Journal of Machine Learning Research 23 (2022) 1-31 Submitted 7/21; Revised 4/22; Published 7/22

# A Forward Approach for Sufficient Dimension Reduction in Binary Classification

**Jongkyeong Kang** J.K@KANGWON.AC.KR
*Department of Information Statistics*
*Kangwon National University*
*Gangwon-do, 24341, Korea*
*and*
*Department of Statistics*
*Korea University*
*Seoul, 02841, Korea*

**Seung Jun Shin** SJSHIN@KOREA.AC.KR
*Department of Statistics*
*Korea University*
*Seoul, 02841, Korea*

**Editor:** Ryan Tibshirani

## Abstract

Since the proposal of the seminal sliced inverse regression (SIR), inverse-type methods have proved to be canonical in sufficient dimension reduction (SDR). However, they often underperform in binary classification because the binary responses yield two slices at most. In this article, we develop a forward SDR approach in binary classification based on weighted large-margin classifiers. First, we show that the gradient of a large-margin classifier is unbiased for SDR as long as the corresponding loss function is Fisher consistent. This leads us to propose the weighted outer-product of gradients (wOPG) estimator. The wOPG estimator can recover the central subspace exhaustively without linearity (or constant variance) conditions, which despite being routinely required, they are untestable assumption. We propose the gradient-based formulation for the large-margin classifier to estimate the gradient function of the classifier directly. We also establish the consistency of the proposed wOPG estimator and demonstrate its promising finite-sample performance through both simulated and real data examples.

**Keywords:** dimension reduction, Fisher consistency, gradient learning, large-margin classifier, outer-product gradient

## 1. Introduction

Sufficient dimension reduction (SDR) seeks a low dimensional subspace of a $p$-dimensional predictor $\mathbf{X}$, referred to as a dimension reduction subspace (DRS), to retain the information of a response $Y$ in $\mathbf{X}$. To be more precise, the space spanned by the columns of $\mathbf{B} \in \mathbb{R}^{p\times d}$ denoted by $\text{span}(\mathbf{B})$ is a DRS if

$$Y \perp \mathbf{X} \mid \mathbf{B}^\top \mathbf{X}, \tag{1}$$

holds where $\perp$ denotes the statistical independence. However, DRS is unidentifiable since any linear transformations of $\mathbf{B}$ satisfying (1) also satisfy (1). As an identifiable target, the central subspace, denoted by $\mathcal{S}_{Y|\mathbf{X}}$, is the intersection of all DRSes and a primal SDR target. It is shown that the central subspace uniquely exists under mild conditions (Yin et al., 2008), and we assume that $\text{span}(\mathbf{B}) = \mathcal{S}_{Y|\mathbf{X}}$ throughout this article.

SDR is first proposed in the regression context where $Y$ is continuous. The seminal sliced inverse regression (SIR, Li, 1991) utilizes the inverse regression function $E(\mathbf{X} \mid Y)$ to estimate $\mathcal{S}_{Y|\mathbf{X}}$. SIR slices data $(y_i, \mathbf{x}_i), i = 1, \cdots, n$ based on the relative size of the observed responses, $y_i$ to estimate $E(\mathbf{X} \mid Y)$. Following the spirit of SIR, many SDR methods are developed based on the conditional moments of $\mathbf{X}$ given $Y$. These are called "inverse" methods, and popular examples include the slice averaged variance estimation (SAVE, Cook and Weisberg, 1991), contour regression (Li et al., 2005), directional regression (Li and Wang, 2007), cumulative slicing estimation (Zhu et al., 2010), and principal support vector machines (PSVM, Li et al., 2011), to name a few. The inverse methods are easy to implement since the inverse moment can be readily obtained by slicing. However, they usually require technical assumptions on $\mathbf{X}$, such as the linearity condition and constant variance condition that assume $E(\mathbf{X} \mid \mathbf{B}^\top \mathbf{X}) = \mathbf{P}\mathbf{X}$ and $\text{Cov}(\mathbf{X} \mid \mathbf{B}^T\mathbf{X}) = \boldsymbol{\Sigma} - \mathbf{P}\boldsymbol{\Sigma}\mathbf{P}^\top$, respectively, where $\mathbf{P} = \boldsymbol{\Sigma}\mathbf{B}(\mathbf{B}^\top\boldsymbol{\Sigma}\mathbf{B})^{-1}\mathbf{B}^\top$ and $\boldsymbol{\Sigma} = \text{Cov}(\mathbf{X})$. It is well established that utilizing these conditions when they are not satisfied introduces bias into estimating the central space. More importantly, employing these conditions leads to a substantial efficiency loss even when the conditions are satisfied (Ma and Zhu, 2013). In other words, when an SDR estimator does not exploit linearity and constant variance conditions, its performance can further be improved.

It is well known that the inverse methods suffer in binary classification, where the slicing becomes inevitably inefficient. For example, SIR cannot recover $\mathcal{S}_{Y|\mathbf{X}}$ exhaustively unless $d = 1$ since there is one slice available at most for the binary $Y$. Most inverse methods based on slicing have similar problems. To overcome this, Shin et al. (2014) proposed the probability-enhanced SDR (PRE-SDR) that slices data based on the class probability $P(Y = 1 \mid \mathbf{X} = \mathbf{x})$. The PRE-SDR estimates the relative order of the class probabilities, rather than their exact values, by training the weighted support vector machine (WSVM), whose loss function is Fisher consistent. Shin et al. (2017) proposed the principal weighted support vector machine (PWSVM) by applying the idea of PRE-SDR to the principal support vector machine proposed by Li et al. (2011). Although PRE-SDR and PWSVM are elegantly designed for the SDR with a binary response, they still depend on the linearity condition and may fail to estimate $\mathcal{S}_{Y|\mathbf{X}}$ exhaustively.

In the meantime, another type of approach, known as the "forward" method, has been developed for the SDR. Namely, the forward method estimates $\mathcal{S}_{Y|\mathbf{X}}$ from the quantities related to the conditional moments of $Y$ given $\mathbf{X}$, not $\mathbf{X}$ given $Y$. The advantage of the forward method is that it does not require additional assumptions on $\mathbf{X}$, such as the linearity and constant variance conditions indispensable to inverse methods. One of the earliest proposals in this regard is the outer product of gradients (OPG) estimator (Xia et al., 2002) to estimate the central mean subspace $\mathcal{S}_{E(Y|\mathbf{X})} \subseteq \mathcal{S}_{Y|\mathbf{X}}$ (Cook and Li, 2002) defined as $\mathcal{S}_{E(Y|\mathbf{X})} = \text{span}(\mathbf{B})$, where $Y \perp E(Y \mid \mathbf{X}) \mid \mathbf{B}^\top\mathbf{X}$. The equation below presents the basis of OPG.

$$\text{span}\left\{E\left[\nabla m(\mathbf{X})\{\nabla m(\mathbf{X})\}^\top\right]\right\} = \mathcal{S}_{E(Y|\mathbf{X})},$$

where $m(\mathbf{x}) = E(Y \mid \mathbf{X} = \mathbf{x})$ denotes the (forward) regression function and $\nabla m(\mathbf{x}) = \partial m(\mathbf{x})/\partial \mathbf{x}$ is its gradient vector. Xia et al. (2002) further proposed the minimum average variance estimation (MAVE), a computationally improved version of the OPG estimator.

Note that $\mathcal{S}_{E(Y|\mathbf{X})}$ is a subspace of $\mathcal{S}_{Y|\mathbf{X}}$ and can be very different. Xia (2007) proposed the dOPG method that extends the idea of the OPG to estimate $\mathcal{S}_{Y|\mathbf{X}}$ by replacing the conditional mean with the corresponding density for which the prefix 'd' stands. That is, it is shown that

$$\operatorname{span}\left\{E\left[\nabla \rho_{Y|X}(\mathbf{X})\{\nabla \rho_{Y|\mathbf{X}}(\mathbf{X})\}^{\top}\right]\right\} = \mathcal{S}_{Y|\mathbf{X}},$$

where $\rho_{Y|X}$ denotes the conditional density of $Y$ given $\mathbf{X}$, and developed an efficient algorithm to compute the sample estimate of $\nabla \rho_{Y|X}(\mathbf{X})$ based on the local regression. In fact, any quantity, other than the density, that determines the conditional distribution of $Y$ given $\mathbf{X}$ can be used to recover $\mathcal{S}_{Y|\mathbf{X}}$. For example, Wang and Xia (2008) proposed the sliced regression that estimates the gradient of the conditional distribution function of $Y$ given $\mathbf{X}$ instead of the density. Kong and Xia (2014) proposed the qOPG method that exploits the gradient of conditional quantile regression functions of $Y \mid \mathbf{X}$ for different quantile levels to recover $\mathcal{S}_{Y|\mathbf{X}}$. We referred to Chapter 11 of Li (2018) for a comprehensive overview of the forward method for SDR.

In this article, we propose a novel forward SDR method for binary classification based on the large-margin classifier. The large-margin classifier is one of the most popular classes for the binary classification, and includes many popular classifiers such as logistic regression (LR) and support vector machine (SVM).

Let $f(\mathbf{x})$ denote a classification function whose sign predicts the class label $y \in \{-1, 1\}$ corresponding to $\mathbf{x}$. Given a set of training samples $(y_i, \mathbf{x}_i), i = 1, \cdots, n$, the weighted large-margin classifier solves

$$\min_{f \in \mathcal{F}} \sum_{i=1}^{n} w_\pi(y_i) L\{y_i f(\mathbf{x}_i)\} + \lambda J(f), \tag{2}$$

where $L$ denotes a loss function of the functional margin $yf(\mathbf{x})$, and the weight function $w_\pi(y)$ equals to $1 - \pi$ when $y = 1$ and $\pi$ if otherwise for a given $\pi \in (0, 1)$. Note that the weight $\pi$ controls the relative importance between the two classes and plays a crucial role in our proposal. The functional $J$ measures the complexity of $f$, and $\lambda$ is the tuning parameter that controls the balance between data fitting and the model complexity.

At the population level, the most natural choice for $L$ is the zero-one loss defined as $L_{0\text{-}1}(m) = 1$ if $m \leq 0$ and 0 otherwise, as its population minimizer yields the Bayes classification rule. Namely, we have

$$\operatorname{sign}\{f_\pi^*(\mathbf{x})\} = \operatorname{sign}\{p(\mathbf{x}) - \pi\},$$

where $f_\pi^*(\mathbf{x}) = \operatorname{argmin}_{f(\mathbf{x}) \in \mathbb{R}} E\left[w_\pi(Y) L_{0\text{-}1}(Y f(\mathbf{x}) \leq 0) \mid \mathbf{X} = \mathbf{x}\right]$ and $p(\mathbf{x}) = P(Y = 1 \mid \mathbf{X} = \mathbf{x})$ denotes the conditional class probability. However, it is not tractable to minimize the sample zero-one loss, and one often replaces it with a convex surrogate function that satisfies the following minimal condition known as Fisher consistency (Lin, 2004; Bartlett et al., 2006).

**Definition 1 (Fisher Consistency)** *A loss function $L$ for the binary classification is said to be Fisher consistent if*

$$sign\{f_\pi(\mathbf{x})\} = sign\{f_\pi^*(\mathbf{x})\} = sign\{p(\mathbf{x}) - \pi\}, \tag{3}$$

*for an arbitrary given $\pi \in (0,1)$, where*

$$f_\pi(\mathbf{x}) = \underset{f(\mathbf{x})\in\mathbb{R}}{\operatorname{argmin}} E\left[w_\pi(Y)L\{Yf(\mathbf{x})\} \mid \mathbf{X} = \mathbf{x}\right]. \tag{4}$$

Well-known examples of Fisher consistent loss include, but are not limited to, the hinge loss for SVM, the negative binomial likelihood loss for logistic regression, the exponential loss for boosting, and the large-margin unified machine loss (Liu et al., 2011). Fisher consistency presents a theoretical connection between $f_\pi(\mathbf{x})$ and $p(\mathbf{x})$. Assuming $f_\pi(\mathbf{x})$ is a continuous and decreasing function of $\pi$, we have $p(\mathbf{x}) = \{\pi : f_\pi(\mathbf{x}) = 0\}$. In other words, $f_\pi(\mathbf{x}), \forall \pi \in (0,1)$ determines the conditional distribution of $Y$ given $\mathbf{X}$ (Wang et al., 2008). This leads us to naturally extend the idea of the OPG estimator for SDR based on the large-margin classifier when the response is binary. We showed that the outer product of the gradient of $f_\pi(\mathbf{x})$ for different values of $\pi \in (0,1)$ exhaustively estimate $\mathcal{S}_{Y|\mathbf{X}}$. The idea is directly motivated by the qOPG estimator (Kong and Xia, 2014) that exploits composite quantile regression (Zou and Yuan, 2008) to extend the concept of "borrowing strength across different quantiles" to the SDR context. However, the qOPG estimator suffers inefficiency due to the irregularity of the population quantile function of a binary random variable. We would like to point out that the weighted large-margin classifier $f_\pi(\mathbf{x})$ plays a similar role to the quantile regression function in qOPG (Kong and Xia, 2014). This motivates us to refer to our method as the weighted outer-product of gradients (wOPG) estimator.

In this article, we propose two approaches to estimate the gradient of $f_\pi(\mathbf{x})$, i.e., $\nabla f_\pi(\mathbf{x})$. First, we estimate $f_\pi(\mathbf{x})$ by solving the conventional large-margin classifier and then compute its gradient $\nabla f_\pi(\mathbf{x})$ by taking derivative with respective $\mathbf{x}$. We call this naive-wOPG estimator to elaborate on its computational simplicity. Although the naive-wOPG may be attractive in practice, it often suffers due to the instability of the differential operator. For instance, the consistency of $f_\pi(\mathbf{x})$ does not imply the consistency of its gradient in general. To overcome this drawback, we propose an alternative approach based on gradient learning (Mukherjee and Wu, 2006) that directly estimates $\nabla f_\pi(\mathbf{x})$ instead of deriving it analytically from the estimated $f_\pi(\mathbf{x})$. We call this the wOPG estimator without any prefix.

The local linear regression (Härdle and Stoker, 1989; Fan, 1993) has been widely used for learning gradients in statistical applications including SDR (Härdle and Stoker, 1989; Fan, 1993; Xia, 2007; Kong and Xia, 2014). In this article, however, we employed the reproducing kernel Hilbert space (Wahba, 1990, RKHS) to estimate the nonlinear gradient function since the kernel trick on RKHS is canonical in the large-margin classifier such as the SVM. The gradient learning in RKHS is also common and has been studied in machine learning communities. See for example, Mukherjee and Wu (2006), Ye and Xie (2012), Yang et al. (2016) and the references therein.

The rest of the article is organized as follows. In Section 2, we provide a foundation of the wOPG estimator by showing that the OPG of the weighted large-margin classifier is unbiased for SDR and then introduce the naive-wOPG estimator. In Section 3, the wOPG

estimator based on gradient learning is introduced in great detail, including its computation and asymptotic properties. Section 4 is devoted to additional issues to complete the proposed method such as the estimation of the structural dimension and tuning parameter selection in learning $f_\pi$. Simulation studies are carried out in Section 5, and the illustration to Breast Cancer Coimbra data is presented in Section 6. Concluding remarks follows in Section 7, and the proofs of theorems and complete computational algorithms are relegated to Appendix.

## 2. Weighted Outer-Product of Gradients (wOPG)

First, we show that the OPG of $f_\pi(\mathbf{x})$ is unbiased for SDR, which serves as a theoretical foundation of the wOPG estimator. This leads us to propose a straightforward approach to estimate $\mathcal{S}_{Y|\mathbf{X}}$.

### 2.1 Foundation of wOPG

For a binary response $Y \in \{-1, 1\}$, let us introduce

$$\tilde{f}_\pi(\mathbf{B}^\top \mathbf{x}) = \underset{f(\mathbf{x})\in\mathbb{R}}{\operatorname{argmin}} E\left[w_\pi(Y) L\{Y f(\mathbf{x})\} \mid \mathbf{B}^\top \mathbf{X} = \mathbf{B}^\top \mathbf{x}\right].$$

Under (1), we have $f_\pi(\mathbf{x}) = \tilde{f}_\pi(\mathbf{B}^\top \mathbf{x})$ which implies $\nabla f_\pi(\mathbf{x}) = \mathbf{B}\nabla \tilde{f}_\pi(\mathbf{B}^\top \mathbf{x})$, where $\nabla f_\pi(\mathbf{x})$ and $\nabla \tilde{f}_\pi(\mathbf{x})$ denote the $p$-dimensional gradient vectors of $f_\pi(\mathbf{x})$ and $\tilde{f}_\pi(\mathbf{x})$ at $\mathbf{x}$, respectively. Motivated by the spirit of the OPG (Xia et al., 2002), we define $\mathbf{M}(\pi)$ as

$$\mathbf{M}(\pi) = E\left[\nabla f_\pi(\mathbf{X})\{\nabla f_\pi(\mathbf{X})\}^\top\right], \tag{5}$$

then we have

$$\operatorname{span}\{\mathbf{M}(\pi)\} \subseteq \mathcal{S}_{Y|\mathbf{X}}. \tag{6}$$

This is because $\mathbf{M}(\pi) = E\left[\nabla f_\pi(\mathbf{X})\{\nabla f_\pi(\mathbf{X})\}^\top\right] = \mathbf{B}E\left[\nabla \tilde{f}_\pi(\mathbf{B}^\top \mathbf{X})\{\nabla \tilde{f}_\pi(\mathbf{B}^\top \mathbf{X})\}^\top\right]\mathbf{B}^\top$, which implies $\mathbf{M}(\pi) = \mathbf{P}_\mathbf{B}\mathbf{M}(\pi)\mathbf{P}_\mathbf{B}$ where $\mathbf{P}_\mathbf{B} = \mathbf{B}(\mathbf{B}^\top\mathbf{B})^{-1}\mathbf{B}^\top$ denotes the projection matrix onto $\mathcal{S}_{Y|\mathbf{X}} = \operatorname{span}(\mathbf{B})$.

Let us define

$$\mathbf{M} = \int_0^1 \mathbf{M}(\pi) d\pi. \tag{7}$$

The following Theorem 2 provides the foundation of the wOPG estimator for $\mathcal{S}_{Y|\mathbf{X}}$.

**Theorem 2** *Suppose $\nabla f_\pi(\mathbf{X})$ exists for all $\pi \in (0, 1)$, then we have*

$$\mathit{span}(\mathbf{M}) = \mathcal{S}_{Y|\mathbf{X}}.$$

We refer to $\mathbf{M}$ in (7) as the wOPG working matrix, and its $d$ leading eigenvectors consistently estimate the basis set of $\mathcal{S}_{Y|\mathbf{X}}$ by Theorem 2. Notably, the wOPG working matrix

recover $\mathcal{S}_{Y|\mathbf{X}}$ exhaustively without any additional assumptions such as linearity and constant variance assumptions. It is also worth pointing out that $\mathbf{M}(\pi)$ tends to zero as $\pi$ goes to zero or one since $f_\pi(\mathbf{x})$ tends to constant functions, which are not worth considering from the classification perspective. In such cases, for a given $\epsilon > 0$ we can choose $\delta > 0$ such that $\|\mathbf{M} - \mathbf{M}_T\| < \epsilon$, where

$$\mathbf{M}_T = \int_\delta^{1-\delta} \mathbf{M}(\pi) d\pi \tag{8}$$

is the truncated version of (7). For this reason, we restrict our attention to the case when $\pi \in (\delta, 1-\delta)$ for a sufficiently small $\delta > 0$ given. Such truncation is not uncommon in the quantile regression context since standard tools based on the average fail to work to uncover the asymptotic properties of extreme quantiles (i.e., minimum and maximum) (Chernozhukov et al., 2017). In addition, Kong and Xia (2014) argued that the truncation dose not have a significant effect since $\mathbf{M}_T$ is expected to be nearly, if not completely, identical to $\mathbf{M}$ for a sufficiently small $\delta$. The truncated version of the estimator is given by

$$\widehat{\mathbf{M}}_T = \int_\delta^{1-\delta} \widehat{\mathbf{M}}(\pi) d\pi, \tag{9}$$

where

$$\widehat{\mathbf{M}}(\pi) = \frac{1}{n} \sum_{i=1}^{n} \nabla \hat{f}_\pi(\mathbf{x}_i) \nabla \hat{f}_\pi(\mathbf{x}_i)^T, \tag{10}$$

and $\nabla \hat{f}_\pi$ is an estimator of $\nabla f_\pi$. Assume that we have $H$ estimates $\nabla \hat{f}_{\pi_h}$ for a given sequence of weights $0 < \delta = \pi_1 \cdots < \pi_H = 1 - \delta < 1$. Then we finally have

$$\widehat{\mathbf{M}} = \frac{1}{H} \sum_{h=1}^{H} \widehat{\mathbf{M}}(\pi_h) = \frac{1}{nH} \sum_{h=1}^{H} \sum_{i=1}^{n} \nabla \hat{f}_{\pi_h}(\mathbf{x}_i) \{\nabla \hat{f}_{\pi_h}(\mathbf{x}_i)\}^\top. \tag{11}$$

Note that $\widehat{\mathbf{M}}$ in (11) is a common estimates for both (7) and (8). Furthermore, (9) in the sample level is equivalent to $\widehat{\mathbf{M}}$ since the integral can always be approximated as accurate as possible by choosing sufficiently fine grid of $\pi$. The basis set of $\mathcal{S}_{Y|\mathbf{X}}$ can be estimated by $d$ leading eigenvectors of $\widehat{\mathbf{M}}$.

As suggested by Kong and Xia (2014), one can modify (11) as

$$\widehat{\mathbf{M}} = \frac{1}{nH} \sum_{h=1}^{H} \sum_{i=1}^{n} u(\pi_h) \nabla \hat{f}_{\pi_h}(\mathbf{x}) \{\nabla \hat{f}_{\pi_h}(\mathbf{x})\}^\top,$$

with the $u(\pi) = \sum_{j=1}^{d} \nu_j / \sum_{j=1}^{p} \nu_j$, where $\nu_j$ denotes the first $j$ eigenvalues of $\widehat{\mathbf{M}}(\pi)$ in order to further improve the performance. However, it turns out in our limited simulation study that the improvement was negligible and we focused on the unweighted version (11) in the rest of article without loss of generality.

### 2.2 A Naive Estimator

Employing the reproducing kernel Hilbert space (RKHS, Wahba, 1990) generated by non-negative definite kernel $K : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$, $\mathcal{H}_K$, the solution of (2) has a finite form

$$f_\pi(\mathbf{x}) = \alpha_{\pi,0} + \sum_{i=1}^{n} \alpha_{\pi,i} K(\mathbf{x}, \mathbf{x}_i), \tag{12}$$

by *Representer Theorem* (Kimeldorf and Wahba, 1971). Plugging (12) into (2) with $J(f) = \|f\|^2_{\mathcal{H}_K}$, we have

$$(\hat{\alpha}_{\pi,0}, \hat{\boldsymbol{\alpha}}_\pi^\top)^\top = \underset{b_\pi, \boldsymbol{\alpha}_\pi}{\operatorname{argmin}} \sum_{i=1}^{n} w_\pi(y_i) L\{y_i f_\pi(\mathbf{x})\} + \frac{\lambda}{2} \boldsymbol{\alpha}_\pi^\top \mathbf{K} \boldsymbol{\alpha}_\pi, \tag{13}$$

where $\boldsymbol{\alpha}_\pi = (\alpha_{\pi,1}, \cdots, \alpha_{\pi,n})^\top$ and $\mathbf{K}$ is the $(n \times n)$-dimensional kernel matrix whose $(i,j)$th element is $K(\mathbf{x}_i, \mathbf{x}_j)$. The corresponding sample estimator $\hat{f}_\pi(\mathbf{x})$ is $\hat{\alpha}_{\pi,0} + \sum_{i=1}^{n} \hat{\alpha}_{\pi,i} K(\mathbf{x}, \mathbf{x}_i)$.

Given $\pi$, the gradient vector estimator $\nabla \hat{f}_\pi(\mathbf{x})$ is given as

$$\nabla \hat{f}_\pi(\mathbf{x}) = \frac{\partial \hat{f}_\pi(\mathbf{x})}{\partial \mathbf{x}} = \sum_{i=1}^{n} \hat{\alpha}_{\pi,i} \nabla k_i(\mathbf{x}),$$

where $\nabla k_i(\mathbf{x}) = \partial K(\mathbf{x}, \mathbf{x}_i)/\partial \mathbf{x}$. Plugging it into (10), we get the empirical estimate of $\widehat{\mathbf{M}}(\pi)$ and estimating $\widehat{\mathbf{M}}$ in (11) is then straightforward.

## 3. wOPG Estimator via Gradient Learning

Although the naive-wOPG estimator is practical, the estimator may not perform well because the differential operator makes it unstable. In particular, the theoretical justification of the naive-wOPG estimator is not straightforward unless stringent conditions are assumed, and its finite-sample performance is often unsatisfactory, as demonstrated in Section 5. To improve the naive-wOPG estimator, we propose to directly estimate $\nabla f(\mathbf{x})$ and refer it as the wOPG estimator without any prefix.

### 3.1 A Gradient-based Formulation

For the sake of brevity, we let $L_\pi\{yf(\mathbf{x})\} = w_\pi(y) L\{yf(\mathbf{x})\}$ and $g_\ell(\mathbf{x})$ be the partial derivative function of $f$ with respect to the $\ell$th predictor, i.e., $g_\ell(\mathbf{x}) = \partial f(\mathbf{x})/\partial x_\ell$, $\ell = 1, \cdots, p$ and $\nabla f(\mathbf{x}) = (g_1(\mathbf{x}), \cdots, g_p(\mathbf{x}))^\top$.

The first order Taylor expansion of $f(\mathbf{x})$ around at $\mathbf{x}'$ closed to $\mathbf{x}$ is given by

$$f(\mathbf{x}) \approx f(\mathbf{x}') + \nabla f(\mathbf{x}')^\top (\mathbf{x} - \mathbf{x}'). \tag{14}$$

Mukherjee and Wu (2006) proposed an empirical risk $\hat{\mathcal{E}}(f, \nabla f)$ based on (14) to directly estimate the gradient of $f(\mathbf{x})$:

$$\hat{\mathcal{E}}(f, \nabla f) := \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \omega_s(\mathbf{x}_i - \mathbf{x}_j) L_\pi \left[ y_i \left\{ f(\mathbf{x}_j) + \nabla f(\mathbf{x}_j)^\top (\mathbf{x}_i - \mathbf{x}_j) \right\} \right], \tag{15}$$

where $\omega_s$ denotes a smoothing kernel with a bandwidth parameter $s$. In this article, we focus on the Gaussian kernel $\omega_s(\mathbf{x}-\mathbf{u}) = (s^2)^{-p/2}\exp\{-\frac{1}{2s^2}\|\mathbf{x}-\mathbf{u}\|^2\}$, a popular choice in practice.

Assuming that $f, g_1, \cdots, g_p$ are living on $\mathcal{H}_K$, the RKHS generated by the kernel function $K$, we propose to solve

$$\min_{f,g_1,\cdots,g_p\in\mathcal{H}_K} \hat{\mathcal{E}}(f,\nabla f) + \frac{\lambda_0}{2}\|f\|^2_{\mathcal{H}_K} + \sum_{\ell=1}^{p}\frac{\lambda_\ell}{2}\|g_\ell\|^2_{\mathcal{H}_K} \tag{16}$$

where $\lambda_0$ and $\lambda_1, \cdots, \lambda_p$ are nonnegative tuning parameters that control the complexity of $f$ and elements of its gradient $\nabla f$, respectively. Note that we adaptively penalize each element of the gradient function to further improve the estimation accuracy as the adaptive LASSO (Zou, 2006; Zhang and Lu, 2007).

By the *Representer Theorem*, the solution of (16) has the following finite dimensional form:

$$f(\mathbf{x}) = \sum_{i=1}^{n}\alpha_{i0}K(\mathbf{x},\mathbf{x}_i), \quad \text{and} \quad g_\ell(\mathbf{x}) = \sum_{i=1}^{n}\alpha_{i\ell}K(\mathbf{x},\mathbf{x}_i), \ell = 1,\cdots,p,$$

where $\boldsymbol{\alpha}_\ell = (\alpha_{1\ell},\cdots,\alpha_{n\ell})^\top, \ell = 0,1,\cdots,p$. This yields

$$f(\mathbf{x}_j) + \nabla f(\mathbf{x}_j)^\top(\mathbf{x}_i - \mathbf{x}_j) = \boldsymbol{\alpha}_0^\top\mathbf{k}_j + \sum_{\ell=1}^{p}\boldsymbol{\alpha}_\ell^\top\mathbf{k}_j(x_{i\ell} - x_{j\ell}),$$

where $\mathbf{k}_j$ is the $j$-th column of the kernel matrix $\mathbf{K}$. Let us define $\delta_{ij\ell} = x_{i\ell} - x_{j\ell}$ for $\ell \neq 0$ and $\delta_{ij\ell} = 1$ for $\ell = 0$. Then we have

$$f(\mathbf{x}_j) + \nabla f(\mathbf{x}_j)^\top(\mathbf{x}_i - \mathbf{x}_j) = \sum_{\ell=0}^{p}\boldsymbol{\alpha}_\ell^\top\mathbf{k}_j\delta_{ij\ell}. \tag{17}$$

Substituting (17) into (16), we have

$$(\hat{\boldsymbol{\alpha}}_{\pi,0}, \hat{\boldsymbol{\alpha}}_{\pi,1}, \cdots, \hat{\boldsymbol{\alpha}}_{\pi,p}) = \underset{\boldsymbol{\alpha}_0,\boldsymbol{\alpha}_1,\cdots,\boldsymbol{\alpha}_p}{\operatorname{argmin}} \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} w_s(\mathbf{x}_i - \mathbf{x}_j)L_\pi\left(y_i\sum_{\ell=0}^{p}\boldsymbol{\alpha}_\ell^\top\mathbf{k}_j\delta_{ij\ell}\right) + \frac{\lambda}{2}\sum_{\ell=0}^{p}\theta_\ell\boldsymbol{\alpha}_\ell^\top\mathbf{K}\boldsymbol{\alpha}_\ell, \tag{18}$$

where $\lambda_\ell = \lambda\theta_\ell, \ell = 0,1,\cdots,p$ are reparameterized to select the tuning parameters in an efficient way. Section 4.2 presents the discussions on how to select $\lambda$ and $\theta_\ell$.

Finally, the gradient is estimated by $\nabla\hat{f}_\pi(\mathbf{x}_i) = \{\hat{g}_{\pi,1}(\mathbf{x}_i),\cdots,\hat{g}_{\pi,p}(\mathbf{x}_i)\}$ with $\hat{g}_{\pi,\ell}(\mathbf{x}_i) = \mathbf{k}_i\hat{\boldsymbol{\alpha}}_{\pi,\ell}, \ell = 1,\cdots,p$ which directly leads the corresponding wOPG estimator (11).

### 3.2 Computation

We consider two most popular loss functions: the negative binomial likelihood loss function for the logistic regression and the hinge loss function for the SVM, which we herein refer to as wOPG-LR and wOPG-SVM, respectively.

- wOPG-LR: $L_{LR}(m) = \log\{1 + \exp(-m)\}$
- wOPG-SVM: $L_{SVM}(m) = \max(0, 1 - m)$

Although it looks canonical, the optimization problem (18) may not be simple to solve even for well-shaped loss functions due to a large number of parameters. To solve (18) with the wOPG-LR, we proposed the blockwise Newton-Raphson algorithm, which can be applied to any convex differentiable loss functions, such as the exponential loss for boosting and the large-margin unified loss (Liu et al., 2011). For the wOPG-SVM with the hinge loss which is not differentiable, we rewrite (18) in a constraint form to apply the Lagrangian method. We first show that the corresponding dual problem turns out to be the quadratic programming with box constraint only and then employ the box-constraint coordinate ascent method (Wright, 2015) to update the parameters coordinately. The complete algorithms for both wOPG-LR and wOPG-SVM are described in Appendix B.

Finally, we remark that the linear approximation (14) is not entirely accurate unless $\|\mathbf{x}_i - \mathbf{x}_j\|$ is small. This leads us to focus only on small number (say $k \ll n$) of data points in the nearest neighborhood of $\mathbf{x}_i$, which further facilitates the computation in practice.

### 3.3 Asymptotic Analysis

In what follows, we prove $\widehat{\mathbf{M}}$ obtained in Section 3.1 via gradient learning converges to its population counter part $\mathbf{M}$. Let $\hat{f}_\pi$ and $\nabla \hat{f}_\pi = (\hat{g}_{\pi,1}, \cdots, \hat{g}_{\pi,p})^\top$ be the minimizer of (16) or equivalently (18), and define

$$f_\pi = \underset{f}{\operatorname{argmin}}\, E\left[L_\pi\left\{Y f(\mathbf{X})\right\}\right] \tag{19}$$

and $\nabla f_\pi$ as the gradient of $f_\pi$. Note that $f_\pi(\mathbf{x})$ that solves (4) is a pointwise solution of (19) for a given $\mathbf{X} = \mathbf{x}$.

Let $\mathrm{P}(\mathbf{x})$ and $\rho(\mathbf{x})$ denote respectively the distribution and density function of $\mathbf{X}$ whose support is denoted by $\mathcal{X}$, respectively. To explore the asymptotic behavior of $(\hat{f}_\pi, \nabla \hat{f}_\pi)$, we assume the following conditions.

(C1) The support $\mathcal{X}$ is a compact subset of $\mathbb{R}^p$, and for some constant $c_\rho > 0$ which depends only on $\rho(\mathbf{x})$,

$$P\left\{d(\mathbf{X}, \partial\mathcal{X}) < s\right\} \le c_\rho s,$$

where $\partial\mathcal{X}$ denotes the boundary of $\mathcal{X}$ and $d(\mathbf{x}, \partial\mathbf{X})$ is the distance between $\mathbf{x} \in \mathcal{X}$ and $\partial\mathcal{X}$.

(C2) The density $\rho(\mathbf{x})$ satisfies

$$\sup_{\mathbf{x}} \rho(\mathbf{x}) \le c_\rho, \quad \text{and} \quad |\rho(\mathbf{x}) - \rho(\mathbf{u})| \le c_\rho |\mathbf{x} - \mathbf{u}|^\tau, \quad \forall \mathbf{u}, \mathbf{x} \in \mathcal{X}$$

for some $0 < \tau \le 1$.

(C3) The gradient of the true classification function $f_\pi$ exists, and $f_\pi$ and each element of $\nabla f_\pi$ reside on $\mathcal{H}_K$, i.e., $(f_\pi, \nabla f_\pi) \in \mathcal{H}_K^{p+1}$.

(C4) There exists a constant $c_f > 0$ that depends only on $f$ such that

$$\left|f_\pi(\mathbf{u}) - f_\pi(\mathbf{x}) + \nabla f_\pi(\mathbf{u})^\top(\mathbf{x}-\mathbf{u})\right| \le c_f(\mathbf{u}-\mathbf{x})^2, \quad \forall \mathbf{u},\mathbf{x} \in \mathcal{X}.$$

In condition (C1), we assume the compactness of support $\mathcal{X}$ for the technical simplicity frequently used in the literature on nonparametric models (Ye and Xie, 2012; Yang et al., 2016). The behavior of the marginal distribution of $\mathbf{X}$ near the boundary $\partial\mathbf{X}$ is also specified, and this is implied by (C2) if the boundary $\partial\mathbf{X}$ is piecewise smooth. Condition (C2) states that the density of the marginal distribution is Hölder continuous with an exponent $\tau$. Condition (C3) is rather a standard assumption in the RKHS theory. Condition (C4) ensures that the linear approximation of the gradient learning does not fail. Reference was made to Mukherjee and Wu (2006) for further details about these conditions.

In order to show the consistency of $\widehat{\mathbf{M}}$, we first compute an error bound of $(\hat{f}_\pi, \nabla\hat{f}_\pi)$. Recall that we introduce two different loss functions, and the following theorem provides an $L_2$-type error bound of $(\hat{f}_\pi, \nabla\hat{f}_\pi)$ for wOPG-LR with the logistic loss $L_{LR}$ that is a direct consequence of Theorem 9 in Mukherjee and Wu (2006). Thus, we omit the proof.

**Theorem 3** *Let $(\hat{f}_\pi, \nabla\hat{f}_\pi)$ and $(f_\pi, \nabla f_\pi)$ be the minimizer of (16) and (19), respectively with $L(m) = L_{LR}(m)$. Choose $s = n^{-\frac{1}{3(p+2+2\tau)}}$ and $\lambda = s^{2+2\tau}$. Under conditions (C1)-(C4), for an arbitrary given $\eta \in (0, 1/2)$, there exists a constant $C_{LR}$ that depends only on a given $\eta$ such that with probability greater than $1-2\eta$,*

$$\max\left\{E\left[(\hat{f}_\pi(\mathbf{X}) - f_\pi(\mathbf{X}))^2\right], E\left[\|\nabla\hat{f}_\pi(\mathbf{X}) - \nabla f_\pi(\mathbf{X})\|_2^2\right]\right\} \le C_{LR}\left(\frac{1}{n}\right)^{\frac{\tau}{3(p+2+2\tau)}}. \quad (20)$$

Theorem 3 provides a justification of the gradient formulation (18) with $L_{LR}$. However its proof depends on the Taylor expansion of the risk and the result is not applicable to non-differentiable function such as $L_{SVM}$. Now, we establish an analogous version of Theorem 3 for the wOPG-SVM with the hinge loss, which is not a trivial extension of Theorem 3. Accordingly, we first need to modify (C2) as follows.

(C2′) (C2) holds for $0 < \tau < \frac{1}{2}$.

Note that (C2′) is slightly stronger than (C2). In addition, we require additional condition (C5) for the wOPG-SVM. Given $\pi \in (\delta, 1-\delta)$, let us define a set $\mathcal{S}_\gamma = \mathcal{L}_{\pi,\gamma} \cup \mathcal{U}_{\pi,\gamma} \cup \mathcal{M}_{\pi,\gamma}$ where

$$\mathcal{L}_{\pi,\gamma} = \{\mathbf{x} : (1-\pi)p(\mathbf{x}) < \gamma\}, \mathcal{U}_{\pi,\gamma} = \{\mathbf{x} : \pi(1-p(\mathbf{x})) < \gamma\} \quad \text{and} \quad \mathcal{M}_{\pi,\gamma} = \{\mathbf{x} : |\pi - p(\mathbf{x})| < \gamma\}.$$

For a sufficiently small $\gamma$, $f_\pi(\mathbf{x})$ is always positive (resp. negative) if $\mathbf{x} \in \mathcal{L}_{\pi,\gamma}$ (resp. $\mathcal{U}_{\pi,\gamma}$), or is a (cost-weighted) random classifier if $\mathbf{x} \in \mathcal{M}_{\pi,\gamma}$.

(C5) Given $\pi \in (\delta, 1-\delta)$, there exists a constant $\gamma > 0$

$$P(\mathbf{X} \in \mathcal{S}_\gamma^c) > 0, \quad (21)$$

where $\mathcal{S}_\gamma = \bigcup_{\pi\in(\delta,1-\delta)} \mathcal{S}_{\pi,\gamma}$ , and

$$k\int_{\mathcal{S}_\gamma} |f(\mathbf{x}) - f_\pi(\mathbf{x})| d\mathrm{P}(\mathbf{x}) \le \int_{\mathcal{S}_\gamma^c} |f(\mathbf{x}) - f_\pi(\mathbf{x})| d\mathrm{P}(\mathbf{x}), \qquad \forall f \in \mathcal{H}_K \quad (22)$$

for some constant $k > 0$.

In (C5), (21) prevents that $\mathbf{X}$ lies only on $\mathcal{S}_{\pi,\gamma}$ for some $\pi$ and (22) is implied by $\int |f(\mathbf{x}) - f_\pi(\mathbf{x})| d\mathrm{P}(\mathbf{x}) < \infty$. Both (21) and (22) are mild for $\pi \in (\delta, 1-\delta)$ even for sufficiently small $\delta > 0$ given.

**Theorem 4** *Let $(\hat{f}_\pi, \nabla\hat{f}_\pi)$ and $(f_\pi, \nabla f_\pi)$ be the minimizer of* (16) *and* (19) *with the hinge loss $L(m) = L_{SVM}(m)$, respectively. Under conditions (C1), (C2′), (C3)–(C5), for $s = n^{-\frac{1}{3(p+1+2\tau)}}$ and $\lambda = s^{1+2\tau}$, for an arbitrary given $\eta \in (0, 1/2)$, there exists a constant $C_{SVM}$ that depends only on a given $\eta$ such that with probability greater than $1-2\eta$,*

$$\max\left\{E\left[|\hat{f}_\pi(\mathbf{X}) - f_\pi(\mathbf{X})|\right], E\left[\|\nabla\hat{f}_\pi(\mathbf{X}) - \nabla f_\pi(\mathbf{X})\|_1\right]\right\} \leq C_{SVM}\left(\frac{1}{n}\right)^{\frac{\tau}{3(p+1+2\tau)}}. \quad (23)$$

We have a couple of remarks on our theoretical results. First, in Theorem 4 (and Theorem 3), we reach identical convergence rate for both $\hat{f}_\pi$ and $\nabla\hat{f}_\pi$, which looks counter-intuitive given that the gradient function $\nabla\hat{f}_\pi$ is less smoother than its original function $f_\pi$. We derive our results based on the standard RKHS theory under the condition (C3) for technical simplicity, which yields similar upper bounds of excess errors for both functions residing on RKHS. In addition, we have very slow rate of convergence in both theorems. this is mainly because we estimate the classification function and its gradient function, and the theorems try to bound all of them simultaneously. As a result, the corresponding excess error depends on the $(p+1)$-dimensional quantity which yields a very slow rate of convergence even when $p$ is moderately large. Similar results can be found in the literature about the gradient learning (Mukherjee and Wu, 2006; Ye and Xie, 2012; Yang et al., 2016). Although the theoretical learning rate of the gradient estimator is slow, our empirical result in Section 5 shows that the proposed method performs reasonably well with moderate sample size $n$.

Second, Theorem 3 and Theorem 4 seem to provide different convergence rates since the former bounds the squared loss while the latter bounds the absolute loss. However, note that Theorem 4 for the hinge loss requires a stronger condition (C2′) that controls the Hölder constant $\tau$ for the marginal density of $\mathbf{x}$, $\rho(\mathbf{x})$. Considering the most favorable scenario for each cases, i.e., $\tau = 1$ for Theorem 3 and $\tau = 1/2$ for Theorem 4, we obtain comparable converge rates in terms of the squared loss, $n^{-\frac{1}{3(p+4)}}$ for the logistic loss and $n^{-\frac{1}{3(p+2)}}$ for the hinge loss. In other words, the hinge loss requires a stronger condition than the logistic loss to achieve a similar rate of convergence, which is sensible given that the hinge loss is less smooth.

Finally, we establish the consistency of the wOPG working matrix as given in Corollary 5.

**Corollary 5** *Assume the conditions in Theorem 3 for wOPG-LR and in Theorem 4 for wOPG-SVM for all $\pi$ on $(\delta, 1-\delta)$. Furthermore, assume that both $\widehat{\mathbf{M}}(\pi)$ and $\mathbf{M}(\pi)$ are Lebesgue measurable with respect to $\pi$ on $(\delta, 1-\delta)$. Then $\widehat{\mathbf{M}}_T$ in* (10) *converges to $\mathbf{M}_T$* (8) *in probability, that is*

$$\widehat{\mathbf{M}}_T \quad \to_p \quad \mathbf{M}_T.$$

## 4. Additional Issues

In the subsequent sections, we describe how to select tuning parameters related to the wOPG estimator.

### 4.1 Structural Dimension Estimation

The structural dimension of $\mathcal{S}_{Y|\mathbf{X}}$, $d$ is another important quantity to be estimated from the data. There are two popular approaches. The first type is based on the eigenvalues of the working matrix. For example, Li (1991) proposed a sequential test based on the cumulative sum of eigenvalues, and Li et al. (2011) proposed the use of a BIC-type criterion; a penalized version of the cumulative sum of eigenvalues. On the other hand, one can use eigenvectors to estimate $d$. For example, Ye and Weiss (2003) proposed to choose the first $d$ eigenvectors that attain the smallest sampling variability, estimated by bootstrapping.

In this article, we propose to employ the ladle estimator (Luo and Li, 2016) that elegantly combines the two ideas as follows. For a given $k < \lceil p/\log p \rceil$ where $\lceil a \rceil$ denotes the largest integer that is not larger than $a$, we set $\widehat{\mathbf{V}}_k = (\mathbf{v}_1, \cdots, \mathbf{v}_k)$ denote $p \times k$-dimensional matrix of the first $k$ leading eigenvectors of $\widehat{\mathbf{M}}$. To measure the variability of $\widehat{\mathbf{M}}$, take $n$ bootstrap samples denoted by $\{(y_i^{j,*}, \mathbf{x}_i^{j,*}), i = 1, \cdots, n\}, j = 1, \cdots, n$, and compute $\widehat{\mathbf{M}}^{j,*}$ and $\widehat{\mathbf{V}}^{j,*}$ for the $j$th bootstrap sample analogously defined as $\widehat{\mathbf{M}}_k$ and $\widehat{\mathbf{V}}_k$ for the original data, respectively. Let us define

$$f_n^0(k) = \frac{1}{n} \sum_{i=1}^{n} \left\{ 1 - \det\left( \widehat{\mathbf{V}}_k^\top \widehat{\mathbf{V}}_k^{j,*} \right) \right\}, k = 1, \cdots, p-1$$

with $f_n^0(0) = 0$. Luo and Li (2016) showed that $f_n^0(k)$ is expected to be minimized at $k = d$ and gets large for $k > d$. Finally, the ladle estimator $\hat{d}_{\text{ladle}}$ is the minimizer of $f_n(k) + \phi_n(k)$ where

$$f_n(k) = \frac{f_n^0(k)}{1 + \sum_{l=0}^{p-1} f_n^0(l)} \text{ and } \phi_n(k) = \frac{\hat{\lambda}_{k+1}}{1 + \sum_{l=1}^{p} \hat{\lambda}_l}.$$

with $\hat{\lambda}_l$ being the $l$th largest eigenvalue of $\widehat{\mathbf{M}}$, $l = 1, \cdots, p$.

### 4.2 Tuning Parameter Selection

Motivated by the adaptive LASSO (Zou, 2006; Zhang and Lu, 2007), there are two layers of regularization parameters in (18), $\lambda$ and $(\theta_0, \cdots, \theta_p)$. The adaptive penalty mitigates the bias of the gradient estimator by less penalizing the variables containing more information about $\mathcal{S}_{Y|\mathbf{X}}$. This leads us to set $\theta_\ell = 1/(\mathbf{P}_{\tilde{\mathbf{B}}})_{ii}$ where $(\mathbf{P}_{\tilde{\mathbf{B}}})_{ii}$ denotes the $i$th diagonal element of the projection matrix on span$\{\tilde{\mathbf{B}}\}$ with $\tilde{\mathbf{B}}$ being the one obtained from the naive-wOPG estimator. The constant term $f$ is not directly related to the SDR context and we propose to set $\theta_0 = 1$. The global tuning parameter $\lambda$ is then selected by the conventional cross-validation to obtain a set of sensible classifiers that minimizes the misclassification cost.

We have another tuning parameter $s$ in the Gaussian smoothing kernel $w_s(\mathbf{x}_i - \mathbf{x}_j)$. We set $s$ to be the median of the standard deviations of $\mathbf{x}_i$ as suggested by Mukherjee and Wu (2006) and Yang et al. (2016).

The choice of $H$ may be critical to controlling the computational complexity of the proposed method. Although in principle, the estimator should perform better as $H$ increases, our limited numerical results show that the finite sample performance of the wOPG estimator is not overly sensitive to the choice of $H$. In this article, we set $\pi_h = h/(H+1), h = 1, \cdots, H$ with $H = 9$, i.e., $\pi_h = h/10, h = 1, \cdots, 0$, and thus $\delta = 0.1$.

## 5. Simulation

In this section, we conduct a simulation study to evaluate the finite-sample performance of the wOPG method. We set $\pi_h = h/10, h = 1, \cdots, 9$, and employ the Gaussian kernel $K(\mathbf{x}, \mathbf{x}') = \exp\{-\|\mathbf{x} - \mathbf{x}'\|^2/(2\sigma^2)\}$ for the RKHS with $\sigma$ being the median of the pairwise distances between the predictors in the positive and negative classes (Jaakkola et al., 1999). Tuning parameters $\lambda$ and $\theta_\ell, \ell = 0, 2, \cdots, p$, and the bandwidth parameter $s$ are chosen as described in Section 4.2.

For the data generation, we assume the following model

$$y_i = \text{sign}\{f(\mathbf{x}_i) + 0.2\epsilon_i\}$$

where $\mathbf{x}_i \overset{\text{iid}}{\sim} N_p(\mathbf{0}_p, \mathbf{I}_p), i = 1, \cdots, n$, and $\epsilon_i \overset{\text{iid}}{\sim} N(0,1), i = 1, \cdots, n$. We consider the following three models

- Model (I) $f(\mathbf{x}) = x_1/\{0.5 + (x_2 + 1)^2\}$
- Model (II) $f(\mathbf{x}) = (x_1 + 0.5)(x_2 - 0.5)^2$
- Model (III) $f(\mathbf{x}) = \sin(x_1)/\exp(x_2)$

All models share a common central subspace, $\text{span}(\mathbf{B})$ where $\mathbf{B} = (\mathbf{e}_1, \mathbf{e}_2)$ with $\mathbf{e}_i$ being a unit vector whose $i$th element is 1 and 0 for others, i.e., $\mathbf{e}_1^\top \mathbf{x} = x_1$ and $\mathbf{e}_2^\top \mathbf{x} = x_2$. Thus we have $d = 2$ which we assume to be known in this section for the fair comparison of different SDR methods.

As competitors, we include the sliced averaged variance estimation (SAVE, Cook and Weisberg, 1991), the principal Hessian direction (pHd, Li, 1992), and the direction regression (DR, Li and Wang, 2007). These are popular SDR methods but they are originally designed for the continuous response. Note that we exclude the most popular SIR since it can at most identify one direction, and thus fail for all models under consideration. We also include recently proposed SDR methods for the binary response: the probability-enhanced SIR (PRE, Shin et al., 2014), and the linear principal weighted support vector machines (PWSVM, Shin et al., 2017).

To evaluate the performance of SDR methods, we use the distance between $\widehat{\mathbf{B}}$ and $\mathbf{B}$ as $d(\widehat{\mathbf{B}}, \mathbf{B}) = \left\|\mathbf{P}_{\widehat{\mathbf{B}}} - \mathbf{P}_{\mathbf{B}}\right\|_F$ where $\|\cdot\|_F$ denotes the Frobenius norm of a matrix. The smaller value of $d(\widehat{\mathbf{B}}, \mathbf{B})$ indicates the better performance for estimating $\mathcal{S}_{Y|\mathbf{X}}$. Table 1 contains the averaged $d(\widehat{\mathbf{B}}, \mathbf{B})$ over 100 independent repetitions under the models (I)–(III) with different combinations of $(n, p) \in \{500, 1000\} \times \{10, 20\}$.

In Table 1, one can observe that SAVE, pHd, and DR developed under the regression context exhibit worse performance than the rest of the methods carefully designed for the binary response. Among the SDR methods for binary classification, the wOPG method

clearly outperforms both the PRE and PWSVM. The naive versions still perform reasonably well but occasionally worse than PWSVM, especially when $n$ is not large. We also highlight that wOPG-SVM shows slightly better results than wOPG-LR for all the scenarios under consideration. We believe that it is possibly connected to the fact that SVM generally outperforms the logistic regression in terms of classification.

Table 1: Averaged $d(\widehat{\mathbf{B}}, \mathbf{B})$ over 100 independent repetitions for Model (I)–(III). Corresponding standard deviations are given in parentheses.

| | $n$ | $p$ | SAVE | PHD | DREG | PRE | PWSVM | Naive wOPG | | wOPG | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | LR | SVM | LR | SVM |
| (I) | 500 | 10 | 1.28 (.16) | 1.54 (.19) | 1.29 (.14) | 1.12 (.22) | 0.74 (.19) | 0.83 (.24) | 0.81 (.24) | 0.51 (.35) | 0.34 (.34) |
| | | 20 | 1.39 (.06) | 1.80 (.11) | 1.35 (.07) | 1.31 (.12) | 1.03 (.15) | 1.32 (.12) | 1.29 (.14) | 0.99 (.41) | 0.84 (.49) |
| | 1000 | 10 | 1.16 (.25) | 1.32 (.25) | 1.30 (.13) | 0.89 (.24) | 0.56 (.13) | 0.49 (.13) | 0.48 (.13) | 0.26 (.09) | 0.10 (.07) |
| | | 20 | 1.35 (.10) | 1.69 (.14) | 1.36 (.07) | 1.24 (.15) | 0.78 (.13) | 1.05 (.21) | 0.94 (.21) | 0.44 (.29) | 0.29 (.31) |
| (II) | 500 | 10 | 1.26 (.19) | 1.38 (.19) | 1.24 (.15) | 1.06 (.23) | 1.02 (.20) | 1.04 (.23) | 0.92 (.25) | 0.65 (.40) | 0.45 (.41) |
| | | 20 | 1.50 (.11) | 1.67 (.14) | 1.32 (.09) | 1.33 (.11) | 1.20 (.13) | 1.40 (.07) | 1.39 (.09) | 1.33 (.22) | 1.16 (.37) |
| | 1000 | 10 | 1.09 (.25) | 1.19 (.28) | 1.14 (.20) | 0.88 (.22) | 0.77 (.17) | 0.57 (.17) | 0.45 (.14) | 0.28 (.15) | 0.11 (.14) |
| | | 20 | 1.36 (.11) | 1.47 (.12) | 1.27 (.12) | 1.27 (.13) | 1.02 (.12) | 1.26 (.14) | 1.19 (.19) | 0.90 (.37) | 0.67 (.48) |
| (III) | 500 | 10 | 1.25 (.19) | 1.49 (.20) | 1.30 (.11) | 1.10 (.22) | 0.79 (.20) | 0.85 (.25) | 0.82 (.27) | 0.51 (.35) | 0.35 (.42) |
| | | 20 | 1.38 (.07) | 1.72 (.15) | 1.36 (.06) | 1.32 (.11) | 1.03 (.13) | 1.29 (.15) | 1.26 (.17) | 0.94 (.46) | 0.82 (.50) |
| | 1000 | 10 | 1.16 (.25) | 1.30 (.26) | 1.31 (.12) | 0.91 (.23) | 0.57 (.13) | 0.50 (.12) | 0.50 (.12) | 0.27 (.10) | 0.10 (.06) |
| | | 20 | 1.36 (.08) | 1.59 (.18) | 1.36 (.07) | 1.25 (.15) | 0.81 (.12) | 1.03 (.19) | 0.94 (.20) | 0.47 (.33) | 0.27 (.29) |

We note that both the PRE and PWSVM are SIR-like SDR methods based on the inverse mean. Hence, they require the linearity condition and fail when the classification boundary is symmetric about the origin. However, the proposed wOPG estimators are the forward method which do not require such conditions and exhaustively estimate $\mathcal{S}_{Y|\mathbf{X}}$ even when the classification boundary is symmetric about the origin. In order to confirm this, we additionally consider the following two models:

- Model (IV) $f(\mathbf{x}) = x_1(x_1 + x_2 + 1)$
- Model (V) $f(\mathbf{x}) = \left(x_1^2 + x_2^2\right)^{1/2} \log\left(x_1^2 + x_2^2\right)^{1/2}$

The central subspaces of these two models are identical to those of the models (I)–(III), yet the class boundaries of models (IV) and (V) are approximately and exactly symmetric about the origin, respectively. Under models (IV) and (V), one can observe that both the PRE and PWSVM clearly fail while the second-moment-based methods such as SAVE and pHd perform quite well compared to models (I)–(III). However, wOPG-SVM and wOPG-LR show much better performance. Again, the wOPG-SVM shows better performance than wOPG-LR under (IV) and (V).

## 6. Illustration to Real Data

We applied both wOPG-LR and wOPG-SVM to Breast Cancer Coimbra (BCC) data available at the UCI machine learning repository (`https://archive.ics.uci.edu/ml/index.`

Table 2: Averaged $d(\widehat{\mathbf{B}}, \mathbf{B})$ over 100 independent repetitions for Model (IV)–(V) whose classification boundaries are nearly or exactly symmetric. Standard deviations are given in parentheses.

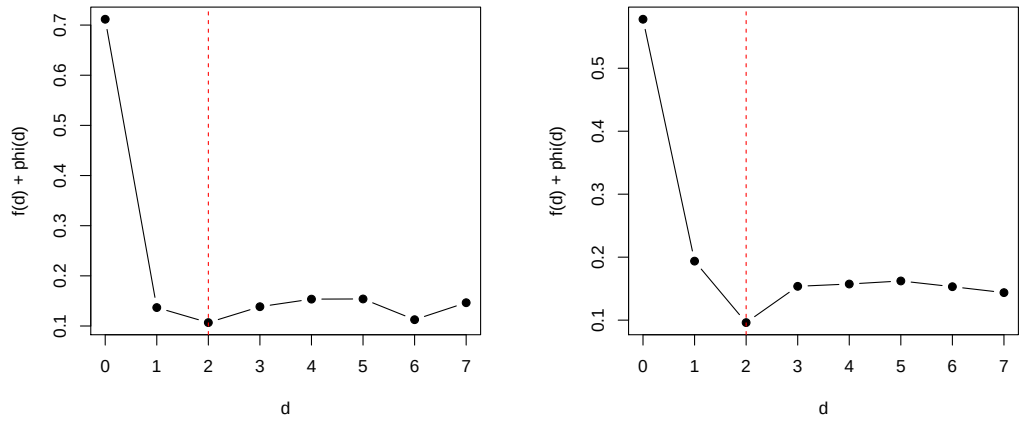
| | $n$ | $p$ | SAVE | PHD | DREG | PRE | PWSVM | Naive wOPG | | wOPG | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | LR | SVM | LR | SVM |
| (IV) | 500 | 10 | 0.77 (.27) | 0.68 (.19) | 0.63 (.16) | 0.58 (.12) | 0.52 (.10) | .41 (.09) | 0.38 (.08) | 0.26 (.07) | 0.11 (.06) |
| | | 20 | 1.18 (.20) | 1.06 (.21) | 0.86 (.14) | 0.95 (.11) | 0.78 (.11) | .77 (.10) | 0.73 (.10) | 0.40 (.10) | 0.19 (.08) |
| | 1000 | 10 | 0.55 (.18) | 0.49 (.14) | 0.48 (.13) | 0.45 (.09) | 0.40 (.08) | .26 (.05) | 0.25 (.05) | 0.18 (.05) | 0.06 (.04) |
| | | 20 | 0.89 (.21) | 0.76 (.16) | 0.68 (.10) | 0.75 (.09) | 0.58 (.06) | .52 (.07) | 0.48 (.07) | 0.31 (.06) | 0.12 (.04) |
| (V) | 500 | 10 | 0.27 (.05) | 0.28 (.05) | 1.36 (.10) | 1.70 (.17) | 1.59 (.13) | .25 (.04) | 0.24 (.04) | 0.14 (.06) | 0.14 (.04) |
| | | 20 | 0.42 (.06) | 0.45 (.07) | 1.41 (.09) | 1.80 (.12) | 1.57 (.07) | .68 (.18) | 0.53 (.08) | 0.29 (.09) | 0.17 (.08) |
| | 1000 | 10 | 0.19 (.04) | 0.20 (.04) | 1.36 (.09) | 1.67 (.17) | 1.62 (.15) | .16 (.03) | 0.16 (.03) | 0.11 (.03) | 0.10 (.03) |
| | | 20 | 0.29 (.04) | 0.29 (.04) | 1.41 (.08) | 1.79 (.12) | 1.58 (.06) | .34 (.04) | 0.31 (.03) | 0.18 (.05) | 0.11 (.03) |

`php`). The BCC data contains breast cancer diagnosis results for 116 patients with nine continuous predictors including age, body mass index, and seven measurements from the blood test, i.e., glucose, insulin, homeostatic model assessment (HOMA), leptin, adiponectin, resistin, and monocyte chemoattractant protein-1 (MCP-1). The goal is to construct a sensible binary classifier that predicts breast cancer based on the given covariates. See Hosni et al. (2019) for more details about the data.

To apply the wOPG methods, we set all the related tuning parameters as done in Section 5. In order to determine the structural dimension $d$, the ladle estimator described in Section 4.1 is employed. Both the wOPG-LR and wOPG-SVM produce similar ladle plots, which yield $d = 2$. See Figure 1.

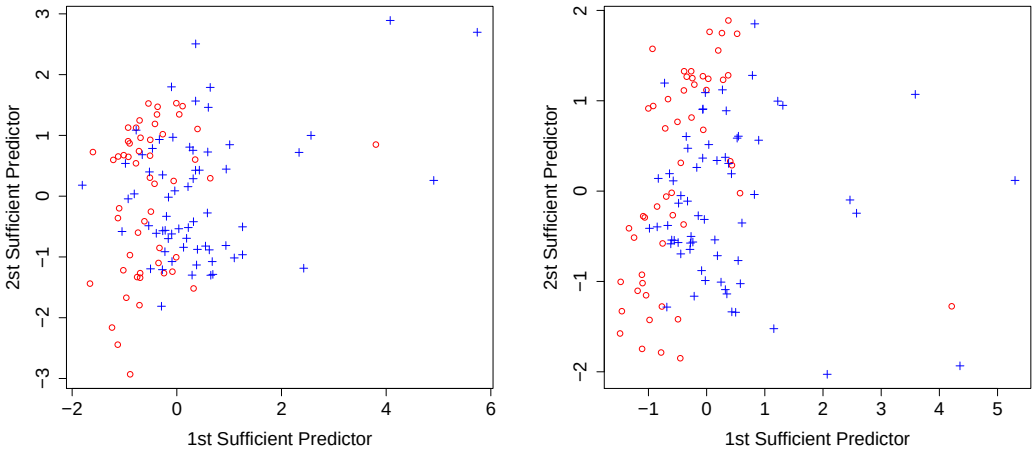
Figure 2 depicts scatter plots of the first two sufficient predictors, $\mathbf{x}_i^\top \hat{\mathbf{b}}_1$ vs $\mathbf{x}_i^\top \hat{\mathbf{b}}_2$ where (red) circles and (blue) cross represent negative and positive classes, respectively. One can conclude that that both methods successfully achieve the dimension reduction from $p$ to 2 since the two classes look well-separated on the estimated $\mathcal{S}_{Y|\mathbf{X}}$.

Finally, we conducted a validation study in order to evaluate the effect of SDR in terms of classification performance. Toward this, we randomly split the data into training and test sets denoted by $\mathcal{D}^{\text{tr}} = \{(y_1^{\text{tr}}, \mathbf{x}_1^{\text{tr}}), \cdots, (y_{58}^{\text{tr}}, \mathbf{x}_{58}^{\text{tr}})\}$ and $\mathcal{D}^{\text{ts}} = \{(y_1^{\text{ts}}, \mathbf{x}_1^{\text{ts}}), \cdots, (y_{58}^{\text{ts}}, \mathbf{x}_{58}^{\text{ts}})\}$, respectively. We then applied various SDR methods to $\mathcal{D}^{\text{tr}}$ and obtained the estimated basis of $\mathcal{S}_{Y|\mathbf{X}}$ denoted by $\widehat{\mathbf{B}}_{\text{tr}}$. In order to check the classification performance on the estimated $\mathcal{S}_{Y|\mathbf{X}} = \text{span}(\widehat{\mathbf{B}}_{\text{tr}})$, we trained the $K$-nearest-neighbor classifier with $K = 5$ from $(y_j^{\text{tr}}, \widehat{\mathbf{B}}_{\text{tr}}^\top \mathbf{x}_j^{\text{tr}})$ and calculated the test error rate to predict $y^{\text{ts}}$ from $\widehat{\mathbf{B}}_{\text{tr}}^\top \mathbf{x}_j^{\text{ts}}$. These steps were repeated independently for a hundred times, and Figure 3 compares the boxplots of test error rates for different SDR methods. The first boxplot represents the benchmark performance when no SDR is applied. Since $d$ is unknown, we select $d$ for other competing methods to have the highest test accuracy. While all methods perform quite well in the sense that the classification performance is comparable to the case when no SDR is applied, the proposed wOPG methods showed the best performance, which is concordant to what we have seen in Section 5




(a) wOPG-LR (b) wOPG-SVM

Figure 1: Structural Dimension Estimation for BCC data: The ladle plot which depicts $f_n(d) + \phi_n(d)$ as a function of $d$ is presented for wOPG-LR (a) and wOP-SVM (b). The Vertical line represents the selected $d = 2$ that minimizes $f_n(d) + \phi_n(d)$.




(a) wOPG-LR (b) wOPG-SVM

Figure 2: wOPG methods applied to BCC data: Scatter plots of the first two sufficient predictors estimated by wOPG-LR and wOPG-SVM are depicted, respectively.

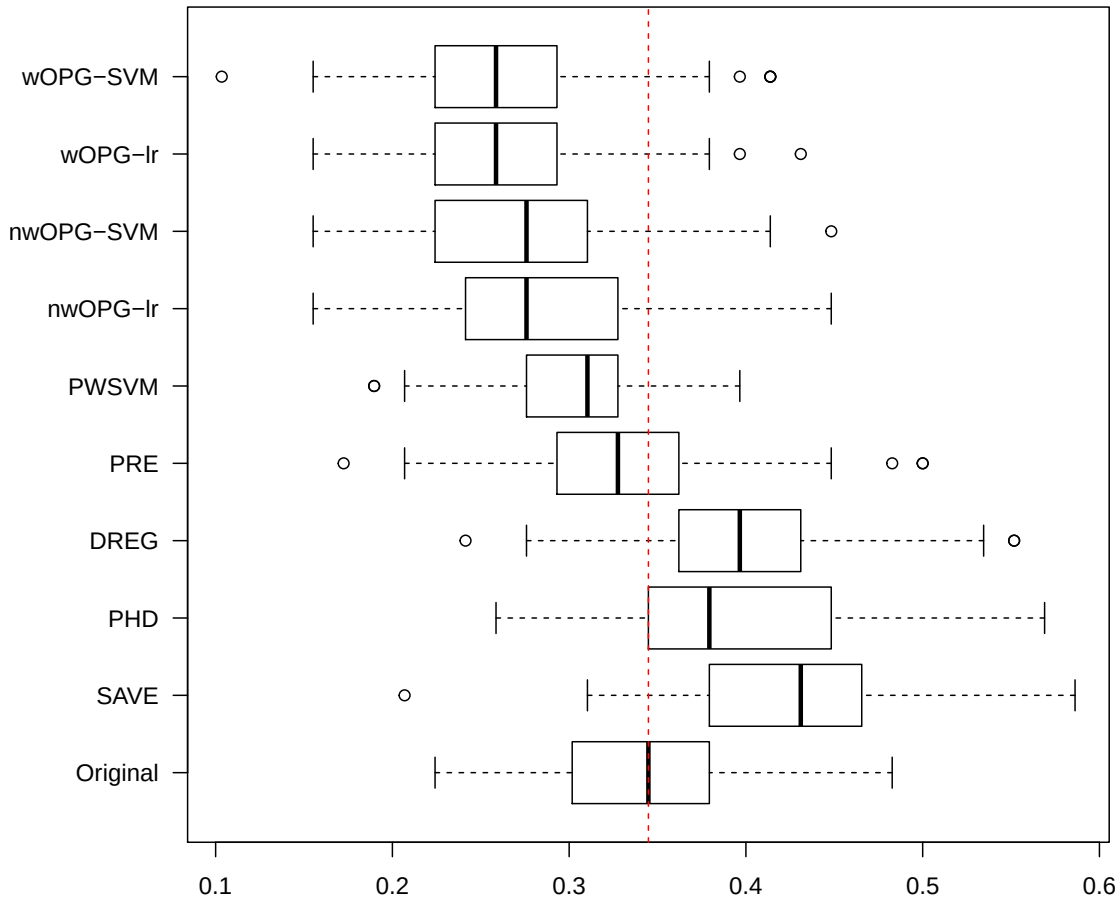



Figure 3: Boxplots of test error rates of KNN classifier with $K = 5$ trained on estimated $\mathcal{S}_{Y|\mathbf{X}}$ by different SDR methods, over 100 random partitioning for the BCC data: The proposed wOPG-LR and wOPG-SVM outperform all other methods. The horizontal (dotted) line corresponds to the average test error rate of the KNN classifier, which uses original predictors without SDR.

## 7. Concluding Remarks

In this study, we propose a novel forward approach for SDR in binary classification, called the wOPG estimator. The proposed estimator exhibits a clear advantage over the popular inverse method since it does not require additional conditions on the predictor, which are essential for the inverse methods. Moreover, wOPG can exhaustively estimate $\mathcal{S}_{Y|\mathbf{X}}$.

For estimating the wOPG working matrix, we employed the kernel trick on RKHS, a quite popular method in machine learning communities, and developed efficient algorithms applicable to a variety of convex loss functions. In addition, the asymptotic behavior of the gradient function of the classifier trained from the hinge loss has been studied, which, to the best of our knowledge, is yet to be explored in the literature.

The proposed wOPG is a general methodology that can be applied to any Fisher consistent loss functions, such as the exponential loss for boosting, the large-marge unified machine loss (Liu et al., 2011), and the $\phi$-loss (Shen et al., 2003), to name a few. In addition, it is possible to extend the wOPG idea to the multiclass problem. Fisher consistency in multiclass classification has been studied for the truncated large-margin classifiers (Wu et al., 2010) and the angle-based large-margin classifiers (Zhang and Liu, 2014). The corresponding wOPG estimator can be naturally developed for the SDR with a categorical response, which warrants further investigation.

## Acknowledgments

We thank the action-editor, Ryan Tibshirani, and two anonymous reviewers for their constructive comments and suggestions which have significantly improved the article. This work is supported by National Research Foundation of Korea (NRF) grant funded by the Korea government (MIST), grant numbers 2018R1D1A1B07043034 and 2019R1A4A1028134.

## Appendix A. Technical Proofs

### A.1 Proof of Theorem 2

Since we have (6), it suffices to show that $\mathcal{S}_{Y|\mathbf{X}} \subseteq \text{span}(\mathbf{M})$. By the definition of $\mathbf{M}$, we have

$$\mathbf{M} = \mathbf{B}\mathbf{W}\mathbf{B}^\top.$$

where

$$\mathbf{W} = \int_0^1 E\left[\nabla \tilde{f}_\pi(\mathbf{B}^\top\mathbf{X})\{\nabla \tilde{f}_\pi(\mathbf{B}^\top\mathbf{X})\}^\top\right] d\pi.$$

Now, we shall show that the matrix $\mathbf{W}$ is of full rank which leads $\mathcal{S}_{Y|\mathbf{X}} \subseteq \text{span}(\mathbf{M})$.

Suppose $\mathbf{W}$ is not a full-rank matrix, then there exists a vector $\mathbf{u}_1 \in \mathbb{R}^d$ such that $\mathbf{u}_1^\top \mathbf{W}\mathbf{u}_1 = 0$, which leads

$$\mathbf{u}_1^\top \nabla \tilde{f}_\pi(\mathbf{B}^\top\mathbf{X}) = 0, \qquad \text{almost sure} \tag{24}$$

for all $\pi \in (0, 1)$.

Let $\mathbf{U} = (\mathbf{u}_1, \cdots, \mathbf{u}_d) \in \mathbb{R}^{p\times d}$ denote a set of orthonormal basis for $\mathbb{R}^d$. Let, for a given $\pi \in (0,1)$,

$$h_\pi(\mathbf{v}) := \tilde{f}_\pi(\mathbf{v}), \qquad \tilde{h}_\pi(\mathbf{v}) = \tilde{f}_\pi(\mathbf{U}\mathbf{v}), \qquad \tilde{\mathbf{B}} = \mathbf{B}\mathbf{U},$$

then we have

$$h_\pi(\mathbf{U}\mathbf{v}) = \tilde{h}_\pi(\mathbf{v}) \quad \text{and} \quad \tilde{h}_\pi(\mathbf{B}^\top\mathbf{X}) = h_\pi(\tilde{\mathbf{B}}^\top\mathbf{X}).$$

The gradient vector of $\tilde{h}_\pi(\mathbf{v})$ can be written as

$$\frac{\partial\tilde{h}_\pi(\mathbf{v})}{\partial\mathbf{v}} = \frac{\partial h_\pi(\mathbf{U}\mathbf{v})}{\partial\mathbf{v}} = \mathbf{U}^\top\frac{\partial h_\pi(\mathbf{U}\mathbf{v})}{\partial\mathbf{U}\mathbf{v}} = \mathbf{U}^\top\nabla h_\pi(\mathbf{U}\mathbf{v}),$$

and thus it is evaluated for $\mathbf{v} = \tilde{\mathbf{B}}^\top\mathbf{X}$ as

$$\frac{\partial\tilde{h}_\pi(\mathbf{v})}{\partial\mathbf{v}}\Big|_{\mathbf{v}=\tilde{\mathbf{B}}^\top\mathbf{X}} = \mathbf{U}^\top\nabla h_\pi(\mathbf{B}^\top\mathbf{v}). \tag{25}$$

The first element of (25) is 0 by (24), which implies $\tilde{h}_\pi(\tilde{\mathbf{B}}^\top\mathbf{X})$ does not change with $\mathbf{u}_1^\top\mathbf{B}^\top\mathbf{X}$. Notice that

$$\tilde{h}_\pi(\tilde{\mathbf{B}}^\top\mathbf{X}) = h_\pi(\mathbf{B}^\top\mathbf{X}) = \tilde{f}_\pi(\mathbf{B}^\top\mathbf{X}) = f_\pi(\mathbf{X}),$$

and hence for any $\pi \in (0,1)$, $f_\pi(\mathbf{X})$ is a function of $(\mathbf{u}_2^\top\mathbf{B}^\top\mathbf{X}, \cdots, \mathbf{u}_d^\top\mathbf{B}^\top\mathbf{X})$, a $(d-1)$-dimensional variable. Since $\{f_\pi(\mathbf{X}) : \pi \in (0,1)\}$ collectively defines $P(Y=1|\mathbf{X})$ which is equivalent to the conditional distribution of $Y$ given $\mathbf{X}$, by the Fisher consistency of the loss function. Finally, $F(\cdot|\mathbf{X})$ is a function of $\{\mathbf{B}(\mathbf{u}_2, \cdots, \mathbf{u}_d)\}^\top\mathbf{X}$, which concludes that $\text{span}\{\mathbf{B}(\mathbf{u}_2, \cdots, \mathbf{u}_d)\}$ is a DRS but has lower dimension then $\mathcal{S}_{Y|\mathbf{X}}$. This leads a contradiction, and thus $\mathbf{W}$ must be of full-rank. ■

### A.2 Proof of Theorem 4

In what follows, we assume (C1)(C2′), and (C3)–(C5). As stated in the following Theorem 6, we first show that both of $E(|\hat{f}(\mathbf{X}) - f_\pi(\mathbf{X})|)$ and $E(\|\nabla\hat{f}(\mathbf{X}) - \nabla f_\pi(\mathbf{X})\|_1)$ are bounded by what we call the excess error defined as:

$$\begin{aligned} Exr(f,\nabla f) = E[w_s(\mathbf{X}-\mathbf{U})L_\pi\{Yf(\mathbf{U}) + \nabla f(\mathbf{U})(\mathbf{X}-\mathbf{U})\}] \\ - E\left[w_s(\mathbf{X}-\mathbf{U})L_\pi\left\{Yf_\pi(\mathbf{X})\right\}\right], \end{aligned}$$

where $\mathbf{U}$ is a random copy of $\mathbf{X}$. For a given $r > 0$, define

$$\mathcal{F}_r = \left\{(f,\nabla f) \in \mathcal{H}_K^{p+1} : J(f,\nabla f) \le r^2\right\},$$

where

$$J(f,\nabla f) = \frac{1}{2}\left(\theta_0\|f\|_{\mathcal{H}_K}^2 + \sum_{\ell=1}^p \theta_\ell\|g_\ell\|_{\mathcal{H}_K}^2\right).$$

**Theorem 6** *For $(f,\nabla f) \in \mathcal{F}_r$ with some $r > 1$, there exists constants $C_1, C_2 > 0$ such that*

$$E(|f(\mathbf{X}) - f_\pi(\mathbf{X})|) \le C_1\left(s^\tau r + s^{2-\tau} + s^{-\tau}Exr(f,\nabla f)\right)$$

*and*

$$E(\|\nabla f(\mathbf{X}) - \nabla f_\pi(\mathbf{X})\|_1) \le C_2\left(s^\tau r + s^{1-\tau} + s^{-\tau-1}Exr(f,\nabla f)\right).$$

In order to prove Theorem 6, we need Lemma 7 and 8. To state the lemmas, let us define the absolute error functional as

$$A(f, \nabla f) = E\left[w_s(\mathbf{X}-\mathbf{U})\left|f(\mathbf{X}) - f_\pi(\mathbf{X}) + \{\nabla f(\mathbf{X}) - \nabla f_\pi(\mathbf{X})\}^\top(\mathbf{U}-\mathbf{X})\right|\right].$$

In addition, we use the following notations for the sake of brevity.

$$N_q = \int_{\{\mathbf{t}\in\mathbb{R}^p, |\mathbf{t}|<1\}} e^{-\frac{|\mathbf{t}|^2}{2}}|\mathbf{t}|^q d\mathbf{t}, \quad \tilde{N}_q = \int_{\mathbf{t}\in\mathbb{R}^p} e^{-\frac{|\mathbf{t}|^2}{2}}|\mathbf{t}|^q d\mathbf{t},$$

and

$$\mathcal{X}_s = \{\mathbf{x}\in\mathbf{X} : d(\mathbf{x}, \partial\mathbf{X}) > s \text{ and } \rho(\mathbf{x}) \geq (1+c_\rho)s^\tau\}.$$

Now, we are ready to state the lemmas.

**Lemma 7** *For $(f, \nabla f) \in \mathcal{F}_r$ with some $r > 1$,*

$$N_0 s^\tau \int_{\mathcal{X}_s} |f(\mathbf{x}) - f_\pi(\mathbf{x})| d\mathrm{P}(\mathbf{x}) + \frac{N_1 s^{\tau+1}}{p}\int_{\mathcal{X}_s} \|\nabla f(\mathbf{x}) - \nabla f_\pi(\mathbf{x})\|_1 d\mathrm{P}(\mathbf{x}) \leq 2A(f, \nabla f).$$

**Proof** For $\mathbf{u}\in\mathcal{X}$ such that $|\mathbf{u}-\mathbf{x}| \leq s$ with $\mathbf{x}\in\mathcal{X}_s$, we have

$$\rho(\mathbf{u}) = \rho(\mathbf{x}) - \{\rho(\mathbf{x}) - \rho(\mathbf{u})\} \geq (1+c_\rho)s^\tau - c_\rho|\mathbf{u}-\mathbf{x}|^\tau \geq s^\tau.$$

The absolute error functional is

$$\begin{aligned}
A(f, \nabla f) =& \int\int w_s(\mathbf{x}-\mathbf{u})\left|f(\mathbf{x}) - f_\pi(\mathbf{x}) + \{\nabla f(\mathbf{x}) - \nabla f_\pi(\mathbf{x})\}^\top(\mathbf{u}-\mathbf{x})\right| d\mathrm{P}(\mathbf{u}) d\mathrm{P}(\mathbf{x}) \\
\geq& \int_{\mathcal{X}_s}\int_{|\mathbf{u}-\mathbf{x}|\leq s} w_s(\mathbf{x}-\mathbf{u})\left|f(\mathbf{x}) - f_\pi(\mathbf{x}) + \{\nabla f(\mathbf{x}) - \nabla f_\pi(\mathbf{x})\}^\top(\mathbf{u}-\mathbf{x})\right| \rho(\mathbf{u}) d\mathbf{u} d\mathrm{P}(\mathbf{x}) \\
\geq& s^\tau \int_{\mathcal{X}_s}\int_{|\mathbf{u}-\mathbf{x}|\leq s} w_s(\mathbf{x}-\mathbf{u})\left|f(\mathbf{x}) - f_\pi(\mathbf{x}) + \{\nabla f(\mathbf{x}) - \nabla f_\pi(\mathbf{x})\}^\top(\mathbf{u}-\mathbf{x})\right| d\mathbf{u} d\mathrm{P}(\mathbf{x}) \\
\geq& s^\tau \int_{\mathcal{X}_s}\int_{|\mathbf{u}-\mathbf{x}|\leq s} w_s(\mathbf{x}-\mathbf{u})\left|f(\mathbf{x}) - f_\pi(\mathbf{x})\right| d\mathbf{u} d\mathrm{P}(\mathbf{x}) \\
&+ s^\tau \int_{\mathcal{X}_s}\int_{|\mathbf{u}-\mathbf{x}|\leq s} w_s(\mathbf{x}-\mathbf{u})\left|\{\nabla f(\mathbf{x}) - \nabla f_\pi(\mathbf{x})\}^\top(\mathbf{u}-\mathbf{x})\right| d\mathbf{u} d\mathrm{P}(\mathbf{x}) \\
&- s^\tau \int_{\mathcal{X}_s}\int_{|\mathbf{u}-\mathbf{x}|\leq s} w_s(\mathbf{x}-\mathbf{u})\left|f(\mathbf{x}) - f_\pi(\mathbf{x}) - \{\nabla f(\mathbf{x}) - \nabla f_\pi(\mathbf{x})\}^\top(\mathbf{u}-\mathbf{x})\right| d\mathbf{u} d\mathrm{P}(\mathbf{x}) \\
:=& J_1 + J_2 - J_3.
\end{aligned}$$

The last inequality comes from the fact that $|a+b| \geq |a| + |b| - |a-b|, \forall a, b \in \mathbb{R}$. Note that the third term, $J_3$ is bounded by $A(f, \nabla f)$ as follows.

$$\begin{aligned}
J_3 =& s^\tau \int_{\mathcal{X}_s}\int_{|\mathbf{u}-\mathbf{x}|\leq s} w_s(\mathbf{x}-\mathbf{u})\left|f(\mathbf{x}) - f_\pi(\mathbf{x}) - \{\nabla f(\mathbf{x}) - \nabla f_\pi(\mathbf{x})\}^\top(\mathbf{u}-\mathbf{x})\right| d\mathbf{u} dP(\mathbf{x}) \\
=& s^\tau \int_{\mathcal{X}_s}\int_{|\mathbf{u}-\mathbf{x}|\leq s} w_s(\mathbf{x}-\mathbf{u})\left|f(\mathbf{x}) - f_\pi(\mathbf{x}) + \{\nabla f(\mathbf{x}) - \nabla f_\pi(\mathbf{x})\}^\top(\mathbf{u}-\mathbf{x})\right| d\mathbf{u} dP(\mathbf{x}) \\
\leq& A(f, \nabla f),
\end{aligned}$$

where the identity holds because both the integration range and $w_s(\mathbf{x}-\mathbf{u})$ are symmetric with respect to $|\mathbf{u}-\mathbf{x}|$. This leads us to have

$$2A(f,\nabla f) \geq J_1 + J_2.$$

For $J_1$, we have

$$J_1 = s^\tau \int_{\mathcal{X}_s} |f(\mathbf{x}) - f_\pi(\mathbf{x})| \int_{|\mathbf{t}|\leq 1} e^{-|\mathbf{t}|^2/2} d\mathbf{t} d\mathrm{P}(\mathbf{x}) = N_0 s^\tau \int_{\mathcal{X}_s} |f(\mathbf{x}) - f_\pi(\mathbf{x})| d\mathrm{P}(\mathbf{x}),$$

and $J_2$ is bounded as follows.

$$\begin{aligned} J_2 &\leq s^\tau \int_{\mathcal{X}_s} \sum_{\ell=1}^{p} \left| g_\ell(\mathbf{x}) - \frac{\partial f_\pi}{x_\ell}(\mathbf{x}) \right| \int_{|t_\ell|\leq 1} e^{-|t_\ell|^2/2} |t_\ell| dt_\ell d\mathrm{P}(\mathbf{x}) \\ &= \frac{N_1 s^{\tau+1}}{p} \int_{\mathcal{X}_s} \|\nabla f(\mathbf{x}) - \nabla f_\pi(\mathbf{x})\|_1 d\mathrm{P}(\mathbf{x}). \end{aligned}$$

which completes the proof. ■

**Lemma 8** *For $(f,\nabla f) \in \mathcal{F}_r$ with some $r > 1$, there exists a constant $C_3 > 0$ satisfying that*

$$A(f,\nabla f) \leq C_3 \left( s^2 + Exr(f,\nabla f) \right).$$

**Proof**

Given $\mathbf{x}$, we have for the hinge loss

$$E[L_\pi(Yf(\mathbf{x})\} \mid \mathbf{x}] = (1-\pi)p(\mathbf{x})[1-f(\mathbf{x})]_+ + \pi\{1-p(\mathbf{x})\}[1+f(\mathbf{x})]_+,$$

which yields

$$0 \leq E[L_\pi\{Yf(\mathbf{x})\} \mid \mathbf{x}] - E[L_\pi\{Yf_\pi(\mathbf{x})\} \mid \mathbf{x}] \leq |f(\mathbf{x}) - f_\pi(\mathbf{x})|.$$

For a given $\mathbf{x} \in \mathcal{S}_\gamma^c$ where $\mathcal{S}_\gamma^c$ is defined in (C5), we have a non-zero lower bound as

$$\gamma |f(\mathbf{x}) - f_\pi(\mathbf{x})| \leq E[L_\pi\{Yf(\mathbf{x})\} \mid \mathbf{x}] - E[L_\pi\{Yf_\pi(\mathbf{x})\} \mid \mathbf{x}], \qquad \forall f \in \mathcal{H}_K \tag{26}$$

where $0 < \gamma \leq \inf_{\mathbf{x} \in S_\gamma^c}\{(1-\pi)p(\mathbf{x}), \pi(1-p(\mathbf{x})), |\pi - p(\mathbf{x})|\} < 1$. Now, we have

$$\begin{aligned}
Exr(f, \nabla f) &= E\left[w_s(\mathbf{X}-\mathbf{U})\left(L_\pi\left\{Y(f(\mathbf{U}) + \nabla f(\mathbf{U})^\top(\mathbf{X}-\mathbf{U}))\right\} - L_\pi\left\{Y f_\pi(\mathbf{X})\right\}\right)\right] \\
&= E_{X,U}\left[w_s(\mathbf{X}-\mathbf{U}) E_{Y|X,U}\left[L_\pi\left\{Y(f(\mathbf{U}) + \nabla f(\mathbf{U})^\top(\mathbf{X}-\mathbf{U}))\right\} - L_\pi\left\{Y f_\pi(\mathbf{X})\right\}\right]\right] \\
&\geq \gamma \int\int_{S_\gamma^c} w_s(\mathbf{x}-\mathbf{u})\left|f(\mathbf{u}) + \nabla f(\mathbf{u})^\top(\mathbf{x}-\mathbf{u}) - f_\pi(\mathbf{x})\right| d\mathrm{P}(\mathbf{x}) d\mathrm{P}(\mathbf{u}) \\
&= \gamma/2 \int\int_{S_\gamma^c} w_s(\mathbf{x}-\mathbf{u})\left|f(\mathbf{u}) + \nabla f(\mathbf{u})^\top(\mathbf{x}-\mathbf{u}) - f_\pi(\mathbf{x})\right| d\mathrm{P}(\mathbf{x}) d\mathrm{P}(\mathbf{u}) \\
&\quad + \gamma/2 \int\int_{S_\gamma^c} w_s(\mathbf{x}-\mathbf{u})\left|f(\mathbf{u}) + \nabla f(\mathbf{u})^\top(\mathbf{x}-\mathbf{u}) - f_\pi(\mathbf{x})\right| d\mathrm{P}(\mathbf{x}) d\mathrm{P}(\mathbf{u}) \\
&\geq \gamma/2 \int\int_{S_\gamma^c} w_s(\mathbf{x}-\mathbf{u})\left|f(\mathbf{u}) + \nabla f(\mathbf{u})^\top(\mathbf{x}-\mathbf{u}) - f_\pi(\mathbf{x})\right| d\mathrm{P}(\mathbf{x}) d\mathrm{P}(\mathbf{u}) \\
&\quad + k\gamma/2 \int\int_{S_\gamma} w_s(\mathbf{x}-\mathbf{u})\left|f(\mathbf{u}) + \nabla f(\mathbf{u})^\top(\mathbf{x}-\mathbf{u}) - f_\pi(\mathbf{x})\right| d\mathrm{P}(\mathbf{x}) d\mathrm{P}(\mathbf{u}) \\
&\geq \gamma' E\left[w(\mathbf{X}-\mathbf{U})\left|f(\mathbf{U}) + \nabla f(\mathbf{U})^\top(\mathbf{X}-\mathbf{U}) - f_\pi(\mathbf{X})\right|\right],
\end{aligned}$$

where $\gamma' = \min\{\gamma/2, k\gamma/2\}$. The first inequality holds by (21) and (26), and the second inequality holds by (22).

Let

$$\begin{aligned}
T_1 = T_1(\mathbf{X}, \mathbf{U}) &= f(\mathbf{U}) - f_\pi(\mathbf{U}) + \{\nabla f(\mathbf{U}) - \nabla f_\pi(\mathbf{U})\}^\top(\mathbf{X}-\mathbf{U}), \\
T_2 = T_2(\mathbf{X}, \mathbf{U}) &= f_\pi(\mathbf{U}) - f_\pi(\mathbf{X}) + \nabla f_\pi(\mathbf{U})^\top(\mathbf{X}-\mathbf{U}).
\end{aligned}$$

Then we have

$$A(f, \nabla f) = E\{w_s(\mathbf{X}-\mathbf{U}) \cdot |T_1|\},$$

and

$$\left|f(\mathbf{U}) + \nabla f(\mathbf{U})^\top(\mathbf{X}-\mathbf{U}) - f_\pi(\mathbf{X})\right| = |T_1 + T_2| \geq |T_1| - |T_2|.$$

This leads

$$A(f, \nabla f) \leq E\{w_s(\mathbf{X}-\mathbf{U}) \cdot |T_2|\} + \frac{1}{\gamma'} Err(f, \nabla f).$$

By (C4), we have

$$|T_2| \leq c_f(\mathbf{X}-\mathbf{U})^2.$$

Together with the assumption $\rho(\mathbf{x}) \leq c_\rho$, we obtain

$$\begin{aligned}
E\{w_s(\mathbf{X}-\mathbf{U}) \cdot |T_2|\} &\leq c_f E\{w_s(\mathbf{X}-\mathbf{U})(\mathbf{X}-\mathbf{U})^2\} \\
&\leq c_f c_F \int\int w_s(\mathbf{x}-\mathbf{u})(\mathbf{x}-\mathbf{u})^2 d\mathbf{u} d\mathrm{P}(\mathbf{x}) \\
&\leq c_f c_F \tilde{N}_2 s^2.
\end{aligned}$$

Taking $C_3 = \max\left\{c_f c_F \tilde{N}_2, \frac{1}{\gamma'}\right\}$ completes the proof. ■

**Proof of Theorem 6.** Write

$$E(|f(\mathbf{X}) - f_\pi(\mathbf{X})|) = \int_{\mathcal{X}_s^c} |f(\mathbf{x}) - f_\pi(\mathbf{x})| d\mathrm{P}(\mathbf{x}) + \int_{\mathcal{X}_s} |f(\mathbf{x}) - f_\pi(\mathbf{x})| d\mathrm{P}(\mathbf{x}) \tag{27}$$
$$:= A_1 + A_2.$$

Since $\mathcal{X}_s^c = \{\mathbf{x} \in \mathbf{X} : d(\mathbf{x}, \partial\mathbf{X}) < s \text{ or } \rho(\mathbf{x}) \leq (1 + c_\rho)s^\tau\}$, by (C1) we have

$$P(\mathbf{X} \in \mathcal{X}_s^c) \leq c_\rho s + (1 + c_\rho)\mu(\mathbf{X})s^\tau \leq \{c_\rho + (1 + c_\rho)\mu(\mathbf{X})\}s^\tau,$$

where $\mu(\mathbf{X})$ denotes the Lebesgue measure of $\mathbf{X}$, and this leads

$$A_1 \leq \kappa(r + \|f_\pi\|_{\mathcal{H}_K})\{c_\rho + (1 + c_\rho)\mu(\mathbf{X})\}s^\tau,$$

where $\kappa = \sup_{\mathbf{x}\in\mathcal{X}} \sqrt{K(\mathbf{x}, \mathbf{x})}$. Here we use the fact $\|f\|_\infty \leq \kappa\|f\|_{\mathcal{H}_K}$ for $f \in \mathcal{H}_K$. By Lemma 7 and Lemma 8, $A_2$ is bounded by

$$A_2 \leq \frac{C_1 s^{-\tau}}{N_0}\{s^2 + Exr(f, \nabla f)\}.$$

which completes the proof of the first part of Theorem 6 with

$$C_2 = \kappa\{1 + \|f_\pi\|_{\mathcal{H}_K})\{c_\rho + (1 + c_\rho)\mu(\mathbf{X})\} + \frac{C_1}{N_0}.$$

The second part of Theorem 6 can be shown in an identical manner to the first part, with

$$C_3 = \kappa\{1 + \|\nabla f_\pi\|_{\mathcal{H}_K})\{c_\rho + (1 + c_\rho)\mu(\mathbf{X})\} + \frac{pC_1}{N_1},$$

which completes the proof. ■

Now we turn to find an upper bound of the excess error $Exr(\hat{f}, \nabla\hat{f})$. To do this, we introduce an intermediate estimator

$$(f_\lambda, \nabla f_\lambda) = \underset{(f,\nabla f)\in\mathcal{H}_K^{p+1}}{\arg\min} \{\mathcal{E}(f, \nabla f) + \lambda J(f, \nabla f)\},$$

where

$$\mathcal{E}(f, \nabla f) = E\left[w(\mathbf{X} - \mathbf{U})L_\pi\left\{y(f(\mathbf{U}) + \nabla f(\mathbf{U})^\top(\mathbf{X} - \mathbf{U}))\right\}\right]. \tag{28}$$

**Theorem 9** *If $(\hat{f}, \nabla\hat{f})$ and $(f_\lambda, \nabla f_\lambda)$ are in $\mathcal{F}_r$ for some $r \geq 1$, then with confidence $1 - \eta$,*

$$Exr(\hat{f}, \nabla\hat{f}) \leq C_4\left(\frac{r(1 + \log\frac{2}{\eta})}{\sqrt{n}s^p} + s^2 + \lambda\right),$$

*where $C_4 > 0$ is a constant depending on $c_f$, $c_\rho$, but not on $r, s$ and $\lambda$.*

**Proof** First of all, we decompose $\mathcal{E}(\hat{f}, \nabla\hat{f})$ as

$$\begin{aligned}
\mathcal{E}(\hat{f}, \nabla\hat{f}) &\leq \mathcal{E}(\hat{f}, \nabla\hat{f}) - \hat{\mathcal{E}}(\hat{f}, \nabla\hat{f}) + \hat{\mathcal{E}}(\hat{f}, \nabla\hat{f}) - \mathcal{E}(f_\lambda, \nabla f_\lambda) + \mathcal{E}(f_\lambda, \nabla f_\lambda) + \lambda J(\hat{f}, \nabla\hat{f}) \\
&\leq [\mathcal{E}(\hat{f}, \nabla\hat{f}) - \hat{\mathcal{E}}(\hat{f}, \nabla\hat{f})] + [\hat{\mathcal{E}}(f_\lambda, \nabla f_\lambda) - \mathcal{E}(f_\lambda, \nabla f_\lambda)] + [\mathcal{E}(f_\lambda, \nabla f_\lambda) + \lambda J(f_\lambda, f_\lambda)] \\
&\leq [\mathcal{E}(\hat{f}, \nabla\hat{f}) - \hat{\mathcal{E}}(\hat{f}, \nabla\hat{f})] + [\hat{\mathcal{E}}(f_\lambda, \nabla f_\lambda) - \mathcal{E}(f_\lambda, \nabla f_\lambda)] + [\mathcal{E}(f_\pi, \nabla f_\pi) + \lambda J(f_\pi, \nabla f_\pi)] \\
&= \underbrace{[\mathcal{E}(\hat{f}, \nabla\hat{f}) - \hat{\mathcal{E}}(\hat{f}, \nabla\hat{f})]}_{E_1} + \underbrace{[\hat{\mathcal{E}}(f_\lambda, \nabla f_\lambda) - \mathcal{E}(f_\lambda, \nabla f_\lambda)]}_{E_2} \\
&\quad + \underbrace{[Exr(f_\pi, \nabla f_\pi) + J(f_\pi, \nabla f_\pi)]}_{E_3} + E\left[w_s(\mathbf{X} - \mathbf{U}) L_\pi \left\{Y f_\pi(\mathbf{X})\right\}\right].
\end{aligned}$$

The first inequality holds because $\lambda J(\hat{f}, \nabla\hat{f}) > 0$, the second inequality holds by the definition of $(\hat{f}, \nabla\hat{f})$, and the third inequality holds by the definition of $(f_\lambda, \nabla f_\lambda)$. If both $(\hat{f}, \nabla\hat{f})$ and $(f_\lambda, \nabla f_\lambda)$ are in $\mathcal{F}_r$ for some $r > 0$, then

$$E_i \leq S(\mathbf{z}, r) := \sup_{(f, \nabla f) \in \mathcal{F}_r} \left| \hat{\mathcal{E}}(f, \nabla f) - \mathcal{E}(f, \nabla f) \right|.$$

for $i = 1, 2$. According to Proposition 27 in Mukherjee and Wu (2006), we have with probability at least $1 - \eta$,

$$E_1 + E_2 \leq C_5 \frac{r(1 + \log\frac{2}{\eta})}{\sqrt{n} s^p}$$

for some $C_5 > 0$. From the proof in Lemma 8,

$$Exr(f_\pi, \nabla f_\pi) \leq c_f c_\rho \tilde{N}_2 s^2,$$

and we have

$$E_3 \leq C_6(s^2 + \lambda),$$

with

$$C_6 = \max\left\{ c_f c_\rho \tilde{N}_2, J(f_\pi, \nabla f_\pi) \right\}.$$

Taking $C_4 = \max\{C_5, C_6\}$, we attain the desired result. ■

To incorporate Theorem 6 and 9, we need to find $r$ such that both $(\hat{f}, \nabla\hat{f})$ and $(f_\lambda, \nabla f_\lambda)$ are in $\mathcal{F}_r$. A natural bound for $r^2$ is $2\lambda^{-1}s^{-p}$ because

$$\lambda J(\hat{f}, \nabla\hat{f}) \leq \hat{\mathcal{E}}(\hat{f}, \nabla\hat{f}) + J(\hat{f}, \nabla\hat{f}) \leq \hat{\mathcal{E}}(0, 0) + 0 = \frac{1}{s^p}$$

and similarly $\lambda J(f_\lambda, \nabla f_\lambda) \leq s^{-p}$. However this bound tends to $\infty$ as $s \to 0$ and $\lambda \to 0$, which is limited to applying to Theorem 6 and 9. The following lemma gives a sharper bound.

**Lemma 10** *Both $(\hat{f}, \nabla\hat{f})$ and $(f_\lambda, \nabla f_\lambda)$ are in $\mathcal{F}_r$ with confidence at least $1 - \eta$ if*

$$r^2 = 2C_4 \left\{ 1 + \frac{s^2}{\lambda} + \left(1 + \log\frac{2}{\eta}\right) \frac{\lambda^{-\frac{3}{2}} s^{-\frac{3p}{2}}}{\sqrt{n}} \right\}.$$

**Proof** From the process in the proof of Theorem 9, we have

$$\lambda J(f_\lambda, \nabla f_\lambda) \leq E_3, \tag{29}$$

and

$$\lambda J(\hat{f}, \nabla \hat{f}) \leq E_1 + E_2 + E_3.$$

Since both $(\hat{f}, \nabla \hat{f})$ and $(f_\lambda, \nabla f_\lambda)$ are in $\mathcal{F}_{\sqrt{2\lambda^{-1}s^{-p}}}$, by Theorem 9 with probability at least $1-\eta$,

$$E_1 + E_2 \leq C_4 \sqrt{\frac{1}{\lambda s^p}} \frac{(1+\log\frac{2}{\eta})}{\sqrt{n}s^p}. \tag{30}$$

Combining (29), (30), we obtain the desired estimate. ■

We now prove Theorem 4.

**Proof** [Proof of Theorem 3] By Theorems 6 and 9 we have with probability at least $1-\eta$, both $E\left[|\hat{f}(\mathbf{X}) - f_\pi(\mathbf{X})|\right]$ and $E\left[\|\nabla \hat{f}(\mathbf{X}) - \nabla f_\pi(\mathbf{X})\|_1\right]$ are bounded by

$$\max\{C_1, C_2\} \left\{ s^\tau r + s^{1-\tau} + C_4 s^{-\tau-1} \left( \frac{r(1+\log\frac{2}{\eta})}{\sqrt{n}s^p} + s^2 + \lambda \right) \right\},$$

if both $(\hat{f}, \nabla \hat{f})$ and $(f_\lambda, \nabla f_\lambda)$ are in $\mathcal{F}_r$ for $r > 1$. By Lemma 10 we have confidence at least $1-\eta$, $(\hat{f}, \nabla \hat{f})$ and $(f_\lambda, \nabla f_\lambda)$ are in $\mathcal{F}_r$ if

$$r^2 = 2C_4 \left\{ 1 + \frac{s^2}{\lambda} + \left(1 + \log\frac{2}{\eta}\right) \frac{\lambda^{-\frac{3}{2}} s^{-\frac{3p}{2}}}{\sqrt{n}} \right\}.$$

Choose $s = n^{-\frac{1}{3(p+2+2\tau)}}$, $\lambda = s^{1+2\tau}$. Then $r > 1$ and $r \leq r_0$ with $r_0 > 1$ an absolute constant. By substituting $r_0$ into this bound, we obtain with confidence at least $1-2\eta$,

$$\max\left\{ E\left[|\hat{f}(\mathbf{X}) - f_\pi(\mathbf{X})|\right], E\left[\|\nabla \hat{f}(\mathbf{X}) - \nabla f_\pi(\mathbf{X})\|_1\right] \right\} \leq C_{SVM} \left(\frac{1}{n}\right)^{\frac{\tau}{3(p+2+2\tau)}},$$

with $C_{SVM} = \max\{C_1, C_2\} r_0(1+C_4)$. ■

### A.3 Proof of Corollary 5

Let $\widehat{\mathbf{M}}(\pi)_{ij}$ be the $(i,j)$th element of $\widehat{\mathbf{M}}(\pi)$. Theorem 3, 4 and the law of large numbers implies that $\widehat{\mathbf{M}}(\pi)_{ij}$ converges to $\mathbf{M}(\pi)_{ij}$ in probability for each $\pi \in (\delta, 1-\delta)$. For any given $\epsilon > 0$, let $\pi^* = \underset{\pi\in(\delta,1-\delta)}{\arg\sup}\, P\left(|\widehat{\mathbf{M}}(\pi)_{ij} - \mathbf{M}(\pi)_{ij}| > \epsilon\right)$, then

$$\begin{aligned}
P\left(\left|\int_\delta^{1-\delta} \widehat{\mathbf{M}}(\pi)_{ij} d\pi - \int_\delta^{1-\delta} \mathbf{M}(\pi)_{ij} d\pi\right| > \epsilon^*\right) &\leq P\left(\int_\delta^{1-\delta} \left|\widehat{\mathbf{M}}(\pi)_{ij} - \mathbf{M}(\pi)_{ij}\right| d\pi > \epsilon^*\right) \\
&\leq P\left((1-2\delta)\left|\widehat{\mathbf{M}}(\pi^*)_{ij} - \mathbf{M}(\pi^*)_{ij}\right| > \epsilon^*\right) \\
&= P\left(\left|\widehat{\mathbf{M}}(\pi^*)_{ij} - \mathbf{M}(\pi^*)_{ij}\right| > \epsilon\right)
\end{aligned}$$

for $\epsilon^* = (1-2\delta)\epsilon$. The last term converges to 0 as $n$ increases to $\infty$, which completes the proof. ■

## Appendix B. Computing Algorithms

In this section, we suppress $\pi$ for the sake of simplicity. Let $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_0^\top, \cdots, \boldsymbol{\alpha}_n^\top)^\top$, and $\omega_{ij} = w_s(\mathbf{x}_i - \mathbf{x}_j)$.

### B.1 wOPG-LR

For the wOPG-LR, we have the following objective function for (18).

$$G(\boldsymbol{\alpha}) = \sum_{i=1}^{n}\sum_{j=1}^{n} \omega_{ij} w(y_i) \log\left[1 + \exp\left(-y_i \boldsymbol{\alpha}^\top \tilde{\mathbf{k}}_{ij}\right)\right] + \frac{1}{2}\sum_{\ell=0}^{p} \lambda_\ell \boldsymbol{\alpha}_\ell^\top \mathbf{K} \boldsymbol{\alpha}_\ell, \tag{31}$$

where $\tilde{\mathbf{k}}_{ij} = (\delta_{ij0}\mathbf{k}_j^\top, \cdots, \delta_{ijp}\mathbf{k}_j^\top)^\top$.

The conventional Newton-Raphson method gives the following updating equation of $\boldsymbol{\alpha}$:

$$\boldsymbol{\alpha}^{\text{new}} = \boldsymbol{\alpha}^{\text{old}} - \mathbf{H}^{-1}(\boldsymbol{\alpha}^{\text{old}}) \nabla G(\boldsymbol{\alpha}^{\text{old}}),$$

where $\nabla G(\boldsymbol{\alpha})$ and $\mathbf{H}(\boldsymbol{\alpha})$ denote the gradient vector and Hessian matrix of $G(\boldsymbol{\alpha})$, respectively. However, the dimension of Hessian matrix $\mathbf{H}$ is $np \times np$, prohibitively large. To circumvent this, we propose update $\boldsymbol{\alpha}$ in a block-wise manner, i.e., update $\boldsymbol{\alpha}_\ell$ only at a time, instead of $\boldsymbol{\alpha}$. For each $\boldsymbol{\alpha}_\ell, \ell = 0, 1, \cdots, p$, we have the following updating equation:

$$\boldsymbol{\alpha}_\ell^{\text{new}} = \boldsymbol{\alpha}_\ell^{\text{old}} - \mathbf{H}_\ell^{-1}(\boldsymbol{\alpha}^{\text{old}}) \nabla G_\ell(\boldsymbol{\alpha}^{\text{old}}). \tag{32}$$

Here the gradient $\nabla G_\ell(\boldsymbol{\alpha})$ is given by

$$\nabla G_\ell(\boldsymbol{\alpha}) = \frac{\partial G(\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}_\ell} = \mathbf{K}(\mathbf{u}_\ell + \lambda_\ell \boldsymbol{\alpha}_\ell),$$

where $\mathbf{u}_\ell = (u_{\ell 1}, \cdots, u_{\ell n})^\top$, $u_{\ell j} = -\sum_{i=1}^{n} \omega_{ij} w(y_i) y_i \mu_{ij} \delta_{ij\ell}, j = 1, \cdots, n$, and

$$\mu_{ij} = \frac{1}{1 + \exp(y_i \boldsymbol{\alpha}^\top \tilde{\mathbf{k}}_{ij})}.$$

The Hessian $H_\ell(\boldsymbol{\alpha})$ is given by

$$\mathbf{H}_\ell(\boldsymbol{\alpha}) = \frac{\partial^2 G(\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}_\ell \partial \boldsymbol{\alpha}_\ell^\top} = \mathbf{K}\mathbf{V}_\ell\mathbf{K} + \lambda_\ell \mathbf{K},$$

where $\mathbf{V}_\ell = \text{diag}\{v_{\ell 1}, \cdots, v_{\ell n}\}$ denotes the $n$-dimensional diagonal matrix with

$$v_{\ell j} = \sum_{i=1}^{n} \omega_{ij} w(y_i) \mu_{ij} (1 - \mu_{ij}) \delta_{ij\ell}^2.$$

Finally, we can estimate $\alpha$ by iterating the following updating equation

$$\boldsymbol{\alpha}_\ell^{\text{new}} = (\mathbf{K}\mathbf{V}_\ell\mathbf{K} + \lambda_\ell \mathbf{K})^{-1} \mathbf{K}\mathbf{V}_\ell (\mathbf{K}\boldsymbol{\alpha}_\ell^{\text{old}} - \mathbf{V}_\ell^{-1}\mathbf{u}_\ell)$$

until convergence. We like to remark that both $\mathbf{u}_\ell$ and $\mathbf{V}_\ell$ evaluated at $\boldsymbol{\alpha}^{\text{old}}$. ■

### B.2 wOPG-SVM

For the wOPG-SVM, (18) with the hinge loss can be equivalently written as

$$\begin{aligned} \min_{\boldsymbol{\alpha},\boldsymbol{\xi}} \quad & \sum_{i=1}^{n}\sum_{j=1}^{n}\omega_{ij}w(y_i)\xi_{ij} + \frac{1}{2}\sum_{\ell=0}^{p}\lambda_\ell\boldsymbol{\alpha}_\ell^\top\mathbf{K}\boldsymbol{\alpha}_\ell \\ \text{subject to} \quad & y_i\sum_{\ell=0}^{p}\boldsymbol{\alpha}_\ell^\top\mathbf{k}_j\delta_{ij\ell} \geq 1-\xi_{ij}, \text{ and } \xi_{ij}\geq 0, \ \forall i,j=1,\cdots,n, \end{aligned} \tag{33}$$

where $\boldsymbol{\Xi} = \{\xi_{ij}\} \in \mathbb{R}^{n\times n}$. Introducing nonnegative Lagrangian multipliers $\mathbf{U} = \{u_{ij}\} \in \mathbb{R}^{n\times n}$ and $\mathbf{V} = \{v_{ij}\} \in \mathbb{R}^{n\times n}$, the Lagrangian $G_1$ associated with (33) is

$$\begin{aligned} G_1(\boldsymbol{\alpha},\boldsymbol{\Xi},\mathbf{U},\mathbf{V}) = & \sum_{i=1}^{n}\sum_{j=1}^{n}\omega_{ij}w(y_i)\xi_{ij} + \frac{1}{2}\sum_{\ell=0}^{p}\lambda_\ell\boldsymbol{\alpha}_\ell^\top\mathbf{K}\boldsymbol{\alpha}_\ell \\ & - \sum_{i=1}^{n}\sum_{j=1}^{n}u_{ij}\left(y_i\sum_{\ell=0}^{p}\boldsymbol{\alpha}_\ell^\top\mathbf{k}_j\delta_{ij\ell} - 1 + \xi_{ij}\right) - \sum_{i=1}^{n}\sum_{j=1}^{n}v_{ij}\xi_{ij}. \end{aligned} \tag{34}$$

Taking the derivative of (34) with respect to $\xi_{ij}$ and setting it to zero, we have the following stationary constraint

$$\omega_{ij}w(y_i) = u_{ij} - v_{ij}, \quad \forall i,j=1,\cdots,n. \tag{35}$$

Plugging (35) into (34), we have the following reduced form of (34):

$$G_2(\mathbf{A},\mathbf{U}) = \frac{1}{2}\sum_{\ell=0}^{p}\lambda_\ell\boldsymbol{\alpha}_\ell^\top\mathbf{K}\boldsymbol{\alpha}_\ell - \sum_{i=1}^{n}\sum_{j=1}^{n}u_{ij}\left(y_i\sum_{\ell=0}^{p}\boldsymbol{\alpha}_\ell^\top\mathbf{k}_j\delta_{ij\ell} - 1\right). \tag{36}$$

Let $d_{j\ell} = \sum_{i=1}^{n} y_i\delta_{ij\ell}u_{ij}$ and $\mathbf{d}_\ell = (d_{1\ell}, d_{2\ell}, \cdots, d_{n\ell})^\top$. Then for each $\ell = 0, 1, ..., p$, we have

$$\sum_{i=1}^{n}\sum_{j=1}^{n}u_{ij}y_i\delta_{ij\ell}\mathbf{k}_j = \mathbf{K}\mathbf{d}_\ell.$$

Taking derivative of (36) with respect to $\boldsymbol{\alpha}_\ell$ yields

$$\lambda_\ell\mathbf{K}\boldsymbol{\alpha}_\ell - \mathbf{K}\mathbf{d}_\ell = \mathbf{0} \quad \Leftrightarrow \quad \boldsymbol{\alpha}_\ell = \frac{1}{\lambda_\ell}\mathbf{d}_\ell. \tag{37}$$

Substituting (37) into (36), we have

$$G(\mathbf{U}) = \sum_{i=1}^{n}\sum_{j=1}^{n}u_{ij} - \frac{1}{2}\sum_{\ell=0}^{p}\frac{1}{\lambda_\ell}\mathbf{d}_\ell^\top\mathbf{K}\mathbf{d}_\ell. \tag{38}$$

Let $\tilde{\mathbf{X}}_{(i)}$ be $n \times (p+1)$ matrix whose $j$th row is $(\frac{1}{\sqrt{\lambda_0}}, \frac{1}{\sqrt{\lambda}}(\mathbf{x}_i - \mathbf{x}_j)^\top)$ and $\mathbf{u}_{(i)} = (u_{i1}, \cdots, u_{in})^\top$.

Let $k_{ij} = K(\mathbf{x}_i, \mathbf{x}_j)$ be the $(i,j)$th element of the kernel matrix $\mathbf{K}$.

$$\begin{aligned}
\sum_{\ell=0}^{p} \frac{1}{\lambda_\ell} \mathbf{d}_\ell^\top \mathbf{K} \mathbf{d}_\ell &= \sum_{\ell=0}^{p} \sum_{j=1}^{n} \sum_{k=1}^{n} \frac{1}{\lambda_\ell} d_{j\ell} k_{jk} d_{k\ell} \\
&= \sum_{\ell=0}^{p} \sum_{j=1}^{n} \sum_{k=1}^{n} \sum_{i=1}^{n} \sum_{h=1}^{n} \frac{1}{\lambda_\ell} y_i y_h \delta_{ij\ell} \delta_{hk\ell} u_{ij} k_{jk} u_{hk} \\
&= \sum_{i=1}^{n} \sum_{h=1}^{n} y_i y_h \sum_{j=1}^{n} \sum_{k=1}^{n} u_{ij} k_{jk} u_{hk} \left( \frac{1}{\lambda_0} + \sum_{\ell=1}^{p} \frac{1}{\lambda_\ell} (x_{i\ell} - x_{j\ell})(x_{h\ell} - x_{k\ell}) \right) \\
&= \sum_{i=1}^{n} \sum_{h=1}^{n} y_i y_h \mathbf{u}_{(i)}^\top \left( \tilde{\mathbf{X}}_{(i)} \tilde{\mathbf{X}}_{(h)}^\top \odot \mathbf{K} \right) \mathbf{u}_{(h)} \\
&= \tilde{\mathbf{u}}^\top \boldsymbol{\Psi} \tilde{\mathbf{u}},
\end{aligned}$$

where $\boldsymbol{\Psi}$ is $n^2 \times n^2$ matrix whose $(i,j)$ block matrix is $y_i y_j \left( \tilde{\mathbf{X}}_{(i)} \tilde{\mathbf{X}}_{(j)}^\top \odot \mathbf{K} \right)$ and $\tilde{\mathbf{u}} = (\mathbf{u}_{(1)}^\top, \cdots, \mathbf{u}_{(n)}^\top)^\top \in \mathbb{R}^{n^2}$. Then we have the following dual program.

$$\max_{\tilde{\mathbf{u}} \in \mathbb{R}^{n^2}} h(\tilde{\mathbf{u}}) = \mathbf{1}^\top \tilde{\mathbf{u}} - \frac{1}{2} \tilde{\mathbf{u}}^\top \boldsymbol{\Psi} \tilde{\mathbf{u}}, \quad \text{subject to} \quad 0 \le u_{ij} \le \omega_{ij} w(y_i),\ i,j = 1, \cdots, n^2. \tag{39}$$

One can solve (39) using the box-constraint coordinate ascent method (Wright, 2015). Note that $\frac{\partial h}{\partial \tilde{u}_i} = 1 - \boldsymbol{\psi}_i \tilde{\mathbf{u}}$ and $\frac{\partial^2 h}{\partial \tilde{u}_i^2} = -\psi_{ii}$, where $\boldsymbol{\psi}_i$ is the $i$th row vector of $\boldsymbol{\Psi}$. Now, one can iteratively update $u_i$ as

$$\tilde{u}_i^{\text{new}} \leftarrow \min \left( \max \left( \tilde{u}_i^{\text{old}} + \frac{1 - \boldsymbol{\psi}_i \tilde{\mathbf{u}}}{\psi_{ii}}, 0 \right), \tilde{\omega}_i \right), \qquad i = 1, \cdots, n^2, \tag{40}$$

until convergence. Here $\tilde{\omega}_i, i = 1, \cdots, n^2$ denotes the upper bound of $\tilde{u}_i$ in the box constraint of (39), i.e., the $i$th element of $\tilde{\boldsymbol{\omega}} = (w(y_1)\boldsymbol{\omega}_1^\top, \cdots, w(y_n)\boldsymbol{\omega}_n^\top)^\top$ where $\boldsymbol{\omega}_j = (\omega_{j1}, \cdots, \omega_{jn})^\top, j = 1, \cdots, n$.

Finally, the solution of (18) with the hinge loss is obtained by

$$\hat{\alpha}_{j\ell} = \frac{1}{\lambda_\ell} \sum_{i=1}^{n} y_i \delta_{ij\ell} \hat{u}_{ij}, \qquad \text{for } j = 1, \cdots, n \text{ and } \ell = 0, 1, \cdots, p,$$

where $\hat{u}_{ij} \in \mathbb{R}^n \times \mathbb{R}^n$ denotes the re-arranged version of $\tilde{\mathbf{u}} \in \mathbb{R}^{n^2}$ obtained from (40).

■

## References

Peter L Bartlett, Michael I Jordan, and Jon D McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156, 2006.

Victor Chernozhukov, Iván Fernández-Val, and Tetsuya Kaji. Extremal quantile regression. *Handbook of Quantile Regression*, pages 333–362, 2017.

R Dennis Cook and Bing Li. Dimension reduction for conditional mean in regression. *The Annals of Statistics*, 30(2):455–474, 2002.

R. Dennis Cook and S. Weisberg. Discussion of "Sliced inverse regression for dimension reduction". *Journal of the American Statistical Association*, 86(414):28–33, 1991.

Jianqing Fan. Local linear regression smoothers and their minimax efficiencies. *The Annals of Statistics*, 21(1):196–216, 1993.

Wolfgang Härdle and Thomas M Stoker. Investigating smooth multiple regression by the method of average derivatives. *Journal of the American statistical Association*, 84(408): 986–995, 1989.

Mohamed Hosni, Ibtissam Abnane, Ali Idri, Juan M Carrillo de Gea, and José Luis Fernández Alemán. Reviewing ensemble classification methods in breast cancer. *Computer Methods and Programs in Biomedicine*, 177:89–112, 2019.

Tommi S Jaakkola, Mark Diekhans, and David Haussler. Using the Fisher kernel method to detect remote protein homologies. In Thomas Lengauer, Reinhard Schneider, Peer Bork, Douglas L. Brutlag, Janice I. Glasgow, Hans-Werner Mewes, and Ralf Zimmer, editors, *Proceedings of the Seventh International Conference on Intelligent Systems for Molecular Biology*, pages 149–158. AAAI Press, 1999.

George Kimeldorf and Grace Wahba. Some results on tchebycheffian spline functions. *Journal of Mathematical Analysis and Applications*, 33(1):82–95, 1971.

Efang Kong and Yingcun Xia. An adaptive composite quantile approach to dimension reduction. *The Annals of Statistics*, 42(4):1657–1688, 2014.

Bing Li. *Sufficient Dimension Reduction: Methods and Applications with R*. CRC Press, 2018.

Bing Li and Shaoli Wang. On directional regression for dimension reduction. *Journal of the American Statistical Association*, 102(479):997–1008, 2007.

Bing Li, Hongyuan Zha, Francesca Chiaromonte, et al. Contour regression: a general approach to dimension reduction. *The Annals of Statistics*, 33(4):1580–1616, 2005.

Bing Li, Andreas Artemiou, and Lexin Li. Principal support vector machines for linear and nonlinear sufficient dimension reduction. *The Annals of Statistics*, 39(6):3182–3210, 2011.

K.-C. Li. Sliced inverse regression for dimension reduction (with discussion). *Journal of the American Statistical Association*, 86(414):316–342, 1991.

K-C Li. On principal Hessian directions for data visualization and dimension reduction: another application of Stein's lemma. *Journal of the American Statistical Association*, 87(415):1025–1039, 1992.

Yi Lin. A note on margin-based loss functions in classification. *Statistics & Probability Letters*, 68(1):73–82, 2004.

Yufeng Liu, Hao Helen Zhang, and Yichao Wu. Hard or soft classification? large-margin unified machines. *Journal of the American Statistical Association*, 106(493):166–177, 2011.

Wei Luo and Bing Li. Combining eigenvalues and variation of eigenvectors for order determination. *Biometrika*, 103(4):875–887, 2016.

Yanyuan Ma and Liping Zhu. Efficiency loss and the linearity condition in dimension reduction. *Biometrika*, 100(2):371–383, 2013.

Sayan Mukherjee and Qiang Wu. Estimation of gradients and coordinate covariation in classification. *Journal of Machine Learning Research*, 7:2481–2514, 2006.

Xiaotong Shen, George C Tseng, Xuegong Zhang, and Wing Hung Wong. On $\psi$-learning. *Journal of the American Statistical Association*, 98(463):724–734, 2003.

Seung Jun Shin, Yichao Wu, Hao Helen Zhang, and Yufeng Liu. Probability enhanced sufficient dimension reduction in binary classification. *Biometrics*, 70(3):546–555, 2014.

Seung Jun Shin, Yichao Wu, Hao Helen Zhang, and Yufeng Liu. Principal weighted support vector machines for sufficient dimension reduction in binary classification. *Biometrika*, 104(1):67–81, 2017.

G. Wahba. *Spline models for observational data*. SIAM. CBMS-NSF Regional Conference Series in Applied Mathematics, v. 59, 1990.

Hansheng Wang and Yingcun Xia. Sliced regression for dimension reduction. *Journal of the American Statistical Association*, 103(482):811–821, 2008.

J. Wang, X. Shen, and Y. Liu. Probability estimation for large-margin classifier. *Biometrika*, 95(1):149–167, 2008.

Stephen J Wright. Coordinate descent algorithms. *Mathematical Programming*, 151(1): 3–34, 2015.

Yichao Wu, Hao Helen Zhang, and Yufeng Liu. Robust model-free multiclass probability estimation. *Journal of the American Statistical Association*, 105(489):424–436, 2010.

Yingcun Xia. A constructive approach to the estimation of dimension reduction directions. *The Annals of Statistics*, 35(6):2654–2690, 2007.

Yingcun Xia, Howell Tong, Wai Keungxs Li, and Li-Xing Zhu. An adaptive estimation of dimension reduction space. *Journal of the Royal Statistical Society, Series B*, 64(3): 363–410, 2002.

Lei Yang, Shaogao Lv, and Junhui Wang. Model-free variable selection in reproducing kernel hilbert space. *The Journal of Machine Learning Research*, 17(1):2885–2908, 2016.

Gui-Bo Ye and Xiaohui Xie. Learning sparse gradients for variable selection and dimension reduction. *Machine Learning*, 87(3):303–355, 2012.

Zhishen Ye and Robert E Weiss. Using the bootstrap to select one of a new class of dimension reduction methods. *Journal of the American Statistical Association*, 98(464): 968–979, 2003.

Xiangrong Yin, Bing Li, and R Dennis Cook. Successive direction extraction for estimating the central subspace in a multiple-index regression. *Journal of Multivariate Analysis*, 99 (8):1733–1757, 2008.

Chong Zhang and Yufeng Liu. Multicategory angle-based large-margin classification. *Biometrika*, 101(3):625–640, 2014.

Hao Helen Zhang and Wenbin Lu. Adaptive lasso for cox's proportional hazards model. *Biometrika*, 94(3):691–703, 2007.

Li-Ping Zhu, Li-Xing Zhu, and Zheng-Hui Feng. Dimension reduction in regressions through cumulative slicing estimation. *Journal of the American Statistical Association*, 105(492): 1455–1466, 2010.

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

Hui Zou and Ming Yuan. Composite quantile regression and the oracle model selection theory. *The Annals of Statistics*, 36(3):1108–1126, 2008.