 1, \ldots, n$, $\mathcal{E}(\mathbf{g})$ is approximated by its empirical version, and then the proposed variable selection method is formulated as

$$\underset{\mathbf{g}\in\mathcal{H}_K^p}{\operatorname{argmin}}\ s(\mathbf{g}) = \frac{1}{n(n-1)} \sum_{i,j=1}^{n} w_{ij}\Big(y_i - y_j - \mathbf{g}(\mathbf{x}_i)^T(\mathbf{x}_i - \mathbf{x}_j)\Big)^2 + J(\mathbf{g}), \tag{2}$$

where $w_{ij} = w(\mathbf{x}_i, \mathbf{x}_j)$, $\mathcal{H}_K$ denotes a RKHS induced by a pre-specified kernel $K(\cdot,\cdot)$, $J(\mathbf{g}) = \lambda_n \sum_{l=1}^p \pi_l J(g_l)$ is a penalty function on the complexity of $\mathbf{g}$, and π_l's are the adaptive tuning parameters to be specified. The representor theorem assures that the minimizer of (2) must be of the following coefficient-based representation,

$$g_l(\mathbf{x}) = \sum_{t=1}^{n} \alpha_t^l K(\mathbf{x}, \mathbf{x}_t);\ l = 1, \ldots, p.$$

Thanks to the explicit form of $g_l(\mathbf{x})$, it is clear that $g_l(\mathbf{x}) \equiv 0$ is equivalent to $\alpha_t^l = 0$ for all t's, or more concisely, $\|\boldsymbol{\alpha}^{(l)}\|_2 = 0$ with $\boldsymbol{\alpha}^{(l)} = (\alpha_1^l, ..., \alpha_n^l)^T$. A similar formulation connecting

between ridge regression with a coefficient-based representation and support vector machine (Cortes and Vapnik, 1995) is also established in Scholköpf and Smola (2002).

Furthermore, to exploit the sparse structure in the regression model, we propose to consider the following sparsity-inducing penalty,

$$J(g_l) = \inf \Big\{ \|\boldsymbol{\alpha}^{(l)}\|_2 : g_l(\cdot) = \sum_{t=1}^{n} \alpha_t^l K(\cdot, \mathbf{x}_t) \Big\}. \tag{3}$$

Here the group Lasso type of penalty $\|\boldsymbol{\alpha}^{(l)}\|_2$ attains the effect of pushing all or none of α_t^l's to be exactly 0 and thus achieves the purpose of variable selection. The infimum is necessary for defining the penalty as the kernel basis $\{K(\cdot, \mathbf{x}_t)\}_{t=1}^n$ may not be linearly independent and thus the representation of g_l in $\mathcal{H}_K$ may not be unique. This penalty term differs from that in Ye and Xie (2012) in that our coefficient-based penalty does not rely on K and usually leads to sparser solutions. On the contrary, the penalty $\|g_l\|_K$ in Ye and Xie (2012) can be sensitive to the choice of K as its minimum eigenvalue can be very small. In addition, the finite dimensional hypothesis space is more flexible than the standard RHKS, and particularly the positive definite K is no longer needed. This relaxation can be critical in scenarios when such kernels are inappropriate.

With the coefficient-based representation and the group Lasso penalty, the proposed variable selection formulation can be rewritten as

$$\underset{\boldsymbol{\alpha}^{(1)},\ldots,\boldsymbol{\alpha}^{(p)}}{\operatorname{argmin}} \ \frac{1}{n(n-1)} \sum_{i,j=1}^{n} w_{ij} \Big(y_i - y_j - \sum_{l=1}^{p} \mathbf{K}_i^T \boldsymbol{\alpha}^{(l)} (x_{il} - x_{jl}) \Big)^2 + \lambda_n \sum_{l=1}^{p} \pi_l \|\boldsymbol{\alpha}^{(l)}\|_2, \tag{4}$$

where $\mathbf{K}_i = (K(\mathbf{x}_i, \mathbf{x}_1), \ldots, K(\mathbf{x}_i, \mathbf{x}_n))^T$ is the i-th column of $\mathbf{K} = (K(\mathbf{x}_i, \mathbf{x}_j))_{n\times n}$, and λ_n is a tuning parameter. The infimum operator in (3) is absorbed in the minimization in (4). Clearly, (4) simplifies the original formulation (2) from a functional space to a finite-dimensional vector space. However, the vector space is of dimension np and thus still requires an efficient large-scale optimization scheme, which will be developed in the next section.

2.3 Computing algorithm

To solve (4), we develop a block coordinate descent algorithm as in Yang and Zou (2015). First, after dropping the $\boldsymbol{\alpha}$-unrelated terms, the cost function in (4) can be simplified as

$$\underset{\boldsymbol{\alpha}}{\operatorname{argmin}} \ -\boldsymbol{\alpha}^T \mathbf{U} + \frac{1}{2} \boldsymbol{\alpha}^T \mathbf{M} \boldsymbol{\alpha} + \lambda_n \sum_{l=1}^{p} \pi_l \|\boldsymbol{\alpha}^{(l)}\|_2, \tag{5}$$

where $\boldsymbol{\alpha}^T = \big((\boldsymbol{\alpha}^{(1)})^T, \ldots, (\boldsymbol{\alpha}^{(p)})^T\big)$, $\mathbf{U} = \frac{2}{n(n-1)} \sum_{i,j=1}^{n} w_{ij} \mathbf{U}_{ij}$, $\mathbf{M} = \frac{2}{n(n-1)} \sum_{i,j=1}^{n} w_{ij} \mathbf{M}_{ij}$,

$$\begin{aligned} \mathbf{U}_{ij} &= (y_i - y_j)(\mathbf{x}_i - \mathbf{x}_j) \otimes \mathbf{K}_i, \\ \mathbf{M}_{ij} &= \Big((\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^T \Big) \otimes \big(\mathbf{K}_i \mathbf{K}_i^T \big), \end{aligned}$$

$\mathbf{K}_i$ is the i-th column of $\mathbf{K} = (K(\mathbf{x}_i, \mathbf{x}_j))_{n\times n}$, $\mathbf{I}_n$ is a n-dimensional identity matrix, and $\otimes$ denotes the kronecker product.

Then we update one $\boldsymbol{\alpha}^{(l)}$ at a time pretending others fixed, and the l-th subproblem becomes

$$\underset{\boldsymbol{\alpha}^{(l)}}{\operatorname{argmin}} \; L(\boldsymbol{\alpha}) + \lambda_n \pi_l \|\boldsymbol{\alpha}^{(l)}\|_2 = -\boldsymbol{\alpha}^T \mathbf{U} + \frac{1}{2}\boldsymbol{\alpha}^T \mathbf{M} \boldsymbol{\alpha} + \lambda_n \pi_l \|\boldsymbol{\alpha}^{(l)}\|_2,$$

To solve the subproblem, a similar approximation as in Yang and Zou (2015) can be employed, where the updated $\boldsymbol{\alpha}^{(l)}$ is obtained by solving

$$\underset{\boldsymbol{\alpha}^{(l)}}{\operatorname{argmin}} \nabla_l L(\widetilde{\boldsymbol{\alpha}})(\boldsymbol{\alpha}^{(l)} - \widetilde{\boldsymbol{\alpha}}^{(l)}) + \frac{\gamma^{(l)}}{2}(\boldsymbol{\alpha}^{(l)} - \widetilde{\boldsymbol{\alpha}}^{(l)})^T(\boldsymbol{\alpha}^{(l)} - \widetilde{\boldsymbol{\alpha}}^{(l)}) + \lambda_n \pi_l \|\boldsymbol{\alpha}^{(l)}\|_2. \tag{6}$$

Here $\widetilde{\boldsymbol{\alpha}}$ is the current estimate for $\boldsymbol{\alpha}$, $\widetilde{\boldsymbol{\alpha}}^{(l)}$ is the l-th column of $\widetilde{\boldsymbol{\alpha}}$,

$$\nabla L(\widetilde{\boldsymbol{\alpha}}) = \frac{2}{n(n-1)} \sum_{i,j=1}^{n} w_{ij}(\mathbf{M}_{ij}\widetilde{\boldsymbol{\alpha}} - \mathbf{U}_{ij}),$$

$\nabla_l L(\widetilde{\boldsymbol{\alpha}})$ denotes the l-th block vector of $\nabla L(\widetilde{\boldsymbol{\alpha}})$, and

$$\nabla_l L(\widetilde{\boldsymbol{\alpha}}) = \frac{2}{n(n-1)} \sum_{i,j=1}^{n} w_{ij} \left(\sum_{s=1}^{p} ((\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^T)_{ls} \mathbf{K}_i \mathbf{K}_i^T \widetilde{\boldsymbol{\alpha}}^{(s)} - (y_i - y_j)(x_{il} - x_{jl})\mathbf{K}_i \right),$$

where $((\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^T)_{ls}$ is the (l,s)-th entry of $(\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^T$. Furthermore, denote $\gamma^{(l)}$ the largest eigenvalue of

$$\mathbf{M}^{(l)} = \frac{2}{n(n-1)} \sum_{i,j=1}^{n} w_{ij}((\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^T)_{ll} \mathbf{K}_i \mathbf{K}_i^T,$$

which is the l-th $n \times n$ block diagonal of $\mathbf{M}$.

It is straightforward to show that (6) has an analytical solution,

$$\boldsymbol{\alpha}^{(l)} = \left(\widetilde{\boldsymbol{\alpha}}^{(l)} - \frac{\nabla_l L(\widetilde{\boldsymbol{\alpha}})}{\gamma^{(l)}} \right) \left(1 - \frac{\lambda_n \pi_l}{\|\gamma^{(l)} \widetilde{\boldsymbol{\alpha}}^{(l)} - \nabla_l L(\widetilde{\boldsymbol{\alpha}})\|_2} \right)_+. \tag{7}$$

The proposed algorithm then iteratively updates $\boldsymbol{\alpha}^{(l)}$ for $l = 1, \ldots, p, 1, \ldots$ until convergence. The algorithm is guaranteed to converge to the global minimum, since the cost function in (5) is convex and its value is decreased in each updating step. Furthermore, the computational complexity of the block coordinate descent algorithm is $O(n^2 p^2 D)$ with D being the number of iterations until convergence, which can be substantially less than the complexity of solving (5) with standard optimization packages.

3. Asymptotic theory

This section presents the asymptotic estimation and variable selection consistencies of the proposed model-free variable selection method. The estimation consistency assures that the distance between $\hat{\mathbf{g}}$ and $\mathbf{g}^*$ converges to 0 at a fast rate, and the variable selection consistency assures that the truly informative variables can be exactly recovered with probability

tending to 1. Both consistency results are established for fixed p. For simplicity, we assume only the first p_0 variables $x^{(1)}, \ldots, x^{(p_0)}$ are truly informative. The following technical assumptions are made.

Assumption A1. The support $\mathcal{X}$ is a non-degenerate compact subset of $\mathcal{R}^p$, and there exists a cosntatnt c_1 such that $\sup_{\mathbf{x}} \| \mathbf{H}^*(\mathbf{x})\|_2 \leq c_1$, where $\mathbf{H}^*(\mathbf{x}) = \nabla^2 f^*(\mathbf{x})$. Also, $\sup_{\mathbf{x}} |K(\mathbf{x}, \mathbf{x})| = 1$, and the largest eigenvalue of $\mathbf{K}$ is of order $O(n^{\psi})$ with $0 \leq \psi \leq 1$.

Assumption A2. For some constants c_2 and $\theta > 0$, the probability density $p(\mathbf{x})$ exists and satisfies

$$|p(\mathbf{x}) - p(\mathbf{u})| \leq c_2 d_X(\mathbf{x}, \mathbf{u})^{\theta}, \text{ for any } \mathbf{x}, \mathbf{u} \in \mathcal{X}, \tag{8}$$

where $d_X(\cdot, \cdot)$ is the Euclidean distance on $\mathcal{X}$.

Assumption A3. There exists some constant c_4 and c_5 such that $c_4 \leq \lim_{n\to\infty} \min_{1\leq l\leq p} \pi_l \leq \lim_{n\to\infty} \max_{1\leq l\leq p_0} \pi_l \leq c_5$ and $n^{-1/2}\lambda_n \min_{l>p_0} \pi_l \to \infty$.

Assumption A4. For any $j \leq p_0$, there exists a constant t such that $\int_{\mathcal{X}\setminus\mathcal{X}_t} \|g_j^*(\mathbf{x})\|_2 d\rho_X(\mathbf{x}) > 0$, and for any $j \geq p_0 + 1$, $g_j^*(\mathbf{x}) \equiv 0$ for any $\mathbf{x} \in \mathcal{X}$, where $\mathcal{X}_t = \{\mathbf{x} \in \mathcal{X} : d_X(\mathbf{x}, \partial\mathcal{X}) < t\}$.

In Assumption A1, the compact support is assumed for the technical simplicity, which has been often used in the literature of nonparametric models (Horowitz and Mammen, 2007; Ye and Xie, 2012). The non-degenerate $\mathcal{X}$ rules out the complete multicollinearity and thus assures the unique minimizer of (4) and the true gradient function $\mathbf{g}^*(\mathbf{x})$. And $\| \mathbf{H}^*(\mathbf{x})\|_2$ is a matrix-2 norm of $\mathbf{H}^*(\mathbf{x})$ for any given $\mathbf{x}$, denoting its largest eigenvalue. The bounded assumption on $\| \mathbf{H}^*(\mathbf{x})\|_2$ implies that $|f^*(\mathbf{u}) - f^*(\mathbf{x}) - (\mathbf{g}^*(\mathbf{x}))^T(\mathbf{u}-\mathbf{x})| \leq c_1 \| \mathbf{u}-\mathbf{x} \|_2^2$ for any $\mathbf{u}$ and $\mathbf{x}$, which is necessary to prevent the loss function from diverging at certain values. Furthermore, for the Gaussian Kernel, the assumption for its largest eigenvalue can be verified with $\psi = 1$. (Gregory et al., 2012). Assumption A2 introduces a Lipschitz condition on the density function, assuring the smoothness of the distribution of $\mathbf{x}$. Assumption A1 and A2 also imply that the probability density $p(\mathbf{x})$ is bounded. Assumption A3 provides some guideline on setting the adaptive weights, and is satisfied with $\pi_l = \|(\tilde{\boldsymbol{\alpha}}^{(l)})^T \mathbf{K}\, \tilde{\boldsymbol{\alpha}}^{(l)}\|_2^{-\gamma}$ and some positive γ. For example, the initializer $\tilde{\boldsymbol{\alpha}}^{(l)}$ can be obtained by solving (4) without the Lasso penalty and $\gamma = 3 - 2\psi$, which can be verified following Lemma 1 and Theorem 14 in Mukherjee and Zhou (2006). Other consistent estimators can also be employed to initialize the weights. Assumption A4 requires that the gradient function along a truly informative variable needs to be substantially different from 0, and that along a non-informative variable is 0 everywhere.

Lemma 1 *Let $\mathbf{g}^0$ be the minimizer of $\mathcal{E}(\mathbf{g})$ over all functionals. If Assumption A1-A2 are met, then as $\tau_n^2 \to 0$, $\mathbf{g}^0(\mathbf{x})$ converges to $\mathbf{g}^*(\mathbf{x})$ in probability, and $\mathcal{E}(\mathbf{g}^0) - 2\sigma_s^2 \to 0$.*

Lemma 1 is analogous to the Fisher consistency established for margin-based classification (Lin, 2004; Liu, 2007). It shows that the error measure $\mathcal{E}(\mathbf{g})$ in (1) is reasonable in the sense that its global minimizer well approximates the true gradient function $\mathbf{g}^*$ as τ_n^2 shrinks to 0. Note that it may not appropriate to set τ_n^2 to be exactly 0 in the gradient learning framework, but a sufficient small τ_n^2 is necessary in order to assure the estimation consistency.

Theorem 2 *Suppose Assumptions A1-A4 are met. Then there exists some constant M_0 and c_6 such that with probability at least $1 - \delta$,*

$$\mathcal{E}(\hat{\mathbf{g}}) - 2\sigma_s^2 \leq c_6 \sqrt{\log(4/\delta)} \Big(n^{-1/4} + n^{\frac{2\psi-1}{2}} \lambda_n^{-2} + \tau_n^{p+4} + n^{-\frac{1}{2(p+2)}} + n^{-\left(1-\frac{1}{2(p+2)}\right)} \lambda_n^2 \tau_n^{-4} \Big).$$

Theorem 2 establishes the estimation consistency of $\hat{\mathbf{g}}$ in terms of its estimation error $\mathcal{E}(\hat{\mathbf{g}}) - 2\sigma_s^2$, which is governed by the choice of λ_n and τ_n. Specifically, let $\lambda_n = n^{\frac{2\psi-1}{4}+\frac{1}{4(p+2)}}$ and $\tau_n = n^{-\frac{\theta}{4p(p+2)(p+4+3\theta)}}$, and we have $\mathcal{E}(\hat{\mathbf{g}}) - 2\sigma_s^2 \to 0$ in probability.

Next, let $\mathcal{A}_T = \{1, \cdots, p_0\}$ consist of all the truly informative variables, and $\widehat{\mathcal{A}} = \{j : \|\widehat{\boldsymbol{\alpha}}^{(j)}\|_2 > 0\}$ consist of all the estimated informative variables, where $\widehat{\boldsymbol{\alpha}}^{(j)}$ is the solution of (4).

Theorem 3 *Suppose all the assumptions in Theorem 2 are met. Let* $\lambda_n = n^{\frac{2\psi-1}{4}+\frac{1}{4(p+2)}}$ *and* $\tau_n = n^{-\frac{\theta}{4p(p+2)(p+4+3\theta)}}$*, then* $P(\widehat{\mathcal{A}} = \mathcal{A}^*) \to 1$ *as* n *diverges.*

Theorem 3 assures that the selected variables by the proposed method can exactly recover the true active set with probability tending to 1. In fact, $P(\widehat{\mathcal{A}} = \mathcal{A}^*)$ can be upper bounded by $1 - O(n^{-1/4})$ with an appropriate choice of δ. This result is particularly interesting given the fact that it is established without assuming any explicit model assumptions.

4. Numerical experiments

This section examines the effectiveness of the proposed model-free variable selection method, and compares it against some popular model based methods in literature, including variable selection with the additive model (Xue, 2009), Cosso (Lin and Zhang, 2006), sparse gradient learning (Ye and Xie, 2012) and multivariate adaptive regression splines (Friedman, 1991), denoted as MF, Add, Cosso, SGL and Mars respectively. In all the experiments, the Gaussian kernel $K(\mathbf{x}, \mathbf{u}) = e^{-\|\mathbf{x}-\mathbf{u}\|_2^2/2\sigma_n^2}$ is used, where the scalar parameters σ_n^2 and τ_n^2 in $w(\mathbf{x}, \mathbf{u})$ are set as the median over the pairwise distances among all the sample points (Mukherjee and Zhou, 2006). Other tuning parameters in these competitors, such as the number of knots in Xue (2009), are set as the default values in the available R and Matlab packages.

The tuning parameters in each method are determined by the stability-based selection criterion in Sun et al. (2013). The idea is to conduct a cross-validation-like scheme, and measure the stability as the agreement between two estimated active sets. It randomly splits the training set into two subsets, applies the candidate variable selection method to each subset, and obtains two estimated active sets, denoted as $\widehat{\mathcal{A}}_{1b}$ and $\widehat{\mathcal{A}}_{2b}$. The variable selection stability can approximated by $s_\lambda = \frac{1}{B}\sum_{b=1}^{B} \kappa(\widehat{\mathcal{A}}_{1b}, \widehat{\mathcal{A}}_{2b})$, where B is the number of splitting in the cross validation scheme, and $\kappa(\cdot, \cdot)$ is the standard Cohen's kappa statistic measuring the agreement between two sets. The tuning parameter λ is then selected as the one maximizing s_λ. Finally, the performance of all methods is evaluated by a number of measures regarding the variable selection accuracy.

4.1 Simulated examples

Two simulated examples are considered. The first example was used in Xue (2009) and Huang et al. (2010), where the true regression model is an additive model. The second example modifies the generating scheme of the first one and includes interaction terms.

Example 1: First generate p-dimensional variables $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^T$ with $x_{ij} = \frac{W_{ij}+\eta U_i}{1+\eta}$, where W_{ij} and U_i are independently from $U(-0.5, 0.5)$, for $i = 1, \ldots, n$ and

$j = 1, \ldots, p$. When $\eta = 0$ all variables are independent, whereas when $\eta = 1$ correlation presents among the variables. Next, set $f^*(\mathbf{x}_i) = 5f_1(x_{i1}) + 3f_2(x_{i2}) + 4f_3(x_{i3}) + 6f_4(x_{i4})$, with $f_1(u) = u$, $f_2(u) = (2u-1)^2$, $f_3(u) = \frac{\sin(\pi u)}{2-\sin(\pi u)}$, and $f_4(u) = 0.1\sin(\pi u) + 0.2\cos(\pi u) + 0.3\sin^2(\pi u) + 0.4\cos^3(\pi u) + 0.5\sin^3(\pi u)$. Finally, generate y_i by $y_i = f(x_i) + \epsilon_i$ with $\epsilon_i \sim N(0, 1.31^2)$. Clearly, the true underlying regression model is additive.

Example 2: The generating scheme is similar as Example 1, except that $f^*(\mathbf{x}_i) = (2x_{i1} - 1)(2x_{i2} - 1)$, W_{ij} and U_i are independently from $N(0, 1)$ and $\epsilon_i \sim N(0, 1)$. It is clear that the underlying regression model includes interaction terms, and thus the additive model assumption is no longer valid.

For each example, different scenarios are considered with $\eta = 0$ or 1, and $(n, p) = (100, 10)$, $(100, 20)$ or $(200, 50)$. Each scenario is replicated 50 times, and the averaged performance measures are summarized in Tables 1 and 2. Specifically, Size represents the averaged number of selected informative variables, TP represents the number of truly informative variables selected, FP represents the number of truly non-informative variables selected and C, U, O are the times of correct-fitting, under-fitting and over-fitting, respectively.

It is evident that the proposed MF method delivers superior selection performance against the other three competitors. In Table 1 where the true model is indeed additive, MF performs similarly as Add and SGL, whereas Cosso and Mars appear more likely to overfit. In Table 2 where the true model consists of interaction terms, the performance of MF becomes competitive, but Add tends to under-fit more frequently, and Cosso, Mars and SGL tend to overfit as the dimension increases. Furthermore, in both examples with $\eta = 1$, it is clear that the correlation among variables increases the difficulty of selecting the truly informative variables, yet the proposed MF method still outperforms its competitors. Furthermore, it is also noted that the estimation accuracy of MF outperforms SGL, but it is omitted here as only MF and SGL estimate the gradient function $\mathbf{g}$, whereas Add, Cosso and Mars estimate the regression function f.

4.2 Real examples

The proposed model-free variable selection method is applied to three real data examples, the Boston housing data, the Ozone concentration data, and the digit recognition data, all of which are publicly available. The Boston housing data concerns the median value of owner-occupied homes in each of the 506 census tracts in the Boston Standard Metropolitan Statistical Area in 1970. It consists of 13 variables, including per capita crime rate by town (CRIM), proportion of residential land zoned for lots over 25,000 square feet (ZN), proportion of non-retail business acres per town (INDUS), Charles River dummy variable (CHAS), nitric oxides concentration (NOX), average number of rooms per dwelling (RM), proportion of owner-occupied units built prior to 1940 (AGE), weighted distances to five Boston employment centers (DIS), index of accessibility to radial highways (RAD), full-value property-tax rate per \$10000 (TAX), pupil-teacher ratio by town (PTRATIO), the proportion of blacks by town (B), lower status of the population (LSTAT), which may affect the housing price. The Ozone concentration data concerns the daily measurements of Ozone concentration in Los Angeles basin in 330 days. The Ozone concertration may be influenced by 11 meteorological quantities, such as month (M), day of month (DM), day of week

(n, p, η)	Method	Size	TP	FP	C	U	O
(100,10,0)	MF	4.000	**4.000**	**0.000**	50	0	0
	Add	4.080	**4.000**	0.080	46	0	4
	Cosso	4.200	3.960	0.240	41	1	8
	SGL	4.020	3.600	0.420	12	20	18
	Mars	5.200	**4.000**	1.200	12	0	38
(100,20,0)	MF	4.040	**4.000**	0.300	35	0	15
	Add	4.040	**4.000**	**0.040**	48	0	2
	Cosso	4.280	**4.000**	0.280	40	0	10
	SGL	4.220	3.620	0.600	16	18	16
	Mars	6.000	**4.000**	2.000	10	0	40
(200,50,0)	MF	4.500	**4.000**	0.500	39	0	11
	Add	5.200	**4.000**	1.200	30	0	20
	Cosso	5.600	**4.000**	1.600	31	0	19
	SGL	3.600	3.400	**0.200**	12	30	8
	Mars	12.400	**4.000**	8.400	0	0	50
(100,10,1)	MF	4.160	3.800	0.360	33	10	7
	Add	3.960	3.960	**0.000**	48	2	0
	Cosso	4.200	3.760	0.440	24	8	18
	SGL	4.200	3.600	0.600	24	12	14
	Mars	5.240	**4.000**	1.240	16	0	34
(100,20,1)	MF	4.080	3.800	0.280	30	10	10
	Add	3.960	3.840	**0.120**	36	8	6
	Cosso	3.960	3.800	0.160	37	5	8
	SGL	4.020	3.500	0.520	10	20	20
	Mars	6.240	**4.000**	2.240	8	0	42
(200,50,1)	MF	4.700	3.900	**0.800**	35	5	10
	Add	5.600	**4.000**	1.600	21	0	29
	Cosso	5.000	3.900	1.100	26	4	20
	SGL	3.720	3.500	0.220	20	20	10
	Mars	13.220	**4.000**	9.220	0	0	50

Table 1: The averaged performance measures of various variable selection methods in Example 1.

(DW), Vandenburg 500 millibar height (VDHT), wind speed (WDSP), humidity (HMDT), temperature at Sandburg (SBTH), inversion base height (IBHT), Daggett pressure gradient (DGPG), inversion base temperature (IBTP) and visibility (VSTY). These two datasets have been widely analyzed in literature, including Breiman and Friedman (1985), Xue (2009), and Lin and Zhang (2006). For the digit recognition data, each digit is described by a 8×8 gray-scale image with each entry ranging from 0 to 16. We focus on digits 3 and 5 due to their similarity, and the resultant dataset consists of 365 observations and 64 attributes.

(n, p, η)	Method	Size	TP	FP	C	U	O
(100,10,0)	MF	1.960	1.920	**0.080**	43	3	4
	Add	2.140	1.760	0.380	25	9	16
	Cosso	2.920	1.920	1.000	15	3	32
	SGL	2.320	1.920	0.400	30	4	16
	Mars	4.000	**2.000**	2.000	8	0	42
(100,20,0)	MF	2.100	**2.000**	**0.100**	45	0	5
	Add	2.200	1.800	0.400	30	8	12
	Cosso	4.320	1.920	2.400	10	3	37
	SGL	2.220	**2.000**	0.220	42	0	8
	Mars	4.240	1.920	2.320	14	2	34
(200,50,0)	MF	2.100	**2.000**	0.100	45	0	5
	Add	2.920	1.920	1.000	28	2	20
	Cosso	2.200	1.800	0.400	25	10	15
	SGL	1.800	1.800	**0.000**	42	8	0
	Mars	8.200	**2.000**	6.200	0	0	50
(100,10,1)	MF	2.160	**2.000**	**0.160**	42	0	8
	Add	2.360	1.560	0.800	16	12	22
	Cosso	3.600	**2.000**	1.600	10	0	40
	SGL	2.300	**2.000**	0.300	35	0	15
	Mars	4.240	**2.000**	2.240	10	0	40
(100,20,1)	MF	2.040	1.920	**0.120**	40	4	6
	Add	2.460	1.920	0.540	34	4	12
	Cosso	3.240	1.800	1.440	9	10	31
	SGL	2.120	1.800	0.320	28	10	12
	Mars	6.740	**2.000**	4.740	0	0	50
(200,50,1)	MF	2.160	**1.960**	**0.200**	40	2	8
	Add	16.200	1.800	14.400	17	9	24
	Cosso	2.340	1.800	0.540	28	10	22
	SGL	2.460	**1.960**	0.500	23	2	25
	Mars	8.160	**1.960**	6.240	0	2	48

Table 2: The averaged performance measures of various variable selection methods in Example 2.

In our analysis, all the variables and responses are standardized and the selected variables are summarized. The selected informative variables by MF, Add, Cosso and Mars are summarized in Tables 3 and 4. As the truly informative variables are unknown in real examples, averaged prediction errors with the selected variables are also reported to compare the performance. To compute the averaged prediction error, each dataset is randomly split into two parts: m observations for testing and the remaining for training. Specifically, $m = 30$ for the Boston housing data, $m = 50$ for the Ozone concentration data, and $m = 35$ for

digit recognition data. Each example is replicated 100 times, and the averaged prediction errors by MF, Add, Cosso and Mars are summarized in Tables 3-5.

Variables	MF	Add	Cosso	SGL	Mars
CRIM	-	√	√	-	√
ZN	-	-	-	-	-
INDUS	-	-	-	-	-
CHAS	-	-	-	-	-
NOX	-	√	-	-	√
RM	√	√	√	√	√
AGE	-	-	-	-	-
DIS	-	√	-	-	√
RAD	-	-	-	-	√
TAX	-	√	-	-	-
PTRATIO	-	√	-	-	√
B	-	-	-	-	√
LSTAT	√	√	√	√	√
Pred. Err.	**1.774(0.0931)**	1.780(0.0916)	1.797(0.0924)	**1.774(0.0931)**	1.956(0.0939)

Table 3: The selected variables as well as the corresponding prediction errors by various selection methods in the Boston housing dataset.

Variables	MF	Add	Cosso	SGL	Mars
M	√	√	√	√	√
DM	-	-	-	-	-
DW	-	-	-	-	-
VDHT	-	√	-	√	√
WDSP	-	-	-	-	√
HMDT	√	-	√	√	√
SBTH	√	√	√	√	√
IBHT	-	√	-	√	√
DGPG	-	√	-	-	√
IBTP	√	-	√	√	√
VSTY	-	√	-	√	√
Pred. Err.	**1.768(0.0416)**	1.769(0.0425)	**1.768(0.0416)**	1.776(0.0426)	1.784(0.0463)

Table 4: The selected variables as well as the corresponding prediction errors by various selection methods in the Ozone concentration dataset.

	MF	Add	Cosso	SGL	Mars
No. of variables	2	48	8	4	18
Prediction error	**1.857(0.0316)**	1.871(0.0310)	1.878(0.0314)	1.875(0.0324)	1.879(0.0310)

Table 5: The number of selected variables and the prediction errors by various selection methods in the digit recognition dataset.

For the Boston housing data, MF and SGL select two informative variables, RM and LSTAT, whereas Add, Cosso and Mars tend to select more variables. However, the corresponding prediction errors of Add, Cosso and Mars appear to be larger than that of MF and SGL, implying that the additional selected variables by Add, Cosso and Mars may hinder the prediction performance. For the Ozone concentration data, both MF and Cosso

select four variables but Add and Mars select more. One discrepancy is the variable IBTP, which is selected by MF, Cosso, SGL and Mars but not by Add. As claimed in Gregory et al. (2012), M, SBTH and IBTP are three most important meteorological variables related to Ozone concentration as all of them describe the temperature changes. Meanwhile, MF and Cosso show smaller prediction error than SGL and Mars, which implies that SGL and Mars may include some non-informative variables. Figure 1 displays scatter plots of the responses against the selected variables by MF in the Boston housing data and the Ozone concentration data. It is clear that all the selected variables show moderate to strong relationship with the responses. For digit recognition data, MF selects much less variables than the other competitors and provides smaller prediction error. Figure 2 shows some randomly selected digits of 3 and 5 and the two selected informative variables, where the left informative variable is always contained in digit 5 and the right one is always contained in digit 3.

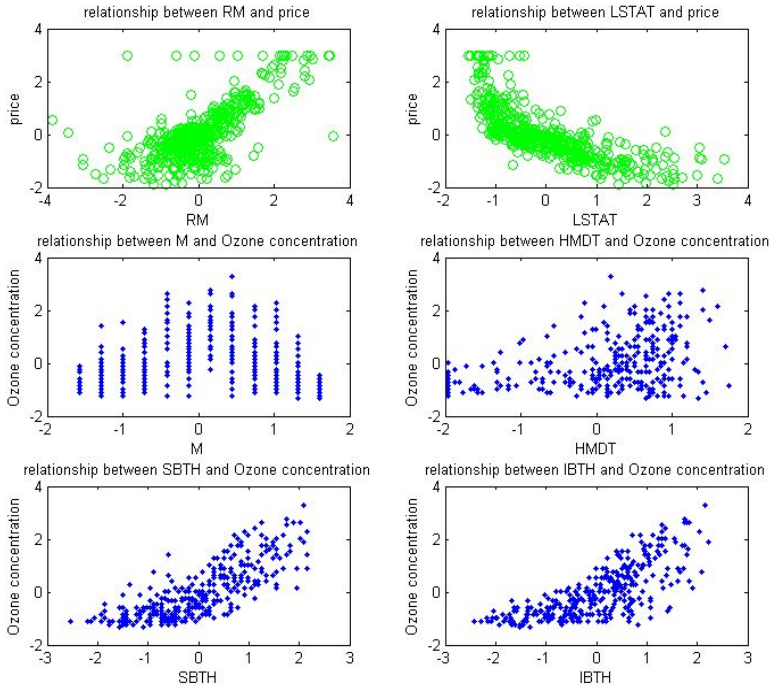

Figure 1: The scatter plots of the responses and the selected variables by MF in the Boston housing data (first row) and the Ozone concentration data (second and third rows).

Figure 2: Some randomly selected digit 3, digit 5 and the two selected informative variables.

5. Summary

This article proposes a model-free variable selection method, which is in sharp contrast to most existing methods relying on various model assumptions. The proposed method makes use of the natural connection between informative variables and sparse gradients, and formulates the variable selection task in a flexible framework of learning gradients. Additionally, we introduce a coefficient-based representation to facilitate variable selection in the learning framework. A block-wise coordinate decent algorithm is developed to make efficient computation for large-scale problems feasible. More importantly, we establish the estimation and variable selection consistencies of the proposed method without assuming any restrictive model assumption. The effectiveness of the proposed method is also supported by numerical experiments on simulated and real examples. It is worth pointing out that the computational cost of the proposed method can be expensive, as it allows for a more flexible modeling framework in RKHS. The extension of the proposed method to diverging dimension is also challenging as a model-free framework with diverging dimension can be too flexible to analyze. One possible remedy is to pre-screen the non-informative variables via some model-free screening methods (Li et al., 2012) to shrink the size of candidate variables.

Acknowledgments

SL's research is partially supported by NSFC-11301421, and JW's research is partially supported by HK GRF-11302615 and CityU SRG-7004244. The authors also thank the Action Editor and two referees for their constructive comments and suggestions, which have led to a significantly improved paper.

Appendix A. technical proofs

Proof of Lemma 1: First, note that under Assumption A1 and A2, the probability density $p(\mathbf{x})$ is bounded, and thus there exists some constant c_7 such that $\sup_{\mathbf{x}\in\mathcal{X}} p(\mathbf{x}) \leq c_7$. Moreover, denote $\mathcal{X}_t = \{\mathbf{x} \in \mathcal{X} : d_X(\mathbf{x}, \partial\mathcal{X}) < t\}$, then we have $\rho_X(\mathcal{X}_t) \leq c_8 t$ for any t given a constant c_8, where $\partial\mathcal{X}$ is the boundary of the compact support $\mathcal{X}$, ρ_X is the marginal distribution and $d_X(\mathbf{x}, \partial\mathcal{X}) = \inf_{\mathbf{u}\in\partial\mathcal{X}} d_X(\mathbf{x}, \mathbf{u})$.

Since $\mathbf{g}^0$ is the minimizer of $\mathcal{E}(\mathbf{g})$, the functional derivative of $\mathcal{E}(\mathbf{g})$ at $\mathbf{g}^0$ yields that for any arbitrary function vector $\delta(\mathbf{x})$,

$$\int_{\mathcal{X}}\int_{\mathcal{X}} w(\mathbf{x},\mathbf{u})\big(f(\mathbf{x}) - f(\mathbf{u}) + \mathbf{g}^0(\mathbf{x})^T(\mathbf{u}-\mathbf{x})\big)(\mathbf{u}-\mathbf{x})^T\delta(\mathbf{x})d\rho_X(\mathbf{u})d\rho_X(\mathbf{x}) = \mathbf{0}_p,$$

where $\mathbf{0}_p$ is a p-dimensional vector with all zeros. As the above equality is true for any $\delta(\mathbf{x})$, it implies that for any given $\mathbf{x}$,

$$\int_{\mathcal{X}} w(\mathbf{x},\mathbf{u})\left(f(\mathbf{x}) - f(\mathbf{u}) + \mathbf{g}^0(\mathbf{x})^T(\mathbf{u}-\mathbf{x})\right)(\mathbf{u}-\mathbf{x})d\rho_X(\mathbf{u}) = \mathbf{0}_p.$$

For simplicity, denote $\mathbf{M}(\mathbf{x}) = \int_{\mathcal{X}} w(\mathbf{x},\mathbf{u})(\mathbf{u}-\mathbf{x})(\mathbf{u}-\mathbf{x})^T d\rho_X(\mathbf{u})$ a function matrix, and $\mathbf{d}(\mathbf{x}) = \int_{\mathcal{X}} w(\mathbf{x},\mathbf{u})(\mathbf{u}-\mathbf{x})\left(f(\mathbf{u}) - f(\mathbf{x})\right)d\rho_X(\mathbf{u})$ a function vector. Then $\mathbf{M}(\mathbf{x})\,\mathbf{g}^0(\mathbf{x}) - \mathbf{d}(\mathbf{x}) = 0$ for any given $\mathbf{x}$. Let $\mathcal{X}_\tau = \{\mathbf{x} : d_X(\mathbf{x}, \partial\mathcal{X}) \geq \tau_n, p(\mathbf{x}) \geq c_2\tau_n^\theta + \tau_n^{1/2}\}$, then by Assumption A2,

$$P(\mathcal{X}_\tau^c) \leq P\left(d_X(\mathbf{x}, \partial\mathcal{X}) < \tau_n\right) + P\left(p(\mathbf{x}) < c_2\tau_n^\theta + \tau_n^{1/2}\right) \leq c_8\tau_n + (c_2\tau_n^\theta + \tau_n^{1/2})|\mathcal{X}|,$$

where $|\mathcal{X}|$ denotes the Lebesgue measure of $\mathcal{X}$. For any $\mathbf{x} \in \mathcal{X}_\tau$,

$$\begin{aligned}
\mathbf{M}(\mathbf{x}) &= \int w(\mathbf{x},\mathbf{u})(\mathbf{u}-\mathbf{x})(\mathbf{u}-\mathbf{x})^T p(\mathbf{u})d\,\mathbf{u} \\
&\geq \tau_n^{1/2}\int_{d_X(\mathbf{x},\mathbf{u})<\tau_n} e^{-\frac{\|\,\mathbf{x}-\mathbf{u}\,\|_2^2}{2\tau_n^2}}(\mathbf{u}-\mathbf{x})(\mathbf{u}-\mathbf{x})^T d\,\mathbf{u} = \tau_n^{p+5/2}\int_{\|\,\mathbf{t}\,\|_2<1} e^{-\frac{\|\,\mathbf{t}\,\|_2^2}{2}}\,\mathbf{t}\,\mathbf{t}^T d\,\mathbf{t},
\end{aligned}$$

where $\mathbf{t} = (\mathbf{u}-\mathbf{x})/\tau_n$. The inequality follows from Assumption A2 and the fact that $p(\mathbf{u}) \geq p(\mathbf{x}) - |p(\mathbf{u}) - p(\mathbf{x})| \geq p(\mathbf{x}) - c_2 d(\mathbf{x},\mathbf{u})^\theta \geq \tau_n^{1/2}$ on $\mathcal{X}_\tau$. As the support $\mathcal{X}$ is non-degenerate by assumption A1, $\int_{\|\,\mathbf{t}\,\|_2<1} e^{-\frac{\|\,\mathbf{t}\,\|_2^2}{2}}\,\mathbf{t}\,\mathbf{t}^T d\,\mathbf{t}$ is always positive definite. So its smallest eigenvalue, denoted as ϕ_{min}, is positive, and thus the smallest eigenvalue of $\mathbf{M}(\mathbf{x})$ must be larger than $\phi_{min}\tau_n^{p+5/2}$, which is also positive.

As $\mathbf{M}(\mathbf{x})$ is invertible for any $\mathbf{x} \in \mathcal{X}_\tau$, we have $\mathbf{g}^0(\mathbf{x}) = \mathbf{M}(\mathbf{x})^{-1}\mathbf{d}(\mathbf{x})$, and thus

$$\|\mathbf{g}^0(\mathbf{x}) - \mathbf{g}^*(\mathbf{x})\|_2 \leq \|(\mathbf{M}(\mathbf{x}))^{-1}\|_2\|\mathbf{d}(\mathbf{x}) - \mathbf{M}(\mathbf{x})\mathbf{g}^*(\mathbf{x})\|_2.$$

Furthermore,

$$\begin{aligned}
\|\mathbf{d}(\mathbf{x}) - \mathbf{M}(\mathbf{x})\mathbf{g}^*(\mathbf{x})\|_2 &= \int_{\mathcal{X}} w(\mathbf{x},\mathbf{u})(\mathbf{u}-\mathbf{x})\left(f(\mathbf{u}) - f(\mathbf{x}) - \mathbf{g}^*(\mathbf{x})^T(\mathbf{u}-\mathbf{x})\right)d\rho_X(\mathbf{u}) \\
&\leq \int \Big|w(\mathbf{x},\mathbf{u})(\mathbf{u}-\mathbf{x})\left(f(\mathbf{u}) - f(\mathbf{x}) - \mathbf{g}^*(\mathbf{x})^T(\mathbf{u}-\mathbf{x})\right)\Big|p(\mathbf{u})d\,\mathbf{u} \\
&\leq c_1c_8\int w(\mathbf{x},\mathbf{u})\|\,\mathbf{u}-\mathbf{x}\,\|_2^3 d\,\mathbf{u} \leq c_1c_8\tau_n^{p+3}\int e^{-\frac{\|\,\mathbf{t}\,\|_2^2}{2}}\|\,\mathbf{t}\,\|_2^3 d\,\mathbf{t}\,.
\end{aligned}$$

Therefore, for any $\mathbf{x} \in \mathcal{X}_\tau$,

$$\|\mathbf{g}^0(\mathbf{x}) - \mathbf{g}^*(\mathbf{x})\|_2 \leq \|\mathbf{M}(\mathbf{x})^{-1}\|_2 \|\mathbf{d}(\mathbf{x}) - \mathbf{M}(\mathbf{x})\mathbf{g}^*(\mathbf{x})\|_2 \leq \frac{c_1 c_8 \tau_n^{1/2}}{\phi_{min}} \int e^{-\frac{\|\mathbf{t}\|_2^2}{2}} \|\mathbf{t}\|_2^3 d\mathbf{t},$$

which converges to 0 for any $\mathbf{x} \in \mathcal{X}_\tau$. Since $P(\mathbf{x} \in \mathcal{X}_\tau) \to 1$ as $\tau_n \to 0$, the desired result follows immediately.

Next, as $\mathcal{E}(\mathbf{g}) - 2\sigma_s^2 > 0$ for any $\mathbf{g}$, we have $0 \leq \mathcal{E}(\mathbf{g}^0) - 2\sigma_s^2 \leq \mathcal{E}(\mathbf{g}^*) - 2\sigma_s^2$. By Proposition 3 in Ye and Xie (2012), $\mathcal{E}(\mathbf{g}^*) - 2\sigma_s^2 \leq O(\tau_n^{p+4}) \to 0$ as $\tau_n \to 0$. Therefore, $\mathcal{E}(\mathbf{g}^0) - 2\sigma_s^2 \to 0$ as $\tau_n \to 0$. ■

To proceed further, we note that the proof of Theorem 2 is substantially different from conventional error analysis as in Mukerjee and Zhou (2006) and Ye and Xie (2012). In our setting, we consider the coefficient-based space $\mathcal{H}_\mathbf{z} = \{\mathbf{g} : \mathbf{g}(x) = \sum_{i=1}^n a_i K(\mathbf{x}_i, \mathbf{x}),\ a_i \in \mathcal{R}\}$ as the candidate functional space, which depends on $\{\mathbf{x}_i\}_{i=1}^n$. One difficulty arises is that $\mathbf{g}^*$ may not be contained in $\mathcal{H}_\mathbf{z}^p$ and thus $J(\mathbf{g}^*)$ can not be defined. To circumvent this difficulty, we introduce an intermediate learning algorithm as a bridge for the error analysis, so that standard empirical process and approximation theories can be used.

Define a vector-valued functional space as $\mathcal{H}_K^p = \{\mathbf{g} = (g_1, ..., g_p)^T, g_j \in \mathcal{H}_K\}$, and $\mathcal{H}_\mathbf{z}^p = \{\mathbf{g} = (g_1, ..., g_p)^T, g_j \in \mathcal{H}_\mathbf{z}\}$. Furthermore, denote the empirical error used in our algorithm as

$$\mathcal{E}_\mathbf{z}(\mathbf{g}) = \frac{1}{n(n-1)} \sum_{i,j=1}^n \omega_{ij} \Big(y_j - y_i - \mathbf{g}(\mathbf{x}_i)^T (\mathbf{x}_j - \mathbf{x}_i) \Big)^2.$$

Clearly, $E(\mathcal{E}_\mathbf{z}(\mathbf{g})) = \mathcal{E}(\mathbf{g})$ for any $\mathbf{g} \in \mathcal{H}_K^p$.

In order to establish the consistency results, we introduce an intermediate learning algorithm,

$$\bar{\mathbf{g}} = \underset{\mathbf{g} \in \mathcal{H}_K^p}{\operatorname{argmin}} \quad \frac{1}{n(n-1)} \sum_{i,j=1}^n w_{ij} \Big(y_i - y_j - \mathbf{g}(\mathbf{x}_i)(\mathbf{x}_i - \mathbf{x}_j) \Big)^2 + \rho_n \sum_{l=1}^p \pi_l \|g_l\|_K^2, \tag{9}$$

where $\rho_n = n^{-\eta}$ with $\eta = \frac{1}{4(p+2)}$. Note that (9) is a weighted version of the original gradient learning in Mukherjee and Zhou (2006). By the representor theorem, each element of $\bar{\mathbf{g}}$ in (9) has a closed solution with the form

$$\bar{g}_l = \sum_{t=1}^n \bar{\alpha}_t^l K(\mathbf{x}, \mathbf{x}_t), \quad \text{for} \quad l = 1, ..., p.$$

Denote $\bar{\boldsymbol{\alpha}}^l = (\bar{\alpha}_1^l, ..., \bar{\alpha}_n^l)^T$ satisfies the linear system

$$\rho_n \pi_l \mathbf{K}\, \bar{\boldsymbol{\alpha}}^l + \frac{1}{n(n-1)} \sum_{i,j=1}^n \omega_{ij} \Big(y_j - y_i - \bar{\mathbf{g}}(\mathbf{x}_i)^T (\mathbf{x}_j - \mathbf{x}_i) \Big) (\mathbf{K})_i^T [\mathbf{x}_j - \mathbf{x}_i]_l = 0, \tag{10}$$

where $(\mathbf{K})_i$ represents the i-th row of $\mathbf{K}$. Without loss of generality, we assume that $\mathbf{K}$ is invertible. In this case, we can solve for $\bar{\alpha}_t^l$ as follows:

$$\rho_n \pi_l \bar{\alpha}_t^l = -\frac{1}{n(n-1)} \sum_{j=1}^n \omega_{tj} \Big(y_j - y_t - \bar{\mathbf{g}}(\mathbf{x}_t)^T (\mathbf{x}_j - \mathbf{x}_t) \Big) [\mathbf{x}_j - \mathbf{x}_t]_l. \tag{11}$$

With these preparations, we are now in the position to decompose the excess error as follows.

Proposition 4 *Let* $\varphi_0(\mathbf{z}) = \mathcal{E}_{\mathbf{z}}(\mathbf{g}^*) - \mathcal{E}(\mathbf{g}^*)$ *and* $\varphi_1(\mathbf{z}) = \mathcal{E}(\hat{\mathbf{g}}) - \mathcal{E}_{\mathbf{z}}(\hat{\mathbf{g}})$. *Then the following inequality holds for any* $0 < \varepsilon \leq 1$,

$$\mathcal{E}(\hat{\mathbf{g}}) + \lambda_n \sum_{l=1}^{p} \pi_l J(\hat{g}_l) \leq \varphi_1(\mathbf{z}) + 2\varphi_0(\mathbf{z}) + \Lambda_n(\varepsilon, \rho, K),$$

where

$$\Lambda_n(\varepsilon, \rho, K) = (1+\varepsilon)\mathcal{E}(\mathbf{g}^*) + \sum_{l=1}^{p_0} \rho_n \pi_l (\|g_l^*\|_K^2 - \|\bar{g}_l\|_K^2) + c_n^2/\varepsilon, \tag{12}$$

with $c_n = \frac{c_x p \lambda_n}{\rho_n \sqrt{n-1}}$ *and* $c_x \geq \sup_{\mathbf{x}} \|\mathbf{x}\|$. *In the literature of statistical learning theory,* $\varphi_0(\mathbf{z})$, $\varphi_1(\mathbf{z})$ *are called the sample error and* $\Lambda(\lambda_n)$ *is the approximation error.*

Proof of Proposition 4: First of all, by the Hölder inequality, it follows from 11 that:

$$J(\bar{g}_l) \leq \frac{c_x}{\rho_n \pi_l \sqrt{n-1}} \sqrt{\mathcal{E}_{\mathbf{z}}(\bar{\mathbf{g}})}, \quad l = 1, ..., p. \tag{13}$$

The above inequality in connection with the definition of $\hat{\mathbf{g}}$ yields that

$$\begin{aligned}
\mathcal{E}(\hat{\mathbf{g}}) + \lambda_n \sum_{l=1}^{p} \pi_l J(\hat{g}_l) &= \mathcal{E}(\hat{\mathbf{g}}) - \mathcal{E}_{\mathbf{z}}(\hat{\mathbf{g}}) + \mathcal{E}_{\mathbf{z}}(\hat{\mathbf{g}}) + \lambda_n \sum_{l=1}^{p} \pi_l J(\hat{g}_l) \\
&\leq \mathcal{E}(\hat{\mathbf{g}}) - \mathcal{E}_{\mathbf{z}}(\hat{\mathbf{g}}) + \mathcal{E}_{\mathbf{z}}(\bar{\mathbf{g}}) + \lambda_n \sum_{l=1}^{p} \pi_l J(\bar{g}_l) \\
&\leq \mathcal{E}(\hat{\mathbf{g}}) - \mathcal{E}_{\mathbf{z}}(\hat{\mathbf{g}}) + \mathcal{E}_{\mathbf{z}}(\bar{\mathbf{g}}) + \Big(\frac{c_x \lambda_n}{\sqrt{n-1}} \sum_{l=1}^{p} \frac{1}{\rho_n}\Big) \sqrt{\mathcal{E}_{\mathbf{z}}(\bar{\mathbf{g}})} \\
&\leq \mathcal{E}(\hat{\mathbf{g}}) - \mathcal{E}_{\mathbf{z}}(\hat{\mathbf{g}}) + (1+\varepsilon)\mathcal{E}_{\mathbf{z}}(\bar{\mathbf{g}}) + c_n^2/\varepsilon \\
&\leq \mathcal{E}(\hat{\mathbf{g}}) - \mathcal{E}_{\mathbf{z}}(\hat{\mathbf{g}}) + (1+\varepsilon)\mathcal{E}_{\mathbf{z}}(\mathbf{g}^*) + 2\sum_{l=1}^{p} \rho_n \pi_l (\|g_l^*\|_K^2 - \|\bar{g}_l\|_K^2) + c_n^2/\varepsilon \\
&\leq \mathcal{E}(\hat{\mathbf{g}}) - \mathcal{E}_{\mathbf{z}}(\hat{\mathbf{g}}) + (1+\varepsilon)\mathcal{E}_{\mathbf{z}}(\mathbf{g}^*) + 2\sum_{l=1}^{p_0} \rho_n \pi_l (\|g_l^*\|_K^2 - \|\bar{g}_l\|_K^2) + c_n^2/\varepsilon,
\end{aligned}$$

where the first inequality follows from the definition of $\hat{\mathbf{g}}$, the second inequality is derived based on 13, the third inequality follows from the fact $\sqrt{xy} \leq \frac{\varepsilon x + y/\varepsilon}{2}$ for any $\varepsilon > 0$, the fourth inequality follows from the definition of $\bar{\mathbf{g}}$, and the last inequality is due to the assumption that $g_l^* = 0$ for any $l > p_0$. ■

Next, For any given value $R > 0$, define the functional subspace with bounded $J(\mathbf{g})$ as

$$\mathcal{H}_R^p = \{\mathbf{g} \in \mathcal{H}_{\mathbf{z}}^p, \text{ with } J(\mathbf{g}) \leq R\},$$

and

$$\mathcal{S}(R,\lambda_n) = \sup_{\mathbf{g}\in\mathcal{H}_R^p} |\mathcal{E}(\mathbf{g}) - \mathcal{E}_{\mathbf{z}}(\mathbf{g})|.$$

Then the quantity $\mathcal{S}(R,\lambda_n)$ can be bounded using the McDiarmid's inequality (McDiarmid, 1989).

Lemma 5 *(McDiarmid's Inequality) Let $Z_1,...,Z_n$ be independent random variables taking values in a set $\mathcal{Z}$, and assume that $\mathbf{f}:\mathcal{Z}^n\to\mathbb{R}$ satisfies*

$$\sup_{z_1,...,z_n,z_i'\in\mathcal{Z}} |\mathbf{f}(z_1,...,z_n) - \mathbf{f}(z_1,...,z_i',...,z_n)| \le C_i,$$

for every $i\in\{1,2,...,n\}$. Then, for every $t>0$,

$$\mathbb{P}\{|\mathbf{f}(z_1,...,z_n) - \mathbb{E}\left(\mathbf{f}(z_1,...,z_n)\right)| \ge t\} \le 2\exp\left(-\frac{2t^2}{\sum_{i=1}^n C_i^2}\right).$$

This result implies that, as soon as one has a function of n independent random variables, whose variation is bounded when only one variable is modified, the function will satisfy a Hoeffding-type inequality.

Lemma 6 *If $|y|\le M_n$ and Assumptions A1-A3 hold, then for any constant $R>0$ and $\varepsilon>0$, there holds*

$$P(|\mathcal{S}(R,\lambda_n) - \mathbb{E}(\mathcal{S}(R,\lambda_n))| \ge \varepsilon) \le 2\exp\left(-\frac{n\varepsilon^2}{8(M_n + \frac{c_x n^{\psi/2}R}{c_4\lambda_n})^4}\right).$$

In addition, there exists a constant c_5, such that

$$P(|\mathcal{E}_{\mathbf{z}}(\mathbf{g}^*) - \mathcal{E}(\mathbf{g}^*)| \ge \varepsilon) \le 2\exp\left(-\frac{n\varepsilon^2}{8(M_n + c_x\sum_{l=1}^p \|g_l^*\|_K)^4}\right). \tag{14}$$

Proof of Lemma 6: It suffices to verify the conditions required by the McDiarmid's inequality. For this purpose, we define $(\mathbf{x}',y')$ as a sample point drawn from the distribution $\rho(\mathbf{x},y)$ and independent of $(\mathbf{x}_i,y_i)$. Denote by $\mathbf{z}'$ the modified training sample which is the same as $\mathbf{z}$ except that the i-th observation $(\mathbf{x}_i,y_i)$ is replaced with $(\mathbf{x}',y')$. Let $h(\mathbf{z}_i,\mathbf{z}_j) = \omega_{ij}(y_j - y_i - \mathbf{g}(\mathbf{x}_i)^T(\mathbf{x}_j-\mathbf{x}_i))^2$ with any fixed $\mathbf{g}\in\mathcal{H}_R^p$, then we decompose $\mathcal{E}_{\mathbf{z}}(\mathbf{g})$ as follows,

$$\mathcal{E}_{\mathbf{z}}(\mathbf{g}) = \frac{1}{n(n-1)}\sum_{k\ne i,j\ne i}^n h(\mathbf{z}_k,\mathbf{z}_j) + \frac{1}{n(n-1)}\sum_{j=1}^n h(\mathbf{z}_i,\mathbf{z}_j) + \frac{1}{n(n-1)}\sum_{k=1}^n h(\mathbf{z}_k,\mathbf{z}_i).$$

Note that if $\mathbf{z}$ is replaced by $\mathbf{z}'$, the difference between $\mathcal{E}_{\mathbf{z}}(\mathbf{g})$ and $\mathcal{E}_{\mathbf{z}'}(\mathbf{g})$ boils down to the differences between the second and third components of the above decomposition. By Assumption A3, we see that $\pi_l > c_4$ for any l. Then it follows that

$$\mathcal{E}_{\mathbf{z}}(\mathbf{g}) - \mathcal{E}_{\mathbf{z}'}(\mathbf{g}) \le \frac{4(M_n + c_x\sum_{l=1}^p \|g_l\|_K)^2}{n} \le \frac{4(M_n + c_x n^{\psi/2}\sum_{l=1}^p \|\boldsymbol{\alpha}^{(l)}\|_2)^2}{n} \le \frac{4(M_n + \frac{c_x n^{\psi/2}R}{c_4\lambda_n})^2}{n},$$

where the second inequality follows from the Hölder inequality and Assumption A1. Interchanging the roles of $\mathbf{z}$ and $\mathbf{z}'$ yields that

$$|\mathcal{E}_{\mathbf{z}}(\mathbf{g}) - \mathcal{E}_{\mathbf{z}'}(\mathbf{g})| \leq \frac{4(M_n + \frac{c_x n^{\psi/2} R}{c_4 \lambda_n})^2}{n}, \quad \forall\ \mathbf{g} \in \mathcal{H}_R^p.$$

Then applying the McDiarmid's inequality, we have

$$P(|\mathcal{S}(R,\lambda_n) - \mathbb{E}(\mathcal{S}(R,\lambda_n))| \geq \varepsilon) \leq 2\exp\left(-\frac{n\varepsilon^2}{8(M_n + \frac{c_x n^{\psi/2} R}{c_4 \lambda_n})^4}\right).$$

In contrast with the first one, it is easier to obtain the second result in Lemma 6, since it only involves the fixed function $\mathbf{g}^*$. As a similar argument to the first one, we can set $C_i = \frac{4(M_n + c_x \sum_{l=1}^p \|g_l^*\|_K)^2}{n}$. Thus plugging C_i into the McDiarmid's inequality, our desired result follows immediately. ■

Proposition 7 *Assume the assumptions of Theorem 2 are met. If* $\mathcal{E}_{\mathbf{z}}(0) = \frac{1}{n(n-1)} \sum_{i,j=1}^n (y_i - y_j)^2$ *is upper bounded by* M_0*, then there exists a constant* c_9 *such that for any* $\delta \in (0,1)$*, with probability at least* $1-\delta$*,*

$$J(\hat{\mathbf{g}}) \leq c_9 \sqrt{\log(4/\delta)} \left\{ M_n^2 n^{-1/2} + n^{\frac{2\psi-1}{2}} M_0^2 \lambda_n^{-2} + \varepsilon \tau_n^p + \max_{l \leq p_0} \rho_n \pi_l \|g_l^* - \bar{g}_l\|_K + c_n^2/\varepsilon \right\}$$

where c_n *is defined as Proposition 4. In addition, there holds*

$$\mathcal{E}(\hat{\mathbf{g}}) - 2\sigma_s^2 \leq c_9 \sqrt{\log(4/\delta)} \left\{ M_n^2 n^{-1/2} + n^{\frac{2\psi-1}{2}} \lambda_n^{-2} + \varepsilon \tau_n^p + \max_{l \leq p_0} \rho_n \pi_l \|g_l^* - \bar{g}_l\|_K + c_n^2/\varepsilon \right\}$$

Proof of Proposition 7: By Lemma 2 of Ye and Xie (2012), we have

$$\mathbb{E}(\mathcal{S}(R,\lambda_n)) \leq \frac{(M_n + \frac{c_x n^{\psi/2} R}{c_4 \lambda_n})^2}{\sqrt{n}},$$

which, together with Lemma 6, implies that with probability at least $1-\delta$,

$$\varphi_1(\mathbf{z}) \leq |\mathcal{S}(R,\lambda_n)| \leq 3\sqrt{\frac{\log(2/\delta)}{n}} \Big(M_n + \frac{c_x n^{\psi/2} R}{c_4 \lambda_n}\Big)^2. \tag{15}$$

By Proposition 4, we recall that

$$J(\hat{\mathbf{g}}) + \mathcal{E}(\hat{\mathbf{g}}) \leq \varphi_1(\mathbf{z}) + 2\varphi_0(\mathbf{z}) + \Lambda_n(\varepsilon, \rho, K),$$

where $\Lambda_n(\varepsilon,\rho,K) = (1+\varepsilon)\mathcal{E}(\mathbf{g}^*) + \sum_{l=1}^{p_0} \rho_n \pi_l (\|g_l^*\|_K^2 - \|\bar{g}_l\|_K^2) + c_n^2/\varepsilon$. In addition, note that the Hessian matrix $\mathbf{H}^*(\mathbf{x})$ of f^* is bounded uniformly on $\mathbf{x}$, it is easy to verify that

$$\mathcal{E}(\mathbf{g}^*) - 2\sigma_s^2 = O(\tau_n^{4+p}),$$

which implies that

$$\Lambda_n(\varepsilon,\rho,K) - 2\sigma_s^2 \leq O(\varepsilon \tau_n^p + \max_{l \leq p_0} \rho_n \pi_l \|g_l^* - \bar{g}_l\|_K + c_n^2/\varepsilon), \tag{16}$$

since $\sigma_s^2 = O(\tau_n^p)$ by definition.

Thus, combining 14 in Lemma 6, 15 with 16 , for some constant c_9, we have

$$\begin{aligned} & J(\hat{\mathbf{g}}) + \mathcal{E}(\hat{\mathbf{g}}) - 2\sigma_s^2 \\ \leq \;& c_9\sqrt{\log(4/\delta)}\left\{M_n^2 n^{-1/2} + n^{\frac{2\psi-1}{2}} R^2 \lambda_n^{-2} + \varepsilon\tau_n^p + \max_{l\leq p_0} \rho_n \pi_l \|g_l^* - \bar{g}_l\|_K + c_n^2/\varepsilon\right\} \end{aligned}$$

with probability at least $1-\delta$. Finally, we give an explicit bound for R. Following the definition of $\hat{\mathbf{g}}$, we have

$$\mathcal{E}_{\mathbf{z}}(\hat{\mathbf{g}}) + J(\hat{\mathbf{g}}) \leq \mathcal{E}_{\mathbf{z}}(0) + J(0) \leq M_0,$$

which implies $\hat{\mathbf{g}} \in \mathcal{H}_{M_0}^p$, and thus $R = M_0$. As a consequence, the first and the second desired inequalities follow immediately after the fact that $\mathcal{E}(\hat{\mathbf{g}}) - 2\sigma_s^2 \geq 0$ and $J(\hat{\mathbf{g}}) \geq 0$. ■

Proof of Theorem 2: For given constant $c_6 > 0$, $\mathcal{C}$ is denoted to be the following event,

$$\mathcal{C} = \left\{\hat{\mathbf{g}} : \mathcal{E}(\hat{\mathbf{g}}) - 2\sigma_s^2 > c_6\sqrt{\log(4/\delta)}\Big(n^{-1/4} + n^{\frac{2\psi-1}{2}}\lambda_n^{-2} + \varepsilon\tau_n^p + n^{-\frac{1}{2(p+2)}} + n^{-\left(1-\frac{1}{2(p+2)}\right)}\lambda_n^2/\varepsilon\Big).\right\}.$$

Then, we split $\mathcal{C}$ into three different events as follows,

$$\begin{aligned} P(\mathcal{C}) &= P\Big(\mathcal{C} \cap \{|y| \leq n^{1/8}, U \leq M_0\}^c\Big) + P\Big(\mathcal{C} \cap \{|y| \leq n^{1/8}, U \leq M_0\}\Big) \\ &\leq P(|y| > n^{1/8}) + P(|y| \leq n^{1/8}, U > M_0) + P\Big(\mathcal{C} \cap \{|y| \leq n^{1/8}, U \leq M_0\}\Big), \end{aligned}$$

where $U = \frac{1}{n(n-1)}\sum_{i,j=1}^n (y_i - y_j)^2$ and $M_0 = 4B^2 + 2\sigma^2 + 1$ with B an upper bound of $f^*(\mathbf{x})$. The existence of B is due to the assumptions that $\mathbf{x}$ has a compact support and f^* is continuous Now we bound these three probabilities one by one.

First, by Chebyshev inequality, $P(|y| > n^{\frac{1}{8}}) = E(P(f^*(\mathbf{x}) + \epsilon > n^{\frac{1}{8}} | \mathbf{x})) \leq O(n^{-\frac{1}{4}})$, where the last inequality is due to bounded $f^*(\mathbf{x})$. For the second probability, note that U is a U-statistic with mean $E(U) = E(E(U|\mathbf{x}_i, \mathbf{x}_j)) = E(f^*(\mathbf{x}_i) - f^*(\mathbf{x}'))^2 + 2\sigma^2 \leq 4B^2 + 2\sigma^2$. By Bernstein's inequality for U-statistic (Janson, 2004),

$$P(|y| \leq n^{1/8}, U > M_0) \leq P(U > E(U) + 1 || y| \leq n^{1/8}) \leq \exp\left\{-\frac{1}{16}n^{3/4}\right\},$$

where $(y_i - y_j)^2$ is upper bounded by $4n^{1/4}$, which completes the second term.

Now we turn to the third term. Within the set $\{|y| \leq n^{1/8}, U \leq M_0\}$, by Proposition 7,

$$\mathcal{E}(\hat{\mathbf{g}}) - 2\sigma_s^2 \leq c_9\sqrt{\log(4/\delta)}\Big(n^{-1/4} + n^{\frac{2\psi-1}{2}} M_0^2 \lambda_n^{-2} + \varepsilon\tau_n^p + \max_{l\leq p_0} \rho_n \pi_l \|g_l^* - \bar{g}_l\|_K + c_n^2/\varepsilon\Big).$$

With the choice of $\rho_l = n^{-\eta}$ with $\eta = \frac{1}{4(p+2)}$ for all l, we have $\|g_l^* - \bar{g}_l\|_K = O(n^{-\frac{1}{4(p+2)}})$ according to theorems 14, 17, 19 in Mukherjee and Zhou (2006). In addition, $c_n = \frac{c_x p \lambda_n}{\rho_n\sqrt{n-1}} = O(n^{-(\frac{1}{2}-\eta)}\lambda_n)$. Thus with probability at least $1-\delta$, for some constant c_6,

$$\mathcal{E}(\hat{\mathbf{g}}) - 2\sigma_s^2 \leq c_6\sqrt{\log(4/\delta)}\Big(n^{-1/4} + n^{\frac{2\psi-1}{2}}\lambda_n^{-2} + \varepsilon\tau_n^p + n^{-\frac{1}{2(p+2)}} + n^{-\left(1-\frac{1}{2(p+2)}\right)}\lambda_n^2/\varepsilon\Big).$$

Specifically, with $\lambda_n = n^{\frac{2\psi-1}{4}+\frac{1}{4(p+2)}}$, $\tau_n = n^{-\frac{\theta}{4p(p+2)(p+4+3\theta)}}$ and $\varepsilon = \tau_n^4$, we have

$$\mathcal{E}(\hat{\mathbf{g}}) - 2\sigma_s^2 \le c_6\sqrt{\log(4/\delta)}n^{-\Theta},$$

where $\Theta = \min\Big\{\frac{\theta(p+4)}{4p(p+2)(p+4+3\theta)}, \frac{1}{2(p+2)}, \frac{p^2(p+4+3\theta)-2\theta}{2p(p+2)(p+4+3\theta)}\Big\}$. ■

Proof of Theorem 3: First, we show that $\|\hat{\boldsymbol{\alpha}}^{(l)}\|_2 = 0$ for any $l > p_0$ by contradiction. Suppose that $\|\hat{\boldsymbol{\alpha}}^{(l)}\|_2 > 0$ for some $l > p_0$. Taking the first derivative of (4) with respect to $\boldsymbol{\alpha}^{(l)}$ yields that

$$\frac{2}{n(n-1)}\sum_{i,j=1}^{n} w_{ij}(y_i - y_j - \hat{\mathbf{g}}(\mathbf{x}_i)^T(\mathbf{x}_i - \mathbf{x}_j))(x_{il} - x_{jl})\mathbf{K}_l = -\frac{\lambda_n\pi_l\hat{\boldsymbol{\alpha}}^{(l)}}{\|\hat{\boldsymbol{\alpha}}^{(l)}\|_2}. \tag{17}$$

Note that the norm of the right-hand side divided by $n^{1/2}$ is $n^{-1/2}\lambda_n\pi_l$, which diverges to ∞ by Assumption A3. Then the contradiction can be concluded by showing the norm of the left-hand side is smaller than $O(n^{1/2})$.

For simplicity, denote $\mathcal{B}_{\mathbf{z}}(\mathbf{g}) = \frac{2}{n(n-1)}\sum_{i,j=1}^{n} w_{ij}(y_i - y_j - \mathbf{g}(\mathbf{x}_i)^T(\mathbf{x}_i - \mathbf{x}_j))$. As the elements in both $\mathbf{x}$ and $\mathbf{K}_l$ are bounded by Assumption A1, it suffices to show $|\mathcal{B}_{\mathbf{z}}(\hat{\mathbf{g}})|$ is bounded. Denote $\mathcal{C} = \Big\{\hat{\mathbf{g}} : |\mathcal{B}_{\mathbf{z}}(\hat{\mathbf{g}})| > c_{10}\sqrt{\log(4/\delta)}(n^{-\Theta/2} + n^{\frac{\psi-1}{2}}\lambda_n^{-1} + n^{-3/8})\Big\}$. As in the proof of Theorem 2, we decompose $P(\mathcal{C})$ as

$$\begin{aligned} P(\mathcal{C}) &= P\Big(\mathcal{C}\cap\{|y| \le n^{1/8}, U \le M_0\}^c\Big) + P\Big(\mathcal{C}\cap\{|y| \le n^{1/8}, U \le M_0\}\Big) \\ &\le P(|y| > n^{1/8}) + P(|y| \le n^{1/8}, U > M_0) + P\Big(\mathcal{C}\cap\{|y| \le n^{1/8}, U \le M_0\}\Big). \end{aligned}$$

where U and M_0 are the same as in Theorem 2.

Following the proof of Theorem 2, the first two probabilities can be bounded as $P(|y| > n^{1/8}) \le O(n^{-1/4})$ and $P(|y| \le n^{1/8}, U > M_0) \le \exp\{-\frac{1}{16}n^{3/4}\}$. To bound the third probability, a slight modification of the proof of Proposition 7 yields that when $|y| \le n^{1/8}$ and $U \le M_0$, we have $J(\hat{\mathbf{g}}) \le M_0$ and with probability at least $1 - \delta/2$,

$$|\mathcal{B}_{\mathbf{z}}(\hat{\mathbf{g}}) - E(\mathcal{B}_{\mathbf{z}}(\hat{\mathbf{g}})| \le 3\sqrt{\frac{\log(4/\delta)}{n}}\Big(n^{1/8} + \frac{c_x n^{\psi/2}M_0}{c_4\lambda_n}\Big).$$

We then bound $|E(\mathcal{B}_{\mathbf{z}}(\hat{\mathbf{g}})|$ as follows,

$$\begin{aligned} |E(\mathcal{B}_{\mathbf{z}}(\hat{\mathbf{g}}))| &= \Big|\int_{\mathcal{X}}\int_{\mathcal{X}} w(\mathbf{x},\mathbf{u})\Big(f^*(\mathbf{x}) - f^*(\mathbf{u}) + \hat{\mathbf{g}}(\mathbf{x})^T(\mathbf{u}-\mathbf{x})\Big)d\rho_X(\mathbf{x})d\rho_X(\mathbf{u})\Big| \\ &\le \Big(\int_{\mathcal{X}}\int_{\mathcal{X}}\Big|w(\mathbf{x},\mathbf{u})\Big(f^*(\mathbf{x}) - f^*(\mathbf{u}) + \hat{\mathbf{g}}(\mathbf{x})^T(\mathbf{u}-\mathbf{x})\Big)\Big|^2 d\rho_X(\mathbf{x})d\rho_X(\mathbf{u})\Big)^{1/2} \\ &\le \Big(\int_{\mathcal{X}}\int_{\mathcal{X}} w(\mathbf{x},\mathbf{u})\Big(f^*(\mathbf{x}) - f^*(\mathbf{u}) + \hat{\mathbf{g}}(\mathbf{x})^T(\mathbf{u}-\mathbf{x})\Big)^2 d\rho_X(\mathbf{x})d\rho_X(\mathbf{u})\Big)^{1/2} \\ &= \Big(\mathcal{E}(\hat{\mathbf{g}}) - 2\sigma_s^2\Big)^{1/2}. \end{aligned}$$

The first inequality follows from Hölder inequality, and the second inequality follows from the fact that $w(\mathbf{x},\mathbf{u}) \le 1$ for any $\mathbf{x}$ and $\mathbf{u}$. Therefore, within the set $\{|y| \le n^{1/8}, U \le M_0\}$,

we have with probability at least $1-\delta/2$ that

$$|\mathcal{B}_{\mathbf{z}}(\hat{\mathbf{g}})| \leq c_{10}\sqrt{\log(4/\delta)}(n^{-\Theta/2} + n^{\frac{\psi-1}{2}}\lambda_n^{-1} + n^{-3/8}). \tag{18}$$

This implies that $P(\mathcal{C}) \leq \delta/2 + O(n^{-1/4})$ for any given δ. Then it is clear that the norm of the left-hand side of (17) divided by $n^{1/2}$ will converge to 0 in probability, which contradicts with the fact that the norm of the right-hand side divided by $n^{1/2}$ diverges to ∞. Therefore, we have $\|\hat{\boldsymbol{\alpha}}^{(l)}\|_2 = 0$ for any $l > p_0$.

Next, we show that $\|\hat{\boldsymbol{\alpha}}^{(l)}\|_2 > 0$ for any $l \leq p_0$. Let $\bar{\mathcal{X}}_\tau = \{\mathbf{x} \in \mathcal{X} : d(\mathbf{x}, \partial\mathcal{X}) > \tau_n, p(\mathbf{x}) > \tau_n + c_2\tau_n^\theta\}$. Same as the proof of Theorem 5 in Ye and Xie (2012), for some given constant c_{11}, we have

$$\int_{\bar{\mathcal{X}}_\tau} \|\hat{\mathbf{g}}(\mathbf{x}) - \mathbf{g}^*(\mathbf{x})\|_2^2 d\rho_X(\mathbf{x}) \leq c_{11}\tau_n^{-(p+3)}\Big(\tau_n^{p+4} + \mathcal{E}(\hat{\mathbf{g}}) - 2\sigma_s^2\Big).$$

According to Theorem 2, it can be showed that

$$\int_{\bar{\mathcal{X}}_\tau} \|\hat{\mathbf{g}}(\mathbf{x}) - \mathbf{g}^*(\mathbf{x})\|_2^2 d\rho_X(\mathbf{x}) \to 0.$$

Now suppose $\|\hat{\boldsymbol{\alpha}}^{(l)}\|_2 = 0$ for some $l \leq p_0$, which implies

$$\int_{\bar{\mathcal{X}}_\tau} \|\hat{\mathbf{g}}(\mathbf{x}) - \mathbf{g}^*(\mathbf{x})\|_2^2 d\rho_X(\mathbf{x}) = \int_{\bar{\mathcal{X}}_\tau} \|g_l^*(\mathbf{x})\|_2^2 d\rho_X(\mathbf{x}).$$

As $\tau_n \to 0$, $\int_{\bar{\mathcal{X}}_\tau} \|g_l^*(\mathbf{x})\|_2^2 d\rho_X(\mathbf{x}) \geq \int_{\mathcal{X}\setminus\mathcal{X}_t} \|g_l^*(\mathbf{x})\|_2^2 d\rho_X(\mathbf{x})$, which is a positive constant by Assumption A4, and then leads to the contradiction. Combining the above two statements implies the desired variable selection consistency. ■

References

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.

K. Brabanter, J. Brabanter and B. Moor. Derivative estimation with local polynomial fitting. *Journal of Machine Learning Research*, 14: 281-301, 2014.

L. Breiman. Better subset regresson using nonnegative garrote. *Technometrics*, 37: 373-384, 1995.

L. Breiman and J. Friedman. Estimating optimal transformations for multiple regression and correlation. *Journal of the American Statistical Association*, 80: 580-598, 1985.

C. Cortes and V. Vapnik. Support vector networks. *Journal of Machine Learning Research*, 20: 273-297, 1995.

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.

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.

J. Friedman. Multivariate Adaptive Regression Splines. *The Annal of Statistics*, 19: 1-67, 1991.

E. Gregory, J. Fred and W. Henryk. On dimension-independent rates of convergence for function approximation with Gaussian Kernels. *SIAM Journal on Numerical Analysis*, 50: 247-271, 2012.

R. Genuer, J. Poggi and C. Tuleau-Malot. Variable selection using random forests. *Pattern Recognition Letters*, 31: 2225-2236, 2010.

N. Hao and H. Zhang. Interaction screening for ultrahigh-dimensional data. *Journal of the American Statistical Association*, 109: 1285-1301, 2014.

W. Härdle and T. Gasser. On robust kernel estimation of derivatives of regression functions. *Scandinavian Journal of Statistics*, 12: 233-240, 1985.

J. L. Horowitz and E. Mammen. Rate-optimal estimation for a general class of nonparametric regression models with unknown link functions. *Annal of Statistics*, 35: 2589-2619, 2007.

J. Huang, J. Horowitz and F. Wei. Variable selection in nonparameteric additive models. *The Annals of Statistics*, 38: 2282-2313, 2010.

J. Huang, and L. Yang. Identification of non-linear additive autoregressive models. *Journal of the Royal Statistical Society: Series B*, 66: 463-477, 2004.

S. Janson. Large deviations for sums of partly dependent random variables. *Random Structures and Algorithms*, 24: 234-248, 2004.

R. Jarrow, D. Ruppert and Y. Yu. Estimating the term structure of corporate debt with a semiparametric penalized spline model. *Journal of the American Statistical Association*, 99: 57-66, 2004.

L. Li, D. Cook and C. Nachtsheim. Model-free variable selection. *Journal of the Royal Statistical Society, Series B.*, 67: 285-299, 2005.

R. Li, W. Zhong and L. Zhu. Feature screening via distance correlation learning. *Journal of the American Statistical Association*, 107: 1129-1139, 2012.

Y. Lin. A note on margin-based loss functions in classification. *Statistics and Probability Letters*, 68: 73-82, 2004.

Y. Lin and H. Zhang. Component selection and smoothing in multivariate non-parametric regression. *Annal of Statistics*, 34: 2272-2297, 2006.

Y. Liu. Fisher consistency of multicategory support vector machines. *Eleventh International Conference on Artificial Intelligence and Statistics*, 289-296, 2006.

Y. Liu and Y. Wu. Variable selection via a combination of the L0 and L1 penalties. *Journal of Computational and Graphical Statistics*, 16: 782-798, 2007.

C. McDiarmid,. On the method of bounded differences. *In Surveys in Combinatorics*, 141: 148-188.

S. Mukherjee and D. Zhou. Learning coordinate covariates via gradient. *Journal of Machine Learning Research*, 7: 519-549, 2006.

B. Scholköpf and A.J. Smola. Learning with kernels: Support vector machines, regularization, optimization, and beyond. MIT Press, 2002.

X. Shen, W. Pan and Y. Zhu. Likelihood-based selection and sharp parameter estimation. *Journal of the American Statistical Association*, 107: 223-232, 2012.

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.

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.

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.

R. Tibshirani. Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society: Series B*, 58: 267-288, 1996.

G. Wahba. Support vector machines, reproducing kernel hilbert spaces, and randomized GACV. *Advances in kernel methods*, 69-88, 1999.

L. Xue. Consistent variable selection in additive models. *Statistica Sinica*, 19: 1281-1296, 2009.

Y. Yang and H. Zou. A fast unified algorithm for solving group-lasso penalized learning problems. *Statistics and Computing*, 25: 1129-1141, 2015

G. Ye and X. Xie. Learning sparse gradients for variable selection and dimension reduction. *Journal of Machine Learning Research*, 87: 303-355, 2012.

M. Yuan and Y. Lin. Model selection and estimation in regression with group variables. *Journal of the Royal Statistical Society: Series B*, 68: 49-67, 2006

L. Zhu, L. Li, R. Li and L. Zhu. Model-free feature screening for ultra-high dimensional data. *Journal of the American Statistical Association*, 106: 1464-1475, 2011.

H. Zou. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101: 1418-1429.