SEMIPARAMETRIC INFERENCE WITH SHAPE CONSTRAINTS

ROHIT KUMAR PATRA

Submitted in partial fulfillment of the
requirements for the degree
of Doctor of Philosophy
in the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2016

ABSTRACT

SEMIPARAMETRIC INFERENCE WITH SHAPE CONSTRAINTS

ROHIT KUMAR PATRA

This thesis deals with estimation and inference in two semiparametric problems: a two-component mixture model and a single index regression model.

For the two-component mixture model, we assume that the distribution of one component is known and develop methods for estimating the mixing proportion and the unknown distribution using ideas from shape restricted function estimation. We establish the consistency of our estimators. We find the rate of convergence and the asymptotic limit of our estimator for the mixing proportion. Furthermore, we develop a completely automated distribution-free honest finite sample lower confidence bound for the mixing proportion. We compare the proposed estimators, which are easily implementable, with some of the existing procedures through simulation studies and analyse two data sets, one arising from an application in astronomy and the other from a microarray experiment.

For the single index model, we consider estimation of the unknown link function and the finite dimensional index parameter. We study the problem when the true link function is assumed to be: (1) smooth or (2) convex. When the link function is just assumed to be smooth, in contrast to standard kernel based methods, we use smoothing splines to estimate the link function. We prove the consistency and find the rates of convergence of the proposed estimators. We establish $n^{-1/2}$-rate of convergence and the semiparametric efficiency of the parametric component under mild assumptions. When the link function is assumed to be convex, we propose a shape constrained penalized least squares estimator and a Lipschitz constrained least squares estimator for the unknown quantities. We prove the consistency and find the rates of convergence for both estimators. For the shape constrained penalized least squares estimator, we establish $n^{-1/2}$-rate of convergence and the semiparametric efficiency of the parametric component under mild assumptions and

conjecture that the parametric component of the Lipschitz constrained least squares estimator is semiparametrically efficient. We develop the R package `simest` that can be used (to compute the proposed estimators) even for moderately large dimensions.

Table of Contents

List of Figures

List of Tables

Acknowledgments

First of all, I would like to thank my advisor Bodhisattva Sen for his continuous support and encouragement. Additionally, I would also like to thank Richard A. Davis, Zhiliang Ying, Moulinath Banerjee, and Kathryn V. Johnston for their helpful discussions and comments. Last but not the least, I am grateful to all my friends for making my experience at Columbia thoroughly enjoyable.

To my parents and my sister.

Chapter 1

Introduction and brief overview

In this thesis we study two models: a two-component mixture model and a single index regression model. Both the models are semiparametric in nature — involve an unknown finite dimensional parameter and an unknown infinite dimensional parameter. We propose consistent estimators for both the finite and the infinite dimensional parameters in the above semiparametric models and study their rates of convergence and, whenever possible, limiting distributions. We assume that the infinite dimensional parameter is smooth (Chapter 3) or satisfies shape constraints such as monotonicity (Chapter 2) or convexity (Chapter 4).

The models considered have applications in genomics (multiple testing problems), economics (utility and production function estimation and binary response models), and astronomy, among other fields. In the following, we briefly introduce the models and give a summary of the results to come.

1.1 Two-component mixture model

Consider a mixture model with two components, i.e.,

$$F(x) = \alpha F_s(x) + (1 - \alpha)F_b(x), \tag{1.1}$$

where the cumulative distribution function (CDF) F_b is known, but the mixing proportion $\alpha \in [0, 1]$ and the CDF F_s ($\neq F_b$) are unknown. Given a random sample from F, we wish to (nonparametrically) estimate F_s and the parameter α.

This model appears in many contexts. In multiple testing problems (microarray analysis, neuroimaging) the p-values, obtained from the numerous (independent) hypotheses tests, are uniformly distributed on [0,1], under H_0, while their distribution associated with H_1 is unknown; see e.g., [Efron, 2010] and [Robin *et al.*, 2007]. Translated to the setting of (1.1), F_b is the uniform distribution and the goal is to estimate the proportion of false null hypotheses α and the distribution of the p-values under the alternative F_s. A reliable estimator of α is important when we want to assess or control multiple error rates, such as the false discovery rate (see [Benjamini and Hochberg, 1995]) or the local false discovery rate; see [Efron, 2010].

More generally, this model arises in contamination problems. Where reasonable assumptions can be made about the contaminant distribution F_b. For example, in astronomy when observing some variable(s) of interest (e.g., metallicity, radial velocity) of stars in a distant galaxy, foreground stars from the Milky Way, in the field of view, contaminate the sample; the galaxy ("signal") stars can be difficult to distinguish from the foreground stars as we can only observe the stereographic projections and not the three dimensional position of the stars (see [Walker *et al.*, 2009]). Known physical models for the foreground stars help us constrain F_b, and the focus is on estimating the distribution of the variable for the signal stars, i.e., F_s. We discuss such an application in more detail in Section 2.9.2.

Most of the existing procedures for estimation in this mixture model assume parametric and/or nonparametric restrictions on the unknown F_s. [Cohen, 1967], [Lindsay, 1983], [Day, 1969], [Lindsay and Basak, 1993], and [Quandt and Ramsey, 1978] assume that F_s belongs to certain parametric models. In multiple testing literature, [Storey, 2002], [Genovese and Wasserman, 2004], [Meinshausen and Rice, 2006], [Meinshausen and Bühlmann, 2005], [Celisse and Robin, 2010], and [Langaas *et al.*, 2005] proposed estimators of α_0 under certain nonparametric assumptions on F_s and its density. [Genovese and Wasserman, 2004] and [Meinshausen and Rice, 2006] also proposed confidence bounds for α.

In Chapter 2, we provide a methodology to estimate the mixing proportion and F_s, without assuming any constraint on the form of F_s. However, without any constraint on

F_s model (1.1) is not identifiable (a trivial solution occurs when $\alpha = 1$ and $F_s = F$). To handle the identifiability issue, we redefine the mixing proportion as

$$\alpha_0 := \inf \{\gamma \in (0, 1] : [F - (1 - \gamma)F_b]/\gamma \text{ is a CDF}\}.$$

Intuitively, this definition makes sure that the "signal" distribution F_s does not include any contribution from the known "background" F_b. A natural question is when is the model (1.1) identifiable. We study this issue in complete generality in Sections 2.2.2 and 2.11; see [Genovese and Wasserman, 2004] for a similar notion of identifiability when F_b is the uniform distribution and F has a density.

Suppose that we observe an independent and identically distributed (i.i.d.) sample $X_1, X_2, \ldots, X_n$ from F as in (1.1). If $\alpha \in (0, 1]$ were known, a naive estimator of F_s would be

$$\hat{F}^{\alpha}_{s,n} := \frac{\mathbb{F}_n - (1 - \alpha)F_b}{\alpha},$$

where $\mathbb{F}_n$ is the empirical CDF of the observed sample, i.e., $\mathbb{F}_n(x) = \sum_{i=1}^{n} \mathbf{1}\{X_i \le x\}/n$. It is easy to see that $\hat{F}^{\alpha}_{s,n}$ does not necessarily satisfy the basic requirements of a CDF: $\hat{F}^{\alpha}_{s,n}$ need not be non-decreasing or lie between 0 and 1. One can obtain an improved estimator by constraining $\hat{F}^{\alpha}_{s,n}$ to be a CDF. We propose the following modification:

$$\check{F}^{\alpha}_{s,n} := \underset{\{W | W \text{ is a CDF}\}}{\arg\min} \frac{1}{n} \sum_{i=1}^{n} \{W(X_i) - \hat{F}^{\alpha}_{s,n}(X_i)\}^2,$$

where the minimization is over the class of all distribution functions. Intuitively, $\check{F}^{\alpha}_{s,n}$ is the "closest" distribution function to $\hat{F}^{\alpha}_{s,n}$; see Section 2.2.1 for a fast algorithm for computing $\check{F}^{\alpha}_{s,n}$.

With the modified estimator in mind, let us turn our attention to model (1.1) with unknown α_0. Here our goal is to estimate α_0. In the estimation procedure described below, the "distance" between $\check{F}^{\gamma}_{s,n}$ and $\hat{F}^{\gamma}_{s,n}$ as γ varies between $[0, 1]$ plays a crucial role. As illustrated in Figure 1.1, the "distance" between $\check{F}^{\gamma}_{s,n}$ and $\hat{F}^{\gamma}_{s,n}$ is large when $\gamma < \alpha_0$ and small when $\gamma > \alpha_0$. In Lemma 6, we formalize this. We show that

$$\gamma d_n(\hat{F}^{\gamma}_{s,n}, \check{F}^{\gamma}_{s,n}) \xrightarrow{a.s.} \begin{cases} 0, & \gamma - \alpha_0 \ge 0, \\ > 0, & \gamma - \alpha_0 < 0, \end{cases}$$

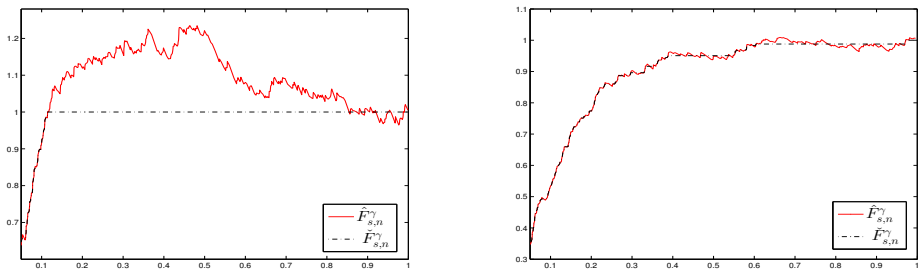

Figure 1.1: Plots of $\hat{F}^{\gamma}_{s,n}$ (in dashed red) $\check{F}^{\gamma}_{s,n}$ (in dot-dashed black) when $n = 300$, $F_b(x) = \Phi(x)$, $\alpha_0 = 0.3$, and $F(x) = .3\Phi(x-2) + .7\Phi(x)$. Left panel: $\gamma = 0.2$; right panel: $\gamma = 0.4$.

as $n \to \infty$, where d_n stands for the $L_2(\mathbb{F}_n)$ distance, i.e., if $g, h : \mathbb{R} \to \mathbb{R}$ are two functions, then $d_n^2(g,h) = \int \{g(x) - h(x)\}^2 \, d\mathbb{F}_n(x)$. We show that $\gamma \mapsto \gamma d_n(\hat{F}^{\gamma}_{s,n}, \check{F}^{\gamma}_{s,n})$ is a decreasing convex function; see Section 2.3. These two observations, lead to the following estimator of α_0 :

$$\hat{\alpha}_0^{c_n} := \inf \left\{ \gamma \in (0,1] : \gamma d_n(\hat{F}^{\gamma}_{s,n}, \check{F}^{\gamma}_{s,n}) \leq \frac{c_n}{\sqrt{n}} \right\},$$

where c_n is sequence of constants. In Figure 1.2, we plot $\gamma \mapsto \gamma d_n(\hat{F}^{\gamma}_{s,n}, \check{F}^{\gamma}_{s,n})$ as n increases from 2000 to 25000.

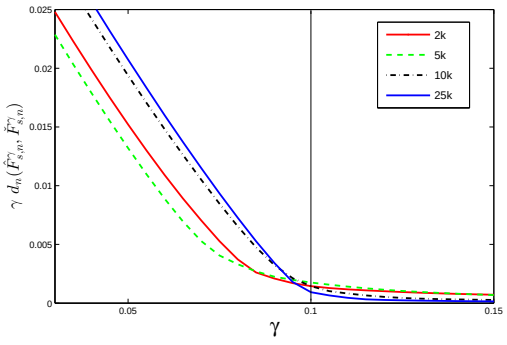

Figure 1.2: Plot of $\gamma \; d_n(\hat{F}^{\gamma}_{s,n}, \check{F}^{\gamma}_{s,n})$ when $F_b(x) = \Phi(x)$, $F_s(x) = \Phi(x-2)$, $\alpha_0 = .1$, and $F(x) = 0.1\Phi(x-2) + .9\Phi(x)$ for $n = 2000, 5000, 10000$, and 25000.

We show that if $c_n \to \infty$ and $c_n = o(n^{1/4})$ then

$$\frac{\sqrt{n}}{c_n}(\hat{\alpha}_0^{c_n} - \alpha_0) \xrightarrow{P} c,$$

where $c < 0$ is a constant that depends only on α_0, F and F_b; see Theorem 4. Observe that the rate of convergence of $\hat{\alpha}_0^{c_n}$ can be made arbitrarily close to $\sqrt{n}$ by choosing an appropriate sequence c_n.[1] We also study the effect c_n on the finite sample performance $\hat{\alpha}_0^{c_n}$ and give recommendations in see Section 2.8.2. We also discuss a tuning parameter free estimator of α_0 in Section 2.5.

A the natural question is weather we can obtain a confidence interval for α_0. As we have a degenerate limit distribution for $\hat{\alpha}_0^{c_n}$, the developed limit theory is of little practical use. Instead, we propose a finite sample (honest) lower confidence bound $\hat{\alpha}_L$ for α_0:

$$\hat{\alpha}_L := \inf\left\{\gamma \in (0,1] : \gamma d_n(\hat{F}_{s,n}^{\gamma}, \check{F}_{s,n}^{\gamma}) \leq \frac{H_n^{-1}(1-\beta)}{\sqrt{n}}\right\},$$

where H_n is the CDF of a distribution-free random variable (does not depend on any of F_s, F_b, or α_0); see Theorem 5 for the specific form of H_n. In Theorem 5, we prove that $\hat{\alpha}_L$ is an honest finite sample lower confidence bound for the mixing proportion α_0, i.e.,

$$\mathbb{P}(\alpha_0 \geq \hat{\alpha}_L) \geq 1 - \beta,$$

for a specified confidence level $(1-\beta)$ $(0 < \beta < 1)$, that is valid for any n. We believe that this is the first distribution-free lower confidence bound for α_0 that is also tuning parameter-free. Furthermore if $\alpha_0 = 0$, then $P(\hat{\alpha}_L = 0) = 1 - \beta$, i.e., it is an exact lower confidence bound.

In many scenarios it is desirable to estimate F_s, the distribution of the signal. When the model is identifiable, we propose a tuning parameter free uniformly consistent estimator of F_s; see Theorem 8. In multiple testing problems, an consistent estimate of f_s (the density of F_s) is required to control the local false discovery rate. However, obtaining a nonparametric estimator of f_s can be difficult as it requires smoothing and

[1] If F_s (defined on $[0,1]$) has a density f_s that vanishes on set of points of Lebesgue measure zero and $F_b(x) = x, \forall x \in [0,1]$, then [Nguyen and Matias, 2013] conjecture that no $\sqrt{n}$-consistent estimator of α_0 can have have finite limiting variance.

usually involves the choice of tuning parameter(s) (e.g., smoothing bandwidths). When f_s is known to be non-increasing (which is a natural assumption for the density of the p-values under the alternative), we find a consistent tuning parameter free estimator of f_s; see Theorem 9.

In Section 2.9, we apply the proposed methodology to two real data examples. The first one is a dataset arising from a microarray experiment. We observe genetic expression of 6033 genes for 50 control subjects and 52 prostate cancer patients. To test for the significance of each gene, we perform a two-sample t-test. The goal of the study is to estimate the proportion of the "interesting" genes (genes with different expression levels in the cancer patients and control subjects) based on the t-statistic values and to estimate the density f_s of the p-values under the alternative; also see [Efron, 2010] for extensive study of this dataset. The second dataset is from the astronomy example discussed in the beginning of this section. We observe radial velocities of 1266 stars from the Carina (a dwarf spheroidal galaxy) contaminated with Milky Way stars in the field of view. The distribution F_b of the the radial velocities of the contaminating stars from the Milky Way in the field of view is known from the Besancon Milky Way model; see [Robin *et al.*, 2003]. We give lower bound for the proportion of stars that belong to Carina and estimate the distribution of their radial velocities.

1.2 Single index models

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 flexible modeling, the price for this flexibility can be high for two main reasons: the estimation precision decreases rapidly as d increases ("curse of dimensionality"; see [Stone, 1980]) and the estimator can be hard to interpret when $d > 1$.

A natural restriction of the nonparametric model that avoids the curse of dimension-

ality while still retaining 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

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

where $X^\top\theta_0$ 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 models; 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].

Formally, we consider the model

$$Y = m_0(\theta_0^\top X) + \epsilon, \quad \mathbb{E}(\epsilon|X) = 0, \quad \text{almost every } X, \tag{1.2}$$

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 error. We assume that both m_0 and θ_0 are unknown and are the parameters of interest. For identifiability of the model we assume that the first coordinate of θ_0 is non-zero and

$$\theta_0 \in \Theta := \{\eta_0 \in \mathbb{R}^d : |\eta_0| = 1 \text{ and } \eta_{0,1} \geq 0\} \subset S^{d-1},$$

where $\eta_{0,1}$ is the first coordinate of η_0, $|\cdot|$ 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]), 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]), and dimension reduction techniques, e.g., sliced inverse regression (see [Li and Duan, 1989] and [Li, 1991]). In direct methods, one tries to estimate θ_0 directly 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 Chapter 3, we consider estimation of (θ_0, m_0) using smoothing splines ([Wahba, 1990]) when m_0 is assumed to be smooth, while in Chapter 4, we consider estimation of (θ_0, m_0) when m_0 is convex. In Sections 1.2.1 and 1.2.2 below, we motivate the two different constraints and give an overview of the proposed estimation procedures.

1.2.1 Smooth single index models

In the last few decades various approaches have been proposed in the statistical literature for estimation in the smooth single index model. [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 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].

All this motivates the use of smoothing splines for estimation in the smooth single index model. Smoothing splines avoid the choice of number of knots and their placement — the number of knots increase to infinity with sample size. Let $\{(x_i, y_i)\}_{1 \le i \le n}$ denote

an i.i.d. sample from (1.2). We propose the following estimator of (m_0, θ_0):

$$(\hat{m}, \hat{\theta}) := \underset{(m,\theta)\in\mathcal{S}\times\Theta}{\arg\min} \left\{ \frac{1}{n}\sum_{i=1}^{n} \left(y_i - m(\theta^\top x_i)\right)^2 + \hat{\lambda}_n^2 \int |m''(t)|^2 dt \right\}, \tag{1.3}$$

where the minimization is over the class

$$\mathcal{S} := \{m : \mathbb{R} \to \mathbb{R} |\, m' \text{ is absolutely continuous}\} \tag{1.4}$$

and the positive half sphere Θ. Here $\hat{\lambda}_n$ is known as the smoothing parameter — high values of $|\hat{\lambda}_n|$ lead to smoother estimators of m_0. To the best of our knowledge, this is the first work that uses smoothing splines in the single index paradigm, under (only) smoothness constraints.

If θ_0 is known and m_0 (unknown) is assumed to be smooth, (1.3) reduces to the familiar penalized least squares problem resulting in a smoothing spline estimator; see [Wahba, 1990] and [Green and Silverman, 1994]. For each $\theta \in \Theta$, define

$$\hat{m}_{\theta,\mathcal{S}} := \underset{m\in\mathcal{S}}{\arg\min}\, \mathcal{L}_n(m, \theta, \hat{\lambda}_n), \tag{1.5}$$

where

$$\mathcal{L}_n(m, \theta, \hat{\lambda}_n) := \frac{1}{n}\sum_{i=1}^{n} \left(y_i - m(\theta^\top x_i)\right)^2 + \hat{\lambda}_n^2 \int |m''(t)|^2 dt.$$

The above minimization is a convex problem and there are fast and efficient algorithms to compute $\hat{m}_{\theta,\mathcal{S}}$; see [Green and Silverman, 1994]. Now, observe that

$$\hat{\theta} = \underset{\theta\in\Theta}{\arg\min}\, \mathcal{L}_n(\hat{m}_{\theta,\mathcal{S}}, \theta, \hat{\lambda}_n).$$

In the R package `simest` ([Kuchibhotla and Patra, 2016]) we have implemented a fast algorithm for computing the estimator $(\hat{\theta}, \hat{m})$ for moderately large dimensions ($d \approx 100$).

To study the theoretical properties of the estimators, along with some distributional assumptions on X, we assume that $\int \{m_0''(t)\}^2 dt < \infty$ and that the errors are sub-Gaussian. If $\hat{\lambda}_n$ goes to 0 at a rate faster than $n^{-1/4}$, but not faster than $n^{-2/5}$, then we show that $\hat{m}$ and $\hat{\theta}$ are consistent estimators of m_0 and θ_0, respectively. Formally, in Theorems 12–14, we show that

$$\|\hat{m}\circ\hat{\theta} - m_0\circ\theta_0\| = O_p(\hat{\lambda}_n), \qquad |\hat{\theta} - \theta_0| = O_p(\hat{\lambda}_n),$$
$$\|\hat{m}\circ\theta_0 - m_0\circ\theta_0\| = O_p(\hat{\lambda}_n),$$

where for any function $g : \mathbb{R}^d \to \mathbb{R}$, $\|g\|^2 := \int g^2 dP_X$, $(m \circ \theta)(x) = m(\theta^\top x)$, and P_X denotes the distribution of X. If we choose $\hat{\lambda}_n = c\, n^{-2/5}$ $(c > 0)$, then both the prediction error for (1.2) and estimation error of $\hat{m}$ are of the optimal order ([Stone, 1980]), i.e.,

$$\|\hat{m} \circ \hat{\theta} - m_0 \circ \theta_0\| = O_p(n^{-2/5}) \quad \text{and} \quad \|\hat{m} \circ \theta_0 - m_0 \circ \theta_0\| = O_p(n^{-2/5}).$$

An obvious question to ask now is how good is the estimator $\hat{\theta}$? Efficiency bounds for the finite dimensional parameter in a semiparametric model give a standard against which any "regular" estimator ([Bickel *et al.*, 1993]) of the finite dimensional parameter can be measured. Efficiency bounds quantify the loss of efficiency that can result from a semiparametric, rather than a parametric, approach. [Stein, 1956] gives the following intuitive way to define the efficiency bounds in semiparametric models. One could imagine that the data are generated by a parametric model that satisfies the semiparametric assumptions and contains the truth. Such a model is often referred to as a parametric submodel. Then any asymptotically normal $\sqrt{n}$-consistent estimator (in the semiparametric model) would have to satisfy the Cramér-Rao lower bound for each of the parametric submodels. Thus one cannot hope to have an estimator with asymptotic variance smaller than the supremum of the Cramér-Rao lower bounds corresponding to all the parametric submodels. This supremum is known as the efficiency bound for a semiparametric model; see Section 3.4.1 for the calculation of the efficiency bound for (1.3).

Under some regularity conditions, we show that $\hat{\theta}$ is an asymptotically normal $\sqrt{n}$-consistent estimator of θ_0; see Section 3.4.2. Furthermore, in Theorem 15, we show that under homoscedastic errors

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0, H_{\theta_0} \tilde{I}^{-1}_{\theta_0, m_0} H^\top_{\theta_0}), \tag{1.6}$$

where $\tilde{I}_{\theta_0, m_0}$ is the efficient information matrix (the inverse of the efficient variance bound) and H_{θ_0} defines a local parametrization[2] of S^{d-1} at θ_0 and depends only on θ_0; see Theorem 16 and Section 3.8.1 for the structure of H_{θ_0}.

[2]Since $\hat{\theta}$ and θ_0 both lie on S^{d-1} (a $d-1$ dimensional manifold in $\mathbb{R}^d$) to study the limiting behavior of $\sqrt{n}(\hat{\theta} - \theta_0)$, we need to consider a local parametrization of S^{d-1} around θ_0 and the limiting variance covariance matrix of $\sqrt{n}(\hat{\theta} - \theta_0)$ is singular.

1.2.2 Convex single index models

In Chapter 4, we consider model (1.2) with the assumption that m_0 is shape constrained, namely, m_0 is convex. This assumption is motivated by the fact that in a wide range of applications in various fields the link function is known to be convex or concave. For example, in microeconomics, production functions are often supposed to be concave and component-wise non-decreasing; concavity indicates decreasing marginal returns. Also utility functions are often assumed to be concave (decreasing marginal utility); see [Li and Racine, 2007].

Shape constrained inference has a long history in the statistical literature dating back to the seminal papers [Grenander, 1956], [Hildreth, 1954], and [Brunk, 1955]. In the case of convex univariate regression the properties of the least squares estimator are well-studied; see [Hildreth, 1954; Hanson and Pledger, 1976; Groeneboom *et al.*, 2001; Dümbgen *et al.*, 2004], and [Guntuboyina and Sen, 2013] for consistency, local and global rates of convergence, and computational aspects of the least squares estimator.

A drawback of the convex shape constrained least squares estimator is that it is piecewise linear. Quite often in practice a smooth estimator is preferred. A natural way to obtain smooth convex is by penalizing the least squares loss with a penalty on the roughness of the convex function through the integrated squared second derivative (as in (1.3)). For univariate convex regression [Elfving and Andersson, 1988] provide a characterization for the constrained penalized least squares estimator while [Mammen and Thomas-Agnan, 1999] provide their rates of convergence. In the following section, we consider the penalized least squares estimator of (m_0, θ_0) and get back to the least squares estimator of (θ_0, m_0) in Section 1.2.2.2

1.2.2.1 Penalized least squares estimator

Let $\{(x_i, y_i)\}_{1\le i\le n}$ denote an i.i.d. sample from (1.2). We propose the following penalized least squares estimator of (m_0, θ_0):

$$(\hat{m}, \hat{\theta}) := \underset{(m,\theta)\in\mathcal{M}_S\times\Theta}{\arg\min} \left\{ \frac{1}{n}\sum_{i=1}^{n} \left(y_i - m(\theta^\top x_i)\right)^2 + \hat{\lambda}_n^2 \int |m''(t)|^2 dt \right\},$$

where the minimization is over the class of all absolutely continuously differentiable convex functions $\mathcal{M}_S$, i.e.,

$$\mathcal{M}_S := \mathcal{S} \cap \{m : \mathbb{R} \to \mathbb{R} | \, m \text{ is convex}\}$$

and $\mathcal{S}$ is defined in (1.4). For each $\theta \in \Theta$, we can define (similar to (1.5)) and compute (see [Elfving and Andersson, 1988]) the "profiled" estimate:

$$\hat{m}_{\theta,\mathcal{M}_S} := \underset{m \in \mathcal{M}_S}{\arg\min}\, \mathcal{L}_n(m, \theta, \hat{\lambda}_n)$$

and then obtain $\hat{\theta}$ by minimizing $\theta \mapsto \mathcal{L}_n(\hat{m}_{\theta,\mathcal{M}_S}, \theta, \hat{\lambda}_n)$ over Θ.

Under conditions similar to those in Section 1.2.1 and the additional assumption that m_0 is convex, in Theorems 21–23, we show that

$$\|\hat{m} \circ \hat{\theta} - m_0 \circ \theta_0\| = O_p(\hat{\lambda}_n), \qquad |\hat{\theta} - \theta_0| = O_p(\hat{\lambda}_n),$$
$$\|\hat{m} \circ \theta_0 - m_0 \circ \theta_0\| = O_p(\hat{\lambda}_n),$$

where $\|\cdot\|$ and $m \circ \theta$ are as defined in Section 1.2.1. In addition to the consistency of $(\hat{m}, \hat{\theta})$, we show that $\hat{m}'$ is a good estimator of m_0' (Theorem 24), i.e.,

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

Under further regularity conditions on m_0, we show that $\hat{\theta}$ is the "best" semiparametric estimator in the sense of the limiting variance. We show that the efficiency bound for any regular estimator ([Bickel *et al.*, 1993]) in model (1.2) when m_0 is convex and smooth is the same as that when m_0 is just assumed to be smooth; see Section 4.5.1. Formally, in Theorem 29, we show that

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0, H_{\theta_0} \tilde{I}^{-1}_{\theta_0, m_0} H^{\top}_{\theta_0}), \tag{1.7}$$

where H_{θ_0} and $\tilde{I}_{\theta_0,m_0}$ are the same as in Section 1.2.1. It must be noted that the penalized estimator discussed here is similar to the one proposed in [Murphy *et al.*, 1999] for the current status regression model where the link function is assumed to be monotone.

1.2.2.2 Least squares estimators

We now consider least squares estimators of (m_0, θ_0) in (1.2) under convexity constraint. First, consider the following least squares estimator,

$$(m^\dagger, \theta^\dagger) := \underset{(m,\theta)\in\mathcal{C}\times\Theta}{\arg\min} \frac{1}{n}\sum_{i=1}^{n}(y_i - m(\theta^\top x_i))^2,$$

where $\mathcal{C}$ is the class of all convex functions, i.e.,

$$\mathcal{C} := \{m : \mathbb{R} \to \mathbb{R} |\, m \text{ is convex}\}.$$

The above minimizer is well-defined and can be computed easily using a quadratic program (with linear constraints). However it is difficult to study the estimator theoretically. The difficulty can be attributed to the inconsistency of $m^\dagger$ at the "boundary" of its domain; it is well-known that shape constrained estimates can be inconsistent at the boundary ([Woodroofe and Sun, 1993]). In single index models the inconsistency of $m^\dagger$ at the boundary affects the estimation of θ_0 as θ_0 and m_0 are intertwined (as opposed to a partially linear model).

To fix the the "boundary problem" of $m^\dagger$, we propose a Lipschitz constrained least squares estimator for (m_0, θ_0), defined as

$$(\check{m}, \check{\theta}) := \underset{(m,\theta)\in\mathcal{M}_L\times\Theta}{\arg\min} \frac{1}{n}\sum_{i=1}^{n}(y_i - m(\theta^\top x_i))^2,$$

where

$$\mathcal{M}_L := \mathcal{C} \cap \{m : \mathbb{R} \to \mathbb{R} |\, m \text{ is uniformly Lipschitz with Lipschitz constant } L\}.$$

As in Sections 1.2.1 and 1.2.2.1, we can compute $(\check{m}, \check{\theta})$ by first computing the "profiled" loss and then by minimizing it over Θ, i.e.,

$$\check{\theta} = \underset{\theta\in\Theta}{\arg\min} \frac{1}{n}\sum_{i=1}^{n}(y_i - m_{\theta,\mathcal{M}_L}(\theta^\top x_i))^2,$$

where

$$m_{\theta,\mathcal{M}_L} := \underset{m\in\mathcal{M}_L}{\arg\min} \frac{1}{n}\sum_{i=1}^{n}(y_i - m(\theta^\top x_i))^2.$$

For each $\theta \in \Theta$, $m_{\theta,\mathcal{M}_L}$ is the solution of a quadratic program with linear constraints and can be computed easily. Even though the function $\theta \mapsto \frac{1}{n}\sum_{i=1}^{n}(y_i - m_{\theta,\mathcal{M}_L}(\theta^\top x_i))^2$

is not convex the algorithm implemented in `simest` performs well in finding the optima for moderately large dimensions ($d \approx 100$).

If the true link function m_0 is uniformly Lipschitz with Lipschitz constant L_0 and $L \geq L_0$ we show that $\check{m}$ and $\check{\theta}$ are consistent estimators of m_0 and θ_0, respectively, i.e.,

$$\|\check{m} \circ \check{\theta} - m_0 \circ \theta_0\| = O_p(n^{-2/5}), \qquad |\check{\theta} - \theta_0| = O_p(n^{-2/5}),$$
$$\|\check{m} \circ \theta_0 - m_0 \circ \theta_0\| = O_p(n^{-2/5});$$

see Theorems 25–27. Note that both the prediction error for (1.2) and estimation error of $\check{m}$ are of the optimal order ([Stone, 1980]) for convex function estimation. Moreover (under additional smoothness assumptions on m_0) we show that the right derivative of $\check{m}$ converges to the derivative of m_0, i.e.,

$$\|\check{m}' \circ \check{\theta} - m_0' \circ \check{\theta}\| = O_p(n^{-2/15});$$

see Theorem 28.

Under further regularity conditions on m_0, in Theorem 34, we show that $\check{\theta}$ is also semiparametrically efficient:

$$\sqrt{n}(\check{\theta} - \theta_0) \xrightarrow{d} N(0, H_{\theta_0} \tilde{I}^{-1}_{\theta_0, m_0} H^{\top}_{\theta_0}), \tag{1.8}$$

where H_{θ_0} and $\tilde{I}_{\theta_0, m_0}$ are the same as in Sections 1.2.1 and 1.2.2.1.

Observe that (1.6), (1.7), and (1.8) suggest that, asymptotically, convexity of m_0 does not help us in estimating θ_0. However, this should not discourage the use of convexity/concavity constraints in single index models for the following two reasons. First, the efficiency bounds are asymptotic in nature and they might not quantify the finite sample performance of the estimators. Secondly, shape constraints can be very useful as they may provide improved finite sample performance and interpretability of $\hat{m}$ or $\check{m}$ (and $\hat{\theta}$ or $\check{\theta}$), especially when the "signal-to-noise" ratio is low; see Figure 1.3 for such an illustration.

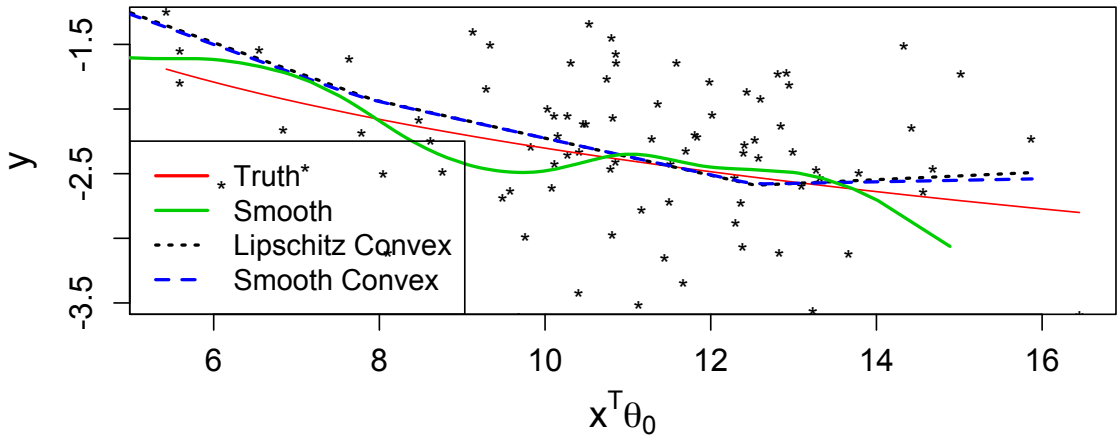

Figure 1.3: Plot of estimators of m_0 proposed in Sections 1.2.1, 1.2.2.1, and 1.2.2.2. Here we have 100 i.i.d. samples from $Y = -\log(\theta_0^\top X) + N(0, .5^2)$, where $X \sim$ Uniform$[2, 10]^3$ and $\theta_0 = (0.58, 0.58, 0.58)$. The estimates for θ_0: in Section 1.2.1 (just under smoothness) is $(0.86, 0.44, 0.25)$, in Section 1.2.2.1 (under smoothness and convexity) is $(0.64, 0.38, 0.66)$, and in Section 1.2.2.2 (just under convexity) is $(0.65, 0.38, 0.66)$.

Part I

Two-component Mixture Model

Chapter 2

Estimation of a Two-component Mixture Model with Applications to Multiple Testing[1]

We consider a two-component mixture model with one known component. We develop methods for estimating the mixing proportion and the unknown distribution nonparametrically, given i.i.d. data from the mixture model, using ideas from shape restricted function estimation. We establish the consistency of our estimators. We find the rate of convergence and asymptotic limit of the estimator for the mixing proportion. Completely automated distribution-free honest finite sample lower confidence bounds are developed for the mixing proportion. Connection to the problem of multiple testing is discussed. The identifiability of the model, and the estimation of the density of the unknown distribution are also addressed. We compare the proposed estimators, which are easily implementable, with some of the existing procedures through simulation studies and analyse two data sets, one arising from an application in astronomy and the other from a microarray experiment.

Keywords: Cramér-von Mises statistic, cross-validation, functional delta method, identifiability, local false discovery rate, lower confidence bound, microarray experiment,

[1] Joint work with Bodhisattva Sen.

projection operator, shape restricted function estimation.

2.1 Introduction

Consider a mixture model with two components, i.e.,

$$F(x) = \alpha F_s(x) + (1 - \alpha)F_b(x), \tag{2.1}$$

where the cumulative distribution function (CDF) F_b is known, but the mixing proportion $\alpha \in [0, 1]$ and the CDF F_s ($\neq F_b$) are unknown. Given a random sample from F, we wish to (nonparametrically) estimate F_s and the parameter α.

This model appears in many contexts. In multiple testing problems (microarray analysis, neuroimaging) the p-values, obtained from the numerous (independent) hypotheses tests, are uniformly distributed on [0,1], under H_0, while their distribution associated with H_1 is unknown; see e.g., [Efron, 2010] and [Robin *et al.*, 2007]. Translated to the setting of (2.1), F_b is the uniform distribution and the goal is to estimate the proportion of false null hypotheses α and the distribution of the p-values under the alternative. In addition, a reliable estimator of α is important when we want to assess or control multiple error rates, such as the false discovery rate of [Benjamini and Hochberg, 1995].

In contamination problems, the distribution F_b, for which reasonable assumptions can be made, may be contaminated by an arbitrary distribution F_s, yielding a sample drawn from F as in (2.1); see e.g., [McLachlan and Peel, 2000]. For example, in astronomy, such situations arise quite often: when observing some variable(s) of interest (e.g., metallicity, radial velocity) of stars in a distant galaxy, foreground stars from the Milky Way, in the field of view, contaminate the sample; the galaxy ("signal") stars can be difficult to distinguish from the foreground stars as we can only observe the stereographic projections and not the three dimensional position of the stars (see [Walker *et al.*, 2009]). Known physical models for the foreground stars help us constrain F_b, and the focus is on estimating the distribution of the variable for the signal stars, i.e., F_s. We discuss such an application in more detail in Section 2.9.2. Such problems also arise in High Energy physics where often the signature of new physics is evidence of a significant-looking peak

at some position on top of a rather smooth background distribution; see e.g., [Lyons, 2008].

Most of the previous work on this problem assume some constraint on the form of the unknown distribution F_s, e.g., it is commonly assumed that the distributions belong to certain parametric models, which lead to techniques based on maximum likelihood (see e.g., [Cohen, 1967] and [Lindsay, 1983]), minimum chi-square (see e.g., [Day, 1969]), method of moments (see e.g., [Lindsay and Basak, 1993]), and moment generating functions (see e.g., [Quandt and Ramsey, 1978]). [Bordes *et al.*, 2006] assume that both the components belong to an unknown symmetric location-shift family. [Jin, 2008] and [Cai and Jin, 2010] use empirical characteristic functions to estimate F_s under a semiparametric normal mixture model. In multiple testing, this problem has been addressed by various authors and different estimators and confidence bounds for α have been proposed in the literature under certain assumptions on F_s and its density, see e.g., [Storey, 2002], [Genovese and Wasserman, 2004], [Meinshausen and Rice, 2006], [Meinshausen and Bühlmann, 2005], [Celisse and Robin, 2010] and [Langaas *et al.*, 2005]. For the sake of brevity, we do not discuss the above references here but come back to this application in Section 2.7.

In this paper we provide a methodology to estimate α and F_s (nonparametrically), without assuming any constraint on the form of F_s. The main contributions of our paper can be summarised in the following.

- We investigate the identifiability of (2.1) in complete generality.
- When F is a continuous CDF, we develop an honest finite sample lower confidence bound for the mixing proportion α. We believe that this is the first attempt to construct a distribution-free lower confidence bound for α that is also tuning parameter-free.
- Two different estimators of α are proposed and studied. We derive the rate of convergence and asymptotic limit for one of the proposed estimators.
- A nonparametric estimator of F_s using ideas from shape restricted function estimation is proposed and its consistency is proved. Further, if F_s has a non-increasing

density f_s, we can also consistently estimate f_s.

The paper is organised as follows. In Section 2.2 we address the identifiability of the model given in (2.1). In Section 2.3 we propose an estimator of α and investigate its theoretical properties, including its consistency, rate of convergence and asymptotic limit. In Section 2.4 we develop a completely automated distribution-free honest finite sample lower confidence bound for α. As the performance of the estimator proposed in Section 2.3 depends on the choice of a tuning parameter, in Section 2.5 we study a tuning parameter-free heuristic estimator of α. We discuss the estimation of F_s and its density f_s in Section 2.6. Connection to the multiple testing problem is developed in Section 2.7. In Section 2.8 we compare the finite sample performance of our procedures, including a plug-in and cross-validated choice of the tuning parameter for the estimator proposed in Section 2.3, with other methods available in the literature through simulation studies, and provide a clear recommendation to the practitioner. Two real data examples, one arising in astronomy and the other from a microarray experiment, are analysed in Section 2.9. Appendix 2.14 gives the proofs of the results in the paper.

2.2 The model and identifiability

2.2.1 When α is known

Suppose that we observe an i.i.d. sample $X_1, X_2, \ldots, X_n$ from F as in (2.1). If $\alpha \in (0, 1]$ were known, a naive estimator of F_s would be

$$\hat{F}^{\alpha}_{s,n} = \frac{\mathbb{F}_n - (1-\alpha)F_b}{\alpha}, \tag{2.2}$$

where $\mathbb{F}_n$ is the empirical CDF of the observed sample, i.e., $\mathbb{F}_n(x) = \sum_{i=1}^{n} \mathbf{1}\{X_i \le x\}/n$. Although this estimator is consistent, it does not satisfy the basic requirements of a CDF: $\hat{F}^{\alpha}_{s,n}$ need not be non-decreasing or lie between 0 and 1. This naive estimator can be improved by imposing the known shape constraint of monotonicity. This can be accomplished by minimising

$$\int \{W(x) - \hat{F}^{\alpha}_{s,n}(x)\}^2 \, d\mathbb{F}_n(x) \equiv \frac{1}{n}\sum_{i=1}^{n} \{W(X_i) - \hat{F}^{\alpha}_{s,n}(X_i)\}^2 \tag{2.3}$$

over all CDFs W. Let $\check{F}^{\alpha}_{s,n}$ be a CDF that minimises (2.3). The above optimisation problem is the same as minimising $\|\boldsymbol{\theta} - \mathbf{V}\|^2$ over $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n) \in \Theta_{inc}$ where

$$\Theta_{inc} = \{\boldsymbol{\theta} \in \mathbb{R}^n : 0 \leq \theta_1 \leq \theta_2 \leq \ldots \leq \theta_n \leq 1\},$$

$\mathbf{V} = (V_1, V_2, \ldots, V_n)$, $V_i := \hat{F}^{\alpha}_{s,n}(X_{(i)})$, $i = 1, 2, \ldots, n$, $X_{(i)}$ being the i-th order statistic of the sample, and $\|\cdot\|$ denotes the usual Euclidean norm in $\mathbb{R}^n$. The estimator $\hat{\boldsymbol{\theta}}$ is uniquely defined by the projection theorem (see e.g., Proposition 2.2.1 on page 88 of [Bertsekas, 2003]); it is the Euclidean projection of $\mathbf{V}$ on the closed convex set $\Theta_{inc} \subset \mathbb{R}^n$. $\hat{\boldsymbol{\theta}}$ is related to $\check{F}^{\alpha}_{s,n}$ via $\check{F}^{\alpha}_{s,n}(X_{(i)}) = \hat{\theta}_i$, and can be easily computed using the pool-adjacent-violators algorithm (PAVA); see Section 1.2 of [Robertson *et al.*, 1988]. Thus, $\check{F}^{\alpha}_{s,n}$ is uniquely defined at the data points X_i, for all $i = 1, \ldots, n$, and can be defined on the entire real line by extending it to a piece-wise constant right continuous function with possible jumps only at the data points. The following result, derived easily from Chapter 1 of [Robertson *et al.*, 1988], characterises $\check{F}^{\alpha}_{s,n}$.

Lemma 1. *Let $\tilde{F}^{\alpha}_{s,n}$ be the isotonic regression (see e.g., page 4 of [Robertson et al., 1988]) of the set of points $\{\hat{F}^{\alpha}_{s,n}(X_{(i)})\}_{i=1}^{n}$. Then $\tilde{F}^{\alpha}_{s,n}$ is characterised as the right-hand slope of the greatest convex minorant of the set of points $\{i/n, \sum_{j=0}^{i} \hat{F}^{\alpha}_{s,n}(X_{(j)})\}_{i=0}^{n}$. The restriction of $\tilde{F}^{\alpha}_{s,n}$ to $[0,1]$, i.e., $\check{F}^{\alpha}_{s,n} = \min\{\max\{\tilde{F}^{\alpha}_{s,n}, 0\}, 1\}$, minimises (2.3) over all CDFs.*

Isotonic regression and the PAVA are very well studied in the statistical literature with many text-book length treatments; see e.g., [Robertson *et al.*, 1988] and [Barlow *et al.*, 1972]. If skillfully implemented, PAVA has a computational complexity of $O(n)$ (see [Grotzinger and Witzgall, 1984]).

2.2.2 Identifiability of F_s

When α is unknown, the problem is considerably harder; in fact, it is non-identifiable. If (2.1) holds for some F_b and α then the mixture model can be re-written as

$$F = (\alpha + \gamma)\left(\frac{\alpha}{\alpha + \gamma} F_s + \frac{\gamma}{\alpha + \gamma} F_b\right) + (1 - \alpha - \gamma) F_b,$$

for $0 \leq \gamma \leq 1-\alpha$, and the term $(\alpha F_s + \gamma F_b)/(\alpha+\gamma)$ can be thought of as the non-parametric component. A trivial solution occurs when we take $\alpha + \gamma = 1$, in which case (2.3) is minimised when $W = \mathbb{F}_n$. Hence, α is not uniquely defined. To handle the identifiability issue, we redefine the mixing proportion as

$$\alpha_0 := \inf \{\gamma \in (0,1] : [F - (1-\gamma)F_b]/\gamma \text{ is a CDF}\}. \tag{2.4}$$

Intuitively, this definition makes sure that the "signal" distribution F_s does not include any contribution from the known "background" F_b.

In this paper we consider the estimation of α_0 as defined in (2.4). Identifiability of mixture models has been discussed in many papers, but generally with parametric assumptions on the model. [Genovese and Wasserman, 2004] discuss identifiability when F_b is the uniform distribution and F has a density. [Hunter *et al.*, 2007] and [Bordes *et al.*, 2006] discuss identifiability for location shift mixtures of symmetric distributions. Most authors try to find conditions for the identifiability of their model, while we go a step further and quantify the non-identifiability by calculating α_0 and investigating the difference between α and α_0. In fact, most of our results are valid even when (2.1) is non-identifiable.

Suppose that we start with a fixed F_s, F_b and α satisfying (2.1). As seen from the above discussion we can only hope to estimate α_0, which, from its definition in (2.4), is smaller than α, i.e., $\alpha_0 \leq \alpha$. A natural question that arises now is: under what condition(s) can we guarantee that the problem is *identifiable*, i.e., $\alpha_0 = \alpha$? The following lemma gives the connection between α and α_0.

Lemma 2. *Let F be as in (2.1) and α_0 as defined in (2.4). Then*

$$\alpha_0 = \alpha - \sup \{0 \leq \epsilon \leq 1 : \alpha F_s - \epsilon F_b \text{ is a sub-CDF}\}, \tag{2.5}$$

where sub-CDF is a non-decreasing right-continuous function taking values between 0 and 1. In particular, $\alpha_0 < \alpha$ if and only if there exists $\epsilon \in (0,1)$ such that $\alpha F_s - \epsilon F_b$ is a sub-CDF. Furthermore, $\alpha_0 = 0$ if and only if $F = F_b$.

In the following we separately identify α_0 for any distribution, be it continuous or discrete or a mixture of the two, with a series of lemmas proved in Appendix 2.11. By

an application of the Lebesgue decomposition theorem in conjunction with the Jordan decomposition theorem (see page 142, Chapter V, Section $3a^*$ of [Feller, 1971]), we have that any CDF G can be uniquely represented as a weighted sum of a piecewise constant CDF $G^{(d)}$, an absolutely continuous CDF $G^{(a)}$, and a continuous but singular CDF $G^{(s)}$, i.e., $G = \eta_1 G^{(a)} + \eta_2 G^{(d)} + \eta_3 G^{(s)}$, where $\eta_i \geq 0$, for $i = 1, 2, 3$, and $\eta_1 + \eta_2 + \eta_3 = 1$. However, from a practical point of view, we can assume $\eta_3 = 0$, since singular functions almost never occur in practice; see e.g., [Parzen, 1960]. Hence, we may assume

$$G = \eta G^{(a)} + (1 - \eta) G^{(d)}, \tag{2.6}$$

where $(1 - \eta)$ is the sum total of all the point masses of G. Let $d(G)$ denote the set of all jump discontinuities of G, i.e., $d(G) = \{x \in \mathbb{R} : G(x) - G(x-) > 0\}$. Let us define $J_G : d(G) \to [0, 1]$ to be a function defined only on the jump points of G such that $J_G(x) = G(x) - G(x-)$ for all $x \in d(G)$. The following result addresses the identifiability issue when both F_s and F_b are discrete CDFs.

Lemma 3. *Let F_s and F_b be discrete CDFs. If $d(F_b) \not\subset d(F_s)$, then $\alpha_0 = \alpha$, i.e., (2.1) is identifiable. If $d(F_b) \subset d(F_s)$, then $\alpha_0 = \alpha \left\{1 - \inf_{x \in d(F_b)} J_{F_s}(x)/J_{F_b}(x)\right\}$. Thus, $\alpha_0 = \alpha$ if and only if $\inf_{x \in d(F_b)} J_{F_s}(x)/J_{F_b}(x) = 0$.*

Next, let us assume that both F_s and F_b are absolutely continuous CDFs.

Lemma 4. *Suppose that F_s and F_b are absolutely continuous, i.e., they have densities f_s and f_b, respectively. Then*

$$\alpha_0 = \alpha \left\{1 - \operatorname{ess\,inf} \frac{f_s}{f_b}\right\},$$

where, for any function g, $\operatorname{ess\,inf} g = \sup\{a \in \mathbb{R} : \mathfrak{m}(\{x : g(x) < a\}) = 0\}$, $\mathfrak{m}$ being the Lebesgue measure. As a consequence, $\alpha_0 < \alpha$ if and only if there exists $c > 0$ such that $f_s \geq c f_b$, almost everywhere w.r.t. $\mathfrak{m}$.

The above lemma states that if there does not exist any $c > 0$ for which $f_s(x) \geq c f_b(x)$, for almost every x, then $\alpha_0 = \alpha$ and we can estimate the mixing proportion correctly. Note that, in particular, if the support of F_s is strictly contained in that of F_b, then the problem is identifiable and we can estimate α.

In Appendix 2.11 we apply the above two lemmas to two discrete (Poisson and binomial) distributions and two absolutely continuous (exponential and normal) distributions to obtain the exact relationship between α and α_0. In the following lemma, proved in greater generality in Appendix 2.11, we give conditions under which a general CDF F, that can be represented as in (2.6), is identifiable.

Lemma 5. *Suppose that $F = \kappa F^{(a)} + (1-\kappa)F^{(d)}$, where $F^{(a)}$ is an absolutely continuous CDF and $F^{(d)}$ is a piecewise constant CDF, for some $\kappa \in (0,1)$. Then (2.1) is identifiable, if either $F^{(a)}$ or $F^{(d)}$ are identifiable.*

2.3 Estimation

2.3.1 Estimation of the mixing proportion α_0

In this section we consider the estimation of α_0 as defined in (2.5). For the rest of the paper, unless otherwise noted, we assume

$$X_1, X_2, \ldots, X_n \text{ is an i.i.d. sample from } F \text{ as in (2.1).}$$

Recall the definitions of $\hat{F}^{\gamma}_{s,n}$ and $\check{F}^{\gamma}_{s,n}$, for $\gamma \in (0,1]$; see (2.2) and (2.3). When $\gamma = 1$, we have $\hat{F}^{\gamma}_{s,n} = \mathbb{F}_n = \check{F}^{\gamma}_{s,n}$ as $\hat{F}^{\gamma}_{s,n}$ (for $\gamma = 1$) is a CDF. Whereas, when γ is much smaller than α_0 the regularisation of $\hat{F}^{\gamma}_{s,n}$ modifies it, and thus $\hat{F}^{\gamma}_{s,n}$ and $\check{F}^{\gamma}_{s,n}$ are quite different. We would like to compare the naive and isotonised estimators $\hat{F}^{\gamma}_{s,n}$ and $\check{F}^{\gamma}_{s,n}$, respectively, and choose the smallest γ for which their distance is still small. This leads to the following estimator of α_0:

$$\hat{\alpha}_0^{c_n} = \inf\left\{\gamma \in (0,1] : \gamma d_n(\hat{F}^{\gamma}_{s,n}, \check{F}^{\gamma}_{s,n}) \leq \frac{c_n}{\sqrt{n}}\right\}, \tag{2.7}$$

where c_n is a sequence of constants and d_n stands for the $L_2(\mathbb{F}_n)$ distance, i.e., if $g, h : \mathbb{R} \to \mathbb{R}$ are two functions, then $d_n^2(g,h) = \int \{g(x) - h(x)\}^2 \, d\mathbb{F}_n(x)$. It is easy to see that

$$d_n(\mathbb{F}_n, \gamma\check{F}^{\gamma}_{s,n} + (1-\gamma)F_b) = \gamma d_n(\hat{F}^{\gamma}_{s,n}, \check{F}^{\gamma}_{s,n}).$$

For simplicity of notation, using (2.3.1), we define $\gamma d_n(\hat{F}^{\gamma}_{s,n}, \check{F}^{\gamma}_{s,n})$ for $\gamma = 0$ as

$$\lim_{\gamma \to 0+} \gamma d_n(\hat{F}^{\gamma}_{s,n}, \check{F}^{\gamma}_{s,n}) = d_n(\mathbb{F}_n, F_b). \tag{2.8}$$

This convention is followed in the rest of the paper.

The choice of c_n is important, and in the following sections we address this issue in detail. We derive conditions on c_n that lead to consistent estimators of α_0. We will also show that particular (distribution-free) choices of c_n will lead to honest lower confidence bounds for α_0.

Next, we prove a result which implies that, in the multiple testing problem, estimators of α_0 do not depend on whether we use p-values or z-values to perform our analysis. Let $\Psi : \mathbb{R} \to \mathbb{R}$ be a known continuous non-decreasing function. We define $\Psi^{-1}(y) := \inf\{t \in \mathbb{R} : y \leq \Psi(t)\}$, and $Y_i := \Psi^{-1}(X_i)$. It is easy to see that $Y_1, Y_2, \ldots, Y_n$ is an i.i.d. sample from $G := \alpha F_s \circ \Psi + (1-\alpha) F_b \circ \Psi$. Suppose now that we work with $Y_1, Y_2, \ldots, Y_n$, instead of $X_1, X_2, \ldots, X_n$, and want to estimate α. We can define α_0^Y as in (2.4) but with $\{G, F_b \circ \Psi\}$ instead of $\{F, F_b\}$. The following result shows that α_0 and its estimators proposed in this paper are invariant under such monotonic transformations.

Theorem 1. *Let $\mathbb{G}_n$ be the empirical CDF of $Y_1, Y_2, \ldots, Y_n$. Also, let $\hat{G}_{s,n}$ and $\check{G}^\gamma_{s,n}$ be as defined in (2.2) and (2.3), respectively, but with $\{\mathbb{G}_n, F_b \circ \Psi\}$ instead of $\{\mathbb{F}_n, F_b\}$. Then $\alpha_0 = \alpha_0^Y$ and $\gamma d_n(\hat{F}^\gamma_{s,n}, \check{F}^\gamma_{s,n}) = \gamma d_n(\hat{G}^\gamma_{s,n}, \check{G}^\gamma_{s,n})$ for all $\gamma \in (0, 1]$.*

2.3.2 Consistency of $\hat{\alpha}_0^{c_n}$

We start with two elementary results on the behaviour of our criterion function $\gamma d_n(\check{F}^\gamma_{s,n}, \hat{F}^\gamma_{s,n})$.

Lemma 6. *For $1 \geq \gamma \geq \alpha_0$, $\gamma d_n(\check{F}^\gamma_{s,n}, \hat{F}^\gamma_{s,n}) \leq d_n(F, \mathbb{F}_n)$. Thus,*

$$\gamma d_n(\hat{F}^\gamma_{s,n}, \check{F}^\gamma_{s,n}) \xrightarrow{a.s.} \begin{cases} 0, & \gamma - \alpha_0 \geq 0, \\ > 0, & \gamma - \alpha_0 < 0. \end{cases} \tag{2.9}$$

Lemma 7. *The set $A_n := \{\gamma \in [0,1] : \sqrt{n}\gamma d_n(\hat{F}^\gamma_{s,n}, \check{F}^\gamma_{s,n}) \leq c_n\}$ is convex. Thus, $A_n = [\hat{\alpha}_0^{c_n}, 1]$.*

The following result shows that for a broad range of choices of c_n, our estimation procedure is consistent.

Theorem 2. *If $c_n = o(\sqrt{n})$ and $c_n \to \infty$, then $\hat{\alpha}_0^{c_n} \xrightarrow{P} \alpha_0$.*

A proper choice of c_n is important and crucial for the performance of $\hat{\alpha}_0^{c_n}$. We suggest doing cross-validation to find the optimal tuning parameter c_n. In Section 2.8.2.1 we detail this approach and illustrate its good finite sample performance through simulation examples; see Tables 2.2-2.5, Section 2.8.2.4, and Appendix 2.12. However, cross-validation can be computationally expensive. Another useful choice for c_n is to take $c_n = 0.1 \log\log n$. After extensive simulations, we observe that $c_n = 0.1 \log\log n$ has good finite sample performance for estimating α_0; see Section 2.8 and Appendix 2.12 for more details.

2.3.3 Rate of convergence and asymptotic limit

We first discuss the case $\alpha_0 = 0$. In this situation, under minimal assumptions, we show that as the sample size grows, $\hat{\alpha}_0^{c_n}$ exactly equals α_0 with probability converging to 1.

Lemma 8. *When $\alpha_0 = 0$, if $c_n \to \infty$ as $n \to \infty$, then $P(\hat{\alpha}_0^{c_n} = 0) \to 1$.*

For the rest of this section we assume that $\alpha_0 > 0$. The following theorem gives the rate of convergence of $\hat{\alpha}_0^{c_n}$.

Theorem 3. *Let $r_n := \sqrt{n}/c_n$. If $c_n \to \infty$ and $c_n = o(n^{1/4})$ as $n \to \infty$, then $r_n(\hat{\alpha}_0^{c_n} - \alpha_0) = O_P(1)$.*

The proof of the above result is involved and we give the details in Appendix 2.14.9.

Remark 1. *[Genovese and Wasserman, 2004] show that the estimators of α_0 proposed by [Hengartner and Stark, 1995] and [Swanepoel, 1999] have convergence rates of $(n/\log n)^{1/3}$ and $n^{2/5}/(\log n)^{\delta}$, for $\delta > 0$, respectively. Morover, both results require smoothness assumptions on F – [Hengartner and Stark, 1995] require F to be concave with a density that is Lipschitz of order 1, while [Swanepoel, 1999] requires even stronger smoothness conditions on the density. [Nguyen and Matias, 2013] prove that when the density of $F_s^{\alpha_0}$ vanishes at a set of points of measure zero and satisfies certain regularity assumptions, then any $\sqrt{n}$-consistent estimator of α_0 will not have finite variance in the limit (if such an estimator exists).*

We can take $r_n = \sqrt{n}/c_n$ arbitrarily close to $\sqrt{n}$ by choosing c_n that increases to infinity very slowly. If we take $c_n = \log\log n$, we get an estimator that has a rate of convergence $\sqrt{n}/\log\log n$. In fact, as the next result shows, $r_n(\hat{\alpha}_0^{c_n} - \alpha_0)$ converges to a degenerate limit. In Section 2.8.2, we analyse the effect of c_n on the finite sample performance of $\hat{\alpha}_0^{c_n}$ for estimating α_0 through simulations and advocate a proper choice of the tuning parameter c_n.

Theorem 4. *When $\alpha_0 > 0$, if $r_n \to \infty$, $c_n = o(n^{1/4})$ and $c_n \to \infty$, as $n \to \infty$, then*

$$r_n(\hat{\alpha}_0^{c_n} - \alpha_0) \xrightarrow{P} c,$$

where $c < 0$ is a constant that depends on α_0, F and F_b.

2.4 Lower confidence bound for α_0

The asymptotic limit of the estimator $\hat{\alpha}_0^{c_n}$ discussed in Section 2.3 depends on unknown parameters (e.g., α_0, F) in a complicated fashion and is of little practical use. Our goal in this sub-section is to construct a finite sample (honest) lower confidence bound $\hat{\alpha}_L$ with the property

$$P(\alpha_0 \geq \hat{\alpha}_L) \geq 1 - \beta, \tag{2.10}$$

for a specified confidence level $(1-\beta)$ $(0 < \beta < 1)$, that is valid for any n and is tuning parameter free. Such a lower bound would allow one to assert, with a specified level of confidence, that the proportion of "signal" is at least $\hat{\alpha}_L$.

It can also be used to test the hypothesis that there is no "signal" at level β by rejecting when $\hat{\alpha}_L > 0$. The problem of no "signal' is known as the homogeneity problem in the statistical literature. It is easy to show that $\alpha_0 = 0$ if and only if $F = F_b$. Thus, the hypothesis of no "signal" or homogeneity can be addressed by testing whether $\alpha_0 = 0$ or not. There has been a considerable amount of work on the homogeneity problem, but most of the papers make parametric model assumptions. [Lindsay, 1995] is an authoritative monograph on the homogeneity problem but the components are assumed to be from a known exponential family. [Walther, 2001] and [Walther, 2002] discuss the homogeneity problem under the assumption that the densities are log-concave. [Donoho

and Jin, 2004] and [Cai and Jin, 2010] discuss the problem of detecting sparse heterogeneous mixtures under parametric settings using the 'higher criticism' statistic; see Appendix 2.13 for more details.

It will be seen that our approach will lead to an exact lower confidence bound when $\alpha_0 = 0$, i.e., $P(\hat{\alpha}_L = 0) = 1 - \beta$. The methods of [Genovese and Wasserman, 2004] and [Meinshausen and Rice, 2006] usually yield conservative lower bounds.

Theorem 5. *Let H_n be the CDF of $\sqrt{n}d_n(\mathbb{F}_n, F)$. Let $\hat{\alpha}_L$ be defined as in (2.7) with $c_n = H_n^{-1}(1-\beta)$. Then (2.10) holds. Furthermore if $\alpha_0 = 0$, then $P(\hat{\alpha}_L = 0) = 1 - \beta$, i.e., it is an exact lower bound.*

The proof of the above theorem can be found in Appendix 2.14.13. Note that H_n is distribution-free (i.e., it does not depend on F_s and F_b) when F is a continuous CDF and can be readily approximated by Monte Carlo simulations using a sample of uniforms. For moderately large n (e.g., $n \geq 500$) the distribution H_n can be very well approximated by that of the Cramér-von Mises statistic, defined as

$$\sqrt{n}d(\mathbb{F}_n, F) := \sqrt{\int n\{\mathbb{F}_n(x) - F(x)\}^2 dF(x)}.$$

Letting G_n be the CDF of $\sqrt{n}d(\mathbb{F}_n, F)$, we have the following result.

Theorem 6. $\sup_{x\in\mathbb{R}} |H_n(x) - G_n(x)| \to 0$ *as* $n \to \infty$.

Hence in practice, for moderately large n, we can take c_n to be the $(1-\beta)$-quantile of G_n or its asymptotic limit, which are readily available (e.g., see [Anderson and Darling, 1952]). When F is a continuous CDF, the asymptotic 95% quantile of G_n is 0.6792, and is used in our data analysis. Note that

$$P(\alpha_0 \geq \hat{\alpha}_L) = P(\sqrt{n}\alpha_0 d_n(\hat{F}_{s,n}^{\alpha_0}, \check{F}_{s,n}^{\alpha_0}) \geq H_n^{-1}(1-\beta)).$$

The following theorem gives the explicit asymptotic limit of $P(\alpha_0 \geq \hat{\alpha}_L)$ but it is not useful for practical purposes as it involves the unknown $F_s^{\alpha_0}$ and F.

Theorem 7. *Assume that $\alpha_0 > 0$. Then $\sqrt{n}\alpha_0 d_n(\hat{F}_{s,n}^{\alpha_0}, \check{F}_{s,n}^{\alpha_0}) \xrightarrow{d} U$, where U is a random variable whose distribution depends only on α_0, F, and F_b.*

The proof of the above theorem and the explicit from of U can be found in Appendix 2.14. The proof of Theorem 6 and a detailed discussion on the performance of the lower confidence bound for detecting heterogeneity in the *moderately sparse* signal regime considered in [Donoho and Jin, 2004] can be found in Appendix 2.13.

2.5 A heuristic estimator of α_0

In simulations, we observe that the finite sample performance of (2.7) is affected by the choice of c_n (for an extensive simulation study on this see Section 2.8.2). This motivates us to propose a method to estimate α_0 that is completely automated and has good finite sample performance. We start with a lemma that describes the shape of our criterion function, and will motivate our procedure.

Lemma 9. *$\gamma d_n(\hat{F}^{\gamma}_{s,n}, \check{F}^{\gamma}_{s,n})$ is a non-increasing convex function of γ in $(0,1)$.*

Writing

$$\hat{F}^{\gamma}_{s,n} = \frac{\mathbb{F}_n - F}{\gamma} + \left\{ \frac{\alpha_0}{\gamma} F_s^{\alpha_0} + \left(1 - \frac{\alpha_0}{\gamma}\right) F_b \right\},$$

we see that for $\gamma \geq \alpha_0$, the second term in the right hand side is a CDF. Thus, for $\gamma \geq \alpha_0$, $\hat{F}^{\gamma}_{s,n}$ is very close to a CDF as $\mathbb{F}_n - F = O_P(n^{-1/2})$, and hence $\check{F}^{\gamma}_{s,n}$ should also be close to $\hat{F}^{\gamma}_{s,n}$. Whereas, for $\gamma < \alpha_0$, $\hat{F}^{\gamma}_{s,n}$ is not close to a CDF, and thus the distance $\gamma d_n(\hat{F}^{\gamma}_{s,n}, \check{F}^{\gamma}_{s,n})$ is appreciably large. Therefore, at α_0, we have a "regime" change: $\gamma d_n(\hat{F}^{\gamma}_{s,n}, \check{F}^{\gamma}_{s,n})$ should have a slowly decreasing segment to the right of α_0 and a steeply non-increasing segment to the left of α_0. Fig. 2.1 shows two typical such plots of the function $\gamma d_n(\hat{F}^{\gamma}_{s,n}, \check{F}^{\gamma}_{s,n})$, where the left panel corresponds to a mixture of $N(2,1)$ with $N(0,1)$ (setting I) and in the right panel we have a mixture of Beta(1,10) and Uniform$(0,1)$ (setting II). We will use these two settings to illustrate our methodology in the rest of this section and also in Section 2.8.1.

Using the above heuristics, we can see that the "elbow" of the function should provide a good estimate of α_0; it is the point that has the maximum curvature, i.e., the point where the second derivative is maximal. We denote this estimator by $\tilde{\alpha}_0$. Notice that both the estimators $\tilde{\alpha}_0$ and $\hat{\alpha}_0^{c_n}$ are derived from $\gamma d_n(\hat{F}^{\gamma}_{s,n}, \check{F}^{\gamma}_{s,n})$, as a function of γ, albeit they look at two different aspects of the function.

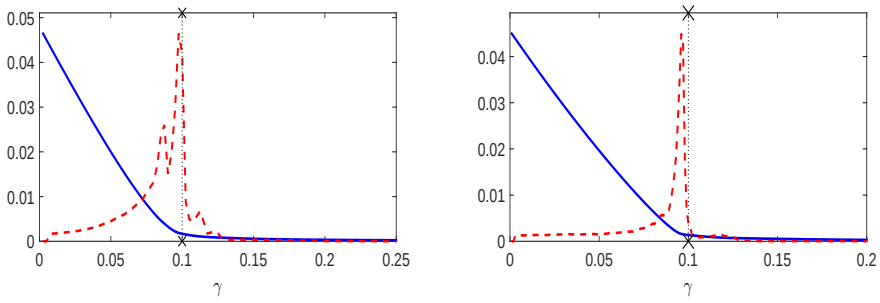

Figure 2.1: Plots of $\gamma d_n(\hat{F}_{s,n}^{\gamma}, \check{F}_{s,n}^{\gamma})$ (in solid blue) overlaid with its (scaled) second derivative (in dashed red) for $\alpha_0 = 0.1$ and $n = 5000$. Left panel: setting I; right panel: setting II.

In the above plots we have used numerical methods to approximate the second derivative of $\gamma d_n(\hat{F}_{s,n}^{\gamma}, \check{F}_{s,n}^{\gamma})$ (using the method of double differencing). We advocate plotting the function $\gamma d_n(\hat{F}_{s,n}^{\gamma}, \check{F}_{s,n}^{\gamma})$ as γ varies between 0 and 1. In most cases, plots similar to Fig. 2.1 would immediately convey to the practitioner the most appropriate choice of $\tilde{\alpha}_0$. In some cases though, there can be multiple peaks in the second derivative, in which case some discretion on the part of the practitioner might be required. It must be noted that the idea of finding the point where the second derivative is large to detect an "elbow" or "knee" of a function is not uncommon; see e.g., [Salvador and Chan, 2004]. However, in Section 2.8.2.4 and Appendix 2.12, we show some simulation examples where $\tilde{\alpha}_0$ fails to consistently estimate the "elbow" of $\gamma d_n(\hat{F}_{s,n}^{\gamma}, \check{F}_{s,n}^{\gamma})$.

2.6 Estimation of the distribution function and its density

2.6.1 Estimation of F_s

Let us assume for the rest of this section that (2.1) is identifiable, i.e., $\alpha = \alpha_0$, and $\alpha_0 > 0$. Thus $F_s^{\alpha_0} = F_s$. Once we have a consistent estimator $\check{\alpha}_n$ (which may or may not be $\hat{\alpha}_0^{c_n}$ as discussed in the previous sections) of α_0, a natural nonparametric estimator of F_s is $\check{F}_{s,n}^{\check{\alpha}_n}$, defined as the minimiser of (2.3). In the following theorem we show that, indeed, $\check{F}_{s,n}^{\check{\alpha}_n}$ is uniformly consistent for estimating F_s. We also derive the rate of convergence of $\check{F}_{s,n}^{\check{\alpha}_n}$.

Theorem 8. *Suppose that* $\check{\alpha}_n \xrightarrow{P} \alpha_0$. *Then, as* $n \to \infty$, $\sup_{x\in\mathbb{R}} |\check{F}^{\check{\alpha}_n}_{s,n}(x) - F_s(x)| \xrightarrow{P} 0$. *Furthermore, if* $q_n(\check{\alpha}_n - \alpha_0) = O_P(1)$, *where* $q_n = o(\sqrt{n})$, *then* $\sup_{x\in\mathbb{R}} q_n|\check{F}^{\check{\alpha}_n}_{s,n}(x) - F_s(x)| = O_P(1)$. *Additionally, for* $\hat{\alpha}_0^{c_n}$ *as defined in* (2.7), *we have*

$$\sup_{x\in\mathbb{R}} |r_n(\hat{F}^{\hat{\alpha}_0^{c_n}}_{s,n} - F_s)(x) - Q(x)| \xrightarrow{P} 0 \quad \textit{and} \quad r_n d(\check{F}^{\hat{\alpha}_0^{c_n}}_{s,n}, F_s) \xrightarrow{P} c$$

for a function $Q : \mathbb{R} \to \mathbb{R}$ *and a constant* $c > 0$ *depending only on* α_0, F, *and* F_b.

An immediate consequence of Theorem 8 is that $d_n(\check{F}^{\check{\alpha}_n}_{s,n}, \hat{F}^{\check{\alpha}_n}_{s,n}) \xrightarrow{P} 0$ as $n \to \infty$. Left panel of Fig. 2.2 shows our estimator $\check{F}^{\check{\alpha}_n}_{s,n}$ along with the true F_s for the same data set used in the right panel of Fig. 2.1.

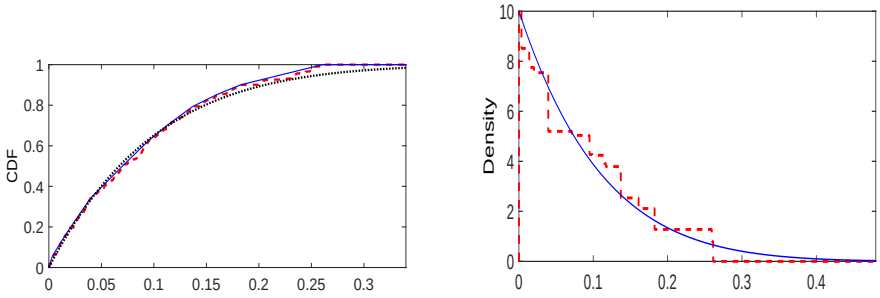

Figure 2.2: Left panel: Plots of $\check{F}^{\check{\alpha}_0}_{s,n}$ (in dashed red), $F^{\dagger}_{s,n}$ (in solid blue) and F_s (in dotted black) for setting II; right panel: plots of $f^{\dagger}_{s,n}$ (in dashed red) and f_s (in solid blue) for setting II.

2.6.2 Estimating the density of F_s

Suppose now that F_s has a density f_s. Obtaining nonparametric estimators of f_s can be difficult as it requires smoothing and usually involves the choice of tuning parameter(s) (e.g., smoothing bandwidths), and especially so in our set-up.

In this sub-section we describe a tuning parameter free approach to estimating f_s, under the additional assumption that f_s is non-increasing. The assumption that f_s is non-increasing, i.e., F_s is concave on its support, is natural in many situations (see Section 2.7 for an application in the multiple testing problem) and has been investigated by several authors, including [Grenander, 1956], [Langaas *et al.*, 2005] and [Genovese

and Wasserman, 2004]. Without loss of generality, we assume that f_s is non-increasing on $[0, \infty)$.

For a bounded function $g : [0, \infty) \to \mathbb{R}$, let us represent the least concave majorant (LCM) of g by $LCM[g]$. Thus, $LCM[g]$ is the smallest concave function that lies above g. Define $F^{\dagger}_{s,n} := LCM[\check{F}^{\check{\alpha}_n}_{s,n}]$. Note that $F^{\dagger}_{s,n}$ is a valid CDF. We can now estimate f_s by $f^{\dagger}_{s,n}$, where $f^{\dagger}_{s,n}$ is the piece-wise constant function obtained by taking the left derivative of $F^{\dagger}_{s,n}$. In the following result we show that both $F^{\dagger}_{s,n}$ and $f^{\dagger}_{s,n}$ are consistent estimators of their population versions.

Theorem 9. *Assume that $F_s(0) = 0$ and that F_s is concave on $[0, \infty)$. If $\check{\alpha}_n \xrightarrow{P} \alpha_0$, then, as $n \to \infty$,*

$$\sup_{x \in \mathbb{R}} |F^{\dagger}_{s,n}(x) - F_s(x)| \xrightarrow{P} 0.$$

Further, if for any $x > 0$, $f_s(x)$ is continuous at x, then, $f^{\dagger}_{s,n}(x) \xrightarrow{P} f_s(x)$.

Computing $F^{\dagger}_{s,n}$ and $f^{\dagger}_{s,n}$ are straightforward, an application of the PAVA gives both the estimators; see e.g., Chapter 1 of [Robertson *et al.*, 1988]. In Fig. 2.2 the left panel shows the LCM $F^{\dagger}_{s,n}$ whereas the right panel shows its derivative $f^{\dagger}_{s,n}$ along with the true density f_s for the same data set used in the right panel of Fig. 2.1.

2.7 Multiple testing problem

The problem of estimating the proportion of false null hypotheses α_0 is of interest in situations where a large number of hypothesis tests are performed. Recently, various such situations have arisen in applications. One major motivation is in estimating the proportion of genes that are differentially expressed in deoxyribonucleic acid (DNA) microarray experiments. However, estimating the proportion of true null hypotheses is also of interest, for example, in functional magnetic resonance imaging (see [Turkheimer *et al.*, 2001]) and source detection in astrophysics (see [Miller *et al.*, 2001]).

Suppose that we wish to test n null hypotheses $H_{01}, H_{02}, \ldots, H_{0n}$ on the basis of a data set $\mathbb{X}$. Let H_i denote the (unobservable) binary variable that is 0 if H_{0i} is true, and 1 otherwise, $i = 1, \ldots, n$. We want a decision rule $\mathcal{D}$ that will produce a decision

of "null" or "non-null" for each of the n cases. In their seminal work, [Benjamini and Hochberg, 1995] argued that an important quantity to control is the false discovery rate (FDR) and proposed a procedure with the property FDR $\leq \beta(1-\alpha_0)$, where β is the user-defined level of the FDR procedure. When α_0 is significantly bigger than 0 an estimate of α_0 can be used to yield a procedure with FDR approximately equal to β and thus will result in an increased power. This is essentially the idea of the adapted control of FDR (see [Benjamini and Hochberg, 2000]). See [Storey, 2002], [Black, 2004], [Langaas *et al.*, 2005], [Benjamini *et al.*, 2006], and [Donoho and Jin, 2004] for a discussion on the importance of efficient estimation of α_0 and some proposed estimators.

Our method can be directly used to yield an estimator of α_0 that does not require the specification of any tuning parameter, as discussed in Section 2.5. We can also obtain a completely nonparametric estimator of F_s, the distribution of the p-values arising from the alternative hypotheses. Suppose that F_b has a density f_b and F_s has a density f_s. To keep the following discussion more general, we allow f_b to be any known density, although in most multiple testing applications we will take f_b to be Uniform$(0,1)$. The *local false discovery rate* (LFDR) is defined as the function $l:(0,1)\to[0,\infty)$, where

$$l(x) = P(H_i = 0 | X_i = x) = \frac{(1-\alpha_0)f_b(x)}{f(x)},$$

and $f(x) = \alpha_0 f_s(x) + (1-\alpha_0)f_b(x)$ is the density of the observed p-values. The estimation of the LFDR l is important because it gives the probability that a particular null hypothesis is true given the observed p-value for the test. The LFDR method can help us get easily interpretable thresholding methods for reporting the "interesting" cases (e.g., $l(x) \leq 0.20$). Obtaining good estimates of l can be tricky as it involves the estimation of an unknown density, usually requiring smoothing techniques; see Section 5 of [Efron, 2010] for a discussion on estimation and interpretation of l. From the discussion in Section 2.6.1, under the additional assumption that f_s is non-increasing, we have a natural tuning parameter free estimator $\hat{l}$ of the LFDR:

$$\hat{l}(x) = \frac{(1-\breve{\alpha}_n)f_b(x)}{\breve{\alpha}_n f^{\dagger}_{s,n}(x) + (1-\breve{\alpha}_n)f_b(x)}, \qquad \text{for } x \in (0,1).$$

The assumption that f_s is non-increasing, i.e., F_s is concave, is quite natural – when the alternative hypothesis is true the p-value is generally small – and has been investigated

Table 2.1: Coverage probabilities of nominal 95% lower confidence bounds for the three methods when $n = 1000$ and $n = 5000$.

	$n = 1000$						$n = 5000$					
	Setting I			Setting II			Setting I			Setting II		
α	$\hat{\alpha}_L$	$\hat{\alpha}_L^{GW}$	$\hat{\alpha}_L^{MR}$	$\hat{\alpha}_L$	$\hat{\alpha}_L^{GW}$	$\hat{\alpha}_L^{MR}$	$\hat{\alpha}_L$	$\hat{\alpha}_L^{GW}$	$\hat{\alpha}_L^{MR}$	$\hat{\alpha}_L$	$\hat{\alpha}_L^{GW}$	$\hat{\alpha}_L^{MR}$
0	0.95	0.98	0.93	0.95	0.98	0.93	0.95	0.97	0.93	0.95	0.97	0.93
0.01	0.97	0.98	0.99	0.97	0.97	0.99	0.98	0.98	0.99	0.98	0.98	0.99
0.03	0.98	0.98	0.99	0.98	0.98	0.99	0.98	0.98	0.99	0.98	0.98	0.99
0.05	0.98	0.98	0.99	0.98	0.98	0.99	0.99	0.99	0.99	0.98	0.98	0.99
0.10	0.99	0.99	1.00	0.99	0.98	0.99	0.99	0.99	1.00	0.99	0.98	0.99

by several authors, including [Genovese and Wasserman, 2004] and [Langaas *et al.*, 2005].

2.8 Simulation

To investigate the finite sample performance of the estimators developed in this paper, we carry out several simulation experiments. We also compare the performance of these estimators with existing methods. The R language ([R Development Core Team, 2008]) codes used to implement our procedures are available at `http://stat.columbia.edu/~rohit/Code/NPMixModelCode.pdf`.

2.8.1 Lower bounds for α_0

Although there has been some work on estimation of α_0 in the multiple testing setting, [Meinshausen and Rice, 2006] and [Genovese and Wasserman, 2004] are the only papers we found that discuss methodology for constructing lower confidence bounds for α_0. These procedures are connected and the methods in [Meinshausen and Rice, 2006] are extensions of those proposed in [Genovese and Wasserman, 2004]. The lower bounds proposed in both the papers approximately satisfy (2.10) and have the form $\sup_{t\in(0,1)}(\mathbb{F}_n(t) - t - \eta_{n,\beta}\delta(t))/(1-t)$, where $\eta_{n,\beta}$ is a *bounding sequence* for the *bounding*

function $\delta(t)$ at level β; see [Meinshausen and Rice, 2006]. [Genovese and Wasserman, 2004] use a constant bounding function, $\delta(t) = 1$, with $\eta_{n,\beta} = \sqrt{\log(2/\beta)/2n}$, whereas [Meinshausen and Rice, 2006] suggest a class of bounding functions but observe that the *standard deviation-proportional* bounding function $\delta(t) = \sqrt{t(1-t)}$ has optimal properties among a large class of possible bounding functions. We use this bounding function and a bounding sequence suggested by the authors. We denote the lower bound proposed in [Meinshausen and Rice, 2006] by $\hat{\alpha}_L^{MR}$, the bound in [Genovese and Wasserman, 2004] by $\hat{\alpha}_L^{GW}$, and the lower bound discussed in Section 2.4 by $\hat{\alpha}_L$. To be able to use the methods of [Meinshausen and Rice, 2006] and [Genovese and Wasserman, 2004] in setting I, introduced in Section 2.5, we transform the data such that F_b is Uniform$(0, 1)$; see Section 2.3.1 for the details.

We take $\alpha \in \{0, 0.01, 0.03, 0.05, 0.10\}$ and compare the performance of the three lower bounds in the two different simulation settings discussed in Section 2.5. For each setting we take the sample size n to be 1000 and 5000. We present the estimated coverage probabilities, obtained by averaging over 5000 independent replications, of the lower bounds for both settings in Table 2.1. We can immediately see from the table that the bounds are usually quite conservative. However, it is worth pointing out that when $\alpha_0 = 0$, our method has exact coverage, as discussed in Section 2.4. Also, the fact that our procedure is simple, easy to implement, and completely automated, makes it very attractive.

2.8.2 Estimation of α_0

In this sub-section, we illustrate and compare the performance of different estimators of α_0 under two sampling scenarios. In scenario A, we proceed as in [Langaas *et al.*, 2005]. Let $\mathbf{X}_j = (X_{1j}, X_{2j}, \ldots, X_{nj})$, for $j = 1, \ldots, J$, and assume that each $\mathbf{X}_j \sim N(\mu_{n\times 1}, \Sigma_{n\times n})$ and that $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_J$ are independent. We test $H_{0i} : \mu_i = 0$ versus $H_{1i} : \mu_i \neq 0$ for each $i = 1, 2, \ldots, n$. We set μ_i to zero for the true null hypotheses, whereas for the false null hypotheses, we draw μ_i from a symmetric bi-triangular density with parameters $a = \log_2(1.2) = 0.263$ and $b = \log_2(4) = 2$; see page 568 of [Langaas *et al.*, 2005] for the details. Let x_{ij} denote a realisation of X_{ij} and α be the proportion

of false null hypotheses. Let $\bar{x}_i = \sum_{j=1}^{J} x_{ij}/J$ and $s_i^2 = \sum_{j=1}^{J}(x_{ij} - \bar{x}_i)^2/(J-1)$. To test H_{0i} versus H_{1i}, we calculate a two-sided p-value based on a one-sample t-test, with $p_i = 2P(T_{J-1} \geq |\bar{x}_i/\sqrt{s_i^2/J}|)$, where T_{J-1} is a t-distributed random variable with $J-1$ degrees of freedom.

In scenario B, we generate $n+L$ independent random variables $w_1, w_2, \ldots, w_{n+L}$ from $N(0,1)$ and set $z_i = \frac{1}{\sqrt{L+1}} \sum_{j=i}^{i+L} w_j$ for $i = 1, 2, \ldots, n$. The dependence structure of the z_i's is determined by L. For example, $L = 0$ corresponds to the case where the z_i's are i.i.d. standard normal. Let $X_i = z_i + m_i$, for $i = 1, 2, \ldots, n$, where $m_i = 0$ under the null, and under the alternative, $|m_i|$ is randomly generated from Uniform$(m^*, m^* + 1)$ and $\text{sgn}(m_i)$, the sign of m_i, is randomly generated from $\{-1, 1\}$ with equal probabilities. Here m^* is a suitable constant that describes the simulation setting. Let $1 - \alpha$ be the proportion of true null hypotheses. Scenario B is inspired by the numerical studies in [Cai and Jin, 2010] and [Jin, 2008].

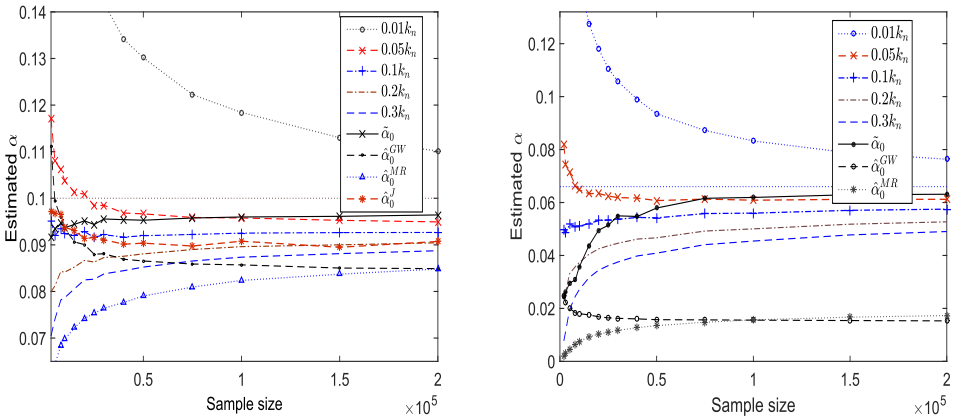

Figure 2.3: Plots of the means of different estimators of α_0, computed over 500 independent replications, as the sample size increases from 3000 to 2×10^5; left panel: scenario A with $\Sigma = I_{n \times n}$; right panel: scenario B with $L = 0$ and $m^* = 1$. The horizontal line (in dotted blue) indicates the value of α_0.

We use $\hat{\alpha}_0^{S,B}$ to denote the estimator proposed by [Storey, 2002] when bootstrapping is used to choose the required tuning parameter, and denote by $\hat{\alpha}_0^{S,\lambda}$ the estimator when the value of the tuning parameter is fixed at λ. [Langaas *et al.*, 2005] proposed an estimator that is tuning parameter free but crucially uses the known shape constraint of a convex

and non-increasing f_s; we denote it by $\hat{\alpha}_0^L$. We evaluate $\hat{\alpha}_0^L$ using the `convest` function in the R library `limma`. We also use the estimator proposed in [Meinshausen and Rice, 2006] for two bounding functions: $\delta(t) = \sqrt{t(1-t)}$ and $\delta(t) = 1$. For its implementation, we must choose a sequence $\{\beta_n\}$ going to zero as $n \to \infty$. [Meinshausen and Rice, 2006] did not specify any particular choice of $\{\beta_n\}$ but required the sequence satisfy some conditions. We choose $\beta_n = 0.05/\sqrt{n}$ and denote the estimators by $\hat{\alpha}_0^{MR}$ when $\delta(t) = \sqrt{t(1-t)}$ and by $\hat{\alpha}_0^{GW}$ when $\delta(t) = 1$ (see [Genovese and Wasserman, 2004]). We also compare our results with $\hat{\alpha}_0^E$, the estimator proposed in [Efron, 2007] using the central matching method, computed using the `locfdr` function in the R library `locfdr`. [Jin, 2008] and [Cai and Jin, 2010] propose estimators when the model is a mixture of Gaussian distributions; we denote the estimator proposed in Section 2.2 of [Jin, 2008] by $\hat{\alpha}_0^J$ and in Section 3.1 of [Cai and Jin, 2010] by $\hat{\alpha}_0^{CJ}$. Some of the competing methods require F_b to be of a specific form (e.g., standard normal) in which case we transform the observed data suitably.

The estimator $\hat{\alpha}_0^{c_n}$ depends on the choice of c_n and in the following we investigate a proper choice of c_n. We take $\alpha_0 = 0.1$ and evaluate the performance of $\hat{\alpha}_0^{\tau \times \log\log n}$ for different values of τ, as n increases, for scenarios A and B. The choice $c_n = \tau \times \log\log n$, for different values of τ, is suggested after extensive simulations. We also include $\tilde{\alpha}_0$, $\hat{\alpha}_0^{GW}$, $\hat{\alpha}_0^{MR}$, and $\hat{\alpha}_0^J$ in the comparison. For scenario A, we fix the sample size n at 5000 and $\Sigma = I_{n\times n}$. For scenario B, we fix $n = 5 \times 10^4$, $L = 0$, and $m^* = 1$. In Fig. 2.3, we illustrate the effect of c_n on estimation of α_0 as n varies from 3000 to 10^5. Recall that $\tilde{\alpha}_0$ denotes the estimator proposed in Section 2.5. For both scenarios, the sample mean of the estimators of α_0 proposed in this paper converge to the true α_0, as the sample size grows. The methods developed in this paper perform favorably in comparison to $\hat{\alpha}_0^{GW}$, $\hat{\alpha}_0^{MR}$, and $\hat{\alpha}_0^J$. Since, the choice of c_n dictates the finite sample performance of $\hat{\alpha}_0^{c_n}$, we propose cross-validation to find an appropriate value of the tuning parameter.

2.8.2.1 Cross-validation

In this sub-section, we use c instead of c_n to simplify the notation. In the following we briefly describe our cross-validation procedure. For a K-fold cross validation, we

Table 2.2: Means×10 and RMSEs×100 (in parentheses) of estimators discussed in Section 2.8.2 for scenario A with $\Sigma = I_{n\times n}$, $J = 10$, $n = 5000$, and $k_n = \log\log n$.

$10\alpha_0$	$\hat{\alpha}_0^{1k_n}$	$\hat{\alpha}_0^{CV}$	$\tilde{\alpha}_0$	$\hat{\alpha}_0^{GW}$	$\hat{\alpha}_0^{MR}$	$\hat{\alpha}_0^{S,0.5}$	$\hat{\alpha}_0^{J}$	$\hat{\alpha}_0^{CJ}$	$\hat{\alpha}_0^{L}$	$\hat{\alpha}_0^{E}$
0.10	0.13	0.15	0.13	0.00	0.01	0.09	0.14	0.05	0.16	0.36
	(1.00)	(1.79)	(0.83)	(1.00)	(0.88)	(1.41)	(1.50)	(5.32)	(1.20)	(3.70)
0.30	0.30	0.35	0.27	0.02	0.12	0.29	0.29	0.15	0.35	0.36
	(1.02)	(1.87)	(1.01)	(2.80)	(1.84)	(1.41)	(1.83)	(5.46)	(1.26)	(3.96)
0.50	0.48	0.51	0.46	0.18	0.26	0.47	0.49	0.26	0.55	0.35
	(1.09)	(1.9)	(1.12)	(3.29)	(2.46)	(1.49)	(1.91)	(5.73)	(1.34)	(3.80)
1.00	0.93	0.97	0.93	0.62	0.65	0.95	0.96	0.51	1.02	0.33
	(1.35)	(1.86)	(1.32)	(3.88)	(3.57)	(1.51)	(1.94)	(7.16)	(1.36)	(3.73)

Table 2.3: Means×10 and RMSEs×100 (in parentheses) of estimators discussed in Section 2.8.2 for scenario B with $L = 0$, $m^* = 1$, $n = 5 \times 10^4$, and $k_n = \log\log n$.

$10\alpha_0$	$\hat{\alpha}_0^{1k_n}$	$\hat{\alpha}_0^{CV}$	$\tilde{\alpha}_0$	$\hat{\alpha}_0^{GW}$	$\hat{\alpha}_0^{MR}$	$\hat{\alpha}_0^{S,B}$	$\hat{\alpha}_0^{J}$	$\hat{\alpha}_0^{CJ}$	$\hat{\alpha}_0^{L}$	$\hat{\alpha}_0^{E}$
0.07	0.03	0.04	0.08	0.00	0.00	0.04	0.11	0.19	0.03	0.06
	(0.44)	(0.67)	(0.28)	(0.66)	(0.66)	(0.65)	(0.96)	(2.96)	(0.38)	(0.77)
0.20	0.14	0.18	0.16	0.00	0.01	0.08	0.28	0.55	0.07	0.05
	(0.73)	(0.79)	(0.62)	(1.98)	(1.89)	(2.25)	(1.33)	(4.41)	(1.26)	(1.28)
0.33	0.25	0.31	0.28	0.02	0.04	0.12	0.48	0.92	0.12	0.05
	(0.89)	(0.85)	(0.95)	(3.15)	(2.91)	(3.83)	(1.77)	(6.48)	(2.14)	(1.90)
0.66	0.55	0.62	0.58	0.12	0.14	0.23	0.95	1.83	0.23	0.05
	(1.21)	(1.00)	(1.48)	(5.38)	(5.25)	(7.73)	(3.04)	(11.98)	(4.34)	(3.84)

randomly partition the data into K sets, say $\mathcal{D}_1, \ldots, \mathcal{D}_K$. Let $\mathbb{F}_n^k$ be the empirical CDF of the data in $\mathcal{D}_k$. Let $\hat{\alpha}_{0,-k}^c$ be the estimator defined in (2.7) using all data except those in $\mathcal{D}_k$ and tuning parameter c. Further, let $\check{F}_{s,n}^{\hat{\alpha}_{0,-k}^c,-k}$ be the estimator of F_s as defined in Lemma 1 using $\hat{\alpha}_{0,-k}^c$ and all data except those in $\mathcal{D}_k$. Define the cross-validated estimator of c as

$$c_{cv} := \arg\min_{c\in\mathbb{R}} \sum_{k=1}^{K} \int (\mathbb{F}_n^k - \hat{F}^k)^2 d\mathbb{F}_n^k, \tag{2.11}$$

where $\hat{F}^k := \hat{\alpha}_{0,-k}^c \check{F}_s^{\hat{\alpha}_{0,-k}^c,-k} + (1-\hat{\alpha}_{0,-k}^c)F_b$. In all simulations in this paper, we use $K = 10$ and denote this estimator by $\hat{\alpha}_0^{CV}$; see Section 7.10 of [Hastie *et al.*, 2009] for a more detailed study of cross-validation and a justification for $K = 10$. Fig. 2.4 illustrates the superior performance of $\hat{\alpha}_0^{CV}$ across different simulation settings; also see Sections 2.8.2.2 and 2.8.2.4, and Appendix 2.12

2.8.2.2 Performance under independence

In this sub-section, we take $\alpha \in \{0.01, 0.03, 0.05, 0.10\}$ and compare the performance of the different estimators under the independence setting of scenarios A and B. In Tables 2.2 and 2.3, we give the mean and root mean squared error (RMSE) of the estimators over 5000 independent replications. For scenario A, we fix the sample size n at 5000 and $\Sigma = I_{n\times n}$. For scenario B, we fix $n = 5\times 10^4$, $L = 0$, and $m^* = 1$. By an application of Lemma 4, it is easy to see that in scenario A, the model is identifiable (i.e., $\alpha_0 = \alpha$), while in scenario B, $\alpha_0 = \alpha \times 0.67$. For scenario A, the sample means of $\hat{\alpha}_0^{CV}$, $\tilde{\alpha}_0$, $\hat{\alpha}_0^J$, $\hat{\alpha}_0^L$, and $\hat{\alpha}_0^{0.1k_n}$ for $k_n = \log\log n$ are comparable. However, the RMSEs of $\tilde{\alpha}_0$ and $\hat{\alpha}_0^{0.1k_n}$ are lower than those of $\hat{\alpha}_0^{CV}$, $\hat{\alpha}_0^J$, and $\hat{\alpha}_0^L$. For scenario B, the sample means of $\tilde{\alpha}_0$, $\hat{\alpha}_0^{CV}$, and $\hat{\alpha}_0^{0.1k_n}$ are comparable. In scenario B, the performances of $\hat{\alpha}_0^J$ and $\hat{\alpha}_0^{CJ}$ are not comparable to the estimators proposed in this paper, as $\hat{\alpha}_0^J$ and $\hat{\alpha}_0^{CJ}$ estimate α, while $\tilde{\alpha}_0$, $\hat{\alpha}_0^{CV}$, and $\hat{\alpha}_0^{c_n}$ estimate α_0. Note that $\hat{\alpha}_0^L$ fails to estimate α_0 because the underlying assumption inherent in their estimation procedure, that f_s be non-increasing, does not hold. In scenario A, $\hat{\alpha}_0^{S,0.5}$ has the best performance among the different values of λ, while in scenario B, $\hat{\alpha}_0^{S,\lambda}$ has poor performance for all values of $\lambda \in [0,1]$. Furthermore, $\hat{\alpha}_0^{GW}$, $\hat{\alpha}_0^{MR}$, $\hat{\alpha}_0^{CJ}$, $\hat{\alpha}_0^{S,B}$ and $\hat{\alpha}_0^E$ perform poorly in both scenarios for all values of α_0.

Table 2.4: Means×10 and RMSEs×100 (in parentheses) of estimators discussed in Section 2.8.2 for scenario A with Σ as described in Section 2.8.2.3, $J = 10$, $n = 5000$, and $k_n = \log\log n$.

$10\alpha_0$	$\hat{\alpha}_0^{1k_n}$	$\hat{\alpha}_0^{CV}$	$\tilde{\alpha}_0$	$\hat{\alpha}_0^{GW}$	$\hat{\alpha}_0^{MR}$	$\hat{\alpha}_0^{S,0.5}$	$\hat{\alpha}_0^{J}$	$\hat{\alpha}_0^{CJ}$	$\hat{\alpha}_0^{L}$	$\hat{\alpha}_0^{E}$
0.10	0.46	0.42	0.33	0.07	0.06	0.28	0.22	0.07	0.32	0.37
	(5.15)	(4.23)	(3.84)	(1.72)	(1.27)	(4.11)	(3.03)	(10.61)	(4.37)	(3.91)
0.30	0.52	0.53	0.41	0.14	0.17	0.65	0.34	0.15	0.49	0.39
	(3.80)	(3.64)	(3.59)	(2.72)	(1.90)	(6.58)	(3.25)	(10.35)	(4.30)	(4.31)
0.50	0.66	0.76	0.54	0.26	0.31	0.54	0.49	0.25	0.66	0.37
	(3.52)	(5.43)	(3.85)	(3.56)	(2.50)	(2.61)	(3.60)	(10.45)	(4.31)	(4.03)
1.00	1.06	1.13	0.97	0.68	0.69	1.15	0.97	0.53	1.11	0.36
	(3.09)	(3.92)	(4.00)	(4.15)	(3.54)	(6.01)	(3.61)	(10.55)	(4.13)	(3.99)

2.8.2.3 Performance under dependence

The simulation settings of this sub-section are designed to investigate the effect of dependence on the performance of the estimators. For scenario A, we use the setting of [Langaas *et al.*, 2005]. We take Σ to be a block diagonal matrix with block size 100. Within blocks, the diagonal elements (i.e., variances) are set to 1 and the off-diagonal elements (within-block correlations) are set to $\rho = 0.5$. Outside of the blocks, all entries are set to 0. Tables 2.4 and 2.5 show that in both scenarios, none of the methods perform well for small values of α_0. However, in scenario A, the performances of $\hat{\alpha}_0^{0.1k_n}$, $\tilde{\alpha}_0$, and α_0^J are comparable, for larger values of α_0. In scenario B, $\hat{\alpha}_0^{0.1k_n}$ performs well for $\alpha_0 = 0.033$ and 0.067. Observe that, as in the independence setting, $\hat{\alpha}_0^{GW}$, $\hat{\alpha}_0^{MR}$, $\hat{\alpha}_0^{S,B}$, $\hat{\alpha}_0^{CJ}$, and $\hat{\alpha}_0^{E}$ perform poorly in both scenarios for all values of α_0.

2.8.2.4 Comparing the performance of $\hat{\alpha}_0^{c_n}$, $\hat{\alpha}_0^{CV}$, and $\tilde{\alpha}_0$

Although the heuristic estimator $\tilde{\alpha}_0$ performs quite well in most of the simulation settings considered, there exists scenarios where $\tilde{\alpha}_0$ can fail to consistently estimate α_0. To illustrate this we consider four different CDFs F_s and fix F_b to be the uniform distribution

Table 2.5: Means×10 and RMSEs×100 (in parentheses) of estimators discussed in Section 2.8.2 for scenario B with $L = 30$, $m^* = 1$, $n = 5 \times 10^4$, and $k_n = \log\log n$.

$10\alpha_0$	$\hat{\alpha}_0^{.1k_n}$	$\hat{\alpha}_0^{CV}$	$\tilde{\alpha}_0$	$\hat{\alpha}_0^{GW}$	$\hat{\alpha}_0^{MR}$	$\hat{\alpha}_0^{S,B}$	$\hat{\alpha}_0^{J}$	$\hat{\alpha}_0^{CJ}$	$\hat{\alpha}_0^{L}$	$\hat{\alpha}_0^{E}$
0.07	0.29	0.38	0.17	0.04	0.05	0.26	0.20	0.21	0.13	0.22
	(2.92)	(3.70)	(1.62)	(1.02)	(1.36)	(3.71)	(2.80)	(9.87)	(1.75)	(2.22)
0.20	0.30	0.42	0.18	0.04	0.04	0.16	0.33	0.55	0.13	0.19
	(1.84)	(2.88)	(1.25)	(1.75)	(1.71)	(2.24)	(3.25)	(10.35)	(1.42)	(2.27)
0.33	0.38	0.52	0.20	0.06	0.06	0.17	0.50	0.93	0.16	0.18
	(1.54)	(2.74)	(1.89)	(2.83)	(2.73)	(3.51)	(3.71)	(11.52)	(2.03)	(2.59)
0.67	0.63	0.77	0.31	0.14	0.15	0.24	0.95	1.82	0.25	0.16
	(1.53)	(2.25)	(4.32)	(5.26)	(5.13)	(7.60)	(4.54)	(15.13)	(4.23)	(4.08)

on $(0, 1)$ (see the top left plot of Fig. 2.4) and compare the performance of $\hat{\alpha}_0^{CV}$, $\tilde{\alpha}_0$, $\hat{\alpha}_0^{0.1k_n}$ with the best performing competing estimators (in each setting).

We see that $\tilde{\alpha}_0$ may fail to estimate the "elbow" of $\gamma d_n(\hat{F}_{s,n}^{\gamma}, \check{F}_{s,n}^{\gamma})$, as a function of γ, when F_s has a multi-modal density (see the middle row of Fig. 2.4). Observe that $\hat{\alpha}_0^{CV}$ and $\hat{\alpha}_0^{0.1k_n}$ perform favorably compared to all competing estimators and in the two scenarios where $\tilde{\alpha}_0$ fails to consistently estimate α_0, all our competing estimators also fail.

The first two toy examples have been carefully constructed to demonstrate situations where the point of maximum curvature ($\tilde{\alpha}_0$) is different from the "elbow" of the function; see the top right plot of Fig. 2.4 (also see Appendix 2.12 for further such examples).

2.8.2.5 Our recommendation

In this paper we study two estimators for α_0. For $\hat{\alpha}_0^{c_n}$, a proper choice of c_n is important for good finite sample performance. We suggest using cross-validation to find the optimal tuning parameter c_n. However, cross-validation can be computationally expensive. An attractive alternative in this situation is to use $\tilde{\alpha}_0$, which is easy to implement and has very good finite sample performance in most scenarios, especially with large sample sizes.

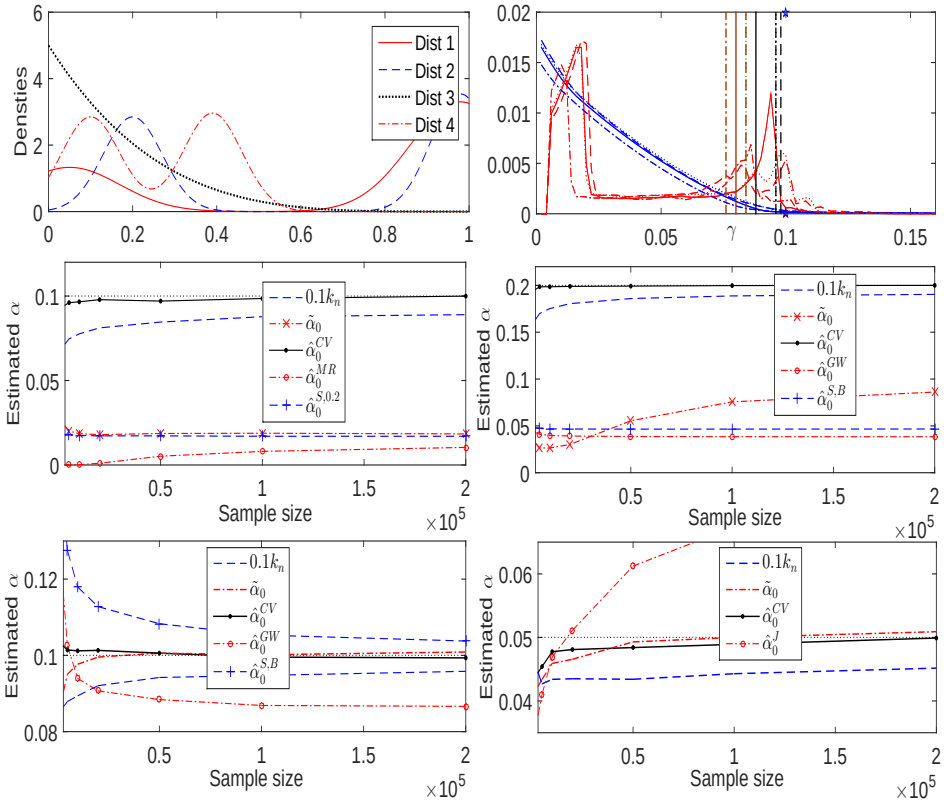

Figure 2.4: Top row left panel: density functions for different choices of F_s; top row right panel: plot of $\gamma d_n(\hat{F}^{\gamma}_{s,n}, \check{F}^{\gamma}_{s,n})$ (in blue), the scaled second derivative (in red), $\hat{\alpha}_0^{CV}$ (in black), and $\hat{\alpha}_0^{0.1k_n}$ (in brown) for 5 independent samples of size 5000 corresponding to "Dist 1"; the blue star denotes α_0. The bottom two rows show the means of different competing estimators of α_0, computed over 500 independent samples for Dist 1-4 (left-right, top-bottom) as sample size increases from 3000 to 2×10^5; in each figure the dotted black line denotes the true α_0.

We feel that a visual analysis of the plot of $\gamma d_n(\hat{F}^{\gamma}_{s,n}, \check{F}^{\gamma}_{s,n})$ can be useful in checking the validity of $\tilde{\alpha}_0$ as an estimator of the "elbow", and thus for α_0.

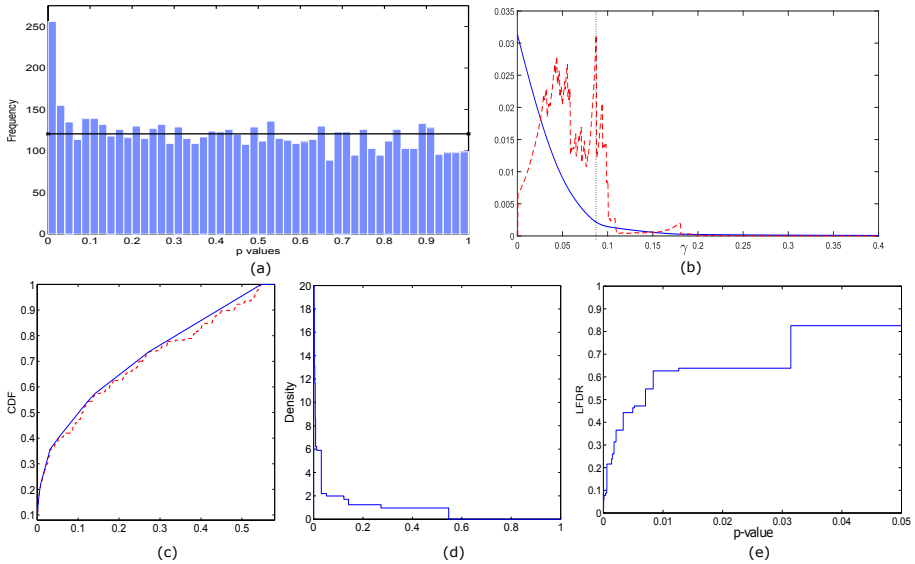

Figure 2.5: Plots for the prostate data: (a) Histogram of the p-values. The horizontal line (in solid black) indicates the Uniform$(0,1)$ distribution. (b) Plot of $\gamma d_n(\hat{F}_{s,n}^{\gamma}, \check{F}_{s,n}^{\gamma})$ (in solid blue) overlaid with its (scaled) second derivative (in dashed red). The vertical line (in dotted black) indicates the point of maximum curvature $\tilde{\alpha}_0 = 0.088$. (c) $\check{F}_{s,n}^{\tilde{\alpha}_0}$ (in dotted red) and $F_{s,n}^{\dagger}$ (in solid blue); (d) $f_{s,n}^{\dagger}$; (e) estimated LFDR $\hat{l}$ for p-values less than 0.05.

2.9 Real data analysis

2.9.1 Prostate data

Genetic expression levels for $n = 6033$ genes were obtained for $m = 102$ men, $m_1 = 50$ normal control subjects and $m_2 = 52$ prostate cancer patients. Without going into the biology involved, the principal goal of the study was to discover a small number of "interesting" genes, that is, genes whose expression levels differ between the cancer and control patients. Such genes, once identified, might be further investigated for a causal link to prostate cancer development. The prostate data is a 6033×102 matrix $\mathbb{X}$ having entries x_{ij} = expression level for gene i on patient j, $i = 1, 2, \ldots, n$, and $j = 1, 2, \ldots, m$, with $j = 1, 2, \ldots, 50$, for the normal controls, and $j = 51, 52, \ldots, 102$, for the cancer patients. Let $\bar{x}_i(1)$ and $\bar{x}_i(2)$ be the averages of x_{ij} for the normal controls and for the cancer patients, respectively, for gene i. The two-sample t-statistic for testing significance

Table 2.6: Estimates of α_0 for the two real data sets.

Data set	$\hat{\alpha}_0^{0.1k_n}$	$\hat{\alpha}_0^{CV}$	$\tilde{\alpha}_0$	$\hat{\alpha}_0^{GW}$	$\hat{\alpha}_0^{MR}$	$\hat{\alpha}_0^{S,B}$	$\hat{\alpha}_0^{J}$	$\hat{\alpha}_0^{CJ}$	$\hat{\alpha}_0^{L}$	$\hat{\alpha}_0^{E}$
Prostate	0.08	0.10	0.09	0.04	0.01	0.19	0.10	0.02	0.11	0.02
Carina	0.36	0.35	0.36	0.31	0.30	0.45	0.61	1.00	0.38	NA

of gene i is $t_i = \{\bar{x}_i(1) - \bar{x}_i(2)\}/s_i$, where s_i is an estimate of the standard error of $\bar{x}_i(1) - \bar{x}_i(2)$, i.e., $s_i^2 = (1/50 + 1/52)[\sum_{j=1}^{50}\{x_{ij} - \bar{x}_i(1)\}^2 + \sum_{j=51}^{102}\{x_{ij} - \bar{x}_i(2)\}^2]/100$.

We work with the p-values obtained from the 6033 two-sided t-tests instead of the "t-values" as then the distribution under the alternative will have a non-increasing density which we can estimate using the method developed in Section 2.6.1. Note that in our analysis we ignore the dependence of the p-values, which is only a moderately risky assumption for the prostate data; see Chapters 2 and 8 of [Efron, 2010] for further analysis and justification. Fig. 2.5 show the plots of various quantities of interest, found using the methodology developed in Section 2.6.1 and Section 2.7, for the prostate data example. The 95% lower confidence bound $\hat{\alpha}_L$ for this data is found to be 0.05. In Table 2.6, we display estimates of α_0 based on the methods considered in this paper for the prostate data and the Carina data (described below).

2.9.2 Carina data – an application in astronomy

In this sub-section we analyse the radial velocity (RV) distribution of stars in Carina, a dwarf spheroidal (dSph) galaxy. The dSph galaxies are low luminosity galaxies that are companions of the Milky Way. The data have been obtained by Magellan and MMT telescopes (see [Walker *et al.*, 2007]) and consist of radial (line of sight) velocity measurements of $n = 1266$ stars from Carina, contaminated with Milky Way stars in the field of view. We would like to understand the distribution of the RV of stars in Carina. For the contaminating stars from the Milky Way in the field of view we assume a non-Gaussian velocity distribution F_b that is known from the Besancon Milky Way model ([Robin *et al.*, 2003]), calculated along the line of sight to Carina.

The 95% lower confidence bound for α_0 is found to be 0.323. The right panel of

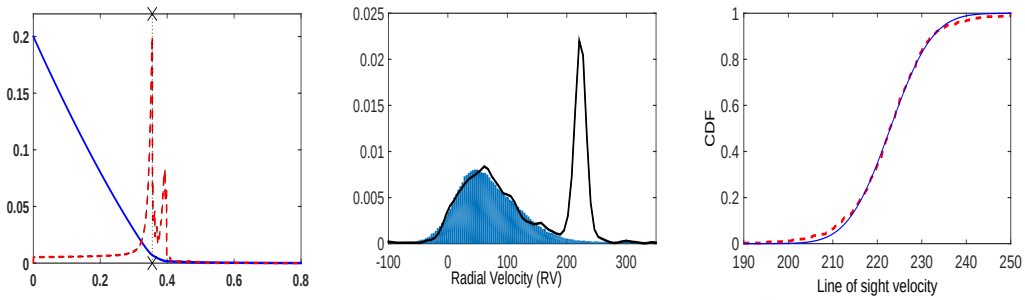

Figure 2.6: Plots for RV data in Carina dSph; left panel: $\gamma d_n(\hat{F}^{\gamma}_{s,n}, \check{F}^{\gamma}_{s,n})$ (in solid blue) overlaid with its (scaled) second derivative (in dashed red); middle panel: density of the RV distribution of the contaminating stars overlaid with the (scaled) kernel density estimator of the observed sample; right panel: $\check{F}^{\tilde{\alpha}_0}_{s,n}$ (in dashed red) overlaid with its closest Gaussian distribution (in solid blue).

Fig. 2.6 shows the estimate of F_s and the closest (in terms of minimising the $L_2(\check{F}^{\tilde{\alpha}_0}_{s,n})$ distance) fitting Gaussian distribution. Astronomers usually assume the distribution of the RVs for these dSph galaxies to be Gaussian. Indeed we see that the estimated F_s is close to a normal distribution (with mean 222.9 and standard deviation 7.51), although a formal test of this hypothesis is beyond the scope of the present paper. The estimate due to [Cai and Jin, 2010], $\hat{\alpha}_0^{CJ}$, is greater than one, while Efron's method (see [Efron, 2007]), implemented using the "locfdr" package in R, fails to estimate α_0.

2.10 Concluding remarks

In this paper we develop procedures for estimating the mixing proportion and the unknown distribution in a two component mixture model using ideas from shape restricted function estimation. We discuss the identifiability of the model and introduce an identifiable parameter α_0, under minimal assumptions on the model. We propose an honest finite sample lower confidence bound of α_0 that is distribution-free. Two point estimators of α_0, $\hat{\alpha}_0^{c_n}$ and $\tilde{\alpha}_0$, are studied. We prove that $\hat{\alpha}_0^{c_n}$ is a consistent estimator of α_0 and show that the rate of convergence of $\hat{\alpha}_0^{c_n}$ can be arbitrarily close to $\sqrt{n}$, for proper choices of c_n. These proposed estimators crucially rely on $\gamma d_n(\hat{F}^{\gamma}_{s,n}, \check{F}^{\gamma}_{s,n})$, as a function of

γ, whose plot provides useful insights about the nature of the problem and performance of the estimators.

We observe that the estimators of α_0 proposed in this paper have superior finite sample performance than most competing methods. In contrast to most previous work on this topic the results discussed in this paper hold true even when (2.1) is not identifiable. Under the assumption that (2.1) is identifiable, we can find an estimator of F_s which is uniformly consistent. Furthermore, if F_s is known to have a non-increasing density f_s we can find a consistent estimator of f_s. All these estimators are tuning parameter free and easily implementable.

We conclude this section by outlining some possible future research directions. Construction of two-sided confidence intervals for α_0 remains a hard problem as the asymptotic distribution of $\hat{\alpha}_0^{c_n}$ depends on the unknown F. We are currently developing estimators of α_0 when we do not exactly know F_b but only have an estimator of F_b (e.g., we observe a second i.i.d. sample from F_b). Investigating consistent alternative ways of detecting the "elbow" of the function $\gamma d_n(\hat{F}_{s,n}^{\gamma}, \check{F}_{s,n}^{\gamma})$, as an estimator of $\tilde{\alpha}_0$, is an interesting future research direction. As we have observed in the astronomy application, formal goodness-of-fit tests for F_s are important – they can guide the practitioner to use appropriate parametric models for further analysis – but are presently unknown. The p-values in the prostate data example, considered in Section 2.9.1, can have slight dependence. Therefore, investigating the performance and properties of the methods introduced in this paper under appropriate dependence assumptions on $X_1, \ldots, X_n$ is another important direction for future research.

2.11 Identifiability of F_s

In this section we continue the discussion on the identifiability of F_s. First, we give some remarks to illustrate Lemmas 3 and 4.

Remark 2. *We consider mixtures of Poisson and binomial distributions to illustrate Lemma 3. If F_s is Poisson(λ_s) and F_b is Poisson(λ_b), then*

$$\inf_{x \in d(F_b)} \frac{J_{F_s(x)}}{J_{F_b(x)}} = \inf_{k \in \mathbb{N} \cup \{0\}} \frac{\lambda_s^k \exp(-\lambda_s)}{\lambda_b^k \exp(-\lambda_b)} = \exp(\lambda_b - \lambda_s) \inf_{k \in \mathbb{N} \cup \{0\}} \left(\frac{\lambda_s}{\lambda_b}\right)^k.$$

By an application of Lemma 3, we have if $\lambda_s < \lambda_b$ *then* $\alpha_0 = \alpha$*; otherwise* $\alpha_0 = \alpha(1 - \exp(\lambda_b - \lambda_s))$.

In the case of a binomial mixture, i.e., $F_s = Bin(n, p_s)$ *and* $F_b = Bin(n, p_b)$,

$$\alpha_0 = \begin{cases} \alpha\left[1 - (\frac{1-p_s}{1-p_b})^n\right], & p_s \geq p_b, \\ \alpha\left[1 - (\frac{p_s}{p_b})^n\right], & p_s < p_b. \end{cases}$$

Remark 3. *If* F_s *is* $N(\mu_s, \sigma_s^2)$ *and* F_b $(\neq F_s)$ *is* $N(\mu_b, \sigma_b^2)$ *then it can be easily shown that the problem is identifiable if and only if* $\sigma_s \leq \sigma_b$*. When* $\sigma_s > \sigma_b$*, the model is not identifiable, an application of Lemma 4 gives* $\alpha_0 = \alpha\big[1 - (\sigma_b/\sigma_s)\exp\big(-\sigma_s\sigma_b(\mu_b - \mu_s)^2/2\big)\big]$*. Thus,* α_0 *increases to* α *as* $|\mu_s - \mu_b|$ *tends to infinity. It should be noted that the problem is actually identifiable if we restrict ourselves to the parametric family of a two-component Gaussian mixture model.*

Remark 4. *Now consider a mixture of exponential random variables, i.e.,* F_s *is* $E(a_s, \sigma_s)$ *and* F_b $(\neq F_s)$ *is* $E(a_b, \sigma_b)$*, where* $E(a, \sigma)$ *is the distribution that has the density* $(1/\sigma)\exp(-(x-a)/\sigma)\mathbf{1}_{(a,\infty)}(x)$*. In this case, the problem is identifiable if* $a_s > a_b$*, as this implies the support of* F_s *is a proper subset of the support of* F_b*. But when* $a_s \leq a_b$*, the problem is identifiable if and only if* $\sigma_s \leq \sigma_b$.

Remark 5. *It is also worth pointing out that even in cases where the problem is not identifiable the difference between the true mixing proportion* α *and the estimand* α_0 *may be very small. Consider the hypothesis test* $H_0 : \theta = 0$ *versus* $H_1 : \theta \neq 0$ *for the model* $N(\theta, 1)$ *with test statistic* $\bar{X}$*. The density of the p-values under* θ *is*

$$f_\theta(p) = \frac{1}{2}e^{-m\theta^2/2}[e^{-\sqrt{m}\theta^2\Phi^{-1}(1-p/2)} + e^{\sqrt{m}\theta^2\Phi^{-1}(1-p/2)}],$$

where m *is the sample size. Here* $f_\theta(1) = e^{-m\theta^2/2} > 0$*, so the model is not identifiable. As* F_b *is uniform, it can be easily verified that* $\alpha_0 = \alpha - \alpha\inf_p f_\theta(p)$*. However, as the value of* f_θ *decreases exponentially with* m*, in many practical situations, where* m *is not too small, the difference between* α *and* α_0 *will be negligible.*

In the following lemma, we try to find the relationship between α and α_0 when F is a general CDF.

Lemma 10. *Suppose that*

$$F = \kappa F^{(a)} + (1-\kappa)F^{(d)}, \tag{2.12}$$

where $F^{(a)}$ is an absolutely continuous CDF and $F^{(d)}$ is a piecewise constant CDF, for some $\kappa \in (0,1)$. Then

$$\alpha_0 = \alpha - \min\left\{\frac{\alpha\kappa_s - \alpha_0^{(a)}\kappa}{\kappa_b}, \frac{\alpha(1-\kappa_s) - \alpha_0^{(d)}(1-\kappa)}{(1-\kappa_b)}\right\},$$

where $\alpha_0^{(a)}$ and $\alpha_0^{(d)}$ are defined as in (2.4), but with $\{F^{(a)}, F_b^{(a)}\}$ and $\{F^{(d)}, F_b^{(d)}\}$, respectively (instead of $\{F, F_b\}$). Similarly, κ_s and κ_b are defined as in (2.12), but for F_s and F_b, respectively.

Proof. From the definition of κ_s and κ_b, we have $F_s = \kappa_s F_s^{(a)} + (1-\kappa_s)F_s^{(d)}$, and $F_b = \kappa_b F_b^{(a)} + (1-\kappa_b)F_b^{(d)}$. Thus from (2.1), we get

$$F = \alpha\kappa_s F_s^{(a)} + (1-\alpha)\kappa_b F_b^{(a)} + \alpha(1-\kappa_s)F_s^{(d)} + (1-\alpha)(1-\kappa_s)F_b^{(d)}.$$

Now using the definition of κ, we see that $\kappa = \alpha\kappa_s + (1-\alpha)\kappa_b$, $1-\kappa = \alpha(1-\kappa_s) + (1-\alpha)(1-\kappa_b)$. If we write

$$F^{(a)} = \alpha^{(a)}F_s^{(a)} + (1-\alpha^{(a)})F_b^{(a)},$$

it can easily seen that $\alpha^{(a)} = \frac{\alpha\kappa_s}{\kappa}$; and similarly, $\alpha^{(d)} = \frac{\alpha(1-\kappa_s)}{1-\kappa}$. Then, we can find $\alpha_0^{(d)}$ and $\alpha_0^{(a)}$ as in Lemmas 3 and 4, respectively. Note that

$$\begin{aligned}
& \sup\{0 \le \epsilon \le 1 : \alpha F_s - \epsilon F_b \text{ is a sub-CDF}\} \\
= \; & \sup\left\{0 \le \epsilon \le 1 : \alpha(\kappa_s F_s^{(a)} + (1-\kappa_s)F_s^{(d)}) - \epsilon(\kappa_b F_b^{(a)} + (1-\kappa_b)F_b^{(d)}) \text{ is a sub-CDF}\right\} \\
= \; & \sup\left\{0 \le \epsilon \le 1 : \text{both } \alpha\kappa_s F_s^{(a)} - \epsilon\kappa_b F_b^{(a)}, \alpha(1-\kappa_s)F_s^{(d)} - \epsilon(1-\kappa_b)F_b^{(d)} \text{ are sub-CDFs}\right\} \\
= \; & \min\left(\sup\left\{0 \le \epsilon \le 1 : \alpha\kappa_s F_s^{(a)} - \epsilon\kappa_b F_b^{(a)} \text{ is a sub-CDF}\right\},\right. \\
& \qquad \left.\sup\left\{0 \le \epsilon \le 1 : \alpha(1-\kappa_s)F_s^{(d)} - \epsilon(1-\kappa_b)F_b^{(d)} \text{ is a sub-CDF}\right\}\right) \\
= \; & \min\left(\frac{\alpha\kappa_s}{\kappa_b}\operatorname{ess\,inf}\frac{f_s^{(a)}}{f_b^{(a)}}, \frac{\alpha(1-\kappa_s)}{(1-\kappa_b)}\inf_{x\in d(F_b^{(d)})}\frac{J_{F_s^{(d)}}(x)}{J_{F_b^{(d)}}(x)}\right) \\
= \; & \min\left(\frac{(\alpha\kappa_s - \alpha_0^{(a)}\kappa)}{\kappa_b}, \frac{(\alpha(1-\kappa_s) - \alpha_0^{(d)}(1-\kappa))}{(1-\kappa_b)}\right),
\end{aligned}$$

where J_G and $d(J_G)$ are defined before Lemma 3 and we use the notion that $\frac{0}{0} = 1$. Hence, by (2.5) the result follows. □

Lemma 5 is now a corollary of this result.

2.12 Performance comparison of $\hat{\alpha}_0^{c_n}$, $\hat{\alpha}_0^{CV}$, and $\tilde{\alpha}_0$

In Figs. 7 and 8 we present further simulation experiments to investigate the finite sample performance of $\hat{\alpha}_0^{c_n}$, $\hat{\alpha}_0^{CV}$, and $\tilde{\alpha}_0$ across different simulation scenarios. In each setting we also include the performance of the best performing competing estimators discussed in Section 2.8.2.

2.13 Detection of sparse heterogeneous mixtures

In this section we draw a connection between the lower confidence bound developed in Section 2.4 and the *Higher Criticism* method of [Donoho and Jin, 2004] for detection of sparse heterogeneous mixtures. The detection of heterogeneity in sparse models arises in many applications, e.g., detection of a disease outbreak (see [Kulldorff *et al.*, 2005]) or early detection of bioweapons use (see [Donoho and Jin, 2004]). Generally, in large scale multiple testing problems, when the non-null effect is sparse it is important to detect the existence of non-null effects (see [Cai *et al.*, 2007]).

[Donoho and Jin, 2004] consider n i.i.d. data from one of the two possible situations:

$$H_0 : X_i \sim F_b, \quad 1 \le i \le n,$$
$$H_1^{(n)} : X_i \sim F^n := \alpha_n F_{n,s} + (1-\alpha_n) F_b, \quad 1 \le i \le n,$$

where $\alpha_n \sim n^{-\lambda}$ and $F_{n,s}$ is such that $d(F_{n,s}, F_b)$ is bounded away from 0. In [Donoho and Jin, 2004] the main focus is on testing H_0, i.e., $\alpha_n = 0$. We can test this hypothesis by rejecting H_0 when $\hat{\alpha}_L > 0$. The following lemma shows that indeed this yields a valid testing procedure for $\lambda < 1/2$.

Theorem 10. *If $\alpha_n \sim n^{-\lambda}$, for $\lambda < 1/2$, then $P_{H_0}(\text{Reject } H_0) = \beta$ and $P_{H_1^{(n)}}(\hat{\alpha}_L > 0) \to 1$ as $n \to \infty$.*

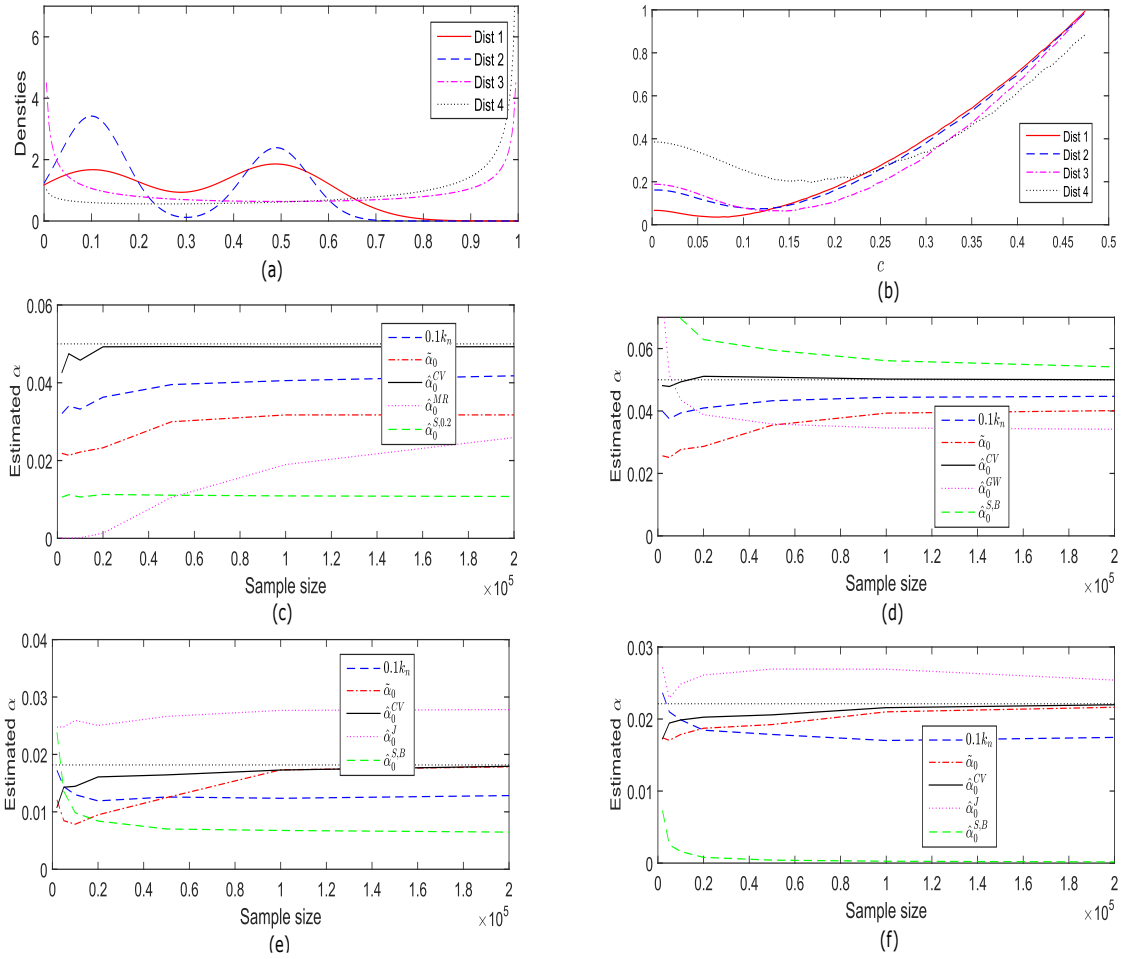

Figure 2.7: Plots comparing the performance of $\hat{\alpha}_0^{c_n}$, $\hat{\alpha}_0^{CV}$, and $\tilde{\alpha}_0$; (a) density functions for four different choices of F_s; (b) plot of the average of $\sum_{k=1}^{K}\int(\mathbb{F}_n^k - \hat{F}^k)^2 d\mathbb{F}_n^k$ (see (2.11)), as a function of c, computed over 500 independent samples of size 50000 corresponding to Dist 1-4; (c)-(f) gives the means of different competing estimators of α_0, computed over 500 independent samples for Dist 1-4 respectively (in each figure the horizontal dotted black line denotes the true α_0).

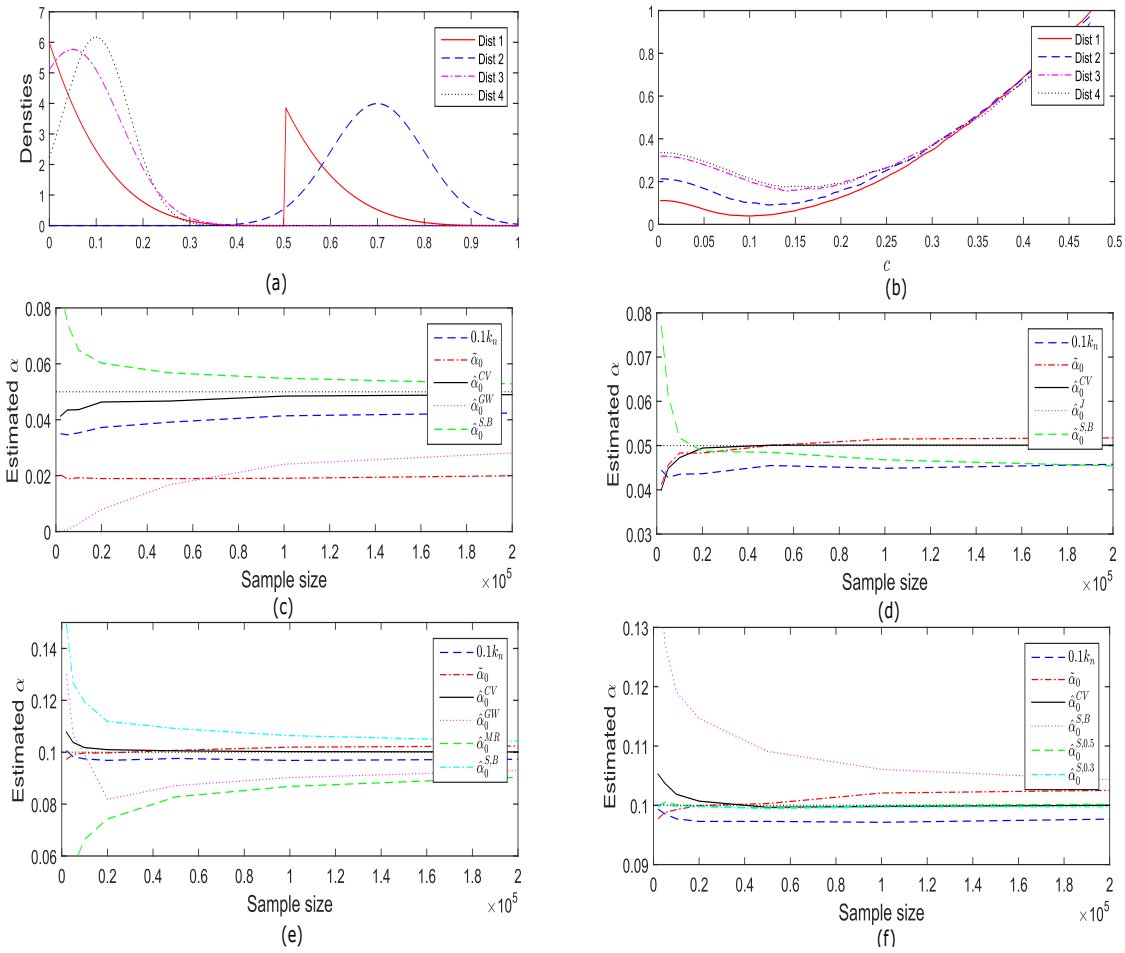

Figure 2.8: Plots comparing the performance of $\hat{\alpha}_0^{c_n}$, $\hat{\alpha}_0^{CV}$, and $\tilde{\alpha}_0$; (a) density functions for four different choices of F_s; (b) plot of the average of $\sum_{k=1}^{K}\int(\mathbb{F}_n^k - \hat{F}^k)^2 d\mathbb{F}_n^k$ (see (2.11)), as a function of c, computed over 500 independent samples of size 50000 corresponding to Dist 1-4; (c)-(f) gives the means of different competing estimators of α_0, computed over 500 independent samples for Dist 1-4 respectively (in each figure the horizontal dotted black line denotes the true α_0).

Proof. Note that $\{\hat{\alpha}_L > 0\}$ is equivalent to $\{c_n \leq \sqrt{n} d_n(\mathbb{F}_n, F_b)\}$ which shows that

$$\begin{aligned} c_n &\leq \sqrt{n} d_n(\mathbb{F}_n, (1-\alpha_n)F_b + \alpha_n F_{n,s}) + \sqrt{n} d_n(\alpha_n F_b, \alpha_n F_{n,s}) \\ &= \sqrt{n} d_n(\mathbb{F}_n, F^n) + \alpha_n \sqrt{n} d_n(F_{n,s}, F_b), \end{aligned}$$

where c_n is chosen as in Theorem 5. It is easy to see that $\sqrt{n} d_n(\mathbb{F}_n, F^n)$ is $O_P(1)$ and $\alpha_n \sqrt{n} d_n(F_{n,s}, F_b) \to \infty$, for $\lambda < 1/2$, which shows that $P_{H_1^{(n)}}(\hat{\alpha}_L > 0) \to 1$. It can be easily seen that $P_{H_0}(\hat{\alpha}_L > 0) = P_{H_0}(\text{Reject } H_0) = \beta$. □

2.14 Proofs of remaining theorems and lemmas

2.14.1 Proof of Lemma 2

From the definition of α_0, we have

$$\begin{aligned} \alpha_0 &= \inf\{0 \leq \gamma \leq \alpha : [F - (1-\gamma)F_b]/\gamma \text{ is a valid CDF}\} \\ &= \inf\{0 \leq \gamma \leq \alpha : [\alpha F_s + (1-\alpha)F_b - (1-\gamma)F_b]/\gamma \text{ is a valid CDF}\} \\ &= \inf\{0 \leq \gamma \leq \alpha : [\alpha F_s - (\alpha-\gamma)F_b]/\gamma \text{ is a valid CDF}\} \\ &= \alpha - \sup\{0 \leq \epsilon \leq \alpha : \alpha F_s - \epsilon F_b \text{ is a sub-CDF}\} \\ &= \alpha - \sup\{0 \leq \epsilon \leq 1 : \alpha F_s - \epsilon F_b \text{ is a sub-CDF}\}, \end{aligned}$$

where the final equality follows from the fact that if $\epsilon > \alpha$, then $\alpha F_s - \epsilon F_b$ will not be a sub-CDF.

To show that $\alpha_0 = 0$ if and only if $F = F_b$ let us define $\delta = \alpha - \epsilon$. Note that $\alpha_0 = 0$, if and only if

$$\begin{aligned} &\sup\{0 \leq \epsilon \leq 1 : \alpha F_s - \epsilon F_b \text{ is a sub-CDF}\} = \alpha \\ \Leftrightarrow\ &\inf\{0 \leq \delta \leq 1 : \alpha(F_s - F_b) + \delta F_b \text{ is a sub-CDF}\} = 0. \end{aligned}$$

However, it is easy to see that the last equality is true if and only if $F_s - F_b \equiv 0$.

2.14.2 Proof of Lemma 3

When $d(F_b) \not\subset d(F_s)$, there exists a $x \in d(F_b) - d(F_s)$, i.e., there exists a x which satisfies $F_b(x) - F_b(x-) > 0$ and $F_s(x) - F_s(x-) = 0$. Then for all $\epsilon > 0$, $F_s(x-) - \epsilon F_b(x-) >$

$F_s(x) - \epsilon F_b(x)$. This shows that $F_s - \epsilon F_b$ cannot be a sub-CDF, and hence by Lemma 2 the model is identifiable. Now let us assume that $d(F_b) \subset d(F_s)$.

$$\begin{aligned}\{0 \le \epsilon \le 1 : \alpha F_s - \epsilon F_b \text{ is a sub-CDF}\} &= \{0 \le \epsilon \le 1 : \alpha J_{F_s}(x) - \epsilon J_{F_b}(x) \ge 0,\ \forall x \in d(J_{F_b})\} \\ &= \left\{0 \le \epsilon \le 1 : \frac{J_{F_s}(x)}{J_{F_b}(x)} \ge \frac{\epsilon}{\alpha},\ \forall x \in d(J_{F_b})\right\} \\ &= \left\{0 \le \epsilon \le 1 : \inf_{x \in d(F_b)} \frac{J_{F_s}(x)}{J_{F_b}(x)} \ge \frac{\epsilon}{\alpha}\right\}.\end{aligned}$$

Therefore, using (2.5), we get the desired result.

2.14.3 Proof of Lemma 4

From (2.5), we have

$$\begin{aligned}\alpha_0 &= \alpha - \sup\{0 \le \epsilon \le 1 : \alpha F_s - \epsilon F_b \text{ is a sub-CDF}\} \\ &= \alpha - \sup\{0 \le \epsilon \le 1 : \alpha f_s(x) - \epsilon f_b(x) \ge 0 \text{ almost every } x\} \\ &= \alpha - \sup\left\{0 \le \epsilon \le 1 : \alpha \frac{f_s}{f_b}(x) \ge \epsilon \text{ almost every } x\right\} \\ &= \alpha\left\{1 - \operatorname{ess\,inf} \frac{f_s}{f_b}\right\}.\end{aligned}$$

2.14.4 Proof of Theorem 1

Without loss of generality, we can assume that F_b is the uniform distribution on $(0, 1)$ and, for clarity, in the following we write U instead of F_b. Let us define

$$\begin{aligned}A &:= \left\{\gamma \in (0, 1] : \frac{F - (1-\gamma)U}{\gamma} \text{ is a valid CDF}\right\}, \\ A^Y &:= \left\{\gamma \in (0, 1] : \frac{G - (1-\gamma)U \circ \Psi}{\gamma} \text{ is a valid CDF}\right\}.\end{aligned}$$

Since $\alpha_0 = \inf A$, and $\alpha_0^Y = \inf A^Y$ for the first part of the theorem it is enough to show that $A = A^Y$. Let us first show that $A^Y \subset A$. Suppose $\eta \in A^Y$. We first show that $(F - (1-\eta)U)/\eta$ is a non-decreasing function. For all $t_1 \le t_2$, we have that

$$\frac{G(t_1) - (1-\eta)U(\Psi(t_1))}{\eta} \le \frac{G(t_2) - (1-\eta)U(\Psi(t_2))}{\eta}.$$

Let $y_1 \le y_2$. Then,

$$\frac{G(\Psi^{-1}(y_1)) - (1-\eta)U(\Psi(\Psi^{-1}(y_1)))}{\eta} \le \frac{G(\Psi^{-1}(y_2)) - (1-\eta)U(\Psi(\Psi^{-1}(y_2)))}{\eta},$$

since $y_1 \leq y_2 \Rightarrow \Psi^{-1}(y_1) \leq \Psi^{-1}(y_2)$. However, as Ψ is continuous, $\Psi(\Psi^{-1}(y)) = y$ and $G(\Psi^{-1}(y)) = \alpha F_s(y) + (1-\alpha)U(y) = F(y)$. Hence, we have

$$\frac{F(y_1) - (1-\eta)U(y_1)}{\eta} \leq \frac{F(y_2) - (1-\eta)U(y_2)}{\eta}.$$

As F and U are CDFs, it is easy to see that $\lim_{x\to-\infty}(F(x) - (1-\eta)U(x))/\eta = 0$, $\lim_{x\to\infty}(F(x) - (1-\eta)U(x))/\eta = 1$ and $(F - (1-\eta)U)/\eta$ is a right continuous function. Hence, for $\eta \in A^Y$, $(F - (1-\eta)U)/\eta$ is a CDF and thus, $\eta \in A$. We can similarly prove $A \subset A^Y$. Therefore, $A = A^Y$ and $\alpha_0 = \alpha_0^Y$.

Note that

$$\gamma d_n(\hat{F}^\gamma_{s,n}, \check{F}^\gamma_{s,n}) = \min_{W\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n \{W(X_i) - \hat{F}^\gamma_{s,n}(X_i)\}^2,$$

where $\mathcal{F}$ is the class of all CDFs. For the second part of theorem it is enough to show that

$$\min_{W\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n \{W(X_i) - \hat{F}^\gamma_{s,n}(X_i)\}^2 = \min_{B\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n \{B(Y_i) - \hat{G}^\gamma_{s,n}(Y_i)\}^2.$$

First note that

$$\begin{aligned}
\mathbb{G}_n(y) &= \frac{1}{n}\sum_{i=1}^n \mathbf{1}\{\Psi^{-1}(X_i) \leq y\} \\
&= \frac{1}{n}\sum_{i=1}^n \mathbf{1}\{X_i \leq \Psi(y)\} \\
&= \mathbb{F}_n(\Psi(y)).
\end{aligned}$$

Thus, from the definition of $\hat{G}^\gamma_{s,n}$, we have

$$\begin{aligned}
\hat{G}^\gamma_{s,n}(Y_i) &= \frac{\mathbb{F}_n(\Psi(Y_i)) - (1-\gamma)U(\Psi(Y_i))}{\gamma} \\
&= \frac{\mathbb{F}_n(X_i) - (1-\gamma)U(X_i)}{\gamma} = \hat{F}^\gamma_{s,n}(X_i).
\end{aligned}$$

Therefore,

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^n \{B(Y_i) - \hat{G}^\gamma_{s,n}(Y_i)\}^2 &= \frac{1}{n}\sum_{i=1}^n \{B(Y_i) - \hat{F}^\gamma_{s,n}(X_i)\}^2 \\
&= \frac{1}{n}\sum_{i=1}^n \{B(\Psi^{-1}(X_i)) - \hat{F}^\gamma_{s,n}(X_i)\}^2 \\
&= \frac{1}{n}\sum_{i=1}^n \{W(X_i) - \hat{F}^\gamma_{s,n}(X_i)\}^2,
\end{aligned}$$

where $W(x) := B(\Psi^{-1}(x))$. W is a valid CDF as Ψ^{-1} is non-decreasing.

2.14.5 Proof of Lemma 6

Letting $F_s^\gamma = (F - (1-\gamma)F_b)/\gamma$, observe that

$$\gamma d_n(\hat{F}_{s,n}^\gamma, F_s^\gamma) = d_n(F, \mathbb{F}_n).$$

Also note that F_s^γ is a valid CDF for $\gamma \geq \alpha_0$. As $\check{F}_{s,n}^\gamma$ is defined as the function that minimises the $L_2(\mathbb{F}_n)$ distance of $\hat{F}_{s,n}^\gamma$ over all CDFs,

$$\gamma d_n(\check{F}_{s,n}^\gamma, \hat{F}_{s,n}^\gamma) \leq \gamma d_n(\hat{F}_{s,n}^\gamma, F_s^\gamma) = d_n(F, \mathbb{F}_n).$$

To prove the second part of the lemma, notice that for $\gamma \geq \alpha_0$ the result follows from above and the fact that $d_n(F, \mathbb{F}_n) \overset{a.s.}{\to} 0$ as $n \to \infty$.

For $\gamma < \alpha_0$, F_s^γ is not a valid CDF, by the definition of α_0. Note that as $n \to \infty$, $\hat{F}_{s,n}^\gamma \overset{a.s.}{\to} F_s^\gamma$ point-wise. So, for large enough n, $\hat{F}_{s,n}^\gamma$ is not a valid CDF, whereas $\check{F}_{s,n}^\gamma$ is always a CDF. Thus, $d_n(\hat{F}_{s,n}^\gamma, \check{F}_{s,n}^\gamma)$ converges to something positive.

2.14.6 Proof of Lemma 7

Assume that $\gamma_1 \leq \gamma_2$ and $\gamma_1, \gamma_2 \in A_n$. If $\gamma_3 = \eta\gamma_1 + (1-\eta)\gamma_2$, for $0 \leq \eta \leq 1$, it is easy to observe from (2.2) that

$$\eta(\gamma_1 \hat{F}_{s,n}^{\gamma_1}) + (1-\eta)(\gamma_2 \hat{F}_{s,n}^{\gamma_2}) = \gamma_3 \hat{F}_{s,n}^{\gamma_3}.$$

Note that $[\eta(\gamma_1 \check{F}_{s,n}^{\gamma_1}) + (1-\eta)(\gamma_2 \check{F}_{s,n}^{\gamma_2})]/\gamma_3$ is a valid CDF, and thus from the definition of $\check{F}_{s,n}^{\gamma_3}$, we have

$$\begin{aligned} d_n(\hat{F}_{s,n}^{\gamma_3}, \check{F}_{s,n}^{\gamma_3}) &\leq d_n\left(\hat{F}_{s,n}^{\gamma_3}, [\eta(\gamma_1 \check{F}_{s,n}^{\gamma_1}) + (1-\eta)(\gamma_2 \check{F}_{s,n}^{\gamma_2})]/\gamma_3\right) \\ &= d_n\Big(\frac{\eta(\gamma_1 \hat{F}_{s,n}^{\gamma_1}) + (1-\eta)(\gamma_2 \hat{F}_{s,n}^{\gamma_2})}{\gamma_3}, \frac{\eta(\gamma_1 \check{F}_{s,n}^{\gamma_1}) + (1-\eta)(\gamma_2 \check{F}_{s,n}^{\gamma_2})}{\gamma_3}\Big) \\ &\leq \frac{\eta\gamma_1}{\gamma_3} d_n(\hat{F}_{s,n}^{\gamma_1}, \check{F}_{s,n}^{\gamma_1}) + \frac{(1-\eta)\gamma_2}{\gamma_3} d_n(\hat{F}_{s,n}^{\gamma_2}, \check{F}_{s,n}^{\gamma_2}) \end{aligned}$$

where the last step follows from the triangle inequality. But as $\gamma_1, \gamma_2 \in A_n$, the above inequality yields

$$d_n(\hat{F}_{s,n}^{\gamma_3}, \check{F}_{s,n}^{\gamma_3}) \leq \frac{\eta\gamma_1}{\gamma_3}\frac{c_n}{\sqrt{n}\gamma_1} + \frac{(1-\eta)\gamma_2}{\gamma_3}\frac{c_n}{\sqrt{n}\gamma_2} = \frac{c_n}{\sqrt{n}\gamma_3}.$$

Thus $\gamma_3 \in A_n$.

2.14.7 Proof of Theorem 2

We need to show that $P(|\hat{\alpha}_0^{c_n} - \alpha_0| > \epsilon) \to 0$ for any $\epsilon > 0$. Let us first show that

$$P(\hat{\alpha}_0^{c_n} - \alpha_0 < -\epsilon) \to 0.$$

The statement is obviously true if $\alpha_0 \leq \epsilon$. So let us assume that $\alpha_0 > \epsilon$. Suppose $\hat{\alpha}_0^{c_n} - \alpha_0 < -\epsilon$, i.e., $\hat{\alpha}_0^{c_n} < \alpha_0 - \epsilon$. Then by the definition of $\hat{\alpha}_0^{c_n}$ and the convexity of A_n, we have $(\alpha_0 - \epsilon) \in A_n$ (as A_n is a convex set in $[0, 1]$ with $1 \in A_n$ and $\hat{\alpha}_0^{c_n} \in A_n$), and thus

$$d_n(\hat{F}_{s,n}^{\alpha_0-\epsilon}, \check{F}_{s,n}^{\alpha_0-\epsilon}) \leq \frac{c_n}{\sqrt{n}(\alpha_0 - \epsilon)}. \tag{2.14}$$

But by (2.9) the left-hand side of (2.14) goes to a non-zero constant in probability. Hence, if $c_n/\sqrt{n} \to 0$,

$$P(\hat{\alpha}_0^{c_n} - \alpha_0 < -\epsilon) \leq P\left(d_n(\hat{F}_{s,n}^{\alpha_0-\epsilon}, \check{F}_{s,n}^{\alpha_0-\epsilon}) \leq \frac{c_n}{\sqrt{n}(\alpha_0 - \epsilon)}\right) \to 0.$$

This completes the proof of the first part of the claim.

Now suppose that $\hat{\alpha}_0^{c_n} - \alpha_0 > \epsilon$. Then,

$$\begin{aligned} \hat{\alpha}_0^{c_n} - \alpha_0 > \epsilon \quad &\Rightarrow \quad \sqrt{n} d_n(\hat{F}_{s,n}^{\alpha_0+\epsilon}, \check{F}_{s,n}^{\alpha_0+\epsilon}) \geq \frac{c_n}{\alpha_0 + \epsilon} \\ &\Rightarrow \quad \sqrt{n} d_n(\mathbb{F}_n, F) \geq c_n. \end{aligned}$$

The first implication follows from the definition of $\hat{\alpha}_0^{c_n}$, while the second implication is true by Lemma 6. The right-hand side of the last inequality is (asymptotically similar to) the Cramér–von Mises statistic for which the asymptotic distribution is well-known and thus if $c_n \to \infty$ the result follows.

2.14.8 Proof of Lemma 8

As $\alpha_0 = 0$,

$$P(\hat{\alpha}_0^{c_n} = 0) = 1 - P(\hat{\alpha}_0^{c_n} > 0) = 1 - P(\sqrt{n} d_n(\mathbb{F}_n, F) > c_n) \to 1,$$

since $\sqrt{n} d_n(\mathbb{F}_n, F) = O_P(1)$ by Theorem 6.

2.14.9 Proof of Theorem 3

As the proof of this result is slightly involved we break it into a number of lemmas (whose proofs are provided later in this sub-section) and give the main arguments below.

We need to show that given any $\epsilon > 0$, we can find an $M > 0$ and $n_0 \in \mathbb{N}$ (depending on ϵ) for which $\sup_{n>n_0} P(r_n|\hat{\alpha}_0^{c_n} - \alpha_0| > M) \leq \epsilon$.

Lemma 11. *If $c_n \to \infty$, then for any $M > 0$, $\sup_{n>n_0} P\left(r_n(\hat{\alpha}_0^{c_n} - \alpha_0) > M\right) < \epsilon$, for large enough $n_0 \in \mathbb{N}$.*

Finding an r_n such that $P\left(r_n(\hat{\alpha}_0^{c_n} - \alpha_0) < -M\right) < \epsilon$ for large enough n is more complicated. We start with some notation. Let $\mathcal{F}$ be the class of all CDFs and $\mathbb{H}$ be the Hilbert space $L_2(F) := \{f : \mathbb{R} \to \mathbb{R} | \int f^2 dF < \infty\}$. For a closed convex subset $\mathcal{K}$ of $\mathbb{H}$ and $h \in \mathbb{H}$, we define the projection of h onto $\mathcal{K}$ as

$$\Pi(h|\mathcal{K}) := \arg\min_{f \in \mathcal{K}} d(f, h), \tag{2.15}$$

where d stands for the $L_2(F)$ distance, i.e., if $g, h \in \mathbb{H}$, then $d^2(g, h) = \int (g - h)^2 dF$. We define the tangent cone of $\mathcal{F}$ at $f_0 \in \mathcal{F}$, as

$$T_{\mathcal{F}}(f_0) := \{\lambda(f - f_0) : \lambda \geq 0, f \in \mathcal{F}\}. \tag{2.16}$$

For any $H \in \mathcal{F}$ and $\gamma > 0$, let us define

$$\hat{H}^{\gamma} := \frac{H - (1-\gamma)F_b}{\gamma}, \quad \check{H}_n^{\gamma} := \arg\min_{G \in \mathcal{F}} \gamma d_n(\hat{H}^{\gamma}, G), \quad \text{and} \quad \bar{H}_n^{\gamma} := \arg\min_{G \in \mathcal{F}} \gamma d(\hat{H}^{\gamma}, G).$$

For $H = \mathbb{F}_n$ and $\gamma = \alpha_0$ we define the three quantities above and call them $\hat{F}_{s,n}^{\alpha_0}$, $\check{F}_{s,n}^{\alpha_0}$, and $\bar{F}_{s,n}^{\alpha_0}$ respectively. Note that

$$P\left(r_n(\hat{\alpha}_0^{c_n} - \alpha_0) < -M\right) = P(\sqrt{n}\gamma_n \; d_n(\hat{F}_{s,n}^{\gamma_n}, \check{F}_{s,n}^{\gamma_n}) < c_n), \tag{2.17}$$

where $\gamma_n = \alpha_0 - M/r_n$. To study the limiting behavior of $d_n(\hat{F}_{s,n}^{\gamma_n}, \check{F}_{s,n}^{\gamma_n})$ we break it as the sum of $d_n(\hat{F}_{s,n}^{\gamma_n}, \check{F}_{s,n}^{\gamma_n}) - d(\hat{F}_{s,n}^{\gamma_n}, \bar{F}_{s,n}^{\gamma_n})$ and $d(\hat{F}_{s,n}^{\gamma_n}, \bar{F}_{s,n}^{\gamma_n})$. The following two lemmas (proved in Sections 2.14.9.2 and 2.14.9.3 respectivley) give the asymptotic behavior of the two terms. The proof of Lemma 13 uses the functional delta method (cf. Theorem 20.8 of [Van der Vaart, 1998a]) for the projection operator; see Theorem 1 of [Fils-Villetard *et al.*, 2008].

Lemma 12. *If* $\sqrt{n}/r_n^2 \to 0$, *then* $U_n := \sqrt{n}\gamma_n d_n(\hat{F}_{s,n}^{\gamma_n}, \check{F}_{s,n}^{\gamma_n}) - \sqrt{n}\gamma_n d(\hat{F}_{s,n}^{\gamma_n}, \bar{F}_{s,n}^{\gamma_n}) \xrightarrow{P} 0$.

Lemma 13. *If* $c_n \to \infty$, *then*

$$\frac{\sqrt{n}\gamma_n}{c_n M} d(\hat{F}_{s,n}^{\gamma_n}, \bar{F}_{s,n}^{\gamma_n}) \xrightarrow{P} \left\{ \int V^2 dF \right\}^{1/2} > 0$$

where

$$V := (F_s^{\alpha_0} - F_b) - \Pi\left(F_s^{\alpha_0} - F_b | \, T_{\mathcal{F}}(F_s^{\alpha_0})\right) \neq 0$$

and

$$F_s^{\alpha_0} := \frac{F - (1-\alpha_0)F_b}{\alpha_0}. \tag{2.18}$$

Using (2.17), and the notation introduced in the above two lemmas we see that

$$P\left(r_n(\hat{\alpha}_0^{c_n} - \alpha_0) < -M\right) \quad = \quad P\left(\frac{1}{c_n}U_n + \frac{\sqrt{n}\gamma_n}{c_n} d(\hat{F}_{s,n}^{\gamma_n}, \bar{F}_{s,n}^{\gamma_n}) < 1\right).$$

However, $U_n \xrightarrow{P} 0$ (by Lemma 12) and $\frac{\sqrt{n}\gamma_n}{c_n M} d(\hat{F}_{s,n}^{\gamma_n}, \bar{F}_{s,n}^{\gamma_n}) \xrightarrow{P} \int V^2 dF$ (by Lemma 13). The result now follows from (2.19), by taking a large enough M.

2.14.9.1 Proof of Lemma 11

Note that

$$\begin{aligned} P(r_n(\hat{\alpha}_0^{c_n} - \alpha_0) > M) \leq P\left(\hat{\alpha}_0^{c_n} > \alpha_0\right) &= P\left(\sqrt{n}\alpha_0 d_n(\hat{F}_{s,n}^{\alpha_0}, \check{F}_{s,n}^{\alpha_0}) > c_n\right) \\ &\leq P\left(\sqrt{n}\alpha_0 d_n(\hat{F}_{s,n}^{\alpha_0}, F_s^{\alpha_0}) > c_n\right) \\ &= P\left(\sqrt{n} d_n(\mathbb{F}_n, F) > c_n\right) \to 0, \end{aligned}$$

as $c_n \to \infty$, since $\sqrt{n} d_n(\mathbb{F}_n, F) = O_P(1)$. Therefore, the result holds for sufficiently large n.

2.14.9.2 Proof of Lemma 12

It is enough to show that

$$W_n := n\gamma_n^2 d_n^2(\hat{F}_{s,n}^{\gamma_n}, \check{F}_{s,n}^{\gamma_n}) - n\gamma_n^2 d^2(\hat{F}_{s,n}^{\gamma_n}, \bar{F}_{s,n}^{\gamma_n}) \quad \xrightarrow{P} \quad 0,$$

since $U_n^2 \leq |W_n|$. Note that

$$\begin{aligned} \check{F}_{s,n}^{\gamma_n} &= \underset{G\in\mathcal{F}}{\arg\min}\, d_n(\mathbb{F}_n, \gamma_n G + (1-\gamma_n)F_b), \\ \bar{F}_{s,n}^{\gamma_n} &= \underset{G\in\mathcal{F}}{\arg\min}\, d(\mathbb{F}_n, \gamma_n G + (1-\gamma_n)F_b). \end{aligned}$$

For each positive integer n and $c > 0$, we introduce the following classes of functions:

$$\begin{aligned} \mathcal{G}_c(n) &= \left\{ \sqrt{n}(G - (1-\gamma_n)F_b - \gamma_n \check{G}_n^{\gamma_n})^2 : G \in \mathcal{F},\ \|G - F\| < \frac{c}{\sqrt{n}} \right\}, \\ \mathcal{H}_c(n) &= \left\{ \sqrt{n}(H - (1-\gamma_n)F_b - \gamma_n \bar{H}_n^{\gamma_n})^2 : H \in \mathcal{F},\ \|H - F\| < \frac{c}{\sqrt{n}} \right\}. \end{aligned}$$

Let us also define

$$D_n := \sup_{t\in\mathbb{R}} \sqrt{n}|\mathbb{F}_n(t) - F(t)| = \|\mathbb{F}_n - F\|.$$

From the definition of the minimisers $\check{F}_{s,n}^{\gamma_n}$ and $\bar{F}_{s,n}^{\gamma_n}$, we see that

$$\begin{aligned} \gamma_n^2\, |d_n^2(\hat{F}_{s,n}^{\gamma_n}, \check{F}_{s,n}^{\gamma_n}) - d^2(\hat{F}_{s,n}^{\gamma_n}, \bar{F}_{s,n}^{\gamma_n})| \leq \max\big\{ &|(d_n^2 - d^2)(\mathbb{F}_n, \gamma_n \check{F}_{s,n}^{\gamma_n} + (1-\gamma_n)F_b)|\,, \\ &|(d_n^2 - d^2)(\mathbb{F}_n, \gamma_n \bar{F}_{s,n}^{\gamma_n} + (1-\gamma_n)F_b)| \big\}. \end{aligned}$$

Observe that

$$n\gamma_n^2\, [(d_n^2 - d^2)(\mathbb{F}_n, \gamma_n \check{F}_{s,n}^{\gamma_n} + (1-\gamma_n)F_b)] = \sqrt{n}(\mathbb{P}_n - P)[g_n] = \nu_n(g_n),$$

where $g_n := \sqrt{n}\{\mathbb{F}_n - \gamma_n \check{F}_{s,n}^{\gamma_n} - (1-\gamma_n)F_b\}^2$, $\mathbb{P}_n$ denotes the empirical measure of the data, and $\nu_n := \sqrt{n}(\mathbb{P}_n - P)$ denotes the usual empirical process. Similarly,

$$n\gamma_n^2\, [(d_n^2 - d^2)(\mathbb{F}_n, \gamma_n \bar{F}_{s,n}^{\gamma_n} + (1-\gamma_n)F_b)] = \sqrt{n}(\mathbb{P}_n - P)[h_n] = \nu_n(h_n),$$

where $h_n := \sqrt{n}\{\mathbb{F}_n - \gamma_n \bar{F}_{s,n}^{\gamma_n} - (1-\gamma_n)F_b\}^2$. Thus, combining (2.20), (2.21) and the above two displays, we get, for any $\delta > 0$,

$$P(|W_n| > \delta) \leq P\left(|\nu_n(g_n)| > \delta\right) + P\left(|\nu_n(h_n)| > \delta\right).$$

The first term in the right hand side of (2.22) can be bounded above as

$$\begin{aligned} P(|\nu_n(g_n)| > \delta) &= P(|\nu_n(g_n)| > \delta, g_n \in \mathcal{G}_c(n)) + P(|\nu_n(g_n)| > \delta, g_n \notin \mathcal{G}_c(n)) \\ &\leq P(|\nu_n(g_n)| > \delta, g_n \in \mathcal{G}_c(n)) + P(g_n \notin \mathcal{G}_c(n)) \\ &\leq P\left(\sup_{g \in \mathcal{G}_c(n)} |\nu_n(g)| > \delta\right) + P(g_n \notin \mathcal{G}_c(n)) \\ &\leq \frac{1}{\delta} E\left(\sup_{g \in \mathcal{G}_c(n)} |\nu_n(g)|\right) + P(g_n \notin \mathcal{G}_c(n)) \\ &\leq J_{[\,]} \frac{P[G_{c,n}^2]}{\delta} + P(g_n \notin \mathcal{G}_c(n)), \end{aligned}$$

where $G_{c,n} := 6c^2/\sqrt{n} + 16\sqrt{n}\frac{M^2}{r_n^2}\|F_s^{\alpha_0} - F_b\|^2$ is an envelope for $\mathcal{G}_c(n)$ and $J_{[\,]}$ is a constant. Note that to derive the last inequality, we have used the maximal inequality in Corollary (4.3) of [Pollard, 1989]; the class $\mathcal{G}_c(n)$ is "manageable" in the sense of [Pollard, 1989] (as a consequence of equation (2.5) of [Van de Geer, 2000a]).

To see that $G_{c,n}$ is an envelope for $\mathcal{G}_c(n)$, observe that for any $G \in \mathcal{F}$,

$$G - (1-\gamma_n)F_b = G - F + \frac{M}{r_n}(F_s^{\alpha_0} - F_b) + \gamma_n F_s^{\alpha_0}.$$

Hence,

$$F_s^{\alpha_0} - \frac{M}{r_n\gamma_n}\|F_s^{\alpha_0} - F_b\| - \frac{\|G-F\|}{\gamma_n} \leq \frac{G-(1-\gamma_n)F_b}{\gamma_n} \leq F_s^{\alpha_0} + \frac{M}{r_n\gamma_n}\|F_s^{\alpha_0} - F_b\| + \frac{\|G-F\|}{\gamma_n}.$$

As the two bounds are monotone, from the properties of isotonic estimators (see e.g., Theorem 1.3.4 of [Robertson *et al.*, 1988]), we can always find a version of $\check{G}_s^{\gamma_n}$ such that

$$F_s^{\alpha_0} - \frac{M}{r_n\gamma_n}\|F_s^{\alpha_0} - F_b\| - \frac{\|G-F\|}{\gamma_n} \leq \check{G}_s^{\gamma_n} \leq F_s^{\alpha_0} + \frac{M}{r_n\gamma_n}\|F_s^{\alpha_0} - F_b\| + \frac{\|G-F\|}{\gamma_n}.$$

Therefore,

$$-2\frac{M}{r_n}\|F_s^{\alpha_0} - F_b\| - \|G-F\| \leq \gamma_n\check{G}_s^{\gamma_n} - \gamma_n F_s^{\alpha_0} - \frac{M}{r_n}(F_s^{\alpha_0} - F_b) \leq 2\frac{M}{r_n}\|F_s^{\alpha_0} - F_b\| + \|G-F\|. \tag{2.24}$$

Thus, for $\sqrt{n}(G-(1-\gamma_n)F_b-\gamma_n\check{G}_s^{\gamma_n})^2 \in \mathcal{G}_c(n)$,

$$\begin{aligned}
(G-(1-\gamma_n)F_b-\gamma_n\check{G}_s^{\gamma_n})^2 &= \left[(G-F)+\left(\gamma_n\check{G}_s^{\gamma_n}-\gamma_nF_s^{\alpha_0}-\frac{M}{r_n}(F_b-F_s^{\alpha_0})\right)\right]^2 \\
&\leq 2(G-F)^2+2\left(\gamma_n\check{G}_s^{\gamma_n}-\gamma_nF_s^{\alpha_0}-\frac{M}{r_n}(F_b-F_s^{\alpha_0})\right)^2 \\
&\leq 2\|G-F\|^2+2\left(2\frac{M}{r_n}\|F_s^{\alpha_0}-F_b\|+\|G-F\|\right)^2 \\
&\leq 6\|G-F\|^2+16\frac{M^2}{r_n^2}\|F_s^{\alpha_0}-F_b\|^2 \\
&\leq 6c^2+16\frac{M^2}{r_n^2}\|F_s^{\alpha_0}-F_b\|^2=\frac{G_{c,n}}{\sqrt{n}},
\end{aligned}$$

where the second inequality follows from (2.24). From the definition of g_n and D_n^2, we have $|g_n(t)| \leq \frac{6}{\sqrt{n}}D_n^2+16\sqrt{n}\frac{M^2}{r_n^2}\|F_s^{\alpha_0}-F_b\|^2$, for all $t\in\mathbb{R}$. As $D_n=O_P(1)$, for any given $\epsilon>0$, there exists $c>0$ (depending on ϵ) such that

$$P(g_n\notin\mathcal{G}_c(n)) = P\left(\|\mathbb{F}_n-F\|\geq\frac{c}{\sqrt{n}}\right)=P(D_n\geq c)\leq\epsilon,$$

for all sufficiently large n.

Therefore, for any given $\delta>0$ and $\epsilon>0$, we can make both $J\{6\frac{c^2}{\sqrt{n}}+16\sqrt{n}\frac{M^2}{r_n^2}\|F_s^{\alpha_0}-F_b\|^2\}^2$ and $P(g_n\notin\mathcal{G}_c(n))$ less than ϵ for large enough n and $c(>0)$, using the fact that $\sqrt{n}/r_n^2\to 0$ and (2.25). Thus, $P(|\nu_n(g_n)|>\delta)\leq 2\epsilon$ by (2.23).

A similar analysis can be done for the second term of (2.22). The result now follows.

2.14.9.3 Proof of Lemma 13

Note that

$$\frac{\sqrt{n}\gamma_n}{c_n}(\hat{F}_{s,n}^{\gamma_n}-\bar{F}_{s,n}^{\gamma_n})=\frac{\sqrt{n}\gamma_n}{c_n}(\hat{F}_{s,n}^{\gamma_n}-F_s^{\alpha_0})-\frac{\sqrt{n}\gamma_n}{c_n}(\bar{F}_{s,n}^{\gamma_n}-F_s^{\alpha_0}).$$

However, a simplification yields

$$\frac{\sqrt{n}\gamma_n}{c_n}(\hat{F}_{s,n}^{\gamma_n}-F_s^{\alpha_0})=\frac{1}{c_n}\sqrt{n}(\mathbb{F}_n-F)+\frac{\sqrt{n}M}{c_nr_n\alpha_0}(F-F_b).$$

Since $\sqrt{n}(\mathbb{F}_n-F)/c_n$ is $o_P(1)$, $\sqrt{n}=c_nr_n$, and $F-F_b=\alpha_0(F_s^{\alpha_0}-F_b)$, we have

$$\frac{\sqrt{n}\gamma_n}{c_nM}(\hat{F}_{s,n}^{\gamma_n}-F_s^{\alpha_0})\xrightarrow{P}F_s^{\alpha_0}-F_b \quad \text{in } \mathbb{H}. \tag{2.26}$$

By applying the functional delta method (see Theorem 20.8 of [Van der Vaart, 1998a]) for the projection operator (see Theorem 1 of [Fils-Villetard *et al.*, 2008]) to (2.26), we have

$$\frac{\sqrt{n}\gamma_n}{c_n M}(\bar{F}_{s,n}^{\gamma_n} - F_s^{\alpha_0}) \xrightarrow{P} \Pi\left(F_s^{\alpha_0} - F_b | T_{\mathcal{F}}(F_s^{\alpha_0})\right) \quad \text{in } \mathbb{H}. \tag{2.27}$$

By combining (2.26) and (2.27), we have

$$\frac{\sqrt{n}\gamma_n}{c_n M}(\hat{F}_{s,n}^{\gamma_n} - \bar{F}_{s,n}^{\gamma_n}) \xrightarrow{P} (F_s^{\alpha_0} - F_b) - \Pi\left(F_s^{\alpha_0} - F_b | T_{\mathcal{F}}(F_s^{\alpha_0})\right) \quad \text{in } \mathbb{H}. \tag{2.28}$$

The result now follows by applying the continuous mapping theorem to (2.28). We prove $V \neq 0$ by contradiction. Suppose that $V = 0$, i.e., $(F_s^{\alpha_0} - F_b) \in T_{\mathcal{F}}(F_s^{\alpha_0})$. Therefore, for some distribution function G and $\eta > 0$, we have $V = (\eta+1)F_s^{\alpha_0} - F_b - \eta G$, by the definition of $T_{\mathcal{F}}(F_s^{\alpha_0})$. By the discussion leading to (2.5), it can be easily seen that ηG is a sub-CDF, while $(\eta+1)F_s^{\alpha_0} - F_b$ is not (as that would contradict (2.5)). Therefore, $V \neq 0$ and thus $\int V^2 dF > 0$.

2.14.10 Proof of Theorem 4

The constant c defined in the statement of the theorem can be explicitly expressed as

$$c = -\left\{\int V^2 dF\right\}^{-\frac{1}{2}},$$

where

$$V = (F_s - F_b) - \Pi(F_s - F_b | T_{\mathcal{F}}(F_s)),$$

and Π and $T_{\mathcal{F}}(\cdot)$ are defined in (2.15) and (2.16), respectively.

Let $x > 0$. Obviously,

$$P(r_n(\hat{\alpha}_0^{c_n} - \alpha_0) \leq x) = 1 - P(r_n(\hat{\alpha}_0^{c_n} - \alpha_0) > x).$$

By Lemma 11, we have that $P(r_n(\hat{\alpha}_0^{c_n} - \alpha_0) > x) \to 0$ if $c_n \to \infty$. Now let $x \leq 0$. In this case the left hand side of the above display equals $P(\sqrt{n}\gamma_n d_n(\hat{F}_{s,n}^{\gamma_n}, \check{F}_{s,n}^{\gamma_n}) \leq c_n)$, where $\gamma_n = \alpha_0 + x/r_n$. A simplification yields

$$\frac{\sqrt{n}}{c_n}\gamma_n(\hat{F}_{s,n}^{\gamma_n} - F_s^{\alpha_0}) \xrightarrow{P} -x(F_s^{\alpha_0} - F_b), \text{ in } \mathbb{H}, \tag{2.29}$$

since $\sqrt{n}(\mathbb{F}_n - F)/c_n$ is $o_P(1)$; see the proof of Lemma 13 (Section 2.14.9.3) for the details. By applying the functional delta method (cf. Theorem 20.8 of [Van der Vaart, 1998a]) for the projection operator (see Theorem 1 of [Fils-Villetard *et al.*, 2008]) to (2.29), we have

$$\frac{\sqrt{n}}{c_n}\gamma_n(\bar{F}_{s,n}^{\gamma_n} - F_s^{\alpha_0}) \xrightarrow{d} \Pi\left(-x(F_s^{\alpha_0} - F_b) \middle| T_{\mathcal{F}}(F_s^{\alpha_0})\right) \quad \text{in } \mathbb{H}. \tag{2.30}$$

Adding (2.29) and (2.30), we get

$$\frac{\sqrt{n}}{c_n}\gamma_n(\hat{F}_{s,n}^{\gamma_n} - \bar{F}_{s,n}^{\gamma_n}) \to -x(F_s^{\alpha_0} - F_b) - \Pi\left(-x(F_s^{\alpha_0} - F_b) \middle| T_{\mathcal{F}}(F_s^{\alpha_0})\right) \quad \text{in } \mathbb{H}.$$

By the continuous mapping theorem, we get $\sqrt{n}/c_n\gamma_n d(\hat{F}_{s,n}^{\gamma_n}, \bar{F}_{s,n}^{\gamma_n}) \xrightarrow{P} |x| \left\{\int V^2 dF\right\}^{1/2}$. Hence, by Lemma 12,

$$P(r_n(\hat{\alpha}_0^{c_n} - \alpha_0) \le x) \to \begin{cases} 1, & \text{if } x > 0, \\ 1, & \text{if } x \le 0 \text{ and } |x| \le \left\{\int V^2 dF\right\}^{-1/2}, \\ 0, & \text{otherwise.} \end{cases}$$

2.14.11 Proof of Theorem 5

Letting $c_n = H_n^{-1}(1-\beta)$, we have

$$\begin{aligned} P(\alpha_0 \ge \hat{\alpha}_L) &= P\left(\sqrt{n}\alpha_0\, d_n(\hat{F}_{s,n}^{\alpha_0}, \check{F}_{s,n}^{\alpha_0}) \le c_n\right) \\ &\ge P\left(\sqrt{n}\alpha_0\, d_n(\hat{F}_{s,n}^{\alpha_0}, F_s^{\alpha_0}) \le c_n\right) = H_n(c_n) = 1-\beta, \end{aligned}$$

where we have used the fact that $\alpha_0 d_n(\hat{F}_{s,n}^{\alpha_0}, F_s^{\alpha_0}) = d_n(\mathbb{F}_n, F)$. Note that, when $\alpha_0 = 0$, $F = F_b$, and using (2.8) we get

$$P(\alpha_0 \ge \hat{\alpha}_L) = P\left(\sqrt{n}\, d_n(\mathbb{F}_n, F_b) \le c_n\right) = P\left(\sqrt{n}\, d_n(\mathbb{F}_n, F) \le c_n\right) = 1-\beta.$$

2.14.12 Proof of Theorem 6

It is enough to show that $\sup_x |H_n(x) - G(x)| \to 0$, where G is the limiting distribution of the Cramér-von Mises statistic, a continuous distribution. As $\sup_x |G_n(x) - G(x)| \to 0$, it is enough to show that

$$\sqrt{n}d_n(\mathbb{F}_n, F) - \sqrt{n}d(\mathbb{F}_n, F) \xrightarrow{P} 0. \tag{2.31}$$

We now prove (2.31). Observe that

$$n(d_n^2 - d^2)(\mathbb{F}_n, F) = \sqrt{n}(\mathbb{P}_n - P)[\hat{g}_n] = \nu_n(\hat{g}_n),$$

where $\hat{g}_n = \sqrt{n}(\mathbb{F}_n - F)^2$, $\mathbb{P}_n$ denotes the empirical measure of the data, and $\nu_n := \sqrt{n}(\mathbb{P}_n - P)$ denotes the usual empirical process. We will show that $\nu_n(\hat{g}_n) \xrightarrow{P} 0$, which will prove (2.32).

For each positive integer n, we introduce the following class of functions

$$\mathcal{G}_c(n) = \left\{ \sqrt{n}(H - F)^2 : H \in \mathcal{F} \text{ and } \sup_{t \in \mathbb{R}} |H(t) - F(t)| < \frac{c}{\sqrt{n}} \right\}.$$

Let us also define

$$D_n := \sup_{t \in \mathbb{R}} \sqrt{n} |\mathbb{F}_n(t) - F(t)|.$$

From the definition of $\hat{g}_n$ and D_n^2, we have $\hat{g}_n(t) \leq \frac{1}{\sqrt{n}} D_n^2$, for all $t \in \mathbb{R}$. As $D_n = O_P(1)$, for any given $\epsilon > 0$, there exists $c > 0$ (depending on ϵ) such that

$$P(\hat{g}_n \notin \mathcal{G}_c(n)) = P(\sqrt{n} \sup_t |\hat{g}_n(t)| \geq c^2) = P(D_n^2 \geq c^2) \leq \epsilon, \tag{2.33}$$

for all sufficiently large n. Therefore, for any $\delta > 0$, using the same sequence of steps as in (2.23),

$$P(|\nu_n(\hat{g}_n)| > \delta) \quad \leq \quad J_{[\,]} \frac{E[G_c^2(n)]}{\delta} + P(\hat{g}_n \notin \mathcal{G}_c(n)),$$

where $G_c(n) := \frac{c^2}{\sqrt{n}}$ is an envelope for $\mathcal{G}_c(n)$ and $J_{[\,]}$ is a constant. Note that to derive the last inequality we have used the maximal inequality in Corollary (4.3) of Pollard (1989); the class $\mathcal{G}_c(n)$ is "manageable" in the sense of [Pollard, 1989] (as a consequence of equation (2.5) of [Van de Geer, 2000a]).

Therefore, for any given $\delta > 0$ and $\epsilon > 0$, for large enough n and $c > 0$ we can make both $J_{[\,]} c^4/(\delta n)$ and $P(\hat{g}_n \notin \mathcal{G}_c(n))$ less than ϵ, using (2.33) and (2.34), and thus, $P(|\nu_n(\hat{g}_n)| > \delta) \leq 2\epsilon$. The result now follows.

2.14.13 Proof of Theorem 7

The random variable U defined in the statement of the theorem can be explicitly expressed as

$$U := \left[\int \{\mathbb{G}_F - \Pi(\mathbb{G}_F | T_{\mathcal{F}}(F_s^{\alpha_0}))\}^2 \, dF \right]^{1/2},$$

where $\mathbb{G}_F$ is the F-Brownian bridge.

By the same line of arguments as in the proof of Lemma 12 (see Section 2.14.9.2), it can be easily seen that $\sqrt{n}\alpha_0\ d_n(\hat{F}_{s,n}^{\alpha_0}, \check{F}_{s,n}^{\alpha_0}) - \sqrt{n}\alpha_0\ d(\hat{F}_{s,n}^{\alpha_0}, \bar{F}_{s,n}^{\alpha_0}) \xrightarrow{P} 0$. Moreover, by Donsker's theorem,

$$\sqrt{n}\alpha_0(\hat{F}_{s,n}^{\alpha_0} - F_s^{\alpha_0}) \xrightarrow{d} \mathbb{G}_F.$$

By applying the functional delta method for the projection operator, in conjunction with the continuous mapping theorem to the previous display, we have

$$\sqrt{n}\alpha_0(\bar{F}_{s,n}^{\alpha_0} - F_s^{\alpha_0}) \xrightarrow{d} \Pi(\mathbb{G}_F|T_{\mathcal{F}}(F_s^{\alpha_0})) \quad \text{in} \quad \mathbb{H},$$

where Π, $T_{\mathcal{F}}(\cdot)$, and $F_s^{\alpha_0}$ are defined in (2.15), (2.16), and (2.18), respectively. Hence, by an application of the continuous mapping theorem, we have $\sqrt{n}\alpha_0 d(\hat{F}_{s,n}^{\alpha_0}, \bar{F}_{s,n}^{\alpha_0}) \xrightarrow{d} U$. The result now follows.

2.14.14 Proof of Lemma 9

Let $0 < \gamma_1 < \gamma_2 < 1$. Then,

$$\begin{aligned}
\gamma_2 d_n(\hat{F}_{s,n}^{\gamma_2}, \check{F}_{s,n}^{\gamma_2}) &\leq \gamma_2 d_n(\hat{F}_{s,n}^{\gamma_2}, (\gamma_1/\gamma_2)\check{F}_{s,n}^{\gamma_1} + (1-\gamma_1/\gamma_2)F_b) \\
&= d_n(\gamma_1\hat{F}_{s,n}^{\gamma_1} + (\gamma_2-\gamma_1)F_b, \gamma_1\check{F}_{s,n}^{\gamma_1} + (\gamma_2-\gamma_1)F_b) \\
&\leq \gamma_1 d_n(\hat{F}_{s,n}^{\gamma_1}, \check{F}_{s,n}^{\gamma_1}),
\end{aligned}$$

which shows that $\gamma d_n(\hat{F}_{s,n}^{\gamma}, \check{F}_{s,n}^{\gamma})$ is a non-increasing function. To show that $\gamma d_n(\hat{F}_{s,n}^{\gamma}, \check{F}_{s,n}^{\gamma})$ is convex, let $0 < \gamma_1 < \gamma_2 < 1$ and $\gamma_3 = \eta\gamma_1 + (1-\eta)\gamma_2$, for $0 \leq \eta \leq 1$. Then, by (2.13) we have the desired result.

2.14.15 Proof of Theorem 8

The constant c and the function Q defined in the statement of the theorem can be explicitly expressed as

$$c = d(Q, \Pi\left(Q|T_{\mathcal{F}}(F_s)\right)),$$

and

$$Q := (F_s - F_b)\left\{\alpha_0^2 \int V^2 dF\right\}^{-1/2},$$

where

$$r_n = \sqrt{n}/c_n, \quad V = (F_s - F_b) - \Pi(F_s - F_b | T_{\mathcal{F}}(F_s)),$$

and Π and $T_{\mathcal{F}}(\cdot)$ are defined in (2.15) and (2.16), respectively.

Recall the notation of Section 2.14.9. Note that from (2.2),

$$\hat{F}_{s,n}^{\check{\alpha}_n}(x) = \frac{\alpha_0}{\check{\alpha}_n} F_s(x) + \frac{\check{\alpha}_n - \alpha_0}{\check{\alpha}_n} F_b(x) + \frac{(\mathbb{F}_n - F)(x)}{\check{\alpha}_n},$$

for all $x \in \mathbb{R}$. Thus we can bound $\hat{F}_{s,n}^{\check{\alpha}_n}(x)$ as follows:

$$\frac{\alpha_0}{\check{\alpha}_n} F_s(x) - \frac{|\check{\alpha}_n - \alpha_0|}{\check{\alpha}_n} - \frac{D_n'}{\check{\alpha}_n} \leq \hat{F}_{s,n}^{\check{\alpha}_n}(x) \leq \frac{\alpha_0}{\check{\alpha}_n} F_s(x) + \frac{|\check{\alpha}_n - \alpha_0|}{\check{\alpha}_n} + \frac{D_n'}{\check{\alpha}_n},$$

where $D_n' = \sup_{x \in \mathbb{R}} |\mathbb{F}_n(x) - F(x)|$. As both the upper and lower bounds are monotone, we can always find a version of $\check{F}_{s,n}^{\check{\alpha}_n}$ such that

$$\frac{\alpha_0}{\check{\alpha}_n} F_s - \frac{|\check{\alpha}_n - \alpha_0|}{\check{\alpha}_n} - \frac{D_n'}{\check{\alpha}_n} \leq \check{F}_{s,n}^{\check{\alpha}_n} \leq \frac{\alpha_0}{\check{\alpha}_n} F_s + \frac{|\check{\alpha}_n - \alpha_0|}{\check{\alpha}_n} + \frac{D_n'}{\check{\alpha}_n}.$$

Therefore,

$$\begin{aligned} |\check{F}_{s,n}^{\check{\alpha}_n} - F_s| &\leq \frac{|\alpha_0 - \check{\alpha}_n|}{\check{\alpha}_n} F_s + \frac{|\check{\alpha}_n - \alpha_0|}{\check{\alpha}_n} + \frac{D_n'}{\check{\alpha}_n} \\ &\leq 2\frac{|\alpha_0 - \check{\alpha}_n|}{\check{\alpha}_n} + \frac{D_n'}{\check{\alpha}_n} \xrightarrow{P} 0, \end{aligned}$$

as $n \to \infty$, using the fact $\check{\alpha}_n \xrightarrow{P} \alpha_0 \in (0,1)$. Furthermore, if $q_n(\check{\alpha}_n - \alpha_0) = O_P(1)$, where $q_n/\sqrt{n} \to 0$, it is easy to see that $q_n|\check{F}_{s,n}^{\check{\alpha}_n} - F_s| = O_P(1)$, as $q_n D_n' = o_P(1)$. Note that

$$r_n \hat{\alpha}_0^{c_n} (\hat{F}_{s,n}^{\hat{\alpha}_0^{c_n}} - F_s) = r_n(\mathbb{F}_n - F) + r_n(\alpha_0 - \hat{\alpha}_0^{c_n}) \, (F_s - F_b)$$

Thus

$$\sup_{x \in \mathbb{R}} |r_n(\hat{F}_{s,n}^{\hat{\alpha}_0^{c_n}} - F_s)(x) - Q(x)| \xrightarrow{P} 0.$$

Hence by an application of functional delta method for the projection operator, in conjunction with the continuous mapping theorem, we have

$$r_n d(\check{F}_{s,n}^{\hat{\alpha}_0^{c_n}}, F_s) \xrightarrow{P} d(Q, \Pi(Q | T_{\mathcal{F}}(F_s))).$$

2.14.16 Proof of Theorem 9

Let $\epsilon_n := \sup_{x\in\mathbb{R}} |\check{F}^{\check{\alpha}_n}_{s,n}(x) - F_s(x)|$. Then the function $F_s + \epsilon_n$ is concave on $[0,\infty)$ and majorises $\check{F}^{\check{\alpha}_n}_{s,n}$. Hence, for all $x \in [0,\infty)$, $\check{F}^{\check{\alpha}_n}_{s,n}(x) \leq F^{\dagger}_{s,n}(x) \leq F_s(x) + \epsilon_n$, as $F^{\dagger}_{s,n}$ is the LCM of $\check{F}^{\check{\alpha}_n}_{s,n}$. Thus,

$$-\epsilon_n \leq \check{F}^{\check{\alpha}_n}_{s,n}(x) - F_s(x) \leq F^{\dagger}_{s,n}(x) - F_s(x) \leq \epsilon_n,$$

and therefore,

$$\sup_{x\in\mathbb{R}} |F^{\dagger}_{s,n}(x) - F_s(x)| \leq \epsilon_n.$$

By Theorem 8, as $\epsilon_n \xrightarrow{P} 0$, we must also have (9).

The second part of the result follows immediately from the lemma is page 330 of [Robertson *et al.*, 1988], and is similar to the result in Theorem 7.2.2 of that book.

Part II

Single Index Model

Chapter 3

Efficient Estimation in Single Index Models using Smoothing splines

We consider estimation and inference in a single index regression model with an unknown but smooth link function. In contrast to standard kernel based 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 i.i.d. data. We prove the consistency and find the rates of convergence of the proposed estimators. We establish $n^{-1/2}$-rate of convergence and the asymptotic efficiency of the parametric component under mild assumptions.

Keywords: interpolation inequality, least favorable submodel, penalized least squares.

3.1 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"; see [Stone, 1980]) 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 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

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

where $X^\top\theta_0$ 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 models; 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].

Formally, in this paper, we consider the model

$$Y = m_0(\theta_0^\top X) + \epsilon, \quad \mathbb{E}(\epsilon|X) = 0, \quad \text{a.e. } X, \tag{3.1}$$

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 the model we assume that the first coordinate of θ_0 is non-zero and

$$\theta_0 \in \Theta := \{\eta_0 \in \mathbb{R}^d : |\eta_0| = 1 \text{ and } \eta_{0,1} \geq 0\} \subset S^{d-1}, \tag{3.2}$$

where $\eta_{0,1}$ is the first coordinate of η_0, $|\cdot|$ 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]), 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 fixed, (3.1) 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) smoothness 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 $\{(y_i, x_i)\}_{1\le i\le n}$ from model (3.1), we propose minimizing the following penalized loss:

$$\frac{1}{n}\sum_{i=1}^{n}\left(y_i - m(\theta^\top x_i)\right)^2 + \lambda^2 \int |m''(t)|^2 dt \quad (\lambda \neq 0) \tag{3.3}$$

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 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]. [Antoniadis *et al.*, 2004] proposed a Bayesian approach for estimation in the single index models using B-splines. However, as in the estimator proposed in [Yu and Ruppert, 2002], the estimator in [Antoniadis *et al.*, 2004] requires a choice of knots for every θ.

All this motivates the use of smoothing splines for estimation in the smooth single index model. Smoothing splines avoid the choice of number of knots and their placement — the number of knots increase to infinity with sample size. Further, smoothness 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 local paths on the sphere to study the semiparametric efficiency of the finite dimensional parameter.

The theory developed in this paper allows for the tuning parameter λ in (3.3) to be data dependent. Thus data-driven procedures such as cross-validation can be used to choose an optimal λ. As opposed to average derivative methods discussed eariler (see [Powell *et al.*, 1989] and [Hristache *et al.*, 2001]), the optimization problem in (3.3) involves only 1-dimensional nonparametric function estimation.

Our exposition is organized as follows. In Section 3.2 we introduce some notation, formally define our estimators and study its existence. We state and discuss our assumptions in Section 3.3 and prove consistency (see Theorem 13) and provide the rates of convergence (see Theorems 12 and 14) for our estimators. We show that the estimator for θ_0 is asymptotically normal (properly scaled) and is semiparametrically efficient; see Theorem 15 in Section 3.4. The Sections 3.6–3.8 contain proofs of the results.

3.2 Preliminaries

Suppose that $\{(y_i, x_i)\}_{1 \le i \le n}$ is an i.i.d. sample from model (3.1). We start with some notation. Let $\mathcal{X} \subset \mathbb{R}^d$ denote the support of X and D be the set of possible index values, i.e.,

$$D := \{\theta^\top x : x \in \mathcal{X}, \theta \in \Theta\}.$$

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 denotes the marginal distribution of X. For $g : \mathcal{X} \to \mathbb{R}$, define

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

Let $P_{\theta,m}$ denote the joint distribution of (Y, X) when $X \sim P_X$ and $Y := m(\theta^\top X) + \epsilon$. In particular, P_{θ_0,m_0} denotes the joint distribution of (Y, X) when $X \sim P_X$ and (Y, X) satisfy (3.1). For any function $m : I \subset \mathbb{R}^p \to \mathbb{R}$, let

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

Moreover, for $I_1 \subset I$, we define $\|m\|_{I_1} := \sup_{u \in I_1} |m(u)|$. 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 $f : \chi \to \mathbb{R}^r, r \geq 1$, let $\{f_i\}_{1 \leq i \leq r}$ denote the each of the components, i.e., $f(x) = (f_1(x), \ldots, f_r(x)), r \geq 1$ 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 class $\mathcal{G}$ of real-valued functions, let $\mathcal{G} \circ \Theta$ denote the set of linear index functions

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

where $(g, \theta) \in \mathcal{G} \times \Theta$. For any function $f : D \subset \mathbb{R} \to \mathbb{R}$ with absolutely continuous first derivative, we define the smoothness 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 3.3). We now state two simple results for functions in the class $\mathcal{S}$. The following two lemmas, proved in Section 3.6, will be useful in the remainder of the paper.

Lemma 14. *Let $m \in \{g \in \mathcal{S} : J(g) < \infty\}$. Then $|m'(s) - m'(s_0)| \leq J(m)|s - s_0|^{1/2}$ for every $s, s_0 \in D$.*

Lemma 15. *For any set $A \in \mathbb{R}^p$ $p \geq 1$, let $\varnothing(A)$ denote the diameter of the set A. Let $m \in \{g \in \mathcal{S} : J(g) < \infty$ and $\|m\|_\infty \leq M\}$, where M is a finite constant. Then*

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

Moreover if $\varnothing(D) < \infty$, then

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

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

The penalized loss for $(m, \theta) \in \mathcal{S} \times \Theta$ (and $\lambda \neq 0$) is defined as

$$\mathcal{L}_n(m, \theta; \lambda) := \frac{1}{n} \sum_{i=1}^{n} \big(y_i - m(\theta^\top x_i)\big)^2 + \lambda^2 J^2(m). \tag{3.4}$$

For simplicity of notation, we define

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

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

$$(\hat{m}, \hat{\theta}) := \underset{(m,\theta) \in \mathcal{S} \times \Theta}{\arg\min}\ \mathcal{L}_n(m, \theta; \lambda). \tag{3.5}$$

Here we suppress the dependence of $(\hat{m}, \hat{\theta})$ on λ, for notational convenience. The following theorem (proved in Section 3.6.3) proves the existence of $(\hat{m}, \hat{\theta})$ for every $\lambda \neq 0$.

Theorem 11. *$\hat{\theta} \in \Theta$ and $\hat{m} \in \mathcal{S}$, where $\hat{\theta}$ and $\hat{m}$ are defined in (3.5). Moreover, $\hat{m}$ is a natural cubic spline with knots at $\{\hat{\theta}^\top x_i\}_{1 \leq i \leq n}$.*

We now outline the identification of the composite population parameter $m_0 \circ \theta_0$. Define $Q(m, \theta) := \mathbb{E}[Y - m(\theta^\top X)]^2$. The following argument shows that (m_0, θ_0) is the minimizer of Q and is well-separated (with respect to the $L_2(P_X)$-norm) from other elements in $\mathcal{S} \times \Theta$. Choose arbitrarily small $\delta > 0$, and pick any $(m, \theta) \in \mathcal{S} \times \Theta$ such that $\|m \circ \theta - m_0 \circ \theta_0\|^2 > \delta^2$. Then

$$Q(m, \theta) = \mathbb{E}[Y - m_0(\theta_0^\top X)]^2 + \mathbb{E}[m_0(\theta_0^\top X) - m(\theta^\top X)]^2,$$

since $\mathbb{E}(\epsilon | X) = 0$. Thus, we have

$$\inf_{\|m \circ \theta - m_0 \circ \theta_0\|^2 > \delta^2} Q(m, \theta) - Q(m_0, \theta_0) > \delta^2.$$

Note that identification of $m_0 \circ \theta_0$ 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) For some integer $d_1 \in (0, d]$, let $(X_1, \ldots, X_{d_1})$ have continuous marginal 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 constants $c_0, c_1, \ldots, c_{d-d_1} \in \mathbb{R}^{d-d_1}$ such that

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

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

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

(A3) The error ϵ in model (3.1) is assumed to be uniformly sub-Gaussian, i.e., there exists $K_1 > 0$ and $K_2 \in \mathbb{R}$ such that

$$K_1^2 \mathbb{E}\big(\exp(\epsilon^2/K_1^2) - 1 | X\big) \leq K_2^2 \text{ a.e. } \mathbb{P}.$$

As stated in (3.1), we also assume that $\mathbb{E}(\epsilon|X) = 0$ a.e. $\mathbb{P}$.

(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:

$$\hat{\lambda}_n^{-1} = O_p(n^{2/5}) \qquad \text{and} \qquad \hat{\lambda}_n = o_p(n^{-1/4}). \tag{3.6}$$

(A5) Define

$$D_0 := \{x^\top \theta_0 : x \in \mathcal{X}\}.$$

We assume that D_0 is the closure of its interior.

(A6) $\text{Var}(X)$ is a positive definite matrix.

(A7) $\mathbb{E}\big[XX^\top |m_0'(\theta_0^\top X)|^2\big]$ 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 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, 2000b] 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 (3.6), $\hat{\theta}$ will be an asymptotically "efficient" estimator of θ_0. We use empirical process methods (e.g., see [van der Vaart, 1998b]) to prove the consistency and to find the rates of convergence of $\hat{m}$ and $\hat{\theta}$. A sufficient condition for **(A5)** is that at least one of the coordinates of X is a continuous random variable and the corresponding index coefficient is non-zero. Assumptions **(A6)** and **(A7)** are mild distributional assumptions on the design. Assumption **(A6)** guarantees that the predictors are not supported on a lower dimensional affine space. Note that **(A7)** 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 12 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 12. *Under assumptions **(A0)–(A6)**, 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 13. *Under assumptions **(A0)–(A5)**,* $\hat{\theta} \xrightarrow{P} \theta_0$, $\|\hat{m} - m_0\|_{D_0}^S \xrightarrow{P} 0$, *and* $\|\hat{m}'\|_\infty = O_p(1)$.

The above result shows that not only is $\hat{m}$ consistent but it's 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 3.4.

Theorem 14. *Under **(A0)–(A7)** 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 12, 13, and 14 are given in Sections 3.7.1, 3.7.3, and 3.7.4, respectively.

3.4 Semiparametric inference

In this section we show that $\hat{\theta}$ is asymptotically normal and 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 exist $r > 0$ such that for all $\theta \in S^{d-1} \cap B_{\theta_0}(r)$ we have

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

For every $\theta \in \Theta$, define $h_\theta : D \to \mathbb{R}^d$,

$$h_\theta(u) := \mathbb{E}[X | \theta^\top X = u]. \tag{3.7}$$

(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,

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

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

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

(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 $(\{\theta_0^\top x : x \in \mathcal{X}\})$ 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 3.4.2.2). For the functions in the path to be in $\mathcal{S}$, we need the smoothness assumptions on h_θ. **(B3)** gives lower and upper bound 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 $f : \mathbb{R} \times \mathcal{X} \to \mathbb{R}$, θ, and m, 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 $f : \mathbb{R} \times \mathcal{X} \to \mathbb{R}$, we define

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

3.4.1 Efficient score

As a first step in showing that $\hat{\theta}$ is an efficient estimator, in the following we find the efficiency bound for the model (3.1). We represent the space of the finite dimensional parameter by Θ. 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 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., [Li and Patilea, 2015]. [Cui *et al.*, 2011] considers the reparameterization $\theta = (1 - |\theta_{-1}|, \theta_{-1})$, where $|\theta_{-1}| < 1$. Under this parameterization the parameter space is bounded; however, calculations of the parametric score becomes unnecessarily tedious. 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 define 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, 2001]. The following lemma proved[1] in Section 3.8.1 shows that the "local parameterization matrix" as function of θ is Lipschitz with respect to the operator norm.

Lemma 16. *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* $|\beta - \theta_0| \le 1/2$

$$\|H_\eta^\top - H_\beta^\top\|_2 \le 8(1 + 8/\sqrt{15})\frac{|\eta - \beta|}{|\eta - \theta_0| + |\beta - \theta_0|}. \tag{3.8}$$

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.

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

$$\zeta_s(\theta, \eta) := \sqrt{1 - s^2|\eta|^2}\,\theta + sH_\theta\eta. \tag{3.9}$$

[1] Our proof is constructive.

Note that $\theta^\top H_\theta = 0$ and $|H_\theta \eta| = |\eta|$ for all $\eta \in \mathbb{R}^{d-1}$. When $|s| \leq 1/|\eta|$ we have $\zeta_s(\theta, \eta) \in S^{d-1}$. For every fixed $s \in \mathbb{R}$, 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

$$Y = m(\theta^\top X) + \epsilon \tag{3.10}$$

for some $(\theta, m) \in (\Theta, \mathcal{S})$ under assumption **(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 6. *Note that under (3.10), 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 $\eta \in S^{d-2}$, consider the path defined in $s \mapsto \zeta_s(\theta, \eta)$. 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

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

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 $s \in \mathbb{R}$, $b : \mathbb{R} \times \mathcal{X} \to \mathbb{R}$ 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 $q : \mathcal{X} \to \mathbb{R}$ is a bounded function such that $\mathbb{E}(q(X)) = 0$. Consider the following parametric submodel of (3.1),

$$s \mapsto (\zeta_s(\theta, \eta),\ m_{s,a},\ p_{\epsilon|X;s,b},\ p_{X;s,q}(x)) \tag{3.11}$$

where $\eta \in S^{d-2}$. Differentiating the log-likelihood of the submodel in (3.11) with respect to s, we get that the score along the submodel in (3.11) 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

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

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 Theorem 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

$$\overline{\text{lin}}\{a \in \mathcal{S} : J(a) < \infty\} = \{a : D \to \mathbb{R} |\, a \in L_2(\mathbf{m})\},$$

$$\overline{\text{lin}}\{q : \mathcal{X} \to \mathbb{R} |\, q \text{ is a bounded function and } \mathbb{E}(q(X) = 0\} = \{q \in L_2(P_X) |\, \mathbb{E}(q(X)) = 0\},$$

and

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

Thus, it is easy to see that under assumptions **(A0)**–**(A7)** and **(B1)**–**(B3)**, the nuisance tangent space of (3.1) is

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

see Theorem 4.1 in [Newey and Stoker, 1993] and Proposition 1 of [Ma and Zhu, 2013a] 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, 2013a] 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, 2013a], it can be shown that

$$\Pi(S_{\theta,m}|\Lambda^{\perp})(y,x) = \frac{(y - m(\theta^{\top}x))}{\sigma^2(x)} m'(\theta^{\top}x) H_{\theta}^{\top} \left\{ x - \frac{\mathbb{E}(\sigma^{-2}(X)X|\theta^{\top}X = \theta^{\top}x)}{\mathbb{E}(\sigma^{-2}(X)|\theta^{\top}X = \theta^{\top}x)} \right\}, \tag{3.12}$$

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 $S^{eff}_{\theta,m}$. 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, we will assume that $\sigma^2(\cdot) \equiv \sigma^2$. Under this assumption the efficient score 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 (3.7). 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.,

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

and class of functions formed by the efficient score indexed by (θ, m) in a "neighborhood" of (θ_0, m_0) satisfies some technical conditions. We formalize this in Theorem 15 below.

3.4.2 Efficiency of $\hat{\theta}$

Theorem 15. *Assume that (Y, X) satisfies (3.1) and assumptions **(A0)–(A7)** and **(B1)–(B3)** hold. Define*

$$\tilde{\ell}_{\theta,m}(y,x) := \big(y - m(\theta^\top x)\big)m'(\theta^\top x)H_\theta^\top \Big\{x - h_\theta(\theta^\top x)\Big\}. \tag{3.13}$$

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

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0, H_{\theta_0}V_{\theta_0,m_0}^{-1}\tilde{I}_{\theta_0,m_0}(H_{\theta_0}V_{\theta_0,m_0}^{-1})^\top), \tag{3.14}$$

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.,

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0, \sigma^4 H_{\theta_0}\tilde{I}_{\theta_0,m_0}^{-1}H_{\theta_0}^\top). \tag{3.15}$$

Remark 7. *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 estimator as*

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

where $\hat{w}(x)$ is a consistent estimator of $V^{-2}(\theta_0^\top x)$. Theorem 15 can be generalized to show that $\tilde{\theta}$ is an efficient estimator of θ_0.

Remark 8. *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 the link function.*

3.4.2.1 Proof of Theorem 15

Now we give a sketch of the proof of (3.14). Some of the steps are proved in the following sections.

Step 1 In Theorem 16 we will show that $(\hat{m}, \hat{\theta})$ satisfy the efficient score equation *approximately*, i.e.,

$$\sqrt{n}\mathbb{P}_n\tilde{\ell}_{\hat{\theta},\hat{m}} = o_p(1). \tag{3.16}$$

Step 2 In Section 3.8.4 we prove that $\tilde{\ell}_{\hat{\theta},\hat{m}}$ is unbiased in the sense of [van der Vaart, 2002], i.e.,

$$P_{\hat{\theta},m_0}\tilde{\ell}_{\hat{\theta},\hat{m}} = 0. \tag{3.17}$$

Step 3 We prove

$$\mathbb{G}_n(\tilde{\ell}_{\hat{\theta},\hat{m}} - \tilde{\ell}_{\theta_0,m_0}) = o_p(1) \tag{3.18}$$

in Theorem 17. In view of (3.16) and (3.17) an equivalent formulation of (3.18) is

$$\sqrt{n}(P_{\hat{\theta},m_0} - P_{\theta_0,m_0})\tilde{\ell}_{\hat{\theta},\hat{m}} = \mathbb{G}_n\tilde{\ell}_{\theta_0,m_0} + o_p(1). \tag{3.19}$$

Step 4 To complete the proof of (3.14), it is enough to show that

$$\sqrt{n}(P_{\hat{\theta},m_0} - P_{\theta_0,m_0})\tilde{\ell}_{\hat{\theta},\hat{m}} = \sqrt{n}V_{\theta_0,m_0}H_{\theta_0}^\top(\hat{\theta} - \theta_0) + o_p(\sqrt{n}|\hat{\theta} - \theta_0|). \tag{3.20}$$

Observe that (3.19) and (3.20) imply

$$\begin{aligned} &\sqrt{n}V_{\theta_0,m_0}H_{\theta_0}^\top(\hat{\theta} - \theta_0) = \mathbb{G}_n\tilde{\ell}_{\theta_0,m_0} + o_p(1 + \sqrt{n}|\hat{\theta} - \theta_0|), \\ &\Rightarrow \sqrt{n}H_{\theta_0}^\top(\hat{\theta} - \theta_0) = V_{\theta_0,m_0}^{-1}\mathbb{G}_n\tilde{\ell}_{\theta_0,m_0} + o_p(1) \xrightarrow{d} V_{\theta_0,m_0}^{-1}N(0, \tilde{I}_{\theta_0,m_0}). \end{aligned} \tag{3.21}$$

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:

$$\hat{\theta} = \sqrt{1 - |\hat{\eta}|^2}\,\theta_0 + H_{\theta_0}\hat{\eta}. \tag{3.22}$$

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

$$\hat{\eta} = H_{\theta_0}^\top(\hat{\theta} - \theta_0). \tag{3.23}$$

Substituting the above expression of $\hat{\eta}$ in (3.22) and subtracting θ_0 from both sides of (3.22) 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 (3.21) 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 (3.20) 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 (3.20) in Section 3.8.3.

Now we prove (3.15). Assume that $\sigma^2(\cdot) \equiv \sigma^2$. Observe that, by (3.12) and (3.13), we have

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

Thus (3.15) follows from (3.14) by observing that

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

3.4.2.2 "Least favorable" path for m

We will now show that **Step 1** holds, i.e., $(\hat{m}, \hat{\theta})$ satisfies (3.16). Recall the definition (3.9). For any $(\theta, m) \in \Theta \times \{m \in \mathcal{S} | J(m) < \infty\}$, 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., $(\zeta_0(\theta, \eta), \xi_0(\cdot; \theta, \eta, m)) = (\theta, m)$. 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

$$\frac{\partial}{\partial t} \mathcal{L}_n(\xi_t(\cdot; \hat{\theta}, \eta, \hat{m}), \zeta_t(\hat{\theta}, \eta), \hat{\lambda}_n)\Big|_{t=0} = 0. \tag{3.24}$$

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

$$\begin{aligned} \frac{\partial}{\partial t}\big(y - \xi_t(\zeta_t(\hat{\theta},\eta)^\top x;\hat{\theta},\eta,\hat{m})\big)^2\Big|_{t=0} &= \eta^\top \tilde{\ell}_{\hat{\theta},\hat{m}}(y,x), \\ \frac{\partial}{\partial t} J^2(\xi_t(\cdot;\hat{\theta},\eta,\hat{m}))\Big|_{t=0} &= O_p(1). \end{aligned} \tag{3.25}$$

for all $\eta \in S^{d-2}$, then we get (3.16) as $\hat{\lambda}_n^2 = o_p(n^{-1/2})$ (see assumption **(A4)**).

Observe that $\hat{\theta}$ is a consistent estimator of θ. 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 $A^\nu := \cup_{a\in A} B_a(\nu)$. 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

$$\begin{aligned} \phi_{\theta,\eta,t}(u) &= u, \quad u \in D_\theta \\ \phi_{\theta,\eta,t}(u + (\theta - \zeta_t(\theta,\eta))^\top h_\theta(u)) &= u, \quad u \in \partial D, \end{aligned} \tag{3.26}$$

where $h_\theta(u)$ and $\zeta_t(\theta,\eta)$ are defined in (3.7) and (3.9), 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:

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

The function $\phi_{\theta,\eta,t}$ helps us control the partial derivative in the second equation of (3.25). In the following theorem (proved in Section 3.8.2) 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 (3.24) and (3.25). Here η is the "direction" for $\zeta_t(\theta,\eta)$ and $(\eta, h_\theta(u))$ defines the "direction" for the path $\xi_t(\cdot;\theta,\eta,m)$.

Theorem 16. *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 (3.25) 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 $\sqrt{n}\mathbb{P}_n\tilde{\ell}_{\hat{\theta},\hat{m}} = o_p(1)$.*

3.4.2.3 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 (3.13) we have

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

Theorem 17. *Under assumptions **(A0)–(A7)** and **(B1)–(B3)**, we have* $\mathbb{G}_n(\tilde{\ell}_{\hat{\theta},\hat{m}} - \tilde{\ell}_{\theta_0,m_0}) = o_p(1)$.

We divide the proof of this theorem into two lemmas. First observe that

$$\begin{aligned}
&\mathbb{G}_n(\tilde{\ell}_{\hat{\theta},\hat{m}} - \tilde{\ell}_{\theta_0,m_0}) \\
&= \mathbb{G}_n\big[\big(Y - \hat{m}(\hat{\theta}^\top X)\big)\hat{m}'(\hat{\theta}^\top X)K_1(X;\hat{\theta}) - \big(Y - m_0(\theta_0^\top X)\big)m_0'(\theta_0^\top X)K_1(X;\theta_0)\big] \\
&= \mathbb{G}_n\big[\big(\epsilon + m_0(\theta_0^\top X) - \hat{m}(\hat{\theta}^\top X)\big)\hat{m}'(\hat{\theta}^\top X)K_1(X;\hat{\theta}) - \epsilon m_0'(\theta_0^\top X)K_1(X;\theta_0)\big] \\
&= \mathbb{G}_n\big[\big(m_0(\theta_0^\top X) - \hat{m}(\hat{\theta}^\top X)\big)\hat{m}'(\hat{\theta}^\top X)K_1(X;\hat{\theta})\big] \\
&\qquad + \mathbb{G}_n\big[\epsilon\big(\hat{m}'(\hat{\theta}^\top X)K_1(X;\hat{\theta}) - m_0'(\theta_0^\top X)K_1(X;\theta_0)\big)\big].
\end{aligned} \tag{3.27}$$

The proof of Theorem 17 will be complete if we can show that both the terms in (3.27) 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 13). For all $n \in \mathbb{N}$, define

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

Let us consider the first term of (3.27). Fix $\delta > 0$. For every fixed M_1, M_2, and M_3,

$$\begin{aligned}
&\mathbb{P}\Big(\big|\mathbb{G}_n\big[(m_0\circ\theta_0 - \hat{m}\circ\hat{\theta})\hat{m}'\circ\hat{\theta}K_1(\cdot;\hat{\theta})\big]\big| > \delta\Big)\\
&\le \mathbb{P}\Big(\big|\mathbb{G}_n\big[(m_0\circ\theta_0 - \hat{m}circ\hat{\theta})\hat{m}'\circ\hat{\theta}K_1(\cdot;\hat{\theta})\big]\big| > \delta, (\hat{m},\hat{\theta})\in\mathcal{C}_{M_1,M_2,M_3}(n)\Big)\\
&\quad + \mathbb{P}\big((\hat{m},\hat{\theta})\notin\mathcal{C}_{M_1,M_2,M_3}(n)\big) \qquad (3.28)\\
&\le \mathbb{P}\Big(\sup_{(m,\theta)\in\mathcal{C}_{M_1,M_2,M_3}(n)}\big|\mathbb{G}_n\big[(m_0\circ\theta_0 - m\circ\theta)m'\circ\theta K_1(\cdot;\theta)\big]\big| > \delta\Big)\\
&\quad + \mathbb{P}\big((\hat{m},\hat{\theta})\notin\mathcal{C}_{M_1,M_2,M_3}(n)\big).
\end{aligned}$$

Recall that $\hat{\theta}$ and $\hat{m}$ are consistent estimators of θ_0 and m_0 with respect to the Euclidean and Sobolev norms, respectively, and $\|\hat{m}'\|_\infty$ is $O_p(1)$ (see Theorem 13). Furthermore, we have that both $\|\hat{m}\|_\infty$ and $J(\hat{m})$ are $O_p(1)$ (see Theorem 12) and $\hat{\lambda}_n^{-1/2}|\hat{\theta}-\theta_0| = o_p(1)$ (see Theorem 14). 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

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

for sufficiently large n. The following lemma (proved in Section 3.8.5) shows this.

Lemma 17. *Fix M_1, M_2, M_3, and $\delta > 0$. For $n\in\mathbb{N}$, let us define*

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

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

$$\sup_{f\in\mathcal{D}_{M_1,M_2,M_3}(n)}\|f\|_{2,\infty}\le 2TM_2(a_n^{-1} + TM_2\hat{\lambda}_n^{1/2}) =: D_{M_1,M_2,M_3}(n). \qquad (3.29)$$

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, 1998b]) 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

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

The following lemma (proved in Section 3.8.6) shows that the second term on the right hand side of (3.27) converges to zero in probability.

Lemma 18. *Let us define* $U_{\theta,m} : \mathcal{X} \to \mathbb{R}^{d-1}$, $U_{\theta,m}(x) := m'(\theta^\top x) K_1(x;\theta)$. *Fix* M_1, M_2, M_3, *and* $\delta > 0$. *For* $n \in \mathbb{N}$, *let us define*

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

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

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

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*

$$\mathbb{P}\Big(\Big| \mathbb{G}_n \big[\epsilon \big(U_{\hat{\theta},\hat{m}} - U_{\theta_0,m_0} \big) \big] \Big| > \delta \Big) \to 0. \tag{3.30}$$

3.5 Simulation Study

To investigate the finite sample performance of the estimator developed in this chapter, we carry out several simulation experiments. We also compare the finite sample performance of the proposed estimator with the estimators proposed in [Cui *et al.*, 2011] and [Hristache *et al.*, 2001]. The code to evaluate the estimates proposed in [Hristache *et al.*, 2001] can be found in the R package `EDR`. Moreover, [Cui *et al.*, 2011] kindly provided us with the R codes to implement their procedure. The codes used to implement our procedure are available in the `simest` package in R; see [Kuchibhotla and Patra, 2016]. In the following, we consider three different data generating mechanisms. It is easy to see that the estimator proposed in this chapter has the best overall performance.

3.5.1 A simple model

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

$$Y = (X^\top \theta_0)^2 + \epsilon, \text{ where } \theta_0 = [1, -1]/\sqrt{2}. \tag{3.31}$$

Observe that for this example, $H_{\theta_0}^\top = [1, 1]/\sqrt{2}$ (see Section 3.8.1) and the analytic expression of the efficient information is

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

Using the above expression, we calculated the asymptotic variance of $\sqrt{n}(\hat{\theta}_1 - \theta_{0,1})$ to be 0.3277. Figure 3.1 shows the box plots of estimator proposed in this paper and compares its performance with the estimators proposed in [Cui *et al.*, 2011] and [Hristache *et al.*, 2001]. We also include the box plot of a sample from the true asymptotic distribution of $\hat{\theta}$ for comparison.

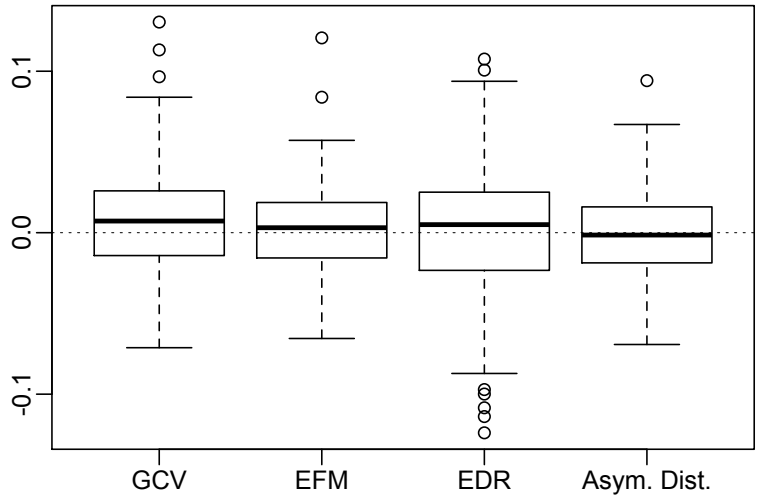

Figure 3.1: Box plots of bias estimates of $\theta_{0,1}$ (from 500 replications) along with the true asymptotic distribution of the $\sqrt{n}(\hat{\theta}_1 - \theta_{0,1})$ (properly scaled) when we have 500 i.i.d. samples from (3.31).

3.5.2 Dependent Covariates

We now consider a simulation scenario, where covariates are not independent and the predictor variable $X \in \mathbb{R}^6$ 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 assume that

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

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 the error distribution:

(2.1) $\epsilon \sim N(0, 1)$ (Homoscedastic, Gaussian Error)

(2.2) $\epsilon \sim N\left(0, \log(2 + (X^\top\theta_0)^2)\right)$ (Heteroscedastic, Gaussian Error)

(2.3) $\epsilon|\xi \sim (-1)^\xi \text{Beta}(2, 3)$, where $\xi \sim \text{Ber}(.5)$ (Homoscedastic, Non-Gaussian Error)

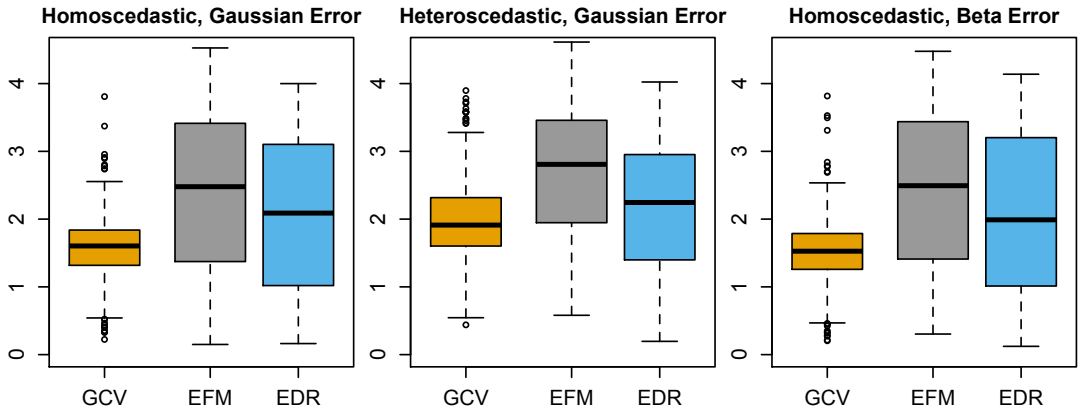

Figure 3.2: Box plots (over 500 replications) of L_1 error of estimates of θ_0 ($\sum_{i=1}^6 |\hat{\theta}_i - \theta_{0,i}|$) based on 200 observations from models **(2.1)**, **(2.2)**, and **(2.3)** in the left, the middle, and the right panel, respectively.

Observe that in all the three scenarios the proposed estimator has improved performance compared to the competitors; see Figure 3.2. Scenarios **(2.1)** and **(2.2)** are similar to simulation scenarios in [Ma and Zhu, 2013b] and [Li and Patilea, 2015].

3.5.3 High Dimensional Covariates

This setup in this simulation setting is similar to one considered in Example 4 of [Cui *et al.*, 2011]; see Section 3.2 of [Cui *et al.*, 2011]. We consider d-variate covariates for $d = 10, 50, 100$. For each d, we assume that $X \sim \text{Uniform}[0,2]^d$, $\epsilon \sim N(0, 0.2^2)$, $\theta_0 = (2, 1, \mathbf{0}_{d-2})^\top/\sqrt{5}$, and have 500 observations from the following model:

$$Y = \sin(aX^\top\theta_0) + \epsilon, \text{ where } a = \pi/2, 3\pi/4, \text{ and } 3\pi/2. \tag{3.32}$$

In Table 3.1, we show the finite sample performance of the estimators developed in this chapter. Here a higher value of a represents a more oscillating function. Observe that proposed estimator performs relatively well, whereas both EFM and EDR perform poorly in certain scenarios, e.g., EFM when $a = 3\pi/2$ and $d = 10$ and EDR when $a = \pi/2$ and $d = 100$.

Table 3.1: Median error (and interquartile range) $\sum_{i=1}^{d} |\hat{\theta}_i - \theta_{0,i}|$ (500 replications) for $n = 400$ from (3.32).

d	$a = \pi/2$			$a = 3\pi/4$			$a = 3\pi/2$		
	GCV	EFM	EDR	GCV	EFM	EDR	GCV	EFM	EDR
10	0.127	0.121	0.135	0.085	0.081	0.092	0.047	2.251	0.069
	(0.034)	(0.030)	(0.033)	(0.025)	(0.020)	(0.023)	(0.019)	(1.734)	(0.018)
50	0.735	0.700	0.896	0.501	0.477	0.613	6.510	6.672	6.562
	(0.109)	(0.092)	(0.113)	(0.082)	(0.061)	(0.074)	(0.368)	(0.372)	(0.284)
100	1.829	1.630	3.163	1.307	1.122	1.812	9.020	9.218	8.994
	(0.327)	(0.182)	(1.010)	(0.272)	(0.112)	(0.299)	(0.313)	(0.280)	(0.287)

3.6 Proof of results in Section 3.2

3.6.1 Proof of Lemma 14

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

$$|m'(s) - m'(s_0)| = \left|\int_{s_0}^{s} m''(t)dt\right| \leq \left|\int_{s_0}^{s} |m''(t)|^2 dt\right|^{1/2} |s - s_0|^{1/2} \leq J(m)|s - s_0|^{1/2},$$

for every $s, s_0 \in D$.

3.6.2 Proof of Lemma 15

Fix $s_0 \in D$. Integrating the inequality

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

with respect to t, we get

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

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

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

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

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

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

3.6.3 Proof of Theorem 11

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 (3.4). For any fixed vector $\theta \in \Theta$, define $t_i^\theta := \theta^\top x_i$, for $i = 1, \ldots, n$. Then we have

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

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

$$T(\theta) := \inf_{m\in\mathcal{S}}\mathcal{L}_n(m,\theta,\lambda) \quad \text{and} \quad m_\theta := \arg\min_{m\in\mathcal{S}}\mathcal{L}_n(m,\theta,\lambda). \tag{3.33}$$

Theorem 2.4 of [Green and Silverman, 1994] proves that the infimum in (3.33) 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 $\sup_{\theta\in\Theta}\max_{i\le n}|\theta^\top x_i| < t_{\max}$. 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 $J_R(m) := \int_{D_R}|m''(t)|^2dt$. 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 $\theta\mapsto T_R(\theta)$ 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 $\theta \in \Theta$,

$$Q_n(m_\theta^R, \theta) + \lambda^2 J_R^2(m_\theta^R) \le K. \tag{3.34}$$

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 (3.34), we have that

$$J_R^2(m_\theta^R) \le K/\lambda^2 \quad \text{and} \quad |m_\theta^R(t_{(i)}^\theta)| \le \sqrt{nK} + \max_{i\le n} |y_i|, \tag{3.35}$$

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 15, 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

$$\left|m_\theta^R(s) - m_\theta^R(t_{(1)}^\theta) - (m_\theta^R)'(t_{(1)}^\theta)(s - t_{(1)}^\theta)\right| \le J_R(m_\theta^R)(t_{\max})^{3/2}. \tag{3.36}$$

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 (3.35)). In view of the bound on $|(m_\theta^R)'(t_{(1)}^\theta)|$, (3.36) 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

$$\begin{aligned}
&|Q_n(m,\theta) - Q_n(m,\eta)| \\
&= \frac{1}{n}\left|\sum_{i=1}^{n}\left[(y_i - m(\theta^\top x_i))^2 - (y_i - m(\eta^\top x_i))^2\right]\right| \\
&= \frac{1}{n}\left|\sum_{i=1}^{n}\left[(m(\eta^\top x_i) - m(\theta^\top x_i))^2 + 2(y_i - m(\eta^\top x_i))(m(\eta^\top x_i) - m(\theta^\top x_i))\right]\right| \\
&\le \max_{1\le i\le n}|m(\eta^\top x_i) - m(\theta^\top x_i)|^2 + \frac{2}{n}\max_{1\le i\le n}|m(\eta^\top x_i) - m(\theta^\top x_i)|\sum_{i=1}^{n}|y_i - m(\eta^\top x_i)|.
\end{aligned} \tag{3.37}$$

In view of Lemma 15, for $i = 1, \ldots, n$ we have

$$|m(\theta^\top x_i) - m(\eta^\top x_i)| \le \|m'\|_\infty |x_i^\top(\theta - \eta)| \le C_2(1 + J_R(m))|\theta - \eta|, \tag{3.38}$$

where C_2 is a constant that depends only on L and $\max_{1\le i\le n}|x_i|$. For every $m \in \{m \in \mathcal{S}_R : \|m\|_\infty \le L \text{ and } J_R(m) \le L\}$, (3.37) and (3.38) 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

$$Q_n(m_\eta^R, \theta) - \varepsilon \le Q_n(m_\eta^R, \eta) \text{ and } Q_n(m_\theta^R, \eta) \le Q_n(m_\theta^R, \theta) + \varepsilon. \tag{3.39}$$

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 (3.39), we have

$$\begin{aligned}
Q_n(m_\eta^R,\theta) - \varepsilon \le Q_n(m_\eta^R,\eta) \;&\Leftrightarrow\; \mathcal{L}_n^R(m_\eta^R,\theta;\lambda) - \varepsilon \le \mathcal{L}_n^R(m_\eta^R,\eta;\lambda) \\
&\Rightarrow\; \mathcal{L}_n^R(m_\theta^R,\theta;\lambda) - \varepsilon \le \mathcal{L}_n^R(m_\eta^R,\eta;\lambda) \;\Rightarrow\; T_R(\theta) - \varepsilon \le T_R(\eta)
\end{aligned} \tag{3.40}$$

and

$$\begin{aligned}
Q_n(m_\theta^R,\eta) \le Q_n(m_\theta^R,\theta) + \varepsilon \;&\Leftrightarrow\; \mathcal{L}_n^R(m_\theta^R,\eta;\lambda) \le \mathcal{L}_n^R(m_\theta^R,\theta;\lambda) + \varepsilon \\
&\Rightarrow\; \mathcal{L}_n^R(m_\eta^R,\eta;\lambda) \le \mathcal{L}_n^R(m_\theta^R,\theta;\lambda) + \varepsilon \;\Rightarrow\; T_R(\eta) \le T_R(\theta) + \varepsilon.
\end{aligned} \tag{3.41}$$

Combining (3.40) and (3.41), we have that $T_R(\theta) - \varepsilon \leq T_R(\eta) \leq T_R(\theta) + \varepsilon$, for all $|\eta - \theta| \leq \delta$. 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 $m^R_{\hat{\theta}}$ 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 $m^R_{\hat{\theta}}$ to D.

3.7 Proofs of results in Section 3.3

3.7.1 Proof of Theorem 12

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

$$Q_n(\hat{m}, \hat{\theta}) + \hat{\lambda}_n^2 J^2(\hat{m}) \leq Q_n(m_0, \theta_0) + \hat{\lambda}_n^2 J^2(m_0). \tag{3.42}$$

Observe that by definition of $Q_n(m, \theta)$, we have that (3.42) implies

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

To find rate the 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

$$\begin{aligned}&Q_n(m_0, \theta_0) - Q_n(\hat{m}, \hat{\theta}) \\ &= \frac{2}{n} \sum_{i=1}^n \epsilon_i (\hat{m}(\hat{\theta}^\top x_i) - m_0(\theta_0^\top x_i)) - \frac{1}{n} \sum_{i=1}^n (\hat{m}(\hat{\theta}^\top x_i) - m_0(\theta_0^\top x_i))^2 \\ &\leq \left(\frac{2}{n} \sum_{i=1}^n \epsilon_i^2 \right)^{1/2} \|\hat{m} \circ \hat{\theta} - m_0 \circ \theta_0\|_n - \|\hat{m} \circ \hat{\theta} - m_0 \circ \theta_0\|_n^2.\end{aligned} \tag{3.43}$$

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

$$Q_n(m_0, \theta_0) - Q_n(\hat{m}, \hat{\theta}) \geq \hat{\lambda}_n^2 (J^2(\hat{m}) - J^2(m_0)) \geq -\hat{\lambda}_n^2 J^2(m_0) \geq o_p(1), \tag{3.44}$$

as $\hat{\lambda}_n = o_p(1)$. Combining (3.43) and (3.44), 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 19. *(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 $\hat{m}_1$ and $\hat{m}_2$ such that

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

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

$$\begin{aligned} \frac{\|\hat{m}_1 \circ \hat{\theta}\|_n}{1 + J(m_0) + J(\hat{m})} &\leq \frac{\|\hat{m} \circ \hat{\theta}\|_n}{1 + J(m_0) + J(\hat{m})} + \frac{\|\hat{m}_2 \circ \hat{\theta}\|_n}{1 + J(m_0) + J(\hat{m})} \\ &\leq \frac{\|\hat{m} \circ \hat{\theta}\|_n}{1 + J(m_0) + J(\hat{m})} + \frac{\|\hat{m}_2\|_\infty}{1 + J(m_0) + J(\hat{m})} = O_p(1). \end{aligned} \tag{3.45}$$

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 **(A6)** and (3.2), 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

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

Thus by (3.45) we have

$$\frac{\max(\hat{\beta}_1, \hat{\beta}_2)}{1 + J(m_0) + J(\hat{m})} = O_p(1). \tag{3.46}$$

Moreover, since D is a bounded set, by (3.46) we have $\|\hat{m}_1\|_\infty/(1 + J(m_0) + J(\hat{m})) = O_p(1)$. Combining this with Lemma 19, we get

$$\frac{\|\hat{m}\|_\infty}{1 + J(m_0) + J(\hat{m})} \leq \frac{\|\hat{m}_1\|_\infty}{1 + J(m_0) + J(\hat{m})} + \frac{\|\hat{m}_2\|_\infty}{1 + J(m_0) + J(\hat{m})} = O_p(1). \tag{3.47}$$

Now define the class of functions

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

Observe that by (3.47), we can find a C_ε such that

$$\mathbb{P}\left(\frac{\hat{m} \circ \hat{\theta} - m_0 \circ \theta_0}{1 + J(m_0) + J(m)} \in \mathcal{B}_{C_\varepsilon} \right) \geq 1 - \varepsilon, \qquad \forall n. \tag{3.48}$$

The following lemma in [van de Geer, 2000b] 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 20. *(Lemma 8.4, [van de Geer, 2000b]) Suppose $\mathbb{G}$ be a class of functions. If $\log N_{[\,]}(\delta, \mathbb{G}, \|\cdot\|_\infty) \leq A\delta^{-\alpha}$, $\sup_{g\in\mathbb{G}} \|g\|_n \leq R$, and ϵ satisfies assumption **(A3)**, for some constants $0 < \alpha < 2$, A, and R. Then for some constant c, we have for all $T \geq c$,*

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

In Lemma 21 (proved in Section 3.7.2) we find the bracketing number for the class of functions $\mathcal{B}_C$.

Lemma 21. *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 (3.48), Lemmas 20 and 21 allow us to conclude

$$\frac{(1/n)\sum_{i=1}^n \epsilon_i(\hat{m}(\hat{\theta}^\top x_i) - m_0(\theta_0^\top x_i))}{\|\hat{m}\circ\hat{\theta} - m_0\circ\theta_0\|_n^{3/4}(1+J(m_0)+J(\hat{m}))^{1/4}} = O_p(n^{-1/2}). \tag{3.49}$$

Together, (3.44) and (3.49) imply

$$\begin{aligned}
&\hat{\lambda}_n^2(J^2(\hat{m}) - J^2(m_0)) \\
&\le Q_n(m_0,\theta_0) - Q_n(\hat{m},\hat{\theta}) \\
&= \frac{2}{n}\sum_{i=1}^n (y_i - m_0(\theta_0^\top x_i))(\hat{m}(\hat{\theta}^\top x_i) - m_0(\theta_0^\top x_i)) - \|\hat{m}\circ\hat{\theta} - m_0\circ\theta_0\|_n^2 \\
&\le \|\hat{m}\circ\hat{\theta} - m_0\circ\theta_0\|_n^{3/4}(1+J(m_0)+J(\hat{m}))^{1/4}O_p(n^{-1/2}) - \|\hat{m}\circ\hat{\theta} - m_0\circ\theta_0\|_n^2.
\end{aligned} \tag{3.50}$$

We will now consider two cases.

Case 1: Suppose $J(\hat{m}) > 1 + J(m_0)$. By (3.50), we have

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

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

$$\|\hat{m}\circ\hat{\theta} - m_0\circ\theta_0\|_n^{3/4} J(\hat{m})^{1/4} n^{-1/2} \le C_1\hat{\lambda}_n^2 J^2(m_0) \tag{3.51}$$

or

$$\hat{\lambda}_n^2 J^2(m_0) < C_2\|\hat{m}\circ\hat{\theta} - m_0\circ\theta_0\|_n^{3/4} J(\hat{m})^{1/4} O_p(n^{-1/2}) \tag{3.52}$$

hold with high probability as $n\to\infty$. Observe that when (3.51) holds we have

$$\|\hat{m}\circ\hat{\theta} - m_0\circ\theta_0\|_n^2 + \hat{\lambda}_n^2 J^2(\hat{m}) \le O_p(1)\hat{\lambda}_n^2 J^2(m_0). \tag{3.53}$$

Now it is easy to see that, (3.53) 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 (3.52) holds, we have

$$\|\hat{m}\circ\hat{\theta} - m_0\circ\theta_0\|_n^2 + \hat{\lambda}_n^2 J^2(\hat{m}) \le \|\hat{m}\circ\hat{\theta} - m_0\circ\theta_0\|_n^{3/4} J(\hat{m})^{1/4} O_p(n^{-1/2}). \tag{3.54}$$

We can bound the first term on the left hand side of (3.54) as

$$\|\hat{m} \circ \hat{\theta} - m_0 \circ \theta_0\|_n \leq \left[J(\hat{m})^{1/4} O_p(n^{-1/2})\right]^{4/5}. \tag{3.55}$$

A similar bound on the second term on the left hand side of (3.54) gives:

$$\begin{aligned}
\hat{\lambda}_n^2 J^2(\hat{m}) &\leq \|\hat{m} \circ \hat{\theta} - m_0 \circ \theta_0\|_n^{3/4} J(\hat{m})^{1/4} O_p(n^{-1/2}) \\
&\leq \left[J(\hat{m})^{1/4} O_p(n^{-1/2})\right]^{3/5} J(\hat{m})^{1/4} O_p(n^{-1/2}) \text{ (by (3.55))} \\
&\leq J(\hat{m})^{2/5} \left[O_p(n^{-1/2})\right]^{8/5},
\end{aligned}$$

which implies that

$$J(\hat{m}) = O_p(n^{-1/2}) \hat{\lambda}_n^{-5/4}. \tag{3.56}$$

Combining (3.55) and (3.56), 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 $J(\hat{m}) \leq 1 + J(m_0)$, (3.50) 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 (3.47), $\|\hat{m}\|_\infty = O_p(1)$. To find the rates of convergence of $\|\hat{m} \circ \hat{\theta} - m_0 \circ \theta_0\|$, we use the following lemma.

Lemma 22. *(Lemma 5.16, [van de Geer, 2000b]) Suppose $\mathbb{G}$ is a class of uniformly bounded functions and for some $0 < \nu < 2$,*

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

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

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

Our proof of Theorem 12 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, 2000b].

3.7.2 Proof of Lemma 21

To prove this lemma, we use the following entropy bound from [van de Geer, 2000b]. We will also use the following result in the proofs of Lemmas 17 and 18 in Sections 3.8.5 and 3.8.6, respectively.

Lemma 23. *(Theorem 2.4, [van de Geer, 2000b]) 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 $M_1, M_2 < \infty$, $\|f\|_\infty \leq M_1$, the first $k-1$ derivatives are absolutely continuous and $\left[\int_I [f^{(k)}(x)]^2 dx\right]^{1/2} \leq M_2$. Then there exists a constant C depending on I such that,*

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

The above lemma says that the class of functions

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

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

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

For all $\theta_1, \theta_2 \in \Theta$, we have that $|\theta_1 - \theta_2| \leq 2$. Thus by Lemma 4.1 of [Pollard, 1990], we have

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

Now define the class of functions

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

We will show that

$$\log N_{[\,]}(\varepsilon, \mathcal{H}_{M_1,M_2}, \|\cdot\|_\infty) \lesssim \left(\frac{M_1+M_2}{\varepsilon}\right)^{1/2}. \tag{3.57}$$

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 15, we have $\|m_i'\|_\infty \lesssim 1 + M_2$.

Now we will show that the set of functions $\{m_i \circ \theta_j\}_{1\le i\le q, 1\le j\le p}$ forms 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

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

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 (3.57).

Now we will use (3.57) to prove Lemma 21. 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 $\mathcal{F}^* \subset \mathcal{G}_{C_2,1}$, where $C_2 = \|m_0\|_\infty / J(m_0)$. 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 $h \in \mathcal{F}_C$ 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 $i \in \{1, \dots q\}$ and $j \in \{1, \dots r\}$ cover the required space. Hence, the bracketing entropy

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

where $C_2 = \|m_0\|_\infty / J(m_0)$.

3.7.3 Proof of Theorem 13

The following lemma, proved in the Section 3.7, is crucial to the proof of Theorem 13.

Lemma 24. *For every fixed M, the set of functions $m \in \mathcal{S}$ with $J(m) \leq M$ and $\|m\|_\infty \leq M$ is precompact relative to $\|\cdot\|_D^S$.*

Proof. By Lemma 14 the class of functions m' is uniformly Lipschitz of order $1/2$. Thus any sequence of functions m_k' is equicontinuous. By Lemma 15, m' is uniformly bounded as soon as $J(m)$ is uniformly bounded. Applying the Arzela-Ascoli theorem, we see that every sequence m_k with $J(m_k) = O(1)$ has a subsequence $\{k_l\}$ such that both m_{k_l} and m_{k_l}' converge uniformly on D. By Lemma 15 and the mean value theorem, we get that m is uniformly bounded. Thus applying the Arzela-Ascoli theorem, we get a subsequence $\{k_{l_j}\}$ of $\{k_l\}$ for which functions converge uniformly. 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 $\|m_k \circ \theta_k - m_0 \circ \theta_0\| \to 0$ and $J(m_k) = O(1)$. By Lemma 24, 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 $\|m_k \circ \theta_k - m \circ \theta\| \to 0$ 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 24 is uniform, we get that $\|m - m_0\|_{D_0} = 0$. Combining this with Theorem 12, 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 14, we have that $|\hat{m}'(s) - \hat{m}'(a)| \leq J(\hat{m})|s-a|^{1/2} = O_p(1)$. Moreover, we have that $|\hat{m}'(a) - m_0'(a)| = o_p(1)$. Thus $\|\hat{m}'\|_\infty = O_p(1)$.

3.7.4 Proof of Theorem 14

We first state and prove a lemma that we will use to prove this theorem.

Lemma 25. *Suppose $m \in \mathcal{S}$, $J(m) < \infty$, and $\theta \in \Theta$. Then*

$$P_X\left|m(\theta^\top X) - m(\theta_0^\top X) - m_0'(\theta_0^\top X)X^\top(\theta - \theta_0)\right|^2 \lesssim |\theta - \theta_0|^3 J^2(m) + |\theta - \theta_0|^2 P_X\left|(m - m_0)'(\theta_0^\top X)\right|^2.$$

Proof.

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

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

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

By Lemma 14, we have

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

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

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

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 $|\hat{\theta} - \theta_0| \xrightarrow{P} 0$, $P_X\big|(\hat{m} - m_0)'(\theta_0^\top X)\big|^2 \xrightarrow{P} 0$ and $J(\hat{m}) = O_p(1)$. Thus by Lemma 25, 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 12, we have that $P_X|A(X)|^2 = O_p(\hat{\lambda}_n^2)$. Thus we have

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

Now define

$$g_1(x) := m_0'(\theta_0^\top x)x^\top(\hat{\theta} - \theta_0) \text{ and } \quad g_2(x) := (\hat{m} - m_0)(\theta_0^\top x) \tag{3.58}$$

and note that by assumption **(A7)** there exists a $\lambda_1 > 0$ such that

$$P_X g_1^2 = (\hat{\theta}-\theta_0)^\top P_X[XX^\top |m_0'(\theta_0^\top X)|^2](\hat{\theta}-\theta_0) \geq \lambda_1(\hat{\theta}-\theta_0)^\top(\hat{\theta}-\theta_0) = \lambda_1|\hat{\theta}-\theta_0|^2. \tag{3.59}$$

With (3.59) in mind, we can see that proof of this theorem will be complete if we can show that

$$P_x g_1^2 + P_X g_2^2 \lesssim P_X\big|m_0'(\theta_0^\top X)X^\top(\hat{\theta} - \theta_0) + (\hat{m} - m_0)(\theta_0^\top X)\big|^2. \tag{3.60}$$

The following theorem gives a sufficient condition for (3.60) to hold.

Lemma 26. *(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 (3.58)) satisfy the condition of Lemma 26. Observe that

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

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.

3.8 Proofs of results in Section 3.4

3.8.1 Proof of Lemma 16

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

$$\theta_d := \frac{\theta_0-\theta}{|\theta-\theta_0|} \quad \text{and} \quad \theta_p := \frac{\theta_0-\theta\theta_0^\top\theta}{|\theta_0-\theta\theta_0^\top\theta|} \tag{3.61}$$

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

$$T_\theta^d := \mathbb{I}_d - 2\theta_d\theta_d^\top \quad \text{and} \quad T_\theta^p := \mathbb{I}_d - 2\theta_p\theta_p^\top. \tag{3.62}$$

Note that for every $x \in \mathbb{R}^d$, $x \mapsto T_\theta^d x$ and $x \mapsto T_\theta^p x$ 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, 2001]. 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

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

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

$$H_{\theta_0} := [e_2, \ldots, e_d] \text{ and } H_\theta := T_\theta^p T_\theta^d H_{\theta_0}, \ \forall \theta \neq \theta_0. \tag{3.63}$$

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 $\|H_\theta - H_{\theta_0}\|_2 \leq |\theta_0 - \theta|$. Observe that

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

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

$$\|T_\theta^p T_\theta^d - \mathbb{I}_d\|_2 = \|Q^\top (T_\theta^p T_\theta^d - \mathbb{I}_d) Q\|_2 = \|Q^\top T_\theta^p T_\theta^d Q - \mathbb{I}_d\|_2. \tag{3.64}$$

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 (3.62) and the fact that $\theta_d \in \text{span}\{\theta, \theta_p\}$, we have

$$Q^\top T_\theta^p T_\theta^d Q = \mathbb{I}_d - 2Q^\top \left[\theta_d \theta_d^\top + \theta_p \theta_p^\top - 2\theta_d \theta_d^\top \theta_p \theta_p^\top\right] Q = \begin{bmatrix} A_\theta & \mathbf{0}_{2\times(d-2)} \\ \mathbf{0}_{(d-2)\times 2} & \mathbb{I}_{(d-2)} \end{bmatrix} \tag{3.65}$$

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

(3.65), we have

$$Q^\top T_\theta^p T_\theta^d Q x - x = A_\theta \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} - \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \quad \text{where } x := (x_1, x_2, \ldots, x_d)^\top \in \mathbb{R}^d. \quad (3.66)$$

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

$$\sup_{y \in S^1} |A_\theta y - y| = |A_\theta z - z| \quad (3.67)$$

for all $z \in S^1$; see Page 326, [Meyer, 2001]. 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 (3.66), we have

$$|A_\theta z^0 - z^0| = |Q^\top T_\theta^p T_\theta^d Q x^0 - Q^\top Q x^0| = |Q^\top (\theta - \theta_0)| = |\theta - \theta_0|, \quad (3.68)$$

here 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 (3.64), (3.66), (3.67), and (3.68), we have $\|T_\theta^p T_\theta^d - \mathbb{I}_d\|_2 = |\theta - \theta_0|$.

Before proving (d), we show that for $x \in \mathbb{R}^{d-1}, |H_\theta x| = |x|$ and for $y \in \mathbb{R}^d, |H_\theta^\top y| \leq |y|$. 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 (3.63). 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 $|H_\theta^\top y| \leq |y|$, it is enough to show that $|H_{\theta_0}^\top y| \leq |y|$. 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 $j \neq i$ and $e_i^\top e_i = 1$. Thus $|H_{\theta_0}^\top y| = \sqrt{\sum_{j=2}^d (e_j^\top y)^2} \leq \sqrt{\sum_{j=1}^d (e_j^\top y)^2} = |y|$.

Now we verify that $\{H_\theta, \theta \in \Theta\}$ defined in (3.63) satisfies condition (d) of Lemma 16.

Let $\eta, \beta \in \Theta \setminus \theta_0$ such that $|\eta - \theta_0| < 1/2$, $|\beta - \theta_0| < 1/2$. Note that

$$\begin{aligned}
\|H_\eta^\top - H_\beta^\top\|_2 &= \|H_{\theta_0}^\top [T_\eta^d T_\eta^p - T_\beta^d T_\beta^p]\|_2 \\
&= \sup_{x \in S^{d-1}} \left| H_{\theta_0}^\top [T_\eta^d T_\eta^p - T_\beta^d T_\beta^p] x \right| \\
&\leq \sup_{x \in S^{d-1}} \left| (T_\eta^d T_\eta^p - T_\beta^d T_\beta^p) x \right| \\
&\leq \sup_{x \in S^{d-1}} \left| (T_\eta^d T_\eta^p - T_\eta^d T_\beta^p) x \right| + \sup_{x \in S^{d-1}} \left| (T_\eta^d T_\beta^p - T_\beta^d T_\beta^p) x \right| \\
&= \sup_{x \in S^{d-1}} \left| T_\eta^d (T_\eta^p - T_\beta^p) x \right| + \sup_{x \in S^{d-1}} \left| (T_\eta^d - T_\beta^d) T_\beta^p x \right| \\
&= \sup_{x \in S^{d-1}} \left| (T_\eta^p - T_\beta^p) x \right| + \sup_{x \in S^{d-1}} \left| (T_\eta^d - T_\beta^d) x \right| \\
&= \|T_\eta^p - T_\beta^p\|_2 + \|T_\eta^d - T_\beta^d\|_2, \qquad (3.69)
\end{aligned}$$

here the first inequality is true as $|H_{\theta_0}^\top x| \leq |x|$ 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

$$\|T_\eta^d - T_\beta^d\|_2 \leq 4|\eta_d - \beta_d| \quad \text{and} \quad \|T_\eta^p - T_\beta^p\|_2 \leq 4|\eta_p - \beta_p|, \qquad (3.70)$$

where η_p, η_d, β_p, and β_d are defined as in (3.61). Observe that

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

A similar calculation will show the second equality in (3.70). The proof of (3.8) will be complete if we can show that

$$|\eta_d - \beta_d| \leq 2 \frac{|\eta - \beta|}{|\eta - \theta_0| + |\beta - \theta_0|} \quad \text{and} \quad |\eta_p - \beta_p| \leq \frac{16|\eta - \beta|/\sqrt{15}}{|\eta - \theta_0| + |\beta - \theta_0|}. \qquad (3.71)$$

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

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

and

$$\begin{aligned} |\eta_p - \beta_p| &= \left| \frac{\theta_0 - \eta\theta_0^\top \eta}{|\theta_0 - \eta\theta_0^\top \eta|} - \frac{\theta_0 - \beta\theta_0^\top \beta}{|\theta_0 - \beta\theta_0^\top \beta|} \right| \\ &\leq \frac{2|\eta\theta_0^\top \eta - \beta\theta_0^\top \beta|}{|\theta_0 - \eta\theta_0^\top \eta| + |\theta_0 - \beta\theta_0^\top \beta|}. \end{aligned} \tag{3.72}$$

We now try to simplify (3.72). First note that $|\eta - \theta_0| \leq 1/2$ implies that $1 + \theta_0^\top \eta \geq 15/8$. Now observe that

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

For the numerator of (3.72), we have

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

Combining the above two displays we havepa

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

Combining (3.69), (3.70), and (3.71), we have that

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

3.8.2 Proof of Theorem 16

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

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

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

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

where $\psi'_{t,\theta,\eta}(u) = \frac{\partial}{\partial u}\psi_{t,\theta,\eta}(u)$. Thus, we have that $J^2(\xi_t(\cdot;\theta,\eta,m)) < \infty$ whenever $J(m) < \infty$ and t in a small neighborhood of θ (as $\psi_{t,\theta,\eta}(\cdot)$ is a strictly increasing function). 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

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

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 $\{t, \xi_t(g,\theta)\}$ 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 (3.26)) and $\partial\zeta_t(\theta,\eta)/\partial t = -2t/\sqrt{1-t^2|\eta|^2}\,\theta + H_\theta\eta$, we get

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

Observe that $(\hat m, \hat\theta)$ minimizes the penalized loss function in (3.5) 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

$$t \mapsto \sum_{i=1}^n \left(y_i - \xi_t(\zeta_t(\hat\theta,\eta)^\top x;\hat\theta,\eta,\hat m)\right)^2 + \hat\lambda_n^2 \int_D \left|\frac{\partial^2}{\partial u^2}\xi_t(u;\hat\theta,\eta,\hat m)\right|^2 du \qquad (3.73)$$

on a some small neighborhood of 0 (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

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

This we have that the function in (3.73) 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 (3.16).

3.8.3 Proof of (3.20) in Theorem 15

To prove (3.20), 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 27. *Under assumptions **(A1)**–**(A6)** and **(B2)**–**(B3)**, the PLSE satisfies*

$$P_{\theta_0,m_0} |\tilde{\ell}_{\hat{\theta},\hat{m}} - \tilde{\ell}_{\theta_0,m_0}|^2 = o_p(1), \tag{3.74}$$

$$P_{\hat{\theta},m_0} |\tilde{\ell}_{\hat{\theta},\hat{m}}|^2 = O_p(1). \tag{3.75}$$

Proof. We start with some notation. Let $P_{\theta,m}^{Y|X}$ 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

$$E_{\theta,m}(f) := \int f dP_{\theta,m},\ E_{\theta,m}^X(f) := \int_{\mathbb{R}} f dP_{\theta,m}^{Y|X},\ \text{and } E_X(f) := \int f dP_X. \tag{3.76}$$

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

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

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

Recall that $K_1(x;\theta) = H_\theta^\top\big(x - h_\theta(\theta^\top x)\big)$. To prove (3.74), observe that

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

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 $|H_\theta^\top a| \leq |a|$; see proof of Lemma 16. We will now

show that $\mathbf{I} = o_p(1)$. Observe that

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

where the second inequality follows from (c) of Lemma 16. Let us define

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

Using Lemma 14 and the fact that $\sup_{x\in\mathcal{X}}|x| \leq T$ (see **(A2)**), we have

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

Recall that both $|\hat{\theta} - \theta_0|$ and $\|\hat{m}' - m_0'\|_{D_0}$ are $o_p(1)$; see Theorem 13. Thus we will have $\mathbf{I} = o_p(1)$, if we can show that $\mathbf{III} = o_p(1)$. First observe that by Theorem 13 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:

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

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 (3.74) will be complete, if we can show that $\mathbf{II} = o_p(1)$. First note that for all $x \in \mathcal{X}$,

$$|K_1(x;\theta)| \leq |H_\theta^\top(x - h_\theta(\theta^\top x))| \leq |x - h_\theta(\theta^\top x)| \leq 2T. \tag{3.77}$$

By Theorem 12 and assumption **(A4)**, we have

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

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 (3.75). Observe that

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

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

Now we prove (3.20). 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 (3.22)), we have that $\zeta_{\hat{\eta},\theta_0} = \hat{\theta}$. As $\hat{\eta} = o_p(1)$ (see Theorem 14 and (3.23)) differentiability in quadratic mean of model (3.1) implies that

$$\int\left(\sqrt{p_{\hat{\theta},m_0}} - \sqrt{p_{\theta_0,m_0}} - \frac{1}{2}\hat{\eta}^\top S_{\theta_0,m_0}\sqrt{p_{\theta_0,m_0}}\right)^2 d\mu = o_p(|\hat{\eta}|^2) = o_p(|\hat{\theta} - \theta_0|^2). \tag{3.78}$$

With Lemma 27 in hand, we now show that (3.20) 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_{\theta_0,m_0}^\top)H_{\theta_0}^\top(\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_{\theta_0,m_0}^\top H_{\theta_0}^\top(\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_{\theta_0,m_0}^\top H_{\theta_0}^\top(\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_{\theta_0,m_0}^\top\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_{\theta_0,m_0}^\top H_{\theta_0}^\top(\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_{\theta_0,m_0}^\top H_{\theta_0}^\top(\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_{\theta_0,m_0}^\top\hat{\eta}p_{\theta_0,m_0}d\mu \quad \text{(by (3.23))}\\
&= \mathbf{IV} + \frac{1}{2}\mathbf{V} + \mathbf{VI},
\end{aligned}
$$

where

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

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

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

where the equality is due to Lemma 27, (3.78), and the fact that $\tilde{\ell}_{\theta_0,m_0} \in L_2(P_{\theta_0,m_0})$ (see **(A1)**, **(A2)**, and Lemma 15).

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

$$|\mathbf{VI}|^2 \le \left\| \int [\tilde{\ell}_{\hat{\theta},\hat{m}} - \tilde{\ell}_{\theta_0,m_0}] S^\top_{\theta_0,m_0} p_{\theta_0,m_0} d\mu \right\|_F^2 |\sqrt{n}\hat{\eta}|^2. \tag{3.79}$$

Let $f = (f_1, \ldots, f_d)$ and $g = (g_1, \ldots, g_d)$ 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

$$\begin{aligned}
\left\| \int_\Re f g^\top d\nu \right\|_F^2 &= \sum_{i,j} \left[\int_\Re f_i g_j d\nu \right]^2 \le \sum_{i,j} \int_\Re f_i^2 d\nu \int_\Re g_j^2 d\nu \\
&= \left[\sum_i \int_\Re f_i^2 d\nu \right] \left[\sum_j \int_\Re g_j^2 d\nu \right] = \int_\Re |f|^2 d\nu \int_\Re |g|^2 d\nu.
\end{aligned} \tag{3.80}$$

Thus from Lemma 27, (3.79), and the fact that $S_{\theta_0,m_0} \in L_2(P_{\theta_0,m_0})$, we have

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

To prove that $|\mathbf{V}|^2$ is $o_p(|\sqrt{n}(\hat{\theta} - \theta_0)|^2)$, first observe that

$$|\mathbf{V}|^2 \le \left\| \int \tilde{\ell}_{\hat{\theta},\hat{m}} (\sqrt{p_{\hat{\theta},m_0}} - \sqrt{p_{\theta_0,m_0}}) S^\top_{\theta_0,m_0} \sqrt{p_{\theta_0,m_0}} d\mu \right\|_F^2 |\sqrt{n}\hat{\eta}|^2. \tag{3.81}$$

We split the integral on the right hand side of the above display into two parts depending on whether $|S_{\theta_0,m_0}| > m_n$ or $|S_{\theta_0,m_0}| \le m_n$ and apply the Cauchy-Schwarz inequality.

Observe that by (3.80), we have

$$
\begin{aligned}
&\left\| \int_{|S_{\theta_0,m_0}|\le m_n} \tilde{\ell}_{\hat{\theta},\hat{m}} S_{\theta_0,m_0}^\top (\sqrt{p_{\hat{\theta},m_0}} - \sqrt{p_{\theta_0,m_0}}) \sqrt{p_{\theta_0,m_0}} d\mu \right\|_F^2 \\
&= \sum_{i,j} \left[\int_{|S_{\theta_0,m_0}|\le m_n} \left\{ \tilde{\ell}_{\hat{\theta},\hat{m}} S_{\theta_0,m_0}^\top \right\}_{i,j} \left(\sqrt{p_{\hat{\theta},m_0}} - \sqrt{p_{\theta_0,m_0}} \right) \sqrt{p_{\theta_0,m_0}} d\mu \right]^2 \\
&\le \left[\int \left(\sqrt{p_{\hat{\theta},m_0}} - \sqrt{p_{\theta_0,m_0}} \right)^2 d\mu \right] \sum_{i,j} \left[\int_{|S_{\theta_0,m_0}|\le m_n} \left\{ \tilde{\ell}_{\hat{\theta},\hat{m}} S_{\theta_0,m_0}^\top \right\}_{i,j}^2 p_{\theta_0,m_0} d\mu \right] \\
&\le 2 \left[\int \left(\frac{1}{2} S_{\theta_0,m_0}^\top (\hat{\theta} - \theta_0) \sqrt{p_{\theta_0,m_0}} \right)^2 + \left(\sqrt{p_{\hat{\theta},m_0}} - \sqrt{p_{\theta_0,m_0}} - \frac{1}{2} S_{\theta_0,m_0}^\top (\hat{\theta} - \theta_0) \sqrt{p_{\theta_0,m_0}} \right)^2 d\mu \right] \\
&\qquad \times \sum_{i,j} \left[\int_{|S_{\theta_0,m_0}|\le m_n} \left\{ \tilde{\ell}_{\hat{\theta},\hat{m}} S_{\theta_0,m_0}^\top \right\}_{i,j}^2 p_{\theta_0,m_0} d\mu \right] \\
&= 2 \left[\int \left(\frac{1}{2} S_{\theta_0,m_0}^\top (\hat{\theta} - \theta_0) \sqrt{p_{\theta_0,m_0}} \right)^2 + \left(\sqrt{p_{\hat{\theta},m_0}} - \sqrt{p_{\theta_0,m_0}} - \frac{1}{2} S_{\theta_0,m_0}^\top (\hat{\theta} - \theta_0) \sqrt{p_{\theta_0,m_0}} \right)^2 d\mu \right] \\
&\qquad \times \int_{|S_{\theta_0,m_0}|\le m_n} |\tilde{\ell}_{\hat{\theta},\hat{m}}|^2 |S_{\theta_0,m_0}^\top|^2 p_{\theta_0,m_0} d\mu \\
&\le 2 m_n^2 P_{\theta_0,m_0} |\tilde{\ell}_{\hat{\theta},\hat{m}}|^2 \left[O_p(|\hat{\theta} - \theta_0|^2) + o_p(|\hat{\theta} - \theta_0|^2) \right] = m_n^2 o_p(1)
\end{aligned}
\tag{3.82}
$$

and by (3.80), we have

$$
\begin{aligned}
&\left\| \int_{|S_{\theta_0,m_0}|> m_n} \tilde{\ell}_{\hat{\theta},\hat{m}} S_{\theta_0,m_0}^\top (\sqrt{p_{\hat{\theta},m_0}} - \sqrt{p_{\theta_0,m_0}}) \sqrt{p_{\theta_0,m_0}} d\mu \right\|_F^2 \\
&\le 2 \int_{|S_{\theta_0,m_0}|> m_n} |S_{\theta_0,m_0}|^2 p_{\theta_0,m_0} d\mu \int |\tilde{\ell}_{\hat{\theta},\hat{m}}|^2 (p_{\hat{\theta},m_0} + p_{\theta_0,m_0}) d\mu \\
&\le O_p(1) \int_{|S_{\theta_0,m_0}|> m_n} |S_{\theta_0,m_0}|^2 p_{\theta_0,m_0} d\mu.
\end{aligned}
\tag{3.83}
$$

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 $\{m_n\}$ such that both (3.82) and (3.83) are $o_p(1)$. Thus by (3.81), we have $|\mathbf{V}|^2 = o_p(|\sqrt{n}(\hat{\theta} - \theta_0)|^2)$, which completes the proof.

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

Theorem 18. *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 (3.76), we have $E^X_{\theta,m_0}\big[Y - m(\theta^\top X)\big] = m_0(\theta^\top X) - m(\theta^\top X)$. Thus

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

□

3.8.5 Proof of Lemma 17

Before proceeding to prove Lemma 17, we find the entropy of the class of matrices $\{H_\theta : \theta \in \Theta\}$, where H_θ satisfies conditions of Lemma 16.

Lemma 28. *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*

$$|\theta - \eta_i| \le \varepsilon \text{ and } \|H_\theta^\top - H_{\eta_i}^\top\|_2 \le \varepsilon. \tag{3.84}$$

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 16, 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 16, we have

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

□

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 $x \in \mathcal{X}$, we have

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

where the first and second inequality follow from the facts that $\sup_{x\in\mathcal{X}}|x| \le T$ and $\|K_1(\cdot;\theta)\|_{2,\infty} \le 2T$, see **(A2)** and (3.77). Next we prove that there exists finite c depending only on M_1, M_2, and M_3, such that

$$N(\varepsilon, \mathcal{D}^*_{M_1,M_2,M_3}, \|\cdot\|_{2,\infty}) \le c\exp\left(\frac{c}{\varepsilon} + \frac{c}{\sqrt{\varepsilon}}\right)\varepsilon^{-2(d-1)}. \tag{3.85}$$

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 23 (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 $r_1,\dots,r_q$ and the functions in the ε-cover of $\{f' : f \in \mathcal{C}^{m*}_{M_1,M_2,M_3}\}$ by $l_1,\dots,l_t$. By Lemma 28, we have that there exists $\theta_1,\dots,\theta_s$ 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 (3.84) (satisfies (3.84)). 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.,

$$\|m - r_1\|_\infty \le \varepsilon, \quad \|m' - l_1\|_\infty \le \varepsilon, \quad \|H_{\theta_1}^\top - H_\theta^\top\|_2 \le \varepsilon^2 \quad \text{and} \quad |\theta_1 - \theta| \le \varepsilon^2. \tag{3.86}$$

Now for every $x \in \mathcal{X}$, observe that

$$\begin{aligned}
&\big|(m_0(\theta_0^\top x) - m(\theta^\top x))m'(\theta^\top x)K_1(x;\theta) - (m_0(\theta_0^\top x) - r_1(\theta_1^\top x))l_1(\theta_1^\top x)K_1(x;\theta_1)\big| \\
&= \big|(m_0(\theta_0^\top x) - m(\theta^\top x))m'(\theta^\top x)K_1(x;\theta) - \\
&\qquad \big(m_0(\theta_0^\top x) - m(\theta^\top x) + m(\theta^\top x) - r_1(\theta_1^\top x)\big)l_1(\theta_1^\top x)K_1(x;\theta_1)\big| \\
&\le \big|m_0(\theta_0^\top x) - m(\theta^\top x)\big|\big|m'(\theta^\top x)K_1(x;\theta) - l_1(\theta_1^\top x)K_1(x;\theta_1)\big| \\
&\qquad + \big|m(\theta^\top x) - r_1(\theta_1^\top x)\big|\big|l_1(\theta_1^\top x)K_1(x;\theta_1)\big| \\
&= \mathbf{A} + \mathbf{B},
\end{aligned} \tag{3.87}$$

where

$$\begin{aligned}
\mathbf{A} &:= \big|m_0(\theta_0^\top x) - m(\theta^\top x)\big|m'(\theta^\top x)K_1(x;\theta) - l_1(\theta_1^\top x)K_1(x;\theta_1)\big| \\
\mathbf{B} &:= \big|m(\theta^\top x) - r_1(\theta_1^\top x)\big|\big|l_1(\theta_1^\top x)K_1(x;\theta_1)\big|.
\end{aligned}$$

We next find an upper bound for $\mathbf{A}$. First, by Lemma 16 and assumption **(B2)**, we have

$$\begin{aligned}
&\big|K_1(x;\theta) - K_1(x;\theta_1)\big| \\
&= \big|H_\theta^\top(x - h_\theta(\theta^\top x)) - H_{\theta_1}^\top(x - h_\theta(\theta^\top x)) + H_{\theta_1}^\top(x - h_\theta(\theta^\top x)) - H_{\theta_1}^\top(x - h_{\theta_1}(\theta_1^\top x))\big| \\
&\le \big|(H_\theta^\top - H_{\theta_1}^\top)(x - h_\theta(\theta^\top x))\big| + \big|H_{\theta_1}^\top\big[(x - h_\theta(\theta^\top x)) - (x - h_{\theta_1}(\theta_1^\top x))\big]\big| \\
&\le \|H_\theta^\top - H_{\theta_1}^\top\|_2 2T + \big|h_\theta(\theta^\top x) - h_{\theta_1}(\theta_1^\top x)\big| \\
&\le T\varepsilon^2 + (\bar{M} + \|h'_{\theta_0}\|_\infty\|)|\theta - \theta_1| \lesssim \varepsilon^2.
\end{aligned} \tag{3.88}$$

Now observe that

$$\begin{aligned}
\mathbf{A} &\leq 2M_2\big|m'(\theta^\top x)K_1(x;\theta) - l_1(\theta_1^\top x)K_1(x;\theta_1)\big| \\
&\leq \big|m'(\theta^\top x)K_1(x;\theta) - l_1(\theta^\top x)K_1(x;\theta)\big| + \big|l_1(\theta^\top x)K_1(x;\theta) - l_1(\theta_1^\top x)K_1(x;\theta)\big| \\
&\qquad + \big|l_1(\theta_1^\top x)K_1(x;\theta) - l_1(\theta_1^\top x)K_1(x;\theta_1)\big| \\
&\leq |K_1(x;\theta)|\big|m'(\theta^\top x) - l_1(\theta^\top x)\big| + |K_1(x;\theta)|\big|l_1(\theta^\top x) - l_1(\theta_1^\top x)\big| \\
&\qquad + \|l_1\|_\infty\big|K_1(x;\theta) - K_1(x;\theta_1)\big| \\
&\leq 2T(\varepsilon + \left[\int_D l_1'^2(z)dz\right]|\theta - \theta_1|^{1/2}T^{1/2}) + M_2(2T + \bar{M})\varepsilon^2 \\
&\leq 2T(\varepsilon + M_3|\theta - \theta_1|^{1/2}T^{1/2}) + (2T + \bar{M})M_2\varepsilon^2 \\
&\lesssim \varepsilon,
\end{aligned} \tag{3.89}$$

where the penultimate inequality follows from (3.86) and the last inequality follows from **(A2)**, (3.88), and Lemma 14. To find an upper bound for $\mathbf{B}$, observe that

$$\begin{aligned}
\mathbf{B} &= \big|m(\theta^\top x) - r_1(\theta_1^\top x)\big|\big|l_1(\theta_1^\top x)K_1(x;\theta_1)\big| \\
&\leq \Big[\big|m(\theta^\top x) - r_1(\theta^\top x)\big| + \big|r_1(\theta^\top x) - r_1(\theta_1^\top x)\big|\Big]\big|l_1(\theta_1^\top x)K_1(x;\theta_1)\big| \\
&\leq \big[\varepsilon + \|r_1'\|_\infty|\theta - \theta_1|T\big]\|l_1\|_\infty 2T \lesssim \varepsilon.
\end{aligned} \tag{3.90}$$

Combining (3.87), (3.89), and (3.90) 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 $1 \leq i \leq q, 1 \leq j \leq t$, and $1 \leq k \leq m$ 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 (3.85). Moreover, as $N_{[]}(\varepsilon, \mathcal{D}^*_{M_1,M_2,M_3}, {}_{2,P_{\theta_0,m_0}}) \lesssim N(\varepsilon, \mathcal{D}^*_{M_1,M_2,M_3}, \|\cdot\|_{2,\infty})$ 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 χ 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

$$\begin{aligned}
&\mathbb{P}\Big(\sup_{f \in \mathcal{D}_{M_1,M_2,M_3}(n)} |\mathbb{G}_n f| > \delta\Big) \\
&\leq \sum_{i=1}^{d-1} \mathbb{P}\Big(\sup_{f \in \mathcal{D}_{M_1,M_2,M_3}(n)} |\mathbb{G}_n f_i| > \delta/\sqrt{d-1}\Big).
\end{aligned} \tag{3.91}$$

We can bound each term in the summation of (3.91) using the maximal inequality in Corollary 19.35 of [van der Vaart, 1998b]. We have

$$\begin{aligned}\mathbb{P}\Big(\sup_{f\in\mathcal{D}_{M_1,M_2,M_3}(n)}|\mathbb{G}_n f_1|>\delta\Big) &\leq \delta^{-1}\mathbb{E}\Big(\sup_{f\in\mathcal{D}_{M_1,M_2,M_3}(n)}|\mathbb{G}_n f_1|\Big)\\ &\leq \delta^{-1}J_{[]}(\|D_{M_1,M_2,M_3}(n)\|,\mathcal{D}^*_{M_1,M_2,M_3}(n),\|\cdot\|_{2,P_{\theta_0,m_0}})\\ &\lesssim \delta^{-1}\|D_{M_1,M_2,M_3}(n)\|^{1/2}\\ &\lesssim \Big[\hat{\lambda}_n^{1/2}+a_n^{-1}\Big]^{1/2}\to 0, \qquad \text{as } n\to\infty. \end{aligned} \tag{3.92}$$

In the last inequality, we have used (3.29) and the fact that $D^2_{M_1,M_2,M_3}(n)$ is non-random. The lemma follows by combining (3.92) and (3.91).

3.8.6 Proof of Lemma 18

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

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

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

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

where for the third term in the penultimate inequality we have used (c) of Lemma 16 and **(B2)**.

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

that

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

As in proof of Lemma 17, 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 23 and Lemma 4.1 of [Pollard, 1990], we have

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

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 28, we have that there exists $\theta_1, \ldots, \theta_s$ for $s \lesssim \varepsilon^{-4d}$ such that $\{\theta_i\}_{1\leq i\leq s}$ form an ε^2-cover of $\Theta \cap B_{\theta_0}(1/2)$ and satisfies (3.84) (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.,

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

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

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

Here the last inequality follows from **(A2)**, (3.88), and Lemma 14. Thus, $\{U_{\theta_i,r_j} - U_{\theta_0,m_0}\}_{1\leq i\leq t, 1\leq j\leq m}$ 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

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

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

$$N_{[]}\big(\varepsilon\|\sigma(\cdot)\|_\infty, \{\epsilon f : f \in \mathcal{W}_{M_1,M_2,M_3}(n)\}, \|\cdot\|_{2,P_{\theta_0,m_0}}\big) \le c\exp(c/\varepsilon)\varepsilon^{-2(d-1)}. \tag{3.93}$$

Now we prove (3.30). As in (3.28), we have

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

By discussion similar to those after Theorem 17, 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 (3.93), we have

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

Thus arguments similar to (3.91) and (3.92), we have

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

Chapter 4

Efficient Estimation in Convex Single Index Models[1]

We consider estimation and inference in a single index regression model with an unknown convex link function. We propose two estimators for the unknown link function: (1) a shape-constrained smoothing spline estimator and (2) a Lipschitz constrained least squares estimator. Moreover, both these procedures lead to estimators for the unknown finite dimensional parameter. We develop methods to compute both the penalized least squares estimator (PLSE) and the Lipschitz constrained least squares estimator (LLSE) of the parametric and the nonparametric components given i.i.d. data. We prove the consistency and find the rates of convergence for both the PLSE and the LLSE. For both the PLSE and the LLSE, we establish $n^{-1/2}$-rate of convergence and semiparametric efficiency of the parametric component under mild assumptions. We develop the R package `simest` *to compute the proposed estimators. Our proposed algorithm works even when n is modest and d is large (e.g., $n = 500$, and $d = 100$).*

Keywords: Approximately least favorable sub-provided models, interpolation inequality, penalized least squares, shape restricted function estimation.

[1] Joint work with Arun Kumar Kuchibhotla and Bodhisattva Sen.

4.1 Introduction

In this chapter we consider the model

$$Y = m_0(\theta_0^\top X) + \epsilon, \quad \mathbb{E}(\epsilon|X) = 0, \quad \text{almost every (a.e.) } X, \tag{4.1}$$

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 error. We assume that both m_0 and θ_0 are unknown and are the parameters of interest. Further, we assume that m_0 is convex. This assumption is motivated by the fact that in a wide range of applications in various fields the link function is known to be convex or concave. For example, in microeconomics, production functions are often supposed to be concave and component-wise nondecreasing; concavity indicates decreasing marginal returns. Also utility functions are often assumed to be concave (decreasing marginal utility); see [Li and Racine, 2007].

Shape constrained inference has a long history in the statistical literature dating back to the seminal papers [Hildreth, 1954], [Brunk, 1955], and [Grenander, 1956]. In the case of convex univariate regression the properties of the least squares estimator are well-studied; see [Hildreth, 1954; Hanson and Pledger, 1976; Groeneboom *et al.*, 2001; Dümbgen *et al.*, 2004], and [Guntuboyina and Sen, 2013] for consistency, local and global rates of convergence, and computational aspects of the least squares estimator.

A drawback of the convex shape constrained least squares estimator is that it is piecewise linear. Quite often in practice a smooth estimator is preferred. A natural way to obtain smooth convex estimator is by penalizing the least squares loss with a penalty on the roughness of the convex function through the integrated squared second derivative. For univariate convex regression [Elfving and Andersson, 1988] provide a characterization of the constrained penalized least squares estimator while [Mammen and Thomas-Agnan, 1999] provide their rates of convergence.

Despite large interest in shape-restricted single index models, estimation and inference is not very well-studied. The earliest reference we could find was the work [Murphy *et al.*, 1999], where the authors consider a constrained penalized likelihood in a current status regression model with a monotone link function. During the preparation of this chapter we became aware of [Chen and Samworth, 2014] and [Groeneboom and Hen-

drickx, 2016]. [Groeneboom and Hendrickx, 2016] provide a $\sqrt{n}$-consistent estimator of the index vector in the current status model under monotonicity constraint on the link function. [Chen and Samworth, 2014] consider maximum likelihood estimation in a generalized additive index model (slightly more general model than (4.1)) and prove consistency of the proposed estimators. However, rates of convergence or asymptotic distributions of the estimators are not studied.

In this chapter we propose a constrained penalized least squares estimator (PLSE) and a Lipschitz constrained least squares estimator (LLSE); see Section 4.3. As in Chapter 3, for identifiability of the model (4.1) we assume that the first coordinate of θ_0 is non-zero and

$$\theta_0 \in \Theta := \{\eta_0 \in \mathbb{R}^d : |\eta_0| = 1 \text{ and } \eta_{0,1} \geq 0\} \subset S^{d-1}, \tag{4.2}$$

where $\eta_{0,1}$ is the first coordinate of η_0, $|\cdot|$ 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.

Our exposition is organized as follows: in Section 4.2, we introduce the model and define some notation. In Section 4.3, we propose two estimators for (m_0, θ_0). In Sections 4.4.1 and 4.4.2, we state our assumptions, prove consistency, and give rates of convergence of the PLSE and LLSE, respectively. In Section 4.5, we use these rates to prove efficiency and asymptotic normality of the PLSE and the LLSE of the index vector. We discuss an algorithm to calculate our estimators in Section 4.6. In Section 4.7, we provide a finite sample simulation study and compare performance with existing methods in the literature. Sections 4.8–4.12 contain proofs omitted from the main text.

4.2 Preliminaries

In what follows, we assume that we have independent and identically distributed (i.i.d.) data $\{(x_i, y_i)\}_{1 \leq i \leq n}$ from (4.1). We start with some notation. Let $\chi \subset \mathbb{R}^d$ denote the support of X, define

$$D := \{\theta^\top x | x \in \chi, \theta \in \Theta\}, \text{ and } Q_n(m, \theta) := \frac{1}{n} \sum_{i=1}^{n} \{y_i - m(\theta^\top x_i)\}^2.$$

Let $\mathcal{C}$ denote the class of real-valued convex functions on D, $\mathcal{S}$ denote the class of functions from D to $\mathbb{R}$ that have an absolutely continuous first derivative, and $\mathfrak{L}_L$ denote the class of uniformly Lipschitz functions from D to $\mathbb{R}$ with Lipschitz bound L. Now, define

$$\mathcal{R} := \mathcal{S} \cap \mathcal{C} \text{ and } \mathcal{M}_L := \mathfrak{L}_L \cap \mathcal{C}.$$

For any $m \in \mathcal{S}$, we define

$$J^2(m) := \int_D \{m''(t)\}^2 dt.$$

For any $m \in \mathcal{M}_L$, let m' denote the nondecreasing right derivative of real-valued the convex function m. As m is an uniformly Lipschitz function with Lipschitz constant L, we can assume that $|m'(t)| \leq L, \forall t \in D$. We use $\mathbb{P}$ to denote the probability of an event, $\mathbb{E}$ to denote expectation of a random quantity, and P_X to denote the marginal distribution of X. For $g : \chi \to \mathbb{R}$, define

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

Let $P_{\theta,m}$ denote the joint distribution of (Y, X) when $X \sim P_X$ and $Y := m(\theta^\top X) + \epsilon$, where ϵ is defined in (4.1). In particular, P_{θ_0,m_0} denotes the joint distribution of (Y, X) when $X \sim P_X$ and (Y, X) satisfies (4.1). For any set $I \subset \mathbb{R}^p$ $(p \geq 1)$ and any function $g : I \to \mathbb{R}$, we define $\|g\|_\infty := \sup_{u\in I} |g(u)|$. Moreover, for $I_1 \subsetneq I$, we define $\|g\|_{I_1} := \sup_{u\in I_1} |g(u)|$. For any function $g : I \to \mathbb{R}$, the Sobolev norm is defined as

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

where $I \subset \mathbb{R}$. The notation $a \lesssim b$ is used to express that a is less than b up to a constant multiple. For any function $f : \chi \to \mathbb{R}^r, r \geq 1$, let $\{f_i\}_{1\leq i\leq r}$ denote each of the components, i.e., $f(x) = (f_1(x), \ldots, f_r(x))$ 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 class $\mathcal{G}$ of functions on the real line, $\mathcal{G} \circ \Theta$ denotes the set of linear index functions

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

with $(g, \theta) \in \mathcal{G} \times \Theta$. We use standard empirical process theory notation. For any function $f : \mathbb{R} \times \chi \mapsto \mathbb{R}$, θ, and m, we define

$$P_{\theta,m} f := \int f(y, x) dP_{\theta,m}(y, x).$$

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

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

The following two lemmas, proved in Section 4.8, will be useful in the remainder of the chapter.

Lemma 29. *Let $m \in \{g \in \mathcal{R} : J(g) < \infty\}$. Then $|m'(s) - m'(s_0)| \leq J(m)|s - s_0|^{1/2}$ for every $s, s_0 \in D$.*

Lemma 30. *For any set $A \in \mathbb{R}^k$ $(k \geq 1)$, let $\varnothing(A)$ denote the diameter of the set A. Let $m \in \{g \in \mathcal{R} : J(g) < \infty$ and $\|g\|_\infty \leq M\}$, where M is a finite constant. Then*

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

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

The following lemma (proved in Section 4.8.3) proves the identification of the composite population parameter $m_0 \circ \theta_0$.

Lemma 31. *Define $Q(m, \theta) := \mathbb{E}[Y - m(\theta^\top X)]^2$. Then*

$$\inf_{\{(m,\theta):\, m\circ\theta \in L_2(P_X) \text{ and } \|m\circ\theta - m_0\circ\theta_0\| > \delta\}} Q(m, \theta) - Q(m_0, \theta_0) > \delta^2. \tag{4.3}$$

Remark 9. *(4.3) tells us that one can hope to consistently estimate (m_0, θ_0) by minimizing $Q_n(m, \theta)$, the sample version of $Q(m, \theta)$.*

Note that identification of $m_0 \circ \theta_0$ 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 almost everywhere differentiable function:

(A0) For some integer $d_1 \in (0, d]$, let $(X_1, \ldots, X_{d_1})$ have continuous marginal distribution and $X_{d_1+1}, \ldots, X_{d-1}$, and X_d are discrete random variables. Furthermore, assume that for each $\theta \in \Theta$, there exists an open interval $\mathcal{I}$ and 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,
- $\mathcal{I} \subset \bigcap_{l=0}^{d-d_1} \left\{\theta^\top x : x \in \mathcal{X} \text{ and } (x_{d_1+1}, \ldots, x_d) = c_l\right\}$.

4.3 Two estimators

As seen in Section 4.2, the least squares is a reasonable loss function to consider, i.e., (m_0, θ_0) minimize the population version of the square error loss. A natural estimator (under the convexity constraint on the link function) would be the following least squares estimator (LSE)

$$(m_n^\dagger, \theta_n^\dagger) := \underset{(m,\theta) \in \mathcal{C} \times \Theta}{\arg\min}\ Q_n(m, \theta). \tag{4.4}$$

The above minimizer is well-defined and can be computed easily using a quadratic program (with linear constraints). However it is difficult to study the estimator theoretically. The difficulty can be attributed to the inconsistency of $m_n^\dagger$ at the "boundary" of its domain; it is well-known that shape constrained estimates can be inconsistent at the boundary; see [Woodroofe and Sun, 1993]. In single index models the inconsistency of $m_n^\dagger$ at the boundary affects the estimation of θ_0 as θ_0 and m_0 are intertwined (as opposed to a partially linear model).

In what follows, we propose two variants of (4.4) and study their asymptotic properties. The first estimator is obtained by forcing the minimizer to be a smooth function through a penalization; see Section 4.3.1. The second estimator is obtained by constraining the set over which we minimize the loss function. Instead of minimizing $Q_n(m, \theta)$ over the class of all convex functions, we minimize the square error loss over the class of uniformly Lipschitz convex function; see Section 4.3.2.

4.3.1 Penalized least squares estimator (PLSE)

With the goal of making the estimator of m smooth, we propose the following penalized loss,

$$\mathcal{L}_n(m, \theta; \lambda) := Q_n(m, \theta) + \lambda^2 \int_D J^2(m), \qquad (\lambda \neq 0).$$

The PLSE can now be defined as

$$(\hat{m}_n, \hat{\theta}_n) := \underset{(m,\theta)\in\mathcal{R}\times\Theta}{\arg\min}\ \mathcal{L}_n(m, \theta; \lambda). \tag{4.5}$$

For notational convenience, we suppress the dependence of $(\hat{m}_n, \hat{\theta}_n)$ on λ. The following theorem, proved in Section 4.9, shows that the joint minimizer is well-defined and $\hat{m}_n$ is a cubic spline.

Theorem 19. *$(\hat{m}_n, \hat{\theta}_n) \in \mathcal{R} \times \Theta$. Moreover, $\hat{m}$ is a natural cubic spline.*

In Sections 4.4.1 and 4.5.2, we study the asymptotic properties of the above estimator.

4.3.2 Lipschitz constrained least squares estimator (LLSE)

Another way to modify (4.4) is to minimize $Q_n(\cdot, \cdot)$ over $\mathcal{M}_L \times \Theta$ instead of $\mathcal{C} \times \Theta$ for some fixed L. We call such an estimator the LLSE. It is defined as

$$(\check{m}_n, \check{\theta}_n) = \underset{(m,\theta)\in\mathcal{M}_L\times\Theta}{\arg\min}\ Q_n(m, \theta). \tag{4.6}$$

The uniform Lipschitz restriction modifies the estimator of m at the boundary of D. As we will show in Section 4.4.2, as long as our class $\mathcal{M}_L$ contains the truth, $(\check{m}_n, \check{\theta}_n)$ are reasonable estimators of the truth. As in the case of PLSE, we suppress the dependence of $(\check{m}_n, \check{\theta}_n)$ on the control parameter L. The following theorem, proved in Section 4.9, shows the existence of the minimizer in (4.6).

Theorem 20. *$(\check{m}_n, \check{\theta}_n) \in \mathcal{M}_L \times \Theta$. Moreover, $\check{m}$ is a piecewise linear convex function.*

In Sections 4.4.2 and 4.5.3, we study the asymptotic properties of the above estimator.

Remark 10. *For every fixed θ, $m(\in \mathcal{R}) \mapsto \mathcal{L}_n(m, \theta; \lambda)$ has a unique minimizer; see Section 2 of [Elfving and Andersson, 1988] and Section 4 of [Utreras, 1985] for algorithms for finding the minimizer of this (constrained) penalized loss function.*

For every fixed θ, $m(\in \mathcal{M}_L) \mapsto Q_n(m,\theta)$ has a unique minimizer. The minimization over class of uniformly Lipschitz function is a quadratic program with linear constraints and can be computed easily. In Section 4.6 we discuss algorithms to compute $(m_n^\dagger, \theta_n^\dagger)$, $(\hat{m}_n, \hat{\theta}_n)$, and $(\check{m}_n, \check{\theta}_n)$.

4.4 Asymptotic analysis

In Sections 4.4.1 and 4.4.2, we study the asymptotic behavior of the estimators proposed in Sections 4.3.1 and 4.3.2, respectively. When there is no scope for confusion, for the rest of the chapter, we use $(m^\dagger, \theta^\dagger)$, $(\hat{m}, \hat{\theta})$, and $(\check{m}, \check{\theta})$ to denote $(m_n^\dagger, \theta_n^\dagger)$, $(\hat{m}_n, \hat{\theta}_n)$, and $(\check{m}_n, \check{\theta}_n)$, respectively. We will now list the assumptions under which we prove the consistency and study the rates of convergence of the estimators proposed in Sections 4.3.1 and 4.3.2.

(A1) The support of X, $\mathcal{X}$, is a compact subset of $\mathbb{R}^d$ and we assume that $\sup_{x\in\mathcal{X}} |x| \leq T$.

(A2) The error variable ϵ in model (4.1) is assumed to be a uniformly sub-Gaussian random random variable, i.e., there exist $K_1 > 0$ and $K_2 \in \mathbb{R}$ such that

$$K_1^2 \mathbb{E}\big(\exp(\epsilon^2/K_1^2) - 1 | X\big) \leq K_2^2 \quad \text{a.e. } X.$$

As stated in (4.1), we also assume that $\mathbb{E}(\epsilon|X) = 0$ a.e $(\mathbb{P})$.

(A3) $\mathbb{E}[XX^\top \{m_0'(\theta_0^\top X)\}^2]$ is a nonsingular matrix.

(A4) $\mathrm{Var}(X)$ is a positive definite matrix.

Define

$$D_0 := \{x^\top \theta_0 : x \in \mathcal{X}\}, \quad D_\theta := \{\theta^\top x : x \in \mathcal{X}\}.$$

(A5) There exists a $r > 0$, such that for every $\theta \in \{\eta \in \Theta : |\eta - \theta_0| \leq r\}$ the density of $\theta^\top X$ with respect to the Lebesgue measure is bounded away from zero and infinity on D_θ. Furthermore, we assume that for every $\theta \in \{\eta \in \Theta : |\eta - \theta_0| \leq r\}$, $D_\theta \subsetneq D$, where $D := \cup_{|\theta-\theta_0|\leq r} D_\theta$.

The assumptions deserve comments. Assumption **(A1)** is the standard bounded support assumption used in the empirical process theory. As the classes of functions $\mathcal{M}_L$ and $\mathcal{R}$ are not uniformly bounded, we need assumption **(A2)** to provide control over the tail behavior of ϵ; see Chapter 8 of [van de Geer, 2000b] for a discussion on this. Observe that **(A2)** allows for heteroscedastic errors. Assumption **(A3)** is similar to that in [Murphy *et al.*, 1999] and helps us obtain the rates of convergence of estimators of m_0 and θ_0 separately from the rate of convergence of the estimators of $m_0 \circ \theta_0$. We use empirical process methods (e.g., see [van der Vaart, 1998b]) methods to prove the consistency and to find the rates of convergence of the proposed estimators. The "size" of function classes with respect to the $\|\cdot\|_n$ norm and the configuration of $\{x_i\}_{1\leq i\leq n}$ play a crucial role in our analysis. Assumption **(A4)** helps us guarantee that the data points are well behaved. Assumption **(A5)** guarantees that the true index set (i.e., $\{\theta_0^\top x : x \in \mathcal{X}\}$) does not lie on the boundary of D. Assumption **(A5)** is needed to find rates of convergence of derivative of the estimators of m_0.

4.4.1 Asymptotic analysis of the PLSE

In this section we give results on the asymptotic properties of $(\hat{m}, \hat{\theta})$. Note that we will study $(\hat{m}, \hat{\theta})$ for a certain (possibly data-driven) choice of λ satisfying some rate conditions; see assumption **(S2)** below. First, we need some smoothness assumption on m_0. We assume:

(S1) The unknown convex link function m_0 is bounded by some constant M_0 on D, has an absolutely continuous first derivative, and satisfies $J(m_0) < \infty$.

(S2) The smoothing parameter λ can be chosen to be a random variable. For the rest of the chapter, we denote it by $\hat{\lambda}_n$. Assume that $\hat{\lambda}_n$ satisfies the rate conditions:

$$\hat{\lambda}_n^{-1} = O_p(n^{2/5}) \qquad \text{and} \qquad \hat{\lambda}_n = o_p(n^{-1/4}). \tag{4.7}$$

Our assumption **(S1)** on m_0 is quite minimal — we essentially require m_0 to have an absolutely continuous derivative. Assumption **(S2)** allows our tuning parameter to be data dependent, as opposed to a sequence of constants. This allows for data driven

choice of $\hat{\lambda}_n$, such as those obtained from cross-validation. We will show that any choice of $\hat{\lambda}_n$ satisfying (4.7) will result in an asymptotically "efficient" estimator of θ_0. Now in a sequence of theorems (proved in Section 4.10), we study the asymptotic properties of $(\hat{m}, \hat{\theta})$. First up is the consistency and rate of convergence of $\hat{m} \circ \hat{\theta}$.

Theorem 21. *Under assumptions **(A0)**-**(A4)** and **(S1)**-**(S2)**, the PLSE satisfies*

$$J(\hat{m}) = O_p(1),\ \|\hat{m}\|_\infty = O_p(1),\ \text{and} \|\hat{m} \circ \hat{\theta} - m_0 \circ \theta_0\| = O_p(\hat{\lambda}_n).$$

Next, we prove the consistency of $\hat{m}$ and $\hat{\theta}$. We prove that $\hat{m}$ is consistent under the Sobolev norm.

Theorem 22. *Under assumptions **(A0)**-**(A4)** and **(S1)**-**(S2)**,*

$$\hat{\theta} \xrightarrow{P} \theta_0,\ \|\hat{m} - m_0\|_{D_0}^S \xrightarrow{P} 0,\ \text{and} \|\hat{m}'\|_\infty = O_p(1).$$

We now provide an upper bound on the rates of convergence of $\hat{\theta}$ and $\hat{m}$.

Theorem 23. *Under assumptions **(A0)**-**(A4)** and **(S1)**-**(S2)**, and the assumption that the conditional distribution of X given $\theta_0^\top X$ is nondegenerate, $\hat{m}$ and $\hat{\theta}$ satisfy*

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

Next we provide an upper bound on the rate of convergence of the derivative of $\hat{m}$. These upper bounds will be useful for computing the asymptotic distribution of $\hat{\theta}$ in Section 4.5.2.

Theorem 24. *Under the assumptions of Theorem 23 and **(A5)**, we have*

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

4.4.2 Asymptotic analysis of LLSE

In this subsection we present the results for the LLSE. The following assumption on m_0 is needed for $\check{m}$ to be a consistent estimator of m_0.

(L1) The unknown convex link function m_0 is bounded by some constant $M_0(\geq 1)$ on D and is uniformly Lipschitz with Lipschitz constant L_0.

Now, as in Section 4.4.1, we give a sequence of theorems (proved in Section 4.11) characterizing the asymptotic properties of $(\check{m}, \check{\theta})$. Theorem 25 below proves the consistency and provides an upper bound on the rate of convergence of $\check{m} \circ \check{\theta}$ to $m_0 \circ \theta_0$ under the $\|\cdot\|$ norm.

Theorem 25. *Let us assume that* ***(A1)-(A4)*** *and* ***(L1)*** *hold. If $L \geq L_0$, then the constrained LSE satisfies*

$$\|\check{m} \circ \check{\theta} - m_0 \circ \theta_0\| = O_p(n^{-2/5}).$$

The next theorem proves the consistency of both $\check{\theta}$ and $\check{m}$.

Theorem 26. *Under assumptions of Theorem 25, we have*

$$|\check{\theta} - \theta_0| = o_p(1), \qquad \|\check{m} - m_0\|_{D_0} = o_p(1), \quad \text{and} \quad \|\check{m}' - m_0'\|_C = o_p(1)$$

for any compact subset C in the interior of D_0.

Now, we will find upper bounds on the rate of convergence of $\check{\theta}$ and $\check{m}$.

Theorem 27. *Under assumptions of Theorem 25, the constrained LSE satisfies*

$$|\check{\theta} - \theta_0| = O_p(n^{-2/5}) \quad \text{and} \quad \|\check{m} \circ \theta_0 - m_0 \circ \theta_0\| = O_p(n^{-2/5}).$$

Under additional smoothness assumption on m_0, we show that the right derivative of $\check{m}$ converges to m_0'.

Theorem 28. *Let us assume that conditions for Theorem 25 and* ***(A5)*** *hold. If m_0 is twice continuously differentiable on D_0, then we have that*

$$\|\check{m}' \circ \theta_0 - m_0' \circ \theta_0\| = O_p(n^{-2/15}) \quad \text{and} \quad \int_{D_0} (\check{m}'(t) - m_0'(t))^2 dt = O_p(n^{-2/15}). \tag{4.8}$$

In fact,

$$\sup_{\theta \in \{\theta \in \Theta:\, |\theta_0 - \theta| \leq n^{-2/15}\}} \|\check{m}' \circ \theta - m_0' \circ \theta\| = O_p(n^{-2/15}).$$

In particular,

$$\|\check{m}' \circ \check{\theta} - m_0' \circ \check{\theta}\| = O_p(n^{-2/15}). \tag{4.9}$$

4.5 Semiparametric inference

In Sections 4.5.2 and 4.5.3, we show that both $\hat{\theta}$ and $\check{\theta}$ are asymptotically normal and semiparametrically efficient estimators of θ_0 under homoscedastic errors, respectively. Before going into the derivation of the limit law of the proposed estimators of θ_0, we need to introduce some further notation and regularity assumptions.

(B1) Assume that m_0 is three times differntiable and that m_0''' is bounded on D. Furthermore, let m_0 be strictly convex on D, i.e., for all $s \in D$ we have $m_0''(s) > \delta > 0$ for some fixed δ.

For every $\theta \in \Theta$, define $h_\theta : D \to \mathbb{R}^d$,

$$h_\theta(u) := \mathbb{E}[X|\theta^\top X = u]. \tag{4.10}$$

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

$$\|h_{\theta_1} - h_{\theta_2}\|_\infty < \bar{M}|\theta_1 - \theta_2|, \tag{4.11}$$

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

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

(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. The function h_θ plays a crucial role in the construction of "least favorable" paths; see Section 4.5.2.1. For the functions in the path to be in $\mathcal{R}$ or $\mathcal{M}_L$, we need the smoothness assumptions **(B2)** on h_θ. We need the lower and upper bound on the variance as we are using a non-weighted least squares method to estimates parameters in a (possibly) heteroscedastic model.

4.5.1 Efficient score

As in Chapter 3, the parameter space Θ is a closed subset of $\mathbb{R}^d$ and the interior of Θ in $\mathbb{R}^d$ is the null set. Thus to compute the score for the model, we construct a path on the sphere; see Section 3.4.1 for a discussion. We use $\mathbb{R}^{d-1}$ to parametrize the paths for model (4.1) on the sphere. For each $\eta \in \mathbb{R}^{d-1}$, $s \in \mathbb{R}$, and $|s| \leq |\eta|^{-1}$, define

$$\zeta_s(\theta, \eta) := \sqrt{1 - s^2|\eta|^2}\,\theta + sH_\theta\eta, \tag{4.12}$$

where H_θ is defined in Lemma 16; see discussion following Lemma 16 for further details.

In Sections 4.5.1.1 and 4.5.1.2 we attempt to calculate the efficient score for model:

$$Y = m(\theta^\top X) + \epsilon, \tag{4.13}$$

where ϵ is defined in (4.1) and satisfies **(B3)**, when $m \in \mathcal{R}$ and $m \in \mathcal{M}_L$, respectively.

4.5.1.1 Efficient score when $(m, \theta) \in \mathcal{R} \times \Theta$

The log-likelihood of the model is

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

For any $\eta \in S^{d-2}$, consider the path defined as $s \mapsto \zeta_s(\theta, \eta)$. 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

$$\left.\frac{\partial l_{\zeta_s(\theta,\eta),m}(y, x)}{\partial s}\right|_{s=0} = \eta^\top S_{\theta,m}(y, x),$$

where

$$S_{\theta,m}(y, x) := -\frac{p'_{\epsilon|X}\big(y - m(\theta^\top x), x\big)}{p_{\epsilon|X}\big(y - m(\theta^\top x), x\big)}m'(\theta^\top x)H_\theta^\top x. \tag{4.14}$$

Remark 11. *Note that under (3.10), 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.*

We now define a parametric submodel for the unknown nonparametric components

$$\begin{aligned} m_{s,a}(t) &= m(t) + sa(t), \\ p_{\epsilon|X;s,b}(e,x) &= p_{\epsilon|X}(e,x)(1+sb(e,x)), \\ p_{X;s,q}(x) &= p_X(x)(1+sq(x)), \end{aligned} \tag{4.15}$$

where $s \in \mathbb{R}$, $b : \mathbb{R} \times \mathcal{X} \to \mathbb{R}$ 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 $m_{s,a} \in \mathcal{R}$ for small enough $s \in \mathbb{R}$ and $q : \mathcal{X} \to \mathbb{R}$ is a bounded function such that $\mathbb{E}(q(X)) = 0$. Consider the following parametric submodel of (4.13),

$$s \mapsto (\zeta_s(\theta, \eta),\ m_{s,a},\ p_{\epsilon|X;s,b},\ p_{X;s,q}(x)) \tag{4.16}$$

where $\eta \in S^{d-2}$. Differentiating the log-likelihood of the submodel in (4.16) with respect to s, we get that the score along the submodel in (4.16) 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_S, of the model is

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

where for any set $A \in L_2(P_{\theta,m})$, $\overline{lin}A$ 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. Now observe that

$$\overline{lin}\{a \in \mathcal{S} : J(a) < \infty \text{ and } m_{s,a} \in \mathcal{R} \text{ for small enough } s\} \subseteq \overline{lin}\{a \in \mathcal{S} : J(a) < \infty\}$$

and

$$\overline{lin}\{q : \mathcal{X} \to \mathbb{R} |\, q \text{ is a bounded function and } \mathbb{E}(q(X) = 0\} = \{q : \mathcal{X} \to \mathbb{R} |\, q \in L_2(P_X)\}.$$

However, by Theorem A.1 of [Györfi *et al.*, 2002], we have that the class of infinitely often differentiable functions of bounded support is dense in $L_2(\mathbf{m})$, where $\mathbf{m}$ denotes the Lebesgue measure on D. Thus we have that

$$\overline{lin}\{a \in \mathcal{S} : J(a) < \infty\} = \{a : D \to \mathbb{R} |\, a \in L_2(\mathbf{m})\}$$

and $\overline{lin}\{b : \mathbb{R} \times \chi \to \mathbb{R} |\, b$ is a bounded function, $\mathbb{E}(\epsilon b(\epsilon, X)|X) = \mathbb{E}(b(\epsilon, X)|X) = 0\} = \{b \in L_2(P_{\epsilon,X}) : \mathbb{E}(\epsilon b(\epsilon, X)|X) = \mathbb{E}(b(\epsilon, X)|X) = 0\}$. Thus, it is easy to see that under assumptions **(A0)**–**(A4)**, **(S1)**, and **(B1)**–**(B3)**, the nuisance tangent space of (4.1) is

$$\Lambda_S \subseteq \Big\{ f \in L_2(P_{\epsilon,X}) : f(e,x) = \frac{p'_{e|X}(e,x)}{p_{e|X}(e,x)} a(\theta^\top x) + b(e,x) + q(x), \tag{4.17}$$
$$\text{where } a \in L_2(\mathbf{m}), b \in L_2(P_{\epsilon,X}), q \in L_2(P_X), \mathbb{E}(\epsilon b(\epsilon, X)|X) = 0,$$
$$\mathbb{E}(b(\epsilon, X)|X) = 0, \text{ and } \mathbb{E}(q(X)) = 0 \Big\} =: \Lambda_0.$$

Observe that the efficient score is the $L_2(P_{\theta,m})$ projection of $S_{\theta,m}(y,x)$ onto $\Lambda_S^\perp$, where $\Lambda_S^\perp$ is the orthogonal complement of Λ_S in $L^2(P_{\theta,m})$. [Newey and Stoker, 1993] and [Ma and Zhu, 2013a] show that

$$\Lambda_0^\perp = \Big\{ f \in L_2(P_{\epsilon,X}) : f(e,x) = \big[g(x) - \mathbb{E}\big(g(X)|\theta^\top X = \theta^\top x\big)\big] e, \tag{4.18}$$
$$\text{for some } g : \chi \to \mathbb{R} \Big\} \subseteq \Lambda_S^\perp$$

Using calculations similar to those in Proposition 1 of [Ma and Zhu, 2013a], it can be shown that

$$\Pi(S_{\theta,m}(y,x)|\Lambda_0^\perp) = \frac{1}{\sigma^2(x)}(y - m(\theta^\top x)) m'(\theta^\top x) H_\theta^\top \left\{ x - \frac{\mathbb{E}(\sigma^{-2}(X)X|\theta^\top X = \theta^\top x)}{\mathbb{E}(\sigma^{-2}(X)|\theta^\top X = \theta^\top x)} \right\},$$

where for any $f : \mathbb{R} \times \chi \to \mathbb{R}$ such that $f \in L_2(P_{\theta,m})$, $\Pi(f|\Lambda_0^\perp)$ denotes the $L_2(P_{\theta,m})$ projection of f onto the space $\Lambda_0^\perp$.

To compute the efficient score, we need to evaluate $\Pi(S_{\theta,m}(y,x)|\Lambda_S^\perp)$, which is hard due to the complicated nature of the set of parametric submodels of m. Note that the efficient information is denoted by

$$\mathcal{I}_{\theta,m}^S := P_{\theta,m}\big[\Pi(S_{\theta,m}(Y,X)|\Lambda_S^\perp)\Pi^\top(S_{\theta,m}(Y,X)|\Lambda_S^\perp)\big].$$

By definition of $\Lambda_0^\top$ (see (4.18), we have

$$\mathcal{I}^S_{\theta,m} \geq P_{\theta,m}\left[\Pi(S_{\theta,m}(Y,X)|\Lambda_0^\perp)\Pi^\top(S_{\theta,m}(Y,X)|\Lambda_0^\perp)\right] =: \mathcal{I}^0_{\theta,m}.$$

In Section 4.5.2 we show that $\mathcal{I}^S_{\theta,m}$ is the efficient information for the model (4.1), when $(m,\theta) \in \mathcal{R}\times\Theta$. Moreover, we construct a path whose score at (m_0,θ_0) is "approximately" (see Section 4.5.2 for a technical definition) equal to $\Pi(S_{\theta_0,m_0}(y,x)|\Lambda_0^\perp)$.

It is important to note that the optimal estimating equation depends on $\sigma^2(x)$. Since in the semiparametric model $\sigma^2(\cdot)$ is left unspecified, it is unknown. Without additional assumptions, estimators of $\sigma^2(\cdot)$ have slow rate of convergence to $\sigma^2(\cdot)$, especially if d is large. Thus if we substitute $\hat{\sigma}(\cdot)$ 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, we will assume that $\sigma^2(\cdot) \equiv \sigma^2$. Under this assumption, we have

$$\Pi(S_{\theta,m}(y,x)|\Lambda_0^\perp) = \frac{1}{\sigma^2}(y - m(\theta^\top x))m'(\theta^\top x)H_\theta^\top\left\{x - h_\theta(\theta^\top x)\right\},$$

where $h_\theta(\theta^\top x)$ is defined in (4.10). 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.,

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

and class of functions formed by the efficient score indexed by (θ,m) in a "neighborhood" of (θ_0,m_0) satisfies some technical conditions. We formalize these in Section 4.5.2.

4.5.1.2 Efficient score when $(m,\theta) \in \mathcal{M}_L \times \Theta$

In Remark 12 below we show that (4.13) is differentiable in quadratic mean in θ, when $m \in \mathcal{M}_L$. Thus the parametric score in this model satisfies (4.14), where m' denotes the right derivative of m; see Section 4.2. Moreover, using parametric submodel as in (4.15) and (4.16) and calculations similar to those in Section 4.5.1.1, it can be shown that the

nuisance tangent space, denoted by Λ_L, of the model is

$$\Lambda_L := \overline{lin}\Big\{ f(e,x) \in L_2(P_{\theta,m}) : f(e,x) = \frac{p'_{e|X}(e,x)}{p_{e|X}(e,x)} a(\theta^\top x) + b(e,x) + q(x),$$
$$\text{where } a \in L^2(\mathbf{m}), m_{s,a} \in \mathcal{M}_L \text{ for small enough } s, b : \mathbb{R} \times \mathcal{X} \to R$$
$$\text{and } q : \mathcal{X} \to \mathbb{R} \text{ are bounded functions}, \mathbb{E}(\epsilon b(\epsilon, X)|X) = 0,$$
$$\mathbb{E}(b(\epsilon, X)|X) = 0, \text{ and } \mathbb{E}(q(X)) = 0 \Big\}.$$

Now using arguments similar to those in Section 4.5.1.1, it can be shown that

$$\mathcal{I}^S_{\theta,m} \geq P_{\theta,m}\big[\Pi(S_{\theta,m}(Y,X)|\Lambda_0^\perp)\Pi^\top(S_{\theta,m}(Y,X)|\Lambda_0^\perp)\big] = \mathcal{I}^0_{\theta,m},$$

where

$$\mathcal{I}^L_{\theta,m} := P_{\theta,m}\big[\Pi(S_{\theta,m}(Y,X)|\Lambda_L^\perp)\Pi^\top(S_{\theta,m}(Y,X)|\Lambda_L^\perp)\big]$$

and $S_{\theta,m}(Y,X)$ and Λ_0 are defined as in (4.14) and (4.17), respectively.

Remark 12. *If the errors are Gaussian random variables then in the following, we show that the model is quadratic mean differentiable in θ. The proof of quadratic mean differentiability for any error distribution satisfying assumption **(B3)** follows similarly. Under Gaussian error, the density of (Y,X) is*

$$f_{\theta,m}(y,x) = \exp\Big(-\frac{1}{2\sigma^2}(y - m(\theta^\top x))^2 \Big).$$

Note that $f_{\theta,m}(y,x)$ is differentiable (with respect to θ) a.e. (y,x). Define

$$\eta(y,x,\theta,m) = \begin{cases} \frac{f'_{\theta,m}(y,x)}{2f^{1/2}_{\theta,m}(y,x)}, & f_{\theta,m}(y,x) > 0 \text{ and } f'_{\theta,m}(y,x) \text{ exists}, \\ 0 & \text{otherwise}, \end{cases}$$

where $f'_{\theta,m}(y,x)$ denotes the derivative with respect to θ. [Hájek, 1972] proves that the family of distributions is quadratic mean diffrentiable (q.m.d) at θ_0 if

$$I_{i,j}(\theta) := \int \eta_i(y,x,\theta,m)\eta_j(y,x,\theta,m) dP_X(x) dy$$

is finite and continuous at θ_0. In the following we prove that $I_{i,j}(\theta)$ is finite and continuous at θ_0. Observe that,

$$\begin{aligned} I_{i,j}(\theta) &= \int_{\chi\times\mathbb{R}} \eta_i(y,x,\theta,m)\eta_j(y,x,\theta,m)dP_X dy \\ &= \int_{\chi\times\mathbb{R}} (y-m(\theta^\top x))^2[m'(\theta^\top x)]^2 x_i x_j \exp\Big(-\frac{1}{2}(y-m(\theta^\top x))^2\Big)dP_X(x)dy \\ &= E_{\theta,m}[(Y-m(\theta^\top X))^2[m'(\theta^\top X)]^2 X_i X_j] \\ &= E_{\theta,m}\big[E[(Y-m(\theta^\top X))^2[m'(\theta^\top X)]^2 X_i X_j|\theta^\top X]\big] \\ &= E_{\theta,m}\big[m'(\theta^\top X)^2 E[X_i X_j|\theta^\top X]\big]. \end{aligned}$$

As both $m(\cdot)$ and $E[X_iX_j|\theta^\top X=\cdot]$ are bounded functions, we have that $I_{i,j}(\theta)$ is finite and continuous at θ_0. Thus, the model is differentiable in quadratic mean in θ.

Remark 13. *Assumptions **(A0)**-**(A4)** and **(S1)** (or **(L1)**) do not guarantee the existence of a least favorable submodel for the model in (4.1), which can be the case when the estimators lie on the "boundary" of the parameter set. [van der Vaart, 2002] introduced the notion of approximately least favorable subprovided model to get around this difficulty. Under the additional assumptions **(B1)**-**(B3)**, we find the approximately least favorable subprovided model and show that $\Pi(S_{\theta_0,m_0}(x,y)|\Lambda_0^\perp)$ is the efficient score at (θ_0,m_0); see Section 4.5.2.1. However, the score corresponding to the approximately least favorable subprovided model does not satisfy the conditions required in [van der Vaart, 2002] for asymptotic normality and efficiency of the finite dimensional parameter in semiparametric models. Thus, we find a well-behaved approximation to the score such that $(\hat\theta,\hat m)$ is an approximate zero of the corresponding estimating equation.*

4.5.2 Efficiency of the PLSE

Theorem 29. *Assume (X,Y) satisfies (4.1) and assumptions **(A0)**–**(A4)**, **(B1)**–**(B2)**, and **(S1)** hold. Define,*

$$\ell_{\theta,m} := \big(y-m(\theta^\top x)\big)m'(\theta^\top x)H_\theta^\top\Big\{x-h_\theta(\theta^\top x)\Big\}.$$

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

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0, H_{\theta_0} V_{\theta_0,m_0}^{-1} I_{\theta_0,m_0} (H_{\theta_0} V_{\theta_0,m_0}^{-1})^\top),$$

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

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

Proof. We now give a sketch of the proof. Some of the steps are proved in the following sections.

Step 1 In Theorem 30 we find *approximately least favorable subprovided models* (see Definition 9.7 of [van der Vaart, 2002]) with score

$$\mathfrak{S}_{\theta,m}(x,y) = \{y - m(\theta^\top x)\} H_\theta^\top \Big[m'(\theta^\top x) x + \int_{s_0}^{\theta^\top x} m'(u) k'(u) du - m'(\theta^\top x) k(\theta^\top x) + m_0'(s_0) k(s_0) - m_0'(s_0) h_{\theta_0}(s_0) \Big] \tag{4.19}$$

where $k : D \to \mathbb{R}^d$ is defined as

$$k(u) := h_{\theta_0}(u) + \frac{m_0'(u)}{m_0''(u)} h_{\theta_0}'(u). \tag{4.20}$$

We prove that there exists a constant $M^* < \infty$ such that

$$\sup_{u \in D} (|k(u)| + |k'(u)|) \le M^*. \tag{4.21}$$

Moreover, $(\hat{\theta}, \hat{m})$ satisfies the score equation approximately, i.e.,

$$\sqrt{n} \mathbb{P}_n \mathfrak{S}_{\hat{\theta},\hat{m}} = o_p(1). \tag{4.22}$$

Furthermore, define $\psi_{\theta,m} : \mathcal{X} \times \mathbb{R} \to \mathbb{R}^{d-1}$ as

$$\psi_{\theta,m}(x,y) := (y - m(\theta^\top x)) H_\theta^\top [m'(\theta^\top x) x - h_{\theta_0}(\theta^\top x) m_0'(\theta^\top x)]. \tag{4.23}$$

Step 2 In Theorem 31 we show that $\psi_{\hat{\theta},\hat{m}}$ is an empirical approximation of the score $\mathfrak{S}_{\hat{\theta},\hat{m}}$, i.e.,

$$\sqrt{n}\mathbb{P}_n(\mathfrak{S}_{\hat{\theta},\hat{m}} - \psi_{\hat{\theta},\hat{m}}) = o_p(1).$$

Thus in view of (4.22) we have that $\hat{\theta}$ is an *approximate* zero of the function $\theta \mapsto \mathbb{P}_n\psi_{\theta,\hat{m}}$, i.e.,

$$\sqrt{n}\mathbb{P}_n\psi_{\hat{\theta},\hat{m}} = o_p(1). \tag{4.24}$$

Step 3 In Theorem 32 we show that $\psi_{\hat{\theta},\hat{m}}$ is approximately unbiased in the sense of [van der Vaart, 2002], i.e.,

$$\sqrt{n}P_{\hat{\theta},m_0}\psi_{\hat{\theta},\hat{m}} = o_p(1). \tag{4.25}$$

Step 4 We prove

$$\mathbb{G}_n(\psi_{\hat{\theta},\hat{m}} - \psi_{\theta_0,m_0}) = o_p(1) \tag{4.26}$$

in Theorem 33. Furthermore, as $\psi_{\theta_0,m_0} = \ell_{\theta_0,m_0}$, we have

$$P_{\theta_0,m_0}[\psi_{\theta_0,m_0}] = 0.$$

Thus, by (4.24) and (4.25), we have that (4.26) is equivalent to

$$\sqrt{n}(P_{\hat{\theta},m_0} - P_{\theta_0,m_0})\psi_{\hat{\theta},\hat{m}} = \mathbb{G}_n\ell_{\theta_0,m_0} + o_p(1). \tag{4.27}$$

Step 5 To complete the proof, it is now enough to show that

$$\sqrt{n}(P_{\hat{\theta},m_0} - P_{\theta_0,m_0})\psi_{\hat{\theta},\hat{m}} = \sqrt{n}V_{\theta_0,m_0}H_{\theta_0}^\top(\hat{\theta} - \theta_0) + o_p(\sqrt{n}|\hat{\theta} - \theta_0|). \tag{4.28}$$

Observe that (4.27) and (4.28) imply

$$\begin{aligned}\sqrt{n}V_{\theta_0,m_0}H_{\theta_0}^\top(\hat{\theta} - \theta_0) &= \mathbb{G}_n\ell_{\theta_0,m_0} + o_p(1 + \sqrt{n}|\hat{\theta} - \theta_0|),\\ \Rightarrow \sqrt{n}H_{\theta_0}^\top(\hat{\theta} - \theta_0) &= V_{\theta_0,m_0}^{-1}\mathbb{G}_n\ell_{\theta_0,m_0} + o_p(1) \xrightarrow{d} V_{\theta_0,m_0}^{-1}N(0, I_{\theta_0,m_0}).\end{aligned}$$

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

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

the proof of which can be found in Section 3.4.2.1. The proof of **Step 5** can be found in proof of Theorem 6.20 of [van der Vaart, 2002]; also see Section 3.8.3.

□

4.5.2.1 An approximately least favorable subprovided path

We now construct a path whose score for any $(\theta, m) \in \Theta \times \{g \in \mathcal{R} | \, J(g) < \infty\}$ is $\mathfrak{S}_{\theta,m}$. Recall (4.12). For any $(\theta, m) \in \Theta \times \{g \in \mathcal{R} | \, J(g) < \infty\}$, let $t \mapsto (\zeta_t(\theta, \eta), \xi_t(\cdot; \theta, \eta, m))$ denote a path in $\Theta \times \{m \in \mathcal{R} | \, J(m) < \infty\}$ that through (θ, m), i.e., $(\zeta_0(\theta, \eta), \xi_0(\cdot; \theta, \eta, m)) = (\theta, m)$. 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

$$\frac{\partial}{\partial t} \mathcal{L}_n(\xi_t(\cdot; \hat{\theta}, \eta, \hat{m}), \zeta_t(\hat{\theta}, \eta), \hat{\lambda}_n)\Big|_{t=0} = 0. \tag{4.29}$$

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

$$\begin{aligned} \frac{\partial}{\partial t} \big(y - \xi_t(\zeta_t(\hat{\theta}, \eta)^\top x; \hat{\theta}, \eta, \hat{m})\big)^2\Big|_{t=0} &= \eta^\top \mathfrak{S}_{\hat{\theta},\hat{m}}(x, y), \\ \hat{\lambda}_n^2 \, \frac{\partial}{\partial t} J^2(\xi_t(\cdot; \hat{\theta}, \eta, \hat{m}))\Big|_{t=0} &= O_p(1), \end{aligned} \tag{4.30}$$

for all $\eta \in S^{d-2}$, then we get (4.22) as $\hat{\lambda}_n^2 = o_p(n^{-1/2})$; see assumption **(S2)**.

In the following we construct a path for (4.13) that satisfies the above requirements. For any set $A \in \mathbb{R}$ and any $\nu > 0$, let us define $A^\nu := \cup_{a \in A} B_a(\nu)$. Fix $\nu > 0$. By assumption **(A5)**, for every $\theta \in \Theta$, $\eta \in S^{d-2}$, 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

$$\begin{aligned} \phi_{\theta,\eta,t}(u) &= u, \quad u \in D_\theta, \\ \phi_{\theta,\eta,t}(u + (\theta - \zeta_t(\theta, \eta))^\top k(u)) &= u, \quad u \in \partial D, \end{aligned} \tag{4.31}$$

where $k(u)$ and $\zeta_t(\theta, \eta)$ are defined in (4.20) and (4.12), respectively. Furthermore, we can ensure that $u \in D \mapsto \phi_{\theta,\eta,t}(u)$ is infinitely differentiable and that $\frac{\partial}{\partial t} \phi_{\theta,\eta,t}\big|_{t=0}$ exists. Note that $\phi_{\theta,\eta,t}(D) = D$. Moreover, $u \mapsto \phi_{\theta,\eta,t}(u)$ cannot be the identity function for $t \neq 0$ if $(\theta - \zeta_t(\theta, \eta))^\top h_\theta(u) \neq 0$ for $u \in \partial D$. Let us now define

$$\urcorner_t(u; \theta, \eta, m) := m' \circ \phi_{\theta,\eta,t}(u + (\theta - \zeta_t(\theta, \eta))^\top k(u)).$$

Now, we can define the submodel

$$\xi_t(u;\theta,\eta,m) := \int_{s_0}^{u} \urcorner_t(y;\theta,\eta,m)dy + (\zeta_t(\theta,\eta)-\theta)^\top \left[(m_0'(s_0)-m'(s_0))k(s_0) - m_0'(s_0)h_{\theta_0}(s_0)\right] + m(s_0), \tag{4.32}$$

where h_{θ_0} is defined in (4.10), k is defined in (4.20), and $s_0 \in \bigcap_{\theta\in\{\eta\in\Theta:|\eta-\theta_0|\le r\}} D_\theta$ with r defined in **(A5)**. The function $\phi_{\theta,\eta,t}$ helps us control the partial derivative in the second equation of (4.30). In the following theorem, proved in Section 4.12.1, 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 (4.29) and (4.30). Here η is the "direction" for $\zeta_t(\theta,\eta)$ and $(\eta, h_\theta(u))$ defines the "direction" for the path $\xi_t(\cdot;\theta,\eta,m)$.

Theorem 30. *[**Step 1**] Under the assumptions of Theorem 29, $(\zeta_t(\hat\theta,\eta), \xi_t(\cdot;\hat\theta,\eta,\hat m))$ is a valid parametric submodel, i.e., $(\zeta_t(\hat\theta,\eta), \xi_t(\cdot;\hat\theta,\eta,\hat m)) \in \Theta \times \{g \in \mathcal{R} | J(g) < \infty\}$ for all t in some neighborhood of 0 and $\mathfrak{S}_{\theta_0,m_0} = \ell_{\theta_0,m_0}$. Moreover, $(\zeta_t(\hat\theta,\eta), \xi_t(\cdot;\hat\theta,\eta,\hat m))$ satisfies (4.30), $\mathcal{L}_n(\xi_t(\cdot;\hat\theta,\eta,\hat m), \zeta_t(\hat\theta,\eta), \hat\lambda_n)$ as function of t is differentiable at 0, and $\mathfrak{S}_{\hat\theta,\hat m}$ satisfies (4.22) and there exists $M^* < \infty$ which satisfies (4.21).*

4.5.2.2 A well behaved approximation

We observe that $\mathfrak{S}_{\theta,m}$ (the score for the approximately least favorable submodel) does not satisfy the conditions required by [van der Vaart, 2002]. In this section we introduce $\psi_{\theta,m}$, a well behaved "approximation" of $\mathfrak{S}_{\theta,m}$. $\psi_{\theta,m}$ is not a score of the (4.13) for any particular path. However, $\psi_{\theta,m}$ is well-behaved in the sense that: (1) $\psi_{\hat\theta,\hat m}$ belongs to a Donsker class of functions (see (4.26)), (2) $\psi_{\theta_0,m_0} = \ell_{\theta_0,m_0}$, (3) $\psi_{\hat\theta,\hat m}$ converges to ψ_{θ_0,m_0} in the $L_2(P_{\theta_0,m_0})$ norm (see Lemma (57)). The following theorem proves that $\mathfrak{S}_{\hat\theta,\hat m}$ and $\psi_{\hat\theta,\hat m}$ are "approximately" the same.

Theorem 31. *[**Step 2**] Under model (4.1) and assumptions **(A0)–(A4)**, **(B1)–(B3)**, and **(S1)**, we have*

$$\sqrt{n}\mathbb{P}_n(\mathfrak{S}_{\hat\theta,\hat m} - \psi_{\hat\theta,\hat m}) = o_p(1). \tag{4.33}$$

We break the proof of this theorem into a number of lemmas proved in Section 4.12. In the following lemma, proved in Section 4.12.2, we find an upper bound for the left hand side of (4.33).

Lemma 32. *Under model (4.1), we have*

$$\begin{aligned}|\sqrt{n}\mathbb{P}_n(\mathfrak{S}_{\hat\theta,\hat m}-\psi_{\hat\theta,\hat m})| \le\ & |\mathbb{G}_n\big[(m_0\circ\theta_0-\hat m\circ\theta_0)U_{\hat\theta,\hat m}\big]| \\ & + |\mathbb{G}_n[(\hat m\circ\theta_0-\hat m\circ\hat\theta)U_{\hat\theta,\hat m}| + |\sqrt{n}\mathbb{P}_n\epsilon U_{\hat\theta,\hat m}| \\ & + \sqrt{n}\big|P_{\theta_0,m_0}[(\hat m\circ\theta_0-\hat m\circ\hat\theta)U_{\hat\theta,\hat m}]\big| \\ & + \sqrt{n}\big|P_{\theta_0,m_0}\big[(m_0\circ\theta_0-\hat m\circ\theta_0)U_{\hat\theta,\hat m}\big]\big|, \end{aligned} \tag{4.34}$$

where $U_{\theta,m}:\chi\to\mathbb{R}^{d-1}$ is defined as

$$U_{\theta,m}(x) := H_\theta^\top\left[\int_{s_0}^{\theta^\top x}\big[m'(u)-m_0'(u)\big]k'(u)du + (m_0{}'(\theta^\top x)-m'(\theta^\top x))k(\theta^\top x)\right]. \tag{4.35}$$

Note that the proof of Theorem 31 will be complete if we show that each of the terms on the right hand side of (4.34) converges 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 $a_n\|\hat m-m_0\|^S_{D_{\theta_0}}=o_p(1)$. Note that we can always find such a sequence a_n, as by Theorem 22 we have $\|\hat m-m_0\|^S_{D_{\theta_0}}=o_p(1)$. For all $n\in\mathbb{N}$, define[2]

$$\begin{aligned}\mathcal{C}^{m*}_{M_1,M_2,M_3} &:= \Big\{m\in\mathcal{R},\ \|m\|_\infty<M_1,\ \|m'\|_\infty<M_2,\ \text{and } J(m)<M_3\Big\},\\ \mathcal{C}^{m}_{M_1,M_2,M_3}(n) &:= \Big\{m\in\mathcal{C}^{m*}_{M_1,M_2,M_3}: a_n\|m-m_0\|^S_{D_{\theta_0}}\le 1\Big\},\\ \mathcal{C}^{*}_{M_1,M_2,M_3} &:= \Big\{(\theta,m):\theta\in\Theta\cap B_{\theta_0}(1/2)\text{ and } m\in\mathcal{C}^{m*}_{M_1,M_2,M_3}\Big\},\\ \mathcal{C}^{\theta}(n) &:= \Big\{\theta\in\Theta\cap B_{\theta_0}(1/2):\hat\lambda_n^{-1/2}|\theta-\theta_0|\le 1\Big\},\\ \mathcal{C}_{M_1,M_2,M_3}(n) &:= \Big\{(\theta,m):\theta\in\mathcal{C}^\theta(n)\text{ and } m\in\mathcal{C}^m_{M_1,M_2,M_3}(n)\Big\},\\ \mathcal{W}^*_{M_1,M_2,M_3} &:= \big\{U_{\theta,m}:(\theta,m)\in\mathcal{C}^*_{M_1,M_2,M_3}\big\},\\ \mathcal{W}_{M_1,M_2,M_3}(n) &:= \big\{U_{\theta,m}:(\theta,m)\in\mathcal{C}_{M_1,M_2,M_3}(n)\big\}. \end{aligned} \tag{4.36}$$

As a first step in proving that each term on the right hand side of (4.34) converges to 0, we try to understand the classes of functions $\mathcal{W}_{M_1,M_2,M_3}(n)$ and $\mathcal{W}^*_{M_1,M_2,M_3}$. In the following lemma, proved later in Section 4.12.3, we find the bracketing numbers and

[2]The notations with $*$ denote the classes that do not depend on n while the ones with n denote shrinking neighborhoods around the truth.

envelope functions for the classes. This lemma will be used in some of the remaining proofs.

Lemma 33. *Fix M_1, M_2, M_3, and $\delta > 0$. Then $\mathcal{W}_{M_1,M_2,M_3}(n)$ is a Donsker class and*

$$\sup_{(\theta,m)\in\mathcal{C}_{M_1,M_2,M_3}(n)} \|U_{\theta,m}\|_{2,\infty} \le W_{M_1,M_2,M_3}(n) := M^*\sqrt{d-1}\big(2(M_3+M_2)T\hat{\lambda}_n^{1/4} + (T+1)\frac{1}{a_n}\big), \tag{4.37}$$

where M^* is defined in (4.21). Moreover, for some c depending only on d, M_1, M_2, and M_3, we have the following upper bound on the bracketing entropy of $\mathcal{W}_{M_1,M_2,M_3}(n)$:

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

see Section 2.1.1 of [van der Vaart and Wellner, 1996] for a definition of $N_{[]}(\cdot,\cdot,\cdot)$.

The study of limiting behaviors of the first three terms on the right hand side of (4.34) are similar. For every fixed M_1, M_2, and M_3 the first term in the right hand side of (4.34) can be bounded above as

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

By Theorem 22 we have that $\hat{\theta}$ and $\hat{m}$ are consistent in the Euclidean and Sobolev norms, respectively and $\|\hat{m}'\|_\infty$ is $O_p(1)$. Furthermore, by Theorem 21, we have that both $\|\hat{m}\|_\infty$ and $J(\hat{m})$ are $O_p(1)$ and by Theorem 23 we have $\hat{\lambda}_n^{-1/2}|\hat{\theta}-\theta_0| = o_p(1)$. Thus, it is easy to see that, for any $\varepsilon > 0$, there exists M_1, M_2, and M_3, (depending on ε) such that

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

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

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

for sufficiently large n. We prove this in Lemma 34.

Lemma 34. *Fix M_1, M_2, M_3, and $\delta > 0$. For $n \in \mathbb{N}$, let us define*

$$\mathcal{D}^*_{M_1,M_2,M_3} := \left\{[m_0 \circ \theta_0 - m \circ \theta_0]U_{\theta,m} : (\theta, m) \in \mathcal{C}^*_{M_1,M_2,M_3}\right\},$$
$$\mathcal{D}_{M_1,M_2,M_3}(n) := \left\{[m_0 \circ \theta_0 - m \circ \theta_0]U_{\theta,m} : (\theta, m) \in \mathcal{C}_{M_1,M_2,M_3}(n)\right\}.$$

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

$$\sup_{f \in \mathcal{D}_{M_1,M_2,M_3}(n)} \|f\|_{2,\infty} \leq D_{M_1,M_2,M_3}(n) := 2M_1 W_{M_1,M_2,M_3}(n). \tag{4.38}$$

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 $\mathcal{F}$, $J_{[]}$ (the entropy integral) is defined as

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

e.g., see [van der Vaart, 1998b]. Hence, as $n \to \infty$, we have

$$\mathbb{P}\left(\sup_{f \in \mathcal{D}_{M_1,M_2,M_3}(n)} |\mathbb{G}_n f| > \delta\right) = o(1).$$

The following two lemmas, proved in Sections 4.12.5 and 4.12.6, complete the proof of Theorem 31 and show that the last four terms on right side of (4.34) converge to zero in probability.

Lemma 35. *Fix M_1, M_2, M_3, and $\delta > 0$. For $n \in \mathbb{N}$, let us define*

$$\mathcal{A}_{M_1,M_2,M_3}(n) := \left\{[m \circ \theta_0 - m \circ \theta]U_{\theta,m} : (\theta, m) \in \mathcal{C}_{M_1,M_2,M_3}(n)\right\},$$
$$\mathcal{A}^*_{M_1,M_2,M_3} := \left\{[m \circ \theta_0 - m \circ \theta]U_{\theta,m} : (\theta, m) \in \mathcal{C}^*_{M_1,M_2,M_3}\right\}.$$

Then $\mathcal{A}_{M_1,M_2,M_3}(n)$ is Donsker class and $\sup_{f \in \mathcal{A}_{M_1,M_2,M_3}(n)} \|f\|_{2,\infty} \leq D_{M_1,M_2,M_3}(n)$. Moreover,

$$J_{[]}(\gamma, \mathcal{A}_{M_1,M_2,M_3}(n), \|\cdot\|_{2,P_{\theta_0,m_0}}) \lesssim \gamma^{1/2}.$$

Hence, as $n \to \infty$, we have

$$\mathbb{P}\left(|\mathbb{G}_n[(\hat{m} \circ \theta_0 - \hat{m} \circ \hat{\theta})U_{\hat{\theta},\hat{m}}]| > \delta\right) = o_p(1).$$

Lemma 36. *If* ***(A0)–(A4)***, ***(B1)–(B3)***, *and* ***(S1)*** *hold, then*

$$\begin{aligned} |\sqrt{n}\mathbb{P}_n[\epsilon U_{\hat{\theta},\hat{m}}]| &= o_p(1) \\ \sqrt{n}\big|P_{\theta_0,m_0}\big[(m_0\circ\theta_0 - \hat{m}\circ\theta_0)U_{\hat{\theta},\hat{m}}\big]\big| &= o_p(1) \\ \sqrt{n}\big|P_{\theta_0,m_0}[(\hat{m}\circ\theta_0 - \hat{m}\circ\hat{\theta})U_{\hat{\theta},\hat{m}}]\big| &= o_p(1) \end{aligned} \tag{4.39}$$

Now that we have shown that $(\hat{\theta},\hat{m})$ is an approximate zero of $(\theta,m)\mapsto\mathbb{P}_n\psi_{\theta,m}$ and $\psi_{\theta_0,m_0}=\ell_{\theta_0,m_0}$, asymptotic normality and efficiency of $\hat{\theta}$ now follows from the theory developed in Section 6.6 of [van der Vaart, 2002]. In the next theorem, we prove that $\psi_{\hat{\theta},\hat{m}}$ satisfies the "no-bias" (see equation 6.6 of [van der Vaart, 2002]) condition.

Theorem 32. *[**Step 3**] Under assumptions* ***(A0)–(A4)*** *and* ***(B2)***,

$$\sqrt{n}P_{\hat{\theta},m_0}\psi_{\hat{\theta},\hat{m}} = o_p(1),$$

In Lemma 57, stated and proved in Section 4.12.8, we prove that $\psi_{\hat{\theta},\hat{m}}$ is a consistent estimator of ψ_{θ_0,m_0} under $L_2(P_{\theta_0,m_0})$ norm. The following theorem completes the proof of Theorem 29.

Theorem 33. *[**Step 4**] Under* ***(A0)–(A4)*** *and* ***(B2)***, *we have*

$$\mathbb{G}_n(\psi_{\hat{\theta},\hat{m}} - \psi_{\theta_0,m_0}) = o_p(1). \tag{4.40}$$

The proof of the above theorem is similar to that of Theorem 31. We first find an upper bound for the left side of (4.40) and then show that each of the terms converge to zero; see Lemmas 59 and 58 in Section 4.12.9.

4.5.3 Efficiency of the LLSE

In this section we show that $\breve{\theta}$ is an asymptotically normal efficient estimator of θ_0. The following theorem is similar to Theorem 29.

Theorem 34. *Assume* (X,Y) *satisfies* (4.1) *and assumptions* ***(A0)–(A4)***, ***(B1)–(B2)***, *and* ***(S1)*** *hold. Let* $\ell_{\theta,m}, V_{\theta_0,m_0}$, *and* I_{θ_0,m_0} *be as defined in Theorem 29. If* V_{θ_0,m_0} *is a*

nonsingular matrix in $\mathbb{R}^{(d-1)\times(d-1)}$*, then*

$$\sqrt{n}(\check{\theta}-\theta_0) \xrightarrow{d} N(0, H_{\theta_0} V_{\theta_0,m_0}^{-1} I_{\theta_0,m_0} (H_{\theta_0} V_{\theta_0,m_0}^{-1})^\top).$$

If we further assume that $\sigma^2(X) \equiv \sigma^2$ *and if the efficient information matrix* (I_{θ_0,m_0}) *is nonsingular, then* $\hat{\theta}$ *is an efficient estimator of* θ_0*, i.e.,*

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

In a series of results, we prove Theorem 34 by showing that $(\check{\theta}, \check{m})$ satisfy the conditions in **Step 1**–**Step 5** of Theorem 29. The following theorem (proved in Section 4.13.1) shows that submodel defined in (4.32) is an approximately least favorable subprovided submodel for the model (4.1).

Theorem 35. *[**Step 1**] Under assumptions of Theorem 34,* $(\zeta_t(\check{\theta},\eta), \xi_t(\cdot;\check{\theta},\eta,\check{m}))$ *is a valid parametric submodel, i.e.,* $(\zeta_t(\check{\theta},\eta), \xi_t(\cdot;\check{\theta},\eta,\check{m})) \in \Theta \times \mathcal{M}_L$ *for all* t *in some neighborhood of* 0 *and* $\mathfrak{S}_{\theta_0,m_0} = \ell_{\theta_0,m_0}$*; see (4.19) for definition of* $\mathfrak{S}_{\theta_0,m_0}$*. Moreover, we have that* $t \mapsto Q_n(\xi_t(\cdot;\check{\theta},\eta,\check{m}), \zeta_t(\check{\theta},\eta), \check{\lambda}_n)$ *is differentiable at 0,*

$$\frac{\partial}{\partial t}\big(y - \xi_t(\zeta_t(\check{\theta},\eta)^\top x;\check{\theta},\eta,\check{m})\big)^2\Big|_{t=0} = \eta^\top \mathfrak{S}_{\check{\theta},\check{m}}(x,y),$$

and

$$\frac{\partial}{\partial t} Q_n(\xi_t(\cdot;\check{\theta},\eta,\check{m}), \zeta_t(\check{\theta},\eta), \check{\lambda}_n) = \mathbb{P}_n \mathfrak{S}_{\check{\theta},\check{m}} = 0.$$

4.5.3.1 A well behaved approximation

As in Section 4.5.2.2, the following theorem (proved in a series of results) shows that the $\mathfrak{S}_{\check{\theta},\check{m}}$ is empirically well approximated by $\psi_{\check{\theta},\check{m}}$; see (4.23) for definition of $\psi_{\check{\theta},\check{m}}$.

Theorem 36. *[**Step 2**] Under assumptions of Theorem 34, we have*

$$\sqrt{n}\mathbb{P}_n(\mathfrak{S}_{\check{\theta},\check{m}} - \psi_{\check{\theta},\check{m}}) = o_p(1).$$

The proof of Theorem 36 is very similar to the proof of 31. As definitions of $\mathfrak{S}_{\theta,m}$ and $\psi_{\theta,m}$ have not changed, Lemma 32 clearly holds with $(\check{\theta}, \check{m})$ instead of $(\hat{\theta}, \hat{m})$. Note that the proof of Theorem 31 will be complete if we show that each of the terms on the right hand side of (4.34) converges to 0 in probability. We begin with some definitions. Let b_n be a sequence of real numbers such that $b_n \to \infty$ as $n \to \infty$, $b_n = o(n^{1/2})$, and $b_n \|\check{m} - m_0\|_{D_0} = o_p(1)$. Note that we can always find such a sequence b_n, as by Theorem 26 we have $\|\hat{m} - m_0\|_{D_0}^S = o_p(1)$. For all $n \in \mathbb{N}$, define

$$
\begin{aligned}
\mathcal{C}_{M_1}^{m*} &:= \Big\{ m \in \mathcal{M}_L : \|m\|_\infty < M_1 \Big\}, \\
\mathcal{C}_{M_1}^{m}(n) &:= \Big\{ m \in \mathcal{C}_{M_1}^{m*} : n^{1/5} \int_{D_0} (m'(t) - m_0'(t))^2 dt \le 1, b_n \|m - m_0\|_{D_0} \le 1 \Big\}, \\
\mathcal{C}_{M_1}^{*} &:= \Big\{ (\theta, m) : \theta \in \Theta \cap B_{\theta_0}(1/2) \text{ and } m \in \mathcal{C}_{M_1}^{m*} \Big\}, \\
\mathcal{C}^{\theta}(n) &:= \Big\{ \theta \in \theta \in \Theta \cap B_{\theta_0}(1/2) : n^{1/10} |\theta - \theta_0| \le 1 \Big\}, \\
\mathcal{C}_{M_1}(n) &:= \Big\{ (\theta, m) : \theta \in \mathcal{C}^{\theta}(n) \text{ and } m \in \mathcal{C}_{M_1}^{m}(n) \Big\}, \\
\mathcal{W}_{M_1}^{*} &:= \left\{ U_{\theta,m} : (\theta, m) \in \mathcal{C}_{M_1}^{*} \right\}, \\
\mathcal{W}_{M_1}(n) &:= \left\{ U_{\theta,m} : (\theta, m) \in \mathcal{C}_{M_1}(n) \right\}.
\end{aligned}
$$

As a first step in proving that each term on the right hand side of (4.34) converges to 0 we try to understand the classes of functions $\mathcal{W}_{M_1}(n)$ and $\mathcal{W}_{M_1}^*$. In the following lemma, proved in Section 4.13.2, we find the bracketing numbers and envelope functions for the classes. This will be used to prove the results that follow.

Lemma 37. *Fix M_1, and $\delta > 0$. Then $\mathcal{W}_{M_1}(n)$ is a Donsker class and there exists a $V^* < \infty$ such that $\sup_{f \in \mathcal{W}_{M_1}^*} \|f\|_{2,\infty} \le V^*$. Moreover, for some c depending only on M_1, we have*

$$
N_{[]}(\varepsilon, \mathcal{W}_{M_1}(n), \|\cdot\|_{2,P_{\theta_0,m_0}}) \le N_{[]}(\varepsilon, \mathcal{W}_{M_1}^*, \|\cdot\|_{2,P_{\theta_0,m_0}}) \le c \exp(c/\varepsilon) \varepsilon^{-2d} \tag{4.41}
$$

and

$$
\sup_{f \in \mathcal{W}_{M_1}(n)} \|f\|_{2,P_{\theta_0,m_0}}^2 \le K_L^2 n^{-1/5}, \tag{4.42}
$$

where $K_L = \sqrt{2\|k'\|_\infty^2 + L^2 \|k'\|_\infty^2 T^2}$ and k is defined in (4.20).

The study of limiting behaviors of the first three terms on the right hand side of (4.34) are similar. For every fixed M_1 the first term in the right hand side of (4.34) can be bounded from above as

$$\begin{aligned}
&\mathbb{P}\Big(|\mathbb{G}_n([m_0\circ\theta_0-\check{m}\circ\theta_0]U_{\check{\theta},\check{m}})| \ > \delta\Big)\\
&\leq \mathbb{P}\Big(|\mathbb{G}_n([m_0\circ\theta_0-\check{m}\circ\theta_0]U_{\check{\theta},\check{m}})| \ > \delta, (\check{\theta},\check{m})\in\mathcal{C}_{M_1}(n)\Big)\\
&\quad + \mathbb{P}\big((\check{\theta},\check{m})\notin\mathcal{C}_{M_1}(n)\big)\\
&\leq \mathbb{P}\Big(\sup_{(\theta,m)\in\mathcal{C}_{M_1}(n)}|\mathbb{G}_n([m_0\circ\theta_0-m\circ\theta_0]U_{\theta,m})|>\delta\Big)\\
&\quad + \mathbb{P}\big((\check{\theta},\check{m})\notin\mathcal{C}_{M_1}(n)\big),
\end{aligned}$$

where $U_{\check{\theta},\check{m}}:\mathcal{X}\mapsto\mathbb{R}^{d-1}$ is defined in (4.35). By Theorem 26 we have that $\check{\theta}$ and $\check{m}$ are consistent in the Euclidean and supremum norms, respectively. Furthermore, by Theorem 27 and 28, we have that $n^{1/10}|\check{\theta}-\theta_0|=o_p(1)$ and $n^{1/5}\int_{D_0}|m'(t)-m_0'(t)|^2dt = o_p(1)$, respectively. Thus, it is easy to see that, for any $\varepsilon>0$, there exists M_1 (depending on ε) such that

$$\mathbb{P}((\check{\theta},\check{m})\notin\mathcal{C}_{M_1}(n))\leq\varepsilon, \quad \text{for all sufficiently large } n.$$

Hence, it is enough to show that for the above choice of M_1 we have

$$\mathbb{P}\left(\sup_{(\theta,m)\in\mathcal{C}_{M_1}(n)}|\mathbb{G}_n([m_0\circ\theta_0-m\circ\theta_0]U_{\theta,m})|>\delta\right)\leq\epsilon$$

for sufficiently large n. We prove this in Lemma 38.

Lemma 38. *Fix M_1, and $\delta>0$. For $n\in\mathbb{N}$, let us define*

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

Then $\mathcal{D}_{M_1}(n)$ is a Donsker class such that

$$\sup_{f\in\mathcal{D}_{M_1}(n)}\|f\|^2_{2,P_{\theta_0,m_0}}\leq D^2_{M_1}n^{-1/5},$$

where $D_{M_1}:=2M_1K_L$. Moreover $J_{[]}(\gamma,\mathcal{D}_{M_1}(n),\|\cdot\|_{2,P_{\theta_0,m_0}})\lesssim\gamma^{1/2}$ and

$$\mathbb{P}\Big(\sup_{f\in\mathcal{D}_{M_1}(n)}|\mathbb{G}_nf|>\delta\Big)\to0, \qquad n\to\infty.$$

The following two lemmas, proved in the Sections 4.13.4 and 4.13.5, complete the proof of Theorem 36.

Lemma 39. *Fix M_1, and $\delta > 0$. For $n \in \mathbb{N}$, let us define*

$$\mathcal{A}_{M_1}(n) := \{[m \circ \theta_0 - m \circ \theta]U_{\theta,m} : (\theta, m) \in \mathcal{C}_{M_1}(n)\},$$
$$\mathcal{A}^*_{M_1} := \left\{[m \circ \theta_0 - m \circ \theta]U_{\theta,m} : (\theta, m) \in \mathcal{C}^*_{M_1}\right\}.$$

Then $\mathcal{A}_{M_1}(n)$ is Donsker class and $D_{M_1} n^{-1/10}$ is an envelope function. Moreover,

$$J_{[]}(\gamma, \mathcal{A}_{M_1}(n), \|\cdot\|_{2,P_{\theta_0,m_0}}) \lesssim \gamma^{1/2}.$$

Moreover, as $n \to \infty$, we have

$$\mathbb{P}\Big(|\mathbb{G}_n[(\check{m} \circ \theta_0 - \check{m} \circ \check{\theta})U_{\check{\theta},\check{m}}| > \delta\Big) \to 0.$$

Lemma 40. *If **(A0)-(A4)**, **(B1)-(B3)**, and **(L1)** hold, then*

$$\begin{aligned} |\sqrt{n}\mathbb{P}_n[\epsilon U_{\check{\theta},\check{m}}]| &= o_p(1) \\ \sqrt{n}\big|P_{\theta_0,m_0}\big[(m_0 \circ \theta_0 - \check{m} \circ \theta_0)U_{\check{\theta},\check{m}}\big]\big| &= o_p(1) \\ \sqrt{n}\big|P_{\theta_0,m_0}[(\check{m} \circ \theta_0 - \check{m} \circ \check{\theta})U_{\check{\theta},\check{m}}]\big| &= o_p(1) \end{aligned} \tag{4.43}$$

Now that we have shown that $(\check{\theta}, \check{m})$ is an approximate zero of $\mathbb{P}_n\psi_{\theta,m}$ and $\psi_{\theta_0,m_0} = \ell_{\theta_0,m_0}$, asymptotic normality and efficiency of $\check{\theta}$ now follows from the theory developed in Section 6.6 of [van der Vaart, 2002]. In the next theorem, we prove that $\psi_{\check{\theta},\check{m}}$ satisfies the "no-bias" (see equation 6.6 of [van der Vaart, 2002]) condition.

Theorem 37. *[**Step 3**] Under assumptions of Theorem 34,*

$$\sqrt{n}P_{\check{\theta},m_0}\psi_{\check{\theta},\check{m}} = o_p(1),$$

In Lemma 61 stated and proved in Appendix 4.12.8, we prove that $\psi_{\check{\theta},\check{m}}$ is a consistent estimator of ψ_{θ_0,m_0} under $L_2(P_{\theta_0,m_0})$ norm. The following theorem completes the proof of Theorem 34.

Theorem 38. *[**Step 4**] Under **(A0)-(A4)** and **(B2)**, we have*

$$\mathbb{G}_n(\psi_{\check{\theta},\check{m}} - \psi_{\theta_0,m_0}) = o_p(1). \tag{4.44}$$

The proof of the above theorem is similar to that of Theorem 36. We first find an upper bound for the left side of (4.44) and then show that each of the terms converge to zero; see Lemmas 62 and 63 in Section 4.13.8.

4.6 Computational algorithms

In this section we describe algorithms for computing the estimators proposed in (4.4), (4.5), and (4.6). As mentioned in Remark 10, in each of the three cases given a θ the minimization of the desired loss function is a convex optimization problem; see Sections 4.6.1.1 and 4.6.1.2 for more details. With the above observation in mind, we propose the following general alternating algorithm to compute the estimators. The algorithms discussed here are implemented in the R package `simest`; see [Kuchibhotla and Patra, 2016].

We now introduce some notation to set up a general framework for all the three estimators proposed in Section 4.3. Let $(m, \theta) \mapsto \mathfrak{C}(m, \theta)$ denote some nonnegative criterion function, e.g., $\mathfrak{C}(m, \theta)$ can be $\mathcal{L}_n(m, \theta; \lambda)$ or $Q_n(m, \theta)$. And suppose, we are interested in finding the minimizer of $\mathfrak{C}(m, \theta)$ over $(m, \theta) \in \mathfrak{A} \times \Theta$, e.g., in our case $\mathfrak{A}$ can be $\mathcal{R}$, $\mathcal{M}_L$ or $\mathcal{C}$. For every $\theta \in \Theta$, let us define

$$m_{\theta,\mathfrak{A}} := \arg\min_{m \in \mathfrak{A}} \mathfrak{C}(m, \theta).$$

Here, we have assumed that for every $\theta \in \Theta$, $m \mapsto \mathfrak{C}(m, \theta)$ has a unique minimizer in $\mathfrak{A}$ and $m_{\theta,\mathfrak{A}}$ exists. The alternating descent algorithm can be described as follows:

1. Start with a initial estimate of θ, say, $\theta^{(0)}$.
2. At iteration k, compute $m^{(k)} := m_{\theta^{(k)},\mathfrak{A}}$.
3. Find a point $\theta^{(k+1)} \in \Theta$ such that

$$\mathfrak{C}(m^{(k)}, \theta^{(k+1)}) \leq \mathfrak{C}(m^{(k)}, \theta^{(k)}).$$

In particular, one can take $\theta^{(k+1)}$ as a minimizer of $\mathfrak{C}(m^{(k)}, \theta)$.

4. Repeat steps 2 and 3 until convergence.

Note that, our assumptions on $\mathfrak{C}$ does not imply that $\theta \mapsto \mathfrak{C}(m_{\theta,\mathfrak{A}}, \theta)$ is a convex function. In fact in our examples the "profiled" criterion function $\theta \mapsto \mathfrak{C}(m_{\theta,\mathfrak{A}}, \theta)$ is not necessarily convex. Thus the algorithm discussed above is not guaranteed to converge to the global minimizer. However, the algorithm guarantees that the criterion value is nonincreasing, i.e., $\mathfrak{C}(m^{(k+1)}, \theta^{(k+1)}) \leq \mathfrak{C}(m^{(k)}, \theta^{(k)})$. In Section 4.6.1.1, we discuss an algorithm to compute $m_{\theta,\mathcal{R}}$ when $\mathfrak{C}(m, \theta) = \mathcal{L}_n(m, \theta; \lambda)$. In Section 4.6.1.2, we discuss algorithms to compute $m_{\theta,\mathcal{M}_L}$ and $m_{\theta,\mathcal{C}}$ when $\mathfrak{C}(m, \theta) = Q_n(m, \theta)$.

4.6.1 Strategy for function estimation: Step 2

In the following subsections we describe the algorithms to compute $m_{\theta,\mathcal{R}}$, $m_{\theta,\mathcal{M}_L}$ and $m_{\theta,\mathcal{C}}$. Before proceeding further, we use the following notation. Fix an arbitrary $\theta \in \Theta$. Let $(t_1, t_2, \cdots, t_n)$ represent the vector $(\theta^\top x_1, \cdots, \theta^\top x_n)$ with sorted values so that $t_1 < t_2 < \cdots < t_n$; in Remark 14 we discuss a solution for the scenarios with ties. Without loss of generality let $y := (y_1, y_2, \ldots, y_n)$ represent the vector of responses corresponding to t_i.

4.6.1.1 Penalized convex least squares

When $\mathfrak{C}(m, \theta) = \mathcal{L}_n(m, \theta; \lambda)$, we need to minimize the objective function

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - m(t_i))^2 + \lambda^2 \int \{m''(t)\}^2 dt.$$

In the following we use m to denote the function $t \mapsto m(t)$ as well as the the vector $(m(t_1), \ldots, m(t_n))$ interchangeably. Consider the objective function

$$(y - m)^\top Q(y - m) + \lambda^2 \int \{m''(t)\}^2 dt,$$

to be minimized over $\mathcal{R}$ and Q is any positive definite matrix. In most cases Q is a $n \times n$ identity matrix; see Remark 14 for other possible scenarios. Theorem 1 of [Elfving and Andersson, 1988] gives the characterization of the minimizer over $\mathcal{R}$. They show that $\hat{m} := \arg\min_{m \in \mathcal{R}} (y - m)^\top Q(y - m) + \lambda^2 \int \{m''(t)\}^2 dt$, will satisfy

$$\hat{m}''(t) = \max\{\hat{\alpha}^\top M(t), 0\} \quad \text{and} \quad \hat{m} = y - \lambda^2 Q^{-1} K^\top \hat{\alpha}.$$

Here $M(t) := (M_1(t), M_2(t), \ldots, M_{n-2}(t))$ and $\{M_i\}_{1\le i\le n-2}$ are real-valued functions defined as

$$M_i(x) := \begin{cases} \frac{1}{t_{i+2}-t_i} \cdot \frac{x-t_i}{t_{i+1}-t_i} & \text{if } t_i \le x < t_{i+1}, \\ \frac{1}{t_{i+2}-t_i} \cdot \frac{t_{i+2}-x}{t_{i+2}-t_{i+1}} & \text{if } t_{i+1} \le x < t_{i+2}, \end{cases}$$

and $\hat{\alpha}$ is a solution of the following equation:

$$[T(\alpha) + \lambda^2 K Q^{-1} K^\top]\alpha = Ky, \tag{4.45}$$

where K is a $(n-2)\times n$ banded matrix containing second order divided differences

$$K_{i,i} = \frac{1}{(t_{i+1}-t_i)(t_{i+2}-t_i)}, \quad K_{i,i+1} = -\frac{1}{(t_{i+2}-t_{i+1})(t_{i+1}-t_i)},$$
$$K_{i,i+2} = \frac{1}{(t_{i+2}-t_i)(t_{i+2}-t_{i+1})}.$$

(all the other entries of K are zeros) and the matrix $T(\alpha)$ is defined by

$$T(\alpha) = \int M(t)M(t)^\top \mathbb{1}_{\{\alpha^\top M(t)>0\}} dt.$$

We use the initial value of α as $\alpha_i = (t_{i+2}-t_i)/4$ based on the empirical evidence suggested by [Elfving and Andersson, 1988] and use Equation (4.45) repeatedly until convergence. This algorithm was shown to have quadratic convergence in [Dontchev *et al.*, 2003].

Remark 14. *The matrices involved in the algorithm have entries depending on fractions such as* $1/(t_{i+1}-t_i)$. *Thus if there are ties in* $\{t_i\}_{1\le i\le n}$, *then the matrix* K *is incomputable. Such fractions can make the matrices ill-conditioned (for the purposes of numerical calculations) if* $t_{i+1}-t_i$ *is very small. Thus to avoid ill-conditioning of matrices, in practice one might have to do pre-bin the data which leads to a diagonal matrix* Q *with different diagonal entries. One common method of pre-binning the data is to take means of all points for which the first coordinates are close. To be more precise, if we choose a tolerance of* $\eta = 10^{-6}$ *and suppose* $0 < t_2 - t_1 < t_3 - t_1 < \eta$, *then we take the mean of* $(t_1, y_1), (t_2, y_2), (t_3, y_3)$ *as the first entry in the sample and put* $Q_{1,1} = 3$ *and the total number of data points is now* $n-2$.

4.6.1.2 Convex Lipschitz least squares

When $\mathfrak{C}(m,\theta) = Q_n(m,\theta)$, we consider the problem of minimizing $\sum_{i=1}^n \{y_i - m(t_i)\}^2$ over $m \in \mathcal{C}$ or $m \in \mathcal{M}_L$. As in Section 4.6.1.1, consider the general problem of minimizing

$$(y-m)Q(y-m) = |Q^{1/2}(y-m)|^2,$$

for some positive definite matrix Q. Here $Q^{1/2}$ denotes the square root of the matrix Q which can be obtained by Cholesky factorization. Observe that any minimizer can only be uniquely determined at the points t_i and so we define the optimum to be a piecewise linear interpolation between $\{t_i\}_{1\le i\le n}$. Let m denote the vector $(m(t_1), m(t_2), \ldots, m(t_n))$ as before. The convexity constraint then can be represented as

$$\frac{m_2 - m_1}{t_2 - t_1} \le \frac{m_3 - m_2}{t_3 - t_2} \le \cdots \le \frac{m_n - m_{n-1}}{t_n - t_{n-1}}. \tag{4.46}$$

The Lipschitz constraint along with convexity can be imposed by adding the constraints,

$$-L \le \frac{m_2 - m_1}{t_2 - t_1} \quad \text{and} \quad \frac{m_n - m_{n-1}}{t_n - t_{n-1}} \le L. \tag{4.47}$$

In particular, the minimization problem at hand can be represented as

$$\text{minimize } |Q^{1/2}(m-y)|^2 \text{ subject to } Am \ge b, \tag{4.48}$$

for A and b written so as to represent (4.46) and/or (4.47). Define $z := Q^{1/2}(m-y)$, so that $m = Q^{-1/2}z + y$. Using this, we have $Am \ge b$ if and only if $AQ^{-1/2}z \ge b - Ay$. Thus, (4.48) is equivalent to

$$\text{minimize } |z|^2 \text{ subject to } Gz \ge h, \tag{4.49}$$

where $G := AQ^{-1/2}$ and $h := b - Ay$. An equivalent problem is to

$$\text{minimize } |Eu - \ell|, \text{ over } u \succeq 0, \text{ where } E := \begin{bmatrix} G^\top \\ h^\top \end{bmatrix} \text{ and } \ell := [0,\ldots,0,1]^\top \in \mathbb{R}^{n+1}. \tag{4.50}$$

Here $\succeq$ represents coordinate-wise inequality. A proof of this equivalence can be found in pages 165-167 of [Lawson and Hanson, 1974]; see [Lawson and Hanson, 1974] and [Chen and Plemmons, 2010] for algorithms to solve (4.49) and (4.50).

If $\hat{u}$ denotes the solution of (4.50), then the solution of (4.49) is given as follows. Define $r := E\hat{u} - \ell$, then $\hat{z}$ (the minimizer of 4.49) is $\hat{z} := (-r_1/r_{n+1}, \ldots, -r_n/r_{n+1})^\top$. Hence the solution to (4.48) is then given by $\hat{y} = Q^{-1/2}\hat{z} + y$.

4.6.2 Algorithm for computing $\theta^{(k+1)}$

In this subsection we describe the algorithm to find the minimizer of $\mathfrak{C}(m^{(k)}, \theta)$ over $\theta \in \Theta$. Recall that Θ is defined to be the "positive" half of the unit sphere, so that θ belongs to a $d-1$ dimensional manifold in $\mathbb{R}^d$. Treating this problem as minimization over a manifold, one can apply a gradient descent algorithm by moving along a geodesic as done in a similar context in Section 3.3 of [Samworth and Yuan, 2012]. But it is computationally expensive to move along a geodesic and so, we follow the approach of [Wen and Yin, 2013] wherein we move along a retraction with the guarantee of descent. To explain the approach of [Wen and Yin, 2013], let the objective function be denoted by $f(\theta)$; in our case $f(\theta) = \mathfrak{C}(m^{(k)}, \theta)$. Let α be a point (initial guess for θ) on the sphere with positive first coordinate and define

$$g := \nabla f(\alpha) \in \mathbb{R}^d \quad \text{and} \quad A := g\alpha^\top - \alpha g^\top,$$

where ∇ denotes the gradient. In the following we use g to denote both the function and the We are trying to find a choice of τ such that $f(\theta(\tau))$ is as much smaller than $f(\alpha)$ as possible; step 3 of the algorithm described in Section 4.6. Thus the next iteration is then given by a point on the path $\tau \mapsto \theta(\tau)$, where

$$\theta(\tau) := \left(I + \frac{\tau}{2}A\right)^{-1}\left(I - \frac{\tau}{2}A\right)\alpha.$$

It is easy to verify that

$$\left.\frac{\partial f(\theta(\tau))}{\partial \tau}\right|_{\tau=0} \leq 0;$$

see Lemma 3 of [Wen and Yin, 2013]. This implies that $\tau \mapsto f(\theta(\tau))$ is a nonincreasing function in the neighborhood of 0. Given a value of g and α, $\theta(\tau)$ has the following closed form expression:

$$\theta(\tau) = \frac{1 + \frac{\tau^2}{4}[(\alpha^\top g)^2 - |g|^2] + \tau\alpha^\top g}{1 - \frac{\tau^2(\alpha^\top g)^2}{4} + \frac{\tau^2|g|^2}{4}}\alpha - \frac{\tau}{1 - \frac{\tau^2(\alpha^\top g)^2}{4} + \frac{\tau^2|g|^2}{4}}g;$$

see Lemma 4 of [Wen and Yin, 2013]. Recall that for every $\eta \in \Theta$, η_1 (the first coordinate of η) is nonnegative. For $\theta(\tau)$ to lie in Θ, τ has to satisfy the following inequality

$$\frac{\tau^2}{4}[(\alpha^\top g)^2 - |g|^2] + \tau\left(\alpha^\top g - \frac{g_1}{\alpha_1}\right) + 1 \geq 0, \tag{4.51}$$

where g_1 and α_1 represent the first coordinates of the vectors g and α.

This implies that a valid choice of τ must lie between the zeros of the quadratic expression on the left hand side of (4.51), given by

$$2\frac{(\alpha^\top g - g_1/\alpha_1) \pm \sqrt{(\alpha^\top g - g_1/\alpha_1)^2 + |g|^2 - (\alpha^\top g)^2}}{|g|^2 - (\alpha^\top g)^2}.$$

Note that this interval always contains zero. Now we can perform a simple line search for $\tau \mapsto f(\theta(\tau))$ when τ in the above mentioned interval to find the value θ for the next iteration in step 3 of the main algorithm.

4.7 Simulation Study

In this section we illustrate the finite sample performance of the estimators defined in (3.5), (4.4), (4.5), and (4.6). We also compare their performance with the EFM estimator (estimating function method; see [Cui *et al.*, 2011]) and the EDR estimator (see [Hristache *et al.*, 2001]). We use `SmoothGCV` to denote the estimator proposed in Section 3.2 where the tuning parameter is chosen by generalized cross-validation; [Wahba, 1990]. For the convex constrained estimators, we use `CvxLSE` to denote the convex LSE estimator proposed in (4.4), `CvxPen` to denote the PLSE proposed in (4.5), and `CvxLip` to denote the LLSE proposed in (4.6). In the following, to compute `CvxPen` we have used $\hat{\lambda}_n = 0.01n^{1/5}$. In what follows, we will use $(\tilde{m}, \tilde{\theta})$ to denote a generic estimator that will help us describe the quantities in the plots and tables; e.g., we will use $\|\tilde{m} \circ \tilde{\theta} - m_0 \circ \theta_0\|_n = [\frac{1}{n}\sum_{i=1}^n (\tilde{m}(\tilde{\theta}^\top x_i) - m_0(\theta_0^\top x_i))^2]^{1/2}$ to denote the root mean squared prediction error of the estimation procedure for all the estimators considered in the simulation study. From the simulation study it is easy to conclude estimators proposed here have superior performance in all sampling scenarios considered.

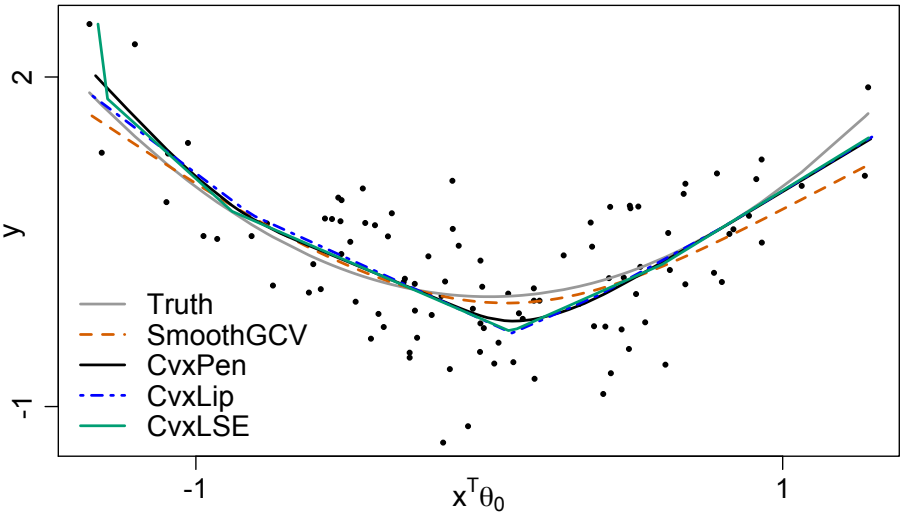

Figure 4.1: Function estimates for the model $Y = (\theta_0^\top X)^2 + N(0,1)$, where $\theta_0 = \mathbf{1}_5/\sqrt{5}, X \sim \text{Uniform}[-1,1]^5$, and $n = 100$.

4.7.1 A simple model

We start with a simple illustrative example. We observe 100 i.i.d. observations from the following homoscedastic model:

$$Y = (\theta_0^\top X)^2 + N(0,1), \text{ where } \theta_0 = \mathbf{1}_5/\sqrt{5} \text{ and } X \sim \text{Uniform}[-1,1]^5.$$

In Figure 4.1, we have a scatter plot of $\{(\theta_0^\top x_i, y_i)\}_{1\le i\le 100}$ overlaid with prediction curves $\{(\tilde{\theta}^\top x_i, \tilde{m}(\tilde{\theta}^\top x_i)\}_{1\le i\le 100}$ for the estimators proposed in Chapters 3 and 4. Table 4.1 displays all the estimates of θ_0 considered in the simulation study. To compute the function estimation error for EFM and EDR approaches we used cross validated smoothing splines to estimate the link function using their estimates for θ_0.

4.7.2 Example 2: Increasing dimension

To illustrate the behavior/performance of the estimators as d grows, we consider the following single-index model:

$$Y = (\theta_0^\top X)^2 + N(0,.2^2), \text{ where } \theta_0 = (2,1,\mathbf{0}_{d-2})^\top/\sqrt{5} \text{ and } X \in \mathbb{R}^d \sim \text{Uniform}[-1,5]^d.$$

Table 4.1: Estimates of θ_0, "Theta Error":= $\sum_{i=1}^{5} |\tilde{\theta}_i - \theta_{0,i}|$, "Func Error":= $\|\tilde{m} \circ \theta_0 - m_0 \circ \theta_0\|_n$, and "Pred Error":= $\|\tilde{m} \circ \tilde{\theta} - m_0 \circ \theta_0\|_n$ for the data used in Figure 4.1.

	θ_1	θ_2	θ_3	θ_4	θ_5	Theta Error	Func Error	Pred Error
Truth	0.45	0.45	0.45	0.45	0.45	—	—	—
SmoothGCV	0.38	0.49	0.41	0.50	0.45	0.21	0.10	0.10
CvxPen	0.36	0.50	0.42	0.47	0.47	0.21	0.12	0.13
CvxLip	0.35	0.50	0.43	0.48	0.46	0.21	0.13	0.15
CvxLSE	0.36	0.50	0.43	0.45	0.48	0.20	0.18	0.15
EFM	0.35	0.49	0.41	0.49	0.47	0.24	0.10	0.11
EDR	0.30	0.48	0.46	0.43	0.53	0.29	0.12	0.15

In each replication we observe 200 samples from the model. It is easy to see that the performance of the all the estimators worsen as dimension increases from 10 to 100 and EDR has the worst overall performance; see Figure 4.2. However when $d = 100$, the convex constrained estimators have significantly better performance. This simulation scenario is similar to the one considered in Example 3 of Section 3.2 in [Cui *et al.*, 2011].

4.7.3 Example 3: Piecewise linear function and dependent covariates

Chapters 3 and 4 study the asymptotic properties of the estimators when the true link function is smooth. To understand the performance of the estimators when the truth is convex but not smooth, we consider the following model:

$$Y = |\theta_0^\top X| + N(0, .1^2), \tag{4.52}$$

where $X \in \mathbb{R}^6$ 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 and θ_0 is $(1.3, -1.3, 1, -0.5, -0.5, -0.5)/\sqrt{5.13}$. The distribution of the covariates is similar to the one considered in Section V.2 of [Li and Patilea, 2015]. Observe that as the truth is not smooth, the convex constrained least squares estimators (`CvxLip` and `CvxLSE`) have

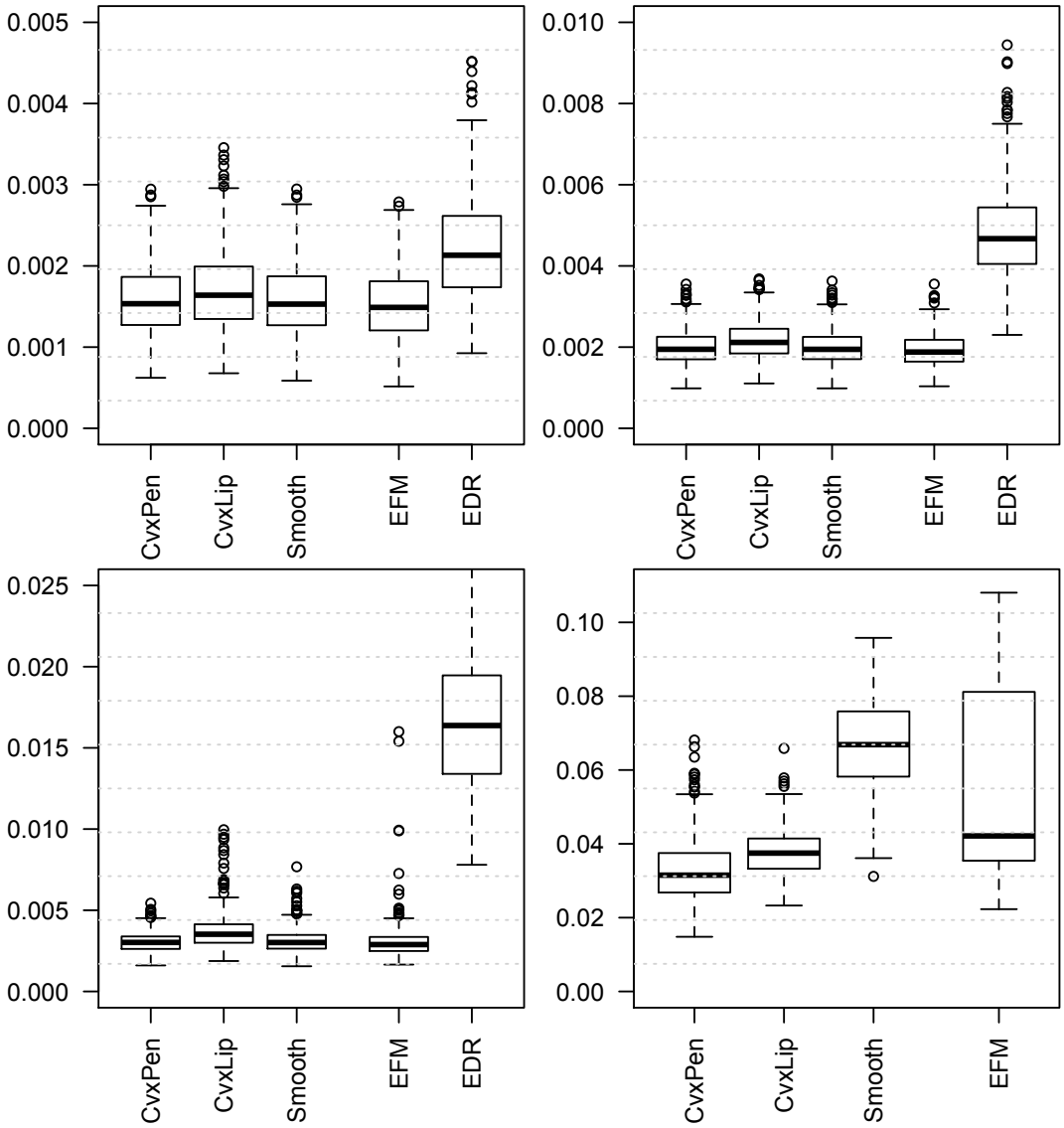

Figure 4.2: Boxplots of $\sum_{i=1}^{d} |\hat{\theta}_i - \theta_{0,i}|/d$ (over 500 replications) based on 200 observations from Example 2 in Section 4.7.2 for dimensions $10, 25, 50$, and 100, shown in the top-left, the top-right, the bottom-left, and the bottom-right panels, respectively. The bottom-right panel doesn't include EDR as the R-package `EDR` does not allow for $d = 100$.

improved performance compared to the (smoothness) penalized least squares estimators (`CvxPen` and `SmoothGCV`). Also observe that both `EFM` and `EDR` fail to estimate the true parameter θ_0.

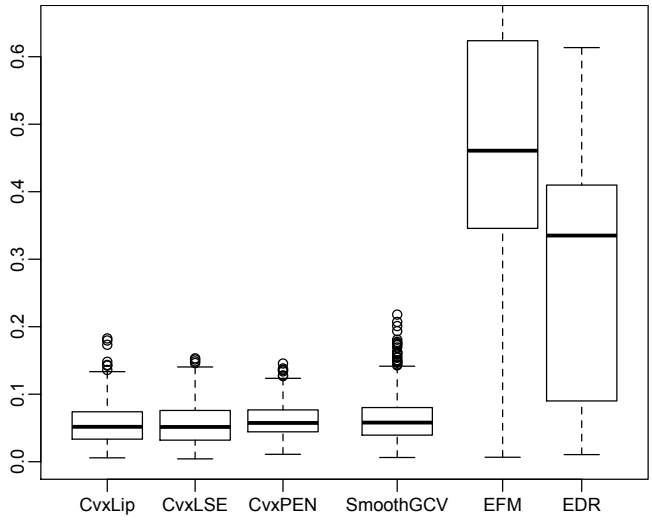

Figure 4.3: Box plots of $\sum_{i=1}^{6} |\tilde{\theta}_i - \theta_{0,i}|$ for the model (4.52) Here $d = 6$, $n = 200$ and we have 500 replications.

4.8 Proof of results in Section 4.2

4.8.1 Proof of Lemma 29

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

$$|m'(s) - m'(s_0)| \quad = \quad \left| \int_{s_0}^{s} m''(t)dt \right| \leq \left| \int_{s_0}^{s} \left| m''(t) \right|^2 dt \right|^{1/2} |s - s_0|^{1/2} \leq J(m)|s - s_0|^{1/2},$$

for every $s, s_0 \in D$.

4.8.2 Proof of Lemma 30

Integrating the inequality

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

with respect to s, we get

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

where $\varnothing(D)$ is the diameter of D, which will be finite since D is a compact subset of $\mathbb{R}$. Since $\|m\|_\infty \leq M$, we get that $|m'(s_0)|$ (by choosing s appropriately) and hence $\|m'\|_\infty$ is bounded by a multiple of $1 + J(m)$ as s_0 is an arbitrary point in D.

4.8.3 Proof of Lemma 31

In the following we show that (m_0, θ_0) is the minimizer of Q and is well-separated, with respect to the $L_2(P_X)$ norm, from $\{(m, \theta) : \ m \circ \theta \in L_2(P_X)\}$. Choose arbitrarily small $\delta > 0$, and pick any $(m, \theta) \in \{(m, \theta) : \ m \circ \theta \in L_2(P_X)\}$ such that $\|m \circ \theta - m_0 \circ \theta_0\|^2 > \delta^2$. Then

$$Q(m, \theta) = \mathbb{E}[Y - m_0(\theta_0^\top X)]^2 + \mathbb{E}[m_0(\theta_0^\top X) - m(\theta^\top X)]^2,$$

since $\mathbb{E}(\epsilon|X) = 0$. Thus we have that $Q(m, \theta) > Q(m_0, \theta_0) + \delta^2$.

4.9 Proof of results in Section 4.3

4.9.1 Proof or Theorem 19

The proof closely follows the proof of existence provided in [Murphy *et al.*, 1999]. The minimization problem considered is

$$\inf_{\theta \in \Theta, m \in \mathcal{R}} \frac{1}{n} \sum_{i=1}^{n} \{y_i - m(x_i^\top \theta)\}^2 + \lambda \int_D \{m''(t)\}^2 dt.$$

This can be equivalently written as

$$\inf_{\theta \in \Theta} \inf_{m \in \mathcal{R}} \frac{1}{n} \sum_{i=1}^{n} \{y_i - m(x_i^\top \theta)\}^2 + \lambda \int_D \{m''(t)\}^2 dt. \tag{4.53}$$

Define for each θ, $\mathcal{T}_\theta := \{p = (p_1, p_2, \ldots, p_n) : (x_1^\top \theta, p_1), (x_2^\top \theta, p_2), \ldots, (x_n^\top \theta, p_n)$ form a convex data/sequence$\}$. Here by convex data we mean that the slopes are nondecreasing with respect to the increasing x-coordinates, i.e, if $x_1^\top \theta < x_2^\top \theta < \cdots < x_n^\top \theta$, then

$$\frac{p_2 - p_1}{x_2^\top \theta - x_1^\top \theta} \leq \frac{p_3 - p_2}{x_3^\top \theta - x_2^\top \theta} \leq \cdots \leq \frac{p_n - p_{n-1}}{x_n^\top \theta - x_{n-1}^\top \theta}.$$

This sequence of inequalities can be written in a matrix form as $Cp \geq \mathbf{0}$, where $C_{(n-2)\times n}$ is a three-banded upper triangular matrix. Hence, $\mathcal{T}_\theta$ is a closed convex subset of $\mathbb{R}^n$. Minimization of (4.53) can be equivalently written as

$$\inf_{\theta \in \Theta} \inf_{p \in \mathcal{T}_\theta} \inf_{m \in \mathcal{M}_{\theta,p}} Q_n(m, \theta) + \lambda \int_D \{m''(t)\}^2 dt.$$

Here $\mathcal{M}_{\theta,p} = \{m \in \mathcal{R} : m(t_i) = p_i \text{ for } 1 \leq i \leq n\}$, with $\{t_i\}$ sequence being the sequence formed by ordering $\{\theta^\top x_i\}$. Observe that $Q_n(m, \theta)$ is constant on $\mathcal{M}_{\theta,p}$. Hence the

third infimum in the previous display is over the integral on $\mathcal{M}_{\theta,p}$. We now consider the minimization

$$\inf_{m\in\mathcal{M}_{\theta,p}} \int_D \{m''(t)\}^2 dt = \inf\left\{ \int_D \{m''(t)\}^2 dt \big| m''(t) \geq 0 \text{ and } m(t_i) = p_i \right\}. \tag{4.54}$$

The restrictions on the function values of m can be expressed in terms of the second derivative of the function m using Peano's theorem (see Chapter IX of [Davis, 1963]) as

$$\int_D m''(t) M(t) dt = Kp,$$

where $M(t) = (M_1(t), M_2(t), \ldots, M_{n-2}(t))^\top$ and M_i is a normalized B-spline supported on $[t_i, t_{i+2}]$ such that $\int_D M_i(t)dt = 0.5$ and Kp is the vector of the second order divided differences of $\{(t_i, p_i)\}_{1\leq i\leq n}$; see Section 4.6.1.1 for details. Using this, we can get a unique solution to minimization in (4.54) by solving

$$\begin{aligned}\inf\left\{ \int_D \{m''(t)\}^2 dt \big| m''(t) \geq 0 \text{ and } \int_D m''(t)M(t)dt = Kp \right\} \\ = \inf\left\{ \int_D \{g(t)\}^2 dt \big| g(t) \geq 0 \text{ and } \int_D g(t)M(t)dt = Kp \right\}.\end{aligned}$$

The right side of above display is a minimum norm problem on a closed convex set which will have a unique minimum if that set is non-empty. For it to be non-empty, we require the existence of at least one convex interpolant which is implied by Theorem 2.2 of [Carnicer and Dahmen, 1994]. Hence infimum in (4.54) is attained for some function in $\mathcal{M}_{\theta,p}$. Let $m_{\theta,p}$ be the minimizer.

Observe that for any p and p' in $\mathcal{T}_\theta$, we have that $\nu m_{\theta,p} + (1-\nu) m_{\theta,p'} \in \mathcal{M}_{\theta,\nu p+(1-\nu)p'}$. As the semi-norm $m \mapsto J(m)$ is convex, we have

$$J(m_{\theta,\nu p_1+(1-\nu)p'}) \leq J(\nu m_{\theta,p} + (1-\nu) m_{\theta,p'}) \leq \nu J(m_{\theta,p}) + (1-\nu) J(m_{\theta,p'}).$$

We conclude that $p \mapsto J(m_{\theta,p})$ is a convex function. Moreover, $p \mapsto Q_n(m_{\theta,p}, \theta)$ is convex and continuous in p and $Q_n(m_{\theta,p}, \theta) + \lambda J(m_{\theta,p}) \to \infty$ as $\|p\|_2 \to \infty$. Hence the infimum with respect to $p \in \mathcal{T}_\theta$ is also attained. It follows that the infimum on the right side of

$$T(\theta) = \inf_{m\in\mathcal{R}} Q_n(m, \theta) + \lambda \int_D \{m''(t)\}^2 dt$$

is attained. Let m_θ denote the minimizer.

Next we show that $\sup_{\theta\in\Theta} T(\theta) < \infty$. Observe that for every $\theta \in \Theta$,

$$T(\theta) \leq Q_n(\mathbf{0}, \theta) + \lambda J^2(\mathbf{0}) \leq \frac{1}{n}\sum_{i=1}^n y_i^2 =: K,$$

where $\mathbf{0}$ denote the constant function that takes the value 0 everywhere. Thus $\sup_{\theta\in\Theta} T(\theta) \leq K$, which implies

$$Q_n(m_\theta, \theta) + \lambda \int_D \{m_\theta''(t)\}^2 dt \leq K \quad \text{and} \quad \int_D \{m_\theta''(t)\}^2 dt \leq K/\lambda$$

implying that there must exist a finite constant L such that

$$T(\theta) = \inf_{m\in\mathcal{R}\cap\{m:J(m)\leq L\}} Q_n(m, \theta) + \lambda \int_D \{m''(t)\}^2 dt.$$

Next we show that the set of functions $\{\theta \mapsto Q_n(m, \theta) : m \in \mathbb{R}, J(m) \leq L\}$ is equicontinuous. Let $\theta, \theta_1 \in \Theta$ such that $|\theta - \theta_1| \leq \delta$ for some ≤ 1, then

$$\begin{aligned}
n|Q_n(m,\theta) - Q_n(m,\theta_1)| &= \left|\sum_{i=1}^n (y_i - m(\theta^\top x_i))^2 - \sum_{i=1}^n (y_i - m(\theta_1^\top x_i))^2\right| \\
&= \left|\sum_{i=1}^n (m(\theta_1^\top x_i) - m(\theta^\top x_i))(2y_i - m(\theta^\top x_i) - m(\theta_1^\top x_i))\right| \\
&\leq \sum_{i=1}^n |m(\theta_1^\top x_i) - m(\theta^\top x_i)| \times |2y_i - m(\theta^\top x_i) - m(\theta_1^\top x_i)| \\
&\leq \max_{1\leq i\leq n} |2y_i| \sum_{i=1}^n |m(\theta_1^\top x_i) - m(\theta^\top x_i)| + \sum_{i=1}^n |m(\theta_1^\top x_i) - m(\theta^\top x_i)|^2 \\
&\lesssim (1 + J(m))|\theta - \theta_1| + [(1 + J(m))|\theta - \theta_1|]^2 \lesssim \delta,
\end{aligned}$$

where the penultimate inequality follows from :

$$|m(\theta^\top x) - m(\theta_1^\top x)| \leq \|m'\|_\infty |x^\top(\theta - \theta_1)| \lesssim (1 + J(m))|\theta - \theta_1|,$$

see Lemma 30. Hence $\theta \mapsto T(\theta)$ is a continuous function (see Lemma 1 of [Jennrich, 1969]) and attains the minimum on the compact set Θ. Hence the existence of a minimizer of the penalized least squares is established.

4.9.2 Proofs of Theorem 20

We consider the estimator

$$(\check{m}_n, \check{\theta}_n) = \operatorname*{arg\,min}_{(m,\theta)\in\mathcal{M}_L\times\Theta} Q_n(m, \theta).$$

Fix $\theta \in \Theta$. Observe that $m \in \mathcal{M}_L \mapsto Q_n(m, \theta)$ is a coercive continuous convex function on a convex domain. Thus for every $\theta \in \Theta$ the minimizer of $m \in \mathcal{M}_L \mapsto Q_n(m, \theta)$ exists. Let us define

$$m_\theta := \operatorname*{arg\,min}_{m \in \mathcal{M}_L} Q_n(m, \theta) \;\text{ and }\; T(\theta) := Q_n(m_\theta, \theta).$$

Observe that $\check{\theta}_n := \arg\min_{\theta \in \Theta} T(\theta)$. As Θ is a compact set, the existence of the minimizer $\theta \mapsto T(\theta)$ will be established if we can show that $T(\theta)$ is a continuous function on Θ; see the Weierstrass extreme value theorem. We will next prove that $\theta \mapsto T(\theta)$ is a continuous function.

But first we will show that $\|m_\theta\|_\infty \leq C_n$ uniformly in θ. Observe that $\sum_{i=1}^n (y_i - m_\theta(\theta^\top x_i))^2 \leq \sum_{i=1}^n y_i^2$. Thus

$$\sum_{i=1}^n \Big[m_\theta(\theta^\top x_i)\Big]^2 \leq 2 \sum_{i=1}^n y_i m_\theta(\theta^\top x_i) \leq 2 \left(\sum_{i=1}^n y_i^2\right)^{1/2} \left(\sum_{i=1}^n \Big[m_\theta(\theta^\top x_i)\Big]^2\right)^{1/2}.$$

Hence, we have $\sum_{i=1}^n m_\theta^2(\theta^\top x_i) \leq 4 \sum_{i=1}^n y_i^2$. Since m_θ is uniformly Lipschitz on a bounded set D, we have that $\|m_\theta\|_\infty \leq C_n$ uniformly in θ.

To complete the proof, we show that $\theta \mapsto T(\theta)$ is a continuous function. It is enough to prove that the class of functions

$$\{\theta \mapsto Q_n(m, \theta) : m \in \mathcal{M}_L,\ \|m\|_\infty \leq 2n^{1/2} \big[\sum_{i=1}^n y_i^2\big]^{1/2}\}$$

is equicontinuous. Observe that for $\theta, \theta_1 \in \Theta$, we have

$$\begin{aligned}
n|Q_n(m, \theta) - Q_n(m, \theta_1)| &= \left|\sum_{i=1}^n (y_i - m(\theta^\top x_i))^2 - \sum_{i=1}^n (y_i - m(\theta_1^\top x_i))^2\right| \\
&= \left|\sum_{i=1}^n (m(\theta_1^\top x_i) - m(\theta^\top x_i))(2y_i - m(\theta^\top x_i) - m(\theta_1^\top x_i))\right| \\
&\leq \sum_{i=1}^n |m(\theta_1^\top x_i) - m(\theta^\top x_i)| \times |2y_i - m(\theta^\top x_i) - m(\theta_1^\top x_i)| \\
&\leq L \sum_{i=1}^n |\theta_1^\top x_i - \theta^\top x_i| \times \left(2|y_i| + 4n^{1/2} \big[\sum_{i=1}^n y_i^2\big]^{1/2}\right) \\
&\leq 2LT \left(2 \max_i |y_i| + 4n^{1/2} \big[\sum_{i=1}^n y_i^2\big]^{1/2}\right) |\theta - \theta_1|.
\end{aligned}$$

4.10 Proofs of results in Section 4.4.1

4.10.1 Proof of Theorem 21

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

$$Q_n(\hat{m}, \hat{\theta}) + \hat{\lambda}_n^2 J^2(\hat{m}) \leq Q_n(m_0, \theta_0) + \hat{\lambda}_n^2 J^2(m_0). \tag{4.55}$$

Observe that by definition of $Q_n(m, \theta)$, we have that (4.55) implies

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

To find rate the 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

$$\begin{aligned}&Q_n(m_0, \theta_0) - Q_n(\hat{m}, \hat{\theta}) \\ &= \frac{2}{n}\sum_{i=1}^n \epsilon_i(\hat{m}(\hat{\theta}^\top x_i) - m_0(\theta_0^\top x_i)) - \frac{1}{n}\sum_{i=1}^n (\hat{m}(\hat{\theta}^\top x_i) - m_0(\theta_0^\top x_i))^2 \\ &\leq \left(\frac{2}{n}\sum_{i=1}^n \epsilon_i^2\right)^{1/2} \|\hat{m} \circ \hat{\theta} - m_0 \circ \theta_0\|_n - \|\hat{m} \circ \hat{\theta} - m_0 \circ \theta_0\|_n^2.\end{aligned} \tag{4.56}$$

Note that by **(A2)**, $(1/n)\sum_{i=1}^n \epsilon_i^2 = O(1)$ 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

$$Q_n(m_0, \theta_0) - Q_n(\hat{m}, \hat{\theta}) \geq \hat{\lambda}_n^2(J^2(\hat{m}) - J^2(m_0)) \geq -\hat{\lambda}_n^2 J^2(m_0) \geq o_p(1), \tag{4.57}$$

as $\hat{\lambda}_n = o_p(1)$. Combining (4.56) and (4.57), 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 $\|m_0 \circ \theta_0\|_\infty < \infty$.

By the Sobolev embedding theorem (see 19), we can find $\hat{m}_1$ and $\hat{m}_2$ such that

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

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

$$\begin{aligned} \frac{\|\hat{m}_1 \circ \hat{\theta}\|_n}{1 + J(m_0) + J(\hat{m})} &\leq \frac{\|\hat{m} \circ \hat{\theta}\|_n}{1 + J(m_0) + J(\hat{m})} + \frac{\|\hat{m}_2 \circ \hat{\theta}\|_n}{1 + J(m_0) + J(\hat{m})} \\ &\leq \frac{\|\hat{m} \circ \hat{\theta}\|_n}{1 + J(m_0) + J(\hat{m})} + \frac{\|\hat{m}_2\|_\infty}{1 + J(m_0) + J(\hat{m})} = O_p(1). \end{aligned} \tag{4.58}$$

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 **(A4)** and (4.2), 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

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

Thus by (4.58) we have

$$\frac{\max(\hat{\beta}_1, \hat{\beta}_2)}{1 + J(m_0) + J(\hat{m})} = O_p(1). \tag{4.59}$$

Moreover, since D is a bounded set, by (4.59) we have $\|\hat{m}_1\|_\infty/(1 + J(m_0) + J(\hat{m})) = O_p(1)$. Combining this with Lemma 19, we get

$$\frac{\|\hat{m}\|_\infty}{1 + J(m_0) + J(\hat{m})} \leq \frac{\|\hat{m}_1\|_\infty}{1 + J(m_0) + J(\hat{m})} + \frac{\|\hat{m}_2\|_\infty}{1 + J(m_0) + J(\hat{m})} = O_p(1). \tag{4.60}$$

Now define the class of functions

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

Observe that by (4.60), we can find a C_ε such that

$$\mathbb{P}\left(\frac{\hat{m} \circ \hat{\theta} - m_0 \circ \theta_0}{1 + J(m_0) + J(m)} \in \mathcal{B}_{C_\varepsilon} \right) \geq 1 - \varepsilon, \qquad \forall n. \tag{4.61}$$

Lemma 8.4 of [van de Geer, 2000b] gives a upper bound for $\sum_{i=1}^n \epsilon_i g(x_i)$, in terms of entropy of the class of functions g. In Lemma 41 we find the bracketing number for the class of functions $\mathcal{B}_C$.

Lemma 41. *For every fixed positive M_1, M_2, and C, we have*

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

Remark 15. *The proof of Lemma 41 follows from the proof of Lemma 21.*

In the view of (4.61), Lemmas 20 and 41 allow us to conclude

$$\frac{(1/n)\sum_{i=1}^n \epsilon_i(\hat{m}(\hat{\theta}^\top x_i) - m_0(\theta_0^\top x_i))}{\|\hat{m} \circ \hat{\theta} - m_0 \circ \theta_0\|_n^{3/4}(1 + J(m_0) + J(\hat{m}))^{1/4}} = O_p(n^{-1/2}). \tag{4.62}$$

Together, (4.57) and (4.62) imply

$$\begin{aligned}
&\hat{\lambda}_n^2(J^2(\hat{m}) - J^2(m_0)) \\
&\leq Q_n(m_0, \theta_0) - Q_n(\hat{m}, \hat{\theta}) \\
&= \frac{2}{n}\sum_{i=1}^n (y_i - m_0(\theta_0^\top x_i))(\hat{m}(\hat{\theta}^\top x_i) - m_0(\theta_0^\top x_i)) - \|\hat{m} \circ \hat{\theta} - m_0 \circ \theta_0\|_n^2 \\
&\leq \|\hat{m} \circ \hat{\theta} - m_0 \circ \theta_0\|_n^{3/4}(1 + J(m_0) + J(\hat{m}))^{1/4} O_p(n^{-1/2}) - \|\hat{m} \circ \hat{\theta} - m_0 \circ \theta_0\|_n^2.
\end{aligned} \tag{4.63}$$

We will now consider two cases.

Case 1: Suppose $J(\hat{m}) > 1 + J(m_0)$. By (4.63), 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

$$\|\hat{m} \circ \hat{\theta} - m_0 \circ \theta_0\|_n^{3/4} J(\hat{m})^{1/4} n^{-1/2} \leq C_1 \hat{\lambda}_n^2 J^2(m_0) \tag{4.64}$$

or

$$\hat{\lambda}_n^2 J^2(m_0) < C_2 \|\hat{m} \circ \hat{\theta} - m_0 \circ \theta_0\|_n^{3/4} J(\hat{m})^{1/4} O_p(n^{-1/2}) \tag{4.65}$$

hold with high probability as $n \to \infty$. Observe that when (4.64) holds we have

$$\|\hat{m} \circ \hat{\theta} - m_0 \circ \theta_0\|_n^2 + \hat{\lambda}_n^2 J^2(\hat{m}) \leq O_p(1) \hat{\lambda}_n^2 J^2(m_0). \tag{4.66}$$

Now it is easy to see that, (4.66) 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 (4.65) holds, we have

$$\|\hat{m} \circ \hat{\theta} - m_0 \circ \theta_0\|_n^2 + \hat{\lambda}_n^2 J^2(\hat{m}) \leq \|\hat{m} \circ \hat{\theta} - m_0 \circ \theta_0\|_n^{3/4} J(\hat{m})^{1/4} O_p(n^{-1/2}). \tag{4.67}$$

We can bound the first term on the left hand side of (4.67) as

$$\|\hat{m} \circ \hat{\theta} - m_0 \circ \theta_0\|_n \leq \left[J(\hat{m})^{1/4} O_p(n^{-1/2}) \right]^{4/5}. \tag{4.68}$$

A similar bound on the second term on the left hand side of (4.67) gives:

$$\begin{aligned}
\hat{\lambda}_n^2 J^2(\hat{m}) &\leq \|\hat{m} \circ \hat{\theta} - m_0 \circ \theta_0\|_n^{3/4} J(\hat{m})^{1/4} O_p(n^{-1/2}) \\
&\leq \left[J(\hat{m})^{1/4} O_p(n^{-1/2}) \right]^{3/5} J(\hat{m})^{1/4} O_p(n^{-1/2}) \text{ (by (4.68))} \\
&\leq J(\hat{m})^{2/5} \left[O_p(n^{-1/2}) \right]^{8/5},
\end{aligned}$$

which implies that

$$J(\hat{m}) = O_p(n^{-1/2}) \hat{\lambda}_n^{-5/4}. \tag{4.69}$$

Combining (4.68) and (4.69), 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 **(S2)**, we have that $\hat{\lambda}_n^{-1} = O_p(n^{2/5})$. Hence the conclusion follows.

Case 2: When $J(\hat{m}) \leq 1 + J(m_0)$, (4.63) 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 (4.60), $\|\hat{m}\|_\infty = O_p(1)$. The rates of convergence of $\|\hat{m} \circ \hat{\theta} - m_0 \circ \theta_0\|$ follows from Lemma 5.16 of [van de Geer, 2000b] (see Lemma 22).

Our proof of Theorem 21 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, 2000b].

4.10.2 Proof of Theorem 22

We first state and prove a lemma crucial to the proof of Theorem 22.

Lemma 42. *For every fixed M, the set of convex functions $m : D \to \mathbb{R}$ with $J(m) \leq M$ and $\|m\|_\infty \leq M$ is precompact relative to $\|\cdot\|_D^S$.*

Proof. By Lemma 29 the class of functions m' is uniformly Lipschitz of order $1/2$. Thus any sequence of functions m'_k is equicontinuous. By Lemma 30, m' is uniformly bounded as soon as $J(m)$ is uniformly bounded. Applying the Arzela-Ascoli theorem, we see that every sequence m_k with $J(m_k) = O(1)$ has a subsequence $\{k_l\}$ such that both m_{k_l} and m'_{k_l} converge uniformly on D. By Lemma 30 and the mean value theorem, we get that m is uniformly bounded. Thus applying the Arzela-Ascoli theorem, we get a subsequence $\{k_{l_j}\}$ of $\{k_l\}$ for which functions converge uniformly. Since these functions converge uniformly on compact set, by applying the dominated convergence theorem, we see that there exists a subsequence of m_k such that functions and derivatives converge. Furthermore, the derivative of the limit equals the limit of the derivative. □

Now, we will prove Theorem 22. Suppose that $\|m_k \circ \theta_k - m_0 \circ \theta_0\| \to 0$ and $J(m_k) = O(1)$. By Lemma 42, every subsequence of (m_k, θ_k) has a further subsequence such that $\theta_k \to \theta$ and $\|m_k - m\|_D^S \to 0$ for some θ and m. Then $\|m_k \circ \theta_k - m \circ \theta\| \to 0$ 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. Under the assumption that D_0 is closure of its interior, this implies that m' and m'_0 agree on D_0. Since the convergence in Lemma

42 is uniform, we get that $\|m - m_0\|_{D_0} \to 0$. Combining this with theorem 21, 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. By Lemma 29, we have that $|\hat{m}'(s) - \hat{m}'(a)| \leq J(\hat{m})|s-a|^{1/2} = O_p(1)$. Moreover, we have that $|\hat{m}'(a) - m_0'(a)| = o_p(1)$. Thus $\|\hat{m}'\|_\infty = O_p(1)$.

4.10.3 Proof of Theorem 23

We first state and prove a lemma that we will use to prove this theorem.

Lemma 43. *Suppose $m \in M_1$, $J(m) \in \infty$, and $\theta \in \Theta$. Then*

$$\mathbb{P}\left[m(\theta^\top X) - m(\theta_0^\top X) - m_0'(\theta_0^\top X)X^\top(\theta - \theta_0)\right]^2 \lesssim |\theta - \theta_0|^3 J^2(m) + |\theta - \theta_0|^2 \mathbb{P}[(m - m_0)'(\theta_0^\top X)]^2.$$

Proof.

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

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

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

By Lemma 29, we have

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

Thus, we have

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

and hence

$$\mathbb{P}\left\{m(\theta^\top X) - m(\theta_0^\top X) - m_0'(\theta^\top X)X^\top(\theta-\theta_0)\right\}^2 \lesssim |\theta-\theta_0|^3 J^2(m) + |\theta-\theta_0|^2 \mathbb{P}\{(m-m_0)'(\theta_0^\top X)\}^2. \qquad \square$$

Now observe that, since $|\hat\theta - \theta_0| \xrightarrow{P} 0$, $\mathbb{P}\{(\hat m - m_0)'(\theta_0^\top X)\}^2 \xrightarrow{P} 0$ and $J(\hat m) = O_p(1)$, we get

$$\mathbb{P}\left[\hat m(\hat\theta^\top X) - m_0(\theta_0^\top X)\right]^2 \gtrsim \mathbb{P}\left[m_0'(\theta_0^\top X)X^\top(\hat\theta-\theta_0) + (\hat m - m_0)(\theta_0^\top X)\right]^2 - o_p(1)|\hat\theta-\theta_0|^2. \tag{4.70}$$

Note that $\mathbb{P}$ is the expectation with respect to X and not with respect to $\hat m$ and $\hat\theta$. If we can now show that the expectation on the right side of (4.70) is bounded below by a multiple of $|\hat\theta-\theta_0| + \mathbb{P}(\hat m - m_0)'(\theta^\top X)$, then we get the rates of convergence of $\hat m$ and $\hat m$ given in Theorem 23. For proving this, we use the following lemma.

Lemma 44. *(Lemma 5.7 of [Murphy et al., 1999]) Let g_1 and g_2 be measurable functions such that $(\mathbb{P}g_1g_2)^2 \le c\mathbb{P}g_1^2\mathbb{P}g_2^2$ for a constant $c<1$. Then*

$$\mathbb{P}(g_1+g_2)^2 \ge (1-\sqrt{c})(\mathbb{P}g_1^2 + \mathbb{P}g_2^2).$$

Taking $g_1 = m_0'(\theta_0^\top X)X^\top(\hat\theta-\theta_0)$ and $g_2(\theta_0^\top X) = (\hat m - m_0)(\theta_0^\top X)$, in Lemma 44 we have

$$\begin{aligned}
&\mathbb{P}[m_0'(\theta_0^\top X)g_2(\theta_0^\top X)X^\top(\hat\theta-\theta_0)]^2 \\
&\qquad = \mathbb{P}m_0'(\theta_0^\top X)g(\theta_0^\top X)E(X^\top(\hat\theta-\theta_0)|\theta_0^\top X)^2 \\
&\qquad \le \mathbb{P}\big[\{m_0'(\theta_0^\top X)\}^2 E^2[X^\top(\hat\theta-\theta_0)|\theta_0^\top X]\big]\mathbb{P}g_2^2(\theta_0^\top X) \\
&\qquad < \mathbb{P}\big[\{m_0'(\theta_0^\top X)\}^2 E[\{X^\top(\hat\theta-\theta_0)\}^2|\theta_0^\top X]\big]\mathbb{P}g_2^2(\theta_0^\top X) \\
&\qquad = \mathbb{P}\big[\mathbb{E}[\{m_0'(\theta_0^\top X)X^\top(\hat\theta-\theta_0)\}^2|\theta_0^\top X)]\big]\mathbb{P}g_2^2(\theta_0^\top X) \\
&\qquad = \mathbb{P}[m_0'(\theta_0^\top X)X^\top(\hat\theta-\theta_0)]^2\mathbb{P}g_2^2(\theta_0^\top X) \\
&\qquad = \mathbb{P}g_1^2\mathbb{P}g_2^2.
\end{aligned}$$

Strict inequality in the above sequence of inequalities holds under the assumption that the conditional distribution of X given $\theta_0^\top X$ is nondegenerate. Thus, the assumption of Lemma 44 hold, for some $c<1$. Hence $\mathbb{P}(g_1+g_2)^2 \gtrsim \mathbb{P}g_1^2 + \mathbb{P}g_2^2$.

Now notice that, with $\hat{l}_n = \hat{\theta} - \theta_0$,

$$\mathbb{P}g_1^2 = \hat{l}_n^\top \mathbb{P}[XX^\top \{m_0'(\theta_0^\top X)\}^2]\hat{l}_n \geq \lambda_1 \hat{l}_n \hat{l}_n = \lambda_1 |\hat{\theta} - \theta_0|^2.$$

All these facts combined prove Theorem 23.

4.10.4 Proof of Theorem 24

We use the following interpolation inequality in [Agmon, 2010] to prove this theorem.

Lemma 45. *(Corollary 3.1, [Agmon, 2010]) Let $f : \mathbb{R} \to \mathbb{R}$ is continuously differentiable function on (a, b) and suppose we can write $f'(x) = f'(\eta) + \int_\eta^x f''(s)ds$ for all $a < \eta \leq x < b$. Furthermore, let $g : \mathbb{R} \to \mathbb{R}$ be a continuous density function and is bounded away from 0, i.e., $g(s) > \delta > 0$ for all $x \in (a, b)$. If $0 < \varepsilon \leq 1$, then*

$$\int_a^b |f'(s)|^2 g(s)ds \leq \gamma \Big[\varepsilon \int_a^b |f''(s)|^2 g(s)ds + \varepsilon^{-1} \int_a^b |f(s)|^2 g(s)ds \Big],$$

where γ depends only on δ, a, b, and $\max_{s\in(a,b)} g(s)$.

Take g to be the density of $\theta_0^\top X$ with respect to Lebesgue measure. By assumption **(A5)**, we have that g is continuous and bounded away from zero on the bounded set $D_{\theta_0} := \{t = \theta_0^\top x : x \in \mathcal{X}\}$. Furthermore, let $f = \hat{m} - m_0$. By assumption **(S1)**, we have that m_0 has an absolutely continuous first derivative. It can also be seen that $\hat{m}$, has an absolutely continuous derivative; see Section 2 of [Elfving and Andersson, 1988]. Thus an easy application of the Lemma 45, we have that

$$\|\hat{m}' \circ \theta_0 - m_0' \circ \theta_0\|^2 \leq \gamma \big[\varepsilon \|\hat{m}'' \circ \theta_0 - m_0'' \circ \theta_0\|^2 + \varepsilon^{-1} \|\hat{m} \circ \theta_0 - m_0 \circ \theta_0\|^2 \big].$$

By Theorem 22, we have that $J(\hat{m}) = O_p(1)$. Because g is bounded away from both zero and infinity, we have that $\int_{D_{\theta_0}} \hat{m}''(s)^2 g(s)ds \lesssim J(\hat{m})$ and $\int_{D_{\theta_0}} m_0''(s)^2 g(s)ds \lesssim J(m_0)$. Fixing $\varepsilon = \hat{\lambda}_n$, by Theorem 22, we have

$$\|\hat{m}' \circ \theta_0 - m_0' \circ \theta_0\|^2 \leq \gamma \big[\hat{\lambda}_n (J^2(\hat{m}) + J^2(m_0)) + \hat{\lambda}_n^{-1} O_p(\hat{\lambda}_n^2) \big] = O_p(\hat{\lambda}_n).$$

4.11 Proofs of results in Section 4.4.2

4.11.1 Proof of Theorem 25

To find the rate of convergence of $\check{m} \circ \check{\theta}$, we use the following modification of Theorem 3.2.5 of [van der Vaart, 1996]. In the following to avoid measurability difficulties, we use $\mathbb{P}^*$ and $\mathbb{E}^*$, outer probability and outer measure.

Lemma 46. *Let $\mathbb{M}_n$ be stochastic processes indexed by a semimetric set Υ and $\mathbb{M} : \Upsilon \to \mathbb{R}$ a deterministic function, such that for every $\eta \in \Upsilon$*

$$\mathbb{M}(\eta) - \mathbb{M}(\eta_0) \lesssim -d^2(\eta, \eta_0), \tag{4.71}$$

where $d(\cdot, \eta_0) : \Upsilon \to \mathbb{R}^+$. Let $\hat{\eta}_n := \operatorname{argmax}_{\eta \in \Upsilon} \mathbb{M}_n(\eta)$. For each $\varepsilon > 0$, suppose that the following hold:

1. *There exists Υ_ε, a subset of Υ, containing η_0 in its interior that satisfies*

$$\mathbb{P}^*(\hat{\eta}_n \notin \Upsilon_\varepsilon) \leq \varepsilon, \qquad \forall n. \tag{4.72}$$

2. *For every n and $\delta > 0$, the centered process $\mathbb{M}_n - \mathbb{M}$ satisfies*

$$\sqrt{n}\mathbb{E}^* \left| \sup_{\substack{d(\eta,\eta_0)<\delta, \\ \eta \in \Upsilon_\varepsilon}} |(\mathbb{M}_n - \mathbb{M})(\eta) - (\mathbb{M}_n - \mathbb{M})(\eta_0)| \right| \leq C_\varepsilon \phi_n(\delta), \tag{4.73}$$

for a constant $C_\varepsilon > 0$ and functions ϕ_n (not depending on ε) such that $\delta \mapsto \phi_n(\delta)/\delta^\alpha$ is decreasing in δ for some constant $\alpha < 2$ (not depending on n).

Then $r_n d(\hat{\eta}_n, \eta_0) = O_p^*(1)$ for every r_n satisfying $r_n^2 \phi_n(1/r_n) \leq \sqrt{n}$ for every n.

Remark 16. *The proof of Lemma 46 is similar to the proof given in Page 290 of [van der Vaart, 1996]. The only difference is that in the "peeling" argument the "shells" are now defined as $S_{j,n} := \{\eta : 2^{j-1} < r_n d(\eta, \eta_0) \leq 2^j$ and $\eta \in \Upsilon_\varepsilon\}$ and the first inequality in the proof now reads*

$$\mathbb{P}^*\big(r_n d(\hat{\eta}, \eta_0) > 2^M\big) \leq \sum_{j \geq M} \mathbb{P}^* \left(\sup_{\eta \in S_{j,n}} (\mathbb{M}_n(\eta) - \mathbb{M}_n(\eta_0)) \geq 0 \right) + \mathbb{P}^*(\hat{\eta} \notin \Upsilon_\varepsilon).$$

We will now obtain the desired rate of convergence in Theorem 25 by verifying conditions of Lemma 46. For the LLSE, $\Upsilon = \mathcal{M}_L \times \Theta$, $\eta = (m, \theta)$, $\eta_0 = (m_0, \theta_0)$, and

$$\hat{\eta}_n = (\check{m}_n, \check{\theta}_n) = \underset{(m,\theta)\in\mathcal{M}_L\times\Theta}{\operatorname{argmax}} -\frac{1}{n}\sum_{i=1}^{n}(y_i - m(\theta^\top x_i))^2.$$

The stochastic processes $\mathbb{M}_n$ and function $\mathbb{M}$ are defined as

$$\mathbb{M}_n(m,\theta) := -\frac{1}{n}\sum_{i=1}^{n}(y_i - m(\theta^\top x_i))^2 \text{ and } \mathbb{M}(m,\theta) := -\mathbb{E}(Y - m(\theta^\top X))^2. \quad (4.74)$$

For any (m_1, θ_1) and (m_2, θ_2) in $\mathcal{M}_L \times \Theta$, we define

$$d((m_1,\theta_1),(m_2,\theta_2)) := \|m_1 \circ \theta_1 - m_2 \circ \theta_2\|. \quad (4.75)$$

We first show that $\mathbb{M}$ defined in (4.74) satisfies (4.71). Observe that $\mathbb{E}(Y|X) = m_0(\theta_0^\top X)$. Thus

$$\begin{aligned}
&\mathbb{M}(m,\theta) - \mathbb{M}(m_0,\theta_0) \\
&\quad = \mathbb{E}[(Y - m_0(\theta_0^\top X))^2 - (Y - m(\theta^\top X))^2] \\
&\quad = -2\mathbb{E}\big[(Y - m_0(\theta_0^\top X))(m_0(\theta_0^\top X) - m(\theta^\top X))\big] - \mathbb{E}(m_0(\theta_0^\top X) - m(\theta^\top X))^2 \\
&\quad = -\mathbb{E}\left[\{m(\theta^\top X) - m_0(\theta_0^\top X)\}^2\right] \\
&\quad = -d^2((m,\theta),(m_0,\theta_0)).
\end{aligned}$$

Next for every $\varepsilon > 0$, we find Υ_ε such that (4.72) is satisfied. The following result (proved in Section 4.11.2) gives the form of Υ_ε.

Lemma 47. *Under assumption **(A2)**, we have that $\|\check{m}_n\|_\infty = O_p(1)$. Moreover, for every $\varepsilon > 0$, there exists a finite M_ε such that*

$$\mathbb{P}(\check{m}_n \notin \mathcal{M}_{M_\varepsilon,L}) \le \varepsilon, \quad \forall n,$$

where for any $M > 0$, we define

$$\mathcal{M}_{M,L} := \{m \in \mathcal{M}_L : \|m\|_\infty \le M\}. \quad (4.76)$$

We can now define $\Upsilon_\varepsilon := \mathcal{M}_{M_\varepsilon,L} \times \Theta$. By Lemma 47, we have

$$\mathbb{P}\big((\check{m}_n, \check{\theta}_n) \notin \Upsilon_\varepsilon\big) \le \varepsilon, \qquad \forall n.$$

To find the rate of convergence of $\check{m} \circ \check{\theta}$, we need to find a function $\phi_n(\delta)$ that satisfies (4.73). Recall that $\epsilon = Y - m_0(\theta_0^\top X)$. By definition of $\mathbb{M}_n$ and $\mathbb{M}$, we have

$$\begin{aligned}
&\sqrt{n}|(\mathbb{M}_n - \mathbb{M})(m, \theta) - (\mathbb{M}_n - \mathbb{M})(m_0, \theta_0)| \\
&= \left|\mathbb{G}_n\big[-2(Y - m_0(\theta_0^\top X))(m_0(\theta_0^\top X) - m(\theta^\top X)) + (m_0(\theta_0^\top X) - m(\theta^\top X))^2\big]\right| \\
&\leq \left|\mathbb{G}_n\big[2\epsilon(m(\theta^\top X) - m_0(\theta_0^\top X))\big]\right| + \left|\mathbb{G}_n\big[(m(\theta^\top X) - m_0(\theta_0^\top X))^2\big]\right|. \qquad (4.77)
\end{aligned}$$

Now, we find the upper bound $\phi_n(\delta)$ by obtaining upper bounds for both the terms in (4.77). Define two classes of functions

$$\begin{aligned}
\mathcal{H}_{M_\varepsilon, L}(\delta) &:= \{m \circ \theta - m_0 \circ \theta_0 : (m, \theta) \in \Upsilon_\varepsilon \text{ and } d((m, \theta), (m_0, \theta_0)) \leq \delta\} \\
\mathfrak{H}_{M_\varepsilon, L}(\delta) &:= \{f^2 : f \in \mathcal{H}_{M_\varepsilon, L}(\delta)\},
\end{aligned} \qquad (4.78)$$

where $d(\cdot, \cdot)$ is defined in (4.75). Thus by (4.77) we have

$$\begin{aligned}
&\mathbb{E}^* \sup_{\substack{d((m,\theta),(m_0,\theta_0))<\delta, \\ (m,\theta)\in\Upsilon_\varepsilon}} \sqrt{n}|(\mathbb{M}_n - \mathbb{M})(m, \theta) - (\mathbb{M}_n - \mathbb{M})(m_0, \theta_0)|, \\
&\leq \mathbb{E}^* \sup_{f \in \mathcal{H}_{M_\varepsilon, L}(\delta)} |\mathbb{G}_n[\epsilon f]| + \mathbb{E}^* \sup_{f \in \mathfrak{H}_{M_\varepsilon, L}(\delta)} |\mathbb{G}_n f|.
\end{aligned} \qquad (4.79)$$

In the following two lemmas (proved in Section 4.11.3) we show that both the terms in the above display are bounded by constant multiples (depending only on $L, \varepsilon, D, M_\varepsilon$ and M_0) of $\delta^{3/4} + n^{-1/2}\delta^{1/2}$. The following lemma (proved in Section 4.11.3.1) shows this for the first term of (4.79).

Lemma 48. *For every $\varepsilon > 0$, we have*

$$\log N_{[\,]}(\nu, \{f : f \in \mathcal{H}_{M_\varepsilon, L}(\delta)\}, L_2(P_{\theta_0, m_0})) \leq C_1^* \nu^{-1/2} \text{ and } \sup_{f \in \mathcal{H}_{M_\varepsilon, L}(\delta)} \|f\|_\infty \leq M_\varepsilon + M_0,$$

where $\mathcal{H}_{M_\varepsilon, L}(\delta)$ is defined in (4.78), C_1^* is a constant that depends only on $M_\varepsilon, L, D, T, M_0, d$, and the distribution of ϵ. Furthermore

$$\mathbb{E}^* \sup_{f \in \mathcal{H}_{M_\varepsilon, L}(\delta)} |\mathbb{G}_n[\epsilon f]| \lesssim C_{\varepsilon,1} \left(\delta^{3/4} + \frac{\delta^{1/2}}{\sqrt{n}}\right), \qquad (4.80)$$

where $C_{\varepsilon,1}$ is a constant depending only on $C_1^*, M_\varepsilon, M_0, d$, and the distribution of ϵ.

The following lemma (proved in Section 4.11.3.2) establishes a similar result for the second term of (4.79).

Lemma 49. *For every $\varepsilon > 0$,*

$$\sup_{f \in \mathfrak{H}_{M_\varepsilon, L}(\delta)} \|f\|_\infty \leq 4(M_\varepsilon + M_0)^2 \quad \textit{and} \quad \sup_{f \in \mathfrak{H}_{M_\varepsilon, L}(\delta)} \|f\| \leq 2(M_\varepsilon + M_0)\delta.$$

Furthermore,

$$\log N_{[\,]}(\nu, \mathfrak{H}_{M_\varepsilon, L}(\delta), L_2(P_{\theta_0, m_0})) \leq \left(\frac{M_\varepsilon + L\varnothing(D)}{\nu}\right)^{1/2}$$

and

$$\mathbb{E}^* \sup_{f \in \mathfrak{H}_{M_\varepsilon, L}(\delta)} |\mathbb{G}_n f| \leq C_{\varepsilon,2} \left(\delta^{3/4} + \frac{\delta^{1/2}}{\sqrt{n}}\right), \tag{4.81}$$

where $C_{\varepsilon,2}$ is a constant that depends only on $M_\varepsilon, L, D, T, M_0$, and d.

Now by applying the upper bounds (4.80) and (4.81) to (4.79), we have $\phi_n(\delta) = (C_{\varepsilon,1} + C_{\varepsilon,2})\left(\delta^{3/4} + n^{-1/2}\delta^{1/2}\right)$. Observe that $\phi(\delta)/\delta^{3/4}$ is a decreasing function of δ and

$$n^{4/5}\phi_n(n^{-2/5}) \leq \sqrt{n}.$$

Thus, by Lemma 46, we have $n^{2/5}\|\check{m}_n \circ \check{\theta}_n - m_0 \circ \theta_0\| = O_p^*(1)$.

4.11.2 Proof of Lemma 47

By the definition of $(\check{m}_n, \check{\theta}_n)$, we have

$$\sum_{i=1}^n (y_i - \check{m}_n(\check{\theta}_n^\top x_i))^2 \leq \sum_{i=1}^n (y_i - m(\check{\theta}_n^\top x_i))^2$$

for all function $m \in \mathcal{M}_L$. Since any constant function belongs to $\mathcal{M}_L$, for any fixed real κ, we have

$$\sum_{i=1}^n (y_i - \check{m}_n(\check{\theta}_n^\top x_i))^2 \leq \sum_{i=1}^n (y_i - \check{m}_n(\check{\theta}_n^\top x_i) + \kappa)^2.$$

A simplification of the above inequality gives us:

$$2\kappa \sum_{i=1}^n (y_i - \check{m}_n(\check{\theta}^\top x_i)) + n\kappa^2 \geq 0, \text{ for all } \kappa \Rightarrow \sum_{i=1}^n (y_i - \check{m}_n(\check{\theta}_n^\top x_i)) = 0. \tag{4.82}$$

Thus for any $t \in D$, we have

$$\begin{aligned}
|\check{m}_n(t)| &\leq \left|\check{m}_n(t) - \frac{1}{n}\sum_{j=1}^n \check{m}_n(\check{\theta}_n^\top x_j)\right| + \left|\frac{1}{n}\sum_{j=1}^n \check{m}_n(\check{\theta}_n^\top x_j)\right| \\
&\leq \frac{1}{n}\sum_{j=1}^n \left|\check{m}_n(t) - \check{m}_n(\check{\theta}_n^\top x_j)\right| + \left|\frac{1}{n}\sum_{j=1}^n \{m_0(\theta_0^\top x_j) + \epsilon_j\}\right| \quad \text{(by (4.82))} \\
&\leq \frac{1}{n}\sum_{j=1}^n L|t - \check{\theta}_n^\top x_j| + \frac{1}{n}\sum_{j=1}^n |m_0(\theta_0^\top x_j)| + \left|\frac{1}{n}\sum_{j=1}^n \epsilon_j\right| \\
&\leq L\varnothing(A) + M_0 + \left|\frac{1}{n}\sum_{j=1}^n \epsilon_j\right|,
\end{aligned}$$

where M_0 is the upper bound on m_0; see **(L1)**. The third inequality in the above display is true because $\check{m}_n$ is L Lipschitz. As ϵ is uniformly sub-gaussian, we have that $|\sum_{j=1}^n \epsilon_j/n| = O_p(1)$. Thus for every $\varepsilon > 0$, there exists a finite c_ε (depending only on the distribution of ϵ and ε) such that $\mathbb{P}(|\sum_{j=1}^n \epsilon_j/n| \geq c_\varepsilon) \leq \varepsilon$, for all n. Define $M_\varepsilon := L\varnothing(A) + M_0 + c_\varepsilon$. The lemma follows as we have

$$\mathbb{P}(\|\check{m}_n\|_\infty > M_\varepsilon) \leq \varepsilon, \quad \forall n.$$

4.11.3 Proofs of Lemmas 48 and 49

To prove Lemmas 48 and 49, we need the following entropy result.

Lemma 50. *Let*

$$\mathcal{H}_{M,L} := \{m \circ \theta - m_0 \circ \theta_0 : m \in \mathcal{M}_{M,L}, \theta \in \Theta\},$$

where $\mathcal{M}_{M,L}$ is defined in (4.76). Then there exists positive constants c and ν_0, such that, for every $M, L > 0$ and $\nu \leq \nu_0(M + L\varnothing(D))$

$$\log N_{[\,]}(\nu, \mathcal{H}_{M,L}, \|\cdot\|_\infty) = \log N_{[\,]}(\nu, \{m \circ \theta : (m, \theta) \in \mathcal{M}_{M,L} \times \Theta\}, \|\cdot\|_\infty) \leq K' \nu^{-1/2},$$

where K' is a constant depending only on M, L, T, D, and d.

Proof. To prove this lemma, we use the covering number for the class of uniformly bounded and uniformly Lipschitz convex functions obtained in [Guntuboyina and Sen, 2013].

Lemma 51 (Theorem 3.2, [Guntuboyina and Sen, 2013]). *Let $\mathcal{F}$ denote the class of real-valued convex functions defined on $[a,b]^d$ that are uniformly bounded in absolute value by B_0 and uniformly Lipschitz with constant L. Then there exists positive constants c and ν_0, depending only on the dimension d, such that for every $B_0, L > 0$ and $b > a$, we have*

$$\log N(\nu, \mathcal{F}, \|\cdot\|_\infty) \leq c\left(\frac{B_0 + L(b-a)}{\nu}\right)^{d/2}$$

for every $\nu \leq \nu_0(B_0 + L(b-a))$.

By Lemma 51 and Lemma 4.1 of [Pollard, 1990] for $\nu \in (0,1)$, we have

$$\log N_{[\,]}(\nu, \mathcal{M}_{M,L}, \|\cdot\|_\infty) \leq c\left(\frac{M + L\varnothing(D)}{\nu}\right)^{1/2},$$

$$\log N(\nu, \Theta, |\cdot|) \leq -c\log(\nu),$$

where c is a constant that depends only on d.

Recall that $\sup_{x\in\chi}|x| \leq T$; see **(A1)**. Let $\{\theta_1, \theta_2, \ldots, \theta_p\}$ be a $\nu/(2LT)$-cover (with respect to the Euclidean norm) of Θ and $\{m_1, m_2, \ldots, m_q\}$ be a $\nu/2$-cover (with respect to the $\|\cdot\|_\infty$-norm) for $\mathcal{M}_{M,L}$. In the following we will show that the set of functions $\{m_i \circ \theta_j - m_0 \circ \theta_0\}_{1\leq i\leq q, 1\leq j\leq p}$ form a ν-cover for $\mathcal{H}_{M,L}$ with respect to the $\|\cdot\|_\infty$-norm. For any given $m\circ\theta - m_0\circ\theta_0 \in \mathcal{H}_{M,L}$, we can get m_i and θ_j such that $\|m - m_i\|_\infty \leq \nu/2$ and $|\theta - \theta_j| \leq \nu/(2LT)$. Therefore, for any $x \in \chi$

$$\begin{aligned}|m(\theta^\top x) - m_i(\theta_j^\top x)| &\leq |m(\theta^\top x) - m(\theta_j^\top x)| + |m(\theta_j^\top x) - m_i(\theta_j^\top x)| \\ &\leq L|x||\theta - \theta_j| + \|m - m_i\|_\infty \leq \frac{L|x|\nu}{2LT} + \frac{\nu}{2} \leq \nu.\end{aligned}$$

Thus we have

$$\log N(\nu, \mathcal{H}_{M,L}, \|\cdot\|_\infty) \leq c\left[-\log(\nu) + \log(2LT) + 2\left(\frac{M + L\varnothing(D)}{\nu}\right)^{1/2}\right] \leq K'\nu^{-1/2}.$$

The result now follows as the covering number is equal to the bracketing number for the sup-norm. □

4.11.3.1 Proof of Lemma 48

Suppose $\mathcal{F}$ is a class of real valued functions defined on χ. We first present a result that gives a maximal inequality for the class of functions $\{\epsilon f : f \in \mathcal{F}\}$ in terms of the

bracketing entropy of $\mathcal{F}$, with respect the $L_2(P_{\theta_0,m_0})$ norm.

Lemma 52. *Suppose $\mathcal{F}$ is a class of functions (defined on $\mathcal{X}$) such that*

$$\sup_{f\in\mathcal{F}}\|f\|_\infty \le \Phi,\ \sup_{f\in\mathcal{F}}\|f\| \le \kappa,\ \textit{and}\ \log N_{[]}(\nu,\mathcal{F},\|\cdot\|) \le \Delta\nu^{-\alpha},$$

for some $0<\alpha<2$, where $\|f\|^2 := \int_{\mathcal{X}} f^2 dP_X$. Then

$$\log N_{[]}(K^*\nu,\epsilon\mathcal{F},\|\cdot\|_B) \le \Delta\nu^{-\alpha},$$

where for any $g\in L_2(P_{\theta_0,m_0})$, $\|g\|_B$ (Bernstein "norm") is defined as

$$\|g\|_B := \Big[2\mathbb{E}\Big(\exp(|g|)-1-|g|\Big)\Big]^{1/2},$$

$K^* := \sup_x\left(\mathbb{E}\left[\epsilon^2\exp(2\Phi|\epsilon|)|X=x\right]\right)^{1/2}$, *and* $\epsilon\mathcal{F} := \{\epsilon f : f\in\mathcal{F}\}$. *Furthermore $f\in\mathcal{F}$, $\|\epsilon f\|_B \le K^*\|f\|$ and*

$$\mathbb{E}^*\sup_{f\in\mathcal{F}}|\mathbb{G}_n\epsilon f| \lesssim \frac{\Delta^{1/2}K^*\kappa^{1-\alpha/2}}{(1-\alpha/2)} + \frac{\Delta\kappa^{-\alpha}}{\sqrt{n}\,(1-\alpha/2)^2}. \tag{4.83}$$

Proof. We will use the $\|\cdot\|$–bracket for $\mathcal{F}$ to form a $\|\cdot\|_B$–bracket for $\mathcal{F}$. Fix $f\in\mathcal{F}$. Observe that there exist $f_1, f_2 : \mathcal{X}\to[-\Phi,\Phi]$, such that

$$\|f_2-f_1\|\le\nu \text{ and } f_1(x)\le f(x)\le f_2(x),\quad \forall x\in\mathcal{X}. \tag{4.84}$$

Define $\epsilon^+ := \max\{\epsilon,0\}$ and $\epsilon^- := \max\{0,-\epsilon\}$. Multiplying ϵ^+ and ϵ^- to the second inequality in (4.84), we have

$$f_1(x)\epsilon^+ \le f(x)\epsilon^+ \le f_2(x)\epsilon^+ \quad\text{and}\quad -f_2(x)\epsilon^- \le -f(x)\epsilon^- \le -f_1(x)\epsilon^-,$$

respectively. Combining the above inequalities, we have

$$f_1(x)\epsilon^+ - f_2(x)\epsilon^- \le f(x)\epsilon \le f_2(x)\epsilon^+ - f_1(x)\epsilon^-.$$

Moreover,

$$\begin{aligned}
&\|f_2(X)\epsilon^+ - f_1(X)\epsilon^- - f_1(X)\epsilon^+ + f_2(X)\epsilon^-\|_B^2 \\
&= \|(f_2(X) - f_1(X))|\epsilon|\|_B^2 \\
&= 2\mathbb{E}\Big\{ \exp(|(f_2(X) - f_1(X))\epsilon|) - 1 - |(f_2(X) - f_1(X))\epsilon| \Big\} \\
&\leq \mathbb{E}\Big\{ (f_2(X) - f_1(X))^2\epsilon^2 \exp\big(|(f_2(X) - f_1(X))\epsilon|\big) \Big\} \\
&\leq \mathbb{E}\Big\{ (f_2(X) - f_1(X))^2\epsilon^2 \exp\big(2\Phi|\epsilon|\big) \Big\} \\
&= \mathbb{E}\Big\{ (f_2(X) - f_1(X))^2 \mathbb{E}\left[\epsilon^2 \exp(2\Phi|\epsilon|)|X\right] \Big\} \\
&\leq (K^*)^2 \|f_2 - f_1\|^2 \leq (K^*\nu)^2 ,
\end{aligned}$$

where K^* is as given in the statement of the lemma.

Thus if (f_1, f_2) is a ν-bracket (with respect to $\|\cdot\|$-norm) for f, then $(f_1\epsilon^+ - f_2\epsilon^-, f_2\epsilon^+ - f_2\epsilon^-)$ is a $K^*\nu$-bracket for ϵf (with respect to $\|\cdot\|_B$-norm). Therefore, we have

$$\log N_{[\,]}(K^*\nu, \epsilon\mathcal{F}, \|\cdot\|_B) \leq \log N_{[\,]}(\nu, \mathcal{F}, \|\cdot\|) \leq \Delta\nu^{-\alpha}.$$

To prove (4.83), we use the following Lemma.

Lemma 53 (Lemma 3.4.3 of [van der Vaart, 1996]). *Let $\mathcal{G}$ be a class of measurable functions such that $\sup_{g\in\mathcal{G}} \|g\|_B \leq \rho$. Then*

$$\mathbb{E}^* \sup_{g\in\mathcal{G}} |\mathbb{G}_n g| \lesssim J_{[\,]}(\rho, \mathcal{G}, \|\cdot\|_B) \left(1 + \frac{J_{[\,]}(\rho, \mathcal{G}, \|\cdot\|_B)}{\rho^2\sqrt{n}}\right). \tag{4.85}$$

We now find an upper bound for $\sup_{f\in\mathcal{F}} \|\epsilon f\|_B$. Observe that

$$\begin{aligned}
\|\epsilon f\|_B^2 &= 2\mathbb{E}\Big\{ \exp(|\epsilon f(X)|) - 1 - |\epsilon f(X)| \Big\} \\
&\leq \mathbb{E}\Big\{ \epsilon^2 f^2(X) \exp\big(|f(X)\epsilon|\big) \Big\} \\
&\leq \mathbb{E}\Big\{ \epsilon^2 f^2(X) \exp\big(\Phi|\epsilon|\big) \Big\} \\
&\leq \mathbb{E}\Big\{ f^2(X) \mathbb{E}\left[\epsilon^2 \exp(2\Phi|\epsilon|)|X\right] \Big\} \leq (K^*)^2 \|f\|^2 \leq (K^*\kappa)^2 .
\end{aligned}$$

Thus, for the class $\epsilon\mathcal{F}$, we can apply Lemma 53 with $\rho = K^*\kappa$. By definition

$$J_{[\,]}(K^*\kappa, \{\epsilon f : f \in \mathcal{F}\}, \|\cdot\|_B) \leq (K^*)^{\alpha/2} \int_0^{K^*\kappa} \sqrt{\Delta\nu^{-\alpha}}\, d\nu = \Delta^{1/2} K^* \kappa^{1-\alpha/2}/(1 - \alpha/2).$$

Therefore by (4.85), we have

$$\mathbb{E}^* \sup_{f \in \mathcal{F}} |\mathbb{G}_n[\epsilon f]| \lesssim \frac{\Delta^{1/2} K^* \kappa^{1-\alpha/2}}{(1-\alpha/2)} + \frac{\Delta \kappa^{-\alpha}}{\sqrt{n}\,(1-\alpha/2)^2}. \qquad \square$$

Th proof of Lemma 48 will now be completed by a simple application of Lemma 52 with $\mathcal{F} = \mathcal{H}_{M_\varepsilon,L}(\delta)$. By definition (4.78), we have

$$\sup_{f \in \mathcal{H}_{M_\varepsilon,L}(\delta)} \|f\|_\infty < M_\varepsilon + M_0 \text{ and } \sup_{f \in \mathcal{H}_{M_\varepsilon,L}(\delta)} \|f\| < \delta.$$

As $\mathcal{H}_{M_\varepsilon,L}(\delta) \subset \mathcal{H}_{M_\varepsilon,L}$, by Lemma 50, we have

$$\log N_{[\,]}(\nu, \mathcal{H}_{M_\varepsilon,L}(\delta), \|\cdot\|_\infty) \le \log N_{[\,]}(\nu, \mathcal{H}_{M_\varepsilon,L}, \|\cdot\|_\infty) \le K' \nu^{-1/2}.$$

Thus

$$\log N_{[\,]}(\nu, \mathcal{H}_{M_\varepsilon,L}(\delta), \|\cdot\|) \le C_1^* \nu^{-1/2},$$

where $C_1^* = \sqrt{2}K'$. By applying Lemma 52 (see (4.83)) with

$$\Phi = M_\varepsilon + M_0,\ \kappa = \delta,\ \Delta = C_1^*, \text{ and } \alpha = 1/2,$$

we have

$$\mathbb{E}^* \sup_{f \in \mathcal{H}_{M_\varepsilon,L}(\delta)} |\mathbb{G}_n[\epsilon f]| \le C_{\varepsilon,1} \left(\delta^{3/4} + \frac{\delta^{-1/2}}{\sqrt{n}} \right),$$

where $C_{\varepsilon,1}$ is constant depending only on $K', M_\varepsilon, M_0, L, d$, and T.

4.11.3.2 Proof of Lemma 49

We proceed as in the proof of Lemma 48. For any function $f \in \mathcal{H}_{M_{\varepsilon,L}}$, there exist functions $f_1, f_2 : \mathcal{X} \to [-M_\varepsilon - M_0, M_0 + M_\varepsilon]$ such that $f_1(x) \le f(x) \le f_2(x)$ and $0 \le f_2(x) - f_1(x) \le \nu$ for each $x \in \mathcal{X}$. Observe that for any two real numbers $x \le y$, we have $x^+ \le y^+$ and $y^- \le x^-$. Thus, we have

$$f_1^+ \le f^+ \le f_2^+ \text{ and } f_2^- \le f^- \le f_1^-.$$

The above inequalities lead to a bracket for f^2. Observe that

$$f_1^+ + f_2^- \le |f| \le f_1^- + f_2^+ \Rightarrow (f_1^+ + f_2^-)^2 \le f^2 \le (f_1^- + f_2^+)^2.$$

The difference between the bracket functions is

$$\begin{aligned}(f_1^- + f_2^+)^2 - (f_1^+ + f_2^-)^2 &= (f_1^- - f_1^+ + f_2^+ - f_2^-)(f_1^- + f_2^+ + f_1^+ + f_2^-) \\ &= (f_2 - f_1)(|f_2| + |f_1|) \\ &\leq 2(M_\varepsilon + M_0)(f_2 - f_1) \leq 2(M_\varepsilon + M_0)\nu.\end{aligned}$$

Thus, if $[f_1, f_2]$ is a ν-bracket (with respect to the $\|\cdot\|_\infty$-norm) for f then $[(f_1^+ + f_2^-)^2, (f_1^- + f_2^+)^2]$ is a $(2M_\varepsilon + 2M_0)\nu$-bracket (with respect to the $\|\cdot\|_\infty$-norm) for f^2. Therefore, we have

$$\log N_{[\,]}(\nu, \mathfrak{H}_{M_\varepsilon,L}(\delta), \|\cdot\|_\infty) \leq \log N_{[\,]}(\nu/(2M_\varepsilon + 2M_0), \mathcal{H}_{M_\varepsilon,L}, \|\cdot\|_\infty) \leq C_2^* \nu^{-1/2}.$$

Thus

$$J_{[\,]}(\rho, \mathfrak{H}_{M_\varepsilon,L}(\delta), \|\cdot\|) \leq \int_0^{2(M_\varepsilon+M_0)\delta} \sqrt{C_2^* \nu^{-1/2}} d\nu \leq \frac{8}{3}\sqrt{C_2^*}\big[(M_\varepsilon + M_0)\delta\big]^{3/4}.$$

To complete the proof we use the following Lemma.

Lemma 54 (Lemma 3.4.2 of [van der Vaart, 1996]). *Let $\mathcal{G}$ be class of measurable functions such that $Pg^2 < \rho^2$ and $\|g\|_\infty \leq M$ for every g in $\mathcal{G}$. Then*

$$\mathbb{E}^* \sup_{g\in\mathcal{G}} |\mathbb{G}_n g| \lesssim J_{[\,]}(\rho, \mathcal{G}, \|\cdot\|) \left(1 + \frac{J_{[\,]}(\rho, \mathcal{G}, L_2(P))}{\rho^2\sqrt{n}} M\right).$$

Note that for every function $f \in \mathfrak{H}_{M_{\varepsilon,L}(\delta)}$, we have $0 \leq f(x) \leq 4(M_\varepsilon + M_0)^2$ for all $x \in \mathcal{X}$. Furthermore, we have

$$\mathbb{E}f^2 \leq \|f\|_\infty \mathbb{E}f \leq 4(M_\varepsilon + M_0)^2\delta^2.$$

Observe that

$$J_{[\,]}(\rho, \mathfrak{H}_{M_\varepsilon,L}(\delta), L_2(P)) \leq \int_0^{2(M_\varepsilon+M_0)\delta} \sqrt{C_2^* \nu^{-1/2}} d\nu \leq \frac{8}{3}\sqrt{C_2^*}\big[(M_\varepsilon + M_0)\delta\big]^{3/4}.$$

Thus by Lemma 54, we have

$$\mathbb{E}^* \sup_{f\in\mathcal{H}_{M_\varepsilon,L}(\delta)} |\mathbb{G}_n f| \leq C_{\varepsilon,2}\left(\delta^{3/4} + \frac{\delta^{-1/2}}{\sqrt{n}}\right).$$

4.11.4 Proof of Theorem 26

By Ascoli-Arzelà theorem, the sequence $\{\check{m}_n\}$ has a uniformly converging subsequence if $\{\check{m}_n\}$ belongs to a class of closed, bounded, and equicontinuous functions. In the proof of Theorem 25, we showed that $\|\check{m}_n\|_\infty = O_p(1)$. Moreover, as $\check{m}_n \in \mathcal{C}_L$ the conditions of Ascoli-Arzelà theorem are satisfied and every subsequence $\{\check{m}_{n_k}\}$ has a further subsequence $\{\check{m}_{n_{k_l}}\}$ such that $\|\check{m}_{n_{k_l}} - m_1\|_{D_0} \to 0$, for some function m_1. Furthermore, as $|\check{\theta}_n| \leq 1$, we have that every subsequence $\{\check{\theta}_{n_k}\}$ has a further subsequence $\{\check{\theta}_{n_{k_l}}\}$ such that $|\check{\theta}_{n_{k_l}} - \theta_1| \to 0$, for some θ_1 in Θ. Now, observe that continuity and almost everywhere differentiability of the link functions imply that $\|m_1 \circ \theta_1 - m_0 \circ \theta_0\| = 0$ is equivalent to $m_1 \equiv m_0$ and $\theta_1 = \theta_0$.Thus we have that $\|\check{m}_n - m_0\|_{D_0} = o_p(1)$. Now the final result follows by an application of the following lemma and a standard subsequence argument.

Lemma 55 (Lemma 3.10, [Seijo and Sen, 2011])**.** *Let $\mathcal{C}$ be an open convex subset of $\mathbb{R}^d$ and f a convex functions which is continuous and differentiable on $\mathcal{C}$. Consider a sequence of convex functions $\{f_n\}$ which are finite on $\mathcal{C}$ such that $f_n \to f$ pointwise on $\mathcal{C}$. Then, if $C \subset \mathcal{C}$ is any compact set,*

$$\sup_{\substack{x \in C \\ \xi \in \partial f_n(x)}} |\xi - \nabla f(x)| \to 0,$$

where $\partial f_n(x)$ represents the subdifferential set of f_n at x.

4.11.5 Proof of Theorem 27

We first state and prove a intermediary lemma.

Lemma 56. *Let m_0 and θ_0 satisfy the assumption **(A1)**, **(A5)**, and **(L1)**. Let $\{\theta_n\} \in \Theta$ and $\{m_n\} \in \mathcal{C}_L$ be two non-random sequences such that*

$$|\theta_n - \theta_0| \to 0, \qquad \|m_n - m_0\|_{D_0} \to 0, \quad \text{and} \quad \|m_n' - m_0'\|_C \to 0 \tag{4.86}$$

for any compact subset C of the interior of D_0. Then

$$P_X \left| m_n(\theta_n^\top X) - m_0(\theta_0^\top X) - \{m_0'(\theta_0^\top X) X^\top (\theta_n - \theta_0) + (m_n - m_0)(\theta_0^\top X)\} \right|^2 = o(|\theta_n - \theta_0|^2).$$

Proof. For any convex function $f \in \mathcal{C}_L$, denote the right derivative of f by f'. Note that f' is a bounded increasing function. First, observe that

$$\begin{aligned} m_n(\theta_n^\top x) - m_0(\theta_0^\top x) - \left[m_0'(\theta_0^\top x)x^\top(\theta_n - \theta_0) + (m_n - m_0)(\theta_0^\top x)\right] \\ = m_n(\theta_n^\top x) - m_n(\theta_0^\top x) - m_0'(\theta_0^\top x)x^\top(\theta_n - \theta_0). \end{aligned}$$

Now,

$$\begin{aligned} &\left|m_n(\theta_n^\top x) - m_n(\theta_0^\top x) - m_0'(\theta_0^\top x)x^\top(\theta_n - \theta_0)\right|^2 \\ &= \left|\int_{\theta_n^\top x}^{\theta_0^\top x} m_n'(t)dt - m_0'(\theta_0^\top x)x^\top(\theta_n - \theta_0)\right|^2 \quad (m_n \text{ is convex and hence absolutely continuous}) \\ &= \left|\int_{\theta_n^\top x}^{\theta_0^\top x} m_n'(t)dt - m_n'(\theta_0^\top x)x^\top(\theta_n - \theta_0) + m_n'(\theta_0^\top x)x^\top(\theta_n - \theta_0) - m_0'(\theta_0^\top x)x^\top(\theta_n - \theta_0)\right|^2 \\ &= \left|\int_{\theta_n^\top x}^{\theta_0^\top x} m_n'(t)dt - m_n'(\theta_0^\top x)x^\top(\theta_n - \theta_0) + (m_n' - m_0')(\theta_0^\top x)x^\top(\theta_n - \theta_0)\right|^2 \\ &\le 2\left|\int_{\theta_n^\top x}^{\theta_0^\top x} m_n'(t)dt - m_n'(\theta_0^\top x)x^\top(\theta_n - \theta_0)\right|^2 + 2\left|(m_n' - m_0')(\theta_0^\top x)x^\top(\theta_n - \theta_0)\right|^2. \quad (4.87) \end{aligned}$$

We will now find an upper bound for the first term on the right hand side of the above display. Observe that m_n' is an increasing function. When $x^\top\theta_n \neq x^\top\theta_0$, we have

$$m_n'(\theta_n^\top x) \wedge m_n'(\theta_0^\top x) \le \frac{\int_{\theta_n^\top x}^{\theta_0^\top x} m_n'(t)dt}{x^\top(\theta_n - \theta_0)} \le m_n'(\theta_n^\top x) \vee m_n'(\theta_0^\top x).$$

Thus for all $x \in \mathcal{X}$, we have

$$\left|\int_{\theta_n^\top x}^{\theta_0^\top x} m_n'(t)dt - m_n'(\theta_0^\top x)x^\top(\theta_n - \theta_0)\right| \le |m_n'(\theta_n^\top x) - m_n'(\theta_0^\top x)||x^\top(\theta_n - \theta_0)|. \quad (4.88)$$

Note that if $x^\top\theta_n = x^\top\theta_0$, then both sides of (4.88) are 0. Combine (4.87) and (4.88), to conclude that

$$\begin{aligned} &P_X\left|m_n(\theta_n^\top X) - m_n(\theta_0^\top X) - m_0'(\theta_0^\top X)X^\top(\theta_n - \theta_0)\right|^2 \quad (4.89) \\ &\le 2P_X\left|(m_n'(\theta_n^\top X) - m_n'(\theta_0^\top X))X^\top(\theta_n - \theta_0)\right|^2 + 2P_X\left|(m_n' - m_0')(\theta_0^\top X)X^\top(\theta_n - \theta_0)\right|^2. \end{aligned}$$

As $\mathcal{X}$ is bounded, the two terms on the right hand side of (4.89) can be bounded as

$$\begin{aligned} P_X\left|(m_n'(\theta_n^\top X) - m_n'(\theta_0^\top X))X^\top(\theta_n - \theta_0)\right|^2 &\le T^2|\theta_n - \theta_0|^2 P_X\left|m_n'(\theta_n^\top X) - m_n'(\theta_0^\top X)\right|^2, \\ P_X\left|(m_n' - m_0')(\theta_0^\top X)x^\top(\theta_n - \theta_0)\right|^2 &\le T^2|\theta_n - \theta_0|^2 P_X\left|(m_n' - m_0')(\theta_0^\top X)\right|^2. \end{aligned}$$

We will now show that both $P_X \left|m_n'(\theta_n^\top X) - m_n'(\theta_0^\top X)\right|^2$ and $P_X \left|(m_n' - m_0')(\theta_0^\top X)\right|^2$ converge to 0 as $n \to \infty$. First observe that

$$P_X \left|m_n'(\theta_n^\top X) - m_n'(\theta_0^\top X)\right|^2 \lesssim P_X \left|m_n'(\theta_n^\top X) - m_0'(\theta_n^\top X)\right|^2 + P_X \left|m_0'(\theta_n^\top X) - m_0'(\theta_0^\top X)\right|^2 + P_X \left|m_0'(\theta_0^\top X) - m_n'(\theta_0^\top X)\right|^2. \tag{4.90}$$

Recall that m_0' is a continuous and bounded function; see assumption **(S1)**. Bounded convergence theorem now implies that $P_X \left|m_0'(\theta_n^\top X) - m_0'(\theta_0^\top X)\right|^2 \to 0$, as $|\theta_n - \theta_0| \to 0$. Now consider the first term on the right hand side of (4.90). As $\theta_0^\top X$ has a density, for any $\varepsilon > 0$, we can define a compact subset C_ε in the interior of D_0 such that $\mathbb{P}(\theta_0^\top X \notin C_\varepsilon) < \varepsilon/4L$. Now note that, by Theorem 26 and the fact that $\mathbb{P}(\theta_n^\top X \notin C_\varepsilon) \to \mathbb{P}(\theta_0^\top X \notin C_\varepsilon)$, we have

$$P_X \left|m_n'(\theta_n^\top X) - m_0'(\theta_n^\top X)\right|^2 \leq \sup_{t \in C_\varepsilon} |m_n'(t) - m_0(t)| + 2LP(\theta_n^\top X \notin C_\varepsilon) \leq \varepsilon,$$

as $n \to \infty$. Similarly, we can see that

$$P_X \left|m_0'(\theta_0^\top X) - m_n'(\theta_0^\top X)\right|^2 \leq \sup_{t \in C_\varepsilon} |m_n'(t) - m_0(t)| + 2LP(\theta_0^\top X \notin C_\varepsilon) \leq \varepsilon,$$

as $n \to \infty$. Combining the results, we have shown that for every $\varepsilon > 0$

$$P_X \left|m_n(\theta_n^\top X) - m(\theta_0^\top X) - m_0'(\theta_0^\top X) X^\top (\theta_n - \theta_0)\right|^2 \leq T^2 |\theta_n - \theta_0|^2 \varepsilon,$$

for all sufficiently large n. Thus the result follows. □

We will now use the above lemma to prove Theorem 23. Let us define, $A_n(x) := \check{m}_n(\check{\theta}_n^\top x) - m_0(\theta_0^\top x)$ and $B_n(x) := m_0'(\theta_0^\top x) x^\top (\check{\theta}_n - \theta_0) + (\check{m}_n - m_0)(\theta_0^\top x)$. Observe that

$$\begin{aligned} A_n(x) - B_n(x) &= \check{m}_n(\check{\theta}_n^\top x) - m_0'(\theta_0^\top x) x^\top (\check{\theta}_n - \theta_0) - \check{m}_n(\theta_0^\top x). \\ &= \check{m}_n(\theta_n^\top x) - m_0(\theta_0^\top x) - \{m_0'(\theta_0^\top x) x^\top (\theta_n - \theta_0) + (\check{m}_n - m_0)(\theta_0^\top x)\}. \end{aligned}$$

We will now show that

$$D_n := \frac{1}{|\check{\theta}_n - \theta_0|^2} P_X |A_n(X) - B_n(X)|^2 = o_p(1). \tag{4.91}$$

It is equivalent to show that for every subsequence $\{D_{n_k}\}$, there exists a further subsequence $\{D_{n_{k_l}}\}$ that converges to 0 almost surely; see Theorem 2.3.2 of [Durrett, 2010].

We showed in Theorem 26, that $\{\check{m}_n, \check{\theta}_n\}$ satisfies (4.86) in probability. Thus by another application of Theorem 2.3.2 of [Durrett, 2010], we have that $\{\check{m}_{n_k}, \check{\theta}_{n_k}\}$ has a further subsequence $\{\check{m}_{n_{k_l}}, \check{\theta}_{n_{k_l}}\}$ that satisfies (4.86) almost surely. Thus by Lemma 43, we have $D_{n_{k_l}} \overset{a.s.}{\to} 0$. Thus $D_n = o_p(1)$.

We will now use (4.91) to find the rate of convergence of $\{\check{m}_n, \check{\theta}_n\}$. We first find an upper bound for $P_X|B_n(X)|^2$. By a simple application of triangle inequality and (4.91), we have

$$P_X|A_n(X)|^2 \geq \frac{1}{2}P_X|B_n(X)|^2 - P_X|A_n(X) - B_n(X)|^2 \geq \frac{1}{2}P_X|B_n(X)|^2 - o_p(|\check{\theta}_n - \theta_0|^2).$$

Note that, by Theorem 25, we have that $P_X|A_n(X)|^2 = O_p(n^{-4/5})$. Thus we have

$$P_X\left|m_0'(\theta_0^\top X)X^\top(\check{\theta}_n - \theta_0) + (\check{m}_n - m_0)(\theta_0^\top X)\right|^2 \leq O_p(n^{-4/5}) + o_p(|\check{\theta}_n - \theta_0|^2).$$

Now define

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

and note that by assumption **(A3)** there exists a $\lambda_1 > 0$ such that

$$P_X g_1^2 = (\check{\theta}_n - \theta_0)^\top P_X[XX^\top |m_0'(\theta_0^\top X)|^2](\check{\theta}_n - \theta_0) \geq \lambda_1|\check{\theta}_n - \theta_0|^2. \tag{4.92}$$

With (4.92) in mind, we can see that proof of this theorem will be complete if we can show that

$$P_x g_1^2 + P_X g_2^2 \lesssim P_X\left|m_0'(\theta_0^\top X)X^\top(\check{\theta}_n - \theta_0) + (\check{m}_n - m_0)(\theta_0^\top X)\right|^2. \tag{4.93}$$

Lemma 44 gives a sufficient condition for (4.93). We now show that g_1 and g_2 satisfy

the condition of Lemma 44. Observe that

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

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

4.11.6 Proof of Theorem 28

We first show (4.8). Let δ_n be a sequence of positive numbers decreasing to 0. Let $a, b \in \mathbb{R}$ such that $D_0 = [a, b]$. Define $C_n := [a + 2\delta_n, b - 2\delta_n]$. In this sub-section, let K and K' denote the minimum and the maximum of the density $f_{\theta_0^\top X}(t)$ over $t \in D_0$. Also, let κ denote the bound on $m_0''(t)$ over $t \in D_0$. As the $\check m$ is a convex function, we have

$$\frac{\check m(t)-\check m(t-\delta_n)}{\delta_n} \le \check m'(t-) \le \check m'(t+) \le \frac{\check m(t+\delta_n)-\check m(t)}{\delta_n},$$

for all $t \in C_n$, where $\check m'(t+)$ and $\check m'(t-)$ denote the right and left derivatives of $\check m$ at t, respectively. By Theorem 23 and assumption **(A5)**, we have

$$\begin{aligned}
&\int_{t\in C_n}\left[\frac{\check m(t+\delta_n)-\check m(t)}{\delta_n}-\frac{m_0(t+\delta_n)-m_0(t)}{\delta_n}\right]^2dt \\
&= \frac{2}{\delta_n^2}\int_{t\in C_n}\{\check m(t+\delta_n)-m_0(t+\delta_n)\}^2dt+\frac{2}{\delta_n^2}\int_{t\in C_n}\{\check m(t)-m_0(t)\}^2dt \\
&= \frac{2}{\delta_n^2}\int_{t\in[a+3\delta_n,b-\delta_n]}\{\check m(t)-m_0(t)\}^2dt+\frac{2}{\delta_n^2}\int_{t\in C_n}\{\check m(t)-m_0(t)\}^2dt \\
&\le \frac{2}{K\delta_n^2}\int_{t\in[a+3\delta_n,b-\delta_n]}\{\check m(t)-m_0(t)\}^2f_{\theta_0^\top X}(t)dt+\frac{2}{K\delta_n^2}\int_{t\in C_n}\{\check m(t)-m_0(t)\}^2f_{\theta_0^\top X}(t)dt \\
&= \frac{1}{\delta_n^2}O_p(n^{-4/5}). \qquad (4.94)
\end{aligned}$$

Similarly, it can be shown that

$$\int_{t\in C_n}\left[\frac{\check{m}(t)-\check{m}(t-\delta_n)}{\delta_n}-\frac{m_0(t)-m_0(t-\delta_n)}{\delta_n}\right]^2 dt=\frac{1}{\delta_n^2}O_p(n^{-4/5}). \tag{4.95}$$

Now observe that

$$\begin{aligned}\Delta_n^+(t):=\left[\frac{\check{m}(t+\delta_n)-\check{m}(t)}{\delta_n}-\frac{m_0(t+\delta_n)-m_0(t)}{\delta_n}\right] &\geq \check{m}'(t+)-m_0'(x_{t_n})\\ &\geq \check{m}'(t+)-m_0'(t)+m_0'(t)-m_0'(x_{t_n})\\ &\geq \check{m}'(t+)-m_0'(t)-\kappa\delta_n\end{aligned}$$

where x_{t_n} lies between t and $t+\delta_n$ and $\kappa\geq\|m_0''\|_{D_0}$. Moreover,

$$\begin{aligned}\Delta_n^-(t):=\left[\frac{\check{m}(t)-\check{m}(t-\delta_n)}{\delta_n}-\frac{m_0(t)-m_0(t-\delta_n)}{\delta_n}\right] &\leq \check{m}'(t+)-m_0'(x'_{t_n})\\ &\leq \check{m}'(t+)-m_0'(t)+m_0'(t)-m_0'(x'_{t_n})\\ &\leq \check{m}'(t+)-m_0'(t)+\kappa\delta_n,\end{aligned}$$

where x'_{t_n} lies between $t-\delta_n$ and t and $\kappa\geq\|m_0''\|_{D_0}$. Combining the above two results, we have

$$\Delta_n^-(t)-\kappa\delta_n\leq\check{m}'(t+)-m_0'(t)\leq\Delta_n^+(t)+\kappa\delta_n$$

Thus for every $t\in C_n$, we have $[\check{m}'(t+)-m_0'(t)]^2\leq 2\kappa^2\delta_n^2+2\max\left\{[\Delta_n^-(t)]^2,[\Delta_n^+(t)]^2\right\}$. By (4.94) and (4.95), we have

$$\int_{t\in C_n}[\check{m}'(t+)-m_0'(t)]^2 f_{\theta_0^\top X}(t)dt\leq 2\kappa^2\delta_n^2+\frac{1}{\delta_n^2}O_p(n^{-4/5})$$

as

$$\begin{aligned}&\int_{t\in C_n}\max\left\{[\Delta_n^-(t)]^2,[\Delta_n^+(t)]^2\right\}f_{\theta_0^\top X}(t)dt\\ &\leq\int_{t\in C_n}\{\Delta_n^-(t)\}^2 f_{\theta_0^\top X}(t)dt+\int_{t\in C_n}\{\Delta_n^+(t)\}^2 f_{\theta_0^\top X}(t)dt\\ &\leq K'\int_{t\in C_n}\{\Delta_n^-(t)\}^2dt+K'\int_{t\in C_n}\{\Delta_n^+(t)\}^2dt\\ &=\frac{1}{\delta_n^2}O_p(n^{-4/5}),\end{aligned}$$

where K' is an upper bound on $f_{\theta^\top X}$. Moreover, note that $\|\check{m}'\|_\infty \leq L$, $\|m_0'\|_\infty \leq L$. Thus

$$\begin{aligned}\int_{t\in D_0}\{\check{m}'(t+)-m_0'(t)\}^2 f_{\theta_0^\top X}(t)dt &= \int_{t\in C_n}\{\check{m}'(t+)-m_0'(t)\}^2 f_{\theta_0^\top X}(t)dt \\ &\quad + \int_{t\in D_0\cap C_n^c}\{\check{m}'(t+)-m_0'(t)\}^2 f_{\theta_0^\top X}(t)dt \\ &= 2\kappa^2\delta_n^2 + \frac{1}{\delta_n^2}O_p(n^{-4/5}) + 4L^2P(\theta_0^\top X\in D_0\cap C^c) \\ &\leq 2\kappa^2\delta_n^2 + \frac{1}{\delta_n^2}O_p(n^{-4/5}) + 8K'L^2\delta_n.\end{aligned}$$

Now choose $\delta_n = n^{-4/15}$. With this choice of δ_n, we have

$$\int_{t\in D_0}\{\check{m}'(t+)-m_0'(t)\}^2 f_{\theta_0^\top X}(t)dt \leq 2\kappa^2 n^{-8/15}+O_p(n^{-4/15})+4K'L^2n^{-4/15} = O_p(n^{-4/15}).$$

We can find a similar upper bound for $\int_{t\in D_0}\{\check{m}'(t+)-m_0'(t)\}^2dt$. Observe,

$$\begin{aligned}\int_{t\in D_0}\{\check{m}'(t+)-m_0'(t)\}^2dt &= \int_{t\in C_n}\{\check{m}'(t+)-m_0'(t)\}^2dt + \int_{t\in D_0\cap C_n^c}\{\check{m}'(t+)-m_0'(t)\}^2dt \\ &\leq 2\kappa^2\delta_n^2 + \frac{1}{\delta_n^2}O_p(n^{-4/5}) + 8L^2\delta_n.\end{aligned}$$

Now choose $\delta_n = n^{-4/15}$. With this choice of δ_n, we have

$$\int_{t\in D_0}\{\check{m}'(t+)-m_0'(t)\}^2dt \leq 2\kappa^2 n^{-8/15} + O_p(n^{-4/15}) + 4K'L^2n^{-4/15} = O_p(n^{-4/15}).$$

Now to prove (4.9), define $D_n := D_{\check{\theta}_n} = [a_n, b_n]$, for some a_n, b_n. As $|\check{\theta}_n - \theta_0| = O_p(n^{-2/5})$, we have that both $|a_n - a|$ and $|b_n - b|$ are $O_p(n^{-2/5})$. Thus $|a_n - a| = o_p(\delta_n)$ and due to assumption **(A4)**, the proof of (4.9) follows similarly to that of (4.8).

4.12 Proofs of results in Section 4.5.2

4.12.1 Proof of Theorem 30

We will first show that $\xi_t(u;\theta,\eta,m)$ is a valid submodel. Let us define

$$\psi_{t,\theta,\eta}(u) := \phi_{\theta,\eta,t}(u + (\theta - \zeta_t(\theta,\eta))^\top k(u)). \tag{4.96}$$

Note that to prove that $\xi_t(u;\theta,\eta,m)$ is a convex function it is enough to show that $\beth_t(\cdot;\theta,\eta,m)$ is an increasing function. Recall that k is a Lipschitz function on D_r. Thus

$u \mapsto u + (\theta - \zeta_t(\theta,\eta))^\top k(u)$ is a strictly increasing function for t in neighborhood of zero. As $\phi_{\theta,\eta,t}(\cdot)$ is a strictly increasing function for t sufficiently close to zero. It now follows that $u \mapsto \psi_{t,\theta,\eta}(u)$ is a nondecreasing function for all $t \in \mathbb{R}^d$ such that $|t-\theta|$ is sufficiently close to zero. Finally, recall that m' is an increasing function and

$$\beth_t(u;\theta,\eta,m) = m' \circ \psi_{t,\theta,\eta}(u).$$

Thus we have that $\beth_t(\cdot;\theta,\eta,m)$ is an increasing function for $t \in \mathbb{R}^d$ such that $|t-\theta|$ is small enough.

Next we show that $\xi_t(u;\theta,\eta,m) = m(u)$ when $t = \theta$. By definition we have

$$\xi_t(s^\top x;\theta,\eta,m) = \int_{s_0}^{s^\top x} m' \circ \psi_{t,\theta,\eta}(u)dy + (\zeta_t(\theta,\eta)-\theta)^\top \Big[(m_0'(s_0) - m'(s_0))k(s_0) - m_0'(s_0)h_{\theta_0}(s_0)\Big] + m(s_0).$$

We have that $\psi_{0,\theta,\eta}(u) = u$, $\forall u \in D$. Hence,

$$\xi_\theta(\theta,m)(\theta^\top x) = \int_{s_0}^{\theta^\top x} m' \circ \psi_{0,\theta,\eta}(y)dy + m(s_0) = \int_{s_0}^{\theta^\top x} m'(y)dy + m(s_0) = m(\theta^\top x).$$

Now we show that $J^2(\xi_t(\cdot;\theta,\eta,m)) < \infty$. Observe that

$$\begin{aligned} J^2(\xi_t(\cdot;\theta,\eta,m)) &= \int_D \{\xi_t''(u;\theta,\eta,m)(u)\}^2 du \\ &= \int_{D_r} \big[\frac{\partial}{\partial u}\beth_t(u;\theta,\eta,m)\big]^2 du \\ &= \int_D \{m'' \circ \psi_{t,\theta,\eta}'(u)\}^2 du \\ &= \int_D \{m''(u)\}^2 \psi_{t,\theta,\eta}' \circ \psi_{t,\theta,\eta}^{-1}(u) du \end{aligned}$$

where $\psi_{t,\theta,\eta}'$ is defined in (4.96) and $\psi_{t,\theta,\eta}'(u) := \frac{\partial}{\partial u}\psi_{t,\theta,\eta}'(u)$. Thus, we have that $J^2(\xi_t(\theta,m)) < \infty$ whenever $J(m) < \infty$.

Next we compute $\partial\beth_t(u;\theta,\eta,m)/\partial t$ and $\partial\xi_t(\zeta_t(\theta,\eta)^\top x;\theta,\eta,m)/\partial t$ to help with the calculation of the score function for the submodel $\{\zeta_t(\theta,\eta), \xi_t(\cdot;\theta,\eta,m)\}$. Observe that

$$\begin{aligned} \frac{\partial}{\partial t}\beth_t(u;\theta,\eta,m) &= \frac{\partial}{\partial t} m' \circ \phi_{\theta,\eta,t}(u + (\theta - \zeta_t(\theta,\eta))^\top k(u)) \\ &= m'' \circ \circ \phi_{\theta,\eta,t}(u + (\theta - \zeta_t(\theta,\eta))^\top k(u)) \Big[\dot\phi_{\theta,\eta,t}(u + (\theta - \zeta_t(\theta,\eta))^\top k(u)) \\ &\qquad - \phi_{\theta,\eta,t}'(u + (\theta - \zeta_t(\theta,\eta))^\top k(u)) \frac{\partial \zeta_t(\theta,\eta)}{\partial t}^\top k(u)\Big], \end{aligned}$$

where $\dot{\phi}_{\theta,\eta,t}(u) := \partial\phi_{\theta,\eta,t}(u)/\partial t$ and $\phi'_{\theta,\eta,t}(u) := \partial\phi_{\theta,\eta,t}(u)/\partial u$. Moreover,

$$
\begin{aligned}
&\frac{\partial}{\partial t}\xi_t(\zeta_t(\theta,\eta)^\top x;\theta,\eta,m)\\
&= \frac{\partial}{\partial t}\left\{\int_{s_0}^{\zeta_t(\theta,\eta)^\top x} \urcorner_t(y;\theta,\eta,m)dy\right\} + \frac{\partial\zeta_t(\theta,\eta)}{\partial t}^\top\Big[(m_0'(s_0)-m'(s_0))k(s_0) - m_0'(s_0)h_{\theta_0}(s_0)\Big]\\
&= \urcorner_t(\zeta_t(\theta,\eta)^\top x;\theta,\eta,m)\frac{\partial\zeta_t(\theta,\eta)}{\partial t}^\top x + \int_{s_0}^{\zeta_t(\theta,\eta)^\top x}\frac{\partial\urcorner_t(y;\theta,\eta,m)}{\partial t}dy\\
&\qquad + \frac{\partial\zeta_t(\theta,\eta)}{\partial t}^\top\Big[(m_0'(s_0)-m'(s_0))k(s_0) - m_0'(s_0)h_{\theta_0}(s_0)\Big]\\
&= \frac{\partial\zeta_t(\theta,\eta)}{\partial t}^\top\Big[\urcorner_t(\zeta_t(\theta,\eta)^\top x;\theta,\eta,m)x + (m_0'(s_0)-m'(s_0))k(s_0) - m_0'(s_0)h_{\theta_0}(s_0)\Big]\\
&\qquad + \int_{s_0}^{\zeta_t(\theta,\eta)^\top x} m''\circ\circ\phi_{\theta,\eta,t}(u+(\theta-\zeta_t(\theta,\eta))^\top k(u))\Big[\dot{\phi}_{\theta,\eta,t}(u+(\theta-\zeta_t(\theta,\eta))^\top k(u))\\
&\qquad\qquad - \phi'_{\theta,\eta,t}(u+(\theta-\zeta_t(\theta,\eta))^\top k(u))\frac{\partial\zeta_t(\theta,\eta)}{\partial t}^\top k(u)\Big]du.
\end{aligned}
$$

The interchange of derivative and the integral is possible by assumptions **(S1)**, **(B1)**, and **(B2)**. Using the fact that $\phi'_{\theta,\eta,t}(u) = 1$ and $\dot{\phi}_{\theta,\eta,t}(u) = 0$ for all $u \in D_\theta$ (follows from the definition (4.31)) and $\partial\zeta_t(\theta,\eta)/\partial t = -2t/\sqrt{1-t^2|\eta|^2}\,\theta + H_\theta\eta$, we get

$$
\begin{aligned}
&\frac{\partial}{\partial t}\xi_t(\zeta_t(\theta,\eta)^\top x;\theta,\eta,m)\Big|_{t=0}\\
&= \eta^\top H_\theta^\top\left[m'(\theta^\top x)x + (m_0'(s_0)-m'(s_0))k(s_0) - m_0'(s_0)h_{\theta_0}(s_0)\right]\\
&\qquad - \eta^\top H_\theta^\top\int_{s_0}^{\theta^\top x} m''(u)k(u)du,
\end{aligned}
$$

and

$$
\begin{aligned}
&-\frac{1}{2}\frac{\partial}{\partial t}(y-\xi_t(\theta,m)(t^\top x))^2\Big|_{t=\theta}\\
&= \big(y-m(\theta^\top x)\big)\eta^\top H_\theta^\top\Big[m'(\theta^\top x)x + \int_{s_0}^{\theta^\top x} m''(u)[-k(u)]du\\
&\qquad\qquad + (m_0'(s_0)-m'(s_0))k(s_0) - m_0'(s_0)h_{\theta_0}(s_0)\Big]\\
&= \big(y-m(\theta^\top x)\big)\eta^\top H_\theta^\top\Big[m'(\theta^\top x)x + \int_{s_0}^{\theta^\top x} m'(u)k'(u)du - m'(\theta^\top x)k(\theta^\top x)\\
&\qquad\qquad + m'(s_0)k(s_0) + (m_0'(s_0)-m'(s_0))k(s_0) - m_0'(s_0)h_{\theta_0}(s_0)\Big]\\
&= \eta^\top\mathfrak{S}_{\theta,m}(x,y).
\end{aligned}
\tag{4.97}
$$

Next, observe that $(\hat{\theta}, \hat{m})$ minimizes $\mathcal{L}_n(m, \theta; \lambda_n)$ and $\xi_0(\zeta_0(\hat{\theta}, \eta)^\top x; \hat{\theta}, \eta, \hat{m}) = \hat{m}(\hat{\theta}^\top x)$. Hence the function

$$t \mapsto \frac{1}{n}\sum_{i=1}^{n}(y_i - \xi_t(\zeta_t(\hat{\theta}, \eta)^\top x; \hat{\theta}, \eta, \hat{m})^2 + \hat{\lambda}_n^2 \int_D \{\xi_t(\hat{\theta}, \eta, \hat{m})''(u)\}^2 du$$

is minimized at $t = \hat{\theta}$ for every $\eta \in S^{d-2}$. Observe that (4.97) and the fact that $J^2(\xi_t(\theta, m))$ is differentiable imply that the above function is differentiable in t on a small neighborhood of 0 (which depends on η). Hence, we have that

$$\mathbb{P}_n \mathfrak{S}_{\hat{\theta},\hat{m}} - \frac{\hat{\lambda}_n^2}{2} \frac{\partial J^2(\xi_t(\hat{\theta}, \hat{m}))}{\partial t}\bigg|_{t=\hat{\theta}} = 0. \tag{4.98}$$

Moreover, after some tedious algebra it can be seen that

$$\frac{\partial}{\partial t} J^2(\xi_t(\theta, m))\bigg|_{t=\theta} \lesssim \int_{D_\theta} k'(p)\{m''(p)\}^2 dp, \tag{4.99}$$

where k is defined in (4.20) and

$$k'(u) = 2h'_{\theta_0}(u) - \frac{m'_0(u)m'''_0(u)}{(m''_0(u))^2} h'_{\theta_0}(u) + \frac{m'_0(u)}{m''_0(u)} h''_{\theta_0}(u).$$

Hence, by assumptions **(S1)**, **(B1)**, and **(B2)**, we can find M^* such that $\|k\|_D^S \le M^*$. Thus, by Theorem 21 and (4.99), we have

$$\frac{\partial}{\partial t} J^2(\xi_t(\hat{\theta}, \hat{m}))\bigg|_{t=\hat{\theta}} \le M^* J(\hat{m}) = O_p(1).$$

Finally, (4.98) and **(S2)** imply $\mathbb{P}_n \mathfrak{S}_{\hat{\theta},\hat{m}} = o_p(n^{-1/2})$.

4.12.2 Proof of Lemma 32

From the definitions of $\mathfrak{S}_{\theta,m}$ and $\psi_{\theta,m}$, we have

$$\begin{aligned}
&\mathfrak{S}_{\theta,m}(x, y) - \psi_{\theta,m}(x, y) \\
&= \{y - m(\theta^\top x)\} H_\theta^\top \Big[\int_{s_0}^{\theta^\top x} m'(u)k'(u)du - m'(\theta^\top x)k(\theta^\top x) + m_0{}'(s_0)k(s_0) \\
&\qquad\qquad - m'_0(s_0)h_{\theta_0}(s_0) + (m'_0\, h_{\theta_0})(\theta^\top x) \Big].
\end{aligned}$$

Observe that

$$\int_{s_0}^{\theta^\top x} m'(u)k'(u)du - m'(\theta^\top x)k(\theta^\top x) + {m_0}'(s_0)k(s_0) - m_0'(s_0)h_{\theta_0}(s_0) + (m_0' \, h_{\theta_0})(\theta^\top x)$$

$$\begin{aligned}
&= \int_{s_0}^{\theta^\top x} m'(u)k'(u)du - \int_{s_0}^{\theta^\top x} m_0'(u)k'(u)du + \int_{s_0}^{\theta^\top x} m_0'(u)k'(u)du - m'(\theta^\top x)k(\theta^\top x) \\
&\quad + {m_0}'(s_0)k(s_0) - m_0'(s_0)h_{\theta_0}(s_0) + (m_0' \, h_{\theta_0})(\theta^\top x) \\
&= \int_{s_0}^{\theta^\top x} \{m'(u) - m_0'(u)\}k'(u)du + \int_{s_0}^{\theta^\top x} m_0'(u)k'(u)du - m'(\theta^\top x)k(\theta^\top x) \\
&\quad + m_0'(s_0)k(s_0) + (m_0' \, h_{\theta_0})(\theta^\top x) - (m_0' \, h_{\theta_0})(s_0). \qquad (4.100)
\end{aligned}$$

We now analyze the terms in the right hand side of the above display. First observe that

$$\begin{aligned}
&\int_{s_0}^{\theta^\top x} m_0'(u)k'(u)du - m'(\theta^\top x)k(\theta^\top x) + m_0'(s_0)k(s_0) \\
&= m_0'(u)k(u)\Big|_{s_0}^{\theta^\top x} - \int_{s_0}^{\theta^\top x} m_0''(u)k(u)du - m'(\theta^\top x)k(\theta^\top x) + m_0'(s_0)k(s_0) \qquad (4.101) \\
&= -\int_{s_0}^{\theta^\top x} m_0''(u)k(u)du + (m_0'(\theta^\top x) - m'(\theta^\top x))k(\theta^\top x).
\end{aligned}$$

Finally, by definition (4.20) and integration by parts, we have

$$\int_{s_0}^{\theta^\top x} m_0''(u)k(u)du = \int_{s_0}^{\theta^\top x} [m_0''(u)h_{\theta_0}(u) + m_0'(u)h_{\theta_0}'(u)]du = m_0'(u)h_{\theta_0}(u)\Big|_{s_0}^{\theta^\top x}. \qquad (4.102)$$

By substituting (4.101) and (4.102) in (4.100), we have that

$$\sqrt{n}\mathbb{P}_n(\mathfrak{S}_{\hat\theta,\hat m} - \psi_{\hat\theta,\hat m}) = \sqrt{n}\mathbb{P}_n[(Y - \hat m(\hat\theta^\top X))U_{\hat\theta,\hat m}(X)].$$

In the following, we find an upper bound of $\sqrt{n}\mathbb{P}_n[(Y - \hat m(\hat\theta^\top X))U_{\hat\theta,\hat m}(X)]$:

$$\begin{aligned}
&|\sqrt{n}\mathbb{P}_n[(Y - \hat m(\hat\theta^\top X))U_{\hat\theta,\hat m}(X)| \\
&= |\sqrt{n}\mathbb{P}_n[(m_0({\theta_0}^\top X) - \hat m(\hat\theta^\top X))U_{\hat\theta,\hat m}(X)] + \sqrt{n}\mathbb{P}_n\epsilon U_{\hat\theta,\hat m}(X)| \\
&\le |\sqrt{n}(\mathbb{P}_n[(m_0 - \hat m)({\theta_0}^\top X)U_{\hat\theta,\hat m}(X)]| + |\sqrt{n}(\mathbb{P}_n[(\hat m({\theta_0}^\top X) - \hat m(\hat\theta^\top X))U_{\hat\theta,\hat m}(X)]| \\
&\quad + |\sqrt{n}\mathbb{P}_n\epsilon U_{\hat\theta,\hat m}(X)| \\
&\le |\mathbb{G}_n[(m_0 - \hat m)({\theta_0}^\top X)U_{\hat\theta,\hat m}(X)]| + \big|\mathbb{G}_n[(\hat m({\theta_0}^\top X) - \hat m(\hat\theta^\top X))U_{\hat\theta,\hat m}(X)]\big| \\
&\quad + |\sqrt{n}\mathbb{P}_n\epsilon U_{\hat\theta,\hat m}(X)| + \sqrt{n}\big|P_{\theta_0,m_0}[(\hat m({\theta_0}^\top X) - \hat m(\hat\theta^\top X))U_{\hat\theta,\hat m}(X)]\big| \\
&\quad + \sqrt{n}\big|P_{\theta_0,m_0}[(m_0 - \hat m)({\theta_0}^\top X)U_{\hat\theta,\hat m}(X)]\big|.
\end{aligned}$$

4.12.3 Proof of Lemma 33

We will first show that

$$N(\varepsilon, \mathcal{W}^*_{M_1,M_2,M_3}, \|\cdot\|_\infty) \leq c \exp(c/\varepsilon)\varepsilon^{-4d}, \tag{4.103}$$

where c depends only on M_1, M_2, and M_3. By Lemma 23, we have

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

where c is a constant depending only on M_1, M_2, and M_3. Let us denote the functions in the ε-cover by $l_1, \ldots, l_t$. By Lemma 28, we have that there exists $\theta_1, \ldots, \theta_s$ for $s \lesssim \varepsilon^{-4d}$ such that $\{\theta_i\}_{1\leq i\leq s}$ form an ε^2-cover of $\Theta \cap B_{\theta_0}(1/2)$ and satisfies (3.84) (with ε^2 instead of ε). Fix $(\theta, m) \in \mathcal{C}^*_{M_1,M_2,M_3}$, without loss of generality assume that the function nearest to m' in the ε-cover 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.,

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

We define r_1 to be the anti-derivative of l_1 i.e., $l_1 = r_1'$. Let us define

$$V_{\theta,m}(x) := \left[\int_{s_0}^{\theta^\top x} \big[m'(u) - m_0'(u)\big] k'(u)du + (m_0{}'(\theta^\top x) - m'(\theta^\top x))k(\theta^\top x)\right].$$

Recall that $U_{\theta,m} = H_\theta^\top V_{\theta,m}$ Now for every $x \in \mathcal{X}$ observe that

$$\begin{aligned}
&\big|U_{\theta,m}(x) - U_{\theta_1,r_1}(x)\big| \\
&\leq \big|U_{\theta,m}(x) - U_{\theta,r_1}(x)\big| + |(H_\theta^\top - H_{\theta_1}^\top)V_{\theta,r_1}| + \big|H_{\theta_1}^\top\big(V_{\theta,r_1}(x) - V_{\theta_1,r_1}(x)\big)\big| \\
&\leq \left|H_\theta^\top \int_{s_0}^{\theta^\top x} [m'(u) - r_1'(u)]k'(u)du\right| + \left|H_\theta^\top (m' - r_1')(\theta^\top x)k(\theta^\top x)\right| \\
&\quad + 4M^*M_2(T+1)\sqrt{d-1}\varepsilon^2 + \left|H_{\theta_1}^\top\left[(r_1' - m_0')(\theta^\top x)\; k(\theta^\top x) - (r_1' - m_0')(\theta_1^\top x)k(\theta_1^\top x)\right]\right| \\
&\quad + \left|H_{\theta_1}^\top \int_{\theta_1{}^\top x}^{\theta^\top x} [r_1'(u) - m_0'(u)]k'(u)du\right| \\
&\leq 2T(1 + |\theta_0|)M^*\|m' - r_1'\|_\infty + M^*\|m' - r_1'\|_\infty + 4M^*M_2(T+1)\sqrt{d-1}\varepsilon^2 \\
&\quad + \left|(r_1' - m_0')(\theta^\top x)k(\theta^\top x) - (r_1' - m_0')(\theta_1^\top x)k(\theta^\top x)\right| + 2M_2M^*T|\theta - \theta_1|.
\end{aligned}$$

Furthermore, note that

$$\begin{aligned}
&\Big|(r_1' - m_0')(\theta^\top x)\, k(\theta^\top x) - (r_1' - m_0')(\theta_1^\top x)k(\theta_1^\top x)\Big| \\
&\le \Big|(r_1' - m_0')(\theta^\top x)k(\theta^\top x) - (r_1' - m_0')(\theta_1^\top x)k(\theta^\top x)\Big| \\
&\qquad + \Big|(r_1' - m_0')(\theta_1^\top x)[k(\theta^\top x) - k(\theta_1^\top x)]\Big| \\
&\le M^*\Big|(r_1' - m_0')(\theta^\top x) - (r_1' - m_0')(\theta_1^\top x)\Big| + 2M_2M^*T|\theta - \theta_1| \\
&\le 2M_3M^*T|\theta - \theta_1|^{1/2} + 2M_2M^*T|\theta - \theta_1|,
\end{aligned}$$

where the last inequality in the previous display follows from Lemma 29. Combining the above two displays, we have

$$\begin{aligned}
\big|U_{\theta,m}(x) - U_{\theta_1,r_1}(x)\big| \le M^*\|m' - r_1'\|_\infty(4T+1) + 4M^*M_2(T+1)\sqrt{d-1}\varepsilon^2 \\
+ 2M_3M^*T|\theta - \theta_1|^{1/2} + 2M_2M^*T|\theta - \theta_1| + 2M_2M^*T|\theta - \theta_1|.
\end{aligned}$$

Thus, $\{U_{\theta_i,l_j}\}$ form an (constant multiple of) ε-cover (with respect to $\|\cdot\|_{2,\infty}$ norm) of $\mathcal{W}^*_{M_1,M_2,M_3}$, and we have (4.103). Moreover, as $N_{[]}(\varepsilon, \mathcal{W}^*_{M_1,M_2,M_3}, \|\cdot\|_{2,P_{\theta_0,m_0}}) \lesssim N(\varepsilon, \mathcal{W}^*_{M_1,M_2,M_3}, \|\cdot\|_\infty)$ 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 $N_{[]}(\varepsilon, \mathcal{W}_{M_1,M_2,M_3}(n), \|\cdot\|_{2,P_{\theta_0,m_0}}) \lesssim N_{[]}(\varepsilon, \mathcal{W}^*_{M_1,M_2,M_3}, \|\cdot\|_{2,P_{\theta_0,m_0}})$. Now we find an envelope function for $\mathcal{W}_{M_1,M_2,M_3}(n)$. Recall that $|H_\theta^\top x| \le |x|$ for all $x \in \mathbb{R}^d$. For every $(\eta, f) \in \mathcal{C}_{M_1,M_2,M_3}(n)$ and $x \in \mathcal{X}$, observe that

$$\begin{aligned}
\big|U_{\theta,m}(x)\big| &\le \Big|\int_{s_0}^{\theta^\top x}[m'(u) - m_0'(u)]k'(u)du\Big| + \Big|(m' - m_0')(\theta^\top x)k(\theta^\top x)\Big| \\
&\le \Big|\int_{s_0}^{\theta_0^\top x}[m'(u) - m_0'(u)]k'(u)du\Big| + \Big|\int_{\theta_0^\top x}^{\theta^\top x}[m'(u) - m_0'(u)]k'(u)du\Big| \\
&\qquad + \Big|(m' - m_0')(\theta^\top x)k(\theta^\top x)\Big| \\
&\le \sqrt{d-1}M^*\big(T\|m - m_0\|^S_{D_{\theta_0}} + 2M_2T|\theta - \theta_0| + 2M_3\sqrt{T}|\theta - \theta_0|^{\frac{1}{2}} + \|m - m_0\|^S_{D_{\theta_0}}\big) \\
&\le W_{M_1,M_2,M_3}(n).
\end{aligned}$$

Thus, $W_{M_1,M_2,M_3}(n)$ satisfies (4.37).

4.12.4 Proof of Lemma 34

For every $(\theta, m) \in \mathcal{C}_{M_1,M_2,M_3}(n)$, note that

$$\big|(m_0 - m)(\theta_0{}^\top x)U_{\theta,m}(x)\big| \le 2M_1\big|U_{\theta,m}(x)\big| \le 2M_1 W_{M_1,M_2,M_3}(n) = D_{M_1,M_2,M_3}(n).$$

Furthermore, we have

$$N(\varepsilon, \big\{(m_0 - m)(\theta_0{}^\top \cdot) : m \in \mathcal{C}^{m*}_{M_1,M_2,M_3}\big\}, \|\cdot\|_\infty) = N(\varepsilon, \mathcal{C}^{m*}_{M_1,M_2,M_3}, \|\cdot\|_\infty) < \exp(c/\sqrt{\varepsilon}),$$

where the inequality follows from Lemma 23 and c is a constant depending only on M_1, M_2, and M_3. By Lemma 9.25 of [Kosorok, 2008] (for entropy of product of uniformly bounded function classes), and Lemma 33, we have that

$$N(\varepsilon, \mathcal{D}^*_{M_1,M_2,M_3}, \|\cdot\|_{2,\infty}) \le c\varepsilon^{-4d}\exp(\frac{c}{\sqrt{\varepsilon}} + \frac{c}{\varepsilon}).$$

Since, $N(\varepsilon, \mathcal{D}_{M_1,M_2,M_3}(n), \|\cdot\|_{2,\infty}) \le N(\varepsilon, \mathcal{D}^*_{M_1,M_2,M_3}, \|\cdot\|_{2,\infty})$ and $N_{[]}(\varepsilon, \mathcal{D}_{M_1,M_2,M_3}(n), \|\cdot\|_{2,P_{\theta_0,m_0}}) \lesssim N(\varepsilon, \mathcal{D}^*_{M_1,M_2,M_3}, \|\cdot\|_{2,\infty})$, we have $J_{[]}(\gamma, \mathcal{D}_{M_1,M_2,M_3}(n), \|\cdot\|_{2,P_{\theta_0,m_0}}) \lesssim c\gamma^{1/2}$. Using arguments similar to (3.91) and (3.92) and the maximal inequality in Corollary 19.35 of [van der Vaart, 1998b]

$$\begin{aligned}
\mathbb{P}\Big(\sup_{f\in\mathcal{D}_{M_1,M_2,M_3}(n)} |\mathbb{G}_n f| > \delta\Big) &\le \delta^{-1}\sqrt{d-1}\sum_{i=1}^{d-1}\mathbb{E}\Big(\sup_{f\in\mathcal{D}_{M_1,M_2,M_3}(n)} |\mathbb{G}_n f_i|\Big) \\
&\le \delta^{-1}\sqrt{d-1}J_{[]}(D_{M_1,M_2,M_3}(n), \mathcal{D}_{M_1,M_2,M_3}(n), \|\cdot\|_{2,P_{\theta_0,m_0}}) \\
&\lesssim \delta^{-1}\|D_{M_1,M_2,M_3}(n)\|^{1/2} \\
&\lesssim \Big[\hat{\lambda}_n^{1/4} + \frac{1}{a_n}\Big]^{1/2} \to 0, \qquad \text{as } n \to \infty,
\end{aligned}$$

where we have used (4.38) and the fact that $D^2_{M_1,M_2,M_3}(n)$ is non-random in the last inequality.

4.12.5 Proof of Lemma 35

First, note that for every $(\theta, m) \in \mathcal{C}_{M_1,M_2,M_3}(n)$, we have

$$\big|(m(\theta_0{}^\top x) - m(\theta^\top x))U_{\theta,m}(x)\big| \le 2M_1\big|U_{\theta,m}(x)\big| \le D_{M_1,M_2,M_3}(n).$$

Observe that the proof of Lemma 35 will be complete (by arguments similar to the proof of Lemma 34) if we can show that

$$\log N(\varepsilon, \mathcal{A}^*_{M_1,M_2,M_3}, \|\cdot\|_{2,\infty}) \leq c \exp\big(\frac{c}{\varepsilon} + \frac{c}{\sqrt{\varepsilon}}\big)\varepsilon^{-4d}, \tag{4.104}$$

where the constant c depends only on M_1, M_2, M_3, and d.

However, arguments similar to the proof of Lemma 21 will show that

$$N(\varepsilon, \{m \circ \theta_0 - m \circ \theta : (\theta, m) \in \mathcal{C}^*_{M_1,M_2,M_3}\}, \|\cdot\|_\infty) < c \exp(c/\sqrt{\varepsilon})\varepsilon^{-d},$$

for some constant c depending only on d, M_1, M_2 and M_3. Thus by Lemma 9.25 of [Kosorok, 2008] and Lemma 33, we have (4.104).

4.12.6 Proof of Lemma 36

Note that, we have

$$\begin{aligned}
&\mathbb{P}(|\sqrt{n}\mathbb{P}_n(\epsilon U_{\hat{\theta},\hat{m}})| > \delta) \\
&\leq \mathbb{P}\Big(\sup_{(\theta,m)\in\mathcal{C}_{M_1,M_2,M_3}(n)} |\sqrt{n}\mathbb{P}_n(\epsilon U_{\theta,m})| > \delta\Big) + \mathbb{P}((\hat{\theta},\hat{m}) \notin \mathcal{C}_{M_1,M_2,M_3}(n)) \\
&\leq \mathbb{P}\Big(\sup_{f\in\mathcal{W}_{M_1,M_2,M_3}(n)} |\sqrt{n}\mathbb{P}_n \epsilon f| > \delta\Big) + \mathbb{P}((\hat{\theta},\hat{m}) \notin \mathcal{C}_{M_1,M_2,M_3}(n)) \\
&= \mathbb{P}\Big(\sup_{f\in\mathcal{W}_{M_1,M_2,M_3}(n)} |\mathbb{G}_n \epsilon f| > \delta\Big) + \mathbb{P}((\hat{\theta},\hat{m}) \notin \mathcal{C}_{M_1,M_2,M_3}(n)),
\end{aligned}$$

where the last equality is due to assumption **(A2)**. Now it is enough 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 $n \to 0$. By Lemma 33, we have

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

Fix $(\theta, m) \in \mathcal{C}_{M_1,M_2,M_3}(n)$. If $[\hbar_1, \hbar_2]$ is a bracket (coordinate wise) for $U_{\theta,m}$, then $[\hbar_1\epsilon^+ - \hbar_2\epsilon^-, \hbar_2\epsilon^+ - \hbar_1\epsilon^-]$ is a bracket for $\epsilon U_{\theta,m}$. Therefore, we have

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

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

$$|\epsilon U_{\theta,m}(x)| \leq |\epsilon| W_{M_1,M_2,M_3}(n).$$

It follows that

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

Thus using arguments similar to (3.91) and (3.92) and the maximal inequality in Corollary 19.35 of [van der Vaart, 1998b], we have

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

Now, we prove the second and third equations in (4.39). First, note that

$$\begin{aligned}
\big|P_{\theta_0,m_0}[(m_0 - \hat{m})(\theta_0{}^\top X) U_{\hat{\theta},\hat{m}}(X)]\big| &\leq \sqrt{P_{\theta_0,m_0}\big[(m_0 - \hat{m})^2 (\theta_0{}^\top X)\big] P_{\theta_0,m_0}\big|U_{\hat{\theta},\hat{m}}(X)\big|^2} \\
&\leq O_p(\hat{\lambda}_n)\Big[P_{\theta_0,m_0}\big|U_{\hat{\theta},\hat{m}}(X)\big|^2\Big]^{1/2},
\end{aligned} \tag{4.105}$$

where the first inequality is an application of the CauchySchwarz inequality and the second inequality is due to Theorem 23. Similarly, using Theorems 23, 22, and the mean value theorem we have

$$\begin{aligned}
\big|P_{\theta_0,m_0}[(\hat{m}(\theta_0{}^\top X) - \hat{m}(\hat{\theta}^\top X)) U_{\hat{\theta},\hat{m}}(X)]\big| &\leq \sqrt{P_{\theta_0,m_0}[\hat{m}(\theta_0{}^\top X) - \hat{m}(\hat{\theta}^\top X)]^2 P_{\theta_0,m_0}\big|U_{\hat{\theta},\hat{m}}(X)\big|^2} \\
&\leq \|\hat{m}'\|_\infty |\hat{\theta} - \theta_0| T\Big[P_{\theta_0,m_0}\big|U_{\hat{\theta},\hat{m}}(X)\big|^2\Big]^{1/2} \\
&\leq O_p(\hat{\lambda}_n)\Big[P_{\theta_0,m_0}\big|U_{\hat{\theta},\hat{m}}(X)\big|^2\Big]^{1/2}.
\end{aligned} \tag{4.106}$$

Now we find an upper bound for $P_{\theta_0,m_0}|U_{\hat{\theta},\hat{m}}(X)|^2$. Note that

$$\begin{aligned}
P_{\theta_0,m_0}\big|U_{\hat{\theta},\hat{m}}(X)\big|^2 &\lesssim P_{\theta_0,m_0}\big|H_{\hat{\theta}}^\top(\hat{m}'-m_0')(\hat{\theta}^\top X)k(\hat{\theta}^\top X)\big|^2 \\
&\quad + P_{\theta_0,m_0}\left|H_{\hat{\theta}}^\top\int_{s_0}^{\hat{\theta}^\top X}[\hat{m}'(u)-m_0'(u)]k'(u)du\right|^2 \\
&\le M^{*2}(d-1)P_{\theta_0,m_0}\Big[(\hat{m}'-m_0')(\hat{\theta}^\top X)\Big]^2 \\
&\quad + P_{\theta_0,m_0}\left[\int_{s_0}^{\hat{\theta}^\top X}[\hat{m}'(u)-m_0'(u)]^2du\int_{s_0}^{\hat{\theta}^\top X}|k'(u)|^2du\right] \\
&\le M^{*2}(d-1)P_{\theta_0,m_0}\Big[(\hat{m}'-m_0')(\hat{\theta}^\top X)\Big]^2 \\
&\quad + M^{*2}(d-1)TP_{\theta_0,m_0}\left[\int_{s_0}^{\hat{\theta}^\top X}[\hat{m}'(u)-m_0'(u)]^2du\right] \\
&\le M^{*2}(d-1)P_{\theta_0,m_0}\Big[(\hat{m}'-m_0')(\hat{\theta}^\top X)\Big]^2 \\
&\quad + M^{*2}(d-1)TP_{\theta_0,m_0}\left[\int_{D_{\hat{\theta}}}[\hat{m}'(u)-m_0'(u)]^2du\right], \qquad (4.107)
\end{aligned}$$

where M^* is defined in (4.21). Since $|\hat{\theta}-\theta_0| = o_p(1)$, by assumption **(A5)**, we have that the density of $\hat{\theta}^\top X$ w.r.t to the Lebesgue measure is bounded away from zero. Thus,

$$\int_{D_{\hat{\theta}}}\{\hat{m}'(u)-m_0'(u)\}^2du \lesssim \|\hat{m}'\circ\hat{\theta}-m_0'\circ\hat{\theta}\|^2 = O_p(\hat{\lambda}_n).$$

The theorem now follows, as

$$\begin{aligned}
&\big|P_{\theta_0,m_0}[(m_0-\hat{m})(\theta_0{}^\top X)U_{\hat{\theta},\hat{m}}(X)]\big| = O_p(\hat{\lambda}_n^{3/2}) = O_p(n^{-3/5}),\\
&\big|P_{\theta_0,m_0}[(\hat{m}(\theta_0{}^\top X)-\hat{m}(\hat{\theta}^\top X))U_{\hat{\theta},\hat{m}}(X)]\big| = O_p(\hat{\lambda}_n^{3/2}) = O_p(n^{-3/5}).
\end{aligned}$$

4.12.7 Proof of Theorem 32

First, note that

$$\begin{aligned}
P_{\hat{\theta},m_0}\psi_{\hat{\theta},\hat{m}} &= H_{\hat{\theta}}^\top E_{\hat{\theta},m_0}\Big[(Y-\hat{m}(\hat{\theta}^\top X))\big[\hat{m}'(\hat{\theta}^\top X)X-(m_0'\,h_{\theta_0})(\hat{\theta}^\top X)\big]\Big] \\
&= H_{\hat{\theta}}^\top E_X\left[E_{\hat{\theta},m_0}\Big[(Y-\hat{m}(\hat{\theta}^\top X))\big[\hat{m}'(\hat{\theta}^\top X)X-(m_0'\,h_{\theta_0})(\hat{\theta}^\top X)\big]\Big|\hat{\theta}^\top X\right]\right] \\
&= H_{\hat{\theta}}^\top E_X\Big[(m_0-\hat{m})(\hat{\theta}^\top X)\big[\hat{m}'(\hat{\theta}^\top X)E(X|\hat{\theta}^\top X)-(m_0'\,h_{\theta_0})(\hat{\theta}^\top X)\big]\Big] \\
&= H_{\hat{\theta}}^\top P_X\Big[(m_0-\hat{m})(\hat{\theta}^\top X)E(X|\hat{\theta}^\top X)(\hat{m}'-m_0')(\hat{\theta}^\top X))\Big] \\
&\quad + H_{\hat{\theta}}^\top P_X\Big[(m_0-\hat{m})(\hat{\theta}^\top X)m_0'(\hat{\theta}^\top X)\big[E(X|\hat{\theta}^\top X)-h_{\theta_0}(\hat{\theta}^\top X))\big]\Big]. \qquad (4.108)
\end{aligned}$$

Now we will show that each of the terms in (4.108) are $o_p(n^{-1/2})$. By assumption **(A1)** and the Cauchy-Schwarz inequality, for the first term in (4.108) we have

$$\begin{aligned}
&\left|P_X[(m_0 - \hat{m})(\hat{\theta}^\top X)E(X|\hat{\theta}^\top X)(\hat{m}' - m_0')(\hat{\theta}^\top X))]\right| \\
&\leq T\sqrt{P_X\left[(m_0 - \hat{m})(\hat{\theta}^\top X)^2\right]P_X\left[(\hat{m}'(\hat{\theta}^\top X) - m_0'(\hat{\theta}^\top X))^2\right]} \\
&\lesssim \|m_0 \circ \hat{\theta} - \hat{m} \circ \hat{\theta}\| \, \|\hat{m}' \circ \hat{\theta} - m_0' \circ \hat{\theta}\|.
\end{aligned} \tag{4.109}$$

We can bound the two terms on the right side above display as follows. For the first term, note that by Theorems 21, 22, and 23, we have

$$\begin{aligned}
\|m_0 \circ \hat{\theta} - \hat{m} \circ \hat{\theta}\| &\leq \|m_0 \circ \theta_0 - m_0 \circ \hat{\theta}\| + \|\hat{m} \circ \hat{\theta} - m_0 \circ \theta_0\| \\
&\leq T\|m_0'\|_\infty|\theta_0 - \hat{\theta}| + \|\hat{m} \circ \hat{\theta} - m_0 \circ \theta_0\| \\
&= O_p(\hat{\lambda}_n).
\end{aligned} \tag{4.110}$$

For the second term in (4.109), observe that by Lemma 29 and Theorems 23 and 24, we have

$$\begin{aligned}
&\|\hat{m}' \circ \hat{\theta} - m_0' \circ \hat{\theta}\| \\
&\leq \|\hat{m}' \circ \hat{\theta} - \hat{m}' \circ \theta_0\| + \|\hat{m}' \circ \theta_0 - m_0' \circ \theta_0\| + \|m_0' \circ \theta_0 - m_0' \circ \hat{\theta}\| \\
&\leq J(\hat{m})|\hat{\theta} - \theta_0|^{\frac{1}{2}}T^{1/2} + \|\hat{m}' \circ \theta_0 - m_0' \circ \theta_0\| + J(m_0)|\hat{\theta} - \theta_0|^{\frac{1}{2}}T^{1/2} \\
&= O_p(\hat{\lambda}_n^{1/2}).
\end{aligned}$$

By the Cauchy-Schwarz inequality, the second term in (4.108) can be bounded as

$$\begin{aligned}
&\left|P_X[(m_0 - \hat{m})(\hat{\theta}^\top X)m_0'(\hat{\theta}^\top X)(E(X|\hat{\theta}^\top X) - h_{\theta_0}(\hat{\theta}^\top X))]\right| \\
&\leq \|m_0'\|_\infty\sqrt{P_X\left[(m_0 - \hat{m})^2(\hat{\theta}^\top X)\right]P_X\left[|h_{\hat{\theta}}(\hat{\theta}^\top X) - h_{\theta_0}(\hat{\theta}^\top X)|^2\right]} \\
&= \|m_0'\|_\infty\|m_0 \circ \hat{\theta} - \hat{m} \circ \hat{\theta}\| \, \|h_{\hat{\theta}} \circ \hat{\theta} - h_{\theta_0} \circ \hat{\theta}\|_{2,P_{\theta_0,m_0}} \\
&\leq \|m_0'\|_\infty O_p(\hat{\lambda}_n)\bar{M}|\hat{\theta} - \theta_0| = O_p(\hat{\lambda}_n^2),
\end{aligned} \tag{4.111}$$

where $\bar{M}$ is defined in (4.11). The last inequality in the above display follows from assumption **(B2)** and (4.110). The theorem now follows by combining these results.

4.12.8 Consistency of $\psi_{\hat{\theta},\hat{m}}$

Lemma 57. *If the conditions in Theorem 29 hold, then*

$$P_{\theta_0,m_0}|\psi_{\hat{\theta},\hat{m}} - \psi_{\theta_0,m_0}|^2 = o_p(1), \tag{4.112}$$

$$P_{\hat{\theta},m_0}|\psi_{\hat{\theta},\hat{m}}|^2 = O_p(1). \tag{4.113}$$

Proof. We first prove (4.112). By assumption **(B2)**, we have

$$\begin{aligned}
&P_{\theta_0,m_0}|\psi_{\hat{\theta},\hat{m}} - \psi_{\theta_0,m_0}|^2 \\
&= P_{\theta_0,m_0}\Big|(y - \hat{m}(\hat{\theta}^\top X))H_{\hat{\theta}}^\top\left[\hat{m}'(\hat{\theta}^\top X)X - (m_0' \; h_{\theta_0})(\hat{\theta}^\top X)\right] \\
&\qquad - (y - m_0(\theta_0^\top X))H_{\theta_0}^\top\left[m_0'(\theta_0^\top X)X - (m_0' \; h_{\theta_0})(\theta_0^\top X)\right]\Big|^2 \\
&= P_{\theta_0,m_0}\Big|\left[(m_0(\theta_0^\top X) - \hat{m}(\hat{\theta}^\top X)) + \epsilon\right]H_{\hat{\theta}}^\top\left[\hat{m}'(\hat{\theta}^\top X)X - (m_0' \; h_{\theta_0})(\hat{\theta}^\top X)\right] \\
&\qquad - \epsilon H_{\theta_0}^\top\left[m_0'(\theta_0^\top X)X - (m_0' \; h_{\theta_0})(\theta_0^\top X)\right]\Big|^2 \\
&= P_{\theta_0,m_0}\Big|\left[m_0(\theta_0^\top X) - \hat{m}(\hat{\theta}^\top X)\right]H_{\hat{\theta}}^\top\left[\hat{m}'(\hat{\theta}^\top X)X - (m_0' \; h_{\theta_0})(\hat{\theta}^\top X)\right]\Big|^2 \\
&\quad + P_{\theta_0,m_0}\Big|\epsilon\Big[H_{\hat{\theta}}^\top\left[\hat{m}'(\hat{\theta}^\top X)X - (m_0' \; h_{\theta_0})(\hat{\theta}^\top X)\right] - H_{\theta_0}^\top\left[m_0'(\theta_0^\top X)X - (m_0' \; h_{\theta_0})(\theta_0^\top X)\right]\Big]\Big|^2 \\
&\le P_{\theta_0,m_0}\Big|\left[m_0(\theta_0^\top X) - \hat{m}(\hat{\theta}^\top X)\right]\left[\hat{m}'(\hat{\theta}^\top X)X - (m_0' \; h_{\theta_0})(\hat{\theta}^\top X)\right]\Big|^2 \\
&\quad + P_{\theta_0,m_0}\Big|\epsilon H_{\hat{\theta}}^\top\Big[\hat{m}'(\hat{\theta}^\top X)X - (m_0' \; h_{\theta_0})(\hat{\theta}^\top X) - m_0'(\theta_0^\top X)X + (m_0' \; h_{\theta_0})(\theta_0^\top X)\Big]\Big|^2 \\
&\quad + P_{\theta_0,m_0}\Big|\epsilon\Big[H_{\hat{\theta}}^\top - H_{\theta_0}^\top\Big]\left[m_0'(\theta_0^\top X)X - (m_0' \; h_{\theta_0})(\theta_0^\top X)\right]\Big|^2 \\
&\le P_{\theta_0,m_0}\Big|\left[m_0(\theta_0^\top X) - \hat{m}(\hat{\theta}^\top X)\right]\left[\hat{m}'(\hat{\theta}^\top X)X - (m_0' \; h_{\theta_0})(\hat{\theta}^\top X)\right]\Big|^2 \\
&\quad + \|\sigma^2(\cdot)\|_\infty P_{\theta_0,m_0}\Big|\hat{m}'(\hat{\theta}^\top X)X - m_0'(\theta_0^\top X)X + (m_0' \; h_{\theta_0})(\theta_0^\top X) - (m_0' \; h_{\theta_0})(\hat{\theta}^\top X)\Big|^2 \\
&\quad + 4M_1^2T^2\|\sigma^2(\cdot)\|_\infty\|H_{\hat{\theta}} - H_{\theta_0}\|_2^2 \\
&\le P_{\theta_0,m_0}\Big|\left[m_0(\theta_0^\top X) - \hat{m}(\hat{\theta}^\top X)\right]\left[\hat{m}'(\hat{\theta}^\top X)X - (m_0' \; h_{\theta_0})(\hat{\theta}^\top X)\right]\Big|^2 \\
&\quad + 2\|\sigma^2(\cdot)\|_\infty P_{\theta_0,m_0}\Big|\hat{m}'(\hat{\theta}^\top X)X - m_0'(\theta_0^\top X)X\Big|^2 \\
&\quad + 2\|\sigma^2(\cdot)\|_\infty P_{\theta_0,m_0}\Big|(m_0' \; h_{\theta_0})(\theta_0^\top X) - (m_0' \; h_{\theta_0})(\hat{\theta}^\top X)\Big|^2 + 4M_1^2T^2|\hat{\theta} - \theta_0|^2\|\sigma^2(\cdot)\|_\infty \\
&= \mathbf{I} + 2\|\sigma^2(\cdot)\|_\infty\ \mathbf{II} + 2\|\sigma^2(\cdot)\|_\infty\ \mathbf{III} + 4M_1^2T^2\|\sigma^2(\cdot)\|_\infty|\hat{\theta} - \theta_0|^2.
\end{aligned} \tag{4.114}$$

We will now show that each of the first three terms in the above display are $o_p(1)$. For the second term, observe that

$$\begin{aligned}
\mathbf{II} &\leq T^2 P_{\theta_0,m_0}\left|\hat{m}'(\hat{\theta}^\top X) - m_0'(\theta_0^\top X)\right|^2 \\
&\leq P_{\theta_0,m_0}\left|(\hat{m}'(\hat{\theta}^\top X) - \hat{m}'(\theta_0^\top X))\right|^2 + P_{\theta_0,m_0}\left|(\hat{m}'(\theta_0^\top X) - m_0'(\theta_0^\top X))\right|^2 \\
&\leq J^2(\hat{m})T|\hat{\theta} - \theta_0| + \|\hat{m}' \circ \theta_0 - m_0' \circ \theta_0\|^2 \\
&= o_p(1).
\end{aligned}$$

Here the last inequality follows from Lemma 29 and the last equality is due to Theorems 23 and 24. For $\mathbf{I}$, recall that by Theorem 21, we have $\|m_0 \circ \theta_0 - \hat{m} \circ \hat{\theta}\| \xrightarrow{P} 0$. Thus,

$$\begin{aligned}
\mathbf{I} &= P_{\theta_0,m_0}\left|(m_0(\theta_0^\top X) - \hat{m}(\hat{\theta}^\top X))(\hat{m}'(\hat{\theta}^\top X)X - (m_0' \, h_{\theta_0})(\hat{\theta}^\top X))\right|^2 \\
&\leq \|m_0 \circ \theta_0 - \hat{m} \circ \hat{\theta}\|^2 (M_2 T + L\|h_{\theta_0}\|_{2,\infty})^2 = o_p(1).
\end{aligned}$$

Finally, we have

$$\begin{aligned}
\mathbf{III} &= P_{\theta_0,m_0}\left|(m_0' \, h_{\theta_0})(\theta_0^\top X) - (m_0' \, h_{\theta_0})(\hat{\theta}^\top X)\right|^2 \\
&\leq P_{\theta_0,m_0}\left[\|m_0'' \, h_{\theta_0} + m_0' \, h_{\theta_0}'\|_{2,\infty}|(\theta_0 - \hat{\theta})^\top X|\right]^2 \\
&\leq \|m_0'' \, h_{\theta_0} + m_0' \, h_{\theta_0}'\|_{2,\infty}^2 T^2 |\theta_0 - \hat{\theta}|^2 = o_p(1).
\end{aligned}$$

All these facts combined show that $P_{\theta_0,m_0}|\psi_{\hat{\theta},\hat{m}} - \psi_{\theta_0,m_0}|^2 = o_p(1)$. We now prove (4.113). Note that

$$\begin{aligned}
&P_{\hat{\theta},m_0}|\psi_{\hat{\theta},\hat{m}}|^2 \\
&\leq P_{\hat{\theta},m_0}\left|(Y - \hat{m}(\hat{\theta}^\top X))^2\left[\hat{m}'(\hat{\theta}^\top X)X - m_0'(\hat{\theta}^\top X)h_{\theta_0}(\hat{\theta}^\top X)\right]\right|^2 \\
&= P_{\hat{\theta},m_0}\left|\left[(m_0(\hat{\theta}^\top X) - \hat{m}(\hat{\theta}^\top X)) + \epsilon\right]\left[\hat{m}'(\hat{\theta}^\top X)X - (m_0' \, h_{\theta_0})(\hat{\theta}^\top X)\right]\right|^2 \\
&= P_{\hat{\theta},m_0}\left|\left[(m_0(\hat{\theta}^\top X) - \hat{m}(\hat{\theta}^\top X))\right]\left[\hat{m}'(\hat{\theta}^\top X)X - (m_0' \, h_{\theta_0})(\hat{\theta}^\top X)\right]\right|^2 \\
&\quad + P_{\hat{\theta},m_0}\left|\hat{m}'(\hat{\theta}^\top X)X - (m_0' \, h_{\theta_0})(\hat{\theta}^\top X)\right|^2 \\
&\leq (\|m_0\|_\infty^2 + \|\hat{m}\|_\infty^2) P_{\hat{\theta},m_0}\left|\hat{m}'(\hat{\theta}^\top X)X - (m_0' \, h_{\theta_0})(\hat{\theta}^\top X)\right|^2 \\
&\quad + P_{\hat{\theta},m_0}|\hat{m}'(\hat{\theta}^\top X)X - (m_0' \, h_{\theta_0})(\hat{\theta}^\top X)|^2 \\
&\leq (\|m_0\|_\infty^2 + \|\hat{m}\|_\infty^2 + 1) P_{\hat{\theta},m_0}|\hat{m}'(\hat{\theta}^\top X)X - (m_0' \, h_{\theta_0})(\hat{\theta}^\top X)|^2.
\end{aligned} \tag{4.115}$$

The result now follows. □

4.12.9 Proof of Theorem 33

Recall the definition (4.23). Under model (4.1),

$$\begin{aligned}\psi_{\hat\theta,\hat m}-\psi_{\theta_0,m_0} &= [\epsilon+m_0(\theta_0^\top x)-\hat m(\hat\theta^\top x)]H_{\hat\theta}^\top[\hat m'(\hat\theta^\top x)x-(m_0'\ h_{\theta_0})(\hat\theta^\top x)]\\ &\quad-\epsilon H_{\theta_0}^\top\big[m_0'(\theta_0^\top x)x-(m_0'h_{\theta_0})(\theta_0^\top x)\big]\\ &= \epsilon H_{\hat\theta}^\top\Big[[\hat m'(\hat\theta^\top x)-m_0'({\theta_0}^\top x)]x+[(m_0'\ h_{\theta_0})({\theta_0}^\top x)-(m_0'\ h_{\theta_0})(\hat\theta^\top x)]\Big]\\ &\quad+\epsilon(H_{\hat\theta}^\top-H_{\theta_0}^\top)\big[m_0'(\theta_0^\top x)x-(m_0'h_{\theta_0})(\theta_0^\top x)\big]\\ &\quad+H_{\hat\theta}^\top\Big[[m_0(\theta_0^\top x)-\hat m(\hat\theta^\top x)][\hat m'(\hat\theta^\top x)x-(m_0'\ h_{\theta_0})(\hat\theta^\top x)]\Big]. \quad (4.116)\end{aligned}$$

For every $(\theta,m)\in\Theta\times\mathcal{R}$, define functions $\upsilon_{\theta,m}:\chi\to\mathbb{R}^{d-1}$ and $\tau_{\theta,m}:\chi\to\mathbb{R}^{d-1}$ as follows:

$$\begin{aligned}\tau_{\theta,m}(x) &:= H_\theta^\top\big\{[m'(\theta^\top x)-m_0'({\theta_0}^\top x)]x+[(m_0'\ h_{\theta_0})({\theta_0}^\top x)-(m_0'\ h_{\theta_0})(\theta^\top x)]\big\}\\ &\quad+(H_\theta^\top-H_{\theta_0}^\top)\big[m_0'(\theta_0^\top x)x-(m_0'h_{\theta_0})(\theta_0^\top x)\big], \quad (4.117)\\ \upsilon_{\theta,m}(x) &:= H_\theta^\top[m_0(\theta_0^\top x)-m(\theta^\top x)][m'(\theta^\top x)x-(m_0'\ h_{\theta_0})(\theta^\top x)],\end{aligned}$$

and the classes of such functions

$$\begin{aligned}\Xi_{M_1,M_2,M_3}(n) &= \big\{\tau_{\theta,m}:(\theta,m)\in\mathcal{C}_{M_1,M_2,M_3}(n)\big\},\\ \Upsilon_{M_1,M_2,M_3}(n) &= \big\{\upsilon_{\theta,m}:(\theta,m)\in\mathcal{C}_{M_1,M_2,M_3}(n)\big\}.\end{aligned}$$

Observe that, for every fixed M_1, M_2, and M_3, we have

$$\begin{aligned}&\mathbb{P}(|\mathbb{G}_n(\psi_{\hat\theta,\hat m}-\psi_{\theta_0,m_0})|>\delta)\\ &\le \mathbb{P}(|\mathbb{G}_n(\epsilon\tau_{\hat\theta,\hat m}+\upsilon_{\hat\theta,\hat m})|>\delta,(\hat\theta,\hat m)\in\mathcal{C}_{M_1,M_2,M_3}(n))+\mathbb{P}((\hat\theta,\hat m)\notin\mathcal{C}_{M_1,M_2,M_3}(n))\\ &\le \mathbb{P}\left(|\mathbb{G}_n(\epsilon\tau_{\hat\theta,\hat m})|>\frac{\delta}{2},(\hat\theta,\hat m)\in\mathcal{C}_{M_1,M_2,M_3}(n)\right)\\ &\quad+\mathbb{P}\left(|\mathbb{G}_n\upsilon_{\hat\theta,\hat m}|>\frac{\delta}{2},(\hat\theta,\hat m)\in\mathcal{C}_{M_1,M_2,M_3}(n)\right)+\mathbb{P}\left((\hat\theta,\hat m)\notin\mathcal{C}_{M_1,M_2,M_3}(n)\right)\\ &\le \mathbb{P}\Big(\sup_{f\in\Xi_{M_1,M_2,M_3}(n)}|\mathbb{G}_n\epsilon f|>\frac{\delta}{2}\Big)\\ &\quad+\mathbb{P}\Big(\sup_{f\in\Upsilon_{M_1,M_2,M_3}(n)}|\mathbb{G}_n f|>\frac{\delta}{2}\Big)+\mathbb{P}\big((\hat\theta,\hat m)\notin\mathcal{C}_{M_1,M_2,M_3}(n)\big). \quad (4.118)\end{aligned}$$

By the discussion following Lemma 32, it is easy to see that to prove Theorem 33 we only need to show that the first two terms in (4.118) are $o(1)$. We prove this in Lemmas 58 and 59.

Lemma 58. *Fix M_1, M_2, M_3, and $\delta > 0$. For $n \in \mathbb{N}$, as $n \to \infty$, we have*

$$\mathbb{P}\Big(\sup_{f \in \Xi_{M_1,M_2,M_3}(n)} |\mathbb{G}_n \epsilon f| > \frac{\delta}{2} \Big) \to 0.$$

Proof. The proof of this lemma is similar to the first part of the proof of Lemma 36. Let us define,

$$\Xi^*_{M_1,M_2,M_3} := \big\{ \tau_{\theta,m} : (\theta, m) \in \mathcal{C}^*_{M_1,M_2,M_3} \big\}.$$

We will prove that

$$N(\varepsilon, \Xi^*_{M_1,M_2,M_3}, \|\cdot\|_{2,\infty}) \leq c \exp(c/\varepsilon) \varepsilon^{-4d}, \tag{4.119}$$

where c depends only on M_1, M_2, and M_3. Fix $(\theta, m) \in \mathcal{C}_{M_1,M_2,M_3}(n)$. By Lemma 23, we have

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

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 $\big\{ m' : (\cdot, m) \in \mathcal{C}^*_{M_1,M_2,M_3} \big\}$ by $l_1, \ldots, l_t$. By Lemma 28, we have that there exists $\theta_1, \ldots, \theta_s$ for $s \lesssim \varepsilon^{-4d}$ such that $\{\theta_i\}_{1 \leq i \leq s}$ form an ε^2-cover of $\Theta \cap B_{\theta_0}(1/2)$ and satisfies (3.84) (with ε^2 instead of ε). Fix $(\theta, m) \in \mathcal{C}^*_{M_1,M_2,M_3}$. Without loss of generality assume that the function nearest to m' in the ε-cover 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.,

$$\|m' - l_1\|_\infty \leq \varepsilon, \quad \|H_\theta^\top - H_{\theta_1}^\top\| \leq \varepsilon^2, \quad \text{and} \quad |\theta - \theta_1| \leq \varepsilon^2. \tag{4.120}$$

We define r_1 to be the anti-derivative of l_1 i.e., $l_1 = r_1'$. Moreover, let us define

$$\varrho_{\theta,m}(x) := [m'(\theta^\top x) - m_0'(\theta_0{}^\top x)]x + [(m_0' \, h_{\theta_0})(\theta_0{}^\top x) - (m_0' \, h_{\theta_0})(\theta^\top x)].$$

Note that to prove (4.119), it is enough to show that $\|\tau_{\theta,m} - \tau_{\theta_1,r_1}\|_{2,\infty} \leq c_1 \varepsilon$, where c_1 is a constant. For every $x \in \mathcal{X}$ observe that

$$\begin{aligned}
&|\tau_{\theta,m}(x) - \tau_{\theta_1,r_1}(x)| \\
&\leq |H_\theta^\top \varrho_{\theta,m}(x) - H_{\theta_1}^\top \varrho_{\theta_1,r_1}(x)| + \big|(H_\theta^\top - H_{\theta_1}^\top)\big[m_0'(\theta_0^\top x)x - (m_0' h_{\theta_0})(\theta_0^\top x)\big]\big| \\
&\leq |(H_\theta^\top - H_{\theta_1}^\top)\varrho_{\theta,m}(x)| + |H_{\theta_1}^\top(\varrho_{\theta,m}(x) - \varrho_{\theta_1,r_1}(x))| + \varepsilon^2\big|m_0'(\theta_0^\top x)x - (m_0' h_{\theta_0})(\theta_0^\top x)\big| \\
&\leq \varepsilon^2 |\varrho_{\theta,m}(x)| + |\varrho_{\theta,m}(x) - \varrho_{\theta_1,r_1}(x)| + 2M_2 T \varepsilon^2 \\
&\leq \varepsilon^2 4M_2 T + |\varrho_{\theta,m}(x) - \varrho_{\theta_1,r_1}(x)| + 2M_2 T \varepsilon^2,
\end{aligned} \tag{4.121}$$

where the last two inequalities follow from properties of H_θ (Lemma 16), (4.120), and definition of $\mathcal{C}^*_{M_1,M_2,M_3}$ (see (4.36)). Furthermore, we have

$$\begin{aligned}
&|\varrho_{\theta,m}(x) - \varrho_{\theta_1,r_1}(x)| \\
&\le |(m'(\theta^\top x) - r_1'({\theta_1}^\top x))x| + |((m_0' \, h_{\theta_0})({\theta_1}^\top x) - (m_0' \, h_{\theta_0})(\theta^\top x))| \\
&\le |(m'(\theta^\top x) - m'({\theta_1}^\top x))x| + |(m'({\theta_1}^\top x) - r_1'({\theta_1}^\top x))x| \\
&\quad + |(m_0'({\theta_1}^\top x) - m_0'(\theta^\top x))h_{\theta_0}({\theta_1}^\top x)| + |m_0'(\theta^\top x)(h_{\theta_0}({\theta_1}^\top x) - h_{\theta_0}(\theta^\top x))| \\
&\le M_3 T^2 |\theta - \theta_1|^{1/2} + \|m - r_1\|_\infty T + \|h_{\theta_0}\|_\infty M_3 |\theta - \theta_1|^{1/2} + M_2 \|h_{\theta_0}'\|_\infty |\theta - \theta_1| T \\
&\lesssim \varepsilon
\end{aligned} \tag{4.122}$$

Thus combining (4.121) and (4.122), we have $\|\tau_{\theta,m} - \tau_{\theta_1,r_1}\|_{2,\infty} \le c_1 \varepsilon$.

However, bracketing entropy for the $\|\cdot\|_{2,P_{\theta_0,m_0}}$-norm is bounded above by a the covering entropy for the uniform norm for a class of function. Thus, we have

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

If $[\hbar_1, \hbar_2]$ is a bracket for $\tau_{\theta,m}$, then $[\hbar_1 \epsilon^+ - \hbar_2 \epsilon^-, \hbar_2 \epsilon^+ - \hbar_1 \epsilon^-]$ is a bracket (coordinate wise) for $\epsilon \tau_{\theta,m}$. Therefore, we have

$$N_{[]}\big(\varepsilon, \{\epsilon f : f \in \Xi^*_{M_1,M_2,M_3}\}, \|\cdot\|_{2,P_{\theta_0,m_0}}\big) \lesssim c \exp(c/\varepsilon).$$

Now, we find the envelope of $\Xi_{M_1,M_2,M_3}(n)$. For every $(\theta, m) \in \mathcal{C}_{M_1,M_2,M_3}(n)$ and $x \in \chi$ note that,

$$\begin{aligned}
|\tau_{\theta,m}(x)| &\le \big[|m'(\theta^\top x) - m'({\theta_0}^\top x)| + |m'({\theta_0}^\top x) - m_0'({\theta_0}^\top x)|\big]|x| \\
&\quad + |m_0'({\theta_0}^\top x) h_{\theta_0}({\theta_0}^\top x) - m_0'(\theta^\top x) h_{\theta_0}({\theta_0}^\top x)| \\
&\quad + |m_0'(\theta^\top x) h_{\theta_0}({\theta_0}^\top x) - m_0'(\theta^\top x) h_{\theta_0}(\theta^\top x)| \\
&\quad + |\theta - \theta_0| |m_0'(\theta^\top x) x - (m_0' h_{\theta_0})(\theta^\top x)| \\
&\le J(m) |\theta^\top x - {\theta_0}^\top x|^{1/2} |x| + \|m' - m_0'\|^S_{D_{\theta_0}} |x| \\
&\quad + |h_{\theta_0}({\theta_0}^\top x)| J(m_0) |{\theta_0}^\top x - \theta^\top x|^{1/2} \\
&\quad + |m_0'(\theta^\top x)| \, |h_{\theta_0}({\theta_0}^\top x) - h_{\theta_0}(\theta^\top x)| + |\theta - \theta_0| 2 M_2 T \\
&\le \hat{\lambda}_n^{-1/4} (M_3 T^2 + \|h_{\theta_0}\|_{2,\infty} M_3 T + M_2 \|h_{\theta_0}'\|_{2,\infty} T + 2 M_2 T) + \frac{1}{a_n} T.
\end{aligned}$$

Hence,

$$|\epsilon\tau_{\theta,m}(x)| \leq |\epsilon|\hat{\lambda}_n^{-1/4}(M_3T^2 + \|h_{\theta_0}\|_{2,\infty}M_3T + M_2\|h'_{\theta_0}\|_{2,\infty}T + 2M_2T) + |\epsilon|\frac{1}{a_n}T.$$

Thus using arguments similar to (3.91) and (3.92) and the maximal inequality in Corollary 19.35 of [van der Vaart, 1998b] (also see proof of Lemma 34), we have

$$\begin{aligned}
&\mathbb{P}\Big(\sup_{f\in\Xi_{M_1,M_2,M_3}(n)}|\mathbb{G}_n\epsilon f| > \frac{\delta}{2}\Big)\\
&\lesssim 2\delta^{-1}\mathbb{E}\Big(\sup_{f\in\Xi_{M_1,M_2,M_3}(n)}|\mathbb{G}_n\epsilon f|\Big)\\
&\lesssim J^*_{[]}\left(\left[\mathbb{P}\Big(|\epsilon|\big(\hat{\lambda}_n^{-1/4}+\frac{1}{a_n}\big)\Big)^2\right]^{1/2}, \Xi_{M_1,M_2,M_3}(n), \|\cdot\|_{2,P_{\theta_0,m_0}}\right)\\
&\lesssim \left[\frac{1}{\hat{\lambda}_n^{1/4}}+\frac{1}{a_n}\right]^{1/2} \to 0 \text{ as } n\to\infty. \qquad \square
\end{aligned}$$

Lemma 59. *Fix M_1, M_2, M_3, and $\delta > 0$. For $n \in \mathbb{N}$, we have*

$$\mathbb{P}\left(\sup_{f\in\Upsilon_{M_1,M_2,M_3}(n)}|\mathbb{G}_n f| > \frac{\delta}{2}\right) = o_p(1).$$

Proof. The proof of this lemma is similar to the proofs of Lemmas 34 and 35. Fix $(\theta, m) \in \mathcal{C}_{M_1,M_2,M_3}(n)$. We first find an envelope of $\Upsilon_{M_1,M_2,M_3}(n)$. Recall that for every $x \in \mathcal{X}$ and $\theta \in \Theta$, we have $|H_\theta^\top x| \leq |x|$. Thus for every $x \in \mathcal{X}$,

$$\begin{aligned}
|\upsilon_{\theta,m}(x)| &\leq |m_0(\theta_0^\top x) - m({\theta_0}^\top x)| \cdot |m'(\theta^\top x)x - m'_0(\theta^\top x)h_{\theta_0}(\theta^\top x)|\\
&\quad + |m(\theta_0^\top x) - m(\theta^\top x)| \cdot |m'(\theta^\top x)x - m'_0(\theta^\top x)h_{\theta_0}(\theta^\top x)|\\
&\leq \|m_0 - m\|^S_{D_{\theta_0}}|m'(\theta^\top x)x - m'_0(\theta^\top x)h_{\theta_0}(\theta^\top x)|\\
&\quad + \|m'\|_\infty T|\theta - \theta_0| \cdot |m'(\theta^\top x)x - m'_0(\theta^\top x)h_{\theta_0}(\theta^\top x)|\\
&\leq \left[\frac{1}{a_n} + TM_2\hat{\lambda}_n^{1/2}\right]2M_2T \leq C\Big[\frac{1}{a_n} + \hat{\lambda}_n^{1/2}\Big],
\end{aligned}$$

where C is a constant depending only on T, M_1, M_2, and M_3. Let us now define

$$\Upsilon^*_{M_1,M_2,M_3} := \big\{\upsilon_{\theta,m} : (\theta, m) \in \mathcal{C}^*_{M_1,M_2,M_3}\big\}.$$

Thus using arguments similar to (3.91) and (3.92) and the maximal inequality in Lemma

19.35 of [van der Vaart, 1998b], we have

$$\begin{aligned}\mathbb{P}\Big(\sup_{f\in\Upsilon_{M_1,M_2,M_3}(n)}|\mathbb{G}_n f|>\frac{\delta}{2}\Big) &\lesssim 2\delta^{-1}\mathbb{E}\Big(\sup_{f\in\Upsilon_{M_1,M_2,M_3}(n)}|\mathbb{G}_n f|\Big)\\ &\lesssim C\delta^{-1}J_{[]}\Big(C\Big[\frac{1}{a_n}+\hat{\lambda}_n^{1/2}\Big],\Upsilon_{M_1,M_2,M_3}(n),\|\cdot\|_{2,\infty}\Big),\\ &\lesssim C\delta^{-1}J_{[]}\Big(C\Big[\frac{1}{a_n}+\hat{\lambda}_n^{1/2}\Big],\Upsilon^*_{M_1,M_2,M_3},\|\cdot\|_{2,\infty}\Big),\end{aligned}$$

where C is a constant depending only on M_1, M_2, and M_3. Here, the last inequality is true because $\Upsilon_{M_1,M_2,M_3}(n)\subset\Upsilon^*_{M_1,M_2,M_3}$. Thus, to prove the theorem is it enough to show that, $J_{[]}(\gamma,\Upsilon^*_{M_1,M_2,M_3},\|\cdot\|_{2,\infty})\leq\gamma^{1/2}$, for all $\gamma>0$, which is implied by

$$N_{[]}(\varepsilon,\Upsilon^*_{M_1,M_2,M_3},\|\cdot\|_{2,\infty})\lesssim\exp\left(\frac{c}{\varepsilon}+\frac{c}{\sqrt{\varepsilon}}\right)\varepsilon^{-5d},\tag{4.123}$$

where c is a constant depending only on d, M_1, M_2, and M_3. In the following, we show (4.123). Observe that by an argument similar to the proof of Lemma 21, we have

$$N(\varepsilon,\{m_0\circ\theta_0-m\circ\theta:(\theta,m)\in\mathcal{C}^*_{M_1,M_2,M_3}\},\|\cdot\|_\infty)\lesssim\exp(c/\varepsilon)\varepsilon^{-d}.$$

For simplicity of notation let us define

$$V_{\theta,m}(x):=m'(\theta^\top x)x-(m_0'\,h_{\theta_0})(\theta^\top x).$$

Observe that by definition of $v_{\theta,m}$ (see (4.117)) and Lemma 9.25 of [Kosorok, 2008] (for the entropy of product of classes of uniformly bounded functions) to prove (4.123), it is enough to show that

$$N(\varepsilon,\{H_\theta^\top V_{\theta,m}:(\theta,m)\in\mathcal{C}^*_{M_1,M_2,M_3}\},\|\cdot\|_{2,\infty})\lesssim\varepsilon^{-4d}\exp(c/\sqrt{\varepsilon}).\tag{4.124}$$

We will prove (4.124) by constructing a cover for $\{H_\theta^\top V_{\theta,m}:(\theta,m)\in\mathcal{C}^*_{M_1,M_2,M_3}\}$. By Lemma 23, we have

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

where c is a constant depending only on M_1, M_2, and M_3. Let us denote the functions in the ε-cover and their anti-derivatives by $l_1,\ldots,l_t$ and $r_1,\ldots,r_t$, i.e., $l_i=r_i'$ for $1\leq i\leq t$.

By Lemma 28, 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 (3.84) (with ε^2 instead of ε). We now show that $\{H_{\theta_i} V_{\theta_i, r_j}\}_{1 \le i \le s, 1 \le j \le t}$ forms a $\|\cdot\|_{2,\infty}$ cover for $\{H_\theta^\top V_{\theta,m}(x) : (\theta, m) \in \mathcal{C}^*_{M_1,M_2,M_3}\}$.

Fix $(\theta, m) \in \mathcal{C}^*_{M_1,M_2,M_3}$, without loss of generality assume that the function nearest to m' in the ε-cover 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.,

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

Observe that

$$|H_\theta^\top V_{\theta,m}(x) - H_{\theta_1}^\top V_{\theta_1,r_1}(x)| \le |H_\theta^\top V_{\theta,m}(x) - H_{\theta_1}^\top V_{\theta,m}(x)| + |H_{\theta_1}^\top V_{\theta,m}(x) - H_{\theta_1}^\top V_{\theta_1,r_1}(x)|$$

$$\le \varepsilon^2 |V_{\theta,m}(x)| + |V_{\theta,m}(x) - V_{\theta_1,r_1}(x)|. \tag{4.125}$$

Furthermore, we have

$$\begin{aligned}
&|V_{\theta,m}(x) - V_{\theta_1,r_1}(x)| \\
&\le \big|m'(\theta^\top x)x - (m_0' \, h_{\theta_0})(\theta^\top x) - r_1'(\theta_1^\top x)x + (m_0' \, h_{\theta_0})(\theta_1^\top x)\big| \\
&\le T\big|m'(\theta^\top x) - r_1'(\theta_1^\top x)\big| + \big|(m_0' \, h_{\theta_0})(\theta^\top x) - (m_0' \, h_{\theta_0})(\theta_1^\top x)\big| \\
&\le T\big|m'(\theta^\top x) - m'(\theta_1^\top x)\big| + T\big|m'(\theta_1^\top x) - r_1'(\theta_1^\top x)\big| \\
&\quad + \big|(m_0' \, h_{\theta_0})(\theta^\top x) - (m_0' \, h_{\theta_0})(\theta_1^\top x)\big| \\
&\le TM_3\big|\theta^\top x - \theta_1^\top x\big|^{1/2} + T\varepsilon \\
&\quad + \big|(m_0' \, h_{\theta_0})(\theta^\top x) - (m_0' \, h_{\theta_0})(\theta_1^\top x)\big| \lesssim \varepsilon.
\end{aligned} \tag{4.126}$$

Thus combining (4.125),(4.126), and the fact that $|V_{\theta,m}(x)| \le 2TM_2$, we have

$$\|H_\theta^\top V_{\theta,m} - H_{\theta_1}^\top V_{\theta_1,r_1}\|_{2,\infty} \le \varepsilon. \qquad \square$$

4.13 Proof of Results in Section 4.5.3

4.13.1 Proof of Theorem 35

Note that $\xi_t(\theta, m)$ is a uniformly Lipschitz convex function for $t \in \mathbb{R}^{d-1}$ such that $|t - \theta|$ sufficiently close to zero, as both m' and ψ are nondecreasing functions. Since $|m'|_\infty$ is

bounded by L, so is $\beth_t$. By definition we have

$$\xi_t(\theta, m)(s^\top x) = \int_{s_0}^{s^\top x} \beth_t(y; \theta, m)dy + (t-\theta)^\top \big[(m'(s_0) - m_0'(s_0))k(s_0) + m_0'(s_0)h_{\theta_0}(s_0)\big] + m(s_0).$$

We have that $\psi_{\theta,\theta}(u + (\theta - \theta)k(u)) = u$, $\forall u \in D$. Hence,

$$\begin{aligned}
\xi_\theta(\theta, m)(\theta^\top x) &= \int_{s_0}^{\theta^\top x} \beth_\theta(y; \theta, m)dy + m(s_0) \\
&= \int_{s_0}^{\theta^\top x} m' \circ \psi_{\theta,\theta}(y)dy + m(s_0) \\
&= \int_{s_0}^{\theta^\top x} m'(y)dy + m(s_0) \\
&= m(\theta^\top x).
\end{aligned}$$

Observe that,

$$\begin{aligned}
&\frac{\partial}{\partial t}\xi_t(\theta, m)(t^\top x) \\
&= \frac{\partial}{\partial t}\left\{\int_{s_0}^{t^\top x} \beth_t(y; \theta, m)dy\right\} + (m'(s_0) - m_0'(s_0))k(s_0) + m_0'(s_0)h_{\theta_0}(s_0) \qquad (4.127) \\
&= \frac{\partial}{\partial t}\left\{\int_{s_0}^{t^\top x} m' \circ \psi_{\theta,t}(y + (\theta - t)k(y))dy\right\} + (m'(s_0) - m_0'(s_0))k(s_0) + m_0'(s_0)h_{\theta_0}(s_0)
\end{aligned}$$

We next evaluate the first term on the right hand side of the above display. But first, we introduce some notations. Let us define,

$$\begin{aligned}
&\psi'_{\theta,t}(y) := \frac{\partial}{\partial y}\psi_{\theta,t}(y), \qquad \psi''_{\theta,t}(y) := \frac{\partial}{\partial y}\psi'_{\theta,t}(y), \qquad \dot{\psi}_{\theta,t}(y) := \frac{\partial}{\partial t}\psi_{\theta,t}(y), \\
&\phi_{\theta,t}(y) := \psi_{\theta,t}(y + (\theta - t)k(y)), \\
&\phi'_{\theta,t}(y) = \psi'_{\theta,t}(y + (\theta - t)k(y))(1 + (\theta - t)k'(y)).
\end{aligned}$$

Now, observe that

$$\begin{aligned}
\frac{\partial \phi_{\theta,t}(y)}{\partial t} &= \dot{\psi}_{\theta,t}(y + (\theta - t)k(y)) - k(y)\psi'_{\theta,t}(y + (\theta - t)k(y)), \\
\frac{\partial \phi'_{\theta,t}(y)}{\partial t} &= (1 + (\theta - t)k'(y))\left[\frac{\partial \psi'_{\theta,t}}{\partial t}(y + (\theta - t)k(y)) - k(y)\psi''_{\theta,t}(y + (\theta - t)k(y))\right], \\
&\quad - k'(y)\psi'_{\theta,t}(y + (\theta - t)k(y)) \\
\frac{\partial \phi_{\theta,t}(t^\top x)}{\partial t} &= \dot{\psi}_{\theta,t}(t^\top x + (\theta - t)k(t^\top x)), \\
&\quad + \psi'_{\theta,t}(t^\top x + (\theta - t)k(t^\top x))(x_2 - k(t^\top x) + (\theta - t)k'(t^\top x)x_2).
\end{aligned}$$

Now, we evaluate the first term on the right hand side of (4.127). Note that

$$\begin{aligned}
&\frac{\partial}{\partial t}\Big\{\int_{s_0}^{t^\top x} m'\circ\psi_{\theta,t}(y+(\theta-t)k(y))dy\Big\}\\
&=\frac{\partial}{\partial t}\Big\{\int_{s_0}^{t^\top x} m'\circ\phi_{\theta,t}(y)dy\Big\}\\
&=\frac{\partial}{\partial t}\Big\{\int_{\phi_{\theta,t}(s_0)}^{\phi_{\theta,t}(t^\top x)} \frac{m'(u)}{\phi'_{\theta,t}\circ\phi^{-1}_{\theta,t}(u)}du\Big\}\\
&=\frac{m'\circ\phi_{\theta,t}(t^\top x)}{\phi'_{\theta,t}(t^\top x)}\frac{\partial\phi_{\theta,t}(t^\top x)}{\partial t}-\frac{m'\circ\phi_{\theta,t}(s_0)}{\phi'_{\theta,t}(s_0)}\frac{\partial\phi_{\theta,t}(s_0)}{\partial t}\\
&\quad-\int_{\phi_{\theta,t}(s_0)}^{\phi_{\theta,t}(t^\top x)}\frac{m'(u)}{\big[\phi'_{\theta,t}\circ\phi^{-1}_{\theta,t}(u)\big]^2}\frac{\partial\phi'_{\theta,t}\circ\phi^{-1}_{\theta,t}(u)}{\partial t}du\\
&=\frac{m'\circ\phi_{\theta,t}(t^\top x)}{\phi'_{\theta,t}(t^\top x)}\frac{\partial\phi_{\theta,t}(t^\top x)}{\partial t}-\frac{m'\circ\phi_{\theta,t}(s_0)}{\phi'_{\theta,t}(s_0)}\frac{\partial\phi_{\theta,t}(s_0)}{\partial t}\\
&\quad-\int_{s_0}^{t^\top x}\frac{m'\circ\phi_{\theta,t}(y)}{\big[\phi'_{\theta,t}(y)\big]^2}\frac{\partial\phi'_{\theta,t}(y)}{\partial t}\phi'_{\theta,t}(y)dy.
\end{aligned}$$

Thus, we have

$$\begin{aligned}
&\frac{\partial}{\partial t}\Big\{\int_{s_0}^{t^\top x} m'\circ\psi_{\theta,t}(y+(\theta-t)k(y))dy\Big\}\Big|_{t=\theta}\\
&=m'(\theta^\top x)(x_2-k(\theta^\top x))-m'(s_0)[-k(s_0)]-\int_{s_0}^{\theta^\top x}m'(y)[-k'(y)]dy.
\end{aligned}$$

We now show that the score function of the sub-model $\{t,\xi_t(\theta,m)\}$ is $\mathfrak{S}_{\theta,m}(x,y)$, i.e.,

$$\begin{aligned}
&-\frac{1}{2}\frac{\partial}{\partial t}\{(y-\xi_t(\theta,m)(t^\top x))^2\}\Big|_{t=\theta}\\
&=(y-\xi_t(\theta,m)(t^\top x))\,\frac{\partial\xi_t(\theta,m)(t^\top x)}{\partial t}\Big|_{t=\theta}\\
&=\big(y-m(\theta^\top x)\big)\Big[m'(\theta^\top x)(x_2-k(\theta^\top x))-m'(s_0)[-k(s_0)]-\int_{s_0}^{\theta^\top x}m'(y)[-k'(y)]dy\\
&\qquad+(m'(s_0)-m'_0(s_0))k(s_0)+m'_0(s_0)h_{\theta_0}(s_0)\Big]\\
&=\big(y-m(\theta^\top x)\big)\Big[m'(\theta^\top x)x_2-m'(\theta^\top x)k(\theta^\top x)+\int_{s_0}^{\theta^\top x}m'(y)k'(y)dy\\
&\qquad+m'_0(s_0)k(s_0)-m'_0(s_0)h_{\theta_0}(s_0)\Big]
\end{aligned}$$

Next, we show that $\mathfrak{S}_{\theta_0,m_0} = \ell_{\theta_0,m_0}$. By definition, it is enough to show that,

$$
\begin{aligned}
& m_0'(\theta_0^\top x)(x_2 - k(\theta^\top x)) + m_0'(s_0)k(s_0) + \int_{s_0}^{\theta^\top x} m_0'(y)k'(y)dy = m_0'(\theta_0^\top x)(x_2 - h_{\theta_0}(\theta_0^\top x)) \\
& \Rightarrow m_0'(\theta_0^\top x)k(\theta_0^\top x) - m_0'(s_0)k(s_0) - \int_{s_0}^{\theta^\top x} m_0'(y)k'(y)dy = m_0'(\theta_0^\top x)(h_{\theta_0}(\theta_0^\top x)) \\
& \Rightarrow \int_{s_0}^{\theta_0^\top x} \frac{\partial m_0'(y)k(y)}{\partial y}dy - \int_{s_0}^{\theta^\top x} m_0'(y)k'(y)dy = m_0'(\theta_0^\top x)(h_{\theta_0}(\theta_0^\top x)) \\
& \Rightarrow \int_{s_0}^{\theta_0^\top x} \frac{\partial m_0'(y)k(y)}{\partial y}dy - \int_{s_0}^{\theta^\top x} m_0'(y)k'(y)dy = m_0'(\theta_0^\top x)(h_{\theta_0}(\theta_0^\top x)) \\
& \Rightarrow \int_{s_0}^{\theta_0^\top x} m_0''(y)k(y)dy = m_0'(\theta_0^\top x)(h_{\theta_0}(\theta_0^\top x)).
\end{aligned}
$$

As the score of the sub-model is the efficient score at the truth, we have that $\zeta_t(\theta, m)$ is an approximately least favorable subprovided model.

4.13.2 Proof of Lemma 37

Recall that $U_{\theta,m}(x) = H_\theta^\top \big[\int_{s_0}^{\theta^\top x} \big[m'(u) - m_0'(u)\big]k'(u)du + (m_0'(\theta^\top x) - m'(\theta^\top x))k(\theta^\top x)\big]$; see (4.35). Observe that D is bounded set, $\sup_{u\in D}(|k(u)| + |k'(u)|) \le M^*$ and $\|m'\|_\infty \le L$. Hence

$$
\begin{aligned}
|U_{\theta,m}(x)| &\le M^* \int_{s_0}^{\theta^\top x} |m'(u) - m_0'(u)|du + M^*|m_0'(\theta^\top x) - m'(\theta^\top x)| \\
&\le 2LM^*|\theta^\top x - s_0| + 2M^*L \le 4LM^*T + 2M^*L := V^*.
\end{aligned}
$$

Now we will try to find the entropy of $\mathcal{W}_{M_1}(n)$. As the definition of $U_{\theta,m}$ involves m' to find entropy of the class of functions $\mathcal{W}_{M_1}^*$, we need the entropy of

$$
\begin{aligned}
\mathcal{H}^* := \{ & f : \chi \to \mathbb{R} |\, f(x) = g(\theta^\top x), \theta \in \Theta \text{ and} \\
& g : \mathcal{D} \to \mathbb{R} \text{ is an increasing function and } \|g\|_\infty \le S\}.
\end{aligned}
$$

The following lemma does this.

Lemma 60. $\log N_{[\,]}(\varepsilon, \mathcal{H}^*, L_2(P_{\theta_0,m_0})) \lesssim \varepsilon^{-1}$.

Proof. Observe that by Lemma 4.1 of [Pollard, 1990] we can get $\theta_1, \theta_2, \ldots, \theta_{N_{\eta_1}}$, with $N_{\eta_1} \lesssim \eta_1^{-d}$ such that for every $\theta \in \Theta$, there exists a j satisfying $|\theta - \theta_j| \le \eta_1/T$ and

$$
|\theta^\top x - \theta_j^\top x| \le |\theta - \theta_j| \cdot |x| \le \eta_1 \quad \forall x \in \chi.
$$

Thus for every $\theta \in \Theta$, we can find a j such that $\theta_j^\top x - \eta_1 \leq \theta^\top x \leq \theta_j^\top x + \eta_1, \forall x \in \mathcal{X}$. For simplicity, define $t_j^{(1)}(x) = \theta_j^\top x - \eta_1$ and $t_j^{(2)}(x) = \theta_j^\top x + \eta_1$. Let us define

$$\mathcal{G}^* := \{g | \, g : \mathcal{D} \to \mathbb{R} \text{ is a uniformly bounded increasing function and } \|g\|_\infty \leq S\}.$$

Recall that $\mathbf{m}$ denotes the Lebesgue measure on D. By Theorem 2.7.5 of [van der Vaart and Wellner, 1996], we have that $N_{[\,]}(\eta_2, \mathcal{G}^*, L_2(\mathbf{m})) \lesssim \exp(\eta_2^{-1})$, i.e., there exists $[\ell_1, u_1], \ldots, [\ell_{M_{\eta_2}}, u_{M_{\eta_2}}]$ with $\ell_i \leq u_i$, $\int_D |u_i(t) - \ell_i(t)|^2 dt \leq \eta_2^2$ and $M_{\eta_2} \lesssim \exp(\eta_2^{-1})$ such that for every $g \in \mathcal{G}^*$, we can find a $k \leq M_{\eta_2}$ such that $\ell_k \leq g \leq u_k$. Without loss of generality we can assume that both ℓ_i, u_i are increasing and bounded for all $1 \leq i \leq M_{\eta_2}$.

Fix any function $g \in \mathcal{G}^*$ and $\theta \in \Theta$. Let $|\theta_j - \theta| \leq \eta_1$ and $[\ell_k, u_k]$ is the η_2-bracket for g, then for every $x \in \mathcal{X}$,

$$\ell_k(\theta_j^\top x - \eta_1) \leq \ell(\theta^\top x) \leq g(\theta^\top x) \leq u_k(\theta^\top x) \leq u_k(\theta_j^\top x + \eta_1),$$

where the outer inequalities follows from the fact that both ℓ_k and u_k are increasing functions. Proof of Lemma 60 will be complete if we can show that

$$\{[\ell_k \circ t_j^{(1)}, u_k \circ t_j^{(2)}] : 1 \leq j \leq N_{\eta_1}, 1 \leq k \leq M_{\eta_2}\},$$

form a $L_2(P_{\theta_0, m_0})$ bracket for $\mathcal{H}^*$. To complete the proof, we now choose η_1 and η_2 such that the $\|\cdot\|$-length of each bracket of $\mathcal{H}^*$ is bounded by ε. Note that by the triangle inequality, we have

$$\|u_k \circ t_j^{(2)} - \ell_k \circ t_j^{(1)}\| \leq \|u_k \circ t_j^{(2)} - \ell_k \circ t_j^{(2)}\| + \|\ell_k \circ t_j^{(2)} - \ell_k \circ t_j^{(1)}\|. \tag{4.128}$$

Assuming that the density (with respect to the Lebesgue measure) of $X^\top \theta$ is uniformly bounded above (by C), we get that

$$\|u_k \circ t_j^{(2)} - \ell_k \circ t_j^{(2)}\|^2 = \int [u_k(r) - \ell_k(r)]^2 \, dP_j(r) \leq C \int [u_k(r) - \ell_k(r)]^2 \, dr \leq C\eta_2^2.$$

For the second term in (4.128), we first approximate the lower bracket ℓ_k by an increasing step (piecewise constant) function. Such an approximation is possible since the set of all simple functions is dense in $L_2(P_{\theta_0, m_0})$; see Lemma 4.2.1 of [Bogachev, 2007]. Since ℓ_k is bounded (by S say), we can get an increasing step function $A : D \to [-S, S]$, such

that $\int\{\ell_k(r) - A(r)\}^2 dr \le \eta_2^2$. Let $v_1 > \cdots > v_{A_d}$ denote an points of discontinuity of A. Then for every $r \in D$, we can write

$$A(r) = -S + \sum_{i=1}^{A_d} c_i \mathbb{1}_{\{r \le v_i\}}, \text{ where } c_i > 0 \text{ and } \sum_{i=1}^{A_d} c_i \le 2S.$$

Using triangle inequality, we get that

$$\begin{aligned}\|\ell_k \circ t_j^{(2)} - \ell_k \circ t_j^{(1)}\| &\le \|\ell_k \circ t_j^{(2)} - A \circ t_j^{(2)}\| + \|A \circ t_j^{(2)} - A \circ t_j^{(1)}\| + \|A \circ t_j^{(1)} - \ell_k \circ t_j^{(1)}\| \\ &\le \sqrt{C}\eta_2 + \|A \circ t_j^{(2)} - A \circ t_j^{(1)}\| + \sqrt{C}\eta_2,\end{aligned}$$

where C is the (uniform) upper bound on the density of $X^\top \theta_j$. Now observe that

$$\begin{aligned}\|A \circ t_j^{(2)} - A \circ t_j^{(1)}\|^2 &= \mathbb{E}\left[\sum_{i=1}^{A_d} c_i \left(\mathbb{1}_{\{X^\top\theta_j + \eta_1 \le v_i\}} - \mathbb{1}_{\{X^\top\theta_j + \eta_1 \le v_i\}}\right)\right]^2 \\ &\le 2S\mathbb{E}\left|\sum_{i=1}^{A_d} c_i \left(\mathbb{1}_{\{X^\top\theta_j + \eta_1 \le v_i\}} - \mathbb{1}_{\{X^\top\theta_j + \eta_1 \le v_i\}}\right)\right| \\ &\le 2S\sum_{i=1}^{A_d} c_i \mathbb{P}(X^\top\theta_j - \eta_1 < v_i \le X^\top\theta_j + \eta_1) \\ &\le 2S\sum_{i=1}^{A_d} c_i \mathbb{P}(v_i - \eta_1 \le X^\top\theta_j < v_i + \eta_1) \\ &\le 2S\sum_{i=1}^{A_d} c_i (2C\eta_1) \le 8CS^2\eta_1.\end{aligned}$$

Therefore, we get that

$$\|u_k \circ t_j^{(2)} - \ell_k \circ t_j^{(1)}\| \le 3\sqrt{C}\eta_2 + 2\sqrt{2C}S\sqrt{\eta_1} \le \varepsilon,$$

by taking $\eta_2 = \varepsilon/(6\sqrt{C})$ and $\eta_1 = \varepsilon^2/(32CS^2)$. Hence the bracketing entropy of $\mathcal{H}^*$ satisfies

$$\log N_{[\,]}(\varepsilon, \mathcal{H}^*, \|\cdot\|) \lesssim \frac{6\sqrt{C}}{\varepsilon} - 2d\log(\varepsilon) - d\log(32CS^2) \lesssim \varepsilon^{-1},$$

for sufficiently small ε. □

We will now use Lemma 60 to prove Lemma 37. Fix $(\theta, m) \in \mathcal{C}_{M_1}(n)$. By definition we have that both $H_\theta^\top k$ and $H_\theta^\top k'$ are coordinate-wise bounded functions; see (4.20) and $H_\theta^\top k(u) + M^*\mathbf{1} \succeq 0$ and $H_\theta^\top k'(u) + M^*\mathbf{1} \succeq 0$ (where $\mathbf{1}$ is the vector of all 1's

and $\succeq$ represents coordinate-wise inequalities). Using these, we can write $U_{\theta,m}(x) = U^{(1)}_{\theta,m}(x) - U^{(2)}_{\theta,m}(x) + U^{(3)}_{\theta,m}(x)$ where

$$\begin{aligned}
U^{(1)}_{\theta,m}(x) &= \int_{s_0}^{\theta^\top x} [m'(u) - m_0'(u)](H_\theta^\top k'(u) + M^*\mathbf{1})du, \\
U^{(2)}_{\theta,m}(x) &= M^*\mathbf{1}\int_{s_0}^{\theta^\top x} [m'(u) - m_0'(u)]du, \\
U^{(3)}_{\theta,m}(x) &= (m_0'(\theta^\top x) - m'(\theta^\top x))H_\theta^\top k(\theta^\top x).
\end{aligned}$$

We will find $c_i\eta$-brackets (with respect to $\|\cdot\|_{2,P_{\theta_0,m_0}}$) for $U^{(i)}_{\theta,m}, i = 1, 2$, and 3 separately and combine them to get a $c\eta$-bracket (with respect to $L_2(P_{\theta_0,m_0})$) bracket for $U_{\theta,m}$, where c, c_1, c_2, and c_3 are constants depending only on S, T, d, M^*, L and L_0. By Lemma 60 there exists a $N'_\eta \le \exp(\eta^{-1})$ such that $\{(\ell_k, u_k)\}_{1\le k\le N'_\eta}$ form a η-bracket (with respect to $L_2(P_{\theta_0,m_0})$ norm) for $\{m'(\theta^\top x) : (\theta, m) \in \mathcal{C}^*_{M_1}\}$, i.e., for all $x \in \mathcal{X}$

$$\ell_k(x) \le m'(\theta^\top x) \le u_k(x), \tag{4.129}$$

and $\|u_k - \ell_k\| \le C\eta$ for some constant C. Similarly by Lemma 28, we can find a $\theta_1, \theta_2, \ldots, \theta_{N_\eta}$ with $N_\eta \le C\eta^{-2d}$ for some constant C such that for every $\theta \in \Theta \cap B_{\theta_0}(1/2)$, there exists a θ_j such that

$$|\theta - \theta_j| \le \eta/T, \ \|H_\theta - H_{\theta_j}\|_2 \le \eta/T, \text{ and } |\theta^\top x - \theta_j^\top x| \le \eta, \ \forall x \in \mathcal{X}.$$

We first find a $\|\cdot\|_{2,P_{\theta_0,m_0}}$ bracket for $U^{(3)}_{\theta,m}$ using Lemma 9.25 of [Kosorok, 2008]. For this application, we need to find bracketing entropy for the class of functions,

$$\{H_\theta^\top k(\theta^\top \cdot) : (\theta, m) \in \mathcal{C}_{M_1}(n)\}, \text{ and } \{m_0'(\theta^\top \cdot) - m'(\theta^\top \cdot) : (\theta, m) \in \mathcal{C}_{M_1}(n)\}.$$

As m_0' is an increasing function bounded by L_0 (see **(L1)**), we have that

$$m_0'(\theta_j^\top x - \eta_1) \le m_0'(\theta^\top x) \le m_0'(\theta_j^\top x + \eta_1).$$

Thus by (4.129), we have

$$m_0'(\theta_j^\top x - \eta_1) - u_k(x) \preceq U^{(4)}_{\theta,m}(x) \preceq m_0'(\theta_j^\top x + \eta_1) - \ell_k(x).$$

The length of the above bracket is given by

$$\begin{aligned}
&\|m_0'(\theta_j^\top \cdot + \eta_1) - \ell_k - m_0'(\theta_j^\top \cdot - \eta_1) + u_k\|_{2,P_{\theta_0,m_0}} \\
&\leq \left[P_{\theta_0,m_0}|m_0'(\theta_j^\top X + \eta_1) - m_0'(\theta_j^\top X - \eta_1)|^2\right]^{1/2} + \|u_k - \ell_k\| \\
&\leq 2\|m_0''\|_\infty \eta + \eta = (2\|m_0''\|_\infty + 1)\eta.
\end{aligned}$$

Thus

$$N_{[]}(\eta, \{m_0'(\theta^\top \cdot) - m'(\theta^\top \cdot) : (\theta, m) \in \mathcal{C}_{M_1}(n)\}, \|\cdot\|) \lesssim \exp(\eta^{-1})\eta^{-2d} \tag{4.130}$$

Recall that $\|k\|_{2,\infty} + \|k'\|_{2,\infty} \leq M^*$. To find the $\|\cdot\|_{2,P_{\theta_0,m_0}}$ bracket for $\{H_\theta^\top k(\theta^\top x) : (\theta, m) \in \mathcal{C}_{M_1}(n)\}$ observe that

$$\begin{aligned}
|H_\theta^\top k(\theta^\top x) - H_{\theta_j}^\top k(\theta_j^\top x)| &\leq |H_\theta^\top k(\theta^\top x) - H_{\theta_j}^\top k(\theta^\top x)| + |H_{\theta_j}^\top k(\theta^\top x) - H_{\theta_j}^\top k(\theta_j^\top x)| \\
&\leq \eta\|k\|_{2,\infty}/T + \|k'\|_{2,\infty}\eta \leq 2\eta M^*.
\end{aligned}$$

This leads to the brackets

$$H_{\theta_j}^\top k(\theta_j^\top x) - 2\eta M^* \mathbf{1} \preceq H_\theta^\top k(\theta^\top x) \preceq H_{\theta_j}^\top k(\theta_j^\top x) + 2\eta M^* \mathbf{1},$$

with $\|\cdot\|_{2,P_{\theta_0,m_0}}$-length $4\eta M^*\sqrt{d-1}$. Thus

$$N_{[]}(\eta, \{H_\theta^\top k(\theta^\top \cdot) : (\theta, m) \in \mathcal{C}_{M_1}(n)\}, \|\cdot\|_{2,P_{\theta_0,m_0}}) \lesssim \exp(\eta^{-1})\eta^{-2d} \tag{4.131}$$

Thus by Lemma 9.25 of [Kosorok, 2008], (4.130) and (4.131), we have that

$$N_{[]}(\eta, \{[m_0'(\theta^\top \cdot) - m'(\theta^\top \cdot)]H_\theta^\top k(\theta^\top \cdot) : (\theta, m) \in \mathcal{C}_{M_1}(n)\}, \|\cdot\|_{2,P_{\theta_0,m_0}}) \lesssim \exp(\eta^{-1})\eta^{-4d}$$

For treating $U_{\theta,m}^{(1)}$ and $U_{\theta,m}^{(2)}$, we take s_0 to be the minimum point of the set $\{\theta^\top x : \theta \in \Theta \cap B_{\theta_0}(1/2), x \in \mathcal{X}\}$. By Theorem 2.7.5 of [van der Vaart and Wellner, 1996], we have

$$\log N_{[]}(\eta, \{m' : m \in \mathcal{C}_{M_1}^{m*}\}, L_2(\mathbf{m})) \lesssim \eta^{-1}.$$

Let $[m_L, m_U]$ be the η-bracket of m', i.e, $m_L(u) \leq m'(u) \leq m_U(u)$ for all u and $\int_D |m_U(t) - m_L(t)|^2 dt \leq \eta^2$. As θ_j satisfies $|\theta - \theta_j| \leq \eta/T$, by Lemma 16 we have

$$|H_\theta^\top k'(u) - H_{\theta_j}^\top k'(u)| \leq |k'(u)|\eta/T \leq M^*\eta/T.$$

This implies

$$H_{\theta_j}^\top k'(u) + M^* \mathbf{1}\,(1-\eta/T) \preceq H_\theta^\top k'(u) + M^* \mathbf{1} \preceq H_{\theta_j}^\top k'(u) + M^* \mathbf{1}\,(1+\eta/T)\,.$$

The $\|\cdot\|_{2,P_{\theta_0,m_0}}$-length of this bracket is given by $2M^*\eta/T$. Since $H_\theta^\top k'(u) + M^*\mathbf{1} \succeq 0$ for all θ, u, we can take the brackets to be

$$\{H_{\theta_j}^\top k'(u) + M^*(1-\eta/T)\mathbf{1}\} \vee 0 \preceq H_\theta^\top k'(u) + M^*\mathbf{1} \preceq \{H_{\theta_j}^\top k'(u) + M^*(1+\eta/T)\mathbf{1}\} \wedge (2M^*).$$

From the brackets $[m_L, m_U]$ of m', we get that

$$m_L(u) - m_0'(u) \le m'(u) - m_0'(u) \le m_U(u) - m_0'(u).$$

Combining the above two displays and the fact that $\theta^\top x > s_0$, we see that for every $x \in \mathcal{X}$ and $\theta \in \Theta \cap B_{\theta_0}(1/2)$,

$$\begin{aligned} \int_{s_0}^{\theta^\top x} [m_L(u) - m_0'(u)](\{H_{\theta_j}^\top k'(u) + M^*(1-\eta/T)\mathbf{1}\} \vee 0)du \preceq U_{\theta,m}^{(1)}(x), \\ U_{\theta,m}^{(1)}(x) \preceq \int_{s_0}^{\theta^\top x} [m_U(u) - m_0'(u)](\{H_{\theta_j}^\top k'(u) + M^*(1+\eta/T)\mathbf{1}\} \wedge (2M^*))du. \end{aligned} \tag{4.132}$$

These bounding functions are not brackets since they depend on θ (in the limits of the integral). Since m_L, m_U, m' are all bounded by L, we get that

$$\int_{\theta_j^\top x}^{\theta^\top x} |m_U(u) - m_0'(u)|(\{H_{\theta_j}^\top k'(u) + M^*(1+\eta/T)\mathbf{1}\} \wedge (2M^*))du \preceq 4M^* L|\theta^\top x - \theta_j^\top x|\mathbf{1} \preceq 4M^* L\eta \mathbf{1},$$

(coordinate-wise) and similarly,

$$\int_{\theta_j^\top x}^{\theta^\top x} |m_L(u) - m_0'(u)|(\{H_{\theta_j}^\top k'(u) + M^*(1-\eta/T)\mathbf{1}\} \vee 0)du \preceq 4M^* L\eta \mathbf{1}.$$

Therefore, from the inequalities (4.132), we get the brackets $[M_L^{(1)}, M_U^{(1)}]$ for $U_{\theta,m}^{(1)}$ as

$$\begin{aligned} M_L^{(1)} &= \int_{s_0}^{\theta_j^\top x} [m_L(u) - m_0'(u)](\{H_{\theta_j}^\top k'(u) + M^*(1-\eta/T)\mathbf{1}\} \vee 0)du - 4M^* L\eta\mathbf{1}, \\ M_U^{(1)} &= \int_{s_0}^{\theta_j^\top x} [m_U(u) - m_0'(u)](\{H_{\theta_j}^\top k'(u) + M^*(1+\eta/T)\mathbf{1}\} \wedge (2M^*))du + 4M^* L\eta\mathbf{1}. \end{aligned}$$

The $\|\cdot\|$-length of this bracket is bounded as follows: by triangle inequality

$$
\begin{aligned}
&\|M_U^{(1)} - M_L^{(1)}\|_{2,P_{\theta_0,m_0}} \\
&\le 8M^*L\eta\sqrt{d-1} + \left\| \int_{s_0}^{\theta_j^\top \cdot} [m_U(u) - m_L(u)](\{H_{\theta_j}^\top k'(u) + M^*(1+\eta/T)\mathbf{1}\} \wedge (2M^*))du \right\|_{2,P_{\theta_0,m_0}} \\
&\quad + \left\| \int_{s_0}^{\theta_j^\top \cdot} [m_U(u) - m_0'(u)] \times \right. \\
&\qquad \left. \Big[(\{H_{\theta_j}^\top k'(u) + M^*(1+\eta/T)\mathbf{1}\} \wedge (2M^*\mathbf{1})) - (\{H_{\theta_j}^\top k'(u) + M^*(1-\eta/T)\mathbf{1}\} \vee 0) \Big] du \right\|_{2,P_{\theta_0,m_0}} \\
&\le 8M^*L\eta\sqrt{d-1} + \left\| 2M^*\mathbf{1} \int_{s_0}^{\theta_j^\top \cdot} [m_U(u) - m_L(u)]du \right\|_{2,P_{\theta_0,m_0}} \\
&\quad + \left\| 2L \int_{s_0}^{\theta_j^\top \cdot} \Big[(H_{\theta_j}^\top k'(u) + M^*(1+\eta/T)\mathbf{1}) - (H_{\theta_j}^\top k'(u) + M^*(1-\eta/T)\mathbf{1}) \Big] du \right\|_{2,P_{\theta_0,m_0}} \\
&\le 8M^*L\eta\sqrt{d-1} + 2M^*\sqrt{d-1} \left(\int_D (m_U(u) - m_L(u))^2 du \right)^{1/2} + 4M^*L\eta\sqrt{d-1}/T \\
&= \sqrt{d-1}(12M^*L\eta + 2M^*\eta).
\end{aligned}
$$

Thus, we get that $[M_L^{(1)}, M_U^{(2)}]$ is a $(12M^*L+2M^*)\sqrt{d-1}\eta$-bracket for $U_{\theta,m}^{(1)}$ with respect to the $\|\cdot\|_{2,P_{\theta_0,m_0}}$ norm.

Following very similar arguments, we get that $[M_L^{(2)}, M_U^{(2)}]$ form a bracket for $U_{\theta,m}^{(2)}$, where

$$
\begin{aligned}
M_L^{(2)}(x) &= \Big[\int_{s_0}^{\theta_j^\top x} [m_L(u) - m_0'(u)]du - 2L\eta \Big] \mathbf{1}, \\
M_U^{(2)}(x) &= \Big[\int_{s_0}^{\theta_j^\top x} [m_U(u) - m_0'(u)]du + 2L\eta \Big] \mathbf{1}.
\end{aligned}
$$

The $\|\cdot\|_{2,P_{\theta_0,m_0}}$-length of this bracket is

$$
\begin{aligned}
\|M_U^{(2)} - M_L^{(2)}\|_{2,P_{\theta_0,m_0}} &\le 4L\eta\sqrt{d-1} + \sqrt{d-1} \left\| \int_{s_0}^{\theta_j^\top \cdot} (m_U(u) - m_L(u))du \right\| \\
&\le 4L\eta\sqrt{d-1} + \eta\sqrt{d-1} = \sqrt{d-1}(4L+1)\eta.
\end{aligned}
$$

Thus for both $U_{\theta,m}^{(1)}$ and $U_{\theta,m}^{(2)}$, the bracketing number is bounded by a constant multiple of $\exp(\eta^{-1})\eta^{-2d}$. Hence we have (4.41).

Next we show (4.42). Observe that

$$\begin{aligned}\big\|U_{\theta,m}(x)\big\|^2_{2,P_{\theta_0,m_0}} &\le \Big\| \int_{s_0}^{\theta^\top \cdot} [m'(u) - m_0'(u)]k'(u)du\Big\|^2_{2,P_{\theta_0,m_0}} + \Big\|(m' - m_0')(\theta^\top \cdot)k(\theta^\top \cdot)\Big\|^2_{2,P_{\theta_0,m_0}} \\ &\le \Big\| \int_{s_0}^{\theta_0^\top \cdot} [m'(u) - m_0'(u)]k'(u)du\Big\|^2_{2,P_{\theta_0,m_0}} + \Big\| \int_{\theta_0^\top \cdot}^{\theta^\top \cdot} [m'(u) - m_0'(u)]k'(u)du\Big\|^2_{2,P_{\theta_0,m_0}} \\ &\quad + \Big\|(m' - m_0')(\theta^\top \cdot)k(\theta^\top \cdot)\Big\|^2_{2,P_{\theta_0,m_0}} \\ &\le \mathbf{I} + \mathbf{II} + \mathbf{III}.\end{aligned}$$

Observe that

$$\begin{aligned}\mathbf{I} &= \int_\chi \left| \int_{s_0}^{\theta_0^\top X} [m'(u) - m_0'(u)]k'(u)du \right|^2 dP_{\theta_0,m_0} \\ &\le \int_\chi \int_{D_0} [m'(u) - m_0'(u)]^2 |k'(u)|^2 du\, dP_{\theta_0,m_0} \\ &\le \|k'\|^2_{2,\infty} \int_{D_0} [m'(u) - m_0'(u)]^2 du \le \|k'\|^2_{2,\infty} n^{-1/5},\end{aligned} \tag{4.133}$$

and

$$\begin{aligned}\mathbf{II} &= \Big\| \int_{\theta_0^\top \cdot}^{\theta^\top \cdot} [m'(u) - m_0'(u)]k'(u)du\Big\|^2_{2,P_{\theta_0,m_0}} \le L^2 \|k'\|^2_{2,\infty} \|(\theta_0 - \theta)^\top \cdot \|^2 \le L^2 \|k'\|^2_{2,\infty} T^2 |\theta_0 - \theta|^2, \\ \mathbf{III} &= \Big\|(m' - m_0')(\theta^\top \cdot)k(\theta^\top \cdot)\Big\|^2_{2,P_{\theta_0,m_0}} \le \|k'\|^2_{2,\infty} \Big\|(m' - m_0')(\theta^\top \cdot)\Big\|^2 = \|k'\|^2_{2,\infty} n^{-1/5}.\end{aligned} \tag{4.134}$$

Combining (4.133) and (4.134), we have

$$\sup_{(\theta,m)\in \mathcal{C}_{M_1}(n)} \big\|U_{\theta,m}\big\|^2_{2,P_{\theta_0,m_0}} \le 2\|k'\|^2_{2,\infty} n^{-1/5} + L^2 \|k'\|^2_{2,\infty} T^2 n^{-1/5} = K_L^2 n^{-1/5}.$$

4.13.3 Proof of Lemma 38

For every $(\theta, m) \in \mathcal{C}_{M_1}(n)$, note that

$$\big\|(m_0 \circ \theta_0 - m \circ \theta_0)U_{\theta,m}\big\|^2_{2,P_{\theta_0,m_0}} \le 4M_1^2 \big\|U_{\theta,m}\big\|^2_{2,P_{\theta_0,m_0}} \le 4M_1^2 K_L^2 n^{-1/5} = D^2_{M_1} n^{-1/5}.$$

Furthermore, note that $\mathcal{D}^*_{M_1}$ is a class of uniformly bounded functions, i.e.,

$$\big|(m_0 - m)(\theta_0{}^\top x)U_{\theta,m}(x)\big| \le 2M_1 \big|U_{\theta,m}(x)\big| \le 2M_1 V^*.$$

and by Lemma 51 there exists a constant c depending only on M_1 and L such that

$$N(\varepsilon, \{(m_0 \circ \theta_0 - m \circ \theta_0 : m \in \mathcal{C}_{M_1}^{m*}\}, \|\cdot\|_\infty) = N(\varepsilon, \mathcal{C}_{M_1}^{m*}, \|\cdot\|_\infty) \le c \exp(c/\sqrt{\varepsilon}).$$

By Lemma 37 and Lemma 9.25 of [Kosorok, 2008] (for bracketing entropy of product of uniformly bounded function classes), we have

$$N_{[]}(\varepsilon, \mathcal{D}_{M_1}(n), \|\cdot\|_{2,P_{\theta_0,m_0}}) \le N_{[]}(\varepsilon, \mathcal{D}_{M_1}^*, \|\cdot\|_{2,P_{\theta_0,m_0}}) \le c\varepsilon^{-2d} \exp\left(\frac{c}{\sqrt{\varepsilon}} + \frac{c}{\varepsilon}\right).$$

It follows that

$$J_{[]}(\gamma, \mathcal{D}_{M_1}(n), \|\cdot\|_{2,P_{\theta_0,m_0}}) \lesssim \gamma^{\frac{1}{2}}.$$

Using arguments similar to (3.91) and the maximal inequality in Lemma 3.4.2 of [van der Vaart and Wellner, 1996] (for uniformly bounded function classes), we have

$$\begin{aligned}
&\mathbb{P}\left(\sup_{f \in \mathcal{D}_{M_1}(n)} |\mathbb{G}_n f| > \delta\right) \\
&\lesssim \delta^{-1} \mathbb{E}\left(\sup_{f \in \mathcal{D}_{M_1}(n)} |\mathbb{G}_n f|\right) \\
&\lesssim \delta^{-1} J_{[]}(D_{M_1} n^{-1/10}, \mathcal{D}_{M_1}(n), \|\cdot\|_{2,P_{\theta_0,m_0}}) \left(1 + \frac{J_{[]}(D_{M_1} n^{-1/10}, \mathcal{D}_{M_1}(n), \|\cdot\|_{2,P_{\theta_0,m_0}})}{D_{M_1}^2 n^{-1/5}\sqrt{n}} 2M_1 V^*\right) \\
&\lesssim \delta^{-1} \left(\sqrt{D_{M_1}} n^{-1/20} + \frac{2M_1 V^* D_{M_1} n^{-1/10}}{D_{M_1}^2 n^{-1/5}\sqrt{n}}\right) \to 0, \qquad \text{as } n \to \infty.
\end{aligned}$$

4.13.4 Proof of Lemma 39

For every $(\theta, m) \in \mathcal{C}_{M_1}(n)$, note that

$$\big\|[m \circ \theta_0 - m \circ \theta] U_{\theta,m}\big\|_{2,P_{\theta_0,m_0}}^2 \le 4M_1^2 \big\|U_{\theta,m}\big\|_{2,P_{\theta_0,m_0}}^2 \le 4M_1^2 K_L n^{-1/5} = D_{M_1}^2 n^{-1/5}.$$

By Lemmas 50 and 51, we have

$$N_{[]}(\varepsilon, \{m \circ \theta_0 - m \circ \theta | (\theta, m) \in \mathcal{C}_{M_1}(n)\}, \|\cdot\|_\infty) \lesssim \exp(1/\sqrt{\varepsilon}).$$

By Lemma 37 and Lemma 9.25 of [Kosorok, 2008] (for bracketing entropy of product of uniformly bounded function classes), we have

$$N_{[]}(\varepsilon, \mathcal{A}_{M_1}(n), \|\cdot\|_{2,P_{\theta_0,m_0}}) \le N_{[]}(\varepsilon, \mathcal{A}_{M_1}^*, \|\cdot\|_{2,P_{\theta_0,m_0}}) \le c\varepsilon^{-2d} \exp\left(\frac{c}{\sqrt{\varepsilon}} + \frac{c}{\varepsilon}\right).$$

It follows that

$$J_{[]}(\gamma, \mathcal{A}_{M_1}(n), \|\cdot\|) \lesssim \gamma^{\frac{1}{2}}.$$

The rest of the proof is similar to the proof of Lemma 38.

4.13.5 Proof of Lemma 40

We first prove the first equation of (4.43). Note that, we have

$$\begin{aligned}
&\mathbb{P}(|\sqrt{n}\mathbb{P}_n \epsilon U_{\check{\theta},\check{m}}(X)| > \delta) \\
&\leq \mathbb{P}\Big(\sup_{(\theta,m)\in\mathcal{C}_{M_1}(n)} |\sqrt{n}\mathbb{P}_n \epsilon U_{\theta,m}(X)| > \delta \Big) + \mathbb{P}((\check{\theta},\check{m}) \notin \mathcal{C}_{M_1}(n)) \\
&\leq \mathbb{P}\Big(\sup_{f\in\mathcal{W}_{M_1}(n)} |\sqrt{n}\mathbb{P}_n \epsilon f| > \delta \Big) + \mathbb{P}((\check{\theta},\check{m}) \notin \mathcal{C}_{M_1}(n)) \\
&= \mathbb{P}\Big(\sup_{f\in\mathcal{W}_{M_1}(n)} |\mathbb{G}_n \epsilon f| > \delta \Big) + \mathbb{P}((\check{\theta},\check{m}) \notin \mathcal{C}_{M_1}(n)),
\end{aligned}$$

where the last equality is due to assumption **(A2)**. Now it is enough to show that for every fixed M_1 and L, we have

$$\mathbb{P}\Big(\sup_{f\in\mathcal{W}_{M_1}(n)} |\mathbb{G}_n \epsilon f| > \delta \Big) = o(1). \tag{4.135}$$

We will prove (4.135) by applying Lemma 52 with $\mathcal{F} = \mathcal{W}_{M_1}(n)$. Observe that by Lemma 37, we have

$$\log N_{[]}(\varepsilon, \mathcal{W}_{M_1}(n), \|\cdot\|_{2,P_{\theta_0,m_0}}) \lesssim \varepsilon^{-1}, \ \sup_{f\in\mathcal{W}^*_{M_1}} \|f\|_{2,\infty} \leq V^*, \text{ and } \sup_{f\in\mathcal{W}_{M_1}(n)} \|f\|^2_{2,P_{\theta_0,m_0}} \leq K_L^2 n^{-1/5}.$$

Now using arguments similar to (3.91), we can apply Lemma 52 with

$$\Phi = V^*, \ \kappa = K_L n^{-1/10}, \text{ and } \alpha = -1,$$

to obtain

$$\begin{aligned}
\mathbb{P}\Big(\sup_{f\in\mathcal{W}_{M_1}(n)} |\mathbb{G}_n \epsilon f| > \delta \Big) &\lesssim \delta^{-1} \mathbb{E}\Big(\sup_{f\in\mathcal{W}_{M_1}(n)} |\mathbb{G}_n \epsilon f| \Big) \\
&\lesssim \delta^{-1} \left(\sqrt{K_L} n^{-1/20} + \frac{1}{n^{1/10}\sqrt{n}} \right) \to 0, \qquad \text{as } n \to \infty.
\end{aligned}$$

We now verify the second and third equations in (4.43). The proof is similar to the proof of Lemma 36. Observe that by (4.105), (4.106), and (4.107) (with $(\check{m},\check{\theta})$ instead of $(\hat{m},\hat{\theta})$), we have

$$\begin{aligned}
\big| P_{\theta_0,m_0}[(m_0 - \check{m})(\theta_0{}^\top X) U_{\check{\theta},\check{m}}(X)] \big| &= O_p(n^{-2/5}) \Big[P_{\theta_0,m_0} \big| U_{\check{\theta},\check{m}}(X) \big|^2 \Big]^{1/2}, \\
\big| P_{\theta_0,m_0}[(\check{m}(\theta_0{}^\top X) - \check{m}(\check{\theta}^\top X)) U_{\check{\theta},\check{m}}(X)] \big| &= O_p(n^{-2/5}) \Big[P_{\theta_0,m_0} \big| U_{\check{\theta},\check{m}}(X) \big|^2 \Big]^{1/2}.
\end{aligned} \tag{4.136}$$

and

$$P_{\theta_0,m_0}|U_{\check{\theta},\check{m}}(X)|^2 \leq M^{*2}(d-1)P_{\theta_0,m_0}\left[(\check{m}'-m_0')(\check{\theta}^\top X)\right]^2 + M^{*2}(d-1)TP_{\theta_0,m_0}\left[\int_{D_{\check{\theta}}}[\check{m}'(u)-m_0'(u)]^2du\right]. \quad (4.137)$$

Finally by (4.9) of Theorem 28, we have that

$$\int_{D_{\check{\theta}}}\{\check{m}'(u)-m_0'(u)\}^2du \lesssim \|\check{m}'\circ\check{\theta}-m_0'\circ\check{\theta}\|^2 = O_p(n^{-4/15}).$$

The required result now follows by combining (4.136) and (4.137).

4.13.6 Proof of Theorem 37

Proof of this theorem follows along the lines of the proof of Theorem 32. By calculations similar to (4.108), (4.109), and (4.111) (with $(\hat{m},\hat{\theta})$ replaced by $(\check{m},\check{\theta})$), we have that

$$|P_{\check{\theta},m_0}\psi_{\check{\theta},\check{m}}| \lesssim \|m_0\circ\check{\theta}-\check{m}\circ\check{\theta}\|\,\|\check{m}'\circ\check{\theta}-m_0'\circ\check{\theta}\| + \|m_0'\|_\infty\|m_0\circ\check{\theta}-\check{m}\circ\check{\theta}\|\,\|h_{\check{\theta}}\circ\check{\theta}-h_{\theta_0}\circ\check{\theta}\|_{2,P_{\theta_0,m_0}}. \quad (4.138)$$

By Theorem 28, we have $\|\check{m}'\circ\check{\theta}-m_0'\circ\check{\theta}\| = O_p(n^{-2/15})$. Furthermore, by Theorems 25 and 27 and assumption **(B2)**, we have

$$\begin{aligned}\|m_0\circ\check{\theta}-\check{m}\circ\check{\theta}\| &\leq \|\check{m}\circ\check{\theta}-m_0\circ\theta_0\| + \|m_0\circ\theta_0-m_0\circ\check{\theta}\|\\ &\leq \|\check{m}\circ\check{\theta}-m_0\circ\theta_0\| + L_0T^2|\theta_0-\check{\theta}|\\ &= O_p(n^{-2/5})\end{aligned}$$

and $\|h_{\check{\theta}}\circ\check{\theta}-h_{\theta_0}\circ\check{\theta}\|_{2,P_{\theta_0,m_0}} \leq \bar{M}|\check{\theta}-\theta_0|$. Thus the first term on the right hand side of (4.138) is $O_p(n^{-8/15})$ and the second term on the right hand side of (4.138) is $O_p(n^{-4/5})$. Thus $|P_{\check{\theta},m_0}\psi_{\check{\theta},\check{m}}| = o_p(n^{-1/2})$.

4.13.7 Consistency of $\psi_{\check{\theta},\check{m}}$

Lemma 61. *If the conditions in Theorem 34 hold, then*

$$P_{\theta_0,m_0}|\psi_{\check{\theta},\check{m}}-\psi_{\theta_0,m_0}|^2 = o_p(1), \quad (4.139)$$

$$P_{\check{\theta},m_0}|\psi_{\check{\theta},\check{m}}|^2 = O_p(1). \quad (4.140)$$

Proof. Observe that the proof of (4.140) is identical to the proof of (4.113) (with $(\hat{\theta}, \hat{m})$ replaced by $(\check{\theta}, \check{m})$); see (4.115).

We now prove (4.139). By assumption **(B2)** and calculations similar to (4.114), we have

$$P_{\theta_0,m_0}|\psi_{\check{\theta},\check{m}} - \psi_{\theta_0,m_0}|^2 \le \mathbf{I} + 2\|\sigma^2(\cdot)\|_\infty \ \mathbf{II} + 2\|\sigma^2(\cdot)\|_\infty \ \mathbf{III} + 4M_1^2T^2\|\sigma^2(\cdot)\|_\infty|\check{\theta} - \theta_0|^2,$$

where

$$\begin{aligned}
\mathbf{I} &= P_{\theta_0,m_0}\Big|\big[m_0(\theta_0^\top X) - \check{m}(\check{\theta}^\top X)\big]\big[\check{m}'(\check{\theta}^\top X)X - (m_0' \ h_{\theta_0})(\check{\theta}^\top X)\big]\Big|^2 \\
\mathbf{II} &= P_{\theta_0,m_0}\Big|\check{m}'(\check{\theta}^\top X)X - m_0'(\theta_0^\top X)X\Big|^2 \\
\mathbf{III} &= P_{\theta_0,m_0}\Big|(m_0' \ h_{\theta_0})(\theta_0^\top X) - (m_0' \ h_{\theta_0})(\check{\theta}^\top X)\Big|^2
\end{aligned}$$

It is enough to show that $\mathbf{I}$, $\mathbf{II}$, and $\mathbf{III}$ are $o_p(1)$. By Theorems 27 and 28, we have

$$\begin{aligned}
\mathbf{II} &\le T^2P_{\theta_0,m_0}\Big|\check{m}'(\check{\theta}^\top X) - m_0'(\theta_0^\top X)\Big|^2 \\
&\le T^2P_{\theta_0,m_0}\big|\check{m}'(\check{\theta}^\top X) - m_0'(\check{\theta}^\top X)\big|^2 + T^2P_{\theta_0,m_0}\big|m_0'(\check{\theta}^\top X) - m_0'(\theta_0^\top X)\big|^2 = o_p(1).
\end{aligned}$$

For $\mathbf{I}$, observe that

$$|\check{m}'(\check{\theta}^\top x)x - m_0'(\check{\theta}^\top x)h_{\theta_0}(\check{\theta}^\top x)| \le |\check{m}'(\check{\theta}^\top x)x| + |m_0'(\check{\theta}^\top x)h_{\theta_0}(\check{\theta}^\top x)| \le LT + L\|h_{\theta_0}\|_{2,\infty}.$$

Moreover, by Theorem 25, we have $\|\check{m} \circ \check{\theta} - m_0 \circ \theta_0\| \xrightarrow{P} 0$. Thus,

$$\begin{aligned}
\mathbf{I} &= P_{\theta_0,m_0}\big|(m_0(\theta_0^\top X) - \check{m}(\check{\theta}^\top X))(\check{m}'(\check{\theta}^\top X)X - (m_0' \ h_{\theta_0})(\check{\theta}^\top X))\big|^2 \\
&\le (LT + L\|h_{\theta_0}\|_{2,\infty})\|m_0 \circ \theta_0 - \check{m} \circ \check{\theta}\|^2 = o_p(1).
\end{aligned}$$

Finally, we have

$$\begin{aligned}
\mathbf{III} &= P_{\theta_0,m_0}\Big|(m_0' \ h_{\theta_0})(\theta_0^\top X) - (m_0' \ h_{\theta_0})(\check{\theta}^\top X)\Big|^2 \\
&\le P_{\theta_0,m_0}\Big[\|m_0'' \ h_{\theta_0} + m_0' \ h_{\theta_0}'\|_{2,\infty}|(\theta_0 - \check{\theta})^\top X|\Big]^2 \\
&\le \|m_0'' \ h_{\theta_0} + m_0' \ h_{\theta_0}'\|_{2,\infty}^2T^2|\theta_0 - \check{\theta}|^2 = o_p(1).
\end{aligned}$$

□

4.13.8 Proof of Theorem 38

Observe that (4.116) and definition (4.117) imply that

$$\psi_{\check{\theta},\check{m}} - \psi_{\theta_0,m_0} = \epsilon\tau_{\check{\theta},\check{m}} + \upsilon_{\check{\theta},\check{m}}.$$

Thus, for every fixed M_1, we have

$$\begin{aligned}
&\mathbb{P}(|\mathbb{G}_n(\psi_{\check{\theta},\check{m}} - \psi_{\theta_0,m_0})| > \delta) \\
&\leq \mathbb{P}(|\mathbb{G}_n(\epsilon\tau_{\check{\theta},\check{m}} + \upsilon_{\check{\theta},\check{m}})| > \delta, (\check{\theta},\check{m}) \in \mathcal{C}_{M_1}(n)) + \mathbb{P}((\check{\theta},\check{m}) \notin \mathcal{C}_{M_1}(n)) \\
&\leq \mathbb{P}\left(|\mathbb{G}_n(\epsilon\tau_{\check{\theta},\check{m}})| > \frac{\delta}{2}, (\check{\theta},\check{m}) \in \mathcal{C}_{M_1}(n)\right) \\
&\quad + \mathbb{P}\left(|\mathbb{G}_n\upsilon_{\check{\theta},\check{m}}| > \frac{\delta}{2}, (\check{\theta},\check{m}) \in \mathcal{C}_{M_1}(n)\right) + \mathbb{P}\left((\check{\theta},\check{m}) \notin \mathcal{C}_{M_1}(n)\right) \\
&\leq \mathbb{P}\Big(\sup_{(\theta,m)\in\mathcal{C}_{M_1}(n)} |\mathbb{G}_n\epsilon\tau_{\theta,m}| > \frac{\delta}{2}\Big) + \mathbb{P}\Big(\sup_{(\theta,m)\in\mathcal{C}_{M_1}(n)} |\mathbb{G}_n\upsilon_{\theta,m}| > \frac{\delta}{2}\Big) \\
&\quad + \mathbb{P}\big((\check{\theta},\check{m}) \notin \mathcal{C}_{M_1}(n)\big). \qquad (4.141)
\end{aligned}$$

Recall that by Theorems 25–28, we have $\mathbb{P}\big((\check{\theta},\check{m}) \notin \mathcal{C}_{M_1}(n)\big) = o(1)$. Thus the proof of Theorem 38 will be complete if we show that the first two terms in (4.141) are $o(1)$. Lemmas 62 and 63 do this.

Lemma 62. *Fix M_1 and $\delta > 0$. For $n \in \mathbb{N}$, as $n \to \infty$, we have*

$$\mathbb{P}\Big(\sup_{(\theta,m)\in\mathcal{C}_{M_1}(n)} |\mathbb{G}_n\epsilon\tau_{\theta,m}| > \frac{\delta}{2}\Big) \to 0.$$

Proof. Recall that

$$\begin{aligned}
\tau_{\theta,m}(x) := H_\theta^\top\big\{[m'(\theta^\top x) - m_0'({\theta_0}^\top x)]x + [(m_0' \, h_{\theta_0})({\theta_0}^\top x) - (m_0' \, h_{\theta_0})(\theta^\top x)]\big\} \\
+ (H_\theta^\top - H_{\theta_0}^\top)\big[m_0'(\theta_0^\top x)x - (m_0' h_{\theta_0})(\theta_0^\top x)\big],
\end{aligned}$$

Let us define,

$$\Xi_{M_1}(n) := \big\{\tau_{\theta,m}\big|(\theta,m) \in \mathcal{C}_{M_1}(n)\big\} \quad \text{and} \quad \Xi^*_{M_1} := \big\{\tau_{\theta,m}\big|(\theta,m) \in \mathcal{C}^*_{M_1}\big\}.$$

We will prove that

$$N(\varepsilon, \Xi^*_{M_1}, \|\cdot\|_\infty) \leq c\exp(c/\varepsilon)\varepsilon^{-10d}, \qquad (4.142)$$

where c depends only on M_1 and d. We will now try to construct a bracket for $\Xi^*_{M_1}$. Recall that by Lemma 60, we have

$$N_{[]}(\varepsilon, \{m'(\theta^\top \cdot)|(\theta, m) \in \mathcal{C}^*_{M_1}\}, \|\cdot\|) \lesssim \exp(1/\varepsilon). \tag{4.143}$$

Moreover, by Lemma 28, we can find a $\theta_1, \theta_2, \ldots, \theta_{N_\varepsilon}$ with $N_\varepsilon \lesssim \varepsilon^{-2d}$ such that for every $\theta \in \Theta \cap B_{\theta_0}(1/2)$, there exists a θ_j such that

$$|\theta - \theta_j| \le \varepsilon/T, \ \|H_\theta - H_{\theta_j}\|_2 \le \varepsilon/T, \text{ and } |\theta^\top x - \theta_j^\top x| \le \varepsilon, \ \forall x \in \chi.$$

Observe that for all $x \in \chi$, we have $H_{\theta_j}^\top x - \varepsilon \preceq H_\theta^\top x \preceq H_{\theta_j}^\top x + \varepsilon$. Thus

$$N_{[]}(\varepsilon, \{f : \chi \to \mathbb{R}^d | f(x) = H_\theta^\top x, \forall x \in \chi, \theta \in \Theta \cap B_{\theta_0}(1/2)\}, \|\cdot\|_{2,\infty}) \lesssim \varepsilon^{-2d} \tag{4.144}$$

Similarly as $|m_0'(\theta^\top x) - m_0'(\theta_j^\top x)| \le L_0\varepsilon$, we have

$$N_{[]}(\varepsilon, \{m_0' \circ \theta : \theta \in \Theta \cap B_{\theta_0}(1/2)\}, \|\cdot\|) \lesssim \varepsilon^{-2d} \tag{4.145}$$

Finally observe that

$$\begin{aligned}
&|H_\theta^\top h_{\theta_0}(\theta^\top x) - H_{\theta_j}^\top h_{\theta_0}(\theta_j^\top x)| \\
&\le |H_\theta^\top h_{\theta_0}(\theta^\top x) - H_\theta^\top h_{\theta_0}(\theta_j^\top x)| + |H_\theta^\top h_{\theta_0}(\theta_j^\top x) - H_{\theta_j}^\top h_{\theta_0}(\theta_j^\top x)| \\
&\le |h_{\theta_0}(\theta^\top x) - h_{\theta_0}(\theta_j^\top x)| + \|H_\theta^\top - H_{\theta_j}^\top\|_2 \|h_{\theta_0}\|_{2,\infty} \\
&\le \|h_{\theta_0}'\|_{2,\infty}|\theta - \theta_j|T + \|H_\theta^\top - H_{\theta_j}^\top\|_2 \|h_{\theta_0}\|_{2,\infty} \le \varepsilon(\|h_{\theta_0}'\|_{2,\infty}| + \|h_{\theta_0}\|_{2,\infty}/T) \lesssim \varepsilon
\end{aligned}$$

and

$$|H_\theta^\top h_{\theta_0}(\theta_0^\top x) - H_{\theta_j}^\top h_{\theta_0}(\theta_0^\top x)| \le \|h_{\theta_0}(\theta_0^\top x)\|_{2,\infty}\varepsilon/T.$$

Thus we have

$$N_{[]}(\varepsilon, \{f : \chi \to \mathbb{R}^d | f(x) = H_\theta^\top h_{\theta_0}(\theta^\top x), \theta \in \Theta \cap B_{\theta_0}(1/2)\}, \|\cdot\|_{2,\infty}) \lesssim \varepsilon^{-2d} \tag{4.146}$$

$$N_{[]}(\varepsilon, \{f : \chi \to \mathbb{R}^d | f(x) = H_\theta^\top h_{\theta_0}(\theta_0^\top x), \theta \in \Theta \cap B_{\theta_0}(1/2)\}, \|\cdot\|_{2,\infty}) \lesssim \varepsilon^{-2d} \tag{4.147}$$

Thus by applying Lemma 9.25 of [Kosorok, 2008] to sums and product of classes of functions in (4.143),(4.144), (4.145), (4.146), and (4.147) we have (4.142).

Now, we find an upper bound for $\sup_{f\in\Xi_{M_1}(n)}\|f\|_{2,\infty}$. For every $(\theta,m)\in\mathcal{C}_{M_1}(n)$ and $x\in\mathcal{X}$ note that,

$$\begin{aligned}
|\tau_{\theta,m}(x)| &\le \big[|m'(\theta^\top x)-m'({\theta_0}^\top x)|+|m'({\theta_0}^\top x)-m_0'({\theta_0}^\top x)|\big]|x| \\
&\quad + |m_0'({\theta_0}^\top x)h_{\theta_0}({\theta_0}^\top x)-m_0'(\theta^\top x)h_{\theta_0}({\theta_0}^\top x)| \\
&\quad + |m_0'(\theta^\top x)h_{\theta_0}({\theta_0}^\top x)-m_0'(\theta^\top x)h_{\theta_0}(\theta^\top x)| \\
&\quad + |\theta-\theta_0||m_0'(\theta^\top x)x-(m_0'h_{\theta_0})(\theta^\top x)| \\
&\le L|\theta^\top x-{\theta_0}^\top x||x|+\|m'-m_0'\|_{D_{\theta_0}}|x| \\
&\quad + |h_{\theta_0}({\theta_0}^\top x)|\|m_0''\|_\infty|{\theta_0}^\top x-\theta^\top x| \\
&\quad + |m_0'(\theta^\top x)|\,\|h_{\theta_0}'\|_{2,\infty}|{\theta_0}^\top x-\theta^\top x| \\
&\quad + |\theta-\theta_0||m_0'(\theta^\top x)x-(m_0'h_{\theta_0})(\theta^\top x)| \\
&\le |\theta-\theta_0|LT^2+n^{-1/5}T+T^2\|m_0''\|_\infty T|\theta-\theta_0|+L_0\|h_{\theta_0}'\|_{2,\infty}T|\theta-\theta_0|+|\theta-\theta_0|L_0T. \\
&\le C_{11}n^{-1/10},
\end{aligned}$$

where C_{11} is constant depending only on L, L_0, T, m_0, and h_{θ_0}. Using arguments similar to (3.91) and Lemma 52 with $\Phi=\kappa=C_{11}n^{-1/10}$, and $\alpha=-1$, we have

$$\mathbb{P}\Big(\sup_{f\in\Xi_{M_1}(n)}|\mathbb{G}_n\epsilon f|>\frac{\delta}{2}\Big)\le 2\delta^{-1}\mathbb{E}\Big(\sup_{f\in\Xi_{M_1}(n)}|\mathbb{G}_n\epsilon f|\Big)\lesssim n^{-1/20}+n^{-4/10}=o(1). \qquad \square$$

Lemma 63. *Fix M_1 and $\delta>0$. For $n\in\mathbb{N}$, we have*

$$\mathbb{P}\left(\sup_{(\theta,m)\in\mathcal{C}_{M_1}(n)}|\mathbb{G}_n\upsilon_{\theta,m}|>\frac{\delta}{2}\right)=o_p(1).$$

Proof. Recall that

$$\upsilon_{\theta,m}(x):=[m_0(\theta_0^\top x)-m(\theta^\top x)][m'(\theta^\top x)H_\theta^\top x-m_0'(\theta^\top x)\,H_\theta^\top h_{\theta_0}(\theta^\top x)].$$

We will first show that

$$J_{[]}(\nu,\{\upsilon_{\theta,m}:(\theta,m)\in\mathcal{C}_{M_1}(n)\},\|\cdot\|_{2,P_{\theta_0,m_0}})\lesssim\nu^{1/2} \tag{4.148}$$

By Lemmas 50 and 60 and (4.144), (4.145), and (4.146), we have

$$\begin{aligned} N_{[]}(\varepsilon, \{m_0(\theta_0^\top \cdot) - m(\theta^\top \cdot) | (\theta, m) \in \mathcal{C}_{M_1}^*\}, \|\cdot\|_\infty) &\lesssim \exp(1/\sqrt{\varepsilon}), \\ N_{[]}(\varepsilon, \{m'(\theta^\top \cdot) | (\theta, m) \in \mathcal{C}_{M_1}^*\}, \|\cdot\|) &\lesssim \exp(1/\varepsilon), \\ N_{[]}(\varepsilon, \{f : \chi \to \mathbb{R}^d | f(x) = H_\theta^\top x, \forall x \in \chi, \theta \in \Theta \cap B_{\theta_0}(1/2)\}, \|\cdot\|_{2,\infty}) &\lesssim \varepsilon^{-2d} \\ N_{[]}(\varepsilon, \{m_0' \circ \theta : \theta \in \Theta \cap B_{\theta_0}(1/2)\}, \|\cdot\|) &\lesssim \varepsilon^{-2d} \\ N_{[]}(\varepsilon, \{f : \chi \to \mathbb{R}^d | f(x) = H_\theta^\top h_{\theta_0}(\theta^\top x), \theta \in \Theta \cap B_{\theta_0}(1/2)\}, \|\cdot\|_{2,\infty}) &\lesssim \varepsilon^{-2d}. \end{aligned} \tag{4.149}$$

Thus by applying Lemma 9.25 of [Kosorok, 2008] to sums and product of classes of functions in (4.149), we have

$$N_{[]}(\varepsilon, \{\upsilon_{\theta,m} : (\theta, m) \in \mathcal{C}_{M_1}^*\}, \|\cdot\|_{2,P_{\theta_0,m_0}}) \lesssim \exp\left(\frac{1}{\varepsilon} + \frac{1}{\sqrt{\varepsilon}}\right) \varepsilon^{-6d}.$$

Now (4.148) follows from the definition of $J_{[]}$ by observing that

$$J_{[]}(\nu, \{\upsilon_{\theta,m} : (\theta, m) \in \mathcal{C}_{M_1}(n)\}, \|\cdot\|_{2,P_{\theta_0,m_0}}) \leq J_{[]}(\nu, \{\upsilon_{\theta,m} : (\theta, m) \in \mathcal{C}_{M_1}^*\}, \|\cdot\|_{2,P_{\theta_0,m_0}}).$$

Now we will find $\sup_{(\theta,m)\in\mathcal{C}_{M_1}(n)} \|\upsilon_{\theta,m}\|_{2,\infty}$. For every $x \in \chi$ observe that,

$$\begin{aligned} |\upsilon_{\theta,m}(x)| &\leq |m_0(\theta_0^\top x) - m(\theta_0{}^\top x)| \cdot |m'(\theta^\top x)x - m_0'(\theta^\top x)h_{\theta_0}(\theta^\top x)| \\ &\quad + |m(\theta_0^\top x) - m(\theta^\top x)| \cdot |m'(\theta^\top x)x - m_0'(\theta^\top x)h_{\theta_0}(\theta^\top x)| \\ &\leq \|m_0 - m\|_{D_{\theta_0}} |m'(\theta^\top x)x - m_0'(\theta^\top x)h_{\theta_0}(\theta^\top x)| \\ &\quad + L|\theta_0^\top x - \theta^\top x||m'(\theta^\top x)x - m_0'(\theta^\top x)h_{\theta_0}(\theta^\top x)| \\ &\leq b_n^{-1} 2LT + 2T^2L^2M_2|\theta - \theta_0 \\ &\leq C[b_n^{-1} + n^{-1/10}], \end{aligned}$$

where C is a constant depending only on T, L, and M_1. Thus

$$\sup_{(\theta,m)\in\mathcal{C}_{M_1}(n)} \|\upsilon_{\theta,m}\|_{2,P_{\theta_0,m_0}} \leq \sup_{(\theta,m)\in\mathcal{C}_{M_1}(n)} \|\upsilon_{\theta,m}\|_{2,\infty} \leq C^2[b_n^{-1} + n^{-1/10}].$$

Using arguments similar to (3.91) and Lemma 3.4.2 of [van der Vaart and Wellner,

1996] (for class of uniformly bounded function, see Lemma 54) we have

$$
\begin{aligned}
&\mathbb{P}\left(\sup_{(\theta,m)\in\mathcal{C}_{M_1}(n)} |\mathbb{G}_n \upsilon_{\theta,m}| > \frac{\delta}{2}\right) \\
&\lesssim 2\delta^{-1}\mathbb{E}\Big(\sup_{(\theta,m)\in\mathcal{C}_{M_1}(n)} |\mathbb{G}_n \upsilon_{\theta,m}|\Big) \\
&\lesssim J_{[]}([b_n^{-1}+n^{-1/10}], \mathcal{W}_{M_1}(n), \|\cdot\|_{2,P_{\theta_0,m_0}}) + \frac{J_{[]}^2([b_n^{-1}+n^{-1/10}], \mathcal{W}_{M_1}(n), \|\cdot\|_{2,P_{\theta_0,m_0}})}{[b_n^{-1}+n^{-1/10}]^2\sqrt{n}} \\
&\lesssim [b_n^{-1}+n^{-1/10}]^{1/2} + \frac{[b_n^{-1}+n^{-1/10}]}{[b_n^{-1}+n^{-1/10}]^2\sqrt{n}} \\
&\lesssim [b_n^{-1}+n^{-1/10}]^{1/2} + \frac{1}{b_n^{-1}\sqrt{n}+n^{4/10}} = o(1) \quad \text{as} \quad b_n = o(n^{1/2}).
\end{aligned}
$$

□

Part III

Bibliography

Bibliography

[Agmon, 2010] Shmuel Agmon. *Lectures on elliptic boundary value problems*. AMS Chelsea Publishing, Providence, RI, 2010. Prepared for publication by B. Frank Jones, Jr. with the assistance of George W. Batten, Jr., Revised edition of the 1965 original.

[Anderson and Darling, 1952] T. W. Anderson and D. A. Darling. Asymptotic theory of certain "goodness of fit" criteria based on stochastic processes. *Ann. Math. Statistics*, 23:193–212, 1952.

[Antoniadis *et al.*, 2004] Anestis Antoniadis, Gérard Grégoire, and Ian W. McKeague. Bayesian estimation in single-index models. *Statist. Sinica*, 14(4):1147–1164, 2004.

[Barlow *et al.*, 1972] R. E. Barlow, D. J. Bartholomew, J. M. Bremner, and H. D. Brunk. *Statistical inference under order restrictions. The theory and application of isotonic regression*. John Wiley & Sons, London-New York-Sydney, 1972. Wiley Series in Probability and Mathematical Statistics.

[Benjamini and Hochberg, 1995] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B*, 57(1):289–300, 1995.

[Benjamini and Hochberg, 2000] Y. Benjamini and Y. Hochberg. On the adaptive control of the false discovery rate in multiple testing with independent statistics. *J. Educational and Behavioral Statistics*, 25:60–83, 2000.

[Benjamini *et al.*, 2006] Y. Benjamini, A. Krieger, and D. Yekutieli. Adaptive linear step-up procedures that control the false discovery rate. *Biometrika*, 93(3):491–507, 2006.

[Bertsekas, 2003] D. P. Bertsekas. *Convex analysis and optimization*. Athena Scientific, Belmont, MA, 2003. With Angelia Nedić and Asuman E. Ozdaglar.

[Bickel *et al.*, 1993] Peter J. Bickel, Chris A. J. Klaassen, Ya'acov Ritov, and John A. Wellner. *Efficient and adaptive estimation for semiparametric models*. Springer-Verlag, New York, 1993.

[Black, 2004] M. A. Black. A note on the adaptive control of false discovery rates. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 66(2):297–304, 2004.

[Bogachev, 2007] V. I. Bogachev. *Measure theory. Vol. I, II*. Springer-Verlag, Berlin, 2007.

[Bordes *et al.*, 2006] L. Bordes, S. Mottelet, and P. Vandekerkhove. Semiparametric estimation of a two-component mixture model. *Ann. Statist.*, 34(3):1204–1232, 2006.

[Brunk, 1955] H. D. Brunk. Maximum likelihood estimates of monotone parameters. *Ann. Math. Statist.*, 26:607–616, 1955.

[Cai and Jin, 2010] T. Tony Cai and J. Jin. Optimal rates of convergence for estimating the null density and proportion of nonnull effects in large-scale multiple testing. *Ann. Statist.*, 38(1):100–145, 2010.

[Cai *et al.*, 2007] T. Cai, J. Jin, and M. G. Low. Estimation and confidence sets for sparse normal mixtures. *Ann. Statist.*, 35(6):2421–2449, 2007.

[Carnicer and Dahmen, 1994] J. M. Carnicer and W. Dahmen. Characterization of local strict convexity preserving interpolation methods by C^1 functions. *J. Approx. Theory*, 77(1):2–30, 1994.

[Carroll *et al.*, 1997] Raymond J Carroll, Jianqing Fan, Irene Gijbels, and Matt P Wand. Generalized partially linear single-index models. *Journal of the American Statistical Association*, 92(438):477–489, 1997.

[Celisse and Robin, 2010] A. Celisse and S. Robin. A cross-validation based estimation of the proportion of true null hypotheses. *J. Statist. Planng. Inf.*, 140(11):3132–3147, 2010.

[Chen and Plemmons, 2010] Donghui Chen and Robert J. Plemmons. Nonnegativity constraints in numerical analysis. In *The birth of numerical analysis*, pages 109–139. World Sci. Publ., Hackensack, NJ, 2010.

[Chen and Samworth, 2014] Y. Chen and R. J. Samworth. Generalised additive and index models with shape constraints. *ArXiv e-prints*, April 2014.

[Cohen, 1967] A. C. Cohen. Estimation in mixtures of two normal distributions. *Technometrics*, 9:15–28, 1967.

[Cui *et al.*, 2011] Xia Cui, Wolfgang Karl Härdle, and Lixing Zhu. The EFM approach for single-index models. *Ann. Statist.*, 39(3):1658–1688, 2011.

[Davis, 1963] Philip J. Davis. *Interpolation and approximation.* Blaisdell Publishing Co. Ginn and Co. New York-Toronto-London, 1963.

[Day, 1969] N. E. Day. Estimating the components of a mixture of normal distributions. *Biometrika*, 56:463–474, 1969.

[Delecroix *et al.*, 2006] Michel Delecroix, Marian Hristache, and Valentin Patilea. On semiparametric m-estimation in single-index regression. *Journal of Statistical Planning and Inference*, 136(3):730–769, 2006.

[Donoho and Jin, 2004] D. Donoho and J. Jin. Higher criticism for detecting sparse heterogeneous mixtures. *Ann. Statist.*, 32(3):962–994, 2004.

[Dontchev *et al.*, 2003] Asen L. Dontchev, Houduo Qi, and Liqun Qi. Quadratic convergence of Newton's method for convex interpolation and smoothing. *Constr. Approx.*, 19(1):123–143, 2003.

[Dümbgen *et al.*, 2004] L. Dümbgen, S. Freitag, and G. Jongbloed. Consistency of concave regression with an application to current-status data. *Math. Methods Statist.*, 13(1):69–81, 2004.

[Durrett, 2010] Rick Durrett. *Probability: theory and examples.* Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge, fourth edition, 2010.

[Efron, 2007] Bradley Efron. Size, power and false discovery rates. *Ann. Statist.*, 35(4):1351–1377, 2007.

[Efron, 2010] B. Efron. *Large-scale inference*, volume 1 of *Institute of Mathematical Statistics Monographs.* Cambridge University Press, Cambridge, 2010. Empirical Bayes methods for estimation, testing, and prediction.

[Elfving and Andersson, 1988] Tommy Elfving and Lars-Erik Andersson. An algorithm for computing constrained smoothing spline functions. *Numer. Math.*, 52(5):583–595, 1988.

[Feller, 1971] William Feller. *An introduction to probability theory and its applications. Vol. II.* Second edition. John Wiley & Sons, Inc., New York-London-Sydney, 1971.

[Fils-Villetard *et al.*, 2008] A. Fils-Villetard, A. Guillou, and J. Segers. Projection estimators of Pickands dependence functions. *Canad. J. Statist.*, 36(3):369–382, 2008.

[Genovese and Wasserman, 2004] C. Genovese and L. Wasserman. A stochastic process approach to false discovery control. *Ann. Statist.*, 32(3):1035–1061, 2004.

[Green and Silverman, 1994] P. J. Green and B. W. Silverman. *Nonparametric regression and generalized linear models*, volume 58 of *Monographs on Statistics and Applied Probability.* Chapman & Hall, London, 1994. A roughness penalty approach.

[Grenander, 1956] U. Grenander. On the theory of mortality measurement. I. *Skand. Aktuarietidskr.*, 39:70–96, 1956.

[Groeneboom and Hendrickx, 2016] Piet Groeneboom and Kim Hendrickx. Current status linear regression. *arXiv preprint arXiv:1601.00202*, 2016.

[Groeneboom *et al.*, 2001] Piet Groeneboom, Geurt Jongbloed, and Jon A. Wellner. Estimation of a convex function: characterizations and asymptotic theory. *Ann. Statist.*, 29(6):1653–1698, 2001.

[Grotzinger and Witzgall, 1984] S. J. Grotzinger and C. Witzgall. Projections onto order simplexes. *Appl. Math. Optim.*, 12(3):247–270, 1984.

[Guntuboyina and Sen, 2013] A. Guntuboyina and B. Sen. Covering numbers for convex functions. *Information Theory, IEEE Transactions on*, 59(4):1957–1965, April 2013.

[Györfi *et al.*, 2002] László Györfi, Michael Kohler, Adam Krzyżak, and Harro Walk. *A distribution-free theory of nonparametric regression*. Springer Series in Statistics. Springer-Verlag, New York, 2002.

[Hájek, 1972] Jaroslav Hájek. Local asymptotic minimax and admissibility in estimation. In *Proceedings of the sixth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 175–194, 1972.

[Hanson and Pledger, 1976] D. L. Hanson and Gordon Pledger. Consistency in concave regression. *Ann. Statist.*, 4(6):1038–1050, 1976.

[Härdle and Liang, 2007] Wolfgang Härdle and Hua Liang. Partially linear models. In *Statistical methods for biostatistics and related fields*, pages 87–103. Springer, Berlin, 2007.

[Härdle *et al.*, 1993] Wolfgang Härdle, Peter Hall, and Hidehiko Ichimura. Optimal smoothing in single-index models. *Ann. Statist.*, 21(1):157–178, 1993.

[Hastie *et al.*, 2009] Trevor Hastie, Robert Tibshirani, Jerome Friedman, T Hastie, J Friedman, and R Tibshirani. *The elements of statistical learning*, volume 2. Springer, 2009.

[Hengartner and Stark, 1995] N. W. Hengartner and P. B. Stark. Finite-sample confidence envelopes for shape-restricted densities. *Ann. Statist.*, 23(2):525–550, 1995.

[Hildreth, 1954] Clifford Hildreth. Point estimates of ordinates of concave functions. *J. Amer. Statist. Assoc.*, 49:598–619, 1954.

[Horowitz, 1998] Joel L. Horowitz. *Semiparametric methods in econometrics*, volume 131 of *Lecture Notes in Statistics*. Springer-Verlag, New York, 1998.

[Horowitz, 2009] Joel L. Horowitz. *Semiparametric and nonparametric methods in econometrics*. Springer Series in Statistics. Springer, New York, 2009.

[Hristache *et al.*, 2001] Marian Hristache, Anatoli Juditsky, and Vladimir Spokoiny. Direct estimation of the index coefficient in a single-index model. *Ann. Statist.*, 29(3):595–623, 2001.

[Hunter *et al.*, 2007] D. R. Hunter, S. Wang, and T. P. Hettmansperger. Inference for mixtures of symmetric distributions. *Ann. Statist.*, 35(1):224–251, 2007.

[Ichimura, 1993] Hidehiko Ichimura. Semiparametric least squares (sls) and weighted (sls) estimation of single-index models. *J. Econometrics*, 58(1-2):71–120, 1993.

[Jennrich, 1969] Robert I. Jennrich. Asymptotic properties of non-linear least squares estimators. *Ann. Math. Statist.*, 40:633–643, 1969.

[Jin, 2008] J. Jin. Proportion of non-zero normal means: universal oracle equivalences and uniformly consistent estimators. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 70(3):461–493, 2008.

[Kalaj *et al.*, 2016] David Kalaj, Matti Vuorinen, and Gendi Wang. On quasi-inversions. *Monatsh. Math.*, 180(4):785–813, 2016.

[Kosorok, 2008] Michael R. Kosorok. *Introduction to empirical processes and semiparametric inference*. Springer Series in Statistics. Springer, New York, 2008.

[Kuchibhotla and Patra, 2016] Arun Kumar Kuchibhotla and Rohit Kumar Patra. *simest: Single Index Model Estimation with Constraints on Link Function*, 2016. R package version 0.6.

[Kulldorff *et al.*, 2005] M. Kulldorff, J. Heffernan, R. Hartman, R. Assuncao, and F. Mostashari. A space-time permutation scan statistic for disease outbreak detection. *PLoS Med.*, 2(3):e59, 2005.

[Langaas *et al.*, 2005] M. Langaas, B. H. Lindqvist, and E. Ferkingstad. Estimating the proportion of true null hypotheses, with application to DNA microarray data. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 67(4):555–572, 2005.

[Lawson and Hanson, 1974] Charles L. Lawson and Richard J. Hanson. *Solving least squares problems*. Prentice-Hall, Inc., Englewood Cliffs, N.J., 1974. Prentice-Hall Series in Automatic Computation.

[Li and Duan, 1989] Ker-Chau Li and Naihua Duan. Regression analysis under link violation. *Ann. Statist.*, 17(3):1009–1052, 1989.

[Li and Patilea, 2015] Weiyu Li and Valentin Patilea. A new inference approach for single-index models. 2015.

[Li and Racine, 2007] Qi Li and Jeffrey Scott Racine. *Nonparametric econometrics*. Princeton University Press, Princeton, NJ, 2007. Theory and practice.

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

[Lindsay and Basak, 1993] B. G. Lindsay and P. Basak. Multivariate normal mixtures: a fast consistent method of moments. *J. Amer. Statist. Assoc.*, 88(422):468–476, 1993.

[Lindsay, 1983] B. G. Lindsay. The geometry of mixture likelihoods: a general theory. *Ann. Statist.*, 11(1):86–94, 1983.

[Lindsay, 1995] B. G. Lindsay. Mixture models: Theory, geometry and applications. *NSF-CBMS Regional Conference Series in Probability and Statistics*, 5:1–163, 1995.

[Lyons, 2008] L. Lyons. Open statistical issues in particle physics. *Ann. Appl. Stat.*, 2(3):887–915, 2008.

[Ma and Zhu, 2013a] Yanyuan Ma and Liping Zhu. 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, 2013.

[Ma and Zhu, 2013b] Yanyuan Ma and Liping Zhu. 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, 2013.

[Mammen and Thomas-Agnan, 1999] E. Mammen and C. Thomas-Agnan. Smoothing splines and shape restrictions. *Scand. J. Statist.*, 26(2):239–252, 1999.

[Mammen and van de Geer, 1997] Enno Mammen and Sara van de Geer. Penalized quasi-likelihood estimation in partial linear models. *Ann. Statist.*, 25(3):1014–1035, 1997.

[McLachlan and Peel, 2000] G. McLachlan and D. Peel. *Finite mixture models*. Wiley Series in Probability and Statistics: Applied Probability and Statistics. Wiley-Interscience, New York, 2000.

[Meinshausen and Bühlmann, 2005] N. Meinshausen and P. Bühlmann. Lower bounds for the number of false null hypotheses for multiple testing of associations under general dependence structures. *Biometrika*, 92(4):893–907, 2005.

[Meinshausen and Rice, 2006] N. Meinshausen and J. Rice. Estimating the proportion of false null hypotheses among a large number of independently tested hypotheses. *Ann. Statist.*, 34(1):373–393, 2006.

[Meyer, 2001] C. D. Meyer. *Matrix Analysis and Applied Linear Algebra*. SIAM, Philadelphia, PA, USA, 2001.

[Miller *et al.*, 2001] C. J. Miller, C. Genovese, R. C. Nichol, L. Wasserman, A. Connolly, D. Reichart, A. Hopkins, and A. Schneider, J.and Moore. Controlling the false-discovery rate in astrophysical data analysis. *Astron. J.*, 122(6):3492–3505, 2001.

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

[Newey and Stoker, 1993] Whitney K. Newey and Thomas M. Stoker. Efficiency of weighted average derivative estimators and index models. *Econometrica*, 61(5):1199–1223, 1993.

[Newey, 1990] Whitney K Newey. Semiparametric efficiency bounds. *Journal of applied econometrics*, 5(2):99–135, 1990.

[Nguyen and Matias, 2013] V. H. Nguyen and C. Matias. On efficient estimators of the proportion of true null hypotheses in a multiple testing setup. arXiv:1205.4097, 2013.

[Oden and Reddy, 2012] John Tinsley Oden and Junuthula Narasimha Reddy. *An introduction to the mathematical theory of finite elements*. Courier Dover Publications, 2012.

[Parzen, 1960] Emanuel Parzen. *Modern probability theory and its applications*. John Wiley & Sons, Incorporated, 1960.

[Patra *et al.*, 2015] Rohit K. Patra, Emilio Seijo, and Bodhisattva Sen. A consistent bootstrap procedure for the maximum score estimator. *J. Econometrics (revision resubmitted)*, 2015.

[Pollard, 1989] David Pollard. Asymptotics via empirical processes. *Statist. Sci.*, 4(4):341–366, 1989. With comments and a rejoinder by the author.

[Pollard, 1990] David Pollard. *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, 1990.

[Powell *et al.*, 1989] James L. Powell, James H. Stock, and Thomas M. Stoker. Semiparametric estimation of index coefficients. *Econometrica*, 57(6):1403–1430, 1989.

[Quandt and Ramsey, 1978] R. E. Quandt and J. B. Ramsey. Estimating mixtures of normal distributions and switching regressions. *J. Amer. Statist. Assoc.*, 73(364):730–752, 1978. With comments and a rejoinder by the authors.

[R Development Core Team, 2008] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2008. ISBN 3-900051-07-0.

[Robertson *et al.*, 1988] T. Robertson, F. T. Wright, and R. L. Dykstra. *Order restricted statistical inference*. Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics. John Wiley & Sons Ltd., Chichester, 1988.

[Robin *et al.*, 2003] A. C. Robin, C. Reyl, S. Derrire, and S. Picaud. A synthetic view on structure and evolution of the milky way. *Astronomy and Astrophysics*, 409(1):523–540, 2003.

[Robin *et al.*, 2007] S. Robin, A. Bar-Hen, J.-J. Daudin, and L. Pierre. A semi-parametric approach for mixture models: application to local false discovery rate estimation. *Comput. Statist. Data Anal.*, 51(12):5483–5493, 2007.

[Salvador and Chan, 2004] S. Salvador and P. Chan. Determining the number of clusters/segments in hierarchical clustering/segmentation algorithms. *Proc. 16th IEEE Intl. Conf. on Tools with AI*, 25:576–584, 2004.

[Samworth and Yuan, 2012] Richard J. Samworth and Ming Yuan. Independent component analysis via nonparametric maximum likelihood estimation. *Ann. Statist.*, 40(6):2973–3002, 2012.

[Seijo and Sen, 2011] Emilio Seijo and Bodhisattva Sen. Nonparametric least squares estimation of a multivariate convex regression function. *The Annals of Statistics*, 39(3):1633–1657, 06 2011.

[Stein, 1956] Charles Stein. Efficient nonparametric testing and estimation. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, 1954–1955, vol. I*, pages 187–195. University of California Press, Berkeley and Los Angeles, 1956.

[Stoker, 1986] Thomas M. Stoker. Consistent estimation of scaled coefficients. *Econometrica*, 54(6):1461–1481, 1986.

[Stone, 1980] Charles J. Stone. Optimal rates of convergence for nonparametric estimators. *Ann. Statist.*, 8(6):1348–1360, 1980.

[Storey, 2002] J. D. Storey. A direct approach to false discovery rates. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 64(3):479–498, 2002.

[Swanepoel, 1999] J. W. H. Swanepoel. The limiting behavior of a modified maximal symmetric $2s$-spacing with applications. *Ann. Statist.*, 27(1):24–35, 1999.

[Tsiatis, 2006] Anastasios A. Tsiatis. *Semiparametric theory and missing data*. Springer Series in Statistics. Springer, New York, 2006.

[Turkheimer *et al.*, 2001] F. Turkheimer, C.B Smith, and K. Schmidt. Estimation of the number of "true null hypotheses in multivariate analysis of neuroimaging data. *NeuroImage*, 13(5):920–930, 2001.

[Utreras, 1985] Florencio I. Utreras. Smoothing noisy data under monotonicity constraint: existence, characterization and convergence rates. *Numer. Math.*, 47(4):611–625, 1985.

[van de Geer, 1990] Sara van de Geer. Estimating a regression function. *The Annals of Statistics*, pages 907–924, 1990.

[Van de Geer, 2000a] Sara A. Van de Geer. *Applications of empirical process theory*, volume 6 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, 2000.

[van de Geer, 2000b] Sara A. van de Geer. *Applications of empirical process theory*, volume 6 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, 2000.

[van der Vaart and Wellner, 1996] Aad W. van der Vaart and Jon A. Wellner. *Weak convergence and empirical processes*. Springer Series in Statistics. Springer-Verlag, New York, 1996. With applications to statistics.

[van der Vaart, 1996] Aad van der Vaart. Efficient maximum likelihood estimation in semiparametric mixture models. *The Annals of Statistics*, 24(2):862–878, 1996.

[Van der Vaart, 1998a] A. W. Van der Vaart. *Asymptotic statistics*, volume 3 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, 1998.

[van der Vaart, 1998b] A. W. van der Vaart. *Asymptotic statistics*, volume 3 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, 1998.

[van der Vaart, 2002] Aad van der Vaart. Semiparametric statistics. In *Lectures on probability theory and statistics (Saint-Flour, 1999)*, volume 1781 of *Lecture Notes in Math.*, pages 331–457. Springer, Berlin, 2002.

[Wahba, 1990] Grace Wahba. *Spline models for observational data*, volume 59 of *CBMS-NSF Regional Conference Series in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1990.

[Walker *et al.*, 2007] M. G. Walker, M. Mateo, E. W. Olszewski, O. Y. Gnedin, X. Wang, B. Sen, and M. Woodroofe. Velocity dispersion profiles of seven dwarf spheroidal galaxies. *Astrophysical J.*, 667(1):L53–L56, 2007.

[Walker *et al.*, 2009] M.G. Walker, M. Mateo, E.W. Olszewski, B. Sen, and M. Woodroofe. Clean kinematic samples in dwarf spheroidals: An algorithm for evaluating membership and estimating distribution parameters when contamination is present. *The Astronomical Journal*, 137:3109, 2009.

[Walther, 2001] G. Walther. Multiscale maximum likelihood analysis of a semiparametric model, with applications. *Ann. Statist.*, 29(5):1297–1319, 2001.

[Walther, 2002] G. Walther. Detecting the presence of mixing with multiscale maximum likelihood. *J. Amer. Statist. Assoc.*, 97(458):508–513, 2002.

[Wen and Yin, 2013] Zaiwen Wen and Wotao Yin. A feasible method for optimization with orthogonality constraints. *Math. Program.*, 142(1-2, Ser. A):397–434, 2013.

[Woodroofe and Sun, 1993] Michael Woodroofe and Jiayang Sun. A penalized maximum likelihood estimate of $f(0+)$ when f is nonincreasing. *Statist. Sinica*, 3(2):501–515, 1993.

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

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

[Yu and Ruppert, 2002] Yan Yu and David Ruppert. Penalized spline estimation for partially linear single-index models. *J. Amer. Statist. Assoc.*, 97(460):1042–1054, 2002.

[Zhou and He, 2008] Jianhui Zhou and Xuming He. Dimension reduction based on constrained canonical correlation and variable filtering. *Ann. Statist.*, 36(4):1649–1668, 2008.