Efficient Estimation in Single Index Models through Smoothing splines

Arun K. Kuchibhotla and Rohit K. Patra*

University of Pennsylvania and University of Florida,
e-mail: arunku@upenn.edu*;* rohitpatra@ufl.edu

Abstract: We consider estimation and inference in a single index regression model with an unknown but smooth link function. In contrast to the standard approach of using kernel methods, we use smoothing splines to estimate the smooth link function. We develop a method to compute the penalized least squares estimators (PLSEs) of the parametric and the nonparametric components given independent and identically distributed (i.i.d.) data. We prove the consistency and find the rates of convergence of the estimators. We establish $n^{-1/2}$-rate of convergence and the asymptotic efficiency of the parametric component under mild assumptions. A finite sample simulation corroborates our asymptotic theory and illustrates the superiority of our procedure over existing procedures. We also analyze a car mileage data set and a ozone concentration data set. The identifiability and existence of the PLSEs are also investigated.

Keywords and phrases: least favorable submodel, penalized least squares, semiparametric model.

1. Introduction

Consider a regression model where one observes i.i.d. copies of the predictor $X \in \mathbb{R}^d$ and the response $Y \in \mathbb{R}$ and is interested in estimating the regression function $\mathbb{E}(Y|X = \cdot)$. In nonparametric regression $\mathbb{E}(Y|X = \cdot)$ is generally assumed to satisfy some smoothness assumptions (e.g., twice continuously differentiable), but no assumptions are made on the form of dependence on X. While nonparametric models offer flexibility in modeling, the price for this flexibility can be high for two main reasons: the estimation precision decreases rapidly as d increases ("curse of dimensionality") and the estimator can be hard to interpret when $d > 1$.

A natural restriction of the nonparametric model that avoids the curse of dimensionality while still retaining some flexibility in the functional form of $\mathbb{E}(Y|X = \cdot)$ is the single index model. In single index models, one assumes the existence of $\theta_0 \in \mathbb{R}^d$ such that

$$\mathbb{E}(Y|X) = \mathbb{E}(Y|\theta_0^\top X), \quad \text{almost every (a.e.) } X,$$

where $\theta_0^\top X$ is called the index; the widely used generalized linear models (GLMs) are special cases. This dimension reduction gives single index models considerable advantages in applications when $d > 1$ compared to the general nonparametric regression model; see Horowitz (2009) and Carroll et al. (1997) for a discussion. The aggregation of dimension by the index enables us to estimate the conditional mean function at a much faster rate than in a general nonparametric model. Since Powell et al. (1989), single index models have become increasingly popular in many scientific fields including biostatistics, economics, finance, and environmental science and have been deployed in a variety of settings; see Li and Racine (2007).

*Corresponding author.

Formally, in this paper, we consider the model

$$Y = m_0(\theta_0^\top X) + \epsilon, \quad \mathbb{E}(\epsilon|X) = 0, \quad \text{a.e.}\, X, \tag{1}$$

where $m_0 : \mathbb{R} \to \mathbb{R}$ is called the link function, $\theta_0 \in \mathbb{R}^d$ is the index parameter, and ϵ is the unobserved mean zero error (with finite variance). We assume that both m_0 and θ_0 are unknown and are the parameters of interest. For identifiability of (1), we assume that the first coordinate of θ_0 is non-zero and

$$\theta_0 \in \Theta := \{\eta = (\eta_1, \dots, \eta_d) \in \mathbb{R}^d : |\eta| = 1 \text{ and } \eta_1 \geq 0\} \subset S^{d-1},$$

where $|\cdot|$ denotes the Euclidean norm, and S^{d-1} is the Euclidean unit sphere in $\mathbb{R}^d$; see Carroll et al. (1997) and Cui et al. (2011) for a similar assumption.

Most of the existing techniques for estimation in single index models can be broadly classified into two groups, namely, M-estimation and "direct" estimation. M-estimation involves a nonparametric regression estimator of m_0, e.g., kernel estimator (Ichimura (1993)), regression splines (Antoniadis et al. (2004)), B-splines (Antoniadis et al. (2004)), and penalized splines (Yu and Ruppert (2002)), and a minimization of a valid criterion function with respect to the index parameter to obtain an estimator of θ_0. The so-called direct estimation methods include average derivative estimators (see e.g., Stoker (1986), Powell et al. (1989), and Hristache et al. (2001)), methods based on the conditional variance of Y (see Xia et al. (2002) and Xia (2006)), and dimension reduction techniques, such as sliced inverse regression (see Li and Duan (1989) and Li (1991)) and partial least squares (see Zhou and He (2008)); Cui et al. (2011) propose a kernel-based fixed point iterative scheme to compute an efficient estimator of θ_0. In these methods one tries to directly estimate θ_0 without estimating m_0, e.g., in Hristache et al. (2001) the authors use the estimate of the derivative of the local linear approximation to $\mathbb{E}(Y|X = \cdot)$ and not the estimate of m_0 to estimate θ_0.

In this paper we propose an M-estimation technique based on smoothing splines to simultaneously estimate the link function m_0 and the index parameter θ_0. When θ_0 is known, (1) reduces to a one-dimensional function estimation problem and smoothing splines offer a fast and easy-to-implement nonparametric estimator of the link function — m_0 is generally estimated by minimizing a penalized least squares criterion with a (natural) roughness penalty of integrated squared second derivative; see Wahba (1990) and Green and Silverman (1994). However, in the case of single index models, the problem is considerably harder as both the link function and the index parameter are unknown and intertwined (unlike in partial linear regression model; see Härdle and Liang (2007)).

In other words, given i.i.d. data $\{(y_i, x_i)\}_{1 \leq i \leq n}$ from model (1), we propose minimizing the following penalized loss:

$$\frac{1}{n}\sum_{i=1}^{n}\left(y_i - m(\theta^\top x_i)\right)^2 + \lambda^2 \int |m''(t)|^2 dt \quad (\lambda \neq 0) \tag{2}$$

over $\theta \in \Theta$ and all differentiable functions m with absolutely continuous derivative. Here λ is known as the smoothing parameter — high values of $|\lambda|$ lead to smoother estimators. To the best of our knowledge, this is the first work that uses smoothing splines in the single index paradigm, under (only) smoothness constraints. We show that the penalized least squares loss leads to a minimizer $(\hat{m}, \hat{\theta})$. We study the asymptotic properties, i.e., consistency, rates of convergence, of the estimator $(\hat{m}, \hat{\theta})$ under data dependent choices of the tuning parameter λ. We show that under sub-Gaussian errors, $\hat{\theta}$ is a $\sqrt{n}$-consistent estimator of θ_0 and, further, under homoscedastic errors $\hat{\theta}$ achieves the optimal semiparametric efficiency bound in the sense of Bickel et al. (1993).

Ichimura (1993) developed a semiparametric least squares estimator of θ_0 using kernel estimates of the link function. However, the choice of tuning parameters (e.g., the bandwidth for estimation of the link function) make this procedure difficult to implement (see Härdle et al. (1993) and Delecroix et al. (2006)) and its numerical instability is well documented; see e.g., Yu and Ruppert (2002). To address these issues Yu and Ruppert (2002) used a penalized spline to estimate m_0. However, in their proposed procedure the practitioner is required to choose the (fixed) number and placement of knots for every θ for fitting a spline to the nonparametric component. Moreover, to prove the consistency of their proposed estimators they assumed that m_0 is a spline and has a fixed (known) number of knots. They note that for consistency of a spline-based estimator (when m_0 is not a spline) one should let the number of knots increase with sample size; see page 1044, Section 3 of Yu and Ruppert (2002). Smoothing splines avoid the choice of number of knots and their placement. Moreover, the number of knots in a smoothing spline estimate increase to infinity with sample size. This motivates us to use smoothing splines for estimation in the single index model.

This paper gives a systematic and rigorous study of a smoothing splines based estimator for the single index model under minimal assumptions and fills an important gap in the literature. The assumptions for m_0 in this paper are weaker than those considered in the literature. We assume that the link function has an absolutely continuous derivative as opposed to the assumed (almost) three times differentiability of m_0, see e.g., Powell et al. (1989), Ichimura (1993), and Cui et al. (2011). Our treatment of the finite dimensional parameter is also novel. In contrast to the existing approaches where the first coordinate of θ is assumed to be 1, we study the model under the assumption that $\theta \in S^{d-1}$. When the first coordinate is assumed to be 1, the parameter space is unbounded and consistent estimation of θ_0 requires further assumptions, see e.g., Li and Patilea (2015). Cui et al. (2011) point out that the assumption $\theta \in S^{d-1}$ makes the parameter space irregular and the construction of paths on the sphere is hard. In this paper we construct paths on the unit sphere to study the semiparametric efficiency of the finite dimensional parameter.

The theory developed in this paper allows for the tuning parameter λ in (2) to be data dependent. Thus data-driven procedures such as cross-validation can be used to choose an optimal λ; see Section 5. As opposed to average derivative methods discussed eariler (see Powell et al. (1989) and Hristache et al. (2001)), the optimization problem in (2) involves only 1-dimensional nonparametric function estimation.

Our exposition is organized as follows. In Section 2 we introduce some notation, formally define our estimator, and study its existence. We state and discuss our assumptions in Section 3 and prove consistency (see Theorem 3) and provide the rates of convergence (see Theorems 2 and 4) for our estimator. We show that the estimator for θ_0 is asymptotically normal (properly normalized) and is semiparametrically efficient; see Theorem 5 in Section 4. In Section 5 we provide finite sample simulation study of the proposed estimator and compare performance with existing methods in the literature. In Section 6, we apply the methodology developed to the car mileage data and the ozone concentration data. In Section 7, we briefly summarize the results in the paper. Sections 8–10 contain proofs of the results in the paper.

2. Preliminaries

Suppose that $\{(y_i, x_i)\}_{1\le i\le n}$ is an i.i.d. sample from model (1). We start with some notation. Let $\chi \subset \mathbb{R}^d$ denote the support of X. Let D be the set of possible index values and D_0 be the set of possible index values at θ_0, i.e.,

$$D := \{\theta^\top x : x \in \chi, \theta \in \Theta\} \quad \text{and} \quad D_0 := \{\theta_0^\top x : x \in \chi\}.$$

We denote the class of all real-valued functions with absolutely continuous first derivative on D by $\mathcal{S}$, i.e.,

$$\mathcal{S} := \{m : D \to \mathbb{R} |\, m' \text{ is absolutely continuous}\}.$$

We use $\mathbb{P}$ to denote the probability of an event, $\mathbb{E}$ for the expectation of a random quantity, and P_X for the distribution of X. For $g : \chi \to \mathbb{R}$, define

$$\|g\|^2 := \int_\chi g^2 dP_X \qquad \text{and} \qquad \|g\|_n^2 = \frac{1}{n}\sum_{i=1}^n g^2(x_i).$$

Let $P_{\epsilon,X}$ denote the joint distribution of (ϵ, X) and $P_{\theta,m}$ denote the joint distribution of (Y, X) when $Y := m(\theta^\top X) + \epsilon$. In particular, P_{θ_0,m_0} denotes the joint distribution of (Y, X) when (Y, X) satisfy (1). For any function $g : I \subset \mathbb{R}^p \to \mathbb{R}$, let

$$\|g\|_\infty := \sup_{u\in I} |g(u)|.$$

Moreover, for $I_1 \subset I$, we define $\|g\|_{I_1} := \sup_{u\in I_1} |g(u)|$. For any set $I \in \mathbb{R}$, $\varnothing(I)$ denotes the diameter of the set I. For any $a \in \mathbb{R}^d$ and $r > 0$, $B_a(r)$ denotes the Euclidean ball of radius r centered at a. The notation $a \lesssim b$ is used to express that a is less than b up to a positive constant multiple. For any function $f : \chi \to \mathbb{R}^r, r \ge 1$, let $\{f_i\}_{1\le i\le r}$ denote each of the components, i.e., $f(x) = (f_1(x), \ldots, f_r(x)), r \ge 1$ and $f_i : \chi \to \mathbb{R}$. We define $\|f\|_{2,P_{\theta_0,m_0}} := \sqrt{\sum_{i=1}^r \|f_i\|^2}$ and $\|f\|_{2,\infty} := \sqrt{\sum_{i=1}^r \|f_i\|_\infty^2}$. For any real-valued function m and $\theta \in \Theta$, we define

$$(m \circ \theta)(x) := m(\theta^\top x), \qquad \text{for all } x \in \chi.$$

For any function $f : D \subset \mathbb{R} \to \mathbb{R}$ with absolutely continuous first derivative, we define the roughness penalty

$$J^2(f) := \int_D |f''(t)|^2 dt.$$

We assume that for the true link function m_0, $J(m_0) < \infty$ (see assumption **(A1)** in Section 3). The penalized loss for $(m, \theta) \in \mathcal{S} \times \Theta$ (and $\lambda \neq 0$) is defined as

$$\mathcal{L}_n(m, \theta; \lambda) := \frac{1}{n}\sum_{i=1}^n \left(y_i - m(\theta^\top x_i)\right)^2 + \lambda^2 J^2(m). \tag{3}$$

For simplicity of notation, we define

$$Q_n(m, \theta) := \frac{1}{n}\sum_{i=1}^n \left(y_i - m(\theta^\top x_i)\right)^2.$$

In this paper we study the following penalized least square estimator (PLSE):

$$(\hat{m}, \hat{\theta}) := \underset{(m,\theta)\in\mathcal{S}\times\Theta}{\arg\min}\ \mathcal{L}_n(m, \theta; \lambda). \tag{4}$$

Here we suppress the dependence of $(\hat{m}, \hat{\theta})$ on λ, for notational convenience. The following theorem (proved in Section 8.1) proves the existence of $(\hat{m}, \hat{\theta})$ for every $\lambda \neq 0$.

Theorem 1. *$\hat{\theta} \in \Theta$ and $\hat{m} \in \mathcal{S}$, where $\hat{\theta}$ and $\hat{m}$ are defined in (4). Moreover, $\hat{m}$ is a natural cubic spline with knots at $\{\hat{\theta}^\top x_i\}_{1\le i\le n}$.*

It is easy to see that the composite population parameter $m_0 \circ \theta_0$ is identifiable. However, this does not guarantee that both m_0 and θ_0 are separately identifiable. Ichimura (1993) (also see Horowitz (1998)) finds sufficient conditions on the distribution/domain of X under which θ_0 and m_0 can be separately identified when m_0 is a non-constant differentiable function:

(A0) Assume that for some integer $d_1 \in [1, d]$, $X_1, \ldots, X_{d_1-1}$, and X_{d_1} have continuous distributions and $X_{d_1+1}, \ldots, X_{d-1}$, and X_d be discrete random variables. Furthermore, assume that for each $\theta \in \Theta$ there exist an open interval $\mathcal{I}$ and constant vectors $c_0, c_1, \ldots, c_{d-d_1} \in \mathbb{R}^{d-d_1}$ such that

- $c_l - c_0$ for $l \in \{1, \ldots, d - d_1\}$ are linearly independent,
- $\mathcal{I} \subset \bigcap_{l=0}^{d-d_1} \left\{\theta^\top x : x \in \chi \text{ and } (x_{d_1+1}, \ldots x_d) = c_l\right\}$.

3. Asymptotic analysis of the PLSE

We now list the assumptions under which we will establish consistency and find the rates of convergence of our estimators. Note that we will study $(\hat{m}, \hat{\theta})$ for a certain (possibly data-driven) choice of λ satisfying two rate conditions; see assumption **(A4)** below.

(A1) The link function m_0 is bounded by some constant M_1 on D and satisfies $J(m_0) < \infty$.

(A2) χ, the support of X, is a compact subset of $\mathbb{R}^d$ and we assume that $\sup_{x\in\chi} |x| \le T$.

(A3) The error ϵ in model (1) is assumed to be uniformly sub-Gaussian, i.e., there exists $K_1 > 0$ such that

$$K_1\mathbb{E}\big(\exp(\epsilon^2/K_1) - 1|X\big) \le 1 \text{ a.e. } X.$$

As stated in (1), we also assume that $\mathbb{E}(\epsilon|X) = 0$ a.e. X.

(A4) The smoothing parameter λ can be chosen to be a random variable. For the rest of the paper, we denote it by $\hat{\lambda}_n$. Assume that $\hat{\lambda}_n$ satisfies the rate condition:

$$\hat{\lambda}_n^{-1} = O_p(n^{2/5}) \qquad \text{and} \qquad \hat{\lambda}_n = o_p(n^{-1/4}). \tag{5}$$

(A5) $\mathrm{Var}(X)$ is a positive definite matrix.

(A6) $\mathbb{E}\big[XX^\top|m_0'(\theta_0^\top X)|^2\big]$ is a nonsingular matrix.

The assumptions deserve comments. In **(A1)** our assumption on m_0 is quite minimal — we essentially require m_0 to have an absolutely continuous derivative. Most previous works assume m_0 to be three times differentiable; see e.g., Powell et al. (1989) and Newey and Stoker (1993). **(A2)** assumes that the support of the covariates is bounded. As the class of functions $\mathcal{S}$ is not uniformly bounded, we need assumption **(A3)** to provide control over the tail behavior of ϵ; see Chapter 8 of van de Geer (2000) for a discussion on this. Observe that **(A3)** allows for heteroscedastic errors. Assumption **(A4)** allows our tuning parameter to be data dependent, as opposed to a sequence of constants. This allows for data driven choices of $\hat{\lambda}_n$, such as cross-validation. We will show that for any choice of $\hat{\lambda}_n$ satisfying (5), $\hat{\theta}$ will be an asymptotically "efficient" estimator of θ_0. We use empirical process methods (e.g., see van der Vaart (1998)) to prove the consistency and to find the rates of convergence of $\hat{m}$ and $\hat{\theta}$. Assumptions **(A5)** and **(A6)** are mild distributional assumptions on the design. Assumption **(A5)** guarantees that the predictors are not supported on a lower dimensional affine space. Moreover, if $\mathrm{Var}(X)$ is singular the model (1) is not identifiable. Note that **(A6)** fails if m_0 is a constant function; however a single index model is not identifiable if m_0 is constant (see **(A0)**).

In Theorem 2 we show that $(\hat{m}, \hat{\theta})$ is a consistent estimator of (m_0, θ_0) and $\hat{m} \circ \hat{\theta}$ converges to $m_0 \circ \theta_0$ at rate $\hat{\lambda}_n$ (with respect to the $L_2(P_X)$-norm).

Theorem 2. *Under assumptions **(A0)–(A5)**, the PLSE satisfies* $J(\hat{m}) = O_p(1)$, $\|\hat{m}\|_\infty = O_p(1)$, *and* $\|\hat{m} \circ \hat{\theta} - m_0 \circ \theta_0\| = O_p(\hat{\lambda}_n)$.

Next we prove the consistency of $\hat{m}$ and $\hat{\theta}$. We prove that $\hat{m}$ is consistent under the Sobolev norm, which for any set $I \subset \mathbb{R}$ and any function $g : I \to \mathbb{R}$ is defined as

$$\|g\|_I^S = \sup_{t \in I} |g(t)| + \sup_{t \in I} |g'(t)|.$$

Theorem 3. *Under assumptions **(A0)–(A6)**,* $\hat{\theta} \xrightarrow{P} \theta_0$, $\|\hat{m} - m_0\|_{D_0}^S \xrightarrow{P} 0$, *and* $\|\hat{m}'\|_\infty = O_p(1)$.

The above result shows that not only is $\hat{m}$ consistent but its derivative $\hat{m}'$ also converges uniformly to m_0'. The following theorem provides an upper bound on the rates of convergence of $\hat{\theta}$ and $\hat{m}$ separately. The following bounds will help us compute the asymptotic distribution of $\hat{\theta}$ in Section 4.

Theorem 4. *Under **(A0)–(A6)** and the assumption that the conditional distribution of X given $\theta_0^\top X$ is non-degenerate, $\hat{m}$ and $\hat{\theta}$ satisfy*

$$|\hat{\theta} - \theta_0| = O_p(\hat{\lambda}_n) \quad \textit{and} \quad \|\hat{m} \circ \theta_0 - m_0 \circ \theta_0\| = O_p(\hat{\lambda}_n).$$

Proofs of Theorems 2, 3, and 4 are given in Sections 9.1, 9.3, and 9.4, respectively.

4. Semiparametric inference

In this section we show that $\hat{\theta}$ is asymptotically normal and is a semiparametrically efficient estimator of θ_0 under homoscedastic errors. Before going into the derivation of the limit law of $\hat{\theta}$, we need to introduce some further notation and some regularity assumptions. For every $\theta \in \Theta$, let us define $D_\theta := \{\theta^\top x : x \in \mathcal{X}\}$.

(B1) Assume that there exists $r > 0$ such that for all $\theta \in S^{d-1} \cap B_{\theta_0}(r)$ we have

$$D_\theta \subsetneq D^{(r)} := \bigcup_{\theta \in S^{d-1} \cap B_{\theta_0}(r)} D_\theta.$$

For the rest of the paper we redefine $D := D^{(r)}$. For every $\theta \in \Theta$, define $h_\theta : D \to \mathbb{R}^d$ as

$$h_\theta(u) := \mathbb{E}[X|\theta^\top X = u]. \tag{6}$$

(B2) Assume that $h_\theta(\cdot)$ is twice continuously differentiable except possibly at a finite number of points, and for every θ_1 and θ_2 in Θ,

$$\|h_{\theta_1} - h_{\theta_2}\|_\infty \le \bar{M}|\theta_1 - \theta_2|,$$

where $\bar{M}$ is a fixed finite constant.

Let $p_{\epsilon,X}$ denote the joint density (with respect to some dominating measure μ on $\mathbb{R} \times \chi$) of (ϵ, X). Let $p_{\epsilon|X}(e,x)$ and $p_X(x)$ denote the corresponding conditional probability density of ϵ given X and the marginal density of X, respectively. We define $\sigma : \chi \to \mathbb{R}$ as $\sigma^2(x) := \mathbb{E}(\epsilon^2|X = x)$.

(B3) Assume that $p_{\epsilon|X}(e,x)$ is differentiable with respect to e, $\|\sigma^2(\cdot)\|_\infty < \infty$ and $\|1/\sigma^2(\cdot)\|_\infty < \infty$.

The assumptions **(B1)**–**(B3)** deserve comments. Assumption **(B1)** guarantees that the true index set $D_0(= \{\theta_0^\top x : x \in \chi\})$ does not lie on the boundary of D. The function h_θ plays a crucial role in the construction of "least favorable" paths; see Section 4.2.2. For the functions in the path to be in $\mathcal{S}$, we need the smoothness assumptions on h_θ. **(B3)** gives lower and upper bounds on the variance of ϵ as we are using a non-weighted least squares method to estimate parameters in a (possibly) heteroscedastic model.

In the sequel we will use standard empirical process theory notation. For any function $f : \mathbb{R} \times \chi \to \mathbb{R}$ and $(m, \theta) \in \mathcal{S} \times \Theta$, we define

$$P_{\theta,m} f = \int f dP_{\theta,m}.$$

Note that $P_{\theta,m} f$ can be a random variable if θ (or m) is random. Moreover, for any function $f : \mathbb{R} \times \chi \to \mathbb{R}$, we define

$$\mathbb{P}_n f := \frac{1}{n}\sum_{i=1}^n f(y_i, x_i) \quad \text{and} \quad \mathbb{G}_n f := \frac{1}{\sqrt{n}}\sum_{i=1}^n \big[f(y_i, x_i) - P_{\theta_0,m_0} f\big].$$

4.1. Efficient score

As a first step in showing that $\hat{\theta}$ is an efficient estimator, in the following we find the efficiency bound for θ_0 in model (1). Recall that Θ denotes the finite dimensional parameter space. Note that Θ is a closed subset of $\mathbb{R}^d$ and the interior of Θ in $\mathbb{R}^d$ is the null set. For any $a \in \mathbb{R}^d$, let a_{-1} denote the last $d-1$ coordinates of a. Another common reparameterization of the finite dimensional parameter in (1) is to write $\theta = (1, \theta_{-1})$, where $\theta_{-1} \in \mathbb{R}^{d-1}$. However in this alternative parameterization, the finite dimensional parameter space is no longer bounded. As most estimators for θ are minimizers/solutions of some criterion function, further assumptions on the estimator of θ_0 are needed to make sure that the estimator does not diverge; see e.g., Section 2 of Li and Zhang (1998) and Li and Patilea (2015). In this paper we consider a local parameterization to construct paths on Θ. The local parameterization maps $\mathbb{R}^{d-1}$ onto Θ and gives a simple form for the parametric scores. First we introduce some notation: for every real matrix $G \in \mathbb{R}^{m\times n}$, we define $\|G\|_2 := \max_{x\in S^{n-1}} |Gx|$. This is sometimes called the operator or matrix 2-norm; see e.g., page 281 of Meyer (2000). The following lemma proved[1] in Section 10.1 shows that the "local parameterization matrix" as a function of θ is Lipschitz at θ_0 with respect to the operator norm.

Lemma 1. *There exists a set of matrices $\{H_\theta \in \mathbb{R}^{d\times(d-1)} : \theta \in \Theta\}$ satisfying the following properties:*

(a) $\xi \mapsto H_\theta \xi$ are bijections from $\mathbb{R}^{d-1}$ to the hyperplanes $\{x \in \mathbb{R}^d : \theta^\top x = 0\}$.
(b) The columns of H_θ form an orthonormal basis for $\{x \in \mathbb{R}^d : \theta^\top x = 0\}$.
(c) $\|H_\theta - H_{\theta_0}\|_2 \le |\theta - \theta_0|$.
(d) For all distinct $\eta, \beta \in \Theta \setminus \theta_0$, such that $|\eta - \theta_0| \le 1/2$ and $|\beta - \theta_0| \le 1/2$,

$$\|H_\eta^\top - H_\beta^\top\|_2 \le 8(1 + 8/\sqrt{15})\frac{|\eta - \beta|}{|\eta - \theta_0| + |\beta - \theta_0|}. \tag{7}$$

Note that for each $\theta \in \Theta$, $H_\theta^\top$ is the Moore-Penrose pseudo-inverse of H_θ, e.g., $H_\theta^\top H_\theta = \mathbb{I}_{d-1}$ where $\mathbb{I}_{d-1}$ is the identity matrix of order $d-1$; see Section 5.2 of Patra et al. (2015) for a similar construction.

For any $\eta \in \mathbb{R}^{d-1}$ and $\theta \in \Theta$, we now define a path $s \mapsto \zeta_s(\theta, \eta)$, for $s \in \mathbb{R}$ and $|s| \le |\eta|^{-1}$, as

$$\zeta_s(\theta, \eta) := \sqrt{1 - s^2|\eta|^2}\,\theta + sH_\theta\eta. \tag{8}$$

Note that $\theta^\top H_\theta = 0$ and $|H_\theta \eta| = |\eta|$ for all $\eta \in \mathbb{R}^{d-1}$. When $|s| \le 1/|\eta|$ we have $\zeta_s(\theta,\eta) \in S^{d-1}$. For every fixed $s \ne 0$, as η varies in $B_0^{d-1}(|s|^{-1})$, $\zeta_s(\theta,\eta)$ takes all values in the set $\{\beta \in S^{d-1} : \theta^\top \beta > 0\}$ and $sH_\theta\eta$ is the orthogonal projection of $\zeta_s(\theta,\eta)$ onto the hyperplane $\{x \in \mathbb{R}^d : \theta^\top x = 0\}$.

We now attempt to calculate the efficient score for

$$Y = m(\theta^\top X) + \epsilon \tag{9}$$

for some $(m, \theta) \in \mathcal{S} \times \Theta$ under assumptions **(A3)** and **(B3)**. The log-likelihood of the model is

$$l_{\theta,m}(y, x) = \log\left[p_{\epsilon|X}\big(y - m(\theta^\top x), x\big) p_X(x)\right].$$

Remark 1. *Note that under (9), we have $\epsilon = Y - m(\theta^\top X)$. For every function $b(e,x) : \mathbb{R}\times\mathcal{X} \to \mathbb{R}$ in $L_2(P_{\epsilon,X})$, there exists an "equivalent" function $\tilde{b}(y,x) : \mathbb{R}\times\mathcal{X} \to \mathbb{R}$ in $L_2(P_{\theta,m})$ defined as $\tilde{b}(y,x) := b(y - m(\theta^\top x), x) \in L_2(P_{\theta,m})$. In this section, we use the function arguments (e,x) $(L_2(P_{\epsilon,X}))$ and (y,x) $(L_2(P_{\theta,m}))$ interchangeably.*

For $\eta \in S^{d-2} \subset \mathbb{R}^{d-1}$, consider the path defined in (8). Note that this is a valid path through θ as $\zeta_0(\theta,\eta) = \theta$. The score function for this submodel (the parametric score) is

$$\left.\frac{\partial l_{\zeta_s(\theta,\eta),m}(y,x)}{\partial s}\right|_{s=0} = \eta^\top S_{\theta,m}(y,x), \text{ where } S_{\theta,m}(y,x) := -\frac{p'_{\epsilon|X}\big(y - m(\theta^\top x), x\big)}{p_{\epsilon|X}\big(y - m(\theta^\top x), x\big)} m'(\theta^\top x) H_\theta^\top x.$$

[1]Our proof is constructive.

We now define a parametric submodel for the unknown nonparametric components:

$$\begin{aligned} m_{s,a}(t) &= m(t) - sa(t), \\ p_{\epsilon|X;s,b}(e,x) &= p_{\epsilon|X}(e,x)(1 + sb(e,x)), \\ p_{X;s,q}(x) &= p_X(x)(1 + sq(x)), \end{aligned}$$

where $s \in \mathbb{R}$, $b : \mathbb{R}\times\chi \to \mathbb{R}$ is a bounded function such that $\mathbb{E}(b(\epsilon,X)|X) = 0$ and $\mathbb{E}(\epsilon b(\epsilon,X)|X) = 0$, $a \in \mathcal{S}$ such that $J(a) < \infty$ and $q : \chi \to \mathbb{R}$ is a bounded function such that $\mathbb{E}(q(X)) = 0$. Consider the following parametric submodel of (1),

$$s \mapsto (\zeta_s(\theta,\eta),\, m_{s,a},\, p_{\epsilon|X;s,b},\, p_{X;s,q}(x)) \tag{10}$$

where $\eta \in S^{d-2}$. Differentiating the log-likelihood of the submodel in (10) with respect to s, we get that the score along the submodel in (10) is

$$\eta^\top S_{\theta,m}(y,x) + \frac{p'_{\epsilon|X}\big(y - m(\theta^\top x), x\big)}{p_{\epsilon|X}\big(y - m(\theta^\top x), x\big)} a(\theta^\top x) + b(y - m(\theta^\top x), x) + q(x).$$

It is now easy to see that the nuisance tangent space, denoted by Λ, of the model is

$$\begin{aligned} \Lambda := \overline{\text{lin}}\Big\{ f \in L_2(P_{\epsilon,X}) : f(e,x) = \frac{p'_{\epsilon|X}(e,x)}{p_{\epsilon|X}(e,x)} a(\theta^\top x) + b(e,x) + q(x),\ \text{where} \\ a \in \mathcal{S}, J(a) < \infty, b : \mathbb{R}\times\chi \to \mathbb{R} \text{ and } q : \chi \to \mathbb{R} \text{ are bounded functions}, \\ \mathbb{E}(\epsilon b(\epsilon,X)|X) = 0,\ \mathbb{E}(b(\epsilon,X)|X) = 0,\ \text{and } \mathbb{E}(q(X)) = 0 \Big\}, \end{aligned}$$

where for any set $A \subset L_2(P_{\theta,m})$, $\overline{\text{lin}}\, A$ denotes the closure in $L^2(P_{\theta,m})$ of the linear span of functions in A; see Newey (1990) for a review of the construction of the nonparametric tangent set as a closure of scores of parametric submodels of the nuisance parameter. By Corollary A.1 of Györfi et al. (2002), we have that the class of infinitely often differentiable functions on D is dense in $L_2(\mathbf{m})$, where $\mathbf{m}$ denotes the Lebesgue measure on D. Thus we have that

$$\begin{aligned} \overline{\text{lin}}\{a \in \mathcal{S} : J(a) < \infty\} &= \{a : D \to \mathbb{R} |\, a \in L_2(\mathbf{m})\}, \\ \overline{\text{lin}}\{q : \chi \to \mathbb{R} |\, q \text{ is a bounded function and } \mathbb{E}(q(X)) = 0\} &= \{q \in L_2(P_X) |\, \mathbb{E}(q(X)) = 0\}, \end{aligned}$$

and

$$\begin{aligned} &\overline{\text{lin}}\{b : \mathbb{R}\times\chi \to \mathbb{R} |\, b \text{ is a bounded function}, \mathbb{E}(\epsilon b(\epsilon,X)|X) = \mathbb{E}(b(\epsilon,X)|X) = 0\} \\ &\qquad = \{b \in L_2(P_{\epsilon,X}) |\, \mathbb{E}(\epsilon b(\epsilon,X)|X) = \mathbb{E}(b(\epsilon,X)|X) = 0\}. \end{aligned}$$

Thus, it is easy to see that under assumptions **(A0)**–**(A6)** and **(B1)**–**(B3)**, the nuisance tangent space of (1) is

$$\begin{aligned} \Lambda = \Big\{ f \in L_2(P_{\epsilon,X}) :\ f(e,x) = \frac{p'_{\epsilon|X}(e,x)}{p_{\epsilon|X}(e,x)} a(\theta^\top x) + b(e,x) + q(x),\ \text{where} \\ a \in L_2(\mathbf{m}), b \in L_2(P_{\epsilon,X}), q \in L_2(P_X), \mathbb{E}(\epsilon b(\epsilon,X)|X) = 0, \\ \mathbb{E}(b(\epsilon,X)|X) = 0,\ \text{and } \mathbb{E}(q(X)) = 0 \Big\}, \end{aligned}$$

see Theorem 4.1 in Newey and Stoker (1993) and Proposition 1 of Ma and Zhu (2013) for a similar nuisance tangent space. Observe that the efficient score is the $L_2(P_{\epsilon,X})$ projection of $S_{\theta,m}(y,x)$ onto $\Lambda^\perp$, where $\Lambda^\perp$ is the orthogonal complement of Λ in $L_2(P_{\epsilon,X})$. Newey and Stoker (1993) and Ma and Zhu (2013) show that

$$\Lambda^\perp = \Big\{ f \in L_2(P_{\epsilon,X}) : \ f(e,x) = \big[g(x) - \mathbb{E}\big(g(X)|\theta^\top X = \theta^\top x\big)\big]e, \text{ for some } g : \mathcal{X} \to \mathbb{R} \Big\}.$$

Using calculations similar those in Proposition 1 in Ma and Zhu (2013), it can be shown that

$$\Pi(S_{\theta,m}|\Lambda^\perp)(y,x) = \frac{(y - m(\theta^\top x))}{\sigma^2(x)} m'(\theta^\top x) H_\theta^\top \left\{ x - \frac{\mathbb{E}(\sigma^{-2}(X)X|\theta^\top X = \theta^\top x)}{\mathbb{E}(\sigma^{-2}(X)|\theta^\top X = \theta^\top x)} \right\}, \tag{11}$$

where for any $f \in L_2(P_{\epsilon,X})$, $\Pi(f|\Lambda^\perp)$ denotes the $L_2(P_{\epsilon,X})$ projection of f onto the space $\Lambda^\perp$. $\Pi(S_{\theta,m}|\Lambda^\perp)$ is sometimes denoted by $S_{\theta,m}^{eff}$. It is important to note that the optimal estimating equation depends on $\sigma^2(\cdot)$. Since in the semiparametric model $\sigma^2(\cdot)$ is left unspecified, it is unknown. Without additional assumptions, nonparametric estimators of $\sigma^2(\cdot)$ have a slow rate of convergence to $\sigma^2(\cdot)$, especially if d is large. Thus if we substitute $\hat{\sigma}(x)$ in the efficient score equation, the solution of the modified score equation would lead to poor finite sample performance; see Tsiatis (2006).

To focus our presentation on the main concepts, briefly consider the case when $\sigma^2(\cdot) \equiv \sigma^2$. In this case the efficient score $\Pi(S_{\theta,m}|\Lambda^\perp)(y,x)$ is

$$\frac{1}{\sigma^2}(y - m(\theta^\top x)) m'(\theta^\top x) H_\theta^\top \left\{ x - h_\theta(\theta^\top x) \right\},$$

where $h_\theta(\theta^\top x)$ is defined in (6). Asymptotic normality and efficiency of $\hat{\theta}$ would follow if we can show that $(\hat{m}, \hat{\theta})$ satisfies the efficient score equation *approximately*, i.e.,

$$\mathbb{P}_n \left[\frac{1}{\sigma^2}(Y - \hat{m}(\hat{\theta}^\top X)) \hat{m}'(\hat{\theta}^\top X) H_{\hat{\theta}}^\top \{ X - h_{\hat{\theta}}(\hat{\theta}^\top X) \} \right] = o_p(n^{-1/2})$$

and a class of functions formed by the efficient score indexed by (θ, m) in a "neighborhood" of (θ_0, m_0) satisfies some "uniformity" conditions, e.g., it is a Donsker class. We formalize this notion of efficiency in Theorem 5 below.

4.2. Efficiency of $\hat{\theta}$

Theorem 5. *Assume that (Y, X) satisfies (1) and assumptions **(A0)–(A6)** and **(B1)–(B3)** hold. Define*

$$\tilde{\ell}_{\theta,m}(y,x) := \big(y - m(\theta^\top x)\big) m'(\theta^\top x) H_\theta^\top \left\{ x - h_\theta(\theta^\top x) \right\}. \tag{12}$$

If $V_{\theta_0,m_0} := P_{\theta_0,m_0}(\tilde{\ell}_{\theta_0,m_0} S_{\theta_0,m_0}^\top)$ is a nonsingular matrix in $\mathbb{R}^{(d-1)\times(d-1)}$, then

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0, H_{\theta_0} V_{\theta_0,m_0}^{-1} \tilde{I}_{\theta_0,m_0} (H_{\theta_0} V_{\theta_0,m_0}^{-1})^\top), \tag{13}$$

where $\tilde{I}_{\theta_0,m_0} := P_{\theta_0,m_0}(\tilde{\ell}_{\theta_0,m_0} \tilde{\ell}_{\theta_0,m_0}^\top)$. If we further assume that $\sigma^2(\cdot) \equiv \sigma^2$ and if the efficient information matrix, $\tilde{I}_{\theta_0,m_0}$, is nonsingular, then $\hat{\theta}$ is an efficient estimator of θ_0, i.e.,

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0, \sigma^4 H_{\theta_0} \tilde{I}_{\theta_0,m_0}^{-1} H_{\theta_0}^\top). \tag{14}$$

Remark 2. *Note that even if* $\mathbb{E}(\epsilon^2|X) \not\equiv \sigma^2$, $\hat{\theta}$ *is a consistent and asymptotically normal estimator of* θ. *When the constant variance assumption provides a good approximation to the truth, estimators similar to* $\hat{\theta}$ *have been known to have high relative efficiency with respect to the optimal semiparametric efficiency bound; see Page 94 of Tsiatis (2006) for a discussion. When* $\sigma^2(x) = V^2(\theta_0^\top x)$ *for some unknown real-valued function* V, *we can define a weighted PLSE as*

$$(\tilde{m}, \tilde{\theta}) := \underset{(m,\theta)\in\mathcal{S}\times\Theta}{\arg\min} \frac{1}{n}\sum_{i=1}^{n} \hat{w}(x_i)\big(y_i - m(\theta^\top x_i)\big)^2 + \hat{\lambda}_n^2 J^2(m),$$

where $\hat{w}(x)$ *is a consistent estimator of* $V^{-2}(\theta_0^\top x)$. *Theorem 5 can be easily generalized to show that* $\tilde{\theta}$ *is an efficient estimator of* θ_0.

Remark 3. *The asymptotic variance of* $\sqrt{n}(\hat{\theta} - \theta_0)$ *is the same as that obtained in Section 2.4 of Härdle et al. (1993). However, Härdle et al. (1993) require stronger smoothness assumptions on* m_0.

4.2.1. Proof of Theorem 5

In the following we give a sketch of the proof of (13). Some of the steps are proved in the following sections.

Step 1 In Theorem 6 we will show that $(\hat{m}, \hat{\theta})$ satisfy the efficient score equation *approximately*, i.e.,

$$\sqrt{n}\mathbb{P}_n \tilde{\ell}_{\hat{\theta},\hat{m}} = o_p(1). \tag{15}$$

Step 2 In Section 10.3 we prove that $\tilde{\ell}_{\hat{\theta},\hat{m}}$ is unbiased in the sense of van der Vaart (2002), i.e.,

$$P_{\hat{\theta},m_0} \tilde{\ell}_{\hat{\theta},\hat{m}} = 0. \tag{16}$$

Similar conditions have appeared before in proofs of asymptotic normality of the MLE (e.g., see Huang (1996)) and the construction of efficient one-step estimators (see Klaassen (1987)). The above condition essentially ensures that $\tilde{\ell}_{\theta_0,\hat{m}}$ is a good "approximation" to $\tilde{\ell}_{\theta_0,m_0}$; see Section 3 of Murphy and van der Vaart (2000) for further discussion.

Step 3 We prove

$$\mathbb{G}_n(\tilde{\ell}_{\hat{\theta},\hat{m}} - \tilde{\ell}_{\theta_0,m_0}) = o_p(1) \tag{17}$$

in Theorem 7. In view of (15) and (16) an equivalent formulation of (17) is

$$\sqrt{n}(P_{\hat{\theta},m_0} - P_{\theta_0,m_0})\tilde{\ell}_{\hat{\theta},\hat{m}} = \mathbb{G}_n \tilde{\ell}_{\theta_0,m_0} + o_p(1). \tag{18}$$

Step 4 To complete the proof of (13), it is enough to show that

$$\sqrt{n}(P_{\hat{\theta},m_0} - P_{\theta_0,m_0})\tilde{\ell}_{\hat{\theta},\hat{m}} = \sqrt{n} V_{\theta_0,m_0} H_{\theta_0}^\top(\hat{\theta} - \theta_0) + o_p(\sqrt{n}|\hat{\theta} - \theta_0|). \tag{19}$$

Observe that (18) and (19) imply

$$\begin{aligned} &\sqrt{n} V_{\theta_0,m_0} H_{\theta_0}^\top(\hat{\theta} - \theta_0) = \mathbb{G}_n \tilde{\ell}_{\theta_0,m_0} + o_p(1 + \sqrt{n}|\hat{\theta} - \theta_0|), \\ &\Rightarrow \sqrt{n} H_{\theta_0}^\top(\hat{\theta} - \theta_0) = V_{\theta_0,m_0}^{-1} \mathbb{G}_n \tilde{\ell}_{\theta_0,m_0} + o_p(1) \stackrel{d}{\to} V_{\theta_0,m_0}^{-1} N(0, \tilde{I}_{\theta_0,m_0}). \end{aligned} \tag{20}$$

The proof of the theorem will be complete if we can show that

$$\sqrt{n}(\hat{\theta} - \theta_0) = H_{\theta_0}\sqrt{n}H_{\theta_0}^\top(\hat{\theta} - \theta_0) + o_p(1).$$

Let $\hat{\eta}$ be the unique vector in $\mathbb{R}^{d-1}$ that satisfies the following equation:

$$\hat{\theta} = \sqrt{1 - |\hat{\eta}|^2}\,\theta_0 + H_{\theta_0}\hat{\eta}, \tag{21}$$

note that such an $\hat{\eta}$ will always exits as $\hat{\theta} \xrightarrow{P} \theta_0$. As $H_{\theta_0}^\top\theta_0 = 0$ and $H_{\theta_0}^\top H_{\theta_0} = \mathbb{I}_{d-1}$, pre-multiplying both sides of the previous equation by $H_{\theta_0}^\top$ we get

$$\hat{\eta} = H_{\theta_0}^\top(\hat{\theta} - \theta_0). \tag{22}$$

Substituting the above expression of $\hat{\eta}$ in (21) and subtracting θ_0 from both sides of (21) we get

$$\hat{\theta} - \theta_0 = \left[\sqrt{1 - |H_{\theta_0}^\top(\hat{\theta} - \theta_0)|^2} - 1\right]\theta_0 + H_{\theta_0}H_{\theta_0}^\top(\hat{\theta} - \theta_0).$$

By (20) we have that $\sqrt{n}H_{\theta_0}^\top(\hat{\theta} - \theta_0) = O_p(1)$. Moreover, note that $\sqrt{1-x^2} - 1 = O(x^2)$, as $x \to 0$. Combining the above facts, we get

$$\begin{aligned}\sqrt{n}(\hat{\theta} - \theta_0) &= \sqrt{n}O_p(|H_{\theta_0}^\top(\hat{\theta} - \theta_0)|^2) + \sqrt{n}H_{\theta_0}H_{\theta_0}^\top(\hat{\theta} - \theta_0)\\ &= H_{\theta_0}\sqrt{n}H_{\theta_0}^\top(\hat{\theta} - \theta_0) + O_p(n^{-1/2}).\end{aligned}$$

A proof of (19) can be found in the proof of Theorem 6.20 of van der Vaart (2002). However, for the sake of completeness we give a proof of (19) in Section 10.4.

Now we prove (14). Assume that $\sigma^2(\cdot) \equiv \sigma^2$. Observe that, by (11) and (12), we have

$$\begin{aligned}S_{\theta_0,m_0} &= \Pi(S_{\theta,m}|\Lambda^\perp) + \big(S_{\theta_0,m_0} - \Pi(S_{\theta,m}|\Lambda^\perp)\big)\\ &= \frac{1}{\sigma^2}\tilde{\ell}_{\theta_0,m_0} + \big(S_{\theta_0,m_0} - \Pi(S_{\theta,m}|\Lambda^\perp)\big).\end{aligned}$$

Thus (14) follows from (13) by observing that

$$V_{\theta_0,m_0} = P_{\theta_0,m_0}\big(\tilde{\ell}_{\theta_0,m_0}S_{\theta_0,m_0}^\top\big) = \frac{1}{\sigma^2}\tilde{I}_{\theta_0,m_0}.$$

4.2.2. "Least favorable" path for m

We will now show that **Step 1** holds, i.e., $(\hat{m}, \hat{\theta})$ satisfies the efficient score equation (15). Recall the definition (8). For any $(\theta, m) \in \Theta \times \{m \in \mathcal{S}|J(m) < \infty\}$ and $\eta \in S^{d-2}$, let $t \mapsto (\zeta_t(\theta,\eta), \xi_t(\cdot;\theta,\eta,m))$ denote a path in $\Theta \times \{m \in \mathcal{S}|J(m) < \infty\}$ that goes through (θ, m), i.e., $(\zeta_0(\theta,\eta), \xi_0(\cdot;\theta,\eta,m)) = (\theta, m)$; see (26) below for definition. Recall that $(\hat{\theta}, \hat{m})$ minimizes $\mathcal{L}_n(m, \theta, \hat{\lambda}_n)$. Hence, for every $\eta \in S^{d-2}$, the function $t \mapsto \mathcal{L}_n(\xi_t(\cdot;\hat{\theta},\eta,\hat{m}), \zeta_t(\hat{\theta},\eta), \hat{\lambda}_n)$ is minimized at $t = 0$. In particular, if the above function is differentiable in a neighborhood of 0, then

$$\frac{\partial}{\partial t}\mathcal{L}_n(\xi_t(\cdot;\hat{\theta},\eta,\hat{m}), \zeta_t(\hat{\theta},\eta), \hat{\lambda}_n)\bigg|_{t=0} = 0. \tag{23}$$

Moreover if $(\zeta_t(\hat{\theta},\eta), \xi_t(\cdot;\hat{\theta},\eta,\hat{m}))$ satisfies

$$\begin{aligned}\frac{\partial}{\partial t}\big(y - \xi_t(\zeta_t(\hat{\theta},\eta)^\top x;\hat{\theta},\eta,\hat{m})\big)^2\bigg|_{t=0} &= \eta^\top\tilde{\ell}_{\hat{\theta},\hat{m}}(y,x),\\ \frac{\partial}{\partial t}J^2(\xi_t(\cdot;\hat{\theta},\eta,\hat{m}))\bigg|_{t=0} &= O_p(1).\end{aligned} \tag{24}$$

for all $\eta \in S^{d-2}$, then we get (15) as $\hat{\lambda}_n^2 = o_p(n^{-1/2})$; see assumption **(A4)**.

Observe that $\hat{\theta}$ is a consistent estimator of θ_0. As we are concerned with the path $t \mapsto \mathcal{L}_n(\xi_t(\cdot;\hat{\theta},\eta,\hat{m}), \zeta_t(\hat{\theta},\eta), \hat{\lambda}_n)$, we will try to construct a path for any $(\theta, m) \in \{\Theta \cap B_{\theta_0}(r)\} \times \{m \in \mathcal{S} | J(m) < \infty\}$ that satisfies the above requirements. For any set $A \subset \mathbb{R}$ and any $\nu > 0$ let us define $A^\nu := \cup_{a \in A} B_a(\nu)$ and let ∂A denote the boundary of A. Fix $\nu > 0$. By assumption **(B1)**, for every $\theta \in \Theta \cap B_{\theta_0}(r)$, $\eta \in S^{d-2}$, and $t \in \mathbb{R}$ sufficiently close to zero, there exists a strictly increasing function $\phi_{\theta,\eta,t} : D^\nu \to \mathbb{R}$ with

$$\begin{aligned} \phi_{\theta,\eta,t}(u) &= u, \quad u \in D_\theta \\ \phi_{\theta,\eta,t}(u + (\theta - \zeta_t(\theta,\eta))^\top h_\theta(u)) &= u, \quad u \in \partial D, \end{aligned} \tag{25}$$

where $h_\theta(u)$ and $\zeta_t(\theta,\eta)$ are defined in (6) and (8), respectively. Furthermore, we can ensure that $\phi_{\theta,\eta,t}(u)$ is infinitely differentiable for $u \in D$ and that $\frac{\partial}{\partial t}\phi_{\theta,\eta,t}\big|_{t=0}$ exists. Note that $\phi_{\theta,\eta,t}(D) = D$. Moreover, $\phi_{\theta,\eta,t}$ cannot be the identity function for $t \neq 0$ if $(\theta - \zeta_t(\theta,\eta))^\top h_\theta(u) \neq 0$ for $u \in \partial D$. Now, we can define the following path through m:

$$\xi_t(u;\theta,\eta,m) := m \circ \phi_{\theta,\eta,t}(u + (\theta - \sqrt{1 - t^2|\eta|^2}\,\theta - tH_\theta\eta)^\top h_\theta(u)). \tag{26}$$

The function $\phi_{\theta,\eta,t}$ helps us control the partial derivative in the second equation of (24). In the following theorem (proved in Appendix 10.2), we show that $(\zeta_t(\hat{\theta},\eta), \xi_t(\cdot;\hat{\theta},\eta,\hat{m}))$ is a path through $(\hat{\theta},\hat{m})$ and satisfies (23) and (24). Here η is the "direction" for the path $t \mapsto \zeta_t(\theta,\eta)$ and $(\eta, h_\theta(u))$ defines the "direction" for the path $t \mapsto \xi_t(\cdot;\theta,\eta,m)$.

Theorem 6. *Under assumptions **(A0)**,**(A1)**, **(A4)**, and **(B1)**–**(B2)**, $(\zeta_t(\hat{\theta},\eta), \xi_t(\cdot;\hat{\theta},\eta,\hat{m}))$ is a valid parametric submodel, i.e., $(\zeta_t(\hat{\theta},\eta), \xi_t(\cdot;\hat{\theta},\eta,\hat{m})) \in \Theta \times \{m \in \mathcal{S}|J(m) < \infty\}$ for all t in some neighborhood of 0. Moreover $(\zeta_t(\hat{\theta},\eta), \xi_t(\cdot;\hat{\theta},\eta,\hat{m}))$ satisfies (24) and $\mathcal{L}_n(\xi_t(\cdot;\hat{\theta},\eta,\hat{m}), \zeta_t(\hat{\theta},\eta), \hat{\lambda}_n)$, as function of t, is differentiable at 0 and $\sqrt{n}\mathbb{P}_n\tilde{\ell}_{\hat{\theta},\hat{m}} = o_p(1)$.*

4.2.3. Asymptotic equicontinuity of $\tilde{\ell}_{\theta,m}$ at (θ_0, m_0)

For notational convenience we define

$$K_1(x;\theta) := H_\theta^\top(x - h_\theta(\theta^\top x)).$$

With the above notation, from (12) we have

$$\tilde{\ell}_{\theta,m}(y,x) = (y - m(\theta^\top x))m'(\theta^\top x)K_1(x;\theta).$$

Theorem 7. *Under assumptions **(A0)**–**(A6)** and **(B1)**–**(B3)**, $\mathbb{G}_n(\tilde{\ell}_{\hat{\theta},\hat{m}} - \tilde{\ell}_{\theta_0,m_0}) = o_p(1)$.*

We divide the proof Theorem 7 into two lemmas. First observe that

$$\begin{aligned} &\mathbb{G}_n(\tilde{\ell}_{\hat{\theta},\hat{m}} - \tilde{\ell}_{\theta_0,m_0}) \\ &= \mathbb{G}_n\big[\big(Y - \hat{m}(\hat{\theta}^\top X)\big)\hat{m}'(\hat{\theta}^\top X)K_1(X;\hat{\theta}) - \big(Y - m_0(\theta_0^\top X)\big)m_0'(\theta_0^\top X)K_1(X;\theta_0)\big] \\ &= \mathbb{G}_n\big[\big(\epsilon + m_0(\theta_0^\top X) - \hat{m}(\hat{\theta}^\top X)\big)\hat{m}'(\hat{\theta}^\top X)K_1(X;\hat{\theta}) - \epsilon m_0'(\theta_0^\top X)K_1(X;\theta_0)\big] \\ &= \mathbb{G}_n\big[\big(m_0(\theta_0^\top X) - \hat{m}(\hat{\theta}^\top X)\big)\hat{m}'(\hat{\theta}^\top X)K_1(X;\hat{\theta})\big] \\ &\qquad + \mathbb{G}_n\big[\epsilon\big(\hat{m}'(\hat{\theta}^\top X)K_1(X;\hat{\theta}) - m_0'(\theta_0^\top X)K_1(X;\theta_0)\big)\big]. \end{aligned} \tag{27}$$

The proof of Theorem 7 will be complete, if we can show that both the terms in (27) converge to 0 in probability. We begin with some definitions. Let a_n be a sequence of real numbers such that $a_n \to \infty$ as $n \to \infty$ and $a_n\|\hat{m} - m_0\|_{D_0}^S = o_p(1)$. We can always find such a sequence a_n, as we have $\|\hat{m} - m_0\|_{D_0}^S = o_p(1)$ (see Theorem 3). For all $n \in \mathbb{N}$, define [2]

$$
\begin{aligned}
\mathcal{C}_{M_1,M_2,M_3}^{m*} &:= \Big\{m \in \mathcal{S} : \|m\|_\infty < M_1,\ \|m'\|_\infty < M_2,\ \text{and } J(m) < M_3\Big\},\\
\mathcal{C}_{M_1,M_2,M_3}^{m}(n) &:= \Big\{m \in \mathcal{C}_{M_1,M_2,M_3}^{m*} : a_n\|m - m_0\|_{D_0}^S \le 1\Big\},\\
\mathcal{C}^{\theta}(n) &:= \Big\{\theta \in \Theta \cap B_{\theta_0}(1/2) : \hat{\lambda}_n^{-1/2}|\theta_0 - \theta| \le 1\Big\},\\
\mathcal{C}_{M_1,M_2,M_3}(n) &:= \Big\{(m,\theta) : \theta \in \mathcal{C}^{\theta}(n) \text{ and } m \in \mathcal{C}_{M_1,M_2,M_3}^{m}(n)\Big\},\\
\mathcal{C}_{M_1,M_2,M_3}^{*} &:= \Big\{(m,\theta) : \theta \in \Theta \cap B_{\theta_0}(1/2) \text{ and } m \in \mathcal{C}_{M_1,M_2,M_3}^{m*}\Big\}.
\end{aligned}
$$

Let us consider the first term of (27). Fix $\delta > 0$. For every fixed M_1, M_2, and M_3,

$$
\begin{aligned}
&\mathbb{P}\Big(\big|\mathbb{G}_n\big[\hat{m}' \circ \hat{\theta}\,(m_0 \circ \theta_0 - \hat{m} \circ \hat{\theta})K_1(\cdot;\hat{\theta})\big]\big| > \delta\Big)\\
&\le \mathbb{P}\Big(\big|\mathbb{G}_n\big[\hat{m}' \circ \hat{\theta}\,(m_0 \circ \theta_0 - \hat{m} \circ \hat{\theta})K_1(\cdot;\hat{\theta})\big]\big| > \delta, (\hat{m},\hat{\theta}) \in \mathcal{C}_{M_1,M_2,M_3}(n)\Big)\\
&\qquad + \mathbb{P}\big((\hat{m},\hat{\theta}) \notin \mathcal{C}_{M_1,M_2,M_3}(n)\big) \qquad (28)\\
&\le \mathbb{P}\Big(\sup_{(m,\theta)\in\mathcal{C}_{M_1,M_2,M_3}(n)}\big|\mathbb{G}_n\big[m' \circ \theta\,(m_0 \circ \theta_0 - m \circ \theta)K_1(\cdot;\theta)\big]\big| > \delta\Big)\\
&\qquad + \mathbb{P}\big((\hat{m},\hat{\theta}) \notin \mathcal{C}_{M_1,M_2,M_3}(n)\big).
\end{aligned}
$$

Recall that $(\hat{m},\hat{\theta})$ is a consistent estimator of (m_0,θ_0) and $\|\hat{m}'\|_\infty$ is $O_p(1)$; see Theorem 3. Furthermore, we have that both $\|\hat{m}\|_\infty$ and $J(\hat{m})$ are $O_p(1)$ (see Theorem 2) and $\hat{\lambda}_n^{-1/2}|\hat{\theta} - \theta_0| = o_p(1)$ (see Theorem 4). Thus for any $\varepsilon > 0$, there exists M_1, M_2, and M_3 (depending on ε) such that

$$
\mathbb{P}\left((\hat{m},\hat{\theta}) \notin \mathcal{C}_{M_1,M_2,M_3}(n)\right) \le \varepsilon,
$$

for all sufficiently large n. Hence, it is enough to show that for the above choice of M_1, M_2, and M_3, we have

$$
\mathbb{P}\Big(\sup_{(m,\theta)\in\mathcal{C}_{M_1,M_2,M_3}(n)}\big|\mathbb{G}_n\big[m' \circ \theta\,(m_0 \circ \theta_0 - m \circ \theta)K_1(\cdot;\theta)\big]\big| > \delta\Big) \le \varepsilon
$$

for sufficiently large n. The following lemma (proved in Section 10.5) shows this.

Lemma 2. *Fix M_1, M_2, M_3, and $\delta > 0$. For $n \in \mathbb{N}$, let us define two classes of functions from χ to $\mathbb{R}^d$*

$$
\begin{aligned}
\mathcal{D}_{M_1,M_2,M_3}(n) &:= \left\{m' \circ \theta(m_0 \circ \theta_0 - m \circ \theta)K_1(\cdot;\theta) : (m,\theta) \in \mathcal{C}_{M_1,M_2,M_3}(n)\right\},\\
\mathcal{D}_{M_1,M_2,M_3}^{*} &:= \left\{m' \circ \theta(m_0 \circ \theta_0 - m \circ \theta)K_1(\cdot;\theta) : (m,\theta) \in \mathcal{C}_{M_1,M_2,M_3}^{*}\right\}.
\end{aligned}
$$

$\mathcal{D}_{M_1,M_2,M_3}(n)$ is a Donsker class and

$$
\sup_{f\in\mathcal{D}_{M_1,M_2,M_3}(n)}\|f\|_{2,\infty} \le 2TM_2(a_n^{-1} + TM_2\hat{\lambda}_n^{1/2}) =: D_{M_1,M_2,M_3}(n). \qquad (29)
$$

[2] The notations with $*$ denote the classes of functions that do not depend on n while the ones with n denote shrinking neighborhoods around (m_0,θ_0).

Moreover, $J_{[\,]}(\gamma, \mathcal{D}_{M_1,M_2,M_3}(n), \|\cdot\|_{2,P_{\theta_0,m_0}}) \lesssim \gamma^{1/2}$, where for any class of functions $\mathcal{F}$, $J_{[\,]}$ is the entropy integral (see e.g., Page 270, van der Vaart (1998)) defined as

$$J_{[\,]}(\delta, \mathcal{F}, \|\cdot\|_{2,P_{\theta_0,m_0}}) := \int_0^{\delta} \sqrt{\log N_{[\,]}(t, \mathcal{F}, \|\cdot\|_{2,P_{\theta_0,m_0}})}dt.$$

Finally, we have

$$\mathbb{P}\Bigg(\sup_{f \in \mathcal{D}_{M_1,M_2,M_3}(n)} |\mathbb{G}_n f| > \delta\Bigg) \to 0 \qquad \text{as } n \to \infty.$$

The following lemma (proved in Section 10.6) shows that the second term on the right hand side of (27) converges to zero in probability.

Lemma 3. *Let us define $U_{\theta,m} : \chi \to \mathbb{R}^{d-1}$, $U_{\theta,m}(x) := m'(\theta^\top x)K_1(x;\theta)$. Fix M_1, M_2, M_3, and $\delta > 0$. For $n \in \mathbb{N}$, let us define*

$$\begin{aligned}\mathcal{W}_{M_1,M_2,M_3}(n) &:= \{U_{\theta,m} - U_{\theta_0,m_0} : (m,\theta) \in \mathcal{C}_{M_1,M_2,M_3}(n)\},\\ \mathcal{W}^*_{M_1,M_2,M_3} &:= \left\{U_{\theta,m} - U_{\theta_0,m_0} : (m,\theta) \in \mathcal{C}^*_{M_1,M_2,M_3}\right\}.\end{aligned}$$

Then $\mathcal{W}_{M_1,M_2,M_3}(n)$ is a Donsker class such that

$$\sup_{f \in \mathcal{W}_{M_1,M_2,M_3}(n)} \|f\|_{2,\infty} \le \left[2T^{3/2}M_3\hat{\lambda}_n^{1/4} + 2Ta_n^{-1} + M_2(2T + \bar{M})\hat{\lambda}_n^{1/2}\right] =: W_{M_1,M_2,M_3}(n).$$

Moreover, $J_{[\,]}(\gamma, \mathcal{W}_{M_1,M_2,M_3}(n), \|\cdot\|_{2,P_{\theta_0,m_0}}) \lesssim \gamma^{1/2}$. Hence, as $n \to \infty$, we have

$$\mathbb{P}\Big(\Big|\mathbb{G}_n\big[\epsilon\big(U_{\hat{\theta},\hat{m}} - U_{\theta_0,m_0}\big)\big]\Big| > \delta\Big) \to 0. \tag{30}$$

5. Simulation study

To investigate the finite sample performance of $(\hat{m}, \hat{\theta})$, we carry out several simulation experiments. We also compare the finite sample performance of the proposed estimator with the EFM estimator (estimating function method; see Cui et al. (2011)) and the EDR estimator (effective dimension reduction; see Hristache et al. (2001)). Cui et al. (2011) compares the performance of the EFM estimator to existing estimators such as the refined minimum average variance estimator (rMAVE) (see Xia et al. (2002)) and the EDR estimator and argues that EFM has improved overall performance compared to existing estimators. Thus we only include the EFM estimator and the EDR estimator in our simulation study. The code to compute the EDR estimates can be found in the R package `EDR`. Moreover, Cui et al. (2011) kindly provided us with the R codes to evaluate the EFM estimate. The codes used to implement our procedure are available in the `simest` package in R; see Kuchibhotla and Patra (2016). In what follows, we chose the penalty parameter $\hat{\lambda}_n$ for the PLSE through generalized cross validation (GCV), i.e., choose $\hat{\lambda}_n$ by minimizing $\text{GCV} : \mathbb{R} \to \mathbb{R}$

$$\text{GCV}(\lambda) := \frac{Q_n(\hat{m}_\lambda, \hat{\theta}_\lambda)}{1 - n^{-1}\text{trace}(A(\lambda))},$$

where $(\hat{m}_\lambda, \hat{\theta}_\lambda) := \arg\min_{(m,\theta)\in\mathcal{S}\times\Theta} \mathcal{L}_n(m, \theta; \lambda)$ and $A(\lambda)$ is the *hat* matrix for $\hat{m}_\lambda$ (see e.g., Sections 3.2 and 3.3 of Green and Silverman (1994) for a detailed description of $A(\lambda)$ and its connection to GCV); see Ruppert et al. (2003) for an extensive discussion on why GCV is an attractive choice for choosing the penalty parameter in the single index model. We choose $\hat{\lambda}_n$ by minimizing GCV score over a grid of values that satisfy assumption **(A4)**. For all the other methods considered in the paper we have used the suggested values of tuning parameters. In the following, we consider three different data generating mechanisms. From the simulation experiments it will be easy to see that the estimator proposed in this paper has the best overall performance. The codes used for the simulation examples can be found at `http://stat.ufl.edu/~rohitpatra/research`.

5.1. *A simple model*

We start with a simple model. Assume that $(X_1, X_2) \in \mathbb{R}^2$, $X_1 \sim$ Uniform$[-2, 2]$, $X_2 \sim$ Uniform$[0, 1]$, $\epsilon \sim N(0, .5^2)$, and

$$Y = (X^\top \theta_0)^2 + \epsilon, \text{ where } \theta_0 = (1, -1)/\sqrt{2}. \tag{31}$$

Observe that for this example, $H_{\theta_0}^\top = [1, 1]/\sqrt{2}$ (see Section 10.1) and the analytic expression of the efficient information is

$$\tilde{I}_{\theta_0, m_0} = 4\mathrm{Var}(\epsilon)\mathbb{E}\Big(\theta_0^\top X H_{\theta_0}^\top \big[X - \mathbb{E}\big(X|\theta_0^\top X\big)\big]\Big)^2 = 4\mathrm{Var}(\epsilon)\mathbb{E}\left|(\theta_0^\top X)^2 \big[H_{\theta_0}^\top \mathrm{Var}\big(X|\theta_0^\top X\big) H_{\theta_0}\big]\right|.$$

Using the above expression, we calculated the asymptotic variance of $\sqrt{n}(\hat{\theta}_1 - \theta_{0,1})$ to be 0.328. Figure 1 shows the box plots of the PLSE and compares its performance with the EFM and the EDR estimators. We also include the box plot of a sample (of size 500) from the true asymptotic distribution of $\hat{\theta}$ for comparison.

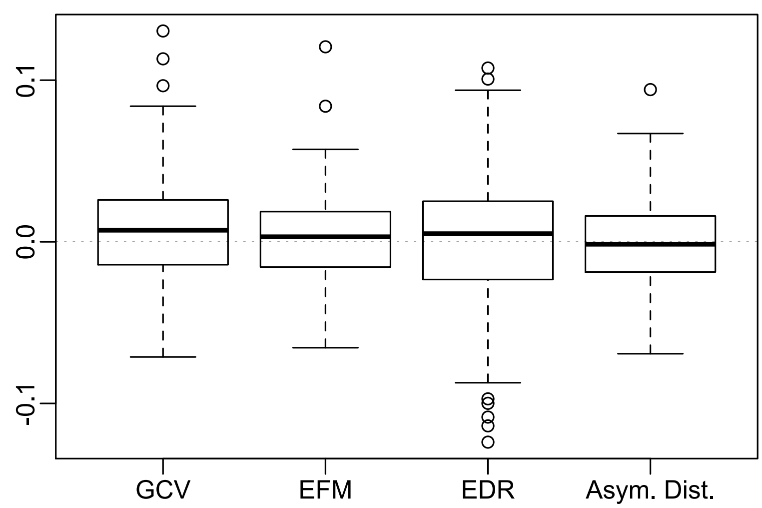

FIG 1. *Box plots of the first coordinate of the estimates (centered at $\theta_{0,1}$) from 500 replications along with the true asymptotic distribution of the $\hat{\theta}_1 - \theta_{0,1}$ when we have 500 i.i.d. samples from (31).*

5.2. *Dependent covariates*

We now consider a simulation scenario where covariates are dependent and the predictor $X \in \mathbb{R}^6$ contains discrete components. More precisely, $(X_1, \ldots, X_6)$ is generated according to the following law: $X_1 \sim \text{Uniform}[-1,1]$, $X_2 \sim \text{Uniform}[-1,1]$, $X_3 := 0.2X_1 + 0.2(X_2+2)^2 + 0.2Z_1$, $X_4 := 0.1 + 0.1(X_1+X_2) + 0.3(X_1+1.5)^2 + 0.2Z_2$, $X_5 \sim \text{Ber}(\exp(X_1)/\{1+\exp(X_1)\})$, and $X_6 \sim \text{Ber}(\exp(X_2)/\{1+\exp(X_2)\})$. Here Z_1 and Z_2 are two Uniform$[-1,1]$ random variables independent of X_1 and X_2. Finally, we let

$$Y = \sin(2X^\top\theta_0) + 2\exp(X^\top\theta_0) + \epsilon,$$

where θ_0 is $(1.3,-1.3,1,-0.5,-0.5,-0.5)/\sqrt{5.13}$. In the following, we consider three different scenarios based on different error distributions:

(2.1) $\epsilon \sim N(0,1)$, (Homoscedastic, Gaussian Error)
(2.2) $\epsilon|X \sim N\left(0, \log(2+(X^\top\theta_0)^2)\right)$, (Heteroscedastic, Gaussian Error)
(2.3) $\epsilon|\xi \sim (-1)^\xi\text{Beta}(2,3)$, where $\xi \sim \text{Ber}(.5)$. (Homoscedastic, Non-Gaussian Error)

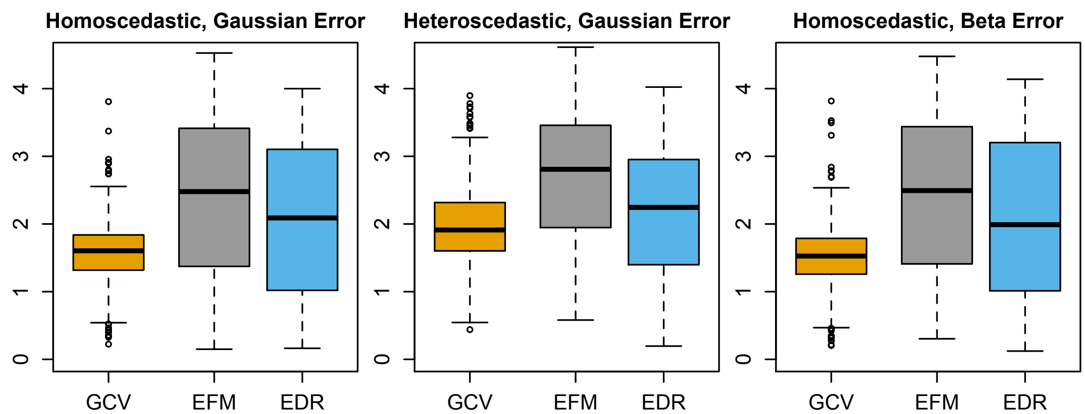

FIG 2. *Box plots (over 500 replications) of L_1 error of estimates of θ_0 ($\sum_{i=1}^6 |\hat\theta_i - \theta_{0,i}|$) based on 200 observations from models (2.1), (2.2), and (2.3) in the left, the middle, and the right panels, respectively.*

Observe that in all the three scenarios the proposed estimator has improved performance compared to the competitors; see Figure 2. The relative poor performance of EDR and EFM can possibly be attributed to the dependency between covariates. Scenarios **(2.1)** and **(2.2)** are similar to simulation scenarios considered in Ma and Zhu (2013) and Li and Patilea (2015). The codes to compute the estimator proposed in Li and Patilea (2015) were not available to us.

TABLE 1
Median (and interquartile range) of $\sum_{i=1}^d |\hat\theta_i - \theta_{0,i}|/d$ from 500 replications for $n = 400$ from (32).

d	$a=\pi/2$			$a=3\pi/4$			$a=3\pi/2$		
	GCV	EFM	EDR	GCV	EFM	EDR	GCV	EFM	EDR
10	0.005	0.006	0.007	0.004	0.279	0.010	0.321	0.326	0.317
	(0.002)	(0.003)	(0.003)	(0.198)	(0.323)	(0.222)	(0.068)	(0.064)	(0.068)
50	0.007	0.134	0.122	0.133	0.134	0.134	0.134	0.136	0.135
	(0.115)	(0.129)	(0.017)	(0.010)	(0.008)	(0.008)	(0.008)	(0.007)	(0.009)
100	0.088	0.092	0.088	0.091	0.092	0.091	0.091	0.092	0.091
	(0.009)	(0.003)	(0.005)	(0.003)	(0.003)	(0.004)	(0.004)	(0.004)	(0.003)

5.3. High Dimensional Covariates

For the final simulation scenario, we consider a setting similar to that of Example 4 in Section 3.2 of Cui et al. (2011). We consider d-variate covariates for $d = 10, 50$, and 100. For each d, we assume that $X \sim \text{Uniform}[0, 5]^d$, $\epsilon \sim N(0, 0.2^2)$, $\theta_0 = (2, 1, \mathbf{0}_{d-2})^\top/\sqrt{5}$, and have 400 observations from the following model:

$$Y = \sin(aX^\top\theta_0) + \epsilon, \text{ where } a = \pi/2, 3\pi/4, \text{ and } 3\pi/2. \tag{32}$$

Note that here a higher value of a represents a more oscillating link function. Table 1 summarizes the finite sample performance of the estimators considered in this paper. Observe that the proposed estimator has the best overall performance, whereas EFM and EDR fail in certain scenarios and it is hard to predict their performance for a particular setting, e.g., both EFM and EDR fail when $a = \pi/2$, $d = 50$ and EFM fails when $a = 3\pi/4$, $d = 10$. This can possibly be attributed to the sensitivity of the procedures towards the multiple tuning parameters involved.

6. Real data analysis

6.1. Car mileage data

In this sub-section, we model the mileages (Y) of 392 cars using the covariates (X): displacement (D), weight (W), acceleration (A), and horsepower (H); see `http://lib.stat.cmu.edu/datasets/cars.data` for the data set. For our data analysis, we have scaled and centered each of covariates to have mean 0 and variance 1. To compare the prediction capabilities of the linear model to that of the single index model for this data set, we randomly split the data set into a training set of size 260 and a test set of size 132 and compute the prediction error for both the linear model fit and the single index model fit. The average prediction error over 1000 such random splits was 4.3 for the linear model fit and 3.8 for the single index model fit. The results indicate that the single index model is a better fit.

In the left panel of Figure 3, we have the scatter plot of $\{(\hat{\theta}^\top x_i, y_i)\}_{i=1}^{392}$ overlaid with the plot of $\hat{m}(\hat{\theta}^\top x)$. In Table 2, we display the estimates of θ_0 based on the methods considered in the paper. The MAVE, the EFM estimator, and the PLSE give similar estimates while the EDR gives a different estimate of the index parameter.

TABLE 2
Estimates of θ_0 for the data sets in Sections 6.1 and 6.2.

Method	Car mileage data				Ozone data		
	D	W	A	H	R	W	T
GCV	0.48	0.17	0.17	0.84	0.32	-0.62	0.71
EFM	0.44	0.18	0.13	0.87	0.29	-0.60	0.75
EDR	0.33	0.11	0.15	0.93	0.22	-0.64	0.73
rMAVE	0.48	0.17	0.17	0.84	0.31	-0.58	0.75

6.2. Ozone concentration data

For the second real data example, we study the relationship between ozone concentration (Y) and three meteorological variables (X): radiation level (R), wind speed (W), and temperature

(T). The data consists of 111 days of complete measurements from May to September, 1973, in New York city. The data set can be found in the `EnvStats` package in R. Yu and Ruppert (2002) fit a linear model, an additive model, and a fully nonparametric model and conclude that the single index model fits the data best. To fit a single index model to the data Yu and Ruppert (2002) fix 10 knots and fit cubic penalized splines to the data. The right panel of Figure 3 shows the scatter plot of $\hat{\theta}^\top X$ and Y overlaid with the plot of $\hat{m}(\hat{\theta}^\top X)$. As in the previous example, we have scaled and centered each of the covariates such that they have mean 0 and variance 1. We see that all the considered methods in the paper give similar estimates for θ_0; see Table 2.

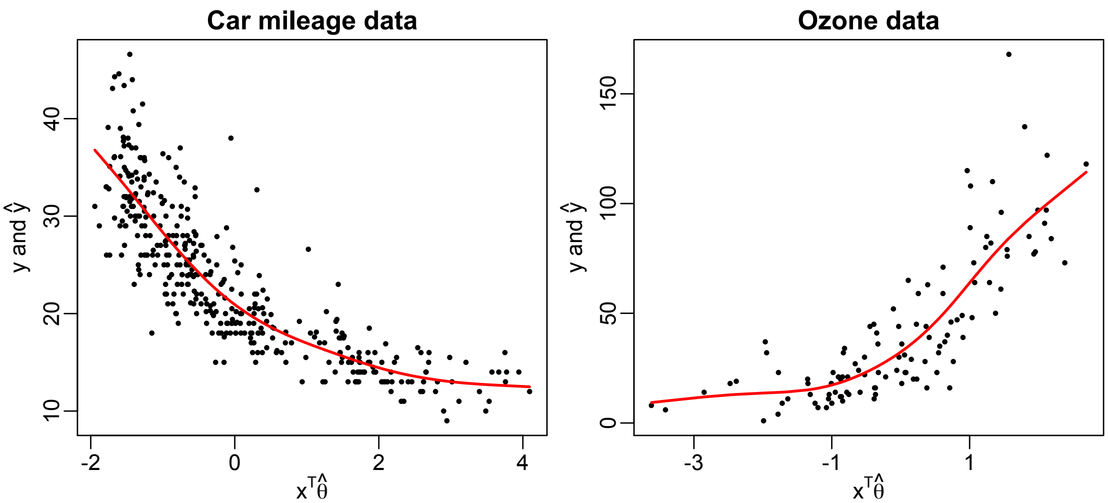

Fig 3. *Scatter plots of $\{(x_i^\top \hat{\theta}, y_i)\}_{i=1}^n$ overlaid with the plots of $\hat{m}$ (in solid red line) for the two real data sets considered. Left panel: the car mileage data (Section 6.1); right panel: ozone concentration data (Section 6.2).*

7. Summary

In this paper we propose a simple penalized least squares based estimator $(\hat{m}, \hat{\theta})$ for the unknown link function, m_0, and the index parameter, θ_0, in the single index model under mild smoothness assumptions on m_0. We prove that $\hat{m}$ is rate optimal (for the given smoothness) and $\hat{\theta}$ is $\sqrt{n}$-consistent and asymptotically normal. Moreover under homoscedastic errors, we show that $\hat{\theta}$ (properly normalized) has the optimal variance in the sense of Bickel et al. (1993). In contrast to existing procedures, our method involves only one tuning parameter. We have developed the R package `simest` to compute the proposed estimators. We observe that the PLSE has superior finite sample performance compared to most competing methods.

Acknowledgments

We would like to thank Bodhisattva Sen for many helpful discussions and for his help in writing this paper. We would also like to thank Promit Ghosal and David A. Hirshberg for helpful discussions.

8. Proof of results in Section 2

We start with two useful lemmas concerning the properties of functions in $\mathcal{S}$.

Lemma 4. *Let $m \in \{g \in \mathcal{S} : J(g) < \infty\}$. Then $|m'(s) - m'(s_0)| \le J(m)|s - s_0|^{1/2}$ for every $s, s_0 \in D$.*

Proof. The proof follows from a simple application of the Cauchy-Schwarz inequality:

$$|m'(s) - m'(s_0)| \quad = \quad \left|\int_{s_0}^{s} m''(t)dt\right| \le \left|\int_{s_0}^{s} \left|m''(t)\right|^2 dt\right|^{1/2} |s - s_0|^{1/2} \le J(m)|s - s_0|^{1/2}, \qquad \forall s, s_0 \in D.$$

□

Lemma 5. *Let $m \in \{g \in \mathcal{S} : J(g) < \infty$ and $\|g\|_\infty \le M\}$, where M is a finite constant. Then*

$$\|m'\|_\infty \le 2M/\varnothing(D) + (1 + J(m))\varnothing(D)^{1/2},$$

where $\varnothing(D)$ is the diameter of D. Moreover if $\varnothing(D) < \infty$, then

$$\|m'\|_\infty \le C(1 + J(m)),$$

where C is a finite constant depending only on M and $\varnothing(D)$.

Proof. Fix $s_0 \in D$. Integrating the inequality

$$-J(m)|t - s_0|^{1/2} \le m'(t) - m'(s_0) \le J(m)|t - s_0|^{1/2}$$

with respect to t, we get

$$|m(s) - m(s_0) - m'(s_0)(s - s_0)| \le J(m)\varnothing(D)^{3/2},$$

where $\varnothing(D)$ is the diameter of D. Since $\|m\|_\infty \le M$, we get that

$$|m'(s_0)(s - s_0)| \le 2M + J(m)\varnothing(D)^{3/2}.$$

If we choose s such that $|s - s_0| = \varnothing(D)/2$, then we have

$$\|m'\|_\infty \le 2M/\varnothing(D) + (1 + J(m))\varnothing(D)^{1/2}.$$

The rest of the lemma follows by choosing $C = 2M/\varnothing(D) + \varnothing(D)^{1/2}$. □

8.1. Proof of Theorem 1

The minimization problem considered is

$$\inf_{\theta \in \Theta, m \in \mathcal{S}} \mathcal{L}_n(m, \theta; \lambda),$$

where $\mathcal{L}_n$ is defined in (3). For any fixed vector $\theta \in \Theta$, define $t_i^\theta := \theta^\top x_i$, for $i = 1, \ldots, n$. Then we have

$$\mathcal{L}_n(m, \theta; \lambda) = \left[\frac{1}{n}\sum_{i=1}^{n} \left(y_i - m(t_i^\theta)\right)^2 + \lambda^2 \int_D \left|m''(t)\right|^2 dt\right]$$

and the minimization can be equivalently written as $\inf_{\theta\in\Theta}\inf_{m\in\mathcal{S}}\mathcal{L}_n(m,\theta;\lambda)$. Let us define

$$T(\theta) := \inf_{m\in\mathcal{S}}\mathcal{L}_n(m,\theta;\lambda) \quad \text{and} \quad m_\theta := \arg\min_{m\in\mathcal{S}}\mathcal{L}_n(m,\theta;\lambda). \tag{33}$$

Theorem 2.4 of Green and Silverman (1994) proves that the infimum in (33) is attained for every $\theta\in\Theta$ and the unique minimizer m_θ is a natural cubic spline with knots at $\{t_i^\theta\}_{i=1}^n$. Furthermore Green and Silverman (1994) note that (see Section 2.3.4), m_θ does not depend on D beyond the condition that $\{t_i^\theta\}_{1\le i\le n}\in D$. Moreover, m_θ'' is zero outside $(t_{(1)}^\theta, t_{(n)}^\theta)$, where for $k=1,\ldots,n$, $t_{(k)}^\theta$ denotes the k-th smallest value in $\{t_i^\theta\}_{i=1}^n$.

For every $\theta\in\Theta$, m_θ is determined by points in a bounded set, namely $D_R := [-t_{\max}, t_{\max}]$, where $t_{\max}$ a finite constant such that $\sup_{\theta\in\Theta}\max_{i\le n}|\theta^\top x_i|\le t_{\max}$. Note that such a constant always exists as $\Theta\subset S^{d-1}$. Define

$$\mathcal{S}_R := \{m : D_R\to\mathbb{R}|\, m' \text{ is absolutely continuous}\},$$

and for all $m\in\mathcal{S}_R$, define $J_R^2(m) := \int_{D_R}|m''(t)|^2dt$. For every $m\in\mathcal{S}_R$ and $\theta\in\Theta$, we define

$$\mathcal{L}_n^R(m,\theta;\lambda) = \left[\frac{1}{n}\sum_{i=1}^n\left(y_i - m(t_i^\theta)\right)^2 + \lambda^2\int_{D_R}\left|m''(t)\right|^2dt\right],$$

$$T_R(\theta) := \inf_{m\in\mathcal{S}_R}\mathcal{L}_n^R(m,\theta;\lambda), \quad \text{and} \quad m_\theta^R := \arg\min_{m\in\mathcal{S}_R}\mathcal{L}_n^R(m,\theta;\lambda).$$

Green and Silverman (1994) observe that (see Section 2.3.4), m_θ is the linear extrapolation of m_θ^R to D. Moreover, as m_θ is a linear function outside D_R, we have

$$\int_{D_R}\left|(m_\theta^R)''(t)\right|^2dt = \int_D\left|m_\theta''(t)\right|^2dt \quad \text{and} \quad T_R(\theta) = T(\theta).$$

Thus we have

$$\inf_{\theta\in\Theta, m\in\mathcal{S}}\mathcal{L}_n(m,\theta;\lambda) = \inf_{\theta\in\Theta}T(\theta) = \inf_{\theta\in\Theta}T_R(\theta) = \inf_{\theta\in\Theta, m\in\mathcal{S}}\mathcal{L}_n^R(m,\theta;\lambda).$$

As Θ is a compact set, the existence of the minimizer of $\theta\mapsto T_R(\theta)$ will be established if we can show that $T_R(\theta)$ is a continuous function on Θ; see the Weierstrass extreme value theorem. We now prove that $\theta\mapsto T_R(\theta)$ is a continuous function. Notice that $\sup_{\theta\in\Theta}T_R(\theta)\le \sup_{\theta\in\Theta}\mathcal{L}_n^R(0,\theta;\lambda) = \sum_{i=1}^n y_i^2/n < \infty$. Hence there is a finite constant K (depending only on $\{y_i\}_{i=1}^n$) such that for all $\theta\in\Theta$,

$$Q_n(m_\theta^R,\theta) + \lambda^2 J_R^2(m_\theta^R)\le K. \tag{34}$$

We will use the above bound to show that there exists a finite L (depending only on λ and $\{(y_i,x_i)\}_{i=1}^n$) such that $\|m_\theta^R\|_\infty\le L$ and $J_R(m_\theta^R)\le L$ for all $\theta\in\Theta$. By (34), we have that

$$J_R^2(m_\theta^R)\le K/\lambda^2 \quad \text{and} \quad |m_\theta^R(t_{(i)}^\theta)|\le\sqrt{nK}+\max_{i\le n}|y_i|, \tag{35}$$

for $i=1,\ldots,n$. If $t_{(1)}^\theta = t_{(n)}^\theta$, then it is easy to see that $m_\theta^R(\cdot)\equiv\sum_{i=1}^n y_i/n$ which implies that $\|m_\theta^R\|_\infty$ is bounded and $J_R(m_\theta^R)=0$. Now let us assume $t_{(1)}^\theta < t_{(n)}^\theta$. By Lemma 5, for any $s\in\mathbb{R}$ such that $|s|\le t_{\max}$, we have

$$\left|(m_\theta^R)'(s) - (m_\theta^R)'(t_{(1)}^\theta)\right|\le J_R(m_\theta^R)\sqrt{t_{\max}}.$$

Integrating the above display with respect to s, we get

$$\left|m_\theta^R(s) - m_\theta^R(t_{(1)}^\theta) - (m_\theta^R)'(t_{(1)}^\theta)(s - t_{(1)}^\theta)\right| \le J_R(m_\theta^R)(t_{\max})^{3/2}. \tag{36}$$

Taking $s = t_{(n)}^\theta$ in the previous display, we have $|(m_\theta^R)'(t_{(1)}^\theta)| \le C$, where the constant C depends only on K, λ, and $\{(x_i, y_i)\}_{i=1}^n$ (see (35)). In view of the bound on $|(m_\theta^R)'(t_{(1)}^\theta)|$, (36) implies that

$$\sup_{|s|\le t_{\max}} |m_\theta^R(s)| \le C_1,$$

where the constant C_1 depends only on K, λ, and $\{(y_i, x_i)\}_{i=1}^n$. Thus, there exists a finite L (depending only on λ and $\{(y_i, x_i)\}_{i=1}^n$) such that $\|m_\theta^R\|_\infty \le L$ and $J_R(m_\theta^R) \le L$. Note that L does not depend on θ. As $\|m_\theta^R\|_\infty \le L$ and $J_R(m_\theta^R) \le L$, we can redefine $T_R(\theta)$ as

$$T_R(\theta) = \inf_{m \in \{m \in \mathcal{S}_R : \|m\|_\infty \le L \text{ and } J_R(m) \le L\}} \left[Q_n(m, \theta) + \lambda^2 \int_{D_R} \left|m''(t)\right|^2 dt\right].$$

We will now show that the class of functions

$$\{Q_n(m, \cdot) : \Theta \to \mathbb{R} |\, m \in \mathcal{S}_R, \|m\|_\infty \le L, \text{ and } J_R(m) \le L\}$$

is uniformly equicontinuous, i.e., for every $\varepsilon > 0$, there exists a $\delta > 0$ such that $|\theta - \eta| \le \delta$ implies that

$$\sup_{m \in \{m \in \mathcal{S}_R : \|m\|_\infty \le L \text{ and } J_R(m) \le L\}} |Q_n(m, \theta) - Q_n(m, \eta)| \le \varepsilon.$$

Note that

$$\begin{aligned}
&|Q_n(m, \theta) - Q_n(m, \eta)| \\
&= \frac{1}{n}\left|\sum_{i=1}^n \left[(y_i - m(\theta^\top x_i))^2 - (y_i - m(\eta^\top x_i))^2\right]\right| \\
&= \frac{1}{n}\left|\sum_{i=1}^n \left[(m(\eta^\top x_i) - m(\theta^\top x_i))^2 + 2(y_i - m(\eta^\top x_i))(m(\eta^\top x_i) - m(\theta^\top x_i))\right]\right| \\
&\le \max_{1\le i\le n} |m(\eta^\top x_i) - m(\theta^\top x_i)|^2 + \frac{2}{n}\max_{1\le i\le n} |m(\eta^\top x_i) - m(\theta^\top x_i)| \sum_{i=1}^n |y_i - m(\eta^\top x_i)|.
\end{aligned} \tag{37}$$

In view of Lemma 5, for $i = 1, \ldots, n$ we have

$$|m(\theta^\top x_i) - m(\eta^\top x_i)| \le \|m'\|_\infty |x_i^\top(\theta - \eta)| \le C_2(1 + J_R(m))|\theta - \eta|, \tag{38}$$

where C_2 is a constant that depends only on L and $\max_{1\le i\le n} |x_i|$. For every $m \in \{m \in \mathcal{S}_R : \|m\|_\infty \le L \text{ and } J_R(m) \le L\}$, (37) and (38) imply that

$$\sup_{m \in \{m \in \mathcal{S}_R : \|m\|_\infty \le L \text{ and } J_R(m) \le L\}} |Q_n(m, \theta) - Q_n(m, \eta)| \le C_3|\theta - \eta|,$$

where the constant C_3 depends only on L and $\max_{1\le i\le n} |x_i|$. Observe that for every $\theta \in \Theta$, $m_\theta^R \in \{m \in \mathcal{S}_R : \|m\|_\infty \le L \text{ and } J_R(m) \le L\}$. Fix $\delta = \varepsilon/C_3$, then uniform equicontinuity of $\{\theta \mapsto Q_n(m, \theta) : m \in \mathcal{S}_R, \|m\|_\infty \le L, \text{ and } J_R(m) \le L\}$ implies that, for all $|\eta - \theta| \le \delta$, we have

$$Q_n(m_\eta^R, \theta) - \varepsilon \le Q_n(m_\eta^R, \eta) \quad \text{and} \quad Q_n(m_\theta^R, \eta) \le Q_n(m_\theta^R, \theta) + \varepsilon. \tag{39}$$

Recall that for every $\beta \in \Theta$ and $m \in \{m \in \mathcal{S}_R : J_R(m) < \infty\}$, we have $\mathcal{L}_n^R(m_\beta^R, \beta; \lambda) \le \mathcal{L}_n^R(m, \beta; \lambda)$. Thus, from (39), we have

$$\begin{aligned} Q_n(m_\eta^R, \theta) - \varepsilon \le Q_n(m_\eta^R, \eta) \;&\Leftrightarrow\; \mathcal{L}_n^R(m_\eta^R, \theta; \lambda) - \varepsilon \le \mathcal{L}_n^R(m_\eta^R, \eta; \lambda) \\ \Rightarrow\; \mathcal{L}_n^R(m_\theta^R, \theta; \lambda) - \varepsilon &\le \mathcal{L}_n^R(m_\eta^R, \eta; \lambda) \;\Rightarrow\; T_R(\theta) - \varepsilon \le T_R(\eta) \end{aligned} \tag{40}$$

and

$$\begin{aligned} Q_n(m_\theta^R, \eta) \le Q_n(m_\theta^R, \theta) + \varepsilon \;&\Leftrightarrow\; \mathcal{L}_n^R(m_\theta^R, \eta; \lambda) \le \mathcal{L}_n^R(m_\theta^R, \theta; \lambda) + \varepsilon \\ \Rightarrow\; \mathcal{L}_n^R(m_\eta^R, \eta; \lambda) &\le \mathcal{L}_n^R(m_\theta^R, \theta; \lambda) + \varepsilon \;\Rightarrow\; T_R(\eta) \le T_R(\theta) + \varepsilon. \end{aligned} \tag{41}$$

Combining (40) and (41), we have that $T_R(\theta) - \varepsilon \le T_R(\eta) \le T_R(\theta) + \varepsilon$, for all $|\eta - \theta| \le \delta$. Thus, it follows that $\theta \mapsto T_R(\theta)$ is uniformly continuous and $T_R(\theta)$ attains a minimum on the compact set Θ (S^{d-1} is compact and Θ is closed subset of S^{d-1}). Thus

$$\hat{\theta} = \arg\min_{\theta \in \Theta} T_R(\theta) = \arg\min_{\theta \in \Theta} T(\theta)$$

is well defined. Moreover by Theorem 2.4 of Green and Silverman (1994) we have that $m_{\hat{\theta}}^R$ is a unique natural cubic spline with knots at $\{t_i^{\hat{\theta}}\}_{i=1}^n$ and

$$\hat{m} = m_{\hat{\theta}},$$

where $m_{\hat{\theta}}$ is the linear extrapolation of $m_{\hat{\theta}}^R$ to D.

9. Proofs of results in Section 3

9.1. Proof of Theorem 2

Since $(\hat{m}, \hat{\theta})$ minimizes $Q_n(m, \theta) + \hat{\lambda}_n^2 J^2(m)$, we have

$$Q_n(\hat{m}, \hat{\theta}) + \hat{\lambda}_n^2 J^2(\hat{m}) \le Q_n(m_0, \theta_0) + \hat{\lambda}_n^2 J^2(m_0). \tag{42}$$

Observe that by definition of $Q_n(m, \theta)$, we have that (42) implies

$$\begin{aligned} \|\hat{m} \circ \hat{\theta} - m_0 \circ \theta_0\|_n^2 + \hat{\lambda}_n^2 J^2(\hat{m}) &\le \frac{2}{n} \sum_{i=1}^n (y_i - m_0(\theta_0^\top x_i))(\hat{m}(\hat{\theta}^\top x_i) - m_0(\theta_0^\top x_i)) + \hat{\lambda}_n^2 J^2(m_0) \\ &= \frac{2}{n} \sum_{i=1}^n \epsilon_i (\hat{m}(\hat{\theta}^\top x_i) - m_0(\theta_0^\top x_i)) + \hat{\lambda}_n^2 J^2(m_0) \end{aligned}$$

To find the rate of convergence of $\|\hat{m} \circ \hat{\theta} - m_0 \circ \theta_0\|_n$ we will try to find upper bounds for $\sum_{i=1}^n \epsilon_i(\hat{m}(\hat{\theta}^\top x_i) - m_0(\theta_0^\top x_i))$ in terms of $\|\hat{m} \circ \hat{\theta} - m_0 \circ \theta_0\|_n$ (modulus of continuity); see Section 1 of van de Geer (1990) for a similar proof technique. To be able to find such a bound, we first study the behavior of $\hat{m} \circ \hat{\theta}$.

Observe that by Cauchy-Schwarz inequality we have

$$\begin{aligned} & Q_n(m_0, \theta_0) - Q_n(\hat{m}, \hat{\theta}) \\ &= \frac{2}{n} \sum_{i=1}^n \epsilon_i (\hat{m}(\hat{\theta}^\top x_i) - m_0(\theta_0^\top x_i)) - \frac{1}{n} \sum_{i=1}^n (\hat{m}(\hat{\theta}^\top x_i) - m_0(\theta_0^\top x_i))^2 \\ &\le \left(\frac{4}{n} \sum_{i=1}^n \epsilon_i^2 \right)^{1/2} \|\hat{m} \circ \hat{\theta} - m_0 \circ \theta_0\|_n - \|\hat{m} \circ \hat{\theta} - m_0 \circ \theta_0\|_n^2. \end{aligned} \tag{43}$$

Note that by **(A3)**, $(1/n)\sum_{i=1}^{n}\epsilon_i^2 = O(1)$ almost surely. On the other hand, since $(\hat{m},\hat{\theta})$ minimizes $Q_n(m,\theta)+\hat{\lambda}_n^2 J^2(m)$, we have

$$Q_n(m_0,\theta_0) - Q_n(\hat{m},\hat{\theta}) \geq \hat{\lambda}_n^2(J^2(\hat{m}) - J^2(m_0)) \geq -\hat{\lambda}_n^2 J^2(m_0) \geq o_p(1), \tag{44}$$

as $\hat{\lambda}_n = o_p(1)$. Combining (43) and (44), we have

$$\|\hat{m}\circ\hat{\theta} - m_0\circ\theta_0\|_n^2 \leq \|\hat{m}\circ\hat{\theta} - m_0\circ\theta_0\|_n O_p(1) + o_p(1).$$

Thus we have $\|\hat{m}\circ\hat{\theta} - m_0\circ\theta_0\|_n = O_p(1)$. We also have $\|\hat{m}\circ\hat{\theta}\|_n = O_p(1)$ as $\|m_0\circ\theta_0\|_\infty < \infty$.

We will now use the Sobolev embedding theorem to get a bound on $\|\hat{m}\|_\infty$ in terms of $J(\hat{m})$.

Lemma 6. *(Sobolev embedding theorem, Page 85, Oden and Reddy (2012)) Let $m : I \to \mathbb{R}$ ($I \subset \mathbb{R}$ is an interval) be a function such that $J(m) < \infty$. We can write*

$$m(t) = m_1(t) + m_2(t),$$

with $m_1(t) = \beta_1 + \beta_2 t$ and $\|m_2\|_\infty \leq J(m)\varnothing(I)$.

Thus, by the above lemma, we can find functions $\hat{m}_1$ and $\hat{m}_2$ such that

$$\hat{m}(t) = \hat{m}_1(t) + \hat{m}_2(t),$$

where $\hat{m}_1 = \hat{\beta}_1 + \hat{\beta}_2 t$, and $\|\hat{m}_2\|_\infty \leq J(\hat{m})\varnothing(D)$. Then

$$\begin{aligned}\frac{\|\hat{m}_1\circ\hat{\theta}\|_n}{1+J(m_0)+J(\hat{m})} &\leq \frac{\|\hat{m}\circ\hat{\theta}\|_n}{1+J(m_0)+J(\hat{m})} + \frac{\|\hat{m}_2\circ\hat{\theta}\|_n}{1+J(m_0)+J(\hat{m})} \\ &\leq \frac{\|\hat{m}\circ\hat{\theta}\|_n}{1+J(m_0)+J(\hat{m})} + \frac{\|\hat{m}_2\|_\infty}{1+J(m_0)+J(\hat{m})} = O_p(1).\end{aligned} \tag{45}$$

Let us define

$$\mathbb{A}_n(\theta) := \frac{1}{n}\sum_{i=1}^{n}\varphi_\theta(X_i)\varphi_\theta^\top(X_i) \qquad \text{and} \qquad A(\theta) := \int \varphi_\theta(x)\varphi_\theta(x)^\top dP_X(x),$$

where $\varphi_\theta(x) := (1,\theta^\top x)^\top$. Furthermore, we denote the smallest eigenvalues of $\mathbb{A}_n(\theta)$ and $A(\theta)$ by $\vartheta_n(\theta)$ and $\vartheta(\theta)$ respectively. Since Θ is a bounded subset of $\mathbb{R}^d$, by the Glivenko-Cantelli Theorem, we have

$$\sup_{\theta\in\Theta}|\vartheta_n(\theta) - \vartheta(\theta)| = o_p(1).$$

Let $\vartheta_0 := \min_{\theta\in\Theta}\vartheta(\theta)$. By assumption **(A5)** and and the fact that $|\theta| = 1$, we have $\det(A(\theta)) = \theta^\top \mathrm{Var}(X)\theta$ and $\inf_{\theta\in\Theta}\det(A(\theta)) > 0$. It follows that $\vartheta_0 > 0$ and

$$\begin{aligned}\|\hat{m}_1\circ\hat{\theta}\|_n^2 &= (\hat{\beta}_1,\hat{\beta}_2)\mathbb{A}_n(\theta)(\hat{\beta}_1,\hat{\beta}_2)^\top \\ &\geq \vartheta_n(\hat{\theta})(\hat{\beta}_1^2+\hat{\beta}_2^2) \\ &= \left[\vartheta_n(\hat{\theta}) - \vartheta(\hat{\theta})\right](\hat{\beta}_1^2+\hat{\beta}_2^2) + \vartheta(\hat{\theta})(\hat{\beta}_1^2+\hat{\beta}_2^2) \\ &\geq o_p(\hat{\beta}_1^2+\hat{\beta}_2^2) + \vartheta_0(\hat{\beta}_1^2+\hat{\beta}_2^2) \\ &\geq o_p(\hat{\beta}_1^2+\hat{\beta}_2^2) + \vartheta_0\max(\hat{\beta}_1,\hat{\beta}_2)^2\end{aligned}$$

Thus by (45) we have

$$\frac{\max(\hat{\beta}_1, \hat{\beta}_2)}{1 + J(m_0) + J(\hat{m})} = O_p(1). \tag{46}$$

Moreover, since D is a bounded set, by (46) we have $\|\hat{m}_1\|_\infty/(1 + J(m_0) + J(\hat{m})) = O_p(1)$. Combining this with Lemma 6, we get

$$\frac{\|\hat{m}\|_\infty}{1 + J(m_0) + J(\hat{m})} \le \frac{\|\hat{m}_1\|_\infty}{1 + J(m_0) + J(\hat{m})} + \frac{\|\hat{m}_2\|_\infty}{1 + J(m_0) + J(\hat{m})} = O_p(1). \tag{47}$$

Now define the class of functions

$$\mathcal{B}_C := \left\{ \frac{m \circ \theta - m_0 \circ \theta_0}{1 + J(m_0) + J(m)} : m \in \mathcal{S},\ \theta \in \Theta,\ \text{and}\ \frac{\|m\|_\infty}{1 + J(m_0) + J(m)} \le C \right\}.$$

Observe that by (47), we can find a C_ε such that

$$\mathbb{P}\left(\frac{\hat{m} \circ \hat{\theta} - m_0 \circ \theta_0}{1 + J(m_0) + J(\hat{m})} \in \mathcal{B}_{C_\varepsilon} \right) \ge 1 - \varepsilon, \qquad \forall n. \tag{48}$$

The following lemma in van de Geer (2000) gives a upper bound for $\sum_{i=1}^n \epsilon_i g(x_i)$, in terms of entropy of the class of functions g.

Lemma 7. *(Lemma 8.4, van de Geer (2000)) Suppose $\mathcal{G}$ be a class of functions. If $\log N_{[\,]}(\delta, \mathcal{G}, \|\cdot\|_\infty) \le A\delta^{-\alpha}$, $\sup_{g\in\mathcal{G}} \|g\|_n \le R$, and ϵ satisfies assumption **(A3)**, for some constants $0 < \alpha < 2$, A, and R. Then for some constant c, we have for all $T \ge c$,*

$$\mathbb{P}\left(\sup_{g\in\mathcal{G}} \frac{|\frac{1}{\sqrt{n}}\sum_{i=1}^n \epsilon_i g(x_i)|}{\|g\|_n^{1-\frac{\alpha}{2}}} \ge T \right) \le c \exp\left[\frac{-T^2}{c^2} \right]$$

Lemma 8, proved in Section 9.2 of the supplementary material, finds the bracketing number for the class of functions $\mathcal{B}_C$.

Lemma 8. *For every fixed positive M_1, M_2, and C, we have*

$$\log N\left(\delta, \mathcal{B}_C, \|\cdot\|_\infty\right) \lesssim \delta^{-1/2}.$$

In the view of (48), Lemmas 7 and 8 allow us to conclude

$$\frac{(1/n)\sum_{i=1}^n \epsilon_i (\hat{m}(\hat{\theta}^\top x_i) - m_0(\theta_0^\top x_i))}{\|\hat{m} \circ \hat{\theta} - m_0 \circ \theta_0\|_n^{3/4} (1 + J(m_0) + J(\hat{m}))^{1/4}} = O_p(n^{-1/2}). \tag{49}$$

Together, (44) and (49) imply

$$\begin{aligned}
&\hat{\lambda}_n^2 (J^2(\hat{m}) - J^2(m_0)) \\
&\le Q_n(m_0, \theta_0) - Q_n(\hat{m}, \hat{\theta}) \\
&= \frac{2}{n} \sum_{i=1}^n (y_i - m_0(\theta_0^\top x_i))(\hat{m}(\hat{\theta}^\top x_i) - m_0(\theta_0^\top x_i)) - \|\hat{m} \circ \hat{\theta} - m_0 \circ \theta_0\|_n^2 \\
&\le \|\hat{m} \circ \hat{\theta} - m_0 \circ \theta_0\|_n^{3/4} (1 + J(m_0) + J(\hat{m}))^{1/4} O_p(n^{-1/2}) - \|\hat{m} \circ \hat{\theta} - m_0 \circ \theta_0\|_n^2.
\end{aligned} \tag{50}$$

We will now consider two cases.

Case 1: Suppose $J(\hat{m}) > 1 + J(m_0)$. By (50), we have

$$\|\hat{m} \circ \hat{\theta} - m_0 \circ \theta_0\|_n^2 + \hat{\lambda}_n^2 J^2(\hat{m}) \leq \|\hat{m} \circ \hat{\theta} - m_0 \circ \theta_0\|_n^{3/4} J(\hat{m})^{1/4} O_p(n^{-1/2}) + \hat{\lambda}_n^2 J^2(m_0).$$

Moreover note that we can find constants C_1 and C_2 such that either

$$\|\hat{m} \circ \hat{\theta} - m_0 \circ \theta_0\|_n^{3/4} J(\hat{m})^{1/4} n^{-1/2} \leq C_1 \hat{\lambda}_n^2 J^2(m_0) \tag{51}$$

or

$$\hat{\lambda}_n^2 J^2(m_0) < C_2 \|\hat{m} \circ \hat{\theta} - m_0 \circ \theta_0\|_n^{3/4} J(\hat{m})^{1/4} n^{-1/2} \tag{52}$$

hold with high probability as $n \to \infty$. Observe that when (51) holds we have

$$\|\hat{m} \circ \hat{\theta} - m_0 \circ \theta_0\|_n^2 + \hat{\lambda}_n^2 J^2(\hat{m}) \leq O_p(1) \hat{\lambda}_n^2 J^2(m_0). \tag{53}$$

Now it is easy to see that, (53) implies that $\|\hat{m} \circ \hat{\theta} - m_0 \circ \theta_0\|_n = O_p(\hat{\lambda}_n) J(m_0)$ and $J(\hat{m}) = O_p(1) J(m_0)$. On the other hand when (52) holds, we have

$$\|\hat{m} \circ \hat{\theta} - m_0 \circ \theta_0\|_n^2 + \hat{\lambda}_n^2 J^2(\hat{m}) \leq \|\hat{m} \circ \hat{\theta} - m_0 \circ \theta_0\|_n^{3/4} J(\hat{m})^{1/4} O_p(n^{-1/2}). \tag{54}$$

We can bound the first term on the left hand side of (54) as

$$\|\hat{m} \circ \hat{\theta} - m_0 \circ \theta_0\|_n \leq \left[J(\hat{m})^{1/4} O_p(n^{-1/2}) \right]^{4/5}. \tag{55}$$

A similar bound on the second term on the left hand side of (54) gives:

$$\begin{aligned}
\hat{\lambda}_n^2 J^2(\hat{m}) &\leq \|\hat{m} \circ \hat{\theta} - m_0 \circ \theta_0\|_n^{3/4} J(\hat{m})^{1/4} O_p(n^{-1/2}) \\
&\leq \left[J(\hat{m})^{1/4} O_p(n^{-1/2}) \right]^{3/5} J(\hat{m})^{1/4} O_p(n^{-1/2}) \text{ (by (55))} \\
&\leq J(\hat{m})^{2/5} \left[O_p(n^{-1/2}) \right]^{8/5},
\end{aligned}$$

which implies that

$$J(\hat{m}) = O_p(n^{-1/2}) \hat{\lambda}_n^{-5/4}. \tag{56}$$

Combining (55) and (56), we have

$$\|\hat{m} \circ \hat{\theta} - m_0 \circ \theta_0\|_n = O_p(n^{-1/2}) \hat{\lambda}_n^{-1/4}.$$

However, by assumption **(A3)**, we have that $\hat{\lambda}_n^{-1} = O_p(n^{2/5})$. Hence the conclusion follows.

Case 2: When $J(\hat{m}) \leq 1 + J(m_0)$, (50) implies,

$$\|\hat{m} \circ \hat{\theta} - m_0 \circ \theta_0\|_n^2 \leq \|\hat{m} \circ \hat{\theta} - m_0 \circ \theta_0\|_n^{3/4} (1 + J(m_0))^{1/4} O_p(n^{-1/2}) + \hat{\lambda}_n^2 J^2(m_0).$$

Therefore, it follows that either

$$\|\hat{m} \circ \hat{\theta} - m_0 \circ \theta_0\|_n \leq (1 + J(m_0))^{1/5} O_p(n^{-2/5}) = O_p(\hat{\lambda}_n)$$

or

$$\|\hat{m} \circ \hat{\theta} - m_0 \circ \theta_0\|_n \leq O_p(1) \hat{\lambda}_n J(m_0) = O_p(\hat{\lambda}_n) J(m_0).$$

Thus we have that $J(\hat{m}) = O_p(1)$, $\|\hat{m} \circ \hat{\theta} - m_0 \circ \theta_0\|_n = O_p(\hat{\lambda}_n)$, and, by (47), $\|\hat{m}\|_\infty = O_p(1)$. To find the rates of convergence of $\|\hat{m} \circ \hat{\theta} - m_0 \circ \theta_0\|$, we use the following lemma.

Lemma 9. *(Lemma 5.16, van de Geer (2000)) Suppose $\mathcal{G}$ is a class of uniformly bounded functions and for some $0 < \nu < 2$,*

$$\sup_{\delta>0} \delta^{\nu} \log N_{[\,]}(\delta, \mathcal{G}, \|\cdot\|_{\infty}) < \infty.$$

Then for every given $\alpha > 0$ there exists a constant $C > 0$ such that

$$\limsup_{n\to\infty} \mathbb{P}\left(\sup_{g\in\mathcal{G}, \|g\|>Cn^{-1/(2+\nu)}} \left| \frac{\|g\|_n}{\|g\|} - 1 \right| > \alpha \right) = 0.$$

Our proof of Theorem 2 is along the lines of the proofs of Lemma 3.1 in Mammen and van de Geer (1997) and Theorem 10.2 in van de Geer (2000).

9.2. Proof of Lemma 8

To prove this lemma, we use the following entropy bound from van de Geer (2000). We will also use the following result in the proofs of Lemmas 2 and 3 in Sections 10.5 and 10.6, respectively.

Lemma 10. *(Theorem 2.4, van de Geer (2000)) Let $\mathcal{F}$ be a class of functions $f : I \to \mathbb{R}$ (for I a compact interval in $\mathbb{R}$) such that for some $M_1, M_2 < \infty$, $\|f\|_{\infty} \le M_1$, the first $k-1$ derivatives are absolutely continuous and $\int_I [f^{(k)}(x)]^2 dx \le M_2^2$. Then there exists a constant C depending only on I such that,*

$$\log N_{[\,]}(\varepsilon, \mathcal{F}, \|\cdot\|_{\infty}) \le C\left(\frac{M_1 + M_2}{\varepsilon}\right)^{1/k}, \quad \textit{for all } \varepsilon > 0.$$

The above lemma says that the class of functions

$$\mathcal{G}_{M_1,M_2} := \{m \in \mathcal{S} : \ \|m\|_{\infty} \le M_1, \ \text{and } J(m) \le M_2\}$$

can be covered by $\exp(C\sqrt{M_1 + M_2}\delta^{-1/2})$ balls with radius δ in the sup-norm, i.e.,

$$\log N_{[\,]}(\delta, \mathcal{G}_{M_1,M_2}, \|\cdot\|_{\infty}) \le C\left(\frac{M_1 + M_2}{\delta}\right)^{1/2}.$$

For all $\theta_1, \theta_2 \in \Theta$, we have that $|\theta_1 - \theta_2| \le 2$. Thus by Lemma 4.1 of Pollard (1990), we have

$$N(\varepsilon, \Theta, |\cdot|) \lesssim \varepsilon^{-d+1}.$$

Now define the class of functions

$$\mathcal{H}_{M_1,M_2} := \{m(\theta^{\top}x) : \ \theta \in \Theta, \ m \in \mathcal{S}, \ \|m\|_{\infty} \le M_1, \ \text{and } J(m) \le M_2\}.$$

We will show that

$$\log N_{[\,]}(\varepsilon, \mathcal{H}_{M_1,M_2}, \|\cdot\|_{\infty}) \lesssim \left(\frac{M_1 + M_2}{\varepsilon}\right)^{1/2}. \tag{57}$$

Note that, with respect to $\|\cdot\|_{\infty}$-norm covering number and bracketing number are the same and we can choose an ε-net from within the function class. Thus $\|\cdot\|_{\infty}$ brackets can be chosen from the function class.

Consider an $\varepsilon/[2(1+M_2)T]$-net of Θ, $\{\theta_1, \theta_2, \ldots, \theta_p\}$, $\chi \subset B_{\mathbf{0}}(T) \subset \mathbb{R}^d$, the Euclidean ball of radius T around the origin. Choose an $\varepsilon/2$-net for $\mathcal{G}_{M_1,M_2}$, $\{m_1, m_2, \ldots, m_q\}$. We can, without loss of generality, assume that $m_i \in \mathcal{G}_{M_1,M_2}$. Thus by Lemma 5, we have $\|m_i'\|_\infty \lesssim 1 + M_2$.

Now we will show that the set of functions $\{m_i \circ \theta_j\}_{1\le i\le q, 1\le j\le p}$ form an ε-net for $\mathcal{H}_{M_1,M_2}$ with respect to $\|\cdot\|_\infty$-norm. For any given $m \circ \theta \in \mathcal{H}_{M_1,M_2}$, we can get m_i and θ_j such that $\|m - m_i\|_\infty < \varepsilon/2$ and $|\theta - \theta_j| < \varepsilon/2(1+M_2)T$. Then

$$
\begin{aligned}
&|m(\theta^\top x) - m_i(\theta_j^\top x)| \\
&\quad \le |m(\theta^\top x) - m(\theta_j^\top x)| + |m(\theta_j^\top x) - m_i(\theta_j^\top x)| \\
&\quad \le \|m'\|_\infty |x| |\theta - \theta_j| + \|m - m_i\|_\infty \le \frac{(1+M_2)|x|\varepsilon}{2T(1+M_2)} + \frac{\varepsilon}{2} \le \varepsilon.
\end{aligned}
$$

Hence, the bracketing entropy number in the $\|\cdot\|_\infty$-norm for the required set is bounded above by a multiple of $(M/\varepsilon)^{1/2} + \log(C2T(1+M_2)\varepsilon^{-d+1})$ for a suitable constant $C > 0$, which is further bounded by a multiple of $(M/\varepsilon)^{1/2}$, where $M = M_1 + M_2$. Thus we have (57).

Now we will use (57) to prove Lemma 8. Let us define,

$$
\mathcal{F}_C := \left\{ f(\theta^\top x) : f = \frac{m}{1 + J(m_0) + J(m)},\ \theta \in \Theta,\ m \in \mathcal{S},\ \text{and}\ \frac{\|m\|_\infty}{1 + J(m_0) + J(m)} \le C \right\}
$$

Since $\mathcal{F}_C \subset \mathcal{H}_{C,1}$, we can choose $\delta/2$ brackets $[g_{1,1}, g_{1,2}], \ldots, [g_{q,1}, g_{q,2}]$ over $\mathcal{F}_C$ such that for every $f(\theta^\top x) \in \mathcal{F}_C$ there exists a i such that $g_{i,1}(x) \le f(\theta^\top x) \le g_{i,2}(x)$. Let us now define,

$$
\mathcal{F}^* := \left\{ h : h = \frac{m_0}{1 + J(m_0) + J(m)} \text{ and } m \in \mathcal{S} \right\}.
$$

Observe that $\mathcal{F}^* \subset \mathcal{G}_{C_1,1}$, where $C_1 = \|m_0\|_\infty / J(m_0)$. Thus we can choose $\delta/2$ brackets $[l_{1,1}, l_{1,2}], \ldots, [l_{r,1}, l_{r,2}]$ over $\mathcal{F}^*$ such that for every $h \in \mathcal{F}^*$ there exists a j such that $l_{j,1}(\theta_0^\top x) \le h(\theta_0^\top x) \le l_{j,2}(\theta_0^\top x)$. Thus we have,

$$
g_{i,1}(x) - l_{j,2}(\theta_0^\top x) \le \frac{m(\theta^\top x)}{1 + J(m_0) + J(m)} - \frac{m_0(\theta_0^\top x)}{1 + J(m_0) + J(m)} \le g_{i,2}(x) - l_{j,1}(\theta_0^\top x),
$$

where i depends on (m, θ) and j on m.

Brackets of the form $[g_{i,1}(x) - l_{j,2}(\theta_0^\top x), g_{i,2}(x) - l_{j,1}(\theta_0^\top x)]$ for $i \in \{1, \ldots q\}$ and $j \in \{1, \ldots r\}$ cover the required space. Hence, the bracketing entropy satisfies

$$
\log N\left(\delta, \mathcal{B}_C,\ \|\cdot\|_\infty\right) \le \frac{(C+1)^{\frac{1}{2}} + (C_1+1)^{\frac{1}{2}}}{\delta^{\frac{1}{2}}},
$$

where $C_1 = \|m_0\|_\infty / J(m_0)$.

9.3. Proof of Theorem 3

The following lemma is crucial to the proof of Theorem 3.

Lemma 11. *For every fixed M, the set of functions $m \in \mathcal{S}$ with $J(m) \le M$ and $\|m\|_\infty \le M$ is precompact relative to $\|\cdot\|_D^S$.*

Proof. Let us define, $\mathcal{D}_M := \{m \in \mathcal{S} : \|m\|_\infty \le M, \text{and}\, J(m) \le M\}$. By Lemma 4 the class of functions $\{m' : m \in \mathcal{D}_M\}$ is uniformly Lipschitz of order $1/2$. Thus any sequence of functions $\{m'_k : m_k \in \mathcal{D}_M\}$ is equicontinuous. By Lemma 5, $\{m'_k\}$ is uniformly bounded. Applying the Arzela-Ascoli theorem, we see that every sequence $\{m_k\}$ has a subsequence $\{m_{k_l}\}$ such that $\{m'_{k_l}\}$ converges uniformly on D. Since $\{m'_{k_l}\}$ is uniformly bounded, we have that $\{m_{k_l}\}$ is equicontinuous. Therefore as $\|m_{k_l}\|_\infty \le M$, by Arzela-Ascoli theorem, there exists a subsequence $\{k_{l_j}\}$ of $\{k_l\}$ such that $\{m_{k_{l_j}}\}$ converge uniformly on D. Since these functions converge uniformly on a compact set, by applying the dominated convergence theorem, we see that there exists a subsequence such that functions and derivatives converge. Furthermore, the derivative of the limit equals the limit of the derivative. □

Suppose that $\|m_k \circ \theta_k - m_0 \circ \theta_0\| \to 0$, $\|m_k\|_\infty = O(1)$, and $J(m_k) = O(1)$. By Lemma 11, every subsequence of (m_k, θ_k) has a further subsequence (m_{k_l}, θ_{k_l}) such that $\theta_{k_l} \to \theta$ and $\|m_{k_l} - m\|_D^S \to 0$ for some θ and m. Then $\|m_{k_l} \circ \theta_{k_l} - m \circ \theta\| \to 0$ by continuity of the map $(m, \theta) \mapsto m \circ \theta$. Thus $\|m \circ \theta - m_0 \circ \theta_0\| = 0$, and hence by assumption **(A0)**, we get $\theta = \theta_0$ and $m = m_0$ on the support D_0. The assumption that D_0 is the closure of its interior implies that m' and m'_0 agree on D_0. Since the convergence in Lemma 11 is uniform, we get that $\|m - m_0\|_{D_0} = 0$. Combining this with Theorem 2, we get that $\hat{\theta} \xrightarrow{P} \theta_0$ and $\|\hat{m} - m_0\|_{D_0}^S \xrightarrow{P} 0$.

Let a be a point in D_0 and $s \in D$. By Lemma 4, we have that $|\hat{m}'(s) - \hat{m}'(a)| \le J(\hat{m})|s - a|^{1/2} = O_p(1)$. Moreover, we have that $|\hat{m}'(a) - m'_0(a)| = o_p(1)$. Thus $\|\hat{m}'\|_\infty = O_p(1)$.

9.4. Proof of Theorem 4

We first state and prove a lemma that we will use to prove Theorem 4.

Lemma 12. *Suppose $m \in \mathcal{S}$, $J(m) < \infty$, and $\theta \in \Theta$. Then*

$$
\begin{aligned}
P_X\big|m(\theta^\top X) - m(\theta_0^\top X) - m_0'(\theta_0^\top X)X^\top(\theta - \theta_0)\big|^2& \\
\lesssim |\theta_0 - \theta|^3 J^2(m) + |\theta - \theta_0|^2 P_X&\big|(m - m_0)'(\theta_0^\top X)\big|^2.
\end{aligned}
$$

Proof. By th mean value theorem, we have

$$
\begin{aligned}
m(\theta^\top x) - m(\theta_0^\top x) - m_0'(\theta_0^\top x)x^\top(\theta - \theta_0) &= m'(\xi^\top x)x^\top(\theta - \theta_0) - m_0'(\theta_0^\top x)x^\top(\theta - \theta_0) \\
&= \{m'(\xi^\top x) - m_0'(\theta_0^\top x)\}x^\top(\theta - \theta_0),
\end{aligned}
$$

where $\xi^\top x$ lies between $\theta^\top x$ and $\theta_0^\top x$. Since χ is bounded (see **(A2)**), by an application of the Cauchy-Schwarz inequality, we have

$$
\begin{aligned}
\big|m(\theta^\top x) - m(\theta_0^\top X) - m_0'(\theta_0^\top x)x^\top(\theta - \theta_0)\big|^2 &\lesssim |\theta - \theta_0|^2\big|m'(\xi^\top x) - m_0'(\theta^\top x)\big|^2 \\
&\lesssim |\theta - \theta_0|^2\big|m'(\xi^\top x) - m'(\theta_0^\top x)\big|^2 \\
&\quad + |\theta - \theta_0|^2\big|m'(\theta_0^\top x) - m_0'(\theta_0^\top x)\big|^2.
\end{aligned}
$$

By Lemma 4, we have

$$
\begin{aligned}
|m'(\xi^\top x) - m'(\theta_0^\top x)| \le J(m)|\xi^\top x - \theta_0^\top x|^{1/2} &\le J(m)|\theta^\top x - \theta_0^\top x|^{1/2} \\
&\lesssim J(m)|\theta - \theta_0|^{1/2}.
\end{aligned}
$$

Thus we have

$$\begin{aligned}&\left|m(\theta^\top x) - m_0(\theta_0^\top x) - m_0'(\theta_0^\top x)x^\top(\theta - \theta_0)\right|^2\\&\qquad\lesssim \left|m'(\theta_0^\top x) - m_0'(\theta_0^\top x)\right|^2|\theta - \theta_0|^2 + J^2(m)|\theta - \theta_0|^3,\end{aligned}$$

and hence

$$\begin{aligned}&P_X\left|m(\theta^\top X) - m(\theta_0^\top X) - m_0'(\theta^\top X)X^\top(\theta - \theta_0)\right|^2\\&\qquad\lesssim |\theta - \theta_0|^2 P_X\left|(m - m_0)'(\theta_0^\top X)\right|^2 + J^2(m)|\theta_0 - \theta|^3. \qquad \square\end{aligned}$$

Let us define $A(x) := \hat{m}(\hat{\theta}^\top x) - m_0(\theta_0^\top x)$ and $B(x) := m_0'(\theta_0^\top x)x^\top(\hat{\theta} - \theta_0) + (\hat{m} - m_0)(\theta_0^\top x)$. Observe that

$$A(x) - B(x) = \hat{m}(\hat{\theta}^\top x) - m_0'(\theta_0^\top x)x^\top(\hat{\theta} - \theta_0) - \hat{m}(\theta_0^\top x).$$

Recall that $|\hat{\theta} - \theta_0| \xrightarrow{P} 0$, $P_X\left|(\hat{m} - m_0)'(\theta_0^\top X)\right|^2 \xrightarrow{P} 0$ and $J(\hat{m}) = O_p(1)$. Thus by Lemma 12, we have that

$$P_X|A(X) - B(X)|^2 \lesssim |\hat{\theta} - \theta_0|^3 J^2(\hat{m}) + |\hat{\theta} - \theta_0|^2 P_X|(\hat{m}' - m_0')(\theta_0^\top X)|^2 = o_p(1)|\hat{\theta} - \theta_0|^2.$$

and

$$P_X|A(X)|^2 \geq \frac{1}{2}P_X|B(X)|^2 - P_X|A(X) - B(X)|^2 \geq \frac{1}{2}P_X|B(X)|^2 - o_p(1)|\hat{\theta} - \theta_0|^2.$$

However by Theorem 2, we have that $P_X|A(X)|^2 = O_p(\hat{\lambda}_n^2)$. Thus we have

$$P_X\left|m_0'(\theta_0^\top X)X^\top(\hat{\theta} - \theta_0) + (\hat{m} - m_0)(\theta_0^\top X)\right|^2 \leq O_p(\hat{\lambda}_n^2) + o_p(1)|\hat{\theta} - \theta_0|^2.$$

Now define

$$g_1(x) := m_0'(\theta_0^\top x)x^\top(\hat{\theta} - \theta_0) \text{ and } \quad g_2(x) := (\hat{m} - m_0)(\theta_0^\top x) \tag{58}$$

and note that by assumption **(A6)** there exists a $\lambda_1 > 0$ such that

$$P_X g_1^2 = (\hat{\theta} - \theta_0)^\top P_X[XX^\top|m_0'(\theta_0^\top X)|^2](\hat{\theta} - \theta_0) \geq \lambda_1(\hat{\theta} - \theta_0)^\top(\hat{\theta} - \theta_0) = \lambda_1|\hat{\theta} - \theta_0|^2. \tag{59}$$

With (59) in mind, we can see that proof of this theorem will be complete if we can show that

$$P_X g_1^2 + P_X g_2^2 \lesssim P_X\left|m_0'(\theta_0^\top X)X^\top(\hat{\theta} - \theta_0) + (\hat{m} - m_0)(\theta_0^\top X)\right|^2. \tag{60}$$

The following theorem gives a sufficient condition for (60) to hold.

Lemma 13. *(Lemma 5.7 of Murphy et al. (1999)) Let g_1 and g_2 be measurable functions such that $|P_X(g_1 g_2)|^2 \leq c P_X g_1^2 P_X g_2^2$ for a constant $c < 1$. Then*

$$P_X(g_1 + g_2)^2 \geq (1 - \sqrt{c})(P_X g_1^2 + P_X g_2^2).$$

The following arguments show that g_1 and g_2 (defined in (58)) satisfy the condition of Lemma

13. Observe that

$$
\begin{aligned}
&P_X[m_0'(\theta_0^\top X)g_2(X)X^\top(\hat\theta-\theta_0)]^2\\
&\qquad\qquad= P_X\left|m_0'(\theta_0^\top X)g_2(X)E(X^\top(\hat\theta-\theta_0)|\theta_0^\top X)\right|^2\\
&\qquad\qquad\le P_X\left[\{m_0'(\theta_0^\top X)\}^2E^2[X^\top(\hat\theta-\theta_0)|\theta_0^\top X]\right]P_Xg_2^2(X)\\
&\qquad\qquad< P_X\left[\{m_0'(\theta_0^\top X)\}^2E[\{X^\top(\hat\theta-\theta_0)\}^2|\theta_0^\top X]\right]P_Xg_2^2(X)\\
&\qquad\qquad= P_X\left[\mathbb{E}[\{m_0'(\theta_0^\top X)X^\top(\hat\theta-\theta_0)\}^2|\theta_0^\top X)]\right]P_Xg_2^2(X)\\
&\qquad\qquad= P_X[m_0'(\theta_0^\top X)X^\top(\hat\theta-\theta_0)]^2P_Xg_2^2(X)\\
&\qquad\qquad= P_Xg_1^2\,P_Xg_2^2.
\end{aligned}
$$

Strict inequality in the above sequence of inequalities holds under the assumption that the conditional distribution of X given $\theta_0^\top X$ is non-degenerate.

10. Proofs of results in Section 4

10.1. Proof of Lemma 1

For every $\theta\in S^{d-1}$ and $\theta\neq\theta_0$, define

$$\theta_d := \frac{\theta_0-\theta}{|\theta_0-\theta|}\quad\text{and}\quad \theta_p := \frac{\theta_0-\theta\theta_0^\top\theta}{|\theta_0-\theta\theta_0^\top\theta|} \tag{61}$$

Observe that $\theta^\top\theta_p=0$ and $\theta_p\in\operatorname{span}\{\theta_0,\theta\}$, where for $a_1,\ldots,a_k\in\mathbb{R}^d$, $\operatorname{span}\{a_1,\ldots,a_k\}$ denotes the linear span of $a_1,\ldots,a_k$. Consider the following symmetric matrices in $\mathbb{R}^{d\times d}$:

$$T_\theta^d := \mathbb{I}_d-2\theta_d\theta_d^\top\quad\text{and}\quad T_\theta^p := \mathbb{I}_d-2\theta_p\theta_p^\top. \tag{62}$$

Note that for every $x\in\mathbb{R}^d$, $x\mapsto T_\theta^dx$ and $x\mapsto T_\theta^px$ define the reflections about the hyperplanes through 0 which are orthogonal to θ_d and θ_p, respectively. More generally, for any $a\in S^{d-1}$, $T_a:=\mathbb{I}_d-2aa^\top$ is known as the Householder transformation or elementary reflector matrix; see Page 324 of Meyer (2000). It is easy to see that T_a is an orthogonal matrix for every $a\in S^{d-1}$ and $\det(T_a)=-1$. As $|\theta_0|=|\theta|=1$, we have

$$1=\theta_d^\top\theta_d=\frac{1}{|\theta_0-\theta|^2}(\theta_0-\theta)^\top(\theta_0-\theta)=\frac{1}{|\theta_0-\theta|^2}\left[2\theta_0^\top\theta_0-2\theta^\top\theta_0\right]=\frac{2}{|\theta_0-\theta|}\theta_d^\top\theta_0.$$

Thus

$$T_\theta^d\theta_0=\theta_0-2\theta_d\theta_d^\top\theta_0=\theta_0-\theta_d|\theta_0-\theta|=\theta$$

and as $\theta_p^\top\theta=0$, we have $T_\theta^p\theta=\theta$. Now, let $\{e_1,\ldots,e_d\}$ be an orthonormal basis of $\mathbb{R}^d$ such that $e_1=\theta_0$. Define

$$H_{\theta_0}:=[e_2,\ldots,e_d]\text{ and }H_\theta:=T_\theta^pT_\theta^dH_{\theta_0},\ \forall\theta\neq\theta_0. \tag{63}$$

As $T_\theta^p T_\theta^d$ is an orthogonal matrix, it is easy to see that H_{θ_0} and H_θ satisfy conditions (a) and (b). Now we will prove that $\|H_\theta - H_{\theta_0}\|_2 \le |\theta_0 - \theta|$. Observe that

$$\begin{aligned}\|H_\theta - H_{\theta_0}\|_2 &= \sup_{\eta\in S^{d-2}} |H_\theta\eta - H_{\theta_0}\eta| \\ &= \sup_{\eta\in S^{d-2}} |T_\theta^p T_\theta^d H_{\theta_0}\eta - H_{\theta_0}\eta| \\ &= \sup_{x^\top\theta_0=0,\, x\in S^{d-1}} |T_\theta^p T_\theta^d x - x| \\ &\le \sup_{x\in S^{d-1}} |T_\theta^p T_\theta^d x - \mathbb{I}_d x| = \|T_\theta^p T_\theta^d - \mathbb{I}_d\|_2.\end{aligned}$$

We will now show that $\|T_\theta^p T_\theta^d - \mathbb{I}_d\|_2 = |\theta_0 - \theta|$. The following argument shows that $T_\theta^p T_\theta^d$ is essentially a rotation operator on $\mathrm{span}\{\theta, \theta_0\}$ that fixes $\mathrm{span}\{\theta, \theta_0\}^\perp$. Fix $\theta \in \Theta$. Observe that for any orthogonal matrix Q, we have

$$\|T_\theta^p T_\theta^d - \mathbb{I}_d\|_2 = \|Q^\top(T_\theta^p T_\theta^d - \mathbb{I}_d)Q\|_2 = \|Q^\top T_\theta^p T_\theta^d Q - \mathbb{I}_d\|_2. \tag{64}$$

We will try to compute the right hand side of the above display by using a convenient choice of Q. Consider any orthogonal matrix Q such that θ and θ_p are the first two columns of Q. Such a Q exists as $\theta \perp \theta_p$ and $|\theta| = |\theta_p| = 1$. By (62) and the fact that $\theta_d \in \mathrm{span}\{\theta, \theta_p\}$, we have

$$Q^\top T_\theta^p T_\theta^d Q = \mathbb{I}_d - 2Q^\top\left[\theta_d\theta_d^\top + \theta_p\theta_p^\top - 2\theta_d\theta_d^\top\theta_p\theta_p^\top\right]Q = \begin{bmatrix} A_\theta & \mathbf{0}_{2\times(d-2)} \\ \mathbf{0}_{(d-2)\times 2} & \mathbb{I}_{(d-2)}\end{bmatrix}, \tag{65}$$

where $A_\theta \in \mathbb{R}^{2\times 2}$. As $Q^\top T_\theta^p T_\theta^d Q$ is an orthogonal matrix and $\det(Q^\top T_\theta^p T_\theta^d Q) = 1$, A_θ is an orthogonal matrix and $\det(A_\theta) = 1$, i.e., A_θ is a rotation matrix for $\mathbb{R}^2$. Note that by (65), we have

$$Q^\top T_\theta^p T_\theta^d Q x - x = A_\theta \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} - \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \quad \text{where } x := (x_1, x_2, \ldots, x_d)^\top \in \mathbb{R}^d. \tag{66}$$

Thus

$$\sup_{x\in S^{d-1}} \left|Q^\top T_\theta^p T_\theta^d Q x - x\right| = \sup_{x\in S^{d-1}} \left| A_\theta \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} - \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}\right| = \sup_{y\in S^1} |A_\theta y - y|.$$

However, as A_θ is a rotation matrix and in two dimension rotation is completely determined by a angle of rotation, we have that

$$\sup_{y\in S^1} |A_\theta y - y| = |A_\theta z - z| \tag{67}$$

for all $z \in S^1$; see Page 326, Meyer (2000). Let $z^0 := (z_1^0, z_2^0)^\top \in S^1$ be such that $\theta_0 = z_1^0\theta + z_2^0\theta_p$. Define $x^0 := (z_1^0, z_2^0, 0, \ldots, 0)^\top \in S^{d-1}$. By (66), we have

$$|A_\theta z^0 - z^0| = |Q^\top T_\theta^p T_\theta^d Q x^0 - Q^\top Q x^0| = |Q^\top(\theta - \theta_0)| = |\theta_0 - \theta|, \tag{68}$$

where the second equality is true due the following observation: as $Qx^0 = z_1^0\theta + z_2^0\theta_p = \theta_0$ and $T_\theta^p T_\theta^d \theta_0 = \theta$, we have $T_\theta^p T_\theta^d Q x^0 = \theta$. The last equality in the above display is true as Q is an orthogonal matrix. Thus combining (64), (66), (67), and (68), we have $\|T_\theta^p T_\theta^d - \mathbb{I}_d\|_2 = |\theta_0 - \theta|$.

Before proving (d), we show that for $x \in \mathbb{R}^{d-1}, |H_\theta x| = |x|$ and for $y \in \mathbb{R}^d, |H_\theta^\top y| \le |y|$. Recall that $T_\theta^p T_\theta^d$ is an orthogonal matrix. For $x \in \mathbb{R}^{d-1}$ observe that $|H_\theta x| = |H_{\theta_0} x| = |\sum_{i=1}^{d-1} x_i e_{i+1}|$, where $e_1, \ldots, e_d$ is defined in (63). As $e_1, \ldots, e_d$ form an orthonormal set, we have that $|H_\theta x| = \sqrt{\sum_{i=1}^{d-1} x_i^2} = |x|$. Recall that $T_\theta^p T_\theta^d$ is an orthogonal matrix. Thus to prove $|H_\theta^\top y| \le |y|$, it is enough to show that $|H_{\theta_0}^\top y| \le |y|$. Let $y \in R^d$, then $y = \sum_{i=1}^d (e_i^\top y) e_i$. Observe that $H_{\theta_0}^\top y = \sum_{j=2}^d \sum_{i=1}^d e_i^\top y e_j^\top e_i$. As $e_1, \ldots, e_d$ form an orthonormal set, we have $e_j^\top e_i = 0$ for all $j \ne i$ and $e_i^\top e_i = 1$. Thus $|H_{\theta_0}^\top y| = \sqrt{\sum_{j=2}^d (e_j^\top y)^2} \le \sqrt{\sum_{j=1}^d (e_j^\top y)^2} = |y|$.

Now we verify that $\{H_\theta : \theta \in \Theta\}$ defined in (63) satisfies condition (d) of Lemma 1. Let $\eta, \beta \in \Theta \setminus \theta_0$ such that $|\eta - \theta_0| < 1/2$, $|\beta - \theta_0| < 1/2$. Note that

$$
\begin{aligned}
\|H_\eta^\top - H_\beta^\top\|_2 &= \|H_{\theta_0}^\top \left[T_\eta^d T_\eta^p - T_\beta^d T_\beta^p\right]\|_2 \\
&= \sup_{x \in S^{d-1}} \left|H_{\theta_0}^\top \left[T_\eta^d T_\eta^p - T_\beta^d T_\beta^p\right] x\right| \\
&\le \sup_{x \in S^{d-1}} \left|\left(T_\eta^d T_\eta^p - T_\beta^d T_\beta^p\right) x\right| \\
&\le \sup_{x \in S^{d-1}} \left|\left(T_\eta^d T_\eta^p - T_\eta^d T_\beta^p\right) x\right| + \sup_{x \in S^{d-1}} \left|\left(T_\eta^d T_\beta^p - T_\beta^d T_\beta^p\right) x\right| \\
&= \sup_{x \in S^{d-1}} \left|T_\eta^d \left(T_\eta^p - T_\beta^p\right) x\right| + \sup_{x \in S^{d-1}} \left|\left(T_\eta^d - T_\beta^d\right) T_\beta^p x\right| \\
&= \sup_{x \in S^{d-1}} \left|\left(T_\eta^p - T_\beta^p\right) x\right| + \sup_{x \in S^{d-1}} \left|\left(T_\eta^d - T_\beta^d\right) x\right| \\
&= \|T_\eta^p - T_\beta^p\|_2 + \|T_\eta^d - T_\beta^d\|_2,
\end{aligned} \tag{69}
$$

here the first inequality is true as $|H_{\theta_0}^\top x| \le |x|$ for all $x \in \mathbb{R}^d$ and the penultimate equality is true as both T_η^d and T_β^p are orthogonal matrices in $\mathbb{R}^{d \times d}$. We will next show that

$$
\|T_\eta^d - T_\beta^d\|_2 \le 4|\eta_d - \beta_d| \quad \text{and} \quad \|T_\eta^p - T_\beta^p\|_2 \le 4|\eta_p - \beta_p|, \tag{70}
$$

where η_p, η_d, β_p, and β_d are defined as in (61). Observe that

$$
\begin{aligned}
\|T_\eta^d - T_\beta^d\|_2 &= 2\|\beta_d \beta_d^\top - \eta_d \eta_d^\top\|_2 \\
&\le 2\|\beta_d \beta_d^\top - \beta_d \eta_d^\top\|_2 + 2\|\beta_d \eta_d^\top - \eta_d \eta_d^\top\|_2 \\
&= 2\|\beta_d (\beta_d^\top - \eta_d^\top)\|_2 + 2\|(\beta_d - \eta_d)\eta_d^\top\|_2 \\
&= 2 \sup_{x \in S^{d-1}} |\beta_d (\beta_d^\top - \eta_d^\top) x| + 2|\beta_d - \eta_d| \sup_{x \in S^{d-1}} |\eta_d^\top x| \\
&= 2 \sup_{x \in S^{d-1}} |(\beta_d^\top - \eta_d^\top) x| + 2|\beta_d - \eta_d| \\
&= 4|\beta_d - \eta_d|.
\end{aligned}
$$

A similar calculation will show the second equality in (70). The proof of (7) will be complete if we can show that

$$
|\eta_d - \beta_d| \le 2\frac{|\eta - \beta|}{|\eta - \theta_0| + |\beta - \theta_0|} \quad \text{and} \quad |\eta_p - \beta_p| \le \frac{16|\eta - \beta|/\sqrt{15}}{|\eta - \theta_0| + |\beta - \theta_0|}. \tag{71}
$$

Observe that by properties of projection onto the unit sphere (see Lemma 3.1 of Kalaj et al. (2016)), we have

$$
|\eta_d - \beta_d| = \left|\frac{\eta - \theta_0}{|\eta - \theta_0|} - \frac{\beta - \theta_0}{|\beta - \theta_0|}\right| \le \frac{2|\eta - \beta|}{|\eta - \theta_0| + |\beta - \theta_0|}.
$$

and

$$|\eta_p - \beta_p| = \left|\frac{\theta_0 - \eta\theta_0^\top\eta}{|\theta_0 - \eta\theta_0^\top\eta|} - \frac{\theta_0 - \beta\theta_0^\top\beta}{|\theta_0 - \beta\theta_0^\top\beta|}\right| \le \frac{2|\eta\theta_0^\top\eta - \beta\theta_0^\top\beta|}{|\theta_0 - \eta\theta_0^\top\eta| + |\theta_0 - \beta\theta_0^\top\beta|}. \tag{72}$$

We now try to simplify (72). First note that $|\eta - \theta_0| \le 1/2$ implies that $1 + \theta_0^\top\eta \ge 15/8$. Now observe that

$$\begin{aligned}|\theta_0 - \eta\theta_0^\top\eta|^2 &= 1 - (\theta_0^\top\eta)^2 = (1-\theta_0^\top\eta)(1+\theta_0^\top\eta)\\ &= \frac{|\eta-\theta_0|^2}{2}(1+\theta_0^\top\eta) \ge \frac{|\eta-\theta_0|^2}{2}\inf_{\eta\in\Theta}(1+\theta_0^\top\eta) \ge \frac{15}{16}|\eta-\theta_0|^2.\end{aligned}$$

For the numerator of (72), we have

$$|\eta\theta_0^\top\eta - \beta\theta_0^\top\beta| \le |\eta\theta_0^\top\eta - \eta\theta_0^\top\beta| + |\eta\theta_0^\top\beta - \beta\theta_0^\top\beta| \le 2|\eta-\beta|.$$

Combining the above two displays, we have

$$|\eta_p - \beta_p| \le \frac{4|\eta-\beta|}{\sqrt{\frac{15}{16}}(|\eta-\theta_0| + |\beta-\theta_0|)} \le \frac{16|\eta-\beta|/\sqrt{15}}{|\eta-\theta_0|+|\beta-\theta_0|}.$$

Combining (69), (70), and (71), we have that

$$\|H_\eta^\top - H_\beta^\top\|_2 \le (8 + 64/\sqrt{15})\frac{|\eta-\beta|}{|\eta-\theta_0|+|\beta-\theta_0|}$$

10.2. Proof of Theorem 6

We will first show that $\xi_t(u;\theta,\eta,m)$ is a valid submodel. Note that $\phi_{\theta,\eta,0}(u + (\theta-\theta)^\top h_\theta(u)) = u$, $\forall u \in D$. Hence,

$$\xi_\theta(\theta^\top x;\theta,\eta,m) = m\circ\phi_{\theta,\eta,0}(\theta^\top x) = m(\theta^\top x).$$

Now we will prove that $J^2(\xi_t(\cdot;\theta,\eta,m)) < \infty$. Let us define

$$\psi_{\theta,\eta,t}(u) := \phi_{\theta,\eta,t}(u + (\theta - \zeta_t(\theta,\eta))^\top h_\theta(u)),$$

then $\xi_t(u;\theta,\eta,m) = m\circ\psi_{\theta,\eta,t}(u)$ Observe that

$$\begin{aligned}J^2(\xi_t(\cdot;\theta,\eta,m)) &= \int_D \left|\xi_t''(u;\theta,\eta,m)\right|^2 du\\ &= \int_D \left[m''\circ\psi_{\theta,\eta,t}(u)\psi_{\theta,\eta,t}'(u)^2 + m'\circ\psi_{\theta,\eta,t}(u)\psi_{\theta,\eta,t}''(u)\right]^2 du\\ &= \int_D \left[m''(u)(\psi_{\theta,\eta,t}'\circ\psi_{\theta,\eta,t}^{-1}(u))^2 + m'(u)\psi_{\theta,\eta,t}''\circ\psi_{\theta,\eta,t}^{-1}(u)\right]^2 \frac{du}{\psi_{\theta,\eta,t}'\circ\psi_{\theta,\eta,t}^{-1}(u)}\end{aligned}$$

where $\psi_{\theta,\eta,t}'(u) = \frac{\partial}{\partial u}\psi_{\theta,\eta,t}(u)$. Thus, we have that $J^2(\xi_t(\cdot;\theta,\eta,m)) = O(1)$ whenever $J(m) = O(1)$, $\|m\|_\infty = O(1)$, and t in a small neighborhood of 0 (as $\psi_{\theta,\eta,t}(\cdot)$ is a strictly increasing function when t is small). Next we evaluate $\partial\xi_t(\zeta_t(\theta,\eta)^\top x;\theta,\eta,m)/\partial t$ to help with the calculation of the score function for the submodel $\{\zeta_t(\theta,\eta), \xi_t(\cdot;\theta,\eta,m)\}$. Note that

$$\begin{aligned}
&\frac{\partial}{\partial t}\xi_t(\zeta_t(\theta,\eta)^\top x;\theta,\eta,m)\\
&= \frac{\partial}{\partial t} m\circ\phi_{\theta,\eta,t}\left(\zeta_t(\theta,\eta)^\top x + (\theta-\zeta_t(\theta,\eta))^\top h_\theta(\zeta_t(\theta,\eta)^\top x)\right)\\
&= m'\circ\phi_{\theta,\eta,t}\big(\zeta_t(\theta,\eta)^\top x + (\theta-\zeta_t(\theta,\eta))^\top h_\theta(\zeta_t(\theta,\eta)^\top x)\big)\\
&\qquad\Big[\dot{\phi}_{\theta,\eta,t}\big[\zeta_t(\theta,\eta)^\top x + [\theta-\zeta_t(\theta,\eta)]^\top h_\theta(\zeta_t(\theta,\eta)^\top x)\big]\\
&\qquad\quad + \phi'_{\theta,\eta,t}\big[\zeta_t(\theta,\eta)^\top + (\theta-\zeta_t(\theta,\eta))^\top h_\theta(\zeta_t(\theta,\eta)^\top x)\big]\frac{\partial\zeta_t(\theta,\eta)}{\partial t}^\top\Big[x\\
&\qquad\qquad\qquad + (\theta-\zeta_t(\theta,\eta))^\top h'_\theta(\zeta_t(\theta,\eta)^\top x)x - h_\theta(\zeta_t(\theta,\eta)^\top x)\Big]\Big],
\end{aligned}$$

where $\dot{\phi}_{t,\theta}(u) = \partial\phi_{\theta,\eta,t}(u)/\partial t$. We will now show that the score function of the submodel $\{t, \xi_t(\cdot;,\theta,\eta,m)\}$ is $\tilde{\ell}_{\theta,m}(y,x)$. Using the facts that $\phi'_{\theta,\eta,t}(u)=1$ and $\dot{\phi}_{\theta,\eta,t}(u)=0$ for all $u\in D$ (follows from the definition (25)) and $\partial\zeta_t(\theta,\eta)/\partial t = (-2t/\sqrt{1-t^2|\eta|^2})\,\theta + H_\theta\eta$, we get

$$\begin{aligned}
\frac{\partial}{\partial t}(y-\xi_t(\zeta_t(\theta,\eta)^\top x;\theta,\eta,m))^2\Big|_{t=0} &= -2(y-\xi_t(\zeta_t(\theta,\eta)^\top x;\theta,\eta,m))\,\frac{\partial\xi_t(\zeta_t(\theta,\eta)^\top x;\theta,\eta,m)}{\partial t}\Big|_{t=0}\\
&= -2(y-m(\theta^\top x))m'(\theta^\top x)\eta^\top H_\theta^\top(x-h_\theta(\theta^\top x))
\end{aligned}$$

Observe that $(\hat{m},\hat{\theta})$ minimizes the penalized loss function in (4) and $\xi_0(\zeta_0(\hat{\theta},\eta)^\top x;\hat{\theta},\eta,\hat{m}) = \hat{m}(\hat{\theta}^\top x)$, where $\zeta_t(\hat{\theta},\eta) = \sqrt{1-t^2|\eta|^2}\,\hat{\theta} + sH_{\hat{\theta}}\eta$. Hence, for every $\eta\in\mathbb{R}^{d-1}$, the function

$$t\mapsto \frac{1}{n}\sum_{i=1}^n\left(y_i - \xi_t(\zeta_t(\hat{\theta},\eta)^\top x;\hat{\theta},\eta,\hat{m})\right)^2 + \hat{\lambda}_n^2\int_D\left|\frac{\partial^2}{\partial u^2}\xi_t(u;\hat{\theta},\eta,\hat{m})\right|^2du \tag{73}$$

on a some small neighborhood of 0 (that depends on η) is minimized at $t=0$. Moreover, using some tedious algebra it can be shown that $J^2(\xi_t(\cdot;\theta,\eta,m))$ is differentiable and

$$\frac{\partial}{\partial t}J^2(\xi_t(\cdot;\theta,\eta,m))\Big|_{t=0} \lesssim \int_D |m''(p)|^2dp.$$

This we have that the function in (73) is differentiable at $t=0$. Conclude that, for all $\eta\in\mathbb{R}^{d-1}$ we have

$$\eta^\top\mathbb{P}_n\tilde{\ell}_{\hat{\theta},\hat{m}} - \hat{\lambda}_n^2\,\frac{\partial J^2(\xi_t(\cdot;\theta,\eta,m))}{\partial t}\Big|_{t=\hat{\theta}} = 0.$$

In the view of assumption **(A4)**, we have (15).

10.3. Unbiasedness of $\tilde{\ell}_{\hat{\theta},\hat{m}}$

We start with some notation. Let $P^{Y|X}_{\theta,m}$ denote the conditional distribution of Y given X, where $Y = m(\theta^\top X)+\epsilon$. For any $(\theta,m)\in\Theta\times\mathcal{S}$ and $f\in L_2(P_{\theta,m})$, define

$$E_{\theta,m}(f) := \int f dP_{\theta,m},\quad E^X_{\theta,m}(f) := \int_{\mathbb{R}} f dP^{Y|X}_{\theta,m},\quad \text{and}\quad E_X(f) := \int f dP_X. \tag{74}$$

For $f : \mathcal{X} \to \mathbb{R}$ we have $P_{\theta_0,m_0}[f(X)] = P_X(f(X))$ and

$$P_{\theta_0,m_0}\left[\big(Y - m_0(\theta_0^\top X)\big)^2 f(X)\right] = E_X\left[E^X_{\theta_0,m_0}\left[f(X)\big(Y - m_0(\theta_0^\top X)\big)^2\right]\right] = E_X\left[f(X)\sigma^2(X)\right],$$

where $\sigma^2(x) = \mathbb{E}(\epsilon^2|X = x)$. For the rest of the paper, we use $E_{\theta,m}$ and $P_{\theta,m}$ interchangeably.

Theorem 8. *Under assumptions (A0)–(A3) and (B1)–(B3),*

$$P_{\theta,m_0}\tilde{\ell}_{\theta,m} = 0,$$

for all $\theta \in \Theta$ and $m \in \{g \in \mathcal{S} : J(g) < \infty\}$.

Proof. Note that by definition (74), we have $E^X_{\theta,m_0}\left[Y - m(\theta^\top X)\right] = m_0(\theta^\top X) - m(\theta^\top X)$. Thus

$$\begin{aligned}
P_{\theta,m_0}\tilde{\ell}_{\theta,m} &= E_{\theta,m_0}[(Y - m(\theta^\top X))m'(\theta^\top X)K_1(X;\theta)] \\
&= E_X\left[E^X_{\theta,m_0}[(Y - m(\theta^\top X))m'(\theta^\top X)K_1(X;\theta)]\right] \\
&= E_{\theta,m_0}[(m_0\, m' - m\, m')(\theta^\top X)K_1(X;\theta)] \\
&= E_{\theta,m_0}\left[\mathbb{E}\big((m_0\, m' - m\, m')(\theta^\top X)K_1(X;\theta)|\theta^\top X\big)\right] \\
&= E_{\theta,m_0}\left[(m_0\, m' - m\, m')(\theta^\top X)\mathbb{E}\big(K_1(X;\theta)|\theta^\top X\big)\right] \\
&= 0.
\end{aligned}$$

□

10.4. Proof of (19) in Theorem 5

To prove (19), we will need some auxiliary results on the asymptotic behavior of $\tilde{\ell}_{\hat{\theta},\hat{m}}$. We summarize them in the following lemma.

Lemma 14. *Under assumptions (A1)–(A5) and (B2)–(B3), the PLSE satisfies*

$$P_{\theta_0,m_0}|\tilde{\ell}_{\hat{\theta},\hat{m}} - \tilde{\ell}_{\theta_0,m_0}|^2 = o_p(1), \tag{75}$$

$$P_{\hat{\theta},m_0}|\tilde{\ell}_{\hat{\theta},\hat{m}}|^2 = O_p(1). \tag{76}$$

Proof. Recall that $K_1(x;\theta) = H_\theta^\top\big(x - h_\theta(\theta^\top x)\big)$. To prove (75), observe that

$$
\begin{aligned}
& P_{\theta_0,m_0}\big|\tilde{\ell}_{\hat\theta,\hat m} - \tilde{\ell}_{\theta_0,m_0}\big|^2 \\
&= P_{\theta_0,m_0}\big|(Y - \hat m(\hat\theta^\top X))\hat m'(\hat\theta^\top X)K_1(X;\hat\theta) \\
&\qquad - (Y - m_0(\theta_0^\top X)){m_0}'(\theta_0^\top X)K_1(X;\theta_0)\big|^2 \\
&= P_{\theta_0,m_0}\big|\{(Y - m_0(\theta_0^\top X)) + (m_0(\theta_0^\top X) - \hat m(\hat\theta^\top X))\}\hat m'(\hat\theta^\top X)K_1(X;\hat\theta) \\
&\qquad - (Y - m_0(\theta_0^\top X)){m_0}'(\theta_0^\top X)K_1(X;\theta_0)\big|^2 \\
&= P_{\theta_0,m_0}\big|(Y - m_0(\theta_0^\top X))\{\hat m'(\hat\theta^\top X)K_1(X;\hat\theta) - {m_0}'(\theta_0^\top X)K_1(X;\theta_0)\} \\
&\qquad + (m_0(\theta_0^\top X) - \hat m(\hat\theta^\top X))\hat m'(\hat\theta^\top X)K_1(X;\hat\theta)\big|^2 \\
&= P_{\theta_0,m_0}[(Y - m_0(\theta_0^\top X))^2\big|\hat m'(\hat\theta^\top X)K_1(X;\hat\theta) - {m_0}'(\theta_0^\top X)K_1(X;\theta_0)\big|^2] \\
&\qquad + P_{\theta_0,m_0}\big|(m_0(\theta_0^\top X) - \hat m(\hat\theta^\top X))\hat m'(\hat\theta^\top X)K_1(X;\hat\theta)\big|^2, \\
&= P_X\big[\sigma^2(X)|\hat m'(\hat\theta^\top X)K_1(X;\hat\theta) - {m_0}'(\theta_0^\top X)K_1(X;\theta_0)|^2\big] \\
&\qquad + P_{\theta_0,m_0}\big|(m_0(\theta_0^\top X) - \hat m(\hat\theta^\top X))\hat m'(\hat\theta^\top X)K_1(X;\hat\theta)\big|^2, \\
&\le \|\sigma^2(\cdot)\|_\infty P_X\big[|\hat m'(\hat\theta^\top X)K_1(X;\hat\theta) - {m_0}'(\theta_0^\top X)K_1(X;\theta_0)|^2\big] \\
&\qquad + P_{\theta_0,m_0}\big|(m_0(\theta_0^\top X) - \hat m(\hat\theta^\top X))\hat m'(\hat\theta^\top X)K_1(X;\hat\theta)\big|^2, \\
&= \|\sigma^2(\cdot)\|_\infty \mathbf{I} + \mathbf{II}
\end{aligned}
$$

where in the fourth equality, the cross product term is zero as $E^X_{\theta_0,m_0}\big(Y - m_0(\theta_0^\top X)\big) = 0$ and

$$
\begin{aligned}
\mathbf{I} &:= P_X\big[|\hat m'(\hat\theta^\top X)K_1(X;\hat\theta) - {m_0}'(\theta_0^\top X)K_1(X;\theta_0)|^2\big], \\
\mathbf{II} &:= P_X\big[|(m_0(\theta_0^\top X) - \hat m(\hat\theta^\top X))\hat m'(\hat\theta^\top X)K_1(X;\hat\theta)|^2\big].
\end{aligned}
$$

Recall that for all $a \in \mathbb{R}^d$, we have $|H_\theta^\top a| \le |a|$; see proof of Lemma 1. We will now show that $\mathbf{I} = o_p(1)$. Observe that

$$
\begin{aligned}
\mathbf{I} &\le 2P_X\Big[\Big|H_{\theta_0}^\top\Big((\hat m'(\hat\theta^\top X) - {m_0}'(\theta_0^\top X))X + ({m_0}'\,h_{\theta_0})(\theta_0^\top X) - (\hat m'\,h_{\hat\theta})(\hat\theta^\top X)\Big)\Big|^2\Big] \\
&\qquad + 2P_X\Big[\Big|\big(H_{\hat\theta}^\top - H_{\theta_0}^\top\big)\hat m'(\hat\theta^\top X)(X - h_{\hat\theta}(\hat\theta^\top X))\Big|^2\Big] \\
&\le 2P_X\Big[\Big|H_{\theta_0}^\top\Big((\hat m'(\hat\theta^\top X) - {m_0}'(\theta_0^\top X))X + ({m_0}'\,h_{\theta_0})(\theta_0^\top X) - (\hat m'\,h_{\hat\theta})(\hat\theta^\top X)\Big)\Big|^2\Big] \\
&\qquad + [4CT(1 + J(\hat m))]^2|\hat\theta - \theta_0|^2 \\
&\le 2P_X\Big[\Big|(\hat m'(\hat\theta^\top X) - {m_0}'(\theta_0^\top X))X + ({m_0}'\,h_{\theta_0})(\theta_0^\top X) - (\hat m'\,h_{\hat\theta})(\hat\theta^\top X)\Big|^2\Big], \\
&\qquad + [4CT(1 + J(\hat m))]^2|\hat\theta - \theta_0|^2,
\end{aligned}
$$

where the second inequality follows from (c) of Lemma 1. Let us define

$$
\mathbf{III} := 4P_X\big|({m_0}'\,h_{\theta_0})(\theta_0^\top X) - (\hat m'\,h_{\hat\theta})(\hat\theta^\top X)\big|^2.
$$

Using Lemma 4 and the fact that $\sup_{x\in\mathcal{X}}|x| \le T$ (see **(A2)**), we have

$$
\begin{aligned}
\mathbf{I} &\le 4T^2P_X|\hat m'(\hat\theta^\top X) - {m_0}'({\theta_0}^\top X)|^2 + \mathbf{III} + o_p(1) \\
&\le 8T^2P_X|\hat m'(\hat\theta^\top X) - \hat m'({\theta_0}^\top X)|^2 + 8T^2P_X|(\hat m' - {m_0}')({\theta_0}^\top X)|^2 + \mathbf{III} + o_p(1) \\
&\le 8T^2J^2(\hat m)P_X\big[|\hat\theta^\top X - {\theta_0}^\top X|\big] + 8T^2\|\hat m' - {m_0}'\|^2_{D_0} + \mathbf{III} + o_p(1) \\
&\le 8T^2J^2(\hat m)T|\hat\theta - \theta_0| + 8T^2\|\hat m' - {m_0}'\|^2_{D_0} + \mathbf{III} + o_p(1).
\end{aligned}
$$

Recall that both $|\hat{\theta}-\theta_0|$ and $\|\hat{m}'-m_0'\|_{D_0}$ are $o_p(1)$; see Theorem 3. Thus we have $\mathbf{I}=o_p(1)$, if we can show that $\mathbf{III}=o_p(1)$. First observe that by Theorem 3 and assumption **(B2)**, we have that $P_X\big|h_{\theta_0}(\theta_0^\top X)-h_{\hat{\theta}}(\hat{\theta}^\top X)\big|^2 \xrightarrow{P} 0$. Hence we can bound $\mathbf{III}$ from above:

$$\begin{aligned}
\mathbf{III} &= 4P_X\big|(m_0'h_{\theta_0})(\theta_0^\top X)-m_0'(\theta_0{}^\top X)h_{\hat{\theta}}(\hat{\theta}^\top X)+m_0'(\theta_0{}^\top X)h_{\hat{\theta}}(\hat{\theta}^\top X)-(\hat{m}'h_{\hat{\theta}})(\hat{\theta}^\top X)\big|^2\\
&\le 8P_X\big|(m_0'\,h_{\theta_0})(\theta_0^\top X)-m_0'(\theta_0{}^\top X)\;h_{\hat{\theta}}(\hat{\theta}^\top X)\big|^2+8P_X\big|m_0'(\theta_0{}^\top X)h_{\hat{\theta}}(\hat{\theta}^\top X)-(\hat{m}'h_{\hat{\theta}})(\hat{\theta}^\top X)\big|^2\\
&\le 8\|m_0'\|_\infty^2P_X\big|h_{\theta_0}(\theta_0^\top X)-h_{\hat{\theta}}(\hat{\theta}^\top X)\big|^2+8\|h_{\hat{\theta}}\|_{2,\infty}^2P_X|m_0'(\theta_0{}^\top X)-\hat{m}'(\hat{\theta}^\top X)|^2\\
&\le 8\|m_0'\|_\infty^2P_X\big|h_{\theta_0}(\theta_0^\top X)-h_{\hat{\theta}}(\hat{\theta}^\top X)\big|^2\\
&\qquad+16\|h_{\hat{\theta}}\|_{2,\infty}^2\Big[P_X\big|(m_0'-\hat{m}')(\theta_0^\top X)\big|^2+P_X|\hat{m}'(\theta_0^\top X)-\hat{m}'(\hat{\theta}^\top X)|^2\Big]\\
&\le 8\|m_0'\|_\infty^2P_X\big|h_{\theta_0}(\theta_0^\top X)-h_{\hat{\theta}}(\hat{\theta}^\top X)\big|^2+16\|h_{\hat{\theta}}\|_{2,\infty}^2\Big[\|m_0'-\hat{m}'\|_{D_0}^2+J^2(\hat{m})T^2|\hat{\theta}-\theta_0|^2\Big].
\end{aligned}$$

As each of the terms in the last inequality of the above display are $o_p(1)$, we have that $\mathbf{III}=o_p(1)$. The proof of (75) will be complete, if we can show that $\mathbf{II}=o_p(1)$. First note that for all $x\in\mathcal{X}$,

$$|K_1(x;\theta)|\le|H_\theta^\top(x-h_\theta(\theta^\top x))|\le|x-h_\theta(\theta^\top x)|\le 2T. \tag{77}$$

By Theorem 2 and assumption **(A4)**, we have

$$\begin{aligned}
\mathbf{II} &= P_X\left[|(m_0(\theta_0^\top X)-\hat{m}(\hat{\theta}^\top X))\hat{m}'(\hat{\theta}^\top X)K_1(X;\hat{\theta})|^2\right]\\
&\le 4T^2\|\hat{m}'\|_\infty^2P_X|(m_0(\theta_0^\top X)-\hat{m}(\hat{\theta}^\top X))|^2\xrightarrow{P}0.
\end{aligned}$$

All these facts combined prove that $P_{\theta_0,m_0}|\tilde{\ell}_{\hat{\theta},\hat{m}}-\tilde{\ell}_{\theta_0,m_0}|^2=o_p(1)$.

Next we prove (76). Observe that

$$\begin{aligned}
P_{\hat{\theta},m_0}|\tilde{\ell}_{\hat{\theta},\hat{m}}|^2 &= P_{\hat{\theta},m_0}\big|(Y-\hat{m}(\hat{\theta}^\top X))\hat{m}'(\hat{\theta}^\top X)K_1(X;\hat{\theta})\big|^2\\
&= P_{\hat{\theta},m_0}\big|(Y-m_0(\hat{\theta}^\top X)+m_0(\hat{\theta}^\top X)-\hat{m}(\hat{\theta}^\top X))\hat{m}'(\hat{\theta}^\top X)K_1(X;\hat{\theta})\big|^2\\
&\le 4T^2\|\hat{m}'\|_\infty^2P_{\hat{\theta},m_0}[(Y-m_0(\hat{\theta}^\top X)+m_0(\hat{\theta}^\top X)-\hat{m}(\hat{\theta}^\top X))]^2\\
&= 4T^2\|\hat{m}'\|_\infty^2P_{\hat{\theta},m_0}[(Y-m_0(\hat{\theta}^\top X))^2+(m_0(\hat{\theta}^\top X)-\hat{m}(\hat{\theta}^\top X))^2]\\
&= 4T^2\|\hat{m}'\|_\infty^2\Big[P_X|\sigma^2(X)|+P_X|m_0(\hat{\theta}^\top X)-\hat{m}(\hat{\theta}^\top X)|^2\Big]=O_p(1),
\end{aligned}$$

where in the penultimate equality, the cross product term is zero as $E^X_{\hat{\theta},m_0}(Y-m_0(\hat{\theta}^\top X))=0$. $\square$

Now we prove (19). For $\theta\in\Theta$ and $m\in\mathcal{S}$, define $p_{\theta,m}(y,x):=p_{\epsilon|X}(y-m(\theta^\top x),x)p_X(x)$ to be the joint density of (Y,X) with respect to the dominating measure μ, where $Y=m(\theta^\top X)+\epsilon$ and $X\sim P_X$. Now consider the following submodel for θ_0:

$$\zeta_{\eta,\theta_0}=\sqrt{1-|\eta|^2}\theta_0+H_{\theta_0}\eta.$$

By definition of $\hat{\eta}$ (see (21)), we have that $\zeta_{\hat{\eta},\theta_0}=\hat{\theta}$. As $\hat{\eta}=o_p(1)$ (see Theorem 4 and (22)) differentiability in quadratic mean of model (1) implies that

$$\int\left(\sqrt{p_{\hat{\theta},m_0}}-\sqrt{p_{\theta_0,m_0}}-\frac{1}{2}\hat{\eta}^\top S_{\theta_0,m_0}\sqrt{p_{\theta_0,m_0}}\right)^2d\mu=o_p(|\hat{\eta}|^2)=o_p(|\hat{\theta}-\theta_0|^2). \tag{78}$$

With Lemma 14 in hand, we now show that (19) holds. Note that

$$
\begin{aligned}
&\sqrt{n}(P_{\hat\theta,m_0} - P_{\theta_0,m_0})\tilde\ell_{\hat\theta,\hat m} - \sqrt{n}P_{\theta_0,m_0}(\tilde\ell_{\theta_0,m_0}S^\top_{\theta_0,m_0})H^\top_{\theta_0}(\hat\theta-\theta_0)\\
&= \sqrt{n}\int \tilde\ell_{\hat\theta,\hat m}(\sqrt{p_{\hat\theta,m_0}} + \sqrt{p_{\theta_0,m_0}})\left(\sqrt{p_{\hat\theta,m_0}} - \sqrt{p_{\theta_0,m_0}} - \frac{1}{2}\hat\eta^\top S_{\theta_0,m_0}\sqrt{p_{\theta_0,m_0}}\right)d\mu\\
&\quad + \sqrt{n}\int \tilde\ell_{\hat\theta,\hat m}(\sqrt{p_{\hat\theta,m_0}} + \sqrt{p_{\theta_0,m_0}})\frac{1}{2}\hat\eta^\top S_{\theta_0,m_0}\sqrt{p_{\theta_0,m_0}}d\mu\\
&\quad - \sqrt{n}\int \tilde\ell_{\theta_0,m_0}S^\top_{\theta_0,m_0}H^\top_{\theta_0}(\hat\theta-\theta_0)p_{\theta_0,m_0}d\mu\\
&= \mathbf{IV} + \sqrt{n}\int \tilde\ell_{\hat\theta,\hat m}(\sqrt{p_{\hat\theta,m_0}} + \sqrt{p_{\theta_0,m_0}})\frac{1}{2}\hat\eta^\top S_{\theta_0,m_0}\sqrt{p_{\theta_0,m_0}}d\mu\\
&\quad - \sqrt{n}\int \tilde\ell_{\hat\theta,\hat m}\sqrt{p_{\theta_0,m_0}}\hat\eta^\top S_{\theta_0,m_0}\sqrt{p_{\theta_0,m_0}}d\mu\\
&\quad + \sqrt{n}\int \tilde\ell_{\hat\theta,\hat m}\sqrt{p_{\theta_0,m_0}}\hat\eta^\top S_{\theta_0,m_0}\sqrt{p_{\theta_0,m_0}}d\mu\\
&\quad - \sqrt{n}\int \tilde\ell_{\theta_0,m_0}S^\top_{\theta_0,m_0}H^\top_{\theta_0}(\hat\theta-\theta_0)p_{\theta_0,m_0}d\mu\\
&= \mathbf{IV} + \sqrt{n}\int \tilde\ell_{\hat\theta,\hat m}(\sqrt{p_{\hat\theta,m_0}} - \sqrt{p_{\theta_0,m_0}})\frac{1}{2}S^\top_{\theta_0,m_0}\hat\eta\sqrt{p_{\theta_0,m_0}}d\mu\\
&\quad + \sqrt{n}\int \tilde\ell_{\hat\theta,\hat m}\hat\eta^\top S_{\theta_0,m_0}p_{\theta_0,m_0}d\mu - \sqrt{n}\int \tilde\ell_{\theta_0,m_0}S^\top_{\theta_0,m_0}H^\top_{\theta_0}(\hat\theta-\theta_0)p_{\theta_0,m_0}d\mu\\
&= \mathbf{IV} + \frac{1}{2}\mathbf{V} + \sqrt{n}\int \tilde\ell_{\hat\theta,\hat m}\hat\eta^\top S_{\theta_0,m_0}p_{\theta_0,m_0}d\mu - \sqrt{n}\int \tilde\ell_{\theta_0,m_0}S^\top_{\theta_0,m_0}H^\top_{\theta_0}(\hat\theta-\theta_0)p_{\theta_0,m_0}d\mu\\
&= \mathbf{IV} + \frac{1}{2}\mathbf{V} + \sqrt{n}\int [\tilde\ell_{\hat\theta,\hat m} - \tilde\ell_{\theta_0,m_0}]S^\top_{\theta_0,m_0}\hat\eta p_{\theta_0,m_0}d\mu \qquad \text{(by (22))}\\
&= \mathbf{IV} + \frac{1}{2}\mathbf{V} + \mathbf{VI},
\end{aligned}
$$

where

$$
\begin{aligned}
\mathbf{IV} &= \sqrt{n}\int \tilde\ell_{\hat\theta,\hat m}(\sqrt{p_{\hat\theta,m_0}} + \sqrt{p_{\theta_0,m_0}})\left(\sqrt{p_{\hat\theta,m_0}} - \sqrt{p_{\theta_0,m_0}} - \frac{1}{2}\hat\eta^\top S_{\theta_0,m_0}\sqrt{p_{\theta_0,m_0}}\right)d\mu,\\
\mathbf{V} &= \sqrt{n}\left[\int \tilde\ell_{\hat\theta,\hat m}(\sqrt{p_{\hat\theta,m_0}} - \sqrt{p_{\theta_0,m_0}})S^\top_{\theta_0,m_0}\sqrt{p_{\theta_0,m_0}}d\mu\right]\hat\eta,\\
\mathbf{VI} &= \sqrt{n}\left[\int [\tilde\ell_{\hat\theta,\hat m} - \tilde\ell_{\theta_0,m_0}]S^\top_{\theta_0,m_0}p_{\theta_0,m_0}d\mu\right]\hat\eta.
\end{aligned}
$$

Observe that $\mathbf{IV}, \mathbf{V}$, and $\mathbf{VI}$ are elements of $\mathbb{R}^d$. In the following, we show that $\mathbf{IV}, \mathbf{V}$, and $\mathbf{VI}$ are $o_p(\sqrt{n}|\hat\theta-\theta_0|)$. Using the Cauchy-Schwarz inequality and the fact that $(a+b)^2 \le 2(a^2+b^2)$, we have

$$
\begin{aligned}
\left|\mathbf{IV}\right|^2 &\le 2n\int |\tilde\ell_{\hat\theta,\hat m}|^2(p_{\hat\theta,m_0} + p_{\theta_0,m_0})d\mu \int\left(\sqrt{p_{\hat\theta,m_0}} - \sqrt{p_{\theta_0,m_0}} - \frac{1}{2}\hat\eta^\top S_{\theta_0,m_0}\sqrt{p_{\theta_0,m_0}}\right)^2 d\mu\\
&\le 2n\left[P_{\hat\theta,m_0}|\tilde\ell_{\hat\theta,\hat m}|^2 + P_{\theta_0,m_0}|\tilde\ell_{\hat\theta,\hat m} - \tilde\ell_{\theta_0,m_0}|^2 + P_{\theta_0,m_0}|\tilde\ell_{\theta_0,m_0}|^2\right]o_p(|\hat\eta|^2)\\
&= o_p(n|\hat\theta-\theta_0|^2),
\end{aligned}
$$

where the equality is due to Lemma 14, (78), and the fact that $\tilde\ell_{\theta_0,m_0} \in L_2(P_{\theta_0,m_0})$ (see **(A1)**, **(A2)**, and Lemma 5).

Now we will show that $|\mathbf{VI}| = o_p(|\sqrt{n}(\theta - \theta_0)|)$. For a matrix $\mathbb{A} \in \mathbb{R}^{d\times d}$, let $\|\mathbb{A}\|_F$ denote the Frobenius norm of $\mathbb{A}$. Then we have

$$\left|\mathbf{VI}\right|^2 \le \left\| \int [\tilde{\ell}_{\hat{\theta},\hat{m}} - \tilde{\ell}_{\theta_0,m_0}] S^\top_{\theta_0,m_0} p_{\theta_0,m_0} d\mu \right\|_F^2 |\sqrt{n}\hat{\eta}|^2. \tag{79}$$

Let $f = (f_1, \dots, f_d)$ and $g = (g_1, \dots, g_d)$ be two functions that map a separable metric space $\Re$ to $\mathbb{R}^d$. If ν is a finite measure on $\Re$ such that $|f|$ and $|g|$ are $L_2(\nu)$, then by the Cauchy-Schwarz inequality, we have

$$\begin{aligned}\left\| \int_\Re f g^\top d\nu \right\|_F^2 &= \sum_{i,j} \left[\int_\Re f_i g_j d\nu \right]^2 \le \sum_{i,j} \int_\Re f_i^2 d\nu \int_\Re g_j^2 d\nu \\ &= \left[\sum_i \int_\Re f_i^2 d\nu \right] \left[\sum_j \int_\Re g_j^2 d\nu \right] = \int_\Re |f|^2 d\nu \int_\Re |g|^2 d\nu.\end{aligned} \tag{80}$$

Thus from Lemma 14, (79), and the fact that $S_{\theta_0,m_0} \in L_2(P_{\theta_0,m_0})$, we have

$$\begin{aligned}\left|\mathbf{VI}\right|^2 &\le |\sqrt{n}\hat{\eta}|^2 \int |\tilde{\ell}_{\hat{\theta},\hat{m}} - \tilde{\ell}_{\theta_0,m_0}|^2 p_{\theta_0,m_0} d\mu \int |S_{\theta_0,m_0}|^2 p_{\theta_0,m_0} d\mu \\ &= |\sqrt{n}\hat{\eta}|^2 P_{\theta_0,m_0} |\tilde{\ell}_{\hat{\theta},\hat{m}} - \tilde{\ell}_{\theta_0,m_0}|^2 P_{\theta_0,m_0} |S_{\theta_0,m_0}|^2 = o_p(|\sqrt{n}\hat{\eta}|^2) = o_p(|\sqrt{n}(\hat{\theta} - \theta_0)|^2).\end{aligned}$$

We will now prove that

$$|\mathbf{V}|^2 = o_p(|\sqrt{n}(\hat{\theta} - \theta_0)|^2). \tag{81}$$

Observe that

$$|\mathbf{V}|^2 \le \left\| \int \tilde{\ell}_{\hat{\theta},\hat{m}} (\sqrt{p_{\hat{\theta},m_0}} - \sqrt{p_{\theta_0,m_0}}) S^\top_{\theta_0,m_0} \sqrt{p_{\theta_0,m_0}} d\mu \right\|_F^2 |\sqrt{n}\hat{\eta}|^2.$$

Thus the proof of (81) will be complete, if we can show that

$$\left\| \int \tilde{\ell}_{\hat{\theta},\hat{m}} (\sqrt{p_{\hat{\theta},m_0}} - \sqrt{p_{\theta_0,m_0}}) S^\top_{\theta_0,m_0} \sqrt{p_{\theta_0,m_0}} d\mu \right\|_F^2 = o_p(1). \tag{82}$$

We will show this by splitting the integral in the above display into two regions that depend on n. More specifically by splitting the integral into $\{(y,x) : |S_{\theta_0,m_0}(y,x)| > r_n\}$ and $\{(y,x) : |S_{\theta_0,m_0}(y,x)| \le r_n\}$, where $\{r_n\}$ is a sequence of constants to be chosen later.

Observe that by (80), we have

$$
\begin{aligned}
&\left\|\int_{|S_{\theta_0,m_0}|\le r_n} \tilde{\ell}_{\hat\theta,\hat m} S_{\theta_0,m_0}^\top(\sqrt{p_{\hat\theta,m_0}} - \sqrt{p_{\theta_0,m_0}})\sqrt{p_{\theta_0,m_0}}d\mu\right\|_F^2 \\
&= \sum_{i,j}\left[\int_{|S_{\theta_0,m_0}|\le r_n}\left\{\tilde{\ell}_{\hat\theta,\hat m} S_{\theta_0,m_0}^\top\right\}_{i,j}\left(\sqrt{p_{\hat\theta,m_0}} - \sqrt{p_{\theta_0,m_0}}\right)\sqrt{p_{\theta_0,m_0}}d\mu\right]^2 \\
&\le \left[\int\left(\sqrt{p_{\hat\theta,m_0}} - \sqrt{p_{\theta_0,m_0}}\right)^2 d\mu\right]\sum_{i,j}\left[\int_{|S_{\theta_0,m_0}|\le r_n}\left\{\tilde{\ell}_{\hat\theta,\hat m} S_{\theta_0,m_0}^\top\right\}_{i,j}^2 p_{\theta_0,m_0}d\mu\right] \\
&\le 2\left[\int\left(\frac{1}{2}S_{\theta_0,m_0}^\top(\hat\theta-\theta_0)\sqrt{p_{\theta_0,m_0}}\right)^2 + \left(\sqrt{p_{\hat\theta,m_0}} - \sqrt{p_{\theta_0,m_0}} - \frac{1}{2}S_{\theta_0,m_0}^\top(\hat\theta-\theta_0)\sqrt{p_{\theta_0,m_0}}\right)^2 d\mu\right] \\
&\qquad\times\sum_{i,j}\left[\int_{|S_{\theta_0,m_0}|\le r_n}\left\{\tilde{\ell}_{\hat\theta,\hat m} S_{\theta_0,m_0}^\top\right\}_{i,j}^2 p_{\theta_0,m_0}d\mu\right] \qquad (83) \\
&= 2\left[\int\left(\frac{1}{2}S_{\theta_0,m_0}^\top(\hat\theta-\theta_0)\sqrt{p_{\theta_0,m_0}}\right)^2 + \left(\sqrt{p_{\hat\theta,m_0}} - \sqrt{p_{\theta_0,m_0}} - \frac{1}{2}S_{\theta_0,m_0}^\top(\hat\theta-\theta_0)\sqrt{p_{\theta_0,m_0}}\right)^2 d\mu\right] \\
&\qquad\times\int_{|S_{\theta_0,m_0}|\le r_n}|\tilde{\ell}_{\hat\theta,\hat m}|^2|S_{\theta_0,m_0}^\top|^2 p_{\theta_0,m_0}d\mu \\
&\le 2r_n^2 P_{\theta_0,m_0}|\tilde{\ell}_{\hat\theta,\hat m}|^2\left[O_p(|\hat\theta-\theta_0|^2) + o_p(|\hat\theta-\theta_0|^2)\right] = r_n^2 o_p(1),
\end{aligned}
$$

where the last equality follows from Theorem 4 and (76). Now to bound the second part of the integral, observe that

$$
\begin{aligned}
&\left\|\int_{|S_{\theta_0,m_0}|> r_n} \tilde{\ell}_{\hat\theta,\hat m} S_{\theta_0,m_0}^\top(\sqrt{p_{\hat\theta,m_0}} - \sqrt{p_{\theta_0,m_0}})\sqrt{p_{\theta_0,m_0}}d\mu\right\|_F^2 \\
&\le 2\int_{|S_{\theta_0,m_0}|> r_n}|S_{\theta_0,m_0}|^2 p_{\theta_0,m_0}d\mu\int|\tilde{\ell}_{\hat\theta,\hat m}|^2(p_{\hat\theta,m_0}+p_{\theta_0,m_0})d\mu \qquad (84) \\
&\le O_p(1)\int_{|S_{\theta_0,m_0}|> r_n}|S_{\theta_0,m_0}|^2 p_{\theta_0,m_0}d\mu.
\end{aligned}
$$

Since $P_{\theta_0,m_0}|S_{\theta_0,m_0}|^2 = O_p(1)$, it is easy to see that we can find a sequence $\{r_n\}$ such that both (83) and (84) are $o_p(1)$. Thus we have (82).

10.5. Proof of Lemma 2

Before proceeding to prove Lemma 2, we find the entropy of the class of matrices $\{H_\theta : \theta \in \Theta\}$, where H_θ satisfies properties of Lemma 1.

Lemma 15. *We can construct a cover $\{\eta_1, \ldots, \eta_{N_\varepsilon}\}$ of $\Theta \cap B_{\theta_0}(1/2)$ such that $N_\varepsilon \lesssim \varepsilon^{-2d}$ and for every $\theta \in \Theta \bigcap B_{\theta_0}(1/2)$, there exists an $i \le N_\varepsilon$ such that*

$$|\theta - \eta_i| \le \varepsilon \text{ and } \|H_\theta^\top - H_{\eta_i}^\top\|_2 \le \varepsilon. \qquad (85)$$

Proof. To find the entropy with respect to the matrix 2-norm, we construct a ε-cover for the set $\{H_\theta^\top : \theta \in \Theta\}$. By Lemma 4.1 of Pollard (1990), we have that

$$N(\varepsilon^2/(8+64/\sqrt{15}), \Theta \cap B_{\theta_0}(1/2) \setminus B_{\theta_0}(\varepsilon/2), |\cdot|) \lesssim \varepsilon^{-2d}.$$

Let $\{\theta_i\}_{1\le i\le N_\varepsilon}$ for $N_\varepsilon \lesssim \varepsilon^{-2d}$ form a cover of $\Theta \cap B_{\theta_0}(1/2) \setminus B_{\theta_0}(\varepsilon/2)$. We can without loss of generality assume that $|\theta_i - \theta_0| \ge \varepsilon/2$ for all $1 \le i \le N_\varepsilon$. We claim that $H_{\theta_0}^\top \cup \{H_{\theta_i}^\top\}_{1\le i\le N_\varepsilon}$ forms a ε-cover for $\{H_\theta^\top : \theta \in \Theta\}$. It is enough to show that for every $\eta \in \Theta$, we can find $i^* \in \{0, 1, \ldots, N_\varepsilon\}$ such that $\|H_\eta^\top - H_{\theta_{i^*}}^\top\|_2 \le \varepsilon$. If $\eta \in B_{\theta_0}(\varepsilon/2)$ then choose $i^* = 0$. By condition (c) of Lemma 1, we have $\|H_\eta^\top - H_{\theta_0}^\top\|_2 \le |\eta - \theta_0| \le \varepsilon$. If $\eta \notin B_{\theta_0}(\varepsilon/2)$ then choose i^* such that $|\eta - \theta_{i^*}| \le \varepsilon^2/(8 + 64/\sqrt{15})$. Thus by condition (d) of Lemma 1, we have

$$\|H_\eta^\top - H_{\theta_{i^*}}^\top\|_2 \le (8 + 64/\sqrt{15})\frac{|\eta - \theta_{i^*}|}{|\eta - \theta_0| + |\theta_{i^*} - \theta_0|} \le (8 + 64/\sqrt{15})\frac{\varepsilon^2/(8 + 64/\sqrt{15})}{\varepsilon} \le \varepsilon. \quad \square$$

Now we will show that $D_{M_1,M_2,M_3}(n)$ is an envelope of $\mathcal{D}_{M_1,M_2,M_3}(n)$. For every $(m, \theta) \in \mathcal{C}_{M_1,M_2,M_3}(n)$ and $x \in \chi$, we have

$$\begin{aligned}
&|(m_0(\theta_0^\top x) - m(\theta^\top x))m'(\theta^\top x)K_1(x;\theta)| \\
&\le \big(|(m_0(\theta_0^\top x) - m(\theta_0^\top x)| + |m(\theta_0^\top x) - m(\theta^\top x))|\big)M_2 2T \\
&\le \big(\|m_0 - m\|_{D_0} + \|m'\|_\infty |\theta_0 - \theta||x|\big)M_2 2T \\
&\le 2TM_2(a_n^{-1} + \|m'\|_\infty|\theta_0 - \theta|T) \\
&\le 2TM_2(a_n^{-1} + TM_2\hat{\lambda}_n^{1/2}) = D_{M_1,M_2,M_3}(n),
\end{aligned}$$

where the first and second inequality follow from the facts that $\sup_{x\in\chi}|x| \le T$ and $\|K_1(\cdot;\theta)\|_{2,\infty} \le 2T$, see **(A2)** and (77). Next we prove that there exists finite c depending only on M_1, M_2, and M_3, such that

$$N(\varepsilon, \mathcal{D}^*_{M_1,M_2,M_3}, \|\cdot\|_{2,\infty}) \le c\exp\left(\frac{c}{\varepsilon} + \frac{c}{\sqrt{\varepsilon}}\right)\varepsilon^{-2(d-1)}. \tag{86}$$

We first find covers for $\mathcal{C}^{m*}_{M_1,M_2,M_3}$, $\{f' : f \in \mathcal{C}^{m*}_{M_1,M_2,M_3}\}$, and $\Theta \cap B_{\theta_0}(1/2)$ and use them to construct a cover for $\mathcal{D}^*_{M_1,M_2,M_3}$. By Lemma 10 (for $k = 1$ and 2, respectively), we have

$$\begin{aligned}
N(\varepsilon, \mathcal{C}^{m*}_{M_1,M_2,M_3}, \|\cdot\|_\infty) &\le \exp(c/\sqrt{\varepsilon}), \\
N(\varepsilon, \{f' : f \in \mathcal{C}^{m*}_{M_1,M_2,M_3}\}, \|\cdot\|_\infty) &\le \exp(c/\varepsilon),
\end{aligned}$$

where c is a constant depending only on M_1, M_2, and M_3. Let us denote the functions in the ε-cover of $\mathcal{C}^{m*}_{M_1,M_2,M_3}$ by $r_1, \ldots, r_q$ and the functions in the ε-cover of $\{f' : f \in \mathcal{C}^{m*}_{M_1,M_2,M_3}\}$ by $l_1, \ldots, l_t$. By Lemma 15, we have that there exists $\theta_1, \ldots, \theta_s$ for $s \lesssim \varepsilon^{-4d}$ such that $\{\theta_i\}_{1\le i\le s}$ form an ε^2-cover of $\Theta \cap B_{\theta_0}(1/2)$ and satisfies (85). Fix $(m, \theta) \in \mathcal{C}_{M_1,M_2,M_3}(n)$. Without loss of generality assume that the function nearest to m in the ε-cover of $\mathcal{C}^{m*}_{M_1,M_2,M_3}$ is r_1, the function nearest to m' in the ε-cover of $\{f' : f \in \mathcal{C}^{m*}_{M_1,M_2,M_3}\}$ is l_1, and the vector nearest to θ in the ε^2-cover of $\Theta \cap B_{\theta_0}(1/2)$ is θ_1, i.e.,

$$\|m - r_1\|_\infty \le \varepsilon, \quad \|m' - l_1\|_\infty \le \varepsilon, \quad \|H_{\theta_1}^\top - H_\theta^\top\|_2 \le \varepsilon^2 \quad \text{and} \quad |\theta_1 - \theta| \le \varepsilon^2. \tag{87}$$

Now for every $x \in \chi$, observe that

$$\begin{aligned}
&\big|(m_0(\theta_0^\top x) - m(\theta^\top x))m'(\theta^\top x)K_1(x;\theta) - (m_0(\theta_0^\top x) - r_1(\theta_1^\top x))l_1(\theta_1^\top x)K_1(x;\theta_1)\big| \\
&= \big|(m_0(\theta_0^\top x) - m(\theta^\top x))m'(\theta^\top x)K_1(x;\theta) - \\
&\qquad \big(m_0(\theta_0^\top x) - m(\theta^\top x) + m(\theta^\top x) - r_1(\theta_1^\top x)\big)l_1(\theta_1^\top x)K_1(x;\theta_1)\big| \\
&\le \big|m_0(\theta_0^\top x) - m(\theta^\top x)\big|\big|m'(\theta^\top x)K_1(x;\theta) - l_1(\theta_1^\top x)K_1(x;\theta_1)\big| \\
&\qquad + \big|m(\theta^\top x) - r_1(\theta_1^\top x)\big|\big|l_1(\theta_1^\top x)K_1(x;\theta_1)\big| \\
&= \mathbf{A} + \mathbf{B},
\end{aligned} \tag{88}$$

where

$$\mathbf{A} := \left|m_0(\theta_0^\top x) - m(\theta^\top x)\right|m'(\theta^\top x)K_1(x;\theta) - l_1(\theta_1^\top x)K_1(x;\theta_1)\right|$$
$$\mathbf{B} := \left|m(\theta^\top x) - r_1(\theta_1^\top x)\right|\left|l_1(\theta_1^\top x)K_1(x;\theta_1)\right|.$$

We next find an upper bound for $\mathbf{A}$. First, by Lemma 1 and assumption **(B2)**, we have

$$\begin{aligned}
&\left|K_1(x;\theta) - K_1(x;\theta_1)\right| \\
&= \left|H_\theta^\top(x - h_\theta(\theta^\top x)) - H_{\theta_1}^\top(x - h_\theta(\theta^\top x)) + H_{\theta_1}^\top(x - h_\theta(\theta^\top x)) - H_{\theta_1}^\top(x - h_{\theta_1}(\theta_1^\top x))\right| \\
&\le \left|(H_\theta^\top - H_{\theta_1}^\top)(x - h_\theta(\theta^\top x))\right| + \left|H_{\theta_1}^\top\left[(x - h_\theta(\theta^\top x)) - (x - h_{\theta_1}(\theta_1^\top x))\right]\right| \\
&\le \|H_\theta^\top - H_{\theta_1}^\top\|_2 2T + \left|h_\theta(\theta^\top x) - h_{\theta_1}(\theta_1^\top x)\right| \\
&\le 2T\varepsilon^2 + (\bar{M} + \|h'_{\theta_0}\|_\infty)|\theta - \theta_1| \lesssim \varepsilon^2.
\end{aligned} \tag{89}$$

Now observe that

$$\begin{aligned}
\mathbf{A} &\le 2M_1\left|m'(\theta^\top x)K_1(x;\theta) - l_1(\theta_1^\top x)K_1(x;\theta_1)\right| \\
&\le 2M_1\left|m'(\theta^\top x)K_1(x;\theta) - l_1(\theta^\top x)K_1(x;\theta)\right| + \left|l_1(\theta^\top x)K_1(x;\theta) - l_1(\theta_1^\top x)K_1(x;\theta)\right| \\
&\qquad + 2M_1\left|l_1(\theta_1^\top x)K_1(x;\theta) - l_1(\theta_1^\top x)K_1(x;\theta_1)\right| \\
&\le 2M_1|K_1(x;\theta)|\left|m'(\theta^\top x) - l_1(\theta^\top x)\right| + |K_1(x;\theta)|\left|l_1(\theta^\top x) - l_1(\theta_1^\top x)\right| \\
&\qquad + 2M_1\|l_1\|_\infty\left|K_1(x;\theta) - K_1(x;\theta_1)\right| \\
&\lesssim 4TM_1\left(\varepsilon + \left[\int_D {l'_1}^2(z)dz\right]|\theta - \theta_1|^{1/2}T^{1/2}\right) + 2M_1M_2(2T + \bar{M})\varepsilon^2 \\
&\le 4TM_1(\varepsilon + M_3|\theta - \theta_1|^{1/2}T^{1/2}) + (2T + \bar{M})2M_1M_2\varepsilon^2 \\
&\lesssim \varepsilon,
\end{aligned} \tag{90}$$

where the penultimate inequality follows from (87) and the last inequality follows from **(A2)**, (89), and Lemma 4. To find an upper bound for $\mathbf{B}$, observe that

$$\begin{aligned}
\mathbf{B} &= \left|m(\theta^\top x) - r_1(\theta_1^\top x)\right|\left|l_1(\theta_1^\top x)K_1(x;\theta_1)\right| \\
&\le \left[\left|m(\theta^\top x) - r_1(\theta^\top x)\right| + \left|r_1(\theta^\top x) - r_1(\theta_1^\top x)\right|\right]\left|l_1(\theta_1^\top x)K_1(x;\theta_1)\right| \\
&\le \left[\varepsilon + \|r'_1\|_\infty|\theta - \theta_1|T\right]\|l_1\|_\infty 2T \lesssim \varepsilon.
\end{aligned} \tag{91}$$

Combining (88), (90), and (91) we get that $\{(m_0(\theta_0^\top x) - r_i(\theta_k^\top x))l'_j(\theta_k^\top x)K_1(x;\theta_k)\}_{i,j,k}$ for $1 \le i \le q, 1 \le j \le t$, and $1 \le k \le s$ form an (constant multiple of) ε-cover (with respect to $\|\cdot\|_{2,\infty}$ norm) of $\mathcal{D}^*_{M_1,M_2,M_3}$. Thus we have (86). Moreover, as $N_{[]}(\varepsilon, \mathcal{D}^*_{M_1,M_2,M_3}, \|\cdot\|_{2,P_{\theta_0,m_0}}) \lesssim N(\varepsilon, \mathcal{D}^*_{M_1,M_2,M_3}, \|\cdot\|_{2,\infty})$ and

$$\mathcal{D}_{M_1,M_2,M_3}(n) \subset \mathcal{D}^*_{M_1,M_2,M_3},$$

for every $n \in \mathbb{N}$, we have $N_{[]}(\varepsilon, \mathcal{D}_{M_1,M_2,M_3}(n), \|\cdot\|_{2,P_{\theta_0,m_0}}) \lesssim N_{[]}(\varepsilon, \mathcal{D}^*_{M_1,M_2,M_3}, \|\cdot\|_{2,P_{\theta_0,m_0}})$ and $J_{[]}(\gamma, \mathcal{D}^*_{M_1,M_2,M_3}(n), \|\cdot\|_{2,P_{\theta_0,m_0}}) \lesssim c\gamma^{1/2}$. Observe that $f \in \mathcal{D}_{M_1,M_2,M_3}(n)$ is a maps $\mathcal{X}$ to $\mathbb{R}^{d-1}$. For any $f \in \mathcal{D}_{M_1,M_2,M_3}(n)$, let $f_1, \ldots, f_{d-1}$ denote each of the real valued components, i.e., $f(\cdot) := (f_1(\cdot), \ldots, f_{d-1}(\cdot))$. With this notation, we have

$$\begin{aligned}
&\mathbb{P}\Bigg(\sup_{f \in \mathcal{D}_{M_1,M_2,M_3}(n)} |\mathbb{G}_n f| > \delta\Bigg) \\
&\le \sum_{i=1}^{d-1}\mathbb{P}\Bigg(\sup_{f \in \mathcal{D}_{M_1,M_2,M_3}(n)} |\mathbb{G}_n f_i| > \delta/\sqrt{d-1}\Bigg).
\end{aligned} \tag{92}$$

We can bound each term in the summation of (92) using the maximal inequality in Corollary 19.35 of van der Vaart (1998). We have

$$\begin{aligned}
\mathbb{P}\left(\sup_{f\in\mathcal{D}_{M_1,M_2,M_3}(n)}|\mathbb{G}_n f_1| > \delta\right) &\le \delta^{-1}\mathbb{E}\left(\sup_{f\in\mathcal{D}_{M_1,M_2,M_3}(n)}|\mathbb{G}_n f_1|\right)\\
&\le \delta^{-1}J_{[\,]}(\|D_{M_1,M_2,M_3}(n)\|, \mathcal{D}^*_{M_1,M_2,M_3}(n), \|\cdot\|_{2,P_{\theta_0,m_0}})\\
&\lesssim \delta^{-1}\|D_{M_1,M_2,M_3}(n)\|^{1/2}\\
&\lesssim \left[\hat{\lambda}_n^{1/2} + a_n^{-1}\right]^{1/2} \to 0, \qquad \text{as } n\to\infty. \qquad (93)
\end{aligned}$$

In the last inequality, we have used (29) and the fact that $D^2_{M_1,M_2,M_3}(n)$ is non-random. The lemma follows by combining (93) and (92).

10.6. Proof of Lemma 3

We will first show that, for every $(m,\theta)\in\mathcal{C}_{M_1,M_2,M_3}(n)$ and $x\in\mathcal{X}$, we have

$$\left|\epsilon\left[U_{\theta,m}(x) - U_{\theta_0,m_0}(x)\right]\right| \le |\epsilon| W_{M_1,M_2,M_3}(n).$$

Observe that for every $(m,\theta)\in\mathcal{C}_{M_1,M_2,M_3}(n)$ and $x\in\mathcal{X}$, we have

$$\begin{aligned}
&|U_{\theta,m}(x) - U_{\theta_0,m_0}(x)|\\
&\le |m'(\theta^\top x)K_1(x;\theta) - m'(\theta_0^\top x)K_1(x;\theta)| + |m'(\theta_0^\top x)K_1(x;\theta) - m_0'(\theta_0^\top x)K_1(x;\theta_0)|\\
&\le |m'(\theta^\top x)K_1(x;\theta) - m'(\theta_0^\top x)K_1(x;\theta)| + |m'(\theta_0^\top x)K_1(x;\theta) - m_0'(\theta_0^\top x)K_1(x;\theta)|\\
&\qquad + |m_0'(\theta_0^\top x)K_1(x;\theta) - m_0'(\theta_0^\top x)K_1(x;\theta_0)|\\
&\le |m'(\theta^\top x) - m'(\theta_0^\top x)||K_1(x;\theta)| + |m'(\theta_0^\top x) - m_0'(\theta_0^\top x)||K_1(x;\theta)|\\
&\qquad + |m_0'(\theta_0^\top x)||K_1(x;\theta) - K_1(x;\theta_0)|\\
&\le J(m)|\theta_0-\theta|^{1/2}T^{1/2}|K_1(x;\theta)| + \|m-m_0\|^S_{D_0}|K_1(x;\theta)| + \|m_0'\|_\infty(2T+\bar{M}+\|h'_{\theta_0}\|_\infty)|\theta_0-\theta|\\
&\le \left[2T^{3/2}M_3\hat{\lambda}_n^{1/4} + 2Ta_n^{-1} + M_2(2T+\bar{M}+\|h'_{\theta_0}\|_\infty)\right]\hat{\lambda}_n^{1/2} = W_{M_1,M_2,M_3}(n),
\end{aligned}$$

where for the third term in the penultimate inequality follows from (89).

Next, we will prove that there exists a constant c depending only on M_1, M_2, and M_3 such that

$$N_{[\,]}(\varepsilon, \mathcal{W}_{M_1,M_2,M_3}(n), \|\cdot\|_{2,P_{\theta_0,m_0}}) \le c\exp(c/\varepsilon)\varepsilon^{-2(d-1)}.$$

As in proof of Lemma 2, we first find covers for the class of functions $\{f' : f\in\mathcal{C}^{m*}_{M_1,M_2,M_3}\}$ and the set $\Theta\cap B_{\theta_0}(1/2)$ and use them to construct a cover for $\mathcal{W}^*_{M_1,M_2,M_3}$. By Lemma 10, we have

$$N(\varepsilon, \{f' : f\in\mathcal{C}^{m*}_{M_1,M_2,M_3}\}, \|\cdot\|_\infty) \le \exp(c/\varepsilon),$$

where c is a constant depending only on d, M_1, M_2, and M_3. We denote the functions in the ε-cover of $\{f' : f\in\mathcal{C}^{m*}_{M_1,M_2,M_3}\}$ by $l_1,\ldots,l_t$. By Lemma 15, we have that there exists $\theta_1,\ldots,\theta_s$ for $s\lesssim\varepsilon^{-4d}$ such that $\{\theta_i\}_{1\le i\le s}$ form an ε^2-cover of $\Theta\cap B_{\theta_0}(1/2)$ and satisfies (85) (with ε^2 instead of ε). Fix $(m,\theta)\in\mathcal{C}_{M_1,M_2,M_3}(n)$. Without loss of generality assume that the function nearest to m' in the ε-cover of $\{f' : f \in \mathcal{C}^{m*}_{M_1,M_2,M_3}\}$ is l_1 and the vector nearest to θ in the ε^2-cover of $\Theta \cap B_{\theta_0}(1/2)$ is θ_1, i.e.,

$$\|m' - l_1\|_\infty \le \varepsilon, \qquad |\theta_1 - \theta| \le \varepsilon^2, \quad \text{and} \quad \|H_\theta^\top - H_{\theta_1}^\top\|_2 \le \varepsilon^2.$$

Let us define $r_1, \ldots, r_t$ to be anti-derivatives of $l_1, \ldots, l_t$, i.e., $l_1 = r_1', \ldots l_t = r_t'$. Then for every $x \in \chi$, observe that

$$\begin{aligned}
&|U_{\theta,m}(x) - U_{\theta_1,r_1}(x)| \\
&\le |U_{\theta,m}(x) - U_{\theta,r_1}(x)| + |U_{\theta,r_1}(x) - U_{\theta_1,r_1}(x)| \\
&\le |m'(\theta^\top x)K_1(x;\theta) - r_1'(\theta^\top x)K_1(x;\theta)| + |r_1'(\theta^\top x)K_1(x;\theta) - r_1'(\theta_1^\top x)K_1(x;\theta_1)| \\
&\le |m'(\theta^\top x) - r_1'(\theta^\top x)||K_1(x;\theta)| + |r_1'(\theta^\top x)K_1(x;\theta) - r_1'(\theta_1^\top x)K_1(x;\theta)| \\
&\qquad + |r_1'(\theta_1^\top x)K_1(x;\theta) - r_1'(\theta_1^\top x)K_1(x;\theta_1)| \\
&\le \varepsilon|K_1(x;\theta)| + |r_1'(\theta^\top x) - r_1'(\theta_1^\top x)||K_1(x;\theta)| + \|r_1'\|_\infty |K_1(x;\theta) - K_1(x;\theta_1)| \\
&\le \varepsilon\|K_1(\cdot;\theta)\|_{2,\infty} + J(r_1)|\theta - \theta_1|^{1/2}T^{1/2}\|K_1(\cdot;\theta)\|_{2,\infty} + M_1(2T + \bar{M})|\theta - \theta_1| \lesssim \varepsilon.
\end{aligned}$$

Here the last inequality follows from **(A2)**, (89), and Lemma 4. Thus, $\{U_{\theta_i,r_j} - U_{\theta_0,m_0}\}_{1\le i\le t, 1\le j\le s}$ form an (constant multiple of) ε-cover (with respect to $\|\cdot\|_{2,\infty}$ norm) of $\mathcal{W}^*_{M_1,M_2,M_3}$. Moreover, as $N_{[\,]}(\varepsilon, \mathcal{W}^*_{M_1,M_2,M_3}, \|\cdot\|_{2,P_{\theta_0,m_0}}) \lesssim N(\varepsilon, \mathcal{W}^*_{M_1,M_2,M_3}, \|\cdot\|_{2,\infty})$ and $\mathcal{W}_{M_1,M_2,M_3}(n) \subset \mathcal{W}^*_{M_1,M_2,M_3}$, for every $n \in \mathbb{N}$, we have

$$N_{[\,]}(\varepsilon, \mathcal{W}_{M_1,M_2,M_3}(n), \|\cdot\|_{2,P_{\theta_0,m_0}}) \lesssim N_{[\,]}(\varepsilon, \mathcal{W}^*_{M_1,M_2,M_3}, \|\cdot\|_{2,P_{\theta_0,m_0}}) \lesssim c\exp(c/\varepsilon)\varepsilon^{-4d}.$$

Observe that if $[\hbar_1, \hbar_2]$ is a bracket for $U_{\theta,m} - U_{\theta_0,m_0}$, then $[\hbar_1\epsilon^+ - \hbar_2\epsilon^-, \hbar_2\epsilon^+ - \hbar_1\epsilon^-]$ is a bracket (here the ordering is coordinate-wise) for $\epsilon(U_{\theta,m} - U_{\theta_0,m_0})$. Therefore, we have

$$N_{[\,]}\big(\varepsilon\|\sigma(\cdot)\|_\infty, \{\epsilon f : f \in \mathcal{W}_{M_1,M_2,M_3}(n)\}, \|\cdot\|_{2,P_{\theta_0,m_0}}\big) \le c\exp(c/\varepsilon)\varepsilon^{-2(d-1)}. \tag{94}$$

Now we prove (30). As in (28), we have

$$\begin{aligned}
&\mathbb{P}(\big|\mathbb{G}_n\big[\epsilon\big(U_{\hat\theta,\hat m}(X) - U_{\theta_0,m_0}(X)\big)\big]\big| > \delta) \\
&\le \mathbb{P}\Big(\sup_{(m,\theta)\in\mathcal{C}_{M_1,M_2,M_3}(n)} \big|\mathbb{G}_n\big[\epsilon\big(U_{\theta,m}(X) - U_{\theta_0,m_0}(X)\big)\big]\big| > \delta\Big) + \mathbb{P}((\hat\theta, \hat m) \notin \mathcal{C}_{M_1,M_2,M_3}(n))
\end{aligned}$$

By discussion similar to those after Theorem 7, we only need to show that for every fixed M_1, M_2, and M_3, we have

$$\mathbb{P}\Big(\sup_{f\in\mathcal{W}_{M_1,M_2,M_3}(n)} |\mathbb{G}_n\epsilon f| > \delta\Big) \to 0,$$

as $n \to 0$. Note that by (94), for $\gamma > 0$ we have

$$J_{[\,]}\Big(\gamma, \{\epsilon f : f \in \mathcal{W}_{M_1,M_2,M_3}(n)\}, \|\cdot\|_{2,P_{\theta_0,m_0}}\Big) \lesssim \gamma^{\frac{1}{2}}.$$

By arguments similar to (92) and (93), we have

$$\begin{aligned}
\mathbb{P}\Big(\sup_{f\in\mathcal{W}_{M_1,M_2,M_3}(n)} |\mathbb{G}_n\epsilon f| > \delta\Big) &\lesssim \delta^{-1}\mathbb{E}\Big(\sup_{f\in\mathcal{W}_{M_1,M_2,M_3}(n)} |\mathbb{G}_n\epsilon f|\Big) \\
&\lesssim J_{[\,]}\Big(P_{\theta_0,m_0}\big(|\epsilon^2|W^2_{M_1,M_2,M_3}(n)\big)^{\frac{1}{2}}, \mathcal{W}_{M_1,M_2,M_3}(n), L_2(P_{\theta_0,m_0})\Big) \\
&\lesssim \Big[\hat\lambda_n^{1/4} + a_n^{-1} + \hat\lambda_n^{1/2}\Big]^{1/2} \to 0, \; as\, n \to \infty.
\end{aligned}$$

References

Antoniadis, A., G. Grégoire, and I. W. McKeague (2004). Bayesian estimation in single-index models. *Statist. Sinica 14*(4), 1147–1164.

Bickel, P. J., C. A. J. Klaassen, Y. Ritov, and J. A. Wellner (1993). *Efficient and adaptive estimation for semiparametric models*. Springer-Verlag, New York.

Carroll, R. J., J. Fan, I. Gijbels, and M. P. Wand (1997). Generalized partially linear single-index models. *J. Amer. Statist. Assoc. 92*(438), 477–489.

Cui, X., W. K. Härdle, and L. Zhu (2011). The EFM approach for single-index models. *Ann. Statist. 39*(3), 1658–1688.

Delecroix, M., M. Hristache, and V. Patilea (2006). On semiparametric M-estimation in single-index regression. *J. Statist. Plann. Inference 136*(3), 730–769.

Green, P. J. and B. W. Silverman (1994). *Nonparametric regression and generalized linear models*, Volume 58 of *Monogr. Statist. Appl. Probab.* Chapman & Hall, London.

Györfi, L., M. Kohler, A. Krzyżak, and H. Walk (2002). *A distribution-free theory of nonparametric regression*. Springer Ser. Statist. Springer-Verlag, New York.

Härdle, W., P. Hall, and H. Ichimura (1993). Optimal smoothing in single-index models. *Ann. Statist. 21*(1), 157–178.

Härdle, W. and H. Liang (2007). Partially linear models. In *Statistical methods for biostatistics and related fields*, pp. 87–103. Springer, Berlin.

Horowitz, J. L. (1998). *Semiparametric methods in econometrics*, Volume 131 of *Lecture Notes in Statistics*. Springer-Verlag, New York.

Horowitz, J. L. (2009). *Semiparametric and nonparametric methods in econometrics*. Springer Ser. Statist. Springer, New York.

Hristache, M., A. Juditsky, and V. Spokoiny (2001). Direct estimation of the index coefficient in a single-index model. *Ann. Statist. 29*(3), 595–623.

Huang, J. (1996). Efficient estimation for the proportional hazards model with interval censoring. *Ann. Statist. 24*(2), 540–568.

Ichimura, H. (1993). Semiparametric least squares (SLS) and weighted SLS estimation of single-index models. *J. Econometrics 58*(1-2), 71–120.

Kalaj, D., M. Vuorinen, and G. Wang (2016). On quasi-inversions. *Monatsh. Math. 180*(4), 785–813.

Klaassen, C. A. J. (1987). Consistent estimation of the influence function of locally asymptotically linear estimators. *Ann. Statist. 15*(4), 1548–1562.

Kuchibhotla, A. K. and R. K. Patra (2016). *simest: Single Index Model Estimation with Constraints on Link Function*. R package version 0.6.

Li, G. and C.-H. Zhang (1998). Linear regression with interval censored data. *Ann. Statist. 26*(4), 1306–1327.

Li, K.-C. (1991). Sliced inverse regression for dimension reduction. *J. Amer. Statist. Assoc. 86*(414), 316–342. With discussion and a rejoinder by the author.

Li, K.-C. and N. Duan (1989). Regression analysis under link violation. *Ann. Statist. 17*(3), 1009–1052.

Li, Q. and J. S. Racine (2007). *Nonparametric econometrics*. Princeton University Press, Princeton, NJ.

Li, W. and V. Patilea (2015). A new inference approach for single-index models.

Ma, Y. and L. Zhu (2013). Doubly robust and efficient estimators for heteroscedastic partially linear single-index models allowing high dimensional covariates. *J. R. Stat. Soc. Ser. B. Stat. Methodol. 75*(2), 305–322.

Mammen, E. and S. van de Geer (1997). Penalized quasi-likelihood estimation in partial linear models. *Ann. Statist. 25*(3), 1014–1035.

Meyer, C. (2000). *Matrix analysis and applied linear algebra*. SIAM.

Murphy, S. A. and A. W. van der Vaart (2000). On profile likelihood. *J. Amer. Statist. Assoc. 95*(450), 449–485. With comments and a rejoinder by the authors.

Murphy, S. A., A. W. van der Vaart, and J. A. Wellner (1999). Current status regression. *Math. Methods Statist. 8*(3), 407–425.

Newey, W. K. (1990). Semiparametric efficiency bounds. *J. Appl. Econometrics 5*(2), 99–135.

Newey, W. K. and T. M. Stoker (1993). Efficiency of weighted average derivative estimators and index models. *Econometrica 61*(5), 1199–1223.

Oden, J. T. and J. N. Reddy (2012). *An introduction to the mathematical theory of finite elements*. Courier Dover Publications.

Patra, R. K., E. Seijo, and B. Sen (2015). A consistent bootstrap procedure for the maximum score estimator. *J. Econometrics (revision resubmitted)*.

Pollard, D. (1990). *Empirical processes: theory and applications*. NSF-CBMS Regional Conference Series in Probability and Statistics, 2. Institute of Mathematical Statistics, Hayward, CA; American Statistical Association, Alexandria, VA.

Powell, J. L., J. H. Stock, and T. M. Stoker (1989). Semiparametric estimation of index coefficients. *Econometrica 57*(6), 1403–1430.

Ruppert, D., M. P. Wand, and R. J. Carroll (2003). *Semiparametric regression*, Volume 12 of *Camb. Ser. Stat. Probab. Math.* Cambridge University Press, Cambridge.

Stoker, T. M. (1986). Consistent estimation of scaled coefficients. *Econometrica 54*(6), 1461–1481.

Tsiatis, A. A. (2006). *Semiparametric theory and missing data*. Springer Ser. Statist. Springer, New York.

van de Geer, S. (1990). Estimating a regression function. *Ann. Statist. 18*(2), 907–924.

van de Geer, S. A. (2000). *Applications of empirical process theory*, Volume 6 of *Camb. Ser. Stat. Probab. Math.* Cambridge University Press, Cambridge.

van der Vaart, A. (2002). Semiparametric statistics. In *Lect. Probab. Theory Stat. St.-Flour, 1999*, Volume 1781 of *Lecture Notes in Math.*, pp. 331–457.

van der Vaart, A. W. (1998). *Asymptotic statistics*, Volume 3 of *Camb. Ser. Stat. Probab. Math.* Cambridge University Press, Cambridge.

Wahba, G. (1990). *Spline models for observational data*, Volume 59 of *CBMS-NSF Regional Conf. Ser. in Appl. Math.*

Xia, Y. (2006). Asymptotic distributions for two estimators of the single-index model. *Econometric Theory 22*(6), 1112–1137.

Xia, Y., H. Tong, W. Li, and L.-X. Zhu (2002). An adaptive estimation of dimension reduction space. *J. R. Stat. Soc. Ser. B. Stat. Methodol. 64*(3), 363–410.

Yu, Y. and D. Ruppert (2002). Penalized spline estimation for partially linear single-index models.

J. Amer. Statist. Assoc. 97(460), 1042–1054.

Zhou, J. and X. He (2008). Dimension reduction based on constrained canonical correlation and variable filtering. *Ann. Statist. 36*(4), 1649–1668.