

# Efficient Estimation in Convex Single Index Models

Arun K. Kuchibhotla, Rohit K. Patra, and Bodhisattva Sen

400 Jon M. Huntsman Hall  
3730 Walnut Street  
Philadelphia, PA-19104.  
e-mail: [arunku@upenn.edu](mailto:arunku@upenn.edu)

Department of Statistics  
University of Florida  
221 Griffin-Floyd Hall  
Gainesville, Florida 32611  
e-mail: [rohitpatra@ufl.edu](mailto:rohitpatra@ufl.edu)

Department of Statistics  
Columbia University  
1255 Amsterdam Avenue  
New York, New York 10027  
e-mail: [bodhi@stat.columbia.edu](mailto:bodhi@stat.columbia.edu)

**Abstract:** 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 Lipschitz constrained least squares estimator and (2) a shape-constrained smoothing spline estimator. Moreover, both of these procedures lead to estimators for the unknown finite dimensional parameter. We develop methods to compute both the Lipschitz constrained least squares estimator (LLSE) and the penalized least squares estimator (PLSE) of the parametric and the nonparametric components given independent and identically distributed (i.i.d.) data. We prove the consistency and find the rates of convergence for both the LLSE and the PLSE. For both the LLSE and the PLSE, we establish  $n^{-1/2}$ -rate of convergence and semiparametric efficiency of the parametric component under mild assumptions. Moreover, both the LLSE and the PLSE readily yield asymptotic confidence sets for the finite dimensional parameter. 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 and phrases:** Approximately least favorable sub-provided models, interpolation inequality, penalized least squares, shape restricted function estimation.

## 1. Introduction

We consider the following single index regression model:

$$Y = m_0(\theta_0^\top X) + \epsilon, \quad \mathbb{E}(\epsilon|X) = 0, \quad \text{almost every (a.e.) } X, \quad (1.1)$$

where  $X \in \mathbb{R}^d$  ( $d \geq 1$ ) is the predictor,  $Y \in \mathbb{R}$  is the response variable,  $m_0 : \mathbb{R} \rightarrow \mathbb{R}$  is the unknown link function,  $\theta_0 \in \mathbb{R}^d$  is the unknown index parameter, and  $\epsilon$  is the unobserved error. The above single index model, a popular choice in many application areas, circumvents the curse of dimensionality encountered in estimating the fully nonparametric regression function  $\mathbb{E}(Y|X = \cdot)$  by assuming that the link function depends on  $X$  only through a one dimensional projection, i.e.,  $\theta_0^\top X$ ; see [45]. Moreover, the coefficient vector  $\theta_0$  provides interpretability; see [36]. The one-dimensional unspecified link function  $m_0$  also offers some flexibility in modeling.

In this paper, we assume further that  $m_0$  is known to be *convex*. This assumption is motivated by the fact that in a wide range of applications in various fields the regression 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;

see e.g., [54]). Utility functions are often assumed to be concave (representing decreasing marginal utility; see e.g., [39, 36]). In finance, theory restricts call option prices to be convex and decreasing functions of the strike price (see e.g., [2]); in stochastic control, value functions are often assumed to be convex (see e.g., [29]).

Given i.i.d. observations  $\{(x_i, y_i) : i = 1, \dots, n\}$  from model (1.1), the goal is to estimate the unknown parameters of interest —  $m_0$  and  $\theta_0$ . In this paper we propose and study two estimation techniques for  $m_0$  and  $\theta_0$  in model (1.1). For both procedures, we conduct a systematic study of the characterization, computation, consistency, rates of convergence and the limiting distribution of the estimator of the finite-dimensional parameter  $\theta_0$ . Moreover, we show that under mild assumptions, the finite dimensional estimators are semiparametrically efficient. Indeed, our paper represents the first work on convexity constrained single index models (without any distributional assumptions on the error and/or design).

Our first estimator, which we call the Lipschitz constrained least squares estimator (LLSE), is defined as

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

where  $\mathcal{M}_L$  denotes the class of all  $L$ -Lipschitz convex functions and

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

As any convex function is Lipschitz in the interior of its domain,  $(\check{m}_n, \check{\theta}_n)$  defines a natural non-parametric least squares estimator (LSE) for model (1.1). Moreover, this leads to a convex piecewise affine estimator for the link function  $m_0$ .

Our second approach, which yields a smooth convex estimator of  $m_0$ , is obtained by penalizing the squared loss with a penalty on the roughness of the convex link:

$$(\hat{m}_n, \hat{\theta}_n) := \arg \min_{(m, \theta) \in \mathcal{R} \times \Theta} \frac{1}{n} \sum_{i=1}^n [y_i - m(\theta^\top x_i)]^2 + \lambda^2 \int [m''(t)]^2 dt,$$

where  $\mathcal{R}$  denotes the class of all convex functions that have absolutely continuous first derivatives. We call this estimator the penalized least squares estimator (PLSE).

Although single index models are well-studied in the statistical literature (e.g., see [45], [35], [28], [21], [25], [12], and [11] among others), estimation and inference in shape-restricted single index models are not very well-studied, despite its numerous applications. The earliest reference we could find was the work of Murphy et al. [41], where the authors considered a penalized likelihood approach in the current status regression model with a monotone link function. During the preparation of this paper we became aware of three relevant papers — [9], [17], and [3]. Chen and Samworth [9] consider maximum likelihood estimation in a generalized additive index model (slightly more general model than (1.1)) and prove consistency of the proposed estimators. However, rates of convergence or asymptotic distributions of the estimators are not studied. Groeneboom and Hendrickx [17] propose a  $\sqrt{n}$ -consistent and asymptotically normal but *inefficient* estimator of the index vector in the current status model based on the (non-smooth) maximum likelihood estimate (MLE) of the nonparametric component under just monotonicity constraint. They also propose two other estimators of the index vector based on kernel smoothed versions of the MLE for the nonparametric component. Although these estimators do not achieve the efficiency bound their asymptotic variances can be made arbitrarily close to the efficient variance. Balabdaoui et al. [3] study model (1.1) under monotonicity constraint but they only prove  $n^{1/3}$ -consistency of the LSE of  $\theta_0$ ; moreover they do not obtain the limiting distribution of the estimator of  $\theta_0$ .

<sup>1</sup>Here  $|\cdot|$  denotes the Euclidean norm, and  $S^{d-1}$  is the Euclidean unit sphere in  $\mathbb{R}^d$ . The norm-1 and the positivity constraints are necessary for identifiability of the model as  $m_0(\theta_0^\top x) \equiv m_1(\theta_1^\top x)$  where  $m_1(t) := m_0(-2t)$  and  $\theta_1 = -\theta_0/2$ ; see [6] and [11] for identifiability of the model (1.1).

In the following we briefly summarize our major contributions and highlight the main novelties.

- Both the proposed penalized and Lipschitz constrained estimators are optimal — the function estimates are minimax rate optimal and the estimates of the index parameter are semiparametrically efficient; see [42] for a brief overview of the notion of semiparametric efficiency. Moreover, our asymptotic results can be immediately used to construct confidence sets for  $\theta_0$ , using a plug-in variance estimator; see Remark 4.4 for details.
- To the best of our knowledge, this is the first work proving semiparametric efficiency for an estimator of the finite dimensional parameter in a *bundled parameter* problem (where the parametric and nonparametric components are intertwined; see [27]) where the nonparametric estimate is shape constrained and non-smooth (in our case, the LLSE of  $m_0$  is a piecewise affine function).
- Due to the imposed shape constraint on  $m_0$ , the parametric submodels for the link function are nonlinear and the nuisance tangent space is intractable. Also, no least favorable submodel exists for the semiparametric model (1.1) for both the PLSE and the LLSE. This behavior can be attributed to the fact that both the estimators lie on the boundary of the parameter space; see [40] for a similar phenomenon. Furthermore, approximation to the least favorable submodels are not well-behaved and require further approximations for both the PLSE and the LLSE.
- Compared to the existing procedures that require the choice of multiple tuning parameters (see [11], [59], and [25] among others), our approaches require just one tuning parameter. Further, as explained in Section 6.4, the choice of the tuning parameter is less crucial (for our estimators) than the selection of the smoothing parameters for typical nonparametric problems. Moreover, the performance of the estimators is robust to the choice of the tuning parameter (see Section 6.4 and Figure 3 for an illustration and discussion), due to the assumed convexity constraint.
- In contrast to the existing approaches in a single index model where it is typically assumed that the index parameter belongs to a (known) bounded set in  $\mathbb{R}^d$  and that the first coordinate of the index parameter is fixed at 1 (see e.g., [41, 37]), we study the model under the (weaker) assumption that  $\theta_0 \in \Theta \subset S^{d-1}$ , a Stiefel manifold; see Hatcher [23, page 301].
- As is typical in single index models, the computation of the estimators is nontrivial: both the LLSE and the PLSE are optimizers of non-convex problems (both the loss function and the constraint set are non-convex) as the parameters  $m$  and  $\theta$  are bundled together. We employ an alternating minimization scheme to compute the estimators — if  $\theta$  is fixed the LLSE is obtained by solving a quadratic program with linear constraints, whereas for the PLSE, the estimator of  $m$  can be shown to be a natural cubic spline; we update  $\theta$  (with  $m$  fixed) by taking a small step along a retraction on the Stiefel manifold  $\Theta$  with a guarantee of descent (see Section 5 for the details; also see [58]). In the R package `simest` ([33]) we provide a fast and efficient implementation of these algorithms; in particular, the computation of the convex constrained spline in the PLSE is implemented in the C programming language. Since our optimization problems are non-convex multiple initializations may be required to find the global minimum. However, the assumed shape constraint appears to increase the size of the *basin of attraction* for both the proposed estimators, thereby ameliorating the problem of multiple local minima. Furthermore, both the LLSE and the PLSE have superior finite sample performance compared to existing procedures, even when  $d$  is large ( $d \approx 100$ ).

Our exposition is organized as follows: in Section 2 we introduce some notation and formally define the LLSE and the PLSE of  $(m_0, \theta_0)$ . In Sections 3.1 and 3.2 we state our assumptions, prove consistency, and give rates of convergence of the LLSE and the PLSE, respectively. In Section 4 we use these rates to prove efficiency and asymptotic normality of the PLSE and the LLSE of  $\theta_0$ . We discuss algorithms to compute the proposed estimators in Section 5. In Section 6 we provide an extensive simulation study and compare the finite sample performance of the proposed estimators with existing

methods in the literature. Section 7 provides a brief summary of the paper and discusses some open problems. Appendices A and B provide additional insights into the proofs of main Theorems 4.1 and 4.2, respectively. Appendix C provides further simulation studies, whereas Appendix D analyzes the Boston housing data and car mileage data. Appendices E-H contain the proofs omitted from the main text.

## 2. Estimation

### 2.1. Preliminaries

In what follows, we assume that we have i.i.d. data  $\{(x_i, y_i)\}_{1 \leq i \leq n}$  from (1.1). We start with some notation. Let  $\chi \subset \mathbb{R}^d$  denote the support of  $X$  and define

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

Let  $\mathcal{C}$  denote the class of real-valued convex functions on  $D$ ,  $\mathcal{S}$  denote the class of real-valued functions on  $D$  that have an absolutely continuous first derivative, and  $\mathfrak{L}_L$  denote the class of uniformly Lipschitz real-valued functions from  $D$  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 the real-valued convex function  $m$ . As  $m$  is a uniformly Lipschitz function with Lipschitz constant  $L$ , we can assume that  $|m'(t)| \leq L$ , for all  $t \in D$ . We use  $\mathbb{P}$  to denote the probability of an event,  $\mathbb{E}$  for the expectation of a random quantity, and  $P_X$  for the distribution of  $X$ . For  $g : \chi \rightarrow \mathbb{R}$ , define

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

Let  $P_{\epsilon, X}$  denote the joint distribution of  $(\epsilon, X)$  and let  $P_{\theta, m}$  denote the joint distribution of  $(Y, X)$  when  $Y := m(\theta^\top X) + \epsilon$ , where  $\epsilon$  is defined in (1.1). In particular,  $P_{\theta_0, m_0}$  denotes the joint distribution of  $(Y, X)$  when  $X \sim P_X$  and  $(Y, X)$  satisfies (1.1). For any set  $I \subseteq \mathbb{R}^p$  ( $p \geq 1$ ) and any function  $g : I \rightarrow \mathbb{R}$ , 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 differentiable function  $g : I \subseteq \mathbb{R} \rightarrow \mathbb{R}$ , the Sobolev norm is defined as

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

The notation  $a \lesssim b$  is used to express that  $a$  is less than  $b$  up to a constant multiple. For any function  $f : \chi \rightarrow \mathbb{R}^r$ ,  $r \geq 1$ , let  $\{f_i\}_{1 \leq i \leq r}$  denote each of the components of  $f$ , i.e.,  $f(x) = (f_1(x), \dots, f_r(x))$  and  $f_i : \chi \rightarrow \mathbb{R}$ . 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 function  $g : \mathbb{R} \rightarrow \mathbb{R}$  and  $\theta \in \Theta$ , we define

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

We use standard empirical process theory notation. For any function  $f : \mathbb{R} \times \chi \rightarrow \mathbb{R}$ ,  $\theta \in \Theta$ , and  $m : \mathbb{R} \rightarrow \mathbb{R}$ , 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  $\theta$  (or  $m$ ) is random. Moreover, for any function  $f : \mathbb{R} \times \mathcal{X} \rightarrow \mathbb{R}$ , we define

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

The following lemma (proved in Appendix E.1) proves the identification of the composite population parameter  $m_0 \circ \theta_0$ .

**Lemma 2.1.** Define  $Q(m, \theta) := \mathbb{E}[Y - m(\theta^\top X)]^2$ . Then

$$\inf_{\{(m, \theta) : m \circ \theta \in L_2(P_X) \text{ and } \|m \circ \theta - m_0 \circ \theta_0\| > \delta\}} Q(m, \theta) - Q(m_0, \theta_0) > \delta^2. \quad (2.1)$$

**Remark 2.1.** (2.1) tells us that one can hope to consistently estimate  $(m_0, \theta_0)$  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  $\theta_0$  are separately identifiable. [28] (also see [24]) finds sufficient conditions on the distribution/domain of  $X$  under which  $\theta_0$  and  $m_0$  can be separately identified when  $m_0$  is a non-constant almost everywhere differentiable function<sup>2</sup>:

(A0) Assume that  $\theta_{0,1} > 0$  and for some integer  $d_1 \in \{1, 2, \dots, d\}$ ,  $X_1, \dots, X_{d_1-1}$ , and  $X_{d_1}$  have continuous distributions and  $X_{d_1+1}, \dots, X_{d-1}$ , and  $X_d$  be discrete random variables. Furthermore, assume that for each  $\theta \in \Theta$  there exist an open interval  $\mathcal{I}$  and constant vectors  $c_0, c_1, \dots, c_{d-d_1} \in \mathbb{R}^{d-d_1}$  such that

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

## 2.2. Lipschitz constrained least squares estimator (LLSE)

The Lipschitz constrained least squares estimator is defined as the minimizer of the sum of squared errors

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

where  $m$  varies over the class of all convex  $L$ -Lipschitz functions  $\mathcal{M}_L$  and  $\theta \in \Theta \subset \mathbb{R}^d$ . Formally,

$$(\check{m}_n, \check{\theta}_n) := \arg \min_{(m, \theta) \in \mathcal{M}_L \times \Theta} Q_n(m, \theta). \quad (2.2)$$

Note that if the true link function  $m_0$  is  $L$ -Lipschitz, then  $(m_0, \theta_0) \in \mathcal{M}_L \times \Theta$ . For notational convenience, we suppress the dependence of  $(\check{m}_n, \check{\theta}_n)$  on  $L$ . The following theorem, proved in Appendix E, shows the existence of the minimizer in (2.2).

**Theorem 2.1.**  $(\check{m}_n, \check{\theta}_n) \in \mathcal{M}_L \times \Theta$ . Moreover,  $\check{m}_n$  is a piecewise affine convex function.

In Sections 3.1 and 4.3 we show that  $(\check{m}_n, \check{\theta}_n)$  is a consistent estimator of  $(m_0, \theta_0)$  and study its asymptotic properties.

**Remark 2.2.** For every fixed  $\theta$ ,  $m \in \mathcal{M}_L \mapsto Q_n(m, \theta)$  has a unique minimizer. The minimization over the class of uniformly Lipschitz functions is a quadratic program with linear constraints and can be computed easily; see Section 5.1.1.

<sup>2</sup>Note that all convex functions are almost everywhere differentiable.

### 2.3. Penalized least squares estimator (PLSE)

With the goal of making the estimator of  $m$  smooth, we propose the following penalized loss,

$$\mathcal{L}_n(m, \theta; \lambda) := Q_n(m, \theta) + \lambda^2 J^2(m), \quad (\lambda \neq 0). \quad (2.3)$$

The PLSE is now defined as

$$(\hat{m}_n, \hat{\theta}_n) := \arg \min_{(m, \theta) \in \mathcal{R} \times \Theta} \mathcal{L}_n(m, \theta; \lambda), \quad (2.4)$$

where  $\mathcal{R}$  denotes the class of all convex functions with absolutely continuous first derivative. As in the case of the LLSE, we suppress the dependence of  $(\hat{m}_n, \hat{\theta}_n)$  on the tuning parameter  $\lambda$ . The following theorem, proved in Appendix E, shows that the joint minimizer is well-defined and that  $\hat{m}_n$  is a natural cubic spline.

**Theorem 2.2.**  $(\hat{m}_n, \hat{\theta}_n) \in \mathcal{R} \times \Theta$ . Moreover,  $\hat{m}_n$  is a natural cubic spline.

In Sections 3.2 and 4.2 we study the asymptotic properties of  $(\hat{m}_n, \hat{\theta}_n)$ .

**Remark 2.3.** For every fixed  $\theta$ ,  $m \in \mathcal{R} \mapsto \mathcal{L}_n(m, \theta; \lambda)$  has a unique minimizer. [13] propose a damped Newton-type algorithm (with quadratic convergence) for finding the minimizer of this constrained penalized loss function (also see Section 2 of [16]); see Section 5.1.2.

### 3. Asymptotic analysis

In Sections 3.1 and 3.2 we study the asymptotic behavior of the estimators proposed in Sections 2.2 and 2.3, respectively. When there is no scope for confusion, for the rest of the paper, we use  $(\check{m}, \check{\theta})$  and  $(\hat{m}, \hat{\theta})$ , to denote  $(\check{m}_n, \check{\theta}_n)$  and  $(\hat{m}_n, \hat{\theta}_n)$ , respectively. We will now list the assumptions under which we prove the consistency and study the rates of convergence of the LLSE and the PLSE.

- (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  $\epsilon$  in model (1.1) is assumed to be uniformly sub-Gaussian, i.e., there exists  $K_1 > 0$  such that

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

As stated in (1.1), we also assume that  $\mathbb{E}(\epsilon | X) = 0$  a.e.  $X$ .

- (A3)  $\mathbb{E}[X X^\top \{m'_0(\theta_0^\top X)\}^2]$  is a nonsingular matrix.
- (A4)  $\text{Var}(X)$  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 an  $r > 0$ , 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 on  $D_\theta$  and bounded above by a finite constant (independent of  $\theta$ ). Furthermore, we assume that for every  $\theta \in \{\eta \in \Theta : |\eta - \theta_0| \leq r\}$ ,  $D_\theta \subsetneq D^{(r)}$ , where  $D^{(r)} := \cup_{|\theta - \theta_0| \leq r} D_\theta$ . For the rest of the paper we redefine  $D := D^{(r)}$ .

The above assumptions deserve comments. (A1) implies that the support of the covariates is bounded. As the classes of functions  $\mathcal{M}_L$  and  $\mathcal{R}$  are not uniformly bounded, we need sub-Gaussian assumption (A2) to provide control over the tail behavior of  $\epsilon$ ; see Chapter 8 of [50] for a discussion on this. Observe that (A2) allows for heteroscedastic errors. Assumptions (A3) and (A4) are mild distributional assumptions on the design. Assumption (A3) is similar to that in [41] and helps us obtain the rates of convergence of estimators of  $m_0$  and  $\theta_0$  separately from the rate of convergence of the estimators of  $m_0 \circ \theta_0$ . Assumption (A4) guarantees that the predictors are not supported on

a lower dimensional affine space. Assumption **(A5)** guarantees that  $D_0$ , the true index set, does not lie on the boundary of  $D$ . Assumption **(A5)** is needed to find rates of convergence of derivative of the estimators of  $m_0$ . If one of the continuous covariates with a nonzero index parameter (e.g.,  $X_1$ ) has a density that is bounded away from zero then assumption **(A5)** is satisfied.

### 3.1. Asymptotic analysis of the LLSE

In this subsection we study the asymptotic properties of the LLSE. The following assumption on  $m_0$  is used to prove that  $\tilde{m}$  is 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 we give a sequence of theorems (proved in Appendix **F**) characterizing the asymptotic properties of  $(\tilde{m}, \tilde{\theta})$ . Theorem **3.1** below proves the consistency and provides an upper bound on the rate of convergence of  $\tilde{m} \circ \tilde{\theta}$  to  $m_0 \circ \theta_0$  under the  $L_2(P_X)$  norm.

**Theorem 3.1.** Assume that **(A1)–(A4)** and **(L1)** hold. If  $L \geq L_0$ , then the LLSE satisfies

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

In the following two theorems, we prove consistency and find upper bounds on the rates of convergence of  $\tilde{\theta}$  and  $\tilde{m}$ .

**Theorem 3.2.** Under the assumptions of Theorem **3.1**, we have

$$|\tilde{\theta} - \theta_0| = o_p(1), \quad \|\tilde{m} - m_0\|_{D_0} = o_p(1), \quad \text{and} \quad \|\tilde{m}' - m_0'\|_C = o_p(1)$$

for any compact subset  $C$  in the interior of  $D_0$ .

**Theorem 3.3.** Under the assumptions of Theorem **3.1**, and the assumption that the conditional distribution of  $X$  given  $\theta_0^\top X$  is nondegenerate, the LLSE satisfies

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

Under additional smoothness assumptions on  $m_0$ , we show that  $\tilde{m}'$ , the right derivative of  $\tilde{m}$ , converges to  $m_0'$ .

**Theorem 3.4.** Assume that **(A1)–(A5)** and **(L1)** hold. If  $m_0$  is twice continuously differentiable on  $D_0$  and  $L \geq L_0$ , then we have that

$$\|\tilde{m}' \circ \theta_0 - m_0' \circ \theta_0\| = O_p(n^{-2/15}) \quad \text{and} \quad \int_{D_0} (\tilde{m}'(t) - m_0'(t))^2 dt = O_p(n^{-2/15}). \quad (3.1)$$

In fact,

$$\sup_{\theta \in \{\theta \in \Theta: |\theta_0 - \theta| \leq n^{-2/15}\}} \|\tilde{m}' \circ \theta - m_0' \circ \theta\| = O_p(n^{-2/15}). \quad (3.2)$$

In particular,

$$\|\tilde{m}' \circ \tilde{\theta} - m_0' \circ \tilde{\theta}\| = O_p(n^{-2/15}). \quad (3.3)$$

The fact that  $\tilde{m}'$  is a step function complicates the proof of the above result (given in Appendix **F.6**). In fact, the obtained rate need not be optimal, but is sufficient for our purposes (in deriving the efficiency of  $\tilde{\theta}$ ; see Section **4.3**).



### 3.2. Asymptotic analysis of the PLSE

In this subsection we give results on the asymptotic properties of  $(\hat{m}, \hat{\theta})$ . Note that we will study  $(\hat{m}, \hat{\theta})$  for any random  $\lambda$  satisfying some rate conditions. The smoothing parameter  $\lambda$  can be chosen to be a random variable. For the rest of the paper, we denote it by  $\hat{\lambda}_n$ . First, we need some smoothness assumption on  $m_0$ . We assume:

- (P1) The unknown convex link function  $m_0$  is bounded by some constant  $M_0$  on  $D$ , has an absolutely continuous first derivative, and satisfies  $J(m_0) < \infty$ .  
(P2)  $\hat{\lambda}_n$  satisfies the rate conditions:

$$\hat{\lambda}_n^{-1} = O_p(n^{2/5}) \quad \text{and} \quad \hat{\lambda}_n = o_p(n^{-1/4}). \quad (3.4)$$

Our assumption (P1) on  $m_0$  is quite minimal — we essentially require  $m_0$  to have an absolutely continuous derivative. Assumption (P2) 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 (3.4) will result in an asymptotically efficient estimator of  $\theta_0$ . Now in a sequence of theorems, 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 3.5.** *Under assumptions (A0)–(A4) and (P1)–(P2), the PLSE satisfies*

$$J(\hat{m}) = O_p(1), \quad \|\hat{m}\|_\infty = O_p(1), \quad \text{and} \quad \|\hat{m} \circ \hat{\theta} - m_0 \circ \theta_0\| = O_p(\hat{\lambda}_n).$$

We now establish the consistency and find the rates of convergence of  $\hat{m}$  (in the Sobolev norm) and  $\hat{\theta}$  (in the Euclidean norm).

**Theorem 3.6.** *Under assumptions (A0)–(A4) and (P1)–(P2),*

$$\hat{\theta} \xrightarrow{P} \theta_0, \quad \|\hat{m} - m_0\|_{D_0}^S \xrightarrow{P} 0, \quad \text{and} \quad \|\hat{m}'\|_\infty = O_p(1).$$

**Theorem 3.7.** *Under assumptions (A0)–(A4) and (P1)–(P2), and the assumption that the conditional distribution of  $X$  given  $\theta_0^\top X$  is nondegenerate,  $\hat{m}$  and  $\hat{\theta}$  satisfy*

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

The proofs of Theorems 3.5, 3.6, and 3.7 follow from proof of Theorems 2, 3, and 4 of [32], respectively. Even though the estimator proposed in [32] is not constrained to be convex, the proofs of [32] can be easily modified for the PLSE; see Appendix G.1 for a brief discussion.

The following theorem, proved in Appendix G.2, provides an upper bound on the rate of convergence of the derivative of  $\hat{m}$ . This upper bound will be useful for computing the asymptotic distribution of  $\hat{\theta}$  in Section 4.2.

**Theorem 3.8.** *Under the assumptions of Theorem 3.7 and (A5), we have*

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

## 4. Semiparametric inference

The main results in this section show that  $\hat{\theta}$  and  $\check{\theta}$  are  $\sqrt{n}$ -consistent and asymptotically normal (see Sections 4.2 and 4.3, respectively). Moreover, both the estimators are shown to be semiparametrically efficient for  $\theta_0$  under homoscedastic errors. The asymptotic analysis of  $\check{\theta}$  is more involved (than that of  $\hat{\theta}$ ) as  $\check{m}$  is a piecewise affine function and hence not differentiable everywhere (while  $\hat{m}$  is a smooth



function). For this reason, we shall at first present the theory for  $\hat{\theta}$  and then proceed to do the same for  $\tilde{\theta}$ .

Before going into the derivation of the limit law of the proposed estimators of  $\theta_0$ , we need to introduce some further notation and regularity assumptions. Let  $p_{\epsilon, X}$  denote the joint density (with respect to some dominating measure  $\mu$  on  $\mathbb{R} \times \mathcal{X}$ ) of  $(\epsilon, X)$ . Let  $p_{\epsilon|X}(\cdot, x)$  and  $p_X(\cdot)$  denote the corresponding conditional probability density of  $\epsilon$  given  $X = x$  and the marginal density of  $X$ , respectively. We define  $\sigma : \mathcal{X} \rightarrow \mathbb{R}^+$  such that

$$\sigma^2(x) := \mathbb{E}(\epsilon^2 | X = x).$$

(B1) Assume that  $m_0$  is three times differentiable and that  $m_0'''$  is bounded on  $D$ . Furthermore, let  $m_0$  be strongly convex on  $D$ , i.e., for all  $s \in D$  we have  $m_0''(s) \geq \delta_0 > 0$  for some fixed  $\delta_0$ .

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

$$h_\theta(u) := \mathbb{E}[X | \theta^\top X = u]. \quad (4.1)$$

(B2) Assume that  $h_\theta(\cdot)$  is twice continuously differentiable except possibly at a finite number of points, and there exists a finite constant  $\bar{M} > 0$  such that for every  $\theta_1, \theta_2 \in \Theta$ ,

$$\|h_{\theta_1} - h_{\theta_2}\|_\infty \leq \bar{M}|\theta_1 - \theta_2|. \quad (4.2)$$

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

Assumptions (B1)–(B3) deserve comments. The function  $h_\theta$  plays a crucial role in the construction of “least favorable” paths and is part of the efficient score function; see Appendix A.1. For the functions in the path to be in  $\mathcal{R}$  or  $\mathcal{M}_L$ , we need the smoothness assumption (B2) on  $h_\theta$ . We need the lower and upper bounds on the variance function as we are using a non-weighted least squares method to estimate parameters in a (possibly) heteroscedastic model.

#### 4.1. Efficient score

First observe that the parameter space  $\Theta$  is a closed subset of  $\mathbb{R}^d$  and the interior of  $\Theta$  in  $\mathbb{R}^d$  is the null set. Thus to compute the score for the model in (1.1), we construct a path on the sphere. We use  $\mathbb{R}^{d-1}$  to parametrize the paths for model (1.1) on the sphere. For each  $\eta \in \mathbb{R}^{d-1}$ ,  $s \in \mathbb{R}$ , and  $|s| \leq |\eta|^{-1}$ , define the following path<sup>3</sup> through  $\theta$  (which lies on the unit sphere)

$$\zeta_s(\theta, \eta) := \sqrt{1 - s^2|\eta|^2} \theta + s H_\theta \eta, \quad (4.3)$$

where for every  $\theta \in \Theta$ ,  $H_\theta \in \mathbb{R}^{d \times (d-1)}$  satisfies the following properties:

- (H1)  $\xi \mapsto H_\theta \xi$  are bijections from  $\mathbb{R}^{d-1}$  to the hyperplanes  $\{x \in \mathbb{R}^d : \theta^\top x = 0\}$ .
- (H2) The columns of  $H_\theta$  form an orthonormal basis for  $\{x \in \mathbb{R}^d : \theta^\top x = 0\}$ .
- (H3)  $\|H_\theta - H_{\theta_0}\|_2 \leq |\theta - \theta_0|$ .
- (H4) For all distinct  $\eta, \beta \in \Theta \setminus \theta_0$ , such that  $|\eta - \theta_0| \leq 1/2$  and  $|\beta - \theta_0| \leq 1/2$ ,

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

See Lemma 1 of [32] for a construction of a class of matrices satisfying the above properties.

In the following two subsections we attempt to calculate the efficient score for the model:

$$Y = m(\theta^\top X) + \epsilon, \quad (4.4)$$

where  $m \in \mathcal{R}$  or  $m \in \mathcal{M}_L$ . We will see that the efficient score is intractable when  $m$  is at the boundary of  $\mathcal{R}$  (or  $\mathcal{M}_L$ ), but we can work with a ‘surrogate’ score.

---

<sup>3</sup>Here  $\eta$  defines the “direction” of the path.

#### 4.1.1. Efficient score when $(m, \theta) \in \mathcal{R} \times \Theta$

The log-likelihood of the model is

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

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 in  $\Theta$  through  $\theta$  as  $\zeta_0(\theta, \eta) = \theta$  and  $\zeta_s(\theta, \eta) \in \Theta$  for every  $s$  in some neighborhood of 0, as  $H_\theta \eta$  is orthogonal to  $\theta$  (by (H1)) and  $|H_\theta \eta| = |\eta|$  (by (H2)). The parametric score for this submodel 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

$$S_{\theta, m}(y, x) := -\frac{p'_{\epsilon|X}(y - m(\theta^\top x), x)}{p_{\epsilon|X}(y - m(\theta^\top x), x)} m'(\theta^\top x) H_\theta^\top x. \quad (4.5)$$

**Remark 4.1.** Note that under (4.4), we have  $\epsilon = Y - m(\theta^\top X)$ . For every function  $b(e, x) : \mathbb{R} \times \mathcal{X} \rightarrow \mathbb{R}$  in  $L_2(P_{\epsilon, X})$  there exists an “equivalent” function  $\tilde{b}(y, x) : \mathbb{R} \times \mathcal{X} \rightarrow \mathbb{R}$  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:

$$\begin{aligned} m_{s, a}(t) &= m(t) - sa(t), \\ p_{\epsilon|X; s, b}(e, x) &= p_{\epsilon|X}(e, x)(1 + sb(e, x)), \\ p_{X; s, q}(x) &= p_X(x)(1 + sq(x)), \end{aligned} \quad (4.6)$$

where  $s \in \mathbb{R}$ ,  $b : \mathbb{R} \times \mathcal{X} \rightarrow \mathbb{R}$  is a bounded function such that  $\mathbb{E}(b(\epsilon, X)|X) = 0$  and  $\mathbb{E}(eb(\epsilon, X)|X) = 0$ ,  $a \in \mathcal{S}$  such that  $J(a) < \infty$  and  $m_{s, a} \in \mathcal{R}$  for every  $s$  in some neighborhood of 0 and  $q : \mathcal{X} \rightarrow \mathbb{R}$  is a bounded function such that  $\mathbb{E}(q(X)) = 0$ . Consider the following parametric submodel of (4.4),

$$s \mapsto (\zeta_s(\theta, \eta), m_{s, a}, p_{\epsilon|X; s, b}, p_{X; s, q}(x)) \quad (4.7)$$

where  $\eta \in S^{d-2}$ . Differentiating the log-likelihood of the submodel in (4.7) with respect to  $s$ , we get that the score along the submodel in (4.7) is

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

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

$$\begin{aligned} \Lambda_S &:= \overline{\text{lin}} \left\{ f \in L_2(P_{\epsilon, X}) : f(e, x) = \frac{p'_{\epsilon|X}(e, x)}{p_{\epsilon|X}(e, x)} a(\theta^\top x) + b(e, x) + q(x), \right. \\ &\quad \text{where } a \in \mathcal{S}, J(a) < \infty \text{ and } m_{s, a} \in \mathcal{R} \text{ for small enough } s, \\ &\quad b : \mathbb{R} \times \mathcal{X} \rightarrow \mathbb{R} \text{ and } q : \mathcal{X} \rightarrow \mathbb{R} \text{ are bounded functions, } \mathbb{E}(eb(\epsilon, X)|X) = 0, \\ &\quad \left. \mathbb{E}(b(\epsilon, X)|X) = 0, \text{ and } \mathbb{E}(q(X)) = 0 \right\}, \end{aligned}$$

where for any set  $A \subseteq L_2(P_{\theta, m})$ ,  $\overline{\text{lin}} A$  denotes the closure in  $L_2(P_{\theta, m})$  of the linear span of functions in  $A$ ; see [42] 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

$$\overline{\text{lin}} \{a \in \mathcal{S} : J(a) < \infty \text{ and } m_{s, a} \in \mathcal{R} \text{ for small enough } s\} \subseteq \overline{\text{lin}} \{a \in \mathcal{S} : J(a) < \infty\} \quad (4.8)$$

and

$$\overline{\text{lin}}\{q : \chi \rightarrow \mathbb{R} \mid q \text{ is a bounded function and } \mathbb{E}(q(X)) = 0\} = \{q : \chi \rightarrow \mathbb{R} \mid q \in L_2(P_X) \text{ and } \mathbb{E}(q(X)) = 0\}.$$

However, by Theorem A.1 of [19], we have that the class of infinitely often differentiable functions on  $D$  (a bounded subset of  $\mathbb{R}$ ) is dense in  $L_2(\mathbf{m})$ , where  $\mathbf{m}$  denotes the Lebesgue measure on  $D$ . Thus we have that

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

and  $\overline{\text{lin}}\{b : \mathbb{R} \times \chi \rightarrow \mathbb{R} \mid b \text{ is a bounded function, } \mathbb{E}(b(\epsilon, X)|X) = \mathbb{E}(b(\epsilon, X)|X) = 0\} = \{b \in L_2(P_{\epsilon, X}) : \mathbb{E}(b(\epsilon, X)|X) = \mathbb{E}(b(\epsilon, X)|X) = 0\}$ . Thus, it is easy to see that under assumptions **(A0)**–**(A4)**, **(P1)**, and **(B1)**–**(B3)**, the nuisance tangent space of (1.1) satisfies

$$\Lambda_S \subseteq \left\{ f \in L_2(P_{\epsilon, X}) : f(e, x) = \frac{p'_{\epsilon|X}(e, x)}{p_{\epsilon|X}(e, x)} a(\theta^\top x) + b(e, x) + q(x), \right. \quad (4.9)$$

$$\left. \text{where } a \in L_2(\mathbf{m}), b \in L_2(P_{\epsilon, X}), q \in L_2(P_X), \mathbb{E}(b(\epsilon, X)|X) = 0, \right. \\ \left. \mathbb{E}(b(\epsilon, X)|X) = 0, \text{ and } \mathbb{E}(q(X)) = 0 \right\} =: \Lambda_0.$$

Note that  $\Lambda_S$  and  $\Lambda_0$  differ as the set inclusion in (4.8) could be strict. However, it can be easily seen that, if  $m$  is strongly convex then  $\Lambda_S = \Lambda_0$ .

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  $\Lambda_S$  in  $L_2(P_{\theta, m})$ . Newey and Stoker [43] and Ma and Zhu [38] show that

$$\Lambda_0^\perp = \left\{ f \in L_2(P_{\epsilon, X}) : f(e, x) = [g(x) - \mathbb{E}(g(X)|\theta^\top X = \theta^\top x)]e, \text{ and } g : \chi \rightarrow \mathbb{R} \right\} \subseteq \Lambda_S^\perp \quad (4.10)$$

where  $\Lambda_0$  is defined in (4.9). Using calculations similar to those in Theorem 4.1 of [43] and Proposition 1 of [38], it can be shown that

$$\Pi(S_{\theta, m}|\Lambda_0^\perp)(y, x) = \frac{1}{\sigma^2(x)}(y - m(\theta^\top x))m'(\theta^\top x)H_\theta^\top \left\{ x - \frac{\mathbb{E}(\sigma^{-2}(X)X|\theta^\top X = \theta^\top x)}{\mathbb{E}(\sigma^{-2}(X)|\theta^\top X = \theta^\top x)} \right\},$$

where for any  $f \in L_2(P_{\theta, m})$ ,  $\Pi(f|\Lambda_0^\perp)$  denotes the  $L_2(P_{\theta, m})$  projection of  $f$  onto the space  $\Lambda_0^\perp$ .

However to compute the efficient score of (4.4) when  $m \in \mathcal{R}$ , we need to evaluate  $\Pi(S_{\theta, m}|\Lambda_S^\perp)(y, x)$ . And computation of  $\Pi(S_{\theta, m}|\Lambda_S^\perp)(y, x)$  is infeasible due to the complicated nature of the set of parametric submodels of  $m$ . Note that the efficient information (of (4.4) when  $m \in \mathcal{R}$ ) is denoted by

$$\mathcal{I}_{\theta, m}^S := P_{\theta, m}[\Pi(S_{\theta, m}|\Lambda_S^\perp)\Pi^\top(S_{\theta, m}|\Lambda_S^\perp)].$$

As  $\Lambda_0^\perp \subseteq \Lambda_S^\perp$  (see (4.10)), we have

$$\mathcal{I}_{\theta, m}^S \geq P_{\theta, m}[\Pi(S_{\theta, m}|\Lambda_0^\perp)\Pi^\top(S_{\theta, m}|\Lambda_0^\perp)] =: \mathcal{I}_{\theta, m}^0.$$

Moreover, we see that at the true parameter values  $(m_0, \theta_0)$ , as  $m_0$  is strongly convex,

$$\Pi(S_{\theta_0, m_0}|\Lambda_S^\perp) \equiv \Pi(S_{\theta_0, m_0}|\Lambda_0^\perp) \quad \text{and} \quad \mathcal{I}_{\theta_0, m_0}^S = \mathcal{I}_{\theta_0, m_0}^0.$$

Once the efficient score is calculated, one usually finds an efficient estimator of  $(m_0, \theta_0)$  by solving the efficient estimating equation, i.e., by finding a  $(m, \theta)$  that satisfy

$$\mathbb{P}_n \Pi(S_{\theta, m}|\Lambda_S^\perp) = 0. \quad (4.11)$$

However since  $\Pi(S_{\theta, m}|\Lambda_S^\perp)$  is intractable when  $m$  is at the boundary of  $\mathcal{R}$ , we use  $\Pi(S_{\theta, m}|\Lambda_0^\perp)$  as its surrogate. In Section 4.2, we show that  $(\hat{m}, \hat{\theta})$  approximately satisfies (4.11) with the surrogate score (see (4.12)) and this enables us to prove that  $\hat{\theta}$  is an efficient estimator of  $\theta_0$ .

Lastly, it is important to note that (4.11), the efficient 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 rates 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 may lead to poor finite sample performance; see Tsiatis [48, page 93].

To focus our presentation on the main concepts, briefly consider the case when  $\sigma^2(\cdot) \equiv \sigma^2$ . In this simplified case, we have

$$\Pi(S_{\theta,m}|\Lambda_0^\perp)(y, x) = \frac{1}{\sigma^2}(y - m(\theta^\top x))m'(\theta^\top x)H_\theta^\top \{x - h_\theta(\theta^\top x)\},$$

where  $h_\theta(\theta^\top x)$  is defined in (4.1). 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.,

$$\sqrt{n}\mathbb{P}_n\Pi(S_{\hat{\theta},\hat{m}}|\Lambda_0^\perp) = \sqrt{n}\mathbb{P}_n\left[\frac{1}{\sigma^2}(Y - \hat{m}(\hat{\theta}^\top X))H_{\hat{\theta}}^\top \hat{m}'(\hat{\theta}^\top X)\{X - h_{\hat{\theta}}(\hat{\theta}^\top X)\}\right] = o_p(1), \quad (4.12)$$

and the class of functions  $\Pi(S_{\theta,m}|\Lambda_0^\perp)$  indexed by  $(\theta, m)$  in a “neighborhood” of  $(\theta_0, m_0)$  satisfies some technical conditions. We formalize these in Section 4.2 and Appendix A.1.

#### 4.1.2. Efficient score when $(m, \theta) \in \mathcal{M}_L \times \Theta$

As  $m \in \mathcal{M}_L$  need not be differentiable everywhere, showing that the underlying class of distributions is differentiable in quadratic mean requires some careful analysis; in Remark H.1 (in the Appendix) we show this for the model with Gaussian errors. We can further show that the parametric score in this model satisfies (4.5), where  $m'$  denotes the right derivative of  $m$ . Moreover, using parametric submodel as in (4.6) and (4.7) and calculations similar to those in Section 4.1.1, it can be shown that the nuisance tangent space, denoted by  $\Lambda_L$ , of the model is

$$\Lambda_L := \overline{\text{lin}}\left\{f \in L_2(P_{e,X}) : f(e, x) = \frac{p'_{e|X}(e, x)}{p_{e|X}(e, x)}a(\theta^\top x) + b(e, x) + q(x),\right.$$

where  $a \in L^2(\mathbf{m})$ ,  $m_{s,a} \in \mathcal{M}_L$  for small enough  $s$ ,  $b : \mathbb{R} \times \mathcal{X} \rightarrow \mathbb{R}$

and  $q : \mathcal{X} \rightarrow \mathbb{R}$  are bounded functions,  $\mathbb{E}(eb(\epsilon, X)|X) = 0$ ,

$$\mathbb{E}(b(\epsilon, X)|X) = 0, \text{ and } \mathbb{E}(q(X)) = 0\}.$$

Now using arguments similar to those in Section 4.1.1, it can be shown that

$$\mathcal{I}_{\theta,m}^L \geq P_{\theta,m}[\Pi(S_{\theta,m}|\Lambda_0^\perp)\Pi^\top(S_{\theta,m}|\Lambda_0^\perp)] = \mathcal{I}_{\theta,m}^0, \quad (4.13)$$

where  $\mathcal{I}_{\theta,m}^L := P_{\theta,m}[\Pi(S_{\theta,m}|\Lambda_L^\perp)\Pi^\top(S_{\theta,m}|\Lambda_L^\perp)]$  and  $S_{\theta,m}(Y, X)$  and  $\Lambda_0$  are defined as in (4.5) and (4.9), respectively. It can easily be seen that  $\Lambda_L \subseteq \Lambda_0$ . In fact, if  $m$  is strongly convex then  $\Lambda_L = \Lambda_0$ . However for a general (non-strongly convex)  $m$ ,  $\Lambda_L$  can be a strict subset of  $\Lambda_0$  and the inequality in (4.13) can be strict.

**Remark 4.2.** Assumptions (A0)–(A4) and (P1) (or (L1)) do not guarantee the existence of a least favorable submodel for the model in (1.1), which can be the case when the estimators lie on the “boundary” of the parameter set. Note that both the estimators  $\hat{m}$  and  $\hat{\theta}$  lie at the “boundary” of the respective parameter sets. van der Vaart [51] 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  $\Pi(S_{\theta_0,m_0}|\Lambda_0^\perp)$  is the efficient score at  $(\theta_0, m_0)$ ; see Appendix A.1 and Theorem B.1 for the PLSE and the LLSE, respectively. However, the score corresponding to the approximately least favorable subprovided model does not satisfy the conditions required in [51] 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  $(\hat{m}, \hat{\theta})$  (or  $(\tilde{m}, \tilde{\theta})$ ) is an approximate zero of the corresponding estimating equation; see (4.21).

#### 4.2. Efficiency of the PLSE

The following result gives the limiting distribution of the PLSE  $\hat{\theta}$  and establishes its semiparametric efficiency (under homoscedasticity).

**Theorem 4.1.** Assume  $(X, Y)$  satisfies (1.1) and assumptions (A0)–(A5), (B1)–(B3), and (P1)–(P2) hold. Define the function

$$\ell_{\theta,m}(y, x) := (y - m(\theta^\top x))m'(\theta^\top x)H_\theta^\top \left\{ x - h_\theta(\theta^\top x) \right\}. \quad (4.14)$$

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

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0, H_{\theta_0} V_{\theta_0, m_0}^{-1} I_{\theta_0, m_0} (H_{\theta_0} V_{\theta_0, m_0}^{-1})^\top),$$

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  $I_{\theta_0, m_0}$  is nonsingular, then  $\hat{\theta}$  is an efficient estimator of  $\theta_0$ , i.e.,

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

**Remark 4.3.** Observe that the variance of the limiting distribution (for both the heteroscedastic and homoscedastic models) is singular. This can be attributed to the fact that  $\Theta$  is a Stiefel manifold of dimension  $\mathbb{R}^{d-1}$  and has an empty interior in  $\mathbb{R}^d$ .

**Remark 4.4** (Construction of confidence sets). Theorem 4.1 shows that (under homoscedastic errors) the PLSE of  $\theta_0$  is  $\sqrt{n}$ -consistent and asymptotically normal with covariance matrix:

$$\Sigma^0 := \sigma^4 H_{\theta_0} P_{\theta_0, m_0} [\ell_{\theta_0, m_0}^\top(Y, X) \ell_{\theta_0, m_0}^\top(Y, X)]^{-1} H_{\theta_0}^\top,$$

where  $\ell_{\theta_0, m_0}$  is defined in (4.14). This result can be used to create confidence sets for  $\theta_0$ . However since  $\Sigma^0$  is unknown, we propose using the following plug-in estimator of  $\Sigma^0$

$$\hat{\Sigma} := \hat{\sigma}^4 H_{\hat{\theta}} P_{\hat{\theta}, \hat{m}} [\ell_{\hat{\theta}, \hat{m}}^\top(Y, X) \ell_{\hat{\theta}, \hat{m}}^\top(Y, X)]^{-1} H_{\hat{\theta}}^\top,$$

where  $\hat{\sigma}^2 := \sum_{i=1}^n [y_i - \hat{m}(\hat{\theta}^\top x_i)]^2 / n$ . One can easily show that Theorems 3.5–3.8 imply consistency of  $\hat{\Sigma}$ . For example one can construct the following  $1 - 2\alpha$  confidence interval for  $\theta_{0,i}$

$$\left[ \hat{\theta}_i - \frac{z_\alpha}{\sqrt{n}} \left( \hat{\Sigma}_{i,i} \right)^{1/2}, \hat{\theta}_i + \frac{z_\alpha}{\sqrt{n}} \left( \hat{\Sigma}_{i,i} \right)^{1/2} \right], \quad (4.15)$$

where  $z_\alpha$  denotes the upper  $\alpha$ th-quantile of the standard normal distribution; see Section 6.2 for a simulation example. A similar analysis can be done for the LLSE using Theorem 4.2.

*Proof.* We give a sketch of the proof below. Some of the steps are proved in Appendix A.

**Step 1** In Theorem A.1 we find an approximately least favorable subprovided model (see Definition 9.7 of [51]) with score

$$\begin{aligned} \mathfrak{S}_{\theta,m}(x, y) = \{y - m(\theta^\top x)\} H_\theta^\top & \left[ m'(\theta^\top x) x + \int_{s_0}^{\theta^\top x} m'(u) k'(u) du - m'(\theta^\top x) k(\theta^\top x) \right. \\ & \left. + m'_0(s_0) k(s_0) - m'_0(s_0) h_{\theta_0}(s_0) \right] \end{aligned} \quad (4.16)$$

where  $k : D \rightarrow \mathbb{R}^d$  is defined as

$$k(u) := h_{\theta_0}(u) + \frac{m'_0(u)}{m''_0(u)} h'_{\theta_0}(u). \quad (4.17)$$

We prove that there exists a constant  $M^* < \infty$  such that

$$\sup_{u \in D} (|k(u)| + |k'(u)|) \leq M^*. \quad (4.18)$$

Moreover,  $(\hat{\theta}, \hat{m})$  satisfies the score equation approximately, i.e.,

$$\sqrt{n} \mathbb{P}_n \mathfrak{S}_{\hat{\theta}, \hat{m}} = o_p(1). \quad (4.19)$$

Furthermore, define  $\psi_{\theta,m} : \mathcal{X} \times \mathbb{R} \rightarrow \mathbb{R}^{d-1}$  as

$$\psi_{\theta,m}(x, y) := (y - m(\theta^\top x))H_\theta^\top [m'(\theta^\top x) - h_{\theta_0}(\theta^\top x)m'_0(\theta^\top x)]. \quad (4.20)$$

Although  $\mathfrak{S}_{\hat{\theta},\hat{m}}$  satisfies the score equation approximately it is quite complicated to deal with. The function  $\psi_{\theta,m}$  is an approximation to  $\mathfrak{S}_{\theta,m}$  and  $\psi_{\theta_0,m_0} = \mathfrak{S}_{\theta_0,m_0} = \ell_{\theta_0,m_0}$  (see (4.14)). Furthermore,  $\psi_{\theta,m}$  is well-behaved in the sense that:  $\psi_{\hat{\theta},\hat{m}}$  belongs to a Donsker class of functions (see (4.23)) and  $\psi_{\hat{\theta},\hat{m}}$  converges to  $\psi_{\theta_0,m_0}$  in the  $L_2(P_{\theta_0,m_0})$  norm; see Lemma H.1.

**Step 2** In Theorem A.2 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 (4.19) we have that  $\hat{\theta}$  is an *approximate* zero of the function  $\theta \mapsto \mathbb{P}_n \psi_{\theta,\hat{m}}$ , i.e.,

$$\sqrt{n}\mathbb{P}_n \psi_{\hat{\theta},\hat{m}} = o_p(1). \quad (4.21)$$

**Step 3** In Theorem A.3 we show that  $\psi_{\hat{\theta},\hat{m}}$  is approximately unbiased in the sense of [51], i.e.,

$$\sqrt{n}P_{\hat{\theta},m_0} \psi_{\hat{\theta},\hat{m}} = o_p(1). \quad (4.22)$$

Similar conditions have appeared before in proofs of asymptotic normality of maximum likelihood estimators (e.g., see [26]) and the construction of efficient one-step estimators (see [30]). The above condition essentially ensures that  $\psi_{\hat{\theta},\hat{m}}$  is a good “approximation” to  $\psi_{\theta_0,m_0}$ ; see Section 3 of [40] for further discussion.

**Step 4** We prove

$$\mathbb{G}_n(\psi_{\hat{\theta},\hat{m}} - \psi_{\theta_0,m_0}) = o_p(1) \quad (4.23)$$

in Theorem A.4. 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 (4.21) and (4.22), we have that (4.23) is equivalent to

$$\sqrt{n}(P_{\hat{\theta},m_0} - P_{\theta_0,m_0})\psi_{\hat{\theta},\hat{m}} = \mathbb{G}_n \ell_{\theta_0,m_0} + o_p(1). \quad (4.24)$$

**Step 5** To complete the proof, it is now enough to show that

$$\sqrt{n}(P_{\hat{\theta},m_0} - P_{\theta_0,m_0})\psi_{\hat{\theta},\hat{m}} = \sqrt{n}V_{\theta_0,m_0}H_{\theta_0}^\top(\hat{\theta} - \theta_0) + o_p(\sqrt{n}|\hat{\theta} - \theta_0|). \quad (4.25)$$

A proof of (4.25) can be found in the proof of Theorem 6.20 in [51]; also see [32, Section 10.4]. Lemma H.1 in Appendix H.8 proves that  $(\hat{\theta}, \hat{m})$  satisfy the required conditions of Theorem 6.20 in [51]. Observe that (4.24) and (4.25) 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 Step 4 of Theorem 5 in [32].  $\square$

### 4.3. Efficiency of the LLSE

In this section we show that  $\check{\theta}$  is an asymptotically normal efficient estimator of  $\theta_0$ . The following theorem is similar to Theorem 4.1.

**Theorem 4.2.** Assume  $(X, Y)$  satisfies (1.1) and assumptions (A0)–(A5), (B1)–(B3), and (L1) hold. Let  $\ell_{\theta, m}$ ,  $V_{\theta_0, m_0}$ , and  $I_{\theta_0, m_0}$  be as defined in Theorem 4.1. If  $V_{\theta_0, m_0}$  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  $(I_{\theta_0, m_0})$  is nonsingular, then  $\check{\theta}$  is an efficient estimator of  $\theta_0$ , i.e.,

$$\sqrt{n}(\check{\theta} - \theta_0) \xrightarrow{d} N(0, \sigma^4 H_{\theta_0} I_{\theta_0, m_0}^{-1} H_{\theta_0}^\top).$$

In Appendix B, we prove Theorem 4.2 via a series of results by showing that  $(\check{\theta}, \check{m})$  satisfy the conditions in Step 1–Step 5 of Theorem 4.1. These proofs/verifications of Step 1–Step 4 for the LLSE are more complicated (when compared to that of the PLSE) as  $\check{m}$  is not differentiable everywhere.

## 5. Computational algorithms

In this section we describe algorithms for computing the estimators defined in (2.2) and (2.4). As mentioned in Remarks 2.2 and 2.3, in each of these cases, the minimization of the desired loss function for a fixed  $\theta$  is a convex optimization problem; see Sections 5.1.1 and 5.1.2 below for more details. With the above observation in mind, we propose the following general alternating minimization algorithm to compute the proposed estimators. The algorithms discussed here are implemented in our R package `simest` [33].

We first introduce some notation. Let  $(m, \theta) \mapsto \mathfrak{C}(m, \theta)$  denote a 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  $\mathcal{R}$  or  $\mathcal{M}_L$ . For every  $\theta \in \Theta$ , let us define

$$m_{\theta, \mathfrak{A}} := \arg \min_{m \in \mathfrak{A}} \mathfrak{C}(m, \theta). \quad (5.1)$$

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 general alternating scheme is described in Algorithm 1.

---

#### Algorithm 1: Alternating minimization algorithm

---

**Input:** Initialize  $\theta$  at  $\theta^{(0)}$ .

**Output:**  $(m^*, \theta^*) := \arg \min_{(m, \theta) \in \mathfrak{A} \times \Theta} \mathfrak{C}(m, \theta)$ .

- 1 At iteration  $k \geq 0$ , compute  $m^{(k)} := m_{\theta^{(k)}, \mathfrak{A}} = \arg \min_{m \in \mathfrak{A}} \mathfrak{C}(m, \theta^{(k)})$ .
- 2 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  $\theta \mapsto \mathfrak{C}(m^{(k)}, \theta)$ .

- 3 Repeat steps 1 and 2 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 convex. Thus the algorithm discussed above is not guaranteed to converge to a global minimizer. However, the algorithm guarantees that the criterion value is nonincreasing over iterations, i.e.,  $\mathfrak{C}(m^{(k+1)}, \theta^{(k+1)}) \leq \mathfrak{C}(m^{(k)}, \theta^{(k)})$  for all  $k \geq 0$ . In Section 5.1.1 we discuss an algorithm to compute  $m_{\theta, \mathcal{M}_L}$ , when  $\mathfrak{C}(m, \theta) = Q_n(m, \theta)$  while in Section 5.1.2 we discuss the computation of  $m_{\theta, \mathcal{R}}$  when  $\mathfrak{C}(m, \theta) = \mathcal{L}_n(m, \theta; \lambda)$ .



### 5.1. Strategy for estimating the link function

In the following subsections we describe algorithms to compute  $m_{\theta, \mathcal{R}}$  and  $m_{\theta, \mathcal{M}_L}$  as defined in (5.1). We use the following notation. Fix an arbitrary  $\theta \in \Theta$ . Let  $(t_1, t_2, \dots, t_n)$  represent the vector  $(\theta^\top x_1, \dots, \theta^\top x_n)$  with sorted entries so that  $t_1 < t_2 < \dots < t_n$ ; in Remark 5.1 we discuss a solution for scenarios with ties. Without loss of generality, let  $y := (y_1, y_2, \dots, y_n)$  represent the vector of responses corresponding to the sorted  $t_i$ .

#### 5.1.1. Lipschitz constrained least squares (LLSE)

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  $m \in \mathcal{M}_L$ . In the following we use  $m$  to denote the function  $t \mapsto m(t)$  as well as the the vector  $(m(t_1), \dots, m(t_n))$  interchangeably. Consider the general problem of minimizing

$$(y - m)Q(y - m) = |Q^{1/2}(y - m)|^2,$$

for some positive definite matrix  $Q$ . In most cases  $Q$  is the  $n \times n$  identity matrix; see Remark 5.1 for other possible scenarios. 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 the piecewise linear interpolation of  $\{m_i\}_{1 \leq i \leq n}$  with possible kinks only at  $\{t_i\}_{1 \leq i \leq n}$ . The Lipschitz constraint along with convexity (i.e.,  $m \in \mathcal{M}_L$ ) reduces to imposing the following linear constraints:

$$-L \leq \frac{m_2 - m_1}{t_2 - t_1} \leq \frac{m_3 - m_2}{t_3 - t_2} \leq \dots \leq \frac{m_n - m_{n-1}}{t_n - t_{n-1}} \leq L. \quad (5.2)$$

In particular, the minimization problem at hand can be represented as

$$\text{minimize } |Q^{1/2}(m - y)|^2 \quad \text{subject to} \quad Am \geq b, \quad (5.3)$$

for  $A$  and  $b$  written so as to represent (5.2).

In the following we reduce the above optimization problem to a nonnegative least squares problem, which can then be solved efficiently using the `nnls` package in R. Define  $z := Q^{1/2}(m - y)$ , so that  $m = Q^{-1/2}z + y$ . Using this, we have  $Am \geq b$  if and only if  $AQ^{-1/2}z \geq b - Ay$ . Thus, (5.3) is equivalent to

$$\text{minimize } |z|^2 \quad \text{subject to } Gz \geq h, \quad (5.4)$$

where  $G := AQ^{-1/2}$  and  $h := b - Ay$ . An equivalent formulation is

$$\text{minimize } |Eu - \ell|, \quad \text{over } u \succeq 0, \quad \text{where } E := \begin{bmatrix} G^\top \\ h^\top \end{bmatrix} \quad \text{and } \ell := [0, \dots, 0, 1]^\top \in \mathbb{R}^{n+1}. \quad (5.5)$$

Here  $\succeq$  represents coordinate-wise inequality. A proof of this equivalence can be found in Lawson and Hanson [34, page 165]; see [8] for an algorithm to solve (5.5).

If  $\hat{u}$  denotes the solution of (5.5) then the solution of (5.4) is given as follows. Define  $r := E\hat{u} - \ell$ . Then  $\hat{z}$ , the minimizer of (5.4), is given by  $\hat{z} := (-r_1/r_{n+1}, \dots, -r_n/r_{n+1})^\top$ <sup>4</sup>. Hence the solution to (5.3) is given by  $\hat{y} = Q^{-1/2}\hat{z} + y$ .

#### 5.1.2. Penalized least squares (PLSE)

When  $\mathfrak{C}(m, \theta) = \mathcal{L}_n(m, \theta; \lambda)$ , we need to minimize the objective function

$$\frac{1}{n} \sum_{i=1}^n (y_i - m(t_i))^2 + \lambda^2 \int \{m''(t)\}^2 dt.$$

As in Section 5.1.2, consider the general objective function

$$(y - m)^\top Q(y - m) + \lambda^2 \int \{m''(t)\}^2 dt,$$

---

<sup>4</sup>Note that (5.4) is a Least Distance Programming (LDP) problem and Lawson and Hanson [34, page 167] prove that  $r_{n+1}$  cannot be zero in an LDP with a feasible constraint set.

to be minimized over  $\mathcal{R}$  and  $Q$  is any positive definite matrix. As in Section 5.1.1, we use  $m$  to denote the function  $t \mapsto m(t)$  as well as the vector  $(m(t_1), \dots, m(t_n))$  interchangeably. Theorem 1 of [16] 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  $M(t) := (M_1(t), M_2(t), \dots, M_{n-2}(t))$  and  $\{M_i(\cdot)\}_{1 \leq i \leq n-2}$  are the real-valued functions defined as

$$M_i(x) := \begin{cases} \frac{1}{t_{i+2}-t_i} \cdot \frac{x-t_i}{t_{i+1}-t_i} & \text{if } t_i \leq x < t_{i+1}, \\ \frac{1}{t_{i+2}-t_i} \cdot \frac{t_{i+1}-x}{t_{i+2}-t_{i+1}} & \text{if } t_{i+1} \leq x < t_{i+2}, \end{cases}$$

and  $\hat{\alpha}$  is a solution of the following equation:

$$[T(\alpha) + \lambda^2 K Q^{-1} K^\top] \alpha = Ky, \quad (5.6)$$

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  $\alpha$  as  $\alpha_i = (t_{i+2} - t_i)/4$  based on empirical evidence suggested by [16] and use (5.6) repeatedly until convergence. This algorithm was shown to have quadratic convergence in [13]. In the `simest` package, we implement the above algorithm in C.

**Remark 5.1** (Pre-binning). *The matrices involved in all these algorithms have entries depending on fractions such as  $1/(t_{i+1} - t_i)$ . Thus if there are ties in  $\{t_i\}_{1 \leq i \leq n}$ , then the matrix  $K$  is incomputable. Moreover, if  $t_{i+1} - t_i$  is very small, then the fractions can force the matrices involved to be ill-conditioned (for the purposes of numerical calculations). Thus to avoid ill-conditioning of these matrices, in practice one might have to 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 the means of all data points for which the  $t_i$ 's are close. To be more precise, if we choose a tolerance of  $\eta = 10^{-6}$  and suppose that  $0 < t_2 - t_1 < t_3 - t_1 < \eta$ , then we combine the data points  $(t_1, y_1), (t_2, y_2), (t_3, y_3)$  by taking their mean and set  $Q_{1,1} = 3$ ; the total number of data points is reduced to  $n - 2$ .*

## 5.2. Algorithm for computing $\theta^{(k+1)}$

In this subsection we describe the algorithm to find the minimizer  $\theta^{(k+1)}$  of  $\mathfrak{C}(m^{(k)}, \theta)$  over  $\theta \in \Theta$ . Recall that  $\Theta$  is defined to be the “positive” half of the unit sphere, 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 [46, Section 3.3]. But it is computationally expensive to move along a geodesic and so we follow the approach of [58] wherein we move along a retraction with the guarantee of descent. To explain the approach of [58], let us denote the objective function by  $f(\theta)$ , i.e., in our case  $f(\theta) = \mathfrak{C}(m^{(k)}, \theta)$ . Let  $\alpha \in \Theta$  be an initial guess for  $\theta^{(k+1)}$  and define

$$g := \nabla f(\alpha) \in \mathbb{R}^d \quad \text{and} \quad A := g\alpha^\top - \alpha g^\top,$$

where  $\nabla$  denotes the gradient operator. Next we choose the path  $\tau \mapsto \theta(\tau)$ , where

$$\theta(\tau) := \left(I + \frac{\tau}{2} A\right)^{-1} \left(I - \frac{\tau}{2} A\right) \alpha = \frac{1 + \frac{\tau^2}{4} [(\alpha^\top g)^2 - |g|^2] + \tau \alpha^\top g}{1 - \frac{\tau^2 (\alpha^\top g)^2}{4} + \frac{\tau^2 |g|^2}{4}} \alpha - \frac{\tau}{1 - \frac{\tau^2 (\alpha^\top g)^2}{4} + \frac{\tau^2 |g|^2}{4}} g,$$

for  $\tau$  in a small neighborhood of 0, and try to find a choice of  $\tau$  such that  $f(\theta(\tau))$  is as much smaller than  $f(\alpha)$  as possible; see step 2 of Algorithm 1. It is easy to verify that

$$\left. \frac{\partial f(\theta(\tau))}{\partial \tau} \right|_{\tau=0} \leq 0;$$

see Lemma 3 of [58]. This implies that  $\tau \mapsto f(\theta(\tau))$  is a nonincreasing function in a neighborhood of 0. Recall that for every  $\eta \in \Theta$ ,  $\eta_1$  (the first coordinate of  $\eta$ ) is nonnegative. For  $\theta(\tau)$  to lie in  $\Theta$ ,  $\tau$  has to satisfy the following inequality

$$\frac{\tau^2}{4}[(\alpha^\top g)^2 - |g|^2] + \tau \left( \alpha^\top g - \frac{g_1}{\alpha_1} \right) + 1 \geq 0, \quad (5.7)$$

where  $g_1$  and  $\alpha_1$  represent the first coordinates of the vectors  $g$  and  $\alpha$ , respectively. This implies that a valid choice of  $\tau$  must lie between the zeros of the quadratic expression on the left hand side of (5.7), 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  $\tau \mapsto f(\theta(\tau))$ , where  $\tau$  is in the above mentioned interval, to find  $\theta^{(k+1)}$ . We implement this step in the R package `simest`.

## 6. Simulation Study

In this section we illustrate the finite sample performance of the two estimators proposed in this paper; see (2.4) and (2.2). We also compare their performance with other existing estimators, namely, the **EFM** estimator (the estimating function method; see [11]), the **EDR** estimator (effective dimension reduction; see [25]), and the estimator proposed in [32] with the tuning parameter chosen by generalized cross-validation (see [55] and [32]; we denote this method by **SmoothGCV**). For the convex constrained estimators, we use **CvxLip** to denote the LLSE, and **CvxPen** to denote the PLSE (to compute **CvxPen** we take  $\hat{\lambda}_n = 0.1 \times n^{-2/5}$ ).

### 6.1. Another convex constrained estimator

Alongside these existing estimators, we also numerically study another natural estimator under the convexity shape constraint — the convex LSE — denoted by **CvxLSE** below. This estimator is obtained by minimizing the sum of squared errors subject to the convexity constraint. Formally, the **CvxLSE** is defined as

$$(m_n^\dagger, \theta_n^\dagger) := \arg \min_{(m, \theta) \in \mathcal{C} \times \Theta} Q_n(m, \theta). \quad (6.1)$$

The convexity constraint (i.e.,  $m \in \mathcal{C}$ ) can be represented by the following set of  $n - 2$  linear constraints:

$$\frac{m_2 - m_1}{t_2 - t_1} \leq \frac{m_3 - m_2}{t_3 - t_2} \leq \dots \leq \frac{m_n - m_{n-1}}{t_n - t_{n-1}},$$

where we use the notation of Section 5. Similar to the LLSE, this reduces the computation of  $m$  (for a given  $\theta$ ) to solving a quadratic program with linear inequalities; see Section 5.1.1. However, theoretically analyzing the performance of this estimator is difficult, because of various reasons; see Section 7 for a brief discussion. In our simulation studies we observe that the performance of **CvxLSE** is very similar to that of **CvxLip**.

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 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 in-sample root mean squared estimation error of  $(\tilde{m}, \tilde{\theta})$ , for all the estimators considered. From the simulation study it is easy to conclude that the proposed estimators have superior finite sample performance in all sampling scenarios considered.

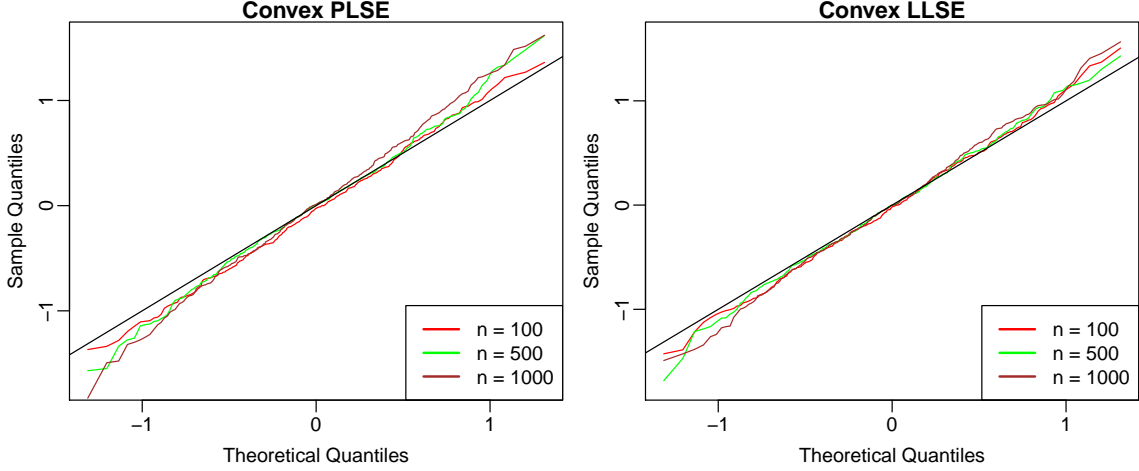


FIG 1. QQ-plots for  $\sqrt{n}(\tilde{\theta}_1 - \theta_{0,1})$  (over 800 replications) based on i.i.d. samples from (6.2) for  $n \in \{100, 500, 1000\}$ . The solid black line corresponds to the  $Y = X$  line. Left panel: *CvxPLSE*; right panel: *CvxLSE*.

## 6.2. Verifying the asymptotics

Theorems 4.1 and 4.2 show that (under homoscedastic error) both *CvxLip* and *CvxPen* are  $\sqrt{n}$ -consistent and asymptotically normal with the following covariance matrix:

$$\Sigma^0 := \sigma^4 H_{\theta_0} P_{\theta_0, m_0} [\ell_{\theta_0, m_0}(Y, X) \ell_{\theta_0, m_0}^\top(Y, X)]^{-1} H_{\theta_0}^\top,$$

where  $\ell_{\theta_0, m_0}(y, x) = (y - m_0(\theta_0^\top x)) m'_0(\theta_0^\top x) H_{\theta_0}^\top \{x - \mathbb{E}[X | \theta_0^\top X = \theta_0^\top x]\}$  and  $\sigma^2 = \mathbb{E}(\epsilon^2)$ . In this section, we give a simulation example that corroborates our theoretical results. We generate i.i.d. samples from the following model:

$$Y = (\theta_0^\top X)^2 + N(0, .3^2), \quad \text{where } X \sim \text{Uniform}[-1, 1]^3 \text{ and } \theta_0 = \mathbf{1}_3 / \sqrt{3}. \quad (6.2)$$

In Figure 1, on the  $y$ -axis we have the empirical quantiles of  $\sqrt{n}(\tilde{\theta}_1 - \theta_{0,1})$  and on the  $x$ -axis we have the theoretical quantiles of the Gaussian distribution with mean 0 and variance  $\Sigma_{1,1}^0$ . For the model (6.2), we computed  $\Sigma_{1,1}^0$  to be 0.22.

In Remark 4.4, we describe a simple plug-in procedure to create confidence sets for  $\theta_0$ ; see (4.15). In Table 1, we present empirical coverages (from 800 replications) of 95% confidence intervals based on *CvxLip* and *CvxPen* as the sample size increases from 50 to 2000.

TABLE 1

The estimated coverage probabilities and average lengths (obtained from 800 replicates) of nominal 95% confidence intervals for the first coordinate of  $\theta_0$  for the model described in Section 6.2.

$n$	CvxLip		CvxPen	
	Coverage	Avg Length	Coverage	Avg Length
50	0.92	0.30	0.94	0.29
100	0.91	0.18	0.92	0.19
200	0.92	0.13	0.93	0.13
500	0.94	0.08	0.92	0.08
1000	0.93	0.06	0.92	0.06
2000	0.92	0.04	0.93	0.04

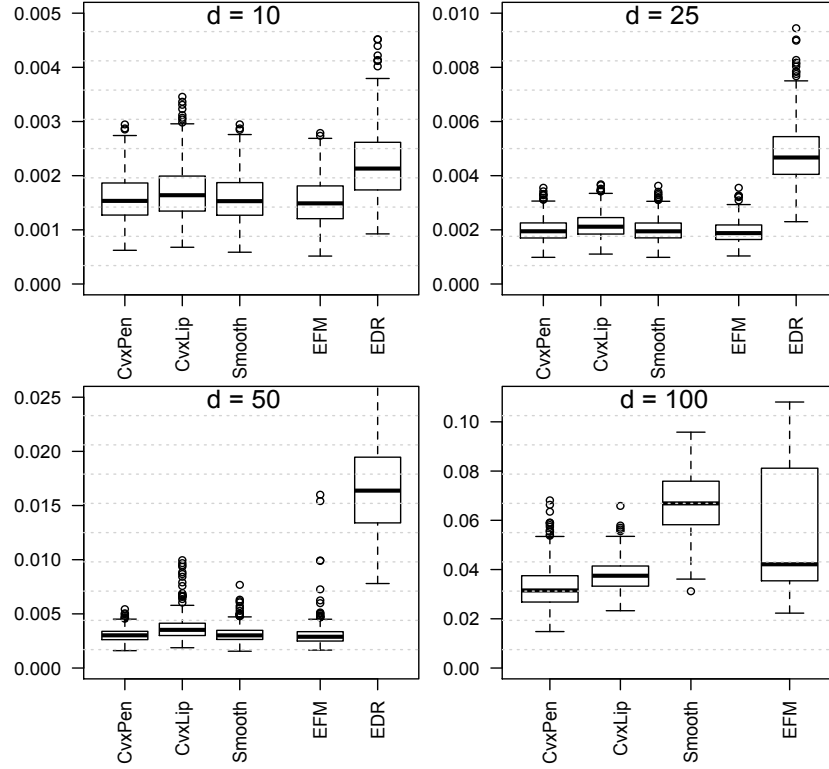


FIG 2. Boxplots of  $\sum_{i=1}^d |\hat{\theta}_i - \theta_{0,i}| / d$  (over 500 replications) based on 200 observations from Example 2 in Section 6.3 for dimensions 10, 25, 50, and 100, shown in the top-left, the top-right, the bottom-left, and the bottom-right panels, respectively. The bottom-right panel doesn't include EDR as the R-package EDR does not allow for  $d = 100$ .

### 6.3. Increasing dimension

To illustrate the behavior/performance of the estimators as  $d$  grows, we consider the following single index model:

$$Y = (\theta_0^\top X)^2 + N(0, .2^2), \text{ where } \theta_0 = (2, 1, \mathbf{0}_{d-2})^\top / \sqrt{5} \text{ and } X \in \mathbb{R}^d \sim \text{Uniform}[-1, 5]^d.$$

In each replication we observe  $n = 200$  i.i.d. samples from the model. It is easy to see that the performance of all the estimators worsens as the dimension increases from 10 to 100 and EDR has the worst overall performance; see Figure 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 [11].

### 6.4. Choice of $\lambda_n$ and $L$

In this subsection, we consider a simple simulation experiment to demonstrate that the finite sample performances of both the PLSE and LLSE are robust to the choice of tuning parameter. We generate an i.i.d. sample (of size  $n = 500$ ) from the following model:

$$Y = (\theta_0^\top X)^2 + N(0, .1^2), \quad \text{where } X \sim \text{Uniform}[-1, 1]^4 \text{ and } \theta_0 = \mathbf{1}_4 / 2. \quad (6.3)$$

Observe that for the above model, we have  $-2 \leq \theta^\top X \leq 2$  and  $L_0 := \sup_{t \in [-2, 2]} m'_0(t) = 4$  as  $m_0(t) = t^2$ . To compare the performances of the proposed estimators as their tuning parameter change, we vary  $\hat{\lambda}_n$  from

$\exp(-7/2) \times n^{-2/5}$  to  $n^{-2/5}$  and vary  $L$  from 3 ( $< L_0$ ) to 10. Figure 3 shows box plots of  $\frac{1}{d} \sum_{i=1}^d |\tilde{\theta}_i - \theta_{0,i}|$  for both **CvxPen** and **CvxLip** as their respective tuning parameter varies. The plots clearly show that the performance of both the estimators are not significantly affected by the particular choice of the tuning parameter. The observed robustness in the behavior of the estimators can be attributed to the stability endowed by the convexity constraint.

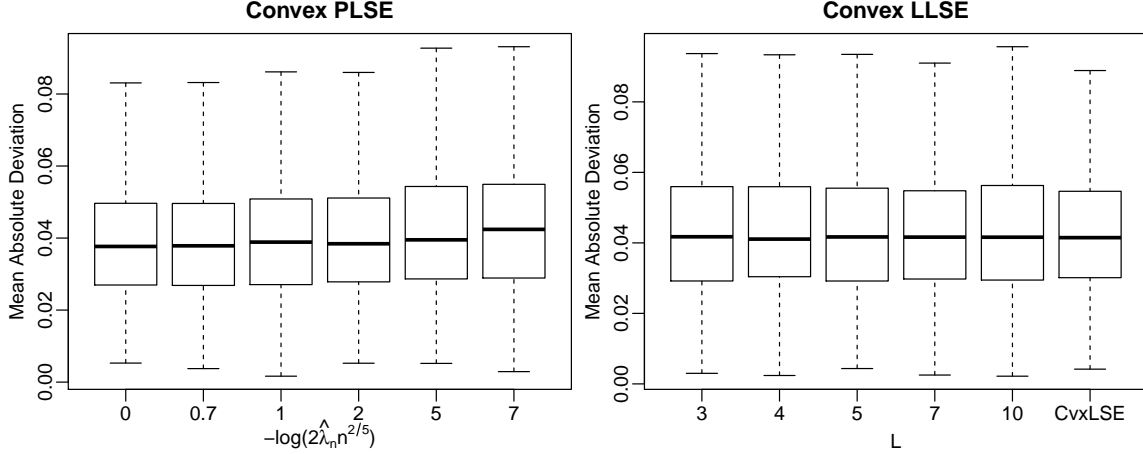


FIG 3. Box plots of  $\frac{1}{4} \sum_{i=1}^4 |\tilde{\theta}_i - \theta_{0,i}|$  (over 1000 replications) for the model (6.3) ( $d = 4$  and  $n = 500$ ) as the tuning parameter varies. Left panel: **CvxPen** when  $\hat{\lambda}_n = \exp(-T/2) \times n^{-2/5}$  for  $T = \{0, 0.7, 1, 2, 5, 7\}$ ; right panel: **CvxLip** for  $L = \{3, 4, 5, 7, 10\}$  and **CvxLSE**.

## 7. Discussion

In this paper we have proposed and studied two estimators in a convex single index model, namely the LLSE and the PLSE. Both the estimators are optimal — the function estimates are minimax rate optimal and the estimates of the index parameter are semiparametrically efficient. We have also introduced another natural estimator in this model, namely the convex LSE (see (6.1)), and have investigated its performance in our simulation studies. However, a thorough study of the theoretical properties of the convex LSE is difficult, and an open research problem. The difficulty can be attributed to the lack of our understanding of the behavior of  $m_n^\dagger$  and its right-derivative near the “boundary” of the covariate domain. In single index models inconsistency of  $m_n^\dagger$  at the boundary affects the estimation of  $\theta_0$ , as  $\theta_0$  and  $m_0$  are intertwined/bundled (as opposed to a partially linear model). Even in the simple univariate convex regression problem there are no existing upper bounds on the value of the LSE at the boundary. It is worth noting that in the recent paper [3] where the authors study the monotone single index model the unboundedness of LSE of the link function at the boundary turned out to be a major hurdle in deriving the asymptotic properties of the estimator (even though there exists closed form expressions for the LSE).

## Appendix A: Proof of Theorem 4.1

In this section we give a detailed discussion of **Step 1–Step 5** in the proof of Theorem 4.1.

### A.1. An approximately least favorable subprovided path [Step 1]

We now construct a path whose score for any  $(\theta, m) \in \Theta \times \{g \in \mathcal{R} | J(g) < \infty\}$  is  $\mathfrak{S}_{\theta,m}$ . Before proceeding further, for notational convenience, let us define

$$\mathcal{R}^* := \{g \in \mathcal{R} | J(g) < \infty\}$$

Recall (4.3). For any  $(\theta, m) \in \Theta \times \mathcal{R}^*$ , let  $t \mapsto (\zeta_t(\theta, \eta), \xi_t(\cdot; \theta, \eta, m))$  denote a path in  $\Theta \times \mathcal{R}^*$  through  $(\theta, m)$ , 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

$$\left. \frac{\partial}{\partial t} \mathcal{L}_n(\xi_t(\cdot; \hat{\theta}, \eta, \hat{m}), \zeta_t(\hat{\theta}, \eta); \hat{\lambda}_n) \right|_{t=0} = 0. \quad (\text{A.1})$$

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

$$\begin{aligned} \left. \frac{\partial}{\partial t} (y - \xi_t(\zeta_t(\hat{\theta}, \eta)^\top x; \hat{\theta}, \eta, \hat{m}))^2 \right|_{t=0} &= \eta^\top \mathfrak{S}_{\hat{\theta}, \hat{m}}(x, y), \\ \left. \frac{\partial}{\partial t} J^2(\xi_t(\cdot; \hat{\theta}, \eta, \hat{m})) \right|_{t=0} &= O_p(1), \quad \forall \eta \in S^{d-2}, \end{aligned} \quad (\text{A.2})$$

then we obtain (4.19) as  $\hat{\lambda}_n^2 = o_p(n^{-1/2})$ ; see assumption (P2).

Observe that  $\hat{\theta}$  is a consistent estimator of  $\theta_0$  and we are concerned with constructing the function  $t \mapsto \mathcal{L}_n(\xi_t(\cdot; \hat{\theta}, \eta, \hat{m}), \zeta_t(\hat{\theta}, \eta); \hat{\lambda}_n)$ , a path through  $(\hat{\theta}, \hat{m})$ . As we know that  $\hat{\theta}$  and  $\hat{m}$  are consistent estimators of  $\theta_0$  and  $m_0$ , it suffices to construct a similar path through any  $(\theta, m) \in \{\Theta \cap B_{\theta_0}(r)\} \times \mathcal{R}^*$  that satisfies the above requirements ( $r$  is as defined in (A5)). For any set  $A \subset \mathbb{R}$  and any  $\nu > 0$ , let us define  $A^\nu := \cup_{a \in A} B_a(\nu)$  and let  $\partial A$  denote the boundary of  $A$ . Fix  $\nu > 0$ . By assumption (A5), 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  $\phi_{\theta, \eta, t} : D^\nu \rightarrow \mathbb{R}$  with

$$\begin{aligned} \phi_{\theta, \eta, t}(u) &= u, \quad u \in D_\theta, \\ \phi_{\theta, \eta, t}(u + (\theta - \zeta_t(\theta, \eta))^\top k(u)) &= u, \quad u \in \partial D, \end{aligned} \quad (\text{A.3})$$

where  $\zeta_t(\theta, \eta)$  and  $k(u)$  are defined in (4.3) and (4.17), respectively. Furthermore, we can ensure that  $u \in D \mapsto \phi_{\theta, \eta, t}(u)$  is infinitely differentiable and that  $\left. \frac{\partial}{\partial t} \phi_{\theta, \eta, t} \right|_{t=0}$  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  $(\theta - \zeta_t(\theta, \eta))^\top k(u) \neq 0$  for some  $u \in \partial D$ . Let us now define

$$\mathfrak{T}_t(u; \theta, \eta, m) := m' \circ \phi_{\theta, \eta, t}(u + (\theta - \zeta_t(\theta, \eta))^\top k(u)).$$

Observe that  $t \mapsto \mathfrak{T}_t(u; \theta, \eta, m)$  is a path through  $m'$ . Thus we can integrate  $\mathfrak{T}_t(u; \theta, \eta, m)$  to construct a path through  $m$ . Let us define

$$\xi_t(u; \theta, \eta, m) := \int_{s_0}^u \mathfrak{T}_t(y; \theta, \eta, m) dy + (\zeta_t(\theta, \eta) - \theta)^\top \left[ (m'_0(s_0) - m'(s_0))k(s_0) - m'_0(s_0)h_{\theta_0}(s_0) \right] + m(\theta), \quad (\text{A.4})$$

where  $h_{\theta_0}$  is defined in (4.1),  $k$  is defined in (4.17), and  $s_0 \in \bigcap_{\theta \in B_{\theta_0}(r)} D_\theta$  where  $r$  satisfies assumption (A5). The function  $\phi_{\theta, \eta, t}$  helps us control the partial derivative in the second equation of (A.2). In the following theorem, proved in Appendix H.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 (A.1) and (A.2). Here  $\eta \in S^{d-2}$  is the “direction” for  $\zeta_t(\theta, \eta)$  and  $(\eta, k(u))$  defines the “direction” for the path  $\xi_t(\cdot; \theta, \eta, m)$ .

**Theorem A.1 (Step 1).** *Under the assumptions of Theorem 4.1,  $(\zeta_t(\hat{\theta}, \eta), \xi_t(\cdot; \hat{\theta}, \eta, \hat{m}))$  is a valid parametric submodel, i.e.,  $(\zeta_t(\hat{\theta}, \eta), \xi_t(\cdot; \hat{\theta}, \eta, \hat{m})) \in \Theta \times \mathcal{R}^*$  for all  $t$  in some neighborhood of 0. Moreover,  $(\zeta_t(\hat{\theta}, \eta), \xi_t(\cdot; \hat{\theta}, \eta, \hat{m}))$  satisfies (A.2),  $t \mapsto \mathcal{L}_n(\xi_t(\cdot; \hat{\theta}, \eta, \hat{m}), \zeta_t(\hat{\theta}, \eta); \hat{\lambda}_n)$  is differentiable at 0,  $\mathfrak{S}_{\hat{\theta}, \hat{m}}$  satisfies (4.19), there exists  $M^* < \infty$  which satisfies (4.18), and  $\mathfrak{S}_{\theta_0, m_0} = \ell_{\theta_0, m_0}$ .*

## A.2. A well-behaved approximation [Step 2]

We observe that  $\mathfrak{S}_{\theta, m}$  (the score for the approximately least favorable subprovided path) does not satisfy the conditions required by [51]. In this section we introduce  $\psi_{\theta, m}$ , a well behaved “approximation” of  $\mathfrak{S}_{\theta, m}$ . Note that  $\psi_{\theta, m}$  is not a score of (4.4) 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 (4.23)), (2)  $\psi_{\theta_0, m_0} = \ell_{\theta_0, m_0} = \mathfrak{S}_{\theta_0, m_0}$ , and (3)  $\psi_{\hat{\theta}, \hat{m}}$  converges to  $\psi_{\theta_0, m_0}$  in the  $L_2(P_{\theta_0, m_0})$  norm; see Lemma H.1. The following theorem proves that  $\mathfrak{S}_{\hat{\theta}, \hat{m}}$  and  $\psi_{\hat{\theta}, \hat{m}}$  are “approximately” the same.



**Theorem A.2 (Step 2).** Under the assumptions of Theorem 4.1, we have

$$\sqrt{n}\mathbb{P}_n(\mathfrak{S}_{\hat{\theta},\hat{m}} - \psi_{\hat{\theta},\hat{m}}) = o_p(1). \quad (\text{A.5})$$

We break the proof of this theorem into a number of lemmas proved in Appendix H. In the following lemma, proved in Appendix H.2, we find an upper bound for the left hand side of (A.5).

**Lemma A.1.** Under model (1.1), we have

$$\begin{aligned} |\sqrt{n}\mathbb{P}_n(\mathfrak{S}_{\hat{\theta},\hat{m}} - \psi_{\hat{\theta},\hat{m}})| &\leq |\mathbb{G}_n[(m_0 \circ \theta_0 - \hat{m} \circ \theta_0)U_{\hat{\theta},\hat{m}}]| + |\mathbb{G}_n[(\hat{m} \circ \theta_0 - \hat{m} \circ \hat{\theta})U_{\hat{\theta},\hat{m}}]| + |\sqrt{n}\mathbb{P}_n \epsilon U_{\hat{\theta},\hat{m}}| \\ &\quad + \sqrt{n}|P_{\theta_0,m_0}[(\hat{m} \circ \theta_0 - \hat{m} \circ \hat{\theta})U_{\hat{\theta},\hat{m}}]| + \sqrt{n}|P_{\theta_0,m_0}[(m_0 \circ \theta_0 - \hat{m} \circ \theta_0)U_{\hat{\theta},\hat{m}}]| \end{aligned} \quad (\text{A.6})$$

where  $U_{\theta,m} : \mathcal{X} \rightarrow \mathbb{R}^{d-1}$  is defined as

$$U_{\theta,m}(x) := H_{\theta}^{\top} \left[ \int_{s_0}^{\theta^{\top} x} [m'(u) - m'_0(u)]k'(u)du + (m'_0(\theta^{\top} x) - m'(\theta^{\top} x))k(\theta^{\top} x) \right]. \quad (\text{A.7})$$

Note that the proof of Theorem A.2 will be complete if we show that each of the terms on the right hand side of (A.6) converges to 0 in probability. We begin with some definitions. Let  $a_n$  be a sequence of real numbers such that  $a_n \rightarrow \infty$  as  $n \rightarrow \infty$  and  $a_n \|\hat{m} - m_0\|_{D_0}^S = o_p(1)$ . Note that we can always find such a sequence  $a_n$ , as by Theorem 3.6 we have  $\|\hat{m} - m_0\|_{D_0}^S = o_p(1)$ . For all  $n \in \mathbb{N}$ , define<sup>5</sup>

$$\begin{aligned} \mathcal{C}_{M_1,M_2,M_3}^{m*} &:= \left\{ m \in \mathcal{R} : \|m\|_{\infty} \leq M_1, \|m'\|_{\infty} \leq M_2, \text{ and } J(m) \leq M_3 \right\}, \\ \mathcal{C}_{M_1,M_2,M_3}^m(n) &:= \left\{ m \in \mathcal{C}_{M_1,M_2,M_3}^{m*} : a_n \|m - m_0\|_{D_0}^S \leq 1 \right\}, \\ \mathcal{C}_{M_1,M_2,M_3}^* &:= \left\{ (\theta, m) : \theta \in \Theta \cap B_{\theta_0}(1/2) \text{ and } m \in \mathcal{C}_{M_1,M_2,M_3}^{m*} \right\}, \\ \mathcal{C}^{\theta}(n) &:= \left\{ \theta \in \Theta \cap B_{\theta_0}(1/2) : \hat{\lambda}_n^{-1/2} |\theta - \theta_0| \leq 1 \right\}, \\ \mathcal{C}_{M_1,M_2,M_3}(n) &:= \left\{ (\theta, m) : \theta \in \mathcal{C}^{\theta}(n) \text{ and } m \in \mathcal{C}_{M_1,M_2,M_3}^m(n) \right\}, \\ \mathcal{W}_{M_1,M_2,M_3}^* &:= \{ U_{\theta,m} : (\theta, m) \in \mathcal{C}_{M_1,M_2,M_3}^* \}, \\ \mathcal{W}_{M_1,M_2,M_3}(n) &:= \{ U_{\theta,m} : (\theta, m) \in \mathcal{C}_{M_1,M_2,M_3}(n) \}. \end{aligned} \quad (\text{A.8})$$

As a first step in proving that each term on the right hand side of (A.6) converges to 0, we analyze 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 in Appendix H.3, we find the bracketing numbers and envelope functions for the classes. The result will be used in some of the remaining proofs.

**Lemma A.2.** 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

$$\sup_{(\theta,m) \in \mathcal{C}_{M_1,M_2,M_3}(n)} \|U_{\theta,m}\|_{2,\infty} \leq W_{M_1,M_2,M_3}(n) := M^* \sqrt{d-1} \left( 2(M_3 + M_2)T\hat{\lambda}_n^{1/4} + (T+1)\frac{1}{a_n} \right), \quad (\text{A.9})$$

where  $M^*$  is defined in (4.18) and  $\|\cdot\|_{2,\infty}$  is defined in Section 2.1. 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  $\mathcal{W}_{M_1,M_2,M_3}(n)$ :

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

see Section 2.1.1 of [53] for a definition of  $N_{[]}(\cdot, \cdot, \cdot)$ .

The study of limiting behaviors of the first three terms on the right hand side of (A.6) are similar. For every fixed  $M_1, M_2$ , and  $M_3$ , the first term in the right hand side of (A.6) can be bounded above as

$$\begin{aligned} &\mathbb{P}\left(|\mathbb{G}_n[(m_0 \circ \theta_0 - \hat{m} \circ \theta_0)U_{\hat{\theta},\hat{m}}]| > \delta\right) \\ &\leq \mathbb{P}\left(|\mathbb{G}_n[(m_0 \circ \theta_0 - \hat{m} \circ \theta_0)U_{\hat{\theta},\hat{m}}]| > \delta, (\hat{\theta}, \hat{m}) \in \mathcal{C}_{M_1,M_2,M_3}(n)\right) + \mathbb{P}((\hat{\theta}, \hat{m}) \notin \mathcal{C}_{M_1,M_2,M_3}(n)) \\ &\leq \mathbb{P}\left(\sup_{(\theta,m) \in \mathcal{C}_{M_1,M_2,M_3}(n)} |\mathbb{G}_n[(m_0 \circ \theta_0 - m \circ \theta_0)U_{\theta,m}]| > \delta\right) + \mathbb{P}((\hat{\theta}, \hat{m}) \notin \mathcal{C}_{M_1,M_2,M_3}(n)). \end{aligned}$$

<sup>5</sup>The notations with \* denote the classes that do not depend on  $n$  while the ones with  $n$  denote shrinking neighborhoods around the truth.

By Theorem 3.6 we have that  $\hat{\theta}$  and  $\hat{m}$  are consistent for  $\theta_0$  and  $m_0$  in the Euclidean and Sobolev norms, respectively and  $\|\hat{m}'\|_\infty$  is  $O_p(1)$ . Furthermore by Theorem 3.5, we have that both  $\|\hat{m}\|_\infty$  and  $J(\hat{m})$  are  $O_p(1)$  and by Theorem 3.7 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  $\varepsilon$ ) such that

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

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

$$\mathbb{P}\left(\sup_{(\theta, m) \in \mathcal{C}_{M_1, M_2, M_3}(n)} |\mathbb{G}_n([m_0 \circ \theta_0 - m \circ \theta_0]U_{\theta, m})| > \delta\right) \leq \varepsilon$$

for sufficiently large  $n$ . We show this in the following lemma (proved in Appendix H.4).

**Lemma A.3.** Fix  $M_1, M_2, M_3$ , and  $\delta > 0$ . For  $n \in \mathbb{N}$ , let us define

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

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

$$\sup_{f \in \mathcal{D}_{M_1, M_2, M_3}(n)} \|f\|_{2, \infty} \leq D_{M_1, M_2, M_3}(n) := 2M_1 W_{M_1, M_2, M_3}(n). \quad (\text{A.10})$$

Moreover,  $J_{[\cdot]}(\delta, \mathcal{D}_{M_1, M_2, M_3}(n), \|\cdot\|_{2, P_{\theta_0, m_0}}) \lesssim \delta^{1/2}$ , where for any class of functions  $\mathcal{F}$ ,  $J_{[\cdot]}$  (the entropy integral) is defined as

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

e.g., see [52]. Hence, we have  $\mathbb{P}(\sup_{f \in \mathcal{D}_{M_1, M_2, M_3}(n)} |\mathbb{G}_n f| > \delta) \rightarrow 0$  as  $n \rightarrow \infty$ .

The following two lemmas, proved in Appendices H.5 and H.6, complete the proof of Theorem A.2 and show that the last four terms on right side of (A.6) converge to zero in probability.

**Lemma A.4.** Fix  $M_1, M_2, M_3$ , and  $\delta > 0$ . For  $n \in \mathbb{N}$ , let us define

$$\begin{aligned} \mathcal{A}_{M_1, M_2, M_3}(n) &:= \{[m \circ \theta_0 - m \circ \theta]U_{\theta, m} : (\theta, m) \in \mathcal{C}_{M_1, M_2, M_3}(n)\}, \\ \mathcal{A}_{M_1, M_2, M_3}^* &:= \{[m \circ \theta_0 - m \circ \theta]U_{\theta, m} : (\theta, m) \in \mathcal{C}_{M_1, M_2, M_3}^*\}. \end{aligned}$$

Then  $\mathcal{A}_{M_1, M_2, M_3}(n)$  is a 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,  $J_{[\cdot]}(\delta, \mathcal{A}_{M_1, M_2, M_3}(n), \|\cdot\|_{2, P_{\theta_0, m_0}}) \lesssim \delta^{1/2}$ , and, as  $n \rightarrow \infty$ , we have

$$\mathbb{P}\left(\left|\mathbb{G}_n[(\hat{m} \circ \theta_0 - \hat{m} \circ \hat{\theta})U_{\hat{\theta}, \hat{m}}]\right| > \delta\right) \rightarrow 0.$$

**Lemma A.5.** If assumptions (A0)–(A4), (B1)–(B3), and (P1)–(P2) hold, then

$$\begin{aligned} |\sqrt{n}\mathbb{P}_n[\epsilon U_{\hat{\theta}, \hat{m}}]| &= o_p(1), \\ \sqrt{n}|P_{\theta_0, m_0}[(m_0 \circ \theta_0 - \hat{m} \circ \theta_0)U_{\hat{\theta}, \hat{m}}]| &= o_p(1), \\ \sqrt{n}|P_{\theta_0, m_0}[(\hat{m} \circ \theta_0 - \hat{m} \circ \hat{\theta})U_{\hat{\theta}, \hat{m}}]| &= o_p(1). \end{aligned} \quad (\text{A.11})$$

Now that we have shown  $(\hat{\theta}, \hat{m})$  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  $\hat{\theta}$  follows from the theory developed in Section 6.6 of [51]. In the next theorem (proved in H.7), we prove that  $\psi_{\hat{\theta}, \hat{m}}$  satisfies the “no-bias” condition; see (6.6) of [51] and Section 3 of [40].

**Theorem A.3 (Step 3).** Under assumptions (A0)–(A4) and (B2),

$$\sqrt{n}P_{\hat{\theta}, m_0}\psi_{\hat{\theta}, \hat{m}} = o_p(1),$$

The following theorem (proved in Appendix H.9) completes the proof of Theorem 4.1.

**Theorem A.4 (Step 4).** *Under assumptions (A0)–(A4) and (B2), we have*

$$\mathbb{G}_n(\psi_{\check{\theta}, \check{m}} - \psi_{\theta_0, m_0}) = o_p(1). \quad (\text{A.12})$$

The proof of the above theorem is similar to that of Theorem A.2. We first find an upper bound for the left side of (A.12) and then show that each of the terms converge to zero; see Lemmas H.2 and H.3 in Appendix H.9.

## Appendix B: Proof of Theorem 4.2

The following theorem (proved in Appendix I.1) shows that submodel defined in (A.4) is an approximately least favorable subprovided submodel for model (1.1). The proof of Theorem B.1 is more complicated (when compared to that of the proof of Theorem A.1) as  $\check{m}$  is not differentiable everywhere.

**Theorem B.1 (Step 1).** *Under assumptions of Theorem 4.2,  $(\zeta_t(\check{\theta}, \eta), \xi_t(\cdot; \check{\theta}, \eta, \check{m}))$  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 (4.16) for definition of  $\mathfrak{S}_{\theta_0, m_0}$ . Moreover, we have that  $t \mapsto Q_n(\xi_t(\cdot; \check{\theta}, \eta, \check{m}), \zeta_t(\check{\theta}, \eta))$  is differentiable at 0,*

$$\left. \frac{\partial}{\partial t} (y - \xi_t(\zeta_t(\check{\theta}, \eta)^\top x; \check{\theta}, \eta, \check{m}))^2 \right|_{t=0} = \eta^\top \mathfrak{S}_{\check{\theta}, \check{m}}(x, y),$$

and

$$\left. \frac{\partial}{\partial t} Q_n(\xi_t(\cdot; \check{\theta}, \eta, \check{m}), \zeta_t(\check{\theta}, \eta)) \right|_{t=0} = \eta^\top \mathbb{P}_n \mathfrak{S}_{\check{\theta}, \check{m}} = 0.$$

### B.1. A well-behaved approximation [Step 2]

As in Appendix A.2, the following theorem (proved in a series of results) shows that  $\mathfrak{S}_{\check{\theta}, \check{m}}$  is empirically well-approximated by  $\psi_{\check{\theta}, \check{m}}$  (defined in (4.20)).

**Theorem B.2 (Step 2).** *Under assumptions of Theorem 4.2, 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 B.2 is very similar to the proof of Theorem A.2. As the definitions of  $\mathfrak{S}_{\theta, m}$  and  $\psi_{\theta, m}$  have not changed, Lemma A.1 clearly holds with  $(\check{\theta}, \check{m})$  instead of  $(\hat{\theta}, \hat{m})$ . Note that the proof of Theorem A.2 will be complete if we show that each of the terms on the right hand side of (A.6) converges to 0 in probability. We begin with some definitions. Let  $b_n$  be a sequence of real numbers such that  $b_n \rightarrow \infty$  as  $n \rightarrow \infty$ ,  $b_n = o(n^{1/2})$ , 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 3.2 we have  $\|\check{m} - m_0\|_{D_0} = o_p(1)$ . For all  $n \in \mathbb{N}$ , define<sup>6</sup>

$$\begin{aligned} \mathcal{C}_{M_1}^{m*} &:= \left\{ m \in \mathcal{M}_L : \|m\|_\infty \leq M_1 \right\}, \\ \mathcal{C}_{M_1}^m(n) &:= \left\{ m \in \mathcal{C}_{M_1}^{m*} : n^{1/5} \int_{D_0} (m'(t) - m'_0(t))^2 dt \leq 1, b_n \|m - m_0\|_{D_0} \leq 1 \right\}, \\ \mathcal{C}_{M_1}^* &:= \left\{ (\theta, m) : \theta \in \Theta \cap B_{\theta_0}(1/2) \text{ and } m \in \mathcal{C}_{M_1}^{m*} \right\}, \\ \mathcal{C}^\theta(n) &:= \left\{ \theta \in \Theta \cap B_{\theta_0}(1/2) : n^{1/10} |\theta - \theta_0| \leq 1 \right\}, \\ \mathcal{C}_{M_1}(n) &:= \left\{ (\theta, m) : \theta \in \mathcal{C}^\theta(n) \text{ and } m \in \mathcal{C}_{M_1}^m(n) \right\}, \\ \mathcal{W}_{M_1}^* &:= \{ U_{\theta, m} : (\theta, m) \in \mathcal{C}_{M_1}^* \}, \\ \mathcal{W}_{M_1}(n) &:= \{ U_{\theta, m} : (\theta, m) \in \mathcal{C}_{M_1}(n) \}, \end{aligned}$$

<sup>6</sup>As in (A.8), the notations with  $*$  denote the classes that do not depend on  $n$  while the ones with  $n$  denote shrinking neighborhoods around the truth.

where  $U_{\theta,m}(\cdot)$  is defined in (A.7). As a first step in proving that each term on the right hand side of (A.6) converges to 0, we study the properties of the classes of functions  $\mathcal{W}_{M_1}(n)$  and  $\mathcal{W}_{M_1}^*$ . In the following lemma, proved in Appendix I.2, we find the bracketing numbers and envelope functions for these two classes of functions. This will be used to prove the results that follow.

**Lemma B.1.** Fix  $M_1$ , and  $\delta > 0$ . Then  $\mathcal{W}_{M_1}(n)$  is a Donsker class and there exists a  $V^* < \infty$  such that  $\sup_{f \in \mathcal{W}_{M_1}^*} \|f\|_{2,\infty} \leq V^*$ . Moreover, for some  $c$  depending only on  $M_1$ , we have

$$N_{[]}(\varepsilon, \mathcal{W}_{M_1}(n), \|\cdot\|_{2,P_{\theta_0,m_0}}) \leq N_{[]}(\varepsilon, \mathcal{W}_{M_1}^*, \|\cdot\|_{2,P_{\theta_0,m_0}}) \leq c \exp(c/\varepsilon) \varepsilon^{-2d} \quad (\text{B.1})$$

and

$$\sup_{f \in \mathcal{W}_{M_1}(n)} \|f\|_{2,P_{\theta_0,m_0}}^2 \leq K_L^2 n^{-1/5}, \quad (\text{B.2})$$

where  $K_L^2 = 2\|k'\|_\infty^2 + L^2\|k'\|_\infty^2 T^2$  and  $k(\cdot)$  is defined in (4.17).

The study of limiting behaviors of the first three terms on the right hand side of (A.6) (with  $(\hat{\theta}, \hat{m})$  replaced by  $(\check{\theta}, \check{m})$ ) are similar. For every fixed  $M_1 > 0$  the first term in the right hand side of (A.6) can be bounded from above as

$$\begin{aligned} & \mathbb{P}\left(|\mathbb{G}_n([m_0 \circ \theta_0 - \check{m} \circ \theta_0]U_{\check{\theta},\check{m}})| > \delta\right) \\ & \leq \mathbb{P}\left(\sup_{(\theta,m) \in \mathcal{C}_{M_1}(n)} |\mathbb{G}_n([m_0 \circ \theta_0 - m \circ \theta_0]U_{\theta,m})| > \delta\right) + \mathbb{P}((\check{\theta}, \check{m}) \notin \mathcal{C}_{M_1}(n)), \end{aligned} \quad (\text{B.3})$$

where  $U_{\check{\theta},\check{m}} : \mathcal{X} \mapsto \mathbb{R}^{d-1}$  is defined in (A.7). By Theorem 3.2 we have that  $\check{\theta}$  and  $\check{m}$  are consistent in the Euclidean and supremum norms, respectively. Furthermore, by Theorems 3.3 and 3.4, we have that  $n^{1/10}|\check{\theta} - \theta_0| = o_p(1)$  and  $n^{1/5} \int_{D_0} |\check{m}'(t) - m_0'(t)|^2 dt = o_p(1)$ , respectively. Thus, it is easy to see that, for any  $\varepsilon > 0$ , there exists  $M_1$  (depending on  $\varepsilon$ ) 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  $M_1 > 0$  we have the first term on the right hand side of (B.3) is smaller than  $\varepsilon$  for sufficiently large  $n$ . We prove this in Lemma B.2.

**Lemma B.2.** Fix  $M_1$ , and  $\delta > 0$ . For  $n \in \mathbb{N}$ , let us define

$$\begin{aligned} \mathcal{D}_{M_1}^* &:= \{[m_0 \circ \theta_0 - m \circ \theta_0]U_{\theta,m} : (\theta, m) \in \mathcal{C}_{M_1}^*\}, \\ \mathcal{D}_{M_1}(n) &:= \{[m_0 \circ \theta_0 - m \circ \theta_0]U_{\theta,m} : (\theta, m) \in \mathcal{C}_{M_1}(n)\}. \end{aligned}$$

Then  $\mathcal{D}_{M_1}(n)$  is a Donsker class such that

$$\sup_{f \in \mathcal{D}_{M_1}(n)} \|f\|_{2,P_{\theta_0,m_0}}^2 \leq D_{M_1}^2 n^{-1/5},$$

where  $D_{M_1} := 2M_1 K_L$ . Moreover  $J_{[]}(\delta, \mathcal{D}_{M_1}(n), \|\cdot\|_{2,P_{\theta_0,m_0}}) \lesssim \delta^{1/2}$  and

$$\mathbb{P}\left(\sup_{f \in \mathcal{D}_{M_1}(n)} |\mathbb{G}_n f| > \delta\right) \rightarrow 0, \quad n \rightarrow \infty.$$

The following two lemmas, proved in the Appendices I.5 and I.6, complete the proof of Theorem B.2.

**Lemma B.3.** Fix  $M_1$ , and  $\delta > 0$ . For  $n \in \mathbb{N}$ , let us define

$$\begin{aligned} \mathcal{A}_{M_1}(n) &:= \{[m \circ \theta_0 - m \circ \theta]U_{\theta,m} : (\theta, m) \in \mathcal{C}_{M_1}(n)\}, \\ \mathcal{A}_{M_1}^* &:= \{[m \circ \theta_0 - m \circ \theta]U_{\theta,m} : (\theta, m) \in \mathcal{C}_{M_1}^*\}. \end{aligned}$$

Then  $\mathcal{A}_{M_1}(n)$  is Donsker class and  $D_{M_1} n^{-1/10}$  is an envelope function with respect to the  $\|\cdot\|_{2,P_{\theta_0,m_0}}$ . Moreover,  $J_{[]}(\delta, \mathcal{A}_{M_1}(n), \|\cdot\|_{2,P_{\theta_0,m_0}}) \lesssim \delta^{1/2}$ , and, as  $n \rightarrow \infty$ , we have

$$\mathbb{P}\left(|\mathbb{G}_n[(\check{m} \circ \theta_0 - \check{m} \circ \check{\theta})U_{\check{\theta},\check{m}}]| > \delta\right) \rightarrow 0.$$

**Lemma B.4.** If (A0)–(A4), (B1)–(B3), and (L1) hold, then

$$\begin{aligned} |\sqrt{n}\mathbb{P}_n[\epsilon U_{\tilde{\theta}, \tilde{m}}]| &= o_p(1), \\ \sqrt{n}|P_{\theta_0, m_0}[(m_0 \circ \theta_0 - \tilde{m} \circ \theta_0)U_{\tilde{\theta}, \tilde{m}}]| &= o_p(1), \\ \sqrt{n}|P_{\theta_0, m_0}[(\tilde{m} \circ \theta_0 - \tilde{m} \circ \tilde{\theta})U_{\tilde{\theta}, \tilde{m}}]| &= o_p(1). \end{aligned} \quad (\text{B.4})$$

Now that we have shown  $(\tilde{\theta}, \tilde{m})$  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  $\tilde{\theta}$  now follows from the theory developed in Section 6.6 of [51]. In the next theorem, we prove that  $\psi_{\tilde{\theta}, \tilde{m}}$  satisfies the “no-bias” (see equation 6.6 of [51]) condition.

**Theorem B.3 (Step 3).** Under assumptions of Theorem 4.2,  $\sqrt{n}P_{\tilde{\theta}, m_0}\psi_{\tilde{\theta}, \tilde{m}} = o_p(1)$ .

In Lemma I.2, stated and proved in Appendix I.8, we prove that  $\psi_{\tilde{\theta}, \tilde{m}}$  is a consistent estimator of  $\psi_{\theta_0, m_0}$  under  $L_2(P_{\theta_0, m_0})$  norm. The following theorem (proved in Appendix I.9) completes the proof of Theorem 4.2.

**Theorem B.4 (Step 4).** Under (A0)–(A4) and (B2), we have

$$\mathbb{G}_n(\psi_{\tilde{\theta}, \tilde{m}} - \psi_{\theta_0, m_0}) = o_p(1). \quad (\text{B.5})$$

The proof of the above theorem is similar to that of Theorem B.2. We first find an upper bound for the left side of (B.5) and then show that each of the terms converge to zero; see Lemmas I.3 and I.4 in Appendix I.9.

## Appendix C: Additional Simulation Studies

### C.1. A simple model

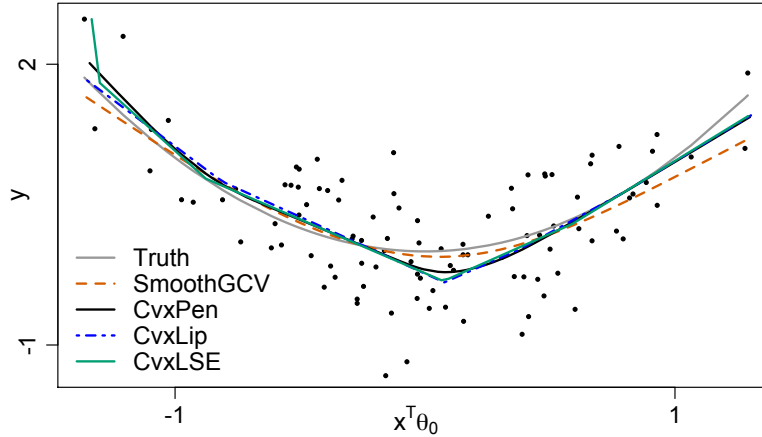


FIG 4. Function estimates for the model  $Y = (\theta_0^\top X)^2 + N(0, 1)$ , where  $\theta_0 = \mathbf{1}_5/\sqrt{5}$ ,  $X \sim \text{Uniform}[-1, 1]^5$ , and  $n = 100$ .

In this section we give a simple illustrative (finite sample) example. We observe 100 i.i.d. observations from the following homoscedastic model:

$$Y = (\theta_0^\top X)^2 + N(0, 1), \text{ where } \theta_0 = \mathbf{1}_5/\sqrt{5} \text{ and } X \sim \text{Uniform}[-1, 1]^5. \quad (\text{C.1})$$

In Figure 4, we have a scatter plot of  $\{(\theta_0^\top x_i, y_i)\}_{1 \leq i \leq 100}$  overlaid with prediction curves  $\{(\tilde{\theta}^\top x_i, \tilde{m}(\tilde{\theta}^\top x_i))\}_{1 \leq i \leq 100}$  for the proposed estimators obtained from one sample from (C.1). Table 2 displays all the corresponding estimates of  $\theta_0$  obtained from the same data set. To compute the function estimates for EFM and EDR approaches we used cross-validated smoothing splines to estimate the link function using their estimates of  $\theta_0$ .

TABLE 2  
*Estimates of  $\theta_0$ , “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 “Est Error”:=  $\|\tilde{m} \circ \tilde{\theta} - m_0 \circ \theta_0\|_n$  for one sample from (C.1).*

Method	$\theta_1$	$\theta_2$	$\theta_3$	$\theta_4$	$\theta_5$	Theta Error	Func Error	Est 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

### C.2. Piecewise linear function and dependent covariates

To understand the performance of the estimators when the truth is convex but not smooth, we consider the following model:

$$Y = |\theta_0^\top X| + N(0, .1^2), \quad (\text{C.2})$$

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  $(Z_1, Z_2) \sim \text{Uniform}[-1, 1]^2$  is independent of  $(X_1, X_2)$  and  $\theta_0$  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 [37]. The performances of the estimators are summarized in Figure 5. Observe that as the truth is not smooth, the convex constrained least squares estimators (CvxLip and CvxLSE) have slightly improved performance compared to the (roughness) penalized least squares estimators (CvxPen and SmoothGCV). Also observe that both EFM and EDR fail to estimate the true parameter  $\theta_0$ .

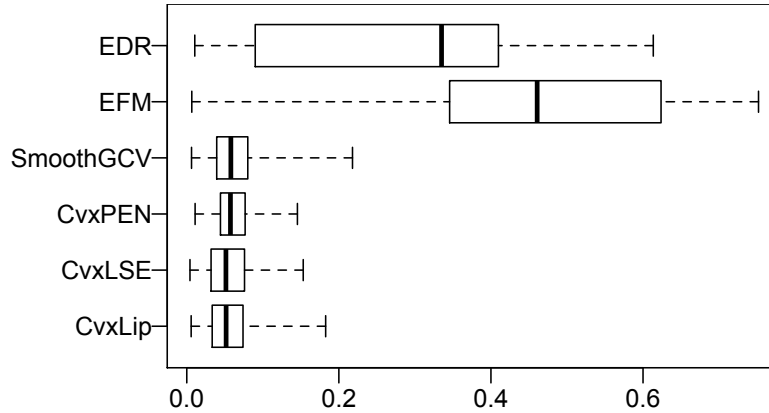


FIG 5. Box plots of  $\sum_{i=1}^6 |\tilde{\theta}_i - \theta_{0,i}|$  for the model (C.2). Here  $d = 6$ ,  $n = 200$  and we have 500 replications.

## Appendix D: Real data analysis

In this following we analyze two real datasets and apply the developed methodology for prediction and estimation.

### D.1. Boston housing data

The Boston housing dataset was collected by [22] to study the effect of different covariates on the real estate price in the greater Boston area. The dependent variable  $Y$  is the median value of owner occupied homes in

each of the 506 census tracts in Boston standard metropolitan statistical areas. [22] observed 13 covariates and fit a linear model after taking log transformation for 3 covariates and power transformations for three other covariates; also see [57] for a discussion of this dataset.

Breiman and Friedman [5] did further analysis to deal with multi-collinearity of the covariates and selected four variables using a penalized stepwise method. The chosen covariates were: average number of rooms per dwelling (RM), full-value property-tax rate per 10,000 USD (TAX), pupil-teacher ratio by town school district (PT), and proportion of population that is of “lower status” in percentage points (LS). As in [56] and [60], we take logarithms of LS and TAX to reduce sparse areas in the dataset. Furthermore, we have scaled and centered each of the covariates to have mean 0 and variance 1. [56] fit a nonparametric additive regression model to the selected variables and obtained an  $R^2$  (the coefficient of determination) of 0.64. [57] fit a single index model to this data using the set of covariates suggested in [7] and obtained a decreasing and approximately convex link function; see Fig. 2 of [57]. Moreover, we think it is logical that the “law of diminishing returns” should apply to the median home prices in this dataset. This motivates us to fit a convex single index model to the Boston housing dataset. Moreover, the convexity constraint adds interpretability to the estimators of both  $\theta_0$  and  $m_0$ . We summarize our results in Table 3. In Figure 6, we plot the scatter plot of  $\{(\hat{\theta}^\top x_i, y_i)\}_{i=1}^{506}$  (recall that  $\hat{\theta}$  denotes the **CvxPen**) overlaid with the plot of  $\tilde{m}(\tilde{\theta}^\top x)$ , for the **SmoothGCV**, **CvxPen**, **CvxLip**, and **CvxLSE**. We also observe that the  $R^2$  for the convexity constrained single index models (**CvxPen**, **CvxLip**, and **CvxLSE**), when using all the available covariates, was approximately 0.79. Inclusion of all the extra variables leads to only a minor increase in  $R^2$  at the cost of interpretability; [57] also reached to a similar conclusion.

TABLE 3  
Estimates of  $\theta_0$  and generalized  $R^2$  for the datasets in Appendices D.1 and D.2.

Method	Boston Data					Car mileage data				
	RM	log(TAX)	PT	log(LS)	$R^2$	D	W	A	H	$R^2$
<b>LM</b> <sup>7</sup>	2.34	-0.37	-1.55	-5.11	0.73	-0.63	-4.49	-0.06	-1.68	0.71
<b>SmoothGCV</b>	0.44	-0.18	-0.27	-0.83	0.77	0.42	0.18	0.11	0.88	0.76
<b>CvxPen</b>	0.48	-0.19	-0.25	-0.82	0.77	0.45	0.15	0.13	0.87	0.76
<b>CvxLip</b>	0.44	-0.14	-0.18	-0.87	0.77	0.44	0.18	0.12	0.87	0.76
<b>CvxLSE</b>	0.44	-0.14	-0.18	-0.87	0.79	0.39	0.14	0.12	0.90	0.77
<b>EFM</b>	0.48	-0.19	-0.21	-0.83	—	0.44	0.18	0.13	0.87	—
<b>EDR</b>	0.44	-0.14	-0.18	-0.87	—	0.33	0.11	0.15	0.93	—

## D.2. Car mileage data

As a second application for the convex single index model, we consider a dataset containing mileages of different cars; which can be found at <http://lib.stat.cmu.edu/datasets/cars.data>. We model the mileages ( $Y$ ) of 392 cars using the covariates ( $X$ ): displacement (D), weight (W), acceleration (A), and horsepower (H). Cheng et al. [10] fit a partial linear model to this dataset, while [32] fit single index model (without any shape constraint).

As in Appendix D.1, we have scaled and centered each of covariates to have mean 0 and variance 1 for our data analysis. It is easy to argue that, as in the previous dataset, “law of diminishing returns” applies to the mileages of cars. The right panel of Figure 6 illustrates this. All index coefficients are positive and the estimates of  $m_0$  are decreasing. We performed a test of significance for  $\theta_0$  using the plug-in variance estimate in Remark 4.4. The covariates acceleration, engine displacement, and power output were found to be significant and each of them had  $p$ -value less than  $10^{-5}$  (for both the PLSE and LLSE). In the right panel of Figure 6, we have the scatter plot of  $\{(\hat{\theta}^\top x_i, y_i)\}_{i=1}^{392}$  overlaid with the plot of  $\tilde{m}(\tilde{\theta}^\top x)$ , for the **SmoothGCV**, **CvxPen**, **CvxLip**, and **CvxLSE**. Table 3 lists different estimators for  $\theta_0$  and their respective  $R^2$ .

## Appendix E: Proof of results in Section 2

The following two lemmas of [32] will be used to prove results of Section 2.

<sup>7</sup>LM denotes the linear regression model.



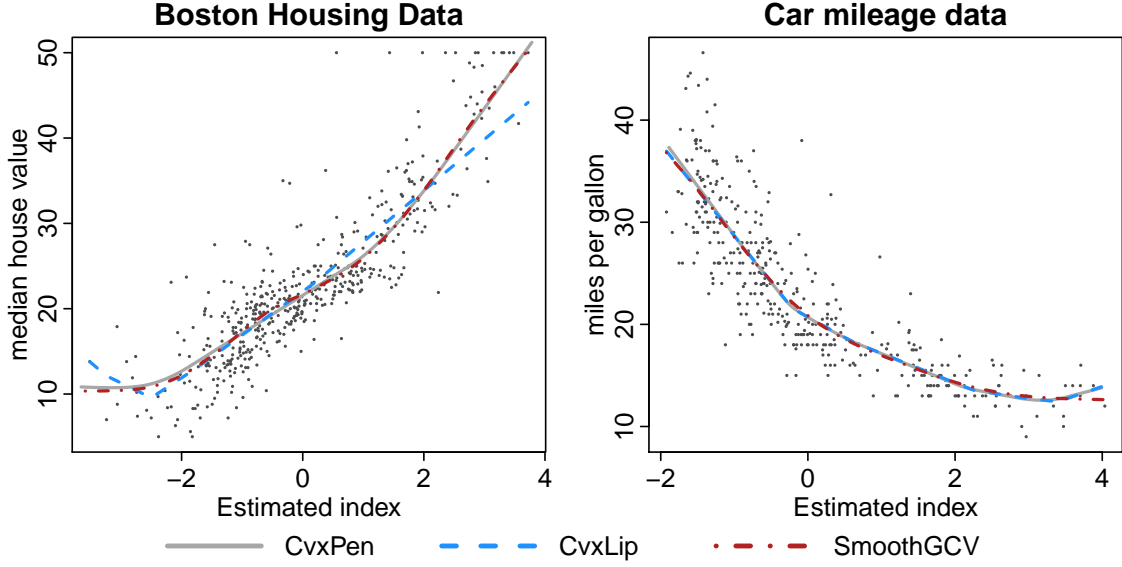


FIG 6. Scatter plots of  $\{(x_i^\top \hat{\theta}, y_i)\}_{i=1}^n$  overlaid with the plots of function estimates corresponding to *CvxPen*, *CvxLip*, and *SmoothGCV* estimators for the two real datasets considered. Left panel: Boston housing data (Appendix D.1); right panel: the car mileage data (Appendix D.2).

**Lemma E.1** (Lemma 4 of [32]). *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 E.2** (Lemma 5 of [32]). *For any set  $A \in \mathbb{R}^k$  ( $k \geq 1$ ), let  $\varnothing(A)$  denote the diameter of the set  $A$ . Let  $m \in \{g \in \mathcal{R} : J(g) < \infty \text{ and } \|g\|_\infty \leq M\}$ , where  $M$  is a finite constant. Then*

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

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

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

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

### E.1. Proof of Lemma 2.1

In the following we show that  $(m_0, \theta_0)$  is the minimizer of  $Q$  and is well-separated, with respect to the  $L_2(P_X)$  norm, from  $\{(m, \theta) : m \circ \theta \in L_2(P_X)\} \setminus m_0 \circ \theta_0$ . Choose arbitrarily small  $\delta > 0$ , and pick any  $(m, \theta)$  such that  $m \circ \theta \in L_2(P_X)$  and  $\|m \circ \theta - m_0 \circ \theta_0\|^2 > \delta^2$ . 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  $Q(m, \theta) > Q(m_0, \theta_0) + \delta^2$ .

### E.2. Proof of Theorem 2.1

We consider the estimator

$$(\tilde{m}_n, \tilde{\theta}_n) = \arg \min_{(m, \theta) \in \mathcal{M}_L \times \Theta} Q_n(m, \theta).$$

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

$$m_\theta := \arg \min_{m \in \mathcal{M}_L} Q_n(m, \theta) \quad \text{and} \quad T(\theta) := Q_n(m_\theta, \theta). \quad (\text{E.1})$$

Observe that  $\tilde{\theta}_n := \arg \min_{\theta \in \Theta} T(\theta)$ . As  $\Theta$  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  $\Theta$ ; see the Weierstrass extreme value theorem. We will now prove that  $\theta \mapsto T(\theta)$  is a continuous function. But first we will show that for every  $\theta \in \Theta$ ,  $\|m_\theta\|_\infty \leq C$ , where  $C$  depends only on  $\{(x_i, y_i)\}_{i=1}^n$ ,  $L$ , and  $T$ . Observe that  $\sum_{i=1}^n (y_i - m_\theta(\theta^\top x_i))^2 \leq \sum_{i=1}^n y_i^2$  and the constant function 0 belongs to  $\mathcal{M}_L$ . Thus

$$\sum_{i=1}^n [m_\theta(\theta^\top x_i)]^2 \leq 2 \sum_{i=1}^n y_i m_\theta(\theta^\top x_i) \leq 2 \left( \sum_{i=1}^n y_i^2 \right)^{1/2} \left( \sum_{i=1}^n [m_\theta(\theta^\top x_i)]^2 \right)^{1/2}.$$

Hence, we have  $|m_\theta(\theta^\top x_1)| \leq 2\sqrt{\sum_{i=1}^n y_i^2}$ . As  $m_\theta$  is uniformly  $L$ -Lipschitz, we have that for any  $t \in D$ ,

$$|m_\theta(t)| \leq |m_\theta(\theta^\top x_1)| + L|t - \theta^\top x_1| \leq \sqrt{4 \sum_{i=1}^n y_i^2 + LT} =: C.$$

As  $C$  does not depend on  $\theta$ , we have that  $\sup_{\theta \in \Theta} \|m_\theta\|_\infty \leq C$ . To show that  $\hat{\theta}$  exists, it suffices to show that  $\theta \mapsto T(\theta)$  is a continuous function. As a first step, we will show that the class of functions

$$\{\theta \mapsto Q_n(m, \theta) : m \in \mathcal{M}_L, \|m\|_\infty \leq C\}$$

is uniformly equicontinuous. Observe that for  $\theta, \eta \in \Theta$ , we have

$$\begin{aligned} n|Q_n(m, \theta) - Q_n(m, \eta)| &= \left| \sum_{i=1}^n (y_i - m(\theta^\top x_i))^2 - \sum_{i=1}^n (y_i - m(\eta^\top x_i))^2 \right| \\ &= \left| \sum_{i=1}^n (m(\eta^\top x_i) - m(\theta^\top x_i))(2y_i - m(\theta^\top x_i) - m(\eta^\top x_i)) \right| \\ &\leq \sum_{i=1}^n |m(\eta^\top x_i) - m(\theta^\top x_i)| \times |2y_i - m(\theta^\top x_i) - m(\eta^\top x_i)| \\ &\leq L \sum_{i=1}^n |\eta^\top x_i - \theta^\top x_i| \times 2(|y_i| + C) \\ &\leq 2nLT \left( \max_i |y_i| + C \right) |\theta - \eta|. \end{aligned}$$

Thus, we have that

$$\sup_{\{m \in \mathcal{M}_L : \|m\|_\infty \leq C\}} |Q_n(m, \theta) - Q_n(m, \eta)| \leq C_3 |\theta - \eta|,$$

where  $C_3$  is a constant depending only on  $\{y_i\}_{i=1}^n$  and  $C$ . Next we show that  $|T(\theta) - T(\eta)| \leq 2C_3 |\theta - \eta|$ . Recall that  $T(\theta) = Q_n(m_\theta, \theta)$ . By (E.1), we have

$$Q_n(m_\theta, \theta) - Q_n(m_\theta, \eta) = T(\theta) - Q_n(m_\theta, \eta) \leq T(\theta) - T(\eta)$$

and

$$T(\theta) - T(\eta) \leq Q_n(m_\eta, \theta) - T(\eta) = Q_n(m_\eta, \theta) - Q_n(m_\eta, \eta).$$

Thus

$$|T(\theta) - T(\eta)| \leq |Q_n(m_\eta, \theta) - Q_n(m_\eta, \eta)| + |Q_n(m_\theta, \theta) - Q_n(m_\theta, \eta)| \leq 2C_3 |\theta - \eta|.$$

### E.3. Proof of Theorem 2.2

The minimization problem considered is

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

where  $\mathcal{L}_n$  is defined in (2.3). For any fixed vector  $\theta \in \Theta$ , define  $t_i^\theta := \theta^\top x_i$ , for  $i = 1, \dots, n$ . Then we have

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

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

$$T(\theta) := \inf_{m \in \mathcal{R}} \mathcal{L}_n(m, \theta; \lambda) \quad \text{and} \quad m_\theta := \arg \min_{m \in \mathcal{R}} \mathcal{L}_n(m, \theta; \lambda). \quad (\text{E.2})$$

Theorem 1 of [16] proves that the infimum in (E.2) is attained for every  $\theta \in \Theta$  and the unique minimizer  $m_\theta$  is a natural cubic spline, see Section 5.1.2 for more details. Furthermore [16] note that  $m_\theta$  does not depend on  $D$  beyond the condition that  $\{t_i^\theta\}_{1 \leq i \leq n} \in D$ . Moreover,  $m_\theta''$  is zero outside  $(t_{(1)}^\theta, t_{(n)}^\theta)$ , where for  $k = 1, \dots, n$ ,  $t_{(k)}^\theta$  denotes the  $k$ -th smallest value in  $\{t_i^\theta\}_{i=1}^n$ . In the proof of Theorem 1 of [33], they show that  $\theta \mapsto T(\theta)$  is a continuous function if  $m_\theta$  satisfies the above properties. This completes the proof as  $\Theta$  is a compact set; see the Weierstrass extreme value theorem.

## Appendix F: Proofs of results in Section 3.1

### F.1. Proof of Theorem 3.1

To find the rate of convergence of  $\tilde{m} \circ \tilde{\theta}$ , we use the following modification of Theorem 3.2.5 of [53]. In the following to avoid measurability difficulties, we use  $\mathbb{P}^*$  and  $\mathbb{E}^*$ , outer probability and outer expectation.

**Lemma F.1.** *Let  $\mathbb{M}_n$  be stochastic processes indexed by a semimetric set  $\Upsilon$  and  $\mathbb{M} : \Upsilon \rightarrow \mathbb{R}$  be a deterministic function, such that for every  $\eta \in \Upsilon$*

$$\mathbb{M}(\eta) - \mathbb{M}(\eta_0) \lesssim -d^2(\eta, \eta_0), \quad (\text{F.1})$$

where  $d(\cdot, \eta_0) : \Upsilon \rightarrow \mathbb{R}^+$ . Let  $\hat{\eta}_n := \arg \max_{\eta \in \Upsilon} \mathbb{M}_n(\eta)$ . For each  $\varepsilon > 0$ , suppose that the following hold:

1. *There exists  $\Upsilon_\varepsilon$ , a subset of  $\Upsilon$ , containing  $\eta_0$  in its interior that satisfies*

$$\mathbb{P}^*(\hat{\eta}_n \notin \Upsilon_\varepsilon) \leq \varepsilon, \quad \forall n. \quad (\text{F.2})$$

2. *For every  $n$  and  $\delta > 0$ , the centered process  $\mathbb{M}_n - \mathbb{M}$  satisfies*

$$\sqrt{n} \mathbb{E}^* \left| \sup_{\substack{d(\eta, \eta_0) < \delta, \\ \eta \in \Upsilon_\varepsilon}} |(\mathbb{M}_n - \mathbb{M})(\eta) - (\mathbb{M}_n - \mathbb{M})(\eta_0)| \right| \leq C_\varepsilon \phi_n(\delta), \quad (\text{F.3})$$

for functions  $\phi_n$  (not depending on  $\varepsilon$ ) such that  $\delta \mapsto \phi_n(\delta)/\delta^\alpha$  is decreasing in  $\delta$  for some constant  $\alpha < 2$  (not depending on  $n$ ) and a constant  $C_\varepsilon > 0$ .

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 F.1.** *The proof of Lemma F.1 is similar to the proof given in Page 290 of [53]. The only difference is that in the “peeling” argument the “shells” are now defined as  $S_{j,n} := \{\eta : 2^{j-1} < r_n d(\eta, \eta_0) \leq 2^j \text{ and } \eta \in \Upsilon_\varepsilon\}$  and the first inequality in the proof now reads*

$$\mathbb{P}^*(r_n d(\hat{\eta}, \eta_0) > 2^M) \leq \sum_{j \geq M} \mathbb{P}^* \left( \sup_{\eta \in S_{j,n}} (\mathbb{M}_n(\eta) - \mathbb{M}_n(\eta_0)) \geq 0 \right) + \mathbb{P}^*(\hat{\eta} \notin \Upsilon_\varepsilon).$$

We will now obtain the desired rate of convergence in Theorem 3.1 by verifying conditions of Lemma F.1. 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

$$\mathbb{M}_n(m, \theta) := -\frac{1}{n} \sum_{i=1}^n (y_i - m(\theta^\top x_i))^2 \text{ and } \mathbb{M}(m, \theta) := -\mathbb{E}(Y - m(\theta^\top X))^2. \quad (\text{F.4})$$

For any  $(m_1, \theta_1)$  and  $(m_2, \theta_2)$  in  $\mathcal{M}_L \times \Theta$ , we define

$$d((m_1, \theta_1), (m_2, \theta_2)) := \|m_1 \circ \theta_1 - m_2 \circ \theta_2\|. \quad (\text{F.5})$$

We first show that  $\mathbb{M}$  defined in (F.4) satisfies (F.1). Observe that

$$\begin{aligned} \mathbb{M}(m, \theta) - \mathbb{M}(m_0, \theta_0) &= \mathbb{E}[(Y - m_0(\theta_0^\top X))^2 - (Y - m(\theta^\top X))^2] \\ &= -2\mathbb{E}[(Y - m_0(\theta_0^\top X))(m_0(\theta_0^\top X) - m(\theta^\top X))] - \mathbb{E}(m_0(\theta_0^\top X) - m(\theta^\top X))^2 \\ &= -\mathbb{E}\left[\{m(\theta^\top X) - m_0(\theta_0^\top X)\}^2\right] \quad (\text{as } \mathbb{E}(Y|X) = m_0(\theta_0^\top X)) \\ &= -d^2((m, \theta), (m_0, \theta_0)). \end{aligned}$$

Next for every  $\varepsilon > 0$ , we find  $\Upsilon_\varepsilon$  that satisfies (F.2). The following result (proved in Appendix F.2) gives the form of  $\Upsilon_\varepsilon$ .

**Lemma F.2.** *Under assumption (A2), we have that  $\|\check{m}_n\|_\infty = O_p(1)$ . Moreover, for every  $\varepsilon > 0$ , there exists a finite  $M_\varepsilon$  such that*

$$\mathbb{P}(\check{m}_n \notin \mathcal{M}_{M_\varepsilon, L}) \leq \varepsilon, \quad \forall n,$$

where for any  $M > 0$ , we define

$$\mathcal{M}_{M, L} := \{m \in \mathcal{M}_L : \|m\|_\infty \leq M\}. \quad (\text{F.6})$$

We can now define  $\Upsilon_\varepsilon := \mathcal{M}_{M_\varepsilon, L} \times \Theta$ . By Lemma F.2, we have

$$\mathbb{P}((\check{m}_n, \check{\theta}_n) \notin \Upsilon_\varepsilon) \leq \varepsilon, \quad \forall n.$$

To find the rate of convergence of  $\check{m}_n \circ \check{\theta}_n$ , we need to find a function  $\phi_n(\delta)$  that satisfies (F.3). Recall that  $\epsilon = Y - m_0(\theta_0^\top X)$ . By definition of  $\mathbb{M}_n$  and  $\mathbb{M}$ , we have

$$\begin{aligned} &\sqrt{n} |(\mathbb{M}_n - \mathbb{M})(m, \theta) - (\mathbb{M}_n - \mathbb{M})(m_0, \theta_0)| \\ &= \left| \mathbb{G}_n \left[ -2(Y - m_0(\theta_0^\top X))(m_0(\theta_0^\top X) - m(\theta^\top X)) + (m_0(\theta_0^\top X) - m(\theta^\top X))^2 \right] \right| \\ &\leq \left| \mathbb{G}_n [2\epsilon(m(\theta^\top X) - m_0(\theta_0^\top X))] \right| + \left| \mathbb{G}_n [(m(\theta^\top X) - m_0(\theta_0^\top X))^2] \right|. \end{aligned} \quad (\text{F.7})$$

Now, we find the upper bound  $\phi_n(\delta)$  by obtaining upper bounds for both the terms in (F.7). Define two classes of functions

$$\begin{aligned} \mathcal{H}_{M_\varepsilon, L}(\delta) &:= \{m \circ \theta - m_0 \circ \theta_0 : (m, \theta) \in \Upsilon_\varepsilon \text{ and } d((m, \theta), (m_0, \theta_0)) \leq \delta\} \\ \mathfrak{H}_{M_\varepsilon, L}(\delta) &:= \{f^2 : f \in \mathcal{H}_{M_\varepsilon, L}(\delta)\}, \end{aligned} \quad (\text{F.8})$$

where  $d(\cdot, \cdot)$  is defined in (F.5). Thus by (F.7), we have

$$\begin{aligned} &\mathbb{E}^* \sup_{\substack{d((m, \theta), (m_0, \theta_0)) < \delta, \\ (m, \theta) \in \Upsilon_\varepsilon}} \sqrt{n} |(\mathbb{M}_n - \mathbb{M})(m, \theta) - (\mathbb{M}_n - \mathbb{M})(m_0, \theta_0)| \\ &\leq \mathbb{E}^* \sup_{f \in \mathcal{H}_{M_\varepsilon, L}(\delta)} |\mathbb{G}_n[\epsilon f]| + \mathbb{E}^* \sup_{f \in \mathfrak{H}_{M_\varepsilon, L}(\delta)} |\mathbb{G}_n f|. \end{aligned} \quad (\text{F.9})$$

In the following two lemmas (proved in Appendices F.3.1 and F.3.2, respectively) we show that both the terms (F.9) are bounded by constant multiples (depending only on  $L, \varepsilon, D, M_\varepsilon$  and  $M_0$ ) of  $\delta^{3/4} + n^{-1/2}\delta^{1/2}$ .

**Lemma F.3.** For every  $\varepsilon, \nu > 0$ , we have

$$\log N_{[]}(\nu, \mathcal{H}_{M_\varepsilon, L}(\delta), L_2(P_{\theta_0, m_0})) \leq C_1^* \nu^{-1/2} \text{ and } \sup_{f \in \mathcal{H}_{M_\varepsilon, L}(\delta)} \|f\|_\infty \leq M_\varepsilon + M_0,$$

where  $\mathcal{H}_{M_\varepsilon, L}(\delta)$  is defined in (F.8) and  $C_1^*$  is a constant that depends only on  $M_\varepsilon, L, D, T, M_0, d$ , and the distribution of  $\epsilon$ . Furthermore

$$\mathbb{E}^* \sup_{f \in \mathcal{H}_{M_\varepsilon, L}(\delta)} |\mathbb{G}_n[\epsilon f]| \lesssim C_{\varepsilon, 1} \left( \delta^{3/4} + \frac{\delta^{1/2}}{\sqrt{n}} \right), \quad (\text{F.10})$$

where  $C_{\varepsilon, 1}$  is a constant depending only on  $C_1^*, M_\varepsilon, M_0, d$ , and the distribution of  $\epsilon$ .

**Lemma F.4.** For every  $\varepsilon > 0$ ,

$$\sup_{f \in \mathfrak{H}_{M_\varepsilon, L}(\delta)} \|f\|_\infty \leq 4(M_\varepsilon + M_0)^2 \text{ and } \sup_{f \in \mathfrak{H}_{M_\varepsilon, L}(\delta)} \|f\| \leq 2(M_\varepsilon + M_0)\delta.$$

Furthermore, for  $\nu > 0$  we have

$$\log N_{[]}(\nu, \mathfrak{H}_{M_\varepsilon, L}(\delta), L_2(P_{\theta_0, m_0})) \leq \left( \frac{M_\varepsilon + L\varnothing(D)}{\nu} \right)^{1/2}$$

and

$$\mathbb{E}^* \sup_{f \in \mathfrak{H}_{M_\varepsilon, L}(\delta)} |\mathbb{G}_n f| \leq C_{\varepsilon, 2} \left( \delta^{3/4} + \frac{\delta^{1/2}}{\sqrt{n}} \right), \quad (\text{F.11})$$

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 (F.10) and (F.11) to (F.9), we have  $\phi_n(\delta) = (C_{\varepsilon, 1} + C_{\varepsilon, 2}) \left( \delta^{3/4} + n^{-1/2} \delta^{1/2} \right)$ . Observe that  $\phi_n(\delta)/\delta^{3/4}$  is a decreasing function of  $\delta$  and

$$n^{4/5} \phi_n(n^{-2/5}) \leq \sqrt{n}.$$

Thus, by Lemma F.1, we have  $n^{2/5} \|\tilde{m}_n \circ \check{\theta}_n - m_0 \circ \theta_0\| = O_p^*(1)$ .

### F.2. Proof of Lemma F.2

By the definition of  $(\tilde{m}_n, \check{\theta}_n)$ , we have

$$\sum_{i=1}^n (y_i - \tilde{m}_n(\check{\theta}_n^\top x_i))^2 \leq \sum_{i=1}^n (y_i - m(\check{\theta}_n^\top x_i))^2$$

for all  $m \in \mathcal{M}_L$ . Since any constant function belongs to  $\mathcal{M}_L$ , for any fixed real  $\kappa$ , we have

$$\sum_{i=1}^n (y_i - \tilde{m}_n(\check{\theta}_n^\top x_i))^2 \leq \sum_{i=1}^n (y_i - \tilde{m}_n(\check{\theta}_n^\top x_i) + \kappa)^2.$$

A simplification of the above inequality gives us:

$$2\kappa \sum_{i=1}^n (y_i - \tilde{m}_n(\check{\theta}_n^\top x_i)) + n\kappa^2 \geq 0, \text{ for all } \kappa \Rightarrow \sum_{i=1}^n (y_i - \tilde{m}_n(\check{\theta}_n^\top x_i)) = 0. \quad (\text{F.12})$$

Thus for any  $t \in D$ , we have

$$\begin{aligned}
|\check{m}_n(t)| &\leq \left| \check{m}_n(t) - \frac{1}{n} \sum_{j=1}^n \check{m}_n(\check{\theta}_n^\top x_j) \right| + \left| \frac{1}{n} \sum_{j=1}^n \check{m}_n(\check{\theta}_n^\top x_j) \right| \\
&\leq \frac{1}{n} \sum_{j=1}^n |\check{m}_n(t) - \check{m}_n(\check{\theta}_n^\top x_j)| + \left| \frac{1}{n} \sum_{j=1}^n \{m_0(\theta_0^\top x_j) + \epsilon_j\} \right| \quad (\text{by (F.12)}) \\
&\leq \frac{1}{n} \sum_{j=1}^n L|t - \check{\theta}_n^\top x_j| + \frac{1}{n} \sum_{j=1}^n |m_0(\theta_0^\top x_j)| + \left| \frac{1}{n} \sum_{j=1}^n \epsilon_j \right| \\
&\leq L\varnothing(A) + M_0 + \left| \frac{1}{n} \sum_{j=1}^n \epsilon_j \right|,
\end{aligned}$$

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  $\epsilon$  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  $c_\varepsilon$  (depending only on the distribution of  $\epsilon$  and  $\varepsilon$ ) 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

$$\mathbb{P}(\|\check{m}_n\|_\infty > M_\varepsilon) \leq \varepsilon, \quad \forall n.$$

### F.3. Proofs of Lemmas F.3 and F.4

To prove Lemmas F.3 and F.4, we need the following entropy result.

**Lemma F.5.** *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 (F.6). Then there exist positive constants  $c$  and  $\nu_0$ , 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 [18].

**Lemma F.6** (Theorem 3.2, [18]). *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 exist positive constants  $c$  and  $\nu_0$ , 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 F.6 and Lemma 4.1 of [44] for  $\nu \in (0, 1)$ , we have

$$\begin{aligned}
\log N_{[]}(\nu, \mathcal{M}_{M,L}, \|\cdot\|_\infty) &\leq c \left( \frac{M + L\varnothing(D)}{\nu} \right)^{1/2}, \\
\log N(\nu, \Theta, |\cdot|) &\leq -c \log(\nu),
\end{aligned}$$

where  $c$  is a constant that depends only on  $d$ .

Recall that  $\sup_{x \in \mathcal{X}} |x| \leq T$ ; see (A1). Let  $\{\theta_1, \theta_2, \dots, \theta_p\}$  be a  $\nu/(2LT)$ -cover (with respect to the Euclidean norm) of  $\Theta$  and  $\{m_1, m_2, \dots, m_q\}$  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  $\nu$ -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  $\theta_j$  such that  $\|m - m_i\|_\infty \leq \nu/2$  and  $|\theta - \theta_j| \leq \nu/(2LT)$ . Therefore, for any  $x \in \mathcal{X}$

$$\begin{aligned} |m(\theta^\top x) - m_i(\theta_j^\top x)| &\leq |m(\theta^\top x) - m(\theta_j^\top x)| + |m(\theta_j^\top x) - m_i(\theta_j^\top x)| \\ &\leq L|x|\|\theta - \theta_j\| + \|m - m_i\|_\infty \leq \frac{L|x|\nu}{2LT} + \frac{\nu}{2} \leq \nu. \end{aligned}$$

Thus for  $\nu \leq \nu_0(M + L\varnothing(D))$ , 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.  $\square$

### F.3.1. Proof of Lemma F.3

Suppose  $\mathcal{F}$  is a class of real valued functions defined on  $\mathcal{X}$ . We first present a result that gives a maximal inequality for the class of functions

$$\epsilon\mathcal{F} := \{\epsilon f : f \in \mathcal{F}\}$$

in terms of the bracketing entropy of  $\mathcal{F}$ , with respect the  $L_2(P_{\theta_0, m_0})$  norm.

**Lemma F.7.** *Suppose  $\mathcal{F}$  is a class of functions (defined on  $\mathcal{X}$ ) such that*

$$\sup_{f \in \mathcal{F}} \|f\|_\infty \leq \Phi, \sup_{f \in \mathcal{F}} \|f\| \leq \kappa, \text{ and } \log N_{[\cdot]}(\nu, \mathcal{F}, \|\cdot\|) \leq \Delta \nu^{-\alpha},$$

for some constant  $\Delta$  and  $0 < \alpha < 2$ , where  $\|f\|^2 := \int_{\mathcal{X}} f^2 dP_X$ . Then

$$\log N_{[\cdot]}(K^* \nu, \epsilon\mathcal{F}, \|\cdot\|_B) \leq \Delta \nu^{-\alpha},$$

where for any  $g \in L_2(P_{\theta_0, m_0})$ ,  $\|g\|_B$  (Bernstein “norm”) is defined as

$$\|g\|_B := \left[ 2\mathbb{E} \left( \exp(|g|) - 1 - |g| \right) \right]^{1/2},$$

$K^* := \sup_x (\mathbb{E} [\epsilon^2 \exp(2\Phi|\epsilon|)|X = x])^{1/2}$ , and  $\epsilon\mathcal{F} := \{\epsilon f : f \in \mathcal{F}\}$ . Furthermore for all  $f \in \mathcal{F}$ ,  $\|\epsilon f\|_B \leq K^* \|f\|$  and

$$\mathbb{E}^* \sup_{f \in \mathcal{F}} |\mathbb{G}_n \epsilon f| \lesssim \frac{\Delta^{1/2} K^* \kappa^{1-\alpha/2}}{(1-\alpha/2)} + \frac{\Delta \kappa^{-\alpha}}{\sqrt{n} (1-\alpha/2)^2}. \quad (\text{F.13})$$

*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  $f_1, f_2 : \mathcal{X} \rightarrow [-\Phi, \Phi]$ , such that

$$\|f_2 - f_1\| \leq \nu \text{ and } f_1(x) \leq f(x) \leq f_2(x), \quad \forall x \in \mathcal{X}. \quad (\text{F.14})$$

Define  $\epsilon^+ := \max\{\epsilon, 0\}$  and  $\epsilon^- := \max\{0, -\epsilon\}$ . Multiplying  $\epsilon^+$  and  $\epsilon^-$  to the second inequality in (F.14), we have

$$f_1(x)\epsilon^+ \leq f(x)\epsilon^+ \leq f_2(x)\epsilon^+ \quad \text{and} \quad -f_2(x)\epsilon^- \leq -f(x)\epsilon^- \leq -f_1(x)\epsilon^-,$$

respectively. Combining the above inequalities, we have

$$f_1(x)\epsilon^+ - f_2(x)\epsilon^- \leq f(x)\epsilon \leq f_2(x)\epsilon^+ - f_1(x)\epsilon^-.$$



Moreover,

$$\begin{aligned}
& \|f_2(X)\epsilon^+ - f_1(X)\epsilon^- - f_1(X)\epsilon^+ + f_2(X)\epsilon^-\|_B^2 \\
&= \|(f_2(X) - f_1(X))\epsilon\|_B^2 \\
&= 2\mathbb{E}\left\{\exp(|(f_2(X) - f_1(X))\epsilon|) - 1 - |(f_2(X) - f_1(X))\epsilon|\right\} \\
&\leq \mathbb{E}\left\{(f_2(X) - f_1(X))^2 \epsilon^2 \exp(|(f_2(X) - f_1(X))\epsilon|)\right\} \\
&\leq \mathbb{E}\left\{(f_2(X) - f_1(X))^2 \epsilon^2 \exp(2\Phi|\epsilon|)\right\} \\
&= \mathbb{E}\left\{(f_2(X) - f_1(X))^2 \mathbb{E}[\epsilon^2 \exp(2\Phi|\epsilon|)|X]\right\} \\
&\leq (K^*)^2 \|f_2 - f_1\|^2 \leq (K^* \nu)^2,
\end{aligned}$$

where  $K^*$  is as given in the statement of the lemma.

Thus if  $(f_1, f_2)$  is a  $\nu$ -bracket (with respect to  $\|\cdot\|$ -norm) for  $f$ , then  $(f_1\epsilon^+ - f_2\epsilon^-, f_2\epsilon^+ - f_1\epsilon^-)$  is a  $K^*\nu$ -bracket for  $\epsilon f$  (with respect to  $\|\cdot\|_B$ -norm). Therefore, we have

$$\log N_{[]} (K^* \nu, \epsilon \mathcal{F}, \|\cdot\|_B) \leq \log N_{[]} (\nu, \mathcal{F}, \|\cdot\|) \leq \Delta \nu^{-\alpha}.$$

Hence, for every  $\nu > 0$ ,

$$\log N_{[]} (\nu, \epsilon \mathcal{F}, \|\cdot\|_B) \leq \Delta \left( \frac{\nu}{K^*} \right)^{-\alpha}$$

To prove (F.13), we use the following Lemma.

**Lemma F.8** (Lemma 3.4.3 of [53]). *Let  $\mathcal{G}$  be a class of measurable functions such that  $\sup_{g \in \mathcal{G}} \|g\|_B \leq \rho$ . Then*

$$\mathbb{E}^* \sup_{g \in \mathcal{G}} |\mathbb{G}_n g| \lesssim J_{[]}(\rho, \mathcal{G}, \|\cdot\|_B) \left( 1 + \frac{J_{[]}(\rho, \mathcal{G}, \|\cdot\|_B)}{\rho^2 \sqrt{n}} \right). \quad (\text{F.15})$$

We now find an upper bound for  $\sup_{f \in \mathcal{F}} \|\epsilon f\|_B$ . Observe that

$$\begin{aligned}
\|\epsilon f\|_B^2 &= 2\mathbb{E}\left\{\exp(|\epsilon f(X)|) - 1 - |\epsilon f(X)|\right\} \\
&\leq \mathbb{E}\left\{\epsilon^2 f^2(X) \exp(|f(X)\epsilon|)\right\} \\
&\leq \mathbb{E}\left\{\epsilon^2 f^2(X) \exp(\Phi|\epsilon|)\right\} \\
&\leq \mathbb{E}\left\{f^2(X) \mathbb{E}[\epsilon^2 \exp(2\Phi|\epsilon|)|X]\right\} \leq (K^*)^2 \|f\|^2 \leq (K^* \kappa)^2.
\end{aligned}$$

Thus, for the class  $\epsilon \mathcal{F}$ , we can apply Lemma F.8 with  $\rho = K^* \kappa$ . By definition

$$J_{[]} (K^* \kappa, \epsilon \mathcal{F}, \|\cdot\|_B) \leq \int_0^{K^* \kappa} \Delta^{1/2} \left( \frac{\nu}{K^*} \right)^{-\alpha/2} d\nu = \Delta^{1/2} K^* \kappa^{1-\alpha/2} / (1 - \alpha/2).$$

Therefore by (F.15), we have

$$\mathbb{E}^* \sup_{f \in \mathcal{F}} |\mathbb{G}_n[\epsilon f]| \lesssim \frac{\Delta^{1/2} K^* \kappa^{1-\alpha/2}}{(1 - \alpha/2)} + \frac{\Delta \kappa^{-\alpha}}{\sqrt{n} (1 - \alpha/2)^2}. \quad \square$$

The proof of Lemma F.3 will now be completed by a simple application of Lemma F.7 with  $\mathcal{F} = \mathcal{H}_{M_\epsilon, L}(\delta)$ . By definition (F.8), we have

$$\sup_{f \in \mathcal{H}_{M_\epsilon, L}(\delta)} \|f\|_\infty < M_\epsilon + M_0 \text{ and } \sup_{f \in \mathcal{H}_{M_\epsilon, L}(\delta)} \|f\| < \delta.$$

As  $\mathcal{H}_{M_\epsilon, L}(\delta) \subset \mathcal{H}_{M_\epsilon, L}$ , by Lemma F.5, we have

$$\log N_{[]} (\nu, \mathcal{H}_{M_\epsilon, L}(\delta), \|\cdot\|_\infty) \leq \log N_{[]} (\nu, \mathcal{H}_{M_\epsilon, L}, \|\cdot\|_\infty) \leq K' \nu^{-1/2}.$$

Thus

$$\log N_{[]}(\nu, \mathcal{H}_{M_\varepsilon, L}(\delta), \|\cdot\|) \leq C_1^* \nu^{-1/2},$$

where  $C_1^* = \sqrt{2}K'$ . By applying Lemma F.7 (see (F.13)) with

$$\Phi = M_\varepsilon + M_0, \quad \kappa = \delta, \quad \Delta = C_1^*, \quad \text{and } \alpha = 1/2,$$

we have

$$\mathbb{E}^* \sup_{f \in \mathcal{H}_{M_\varepsilon, L}(\delta)} |\mathbb{G}_n[\epsilon f]| \leq C_{\varepsilon, 1} \left( \delta^{3/4} + \frac{\delta^{-1/2}}{\sqrt{n}} \right),$$

where  $C_{\varepsilon, 1}$  is constant depending only on  $K', M_\varepsilon, M_0, L, d$ , and  $T$ .

### F.3.2. Proof of Lemma F.4

We proceed as in the proof of Lemma F.3. For any function  $f \in \mathcal{H}_{M_\varepsilon, L}$ , there exist functions  $f_1, f_2 : \mathcal{X} \rightarrow [-M_\varepsilon - M_0, M_0 + M_\varepsilon]$  such that  $f_1(x) \leq f(x) \leq f_2(x)$  and  $0 \leq f_2(x) - f_1(x) \leq \nu$ , for each  $x \in \mathcal{X}$ ; see Lemma F.5. Observe that for any two real numbers  $x \leq y$ , we have  $x^+ \leq y^+$  and  $y^- \leq x^-$ . Thus, we have

$$f_1^+ \leq f^+ \leq f_2^+ \quad \text{and} \quad f_2^- \leq f^- \leq f_1^-.$$

The above inequalities lead to a bracket for  $f^2$ . Observe that

$$f_1^+ + f_2^- \leq |f| \leq f_1^- + f_2^+ \Rightarrow (f_1^+ + f_2^-)^2 \leq f^2 \leq (f_1^- + f_2^+)^2,$$

and the length of the above bracket 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  $\nu$ -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

$$J_{[]}(\rho, \mathfrak{H}_{M_\varepsilon, L}(\delta), \|\cdot\|) \leq \int_0^{2(M_\varepsilon + M_0)\delta} \sqrt{C_2^* \nu^{-1/2}} d\nu \leq \frac{8}{3} \sqrt{C_2^*} [(M_\varepsilon + M_0)\delta]^{3/4}.$$

To complete the proof we use the following lemma.

**Lemma F.9** (Lemma 3.4.2 of [53]). *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,

$$0 \leq f(\cdot) \leq 4(M_\varepsilon + M_0)^2 \quad \text{and} \quad \mathbb{E} f^2 \leq \|f\|_\infty \mathbb{E} f \leq 4(M_\varepsilon + M_0)^2 \delta^2 =: \rho^2.$$

Moreover

$$J_{[]}(\rho, \mathfrak{H}_{M_\varepsilon, L}(\delta), L_2(P)) \leq \int_0^{2(M_\varepsilon + M_0)\delta} \sqrt{C_2^* \nu^{-1/2}} d\nu \leq \frac{8}{3} \sqrt{C_2^*} [(M_\varepsilon + M_0)\delta]^{3/4}.$$

Thus by Lemma F.9,

$$\mathbb{E}^* \sup_{f \in \mathfrak{H}_{M_\varepsilon, L}(\delta)} |\mathbb{G}_n f| \leq C_{\varepsilon, 2} \left( \delta^{3/4} + \frac{\delta^{-1/2}}{\sqrt{n}} \right).$$

#### F.4. Proof of Theorem 3.2

Let  $(m_n, \theta_n)$  be any sequence in  $\mathcal{M}_L \times \Theta$ . Recall that  $\mathcal{M}_L$  is a class of closed, bounded, and equicontinuous functions and  $\Theta$  is a compact set. Thus, by Ascoli-Arzelà theorem, there exists a subsequence  $\{(m_{n_k}, \theta_{n_k})\}$ ,  $\theta \in \Theta$ , and  $m \in \mathcal{M}_L$  such that  $|\theta_{n_k} - \theta| \rightarrow 0$  and  $\|m_{n_k} - m\|_{D_0} \rightarrow 0$ . Now suppose that  $\|m_n \circ \theta_n - m_0 \circ \theta_0\| \rightarrow 0$ . This implies that  $\|m \circ \theta - m_0 \circ \theta_0\| = 0$ . Then the continuity and almost everywhere differentiability of  $\{m_n\}$  implies that  $m \equiv m_0$  and  $\theta = \theta_0$ . Now recall that in Theorem 3.1 and Lemma F.2, we showed that  $\|\tilde{m} \circ \tilde{\theta} - m_0 \circ \theta_0\| = o_p(1)$  and  $\|\tilde{m}_n\|_\infty = O_p(1)$ , respectively. Thus by taking  $m_n = \tilde{m}_n$  and  $\theta = \tilde{\theta}_n$ , we have that  $|\tilde{\theta} - \theta_0| = o_p(1)$  and  $\|\tilde{m} - m_0\|_{D_0} = o_p(1)$ . The following lemma applied to  $\{\tilde{m}_n\}$  completes the proof of the theorem by showing that  $\|\tilde{m}'_n - m'_0\|_C = o_p(1)$  for any compact subset  $C$  in the interior of  $D_0$ .

**Lemma F.10** (Lemma 3.10, [47]). *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  $f_n \rightarrow f$  pointwise on  $\mathcal{C}$ . Then, if  $C \subset \mathcal{C}$  is any compact set,*

$$\sup_{\substack{x \in C \\ \xi \in \partial f_n(x)}} |\xi - \nabla f(x)| \rightarrow 0,$$

where  $\partial f_n(x)$  represents the sub-differential set of  $f_n$  at  $x$ .

#### F.5. Proof of Theorem 3.3

We first state and prove a intermediary lemma.

**Lemma F.11.** *Let  $m_0$  and  $\theta_0$  satisfy the assumptions (A1), (A5), and (L1). Furthermore, let  $\{\theta_n\} \in \Theta$  and  $\{m_n\} \in \mathcal{M}_L$  be two non-random sequences such that*

$$|\theta_n - \theta_0| \rightarrow 0, \quad \|m_n - m_0\|_{D_0} \rightarrow 0, \quad \text{and} \quad \|m'_n - m'_0\|_C \rightarrow 0 \quad (\text{F.16})$$

for any compact subset  $C$  of the interior of  $D_0$ . Then

$$P_X |m_n(\theta_n^\top X) - m_0(\theta_0^\top X) - \{m'_0(\theta_0^\top X)X^\top(\theta_n - \theta_0) + (m_n - m_0)(\theta_0^\top X)\}|^2 = o(|\theta_n - \theta_0|^2).$$

*Proof.* For any convex function  $f \in \mathcal{M}_L$ , denote the right derivative of  $f$  by  $f'$ . Note that  $f'$  is a bounded increasing function. First, observe that

$$\begin{aligned} m_n(\theta_n^\top x) - m_0(\theta_0^\top x) &- [m'_0(\theta_0^\top x)x^\top(\theta_n - \theta_0) + (m_n - m_0)(\theta_0^\top x)] \\ &= m_n(\theta_n^\top x) - m_n(\theta_0^\top x) - m'_0(\theta_0^\top x)x^\top(\theta_n - \theta_0). \end{aligned}$$

Now,

$$\begin{aligned} &|m_n(\theta_n^\top x) - m_n(\theta_0^\top x) - m'_0(\theta_0^\top x)x^\top(\theta_n - \theta_0)|^2 \\ &= \left| \int_{\theta_n^\top x}^{\theta_0^\top x} m'_n(t) dt - m'_0(\theta_0^\top x)x^\top(\theta_n - \theta_0) \right|^2 \quad (m_n \text{ is absolutely continuous}) \\ &= \left| \int_{\theta_n^\top x}^{\theta_0^\top x} m'_n(t) dt - m'_n(\theta_0^\top x)x^\top(\theta_n - \theta_0) + m'_n(\theta_0^\top x)x^\top(\theta_n - \theta_0) - m'_0(\theta_0^\top x)x^\top(\theta_n - \theta_0) \right|^2 \\ &= \left| \int_{\theta_n^\top x}^{\theta_0^\top x} m'_n(t) dt - m'_n(\theta_0^\top x)x^\top(\theta_n - \theta_0) + (m'_n - m'_0)(\theta_0^\top x)x^\top(\theta_n - \theta_0) \right|^2 \\ &\leq 2 \left| \int_{\theta_n^\top x}^{\theta_0^\top x} m'_n(t) dt - m'_n(\theta_0^\top x)x^\top(\theta_n - \theta_0) \right|^2 + 2 \left| (m'_n - m'_0)(\theta_0^\top x)x^\top(\theta_n - \theta_0) \right|^2. \quad (\text{F.17}) \end{aligned}$$

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

$$m'_n(\theta_n^\top x) \wedge m'_n(\theta_0^\top x) \leq \frac{\int_{\theta_n^\top x}^{\theta_0^\top x} m'_n(t) dt}{x^\top(\theta_n - \theta_0)} \leq m'_n(\theta_n^\top x) \vee m'_n(\theta_0^\top x).$$

Thus for all  $x \in \mathcal{X}$ , we have

$$\left| \int_{\theta_n^\top x}^{\theta_0^\top x} m'_n(t) dt - m'_n(\theta_0^\top x) x^\top (\theta_n - \theta_0) \right| \leq |m'_n(\theta_n^\top x) - m'_n(\theta_0^\top x)| |x^\top (\theta_n - \theta_0)|. \quad (\text{F.18})$$

Note that if  $x^\top \theta_n = x^\top \theta_0$ , then both sides of (F.18) are 0. Combine (F.17) and (F.18), to conclude that

$$\begin{aligned} & P_X |m_n(\theta_n^\top X) - m_n(\theta_0^\top X) - m'_0(\theta_0^\top X) X^\top (\theta_n - \theta_0)|^2 \\ & \leq 2P_X |(m'_n(\theta_n^\top X) - m'_n(\theta_0^\top X)) X^\top (\theta_n - \theta_0)|^2 + 2P_X |(m'_n - m'_0)(\theta_0^\top X) X^\top (\theta_n - \theta_0)|^2. \end{aligned} \quad (\text{F.19})$$

As  $\mathcal{X}$  is bounded, the two terms on the right hand side of (F.19) can be bounded as

$$\begin{aligned} P_X |(m'_n(\theta_n^\top X) - m'_n(\theta_0^\top X)) X^\top (\theta_n - \theta_0)|^2 & \leq T^2 |\theta_n - \theta_0|^2 P_X |m'_n(\theta_n^\top X) - m'_n(\theta_0^\top X)|^2, \\ P_X |(m'_n - m'_0)(\theta_0^\top X) X^\top (\theta_n - \theta_0)|^2 & \leq T^2 |\theta_n - \theta_0|^2 P_X |(m'_n - m'_0)(\theta_0^\top X)|^2. \end{aligned}$$

We will now show that both  $P_X |m'_n(\theta_n^\top X) - m'_n(\theta_0^\top X)|^2$  and  $P_X |(m'_n - m'_0)(\theta_0^\top X)|^2$  converge to 0 as  $n \rightarrow \infty$ . First observe that

$$\begin{aligned} P_X |m'_n(\theta_n^\top X) - m'_n(\theta_0^\top X)|^2 & \lesssim P_X |m'_n(\theta_n^\top X) - m'_0(\theta_n^\top X)|^2 + P_X |m'_0(\theta_n^\top X) - m'_0(\theta_0^\top X)|^2 \\ & \quad + P_X |m'_0(\theta_0^\top X) - m'_n(\theta_0^\top X)|^2. \end{aligned} \quad (\text{F.20})$$

Recall that  $m'_0$  is a continuous and bounded function; see assumption (P1). Bounded convergence theorem now implies that  $P_X |m'_0(\theta_n^\top X) - m'_0(\theta_0^\top X)|^2 \rightarrow 0$ , as  $|\theta_n - \theta_0| \rightarrow 0$ . Now consider the first term on the right hand side of (F.20). As  $\theta_0^\top X$  has a density, for any  $\varepsilon > 0$ , we can define a compact subset  $C_\varepsilon$  in the interior of  $D_0$  such that  $\mathbb{P}(\theta_0^\top X \notin C_\varepsilon) < \varepsilon/4L$ . Now note that, by Theorem 3.2 and the fact that  $\mathbb{P}(\theta_n^\top X \notin C_\varepsilon) \rightarrow \mathbb{P}(\theta_0^\top X \notin C_\varepsilon)$ , we have

$$P_X |m'_n(\theta_n^\top X) - m'_0(\theta_n^\top X)|^2 \leq \sup_{t \in C_\varepsilon} |m'_n(t) - m'_0(t)| + 2LP(\theta_n^\top X \notin C_\varepsilon) \leq \varepsilon,$$

as  $n \rightarrow \infty$ . Similarly, we can see that

$$P_X |m'_0(\theta_0^\top X) - m'_n(\theta_0^\top X)|^2 \leq \sup_{t \in C_\varepsilon} |m'_n(t) - m'_0(t)| + 2LP(\theta_0^\top X \notin C_\varepsilon) \leq \varepsilon,$$

as  $n \rightarrow \infty$ . Combining the results, we have shown that for every  $\varepsilon > 0$

$$P_X |m_n(\theta_n^\top X) - m(\theta_0^\top X) - m'_0(\theta_0^\top X) X^\top (\theta_n - \theta_0)|^2 \leq T^2 |\theta_n - \theta_0|^2 \varepsilon,$$

for all sufficiently large  $n$ . Thus the result follows.  $\square$

We will now use the above lemma to prove Theorem 3.3. 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

$$\begin{aligned} A_n(x) - B_n(x) &= \check{m}_n(\check{\theta}_n^\top x) - m'_0(\theta_0^\top x) x^\top (\check{\theta}_n - \theta_0) - \check{m}_n(\theta_0^\top x) \\ &= \check{m}_n(\check{\theta}_n^\top x) - m_0(\theta_0^\top x) - \{m'_0(\theta_0^\top x) x^\top (\check{\theta}_n - \theta_0) + (\check{m}_n - m_0)(\theta_0^\top x)\}. \end{aligned}$$

We will now show that

$$D_n := \frac{1}{|\check{\theta}_n - \theta_0|^2} P_X |A_n(X) - B_n(X)|^2 = o_p(1). \quad (\text{F.21})$$

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 [15]. We showed in Theorem 3.2, that  $\{\check{m}_n, \check{\theta}_n\}$  satisfies (F.16) in probability. Thus by another application of Theorem 2.3.2 of [15], 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 (F.16) almost surely. Thus by Lemma F.11, we have  $D_{n_{k_l}} \xrightarrow{a.s.} 0$ . Thus  $D_n = o_p(1)$ .

We will now use (F.21) to find the rate of convergence of  $\{\tilde{m}_n, \check{\theta}_n\}$ . We first find an upper bound for  $P_X|B_n(X)|^2$ . By a simple application of triangle inequality and (F.21), 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 3.1, we have that  $P_X|A_n(X)|^2 = O_p(n^{-4/5})$ . Thus we have

$$P_X|m'_0(\theta_0^\top X)X^\top(\check{\theta}_n - \theta_0) + (\tilde{m}_n - m_0)(\theta_0^\top X)|^2 \leq O_p(n^{-4/5}) + o_p(|\check{\theta}_n - \theta_0|^2).$$

Now define

$$g_1(x) := m'_0(\theta_0^\top x)x^\top(\check{\theta}_n - \theta_0) \text{ and } g_2(x) := (\tilde{m}_n - m_0)(\theta_0^\top x) \quad (\text{F.22})$$

and note that by assumption (A3) there exists a  $\lambda_1 > 0$  such that

$$P_X g_1^2 = (\check{\theta}_n - \theta_0)^\top P_X [X X^\top |m'_0(\theta_0^\top X)|^2] (\check{\theta}_n - \theta_0) \geq \lambda_1 |\check{\theta}_n - \theta_0|^2. \quad (\text{F.23})$$

With (F.23) in mind, we can see that proof of this theorem will be complete if we can show that

$$P_X g_1^2 + P_X g_2^2 \lesssim P_X |m'_0(\theta_0^\top X)X^\top(\check{\theta}_n - \theta_0) + (\tilde{m}_n - m_0)(\theta_0^\top X)|^2. \quad (\text{F.24})$$

The following lemma from [41] gives a sufficient condition for (F.24).

**Lemma F.12** (Lemma 5.7 of [41]). *Let  $g_1$  and  $g_2$  be measurable functions such that  $(\mathbb{P}g_1 g_2)^2 \leq c \mathbb{P}g_1^2 \mathbb{P}g_2^2$  for a constant  $c < 1$ . Then*

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

We now show that  $g_1$  and  $g_2$  (defined in (F.22)) satisfy the condition of the above lemma. Observe that

$$\begin{aligned} P_X [g_1(X)g_2(X)]^2 &= P_X |m'_0(\theta_0^\top X)g_2(X)E(X^\top(\check{\theta} - \theta_0)|\theta_0^\top X)|^2 \\ &\leq P_X [\{m'_0(\theta_0^\top X)\}^2 E[X^\top(\check{\theta} - \theta_0)|\theta_0^\top X]] P_X g_2^2(X) \\ &< P_X [\{m'_0(\theta_0^\top X)\}^2 E[X^\top(\check{\theta} - \theta_0)\}^2 |\theta_0^\top X|] P_X g_2^2(X) \\ &= P_X [\mathbb{E}[\{m'_0(\theta_0^\top X)X^\top(\check{\theta} - \theta_0)\}^2 |\theta_0^\top X|]] P_X g_2^2(X) \\ &= P_X [m'_0(\theta_0^\top X)X^\top(\check{\theta} - \theta_0)]^2 P_X g_2^2(X) \\ &= P_X g_1^2 P_X g_2^2. \end{aligned}$$

The 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.

### F.6. Proof of Theorem 3.4

We first show (3.1). Let  $\delta_n$  be a sequence of positive numbers decreasing to 0. Let  $a, b \in \mathbb{R}$  such that  $D_0 = [a, b]$ . Define  $C_n := [a + 2\delta_n, b - 2\delta_n]$ . By (A5),  $f_{\theta_0^\top X}$ , the density of  $\theta_0^\top X$  is bounded away from 0 and  $\infty$ . Let  $K$  and  $K'$  denote the minimum and the maximum of  $f_{\theta_0^\top X}(\cdot)$  in  $D_0$ . Also, let  $\kappa$  denote the bound on  $m''_0(t)$  over  $t \in D_0$ . As the  $\tilde{m}$  is a convex function, we have

$$\frac{\tilde{m}(t) - \tilde{m}(t - \delta_n)}{\delta_n} \leq \tilde{m}'(t-) \leq \tilde{m}'(t+) \leq \frac{\tilde{m}(t + \delta_n) - \tilde{m}(t)}{\delta_n},$$

for all  $t \in C_n$ , where  $\tilde{m}'(t+)$  and  $\tilde{m}'(t-)$  denote the right and left derivatives of  $\tilde{m}$  at  $t$ , respectively. Observe that

$$\begin{aligned}
& \int_{t \in C_n} \left[ \frac{\tilde{m}(t + \delta_n) - \tilde{m}(t)}{\delta_n} - \frac{m_0(t + \delta_n) - m_0(t)}{\delta_n} \right]^2 dt \\
&= \frac{2}{\delta_n^2} \int_{t \in C_n} \{\tilde{m}(t + \delta_n) - m_0(t + \delta_n)\}^2 dt + \frac{2}{\delta_n^2} \int_{t \in C_n} \{\tilde{m}(t) - m_0(t)\}^2 dt \\
&= \frac{2}{\delta_n^2} \int_{t \in [a+3\delta_n, b-\delta_n]} \{\tilde{m}(t) - m_0(t)\}^2 dt + \frac{2}{\delta_n^2} \int_{t \in C_n} \{\tilde{m}(t) - m_0(t)\}^2 dt \\
&\leq \frac{2}{K\delta_n^2} \int_{t \in [a+3\delta_n, b-\delta_n]} \{\tilde{m}(t) - m_0(t)\}^2 f_{\theta_0^\top X}(t) dt + \frac{2}{K\delta_n^2} \int_{t \in C_n} \{\tilde{m}(t) - m_0(t)\}^2 f_{\theta_0^\top X}(t) dt \\
&= \frac{1}{\delta_n^2} O_p(n^{-4/5}),
\end{aligned} \tag{F.25}$$

where the last equality follows from Theorem 3.7. Similarly, it can be shown that

$$\int_{t \in C_n} \left[ \frac{\tilde{m}(t) - \tilde{m}(t - \delta_n)}{\delta_n} - \frac{m_0(t) - m_0(t - \delta_n)}{\delta_n} \right]^2 dt = \frac{1}{\delta_n^2} O_p(n^{-4/5}). \tag{F.26}$$

Now observe that, as  $\kappa \geq \|m_0''\|_{D_0}$ , we have

$$\begin{aligned}
\Delta_n^+(t) &:= \left[ \frac{\tilde{m}(t + \delta_n) - \tilde{m}(t)}{\delta_n} - \frac{m_0(t + \delta_n) - m_0(t)}{\delta_n} \right] \geq \tilde{m}'(t+) - m_0'(x_{t_n}) \\
&\geq \tilde{m}'(t+) - m_0'(t) + m_0'(t) - m_0'(x_{t_n}) \\
&\geq \tilde{m}'(t+) - m_0'(t) - \kappa\delta_n,
\end{aligned}$$

where  $x_{t_n}$  lies between  $t$  and  $t + \delta_n$ . Moreover,

$$\begin{aligned}
\Delta_n^-(t) &:= \left[ \frac{\tilde{m}(t) - \tilde{m}(t - \delta_n)}{\delta_n} - \frac{m_0(t) - m_0(t - \delta_n)}{\delta_n} \right] \leq \tilde{m}'(t+) - m_0'(x'_{t_n}) \\
&\leq \tilde{m}'(t+) - m_0'(t) + m_0'(t) - m_0'(x'_{t_n}) \\
&\leq \tilde{m}'(t+) - m_0'(t) + \kappa\delta_n,
\end{aligned}$$

where  $x'_{t_n}$  lies between  $t - \delta_n$  and  $t$ . Combining the above two results, we have

$$\Delta_n^-(t) - \kappa\delta_n \leq \tilde{m}'(t+) - m_0'(t) \leq \Delta_n^+(t) + \kappa\delta_n;$$

see proof of Corollary 1 of [14] for a similar inequality. Thus for every  $t \in C_n$ , we have  $[\tilde{m}'(t+) - m_0'(t)]^2 \leq 2\kappa^2\delta_n^2 + 2 \max\{[\Delta_n^-(t)]^2, [\Delta_n^+(t)]^2\}$ . By (F.25) and (F.26), we have

$$\int_{t \in C_n} [\tilde{m}'(t+) - m_0'(t)]^2 f_{\theta_0^\top X}(t) dt \leq 2\kappa^2\delta_n^2 + \frac{1}{\delta_n^2} O_p(n^{-4/5}), \tag{F.27}$$

as

$$\begin{aligned}
& \int_{t \in C_n} \max\{[\Delta_n^-(t)]^2, [\Delta_n^+(t)]^2\} f_{\theta_0^\top X}(t) dt \\
&\leq \int_{t \in C_n} \{\Delta_n^-(t)\}^2 f_{\theta_0^\top X}(t) dt + \int_{t \in C_n} \{\Delta_n^+(t)\}^2 f_{\theta_0^\top X}(t) dt \\
&\leq K' \int_{t \in C_n} \{\Delta_n^-(t)\}^2 dt + K' \int_{t \in C_n} \{\Delta_n^+(t)\}^2 dt \\
&= \frac{1}{\delta_n^2} O_p(n^{-4/5}),
\end{aligned}$$

where  $K'$  is an upper bound on  $f_{\theta^\top X}(\cdot)$  in  $D_0$ . Moreover, note that  $\|\tilde{m}'\|_\infty \leq L$  and  $\|m'_0\|_\infty \leq L_0 \leq L$ . Thus

$$\begin{aligned} \int_{t \in D_0} \{\tilde{m}'(t+) - m'_0(t)\}^2 f_{\theta_0^\top X}(t) dt &= \int_{t \in C_n} \{\tilde{m}'(t+) - m'_0(t)\}^2 f_{\theta_0^\top X}(t) dt \\ &\quad + \int_{t \in D_0 \cap C_n^c} \{\tilde{m}'(t+) - m'_0(t)\}^2 f_{\theta_0^\top X}(t) dt \\ &= 2\kappa^2 \delta_n^2 + \frac{1}{\delta_n^2} O_p(n^{-4/5}) + 4L^2 P(\theta_0^\top X \in D_0 \cap C_n^c) \\ &\leq 2\kappa^2 \delta_n^2 + \frac{1}{\delta_n^2} O_p(n^{-4/5}) + 16K' L^2 \delta_n. \end{aligned}$$

The tightest upper bound for the left hand side of the above display is achieved when  $\delta_n = n^{-4/15}$ . With this choice of  $\delta_n$ , we have

$$\int_{t \in D_0} \{\tilde{m}'(t+) - m'_0(t)\}^2 f_{\theta_0^\top X}(t) dt \leq 2\kappa^2 n^{-8/15} + O_p(n^{-4/15}) + 16K' L^2 n^{-4/15} = O_p(n^{-4/15}).$$

We can find a similar upper bound for  $\int_{t \in D_0} \{\tilde{m}'(t+) - m'_0(t)\}^2 dt$ :

$$\begin{aligned} \int_{t \in D_0} \{\tilde{m}'(t+) - m'_0(t)\}^2 dt &= \int_{t \in C_n} \{\tilde{m}'(t+) - m'_0(t)\}^2 dt + \int_{t \in D_0 \cap C_n^c} \{\tilde{m}'(t+) - m'_0(t)\}^2 dt \quad (\text{F.28}) \\ &\leq 2\frac{\kappa^2 \delta_n^2}{K} + \frac{1}{K \delta_n^2} O_p(n^{-4/5}) + 16L^2 \delta_n, \end{aligned}$$

where the first term on the right hand side of (F.28) is bounded above via (F.27) and  $K$  is the minimum of  $f_{\theta_0^\top X}$ . When  $\delta_n = n^{-4/15}$ , we have

$$\int_{t \in D_0} \{\tilde{m}'(t+) - m'_0(t)\}^2 dt \leq 2\kappa^2 n^{-8/15} + O_p(n^{-4/15}) + 16L^2 n^{-4/15} = O_p(n^{-4/15}). \quad (\text{F.29})$$

Recall that by assumption (A5), we have

$$\sup_{|\theta - \theta_0| \leq n^{-2/15}} \|f_{\theta^\top X}\|_D \leq c < \infty,$$

for large enough  $n$ . Now the proof of (3.2) easily follows from (F.28), since

$$\begin{aligned} \sup_{|\theta - \theta_0| \leq n^{-2/15}} \int_{t \in D_0} \{\tilde{m}'(t+) - m'_0(t)\}^2 f_{\theta^\top X}(t) dt &\leq \sup_{|\theta - \theta_0| \leq n^{-2/15}} \|f_{\theta^\top X}\|_D \int_{t \in D_0} \{\tilde{m}'(t+) - m'_0(t)\}^2 dt \\ &\leq O_p(n^{-4/15}), \end{aligned}$$

where the last inequality follows from (F.29). Note that the upper bound in (F.29) is independent of  $\theta$ . The proof of (3.3) follows from (3.2) as  $|\tilde{\theta} - \theta_0| = o_p(n^{-2/15})$ .

## Appendix G: Proofs of results in Section 3.2

### G.1. Discussion for Proofs of Theorems 3.5, 3.6, and 3.7

Proof of Theorem 3.5 is almost identical to the proof of Theorem 2 of [32]. They propose following estimator for  $(m_0, \theta_0)$  in a similar single index model:

$$(m^{kp}, \theta^{kp}) := \arg \min_{(m, \theta) \in \mathcal{S} \times \Theta} \mathcal{L}_n(m, \theta; \lambda),$$

where  $\mathcal{S}$  is defined in Section 2.1. The single index model in [32] does not assume any shape constraint on  $m_0$  and  $m^{kp}$  is a (possibly non-convex) cubic spline. Under assumptions (A1)–(A4), they prove that  $(m^{kp}, \theta^{kp})$  satisfies the properties of Theorem 3.5. The only modification in the proof of Theorem 2 of [32] needed for it

to be applicable to  $(\hat{m}, \hat{\theta})$  is in the definition of  $\mathcal{B}_C$  and Lemma 8 of [32]. For our purposes, we can redefine  $\mathcal{B}_C$  as follow:

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

The rest of the proof of Theorem 2 of [32] will follow if we can show that

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

The proof of (G.1) follows from Lemma 8 of [32]. Proofs of Theorems 3.6 and 3.7 are identical to the proof of Theorems 3 and 4, respectively.

### G.2. Proof of Theorem 3.8

We use the following interpolation inequality in [1] to prove this theorem.

**Lemma G.1.** (Corollary 3.1, [1]) *Let  $f : \mathbb{R} \rightarrow \mathbb{R}$  be a 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  $g : \mathbb{R} \rightarrow \mathbb{R}$  be a continuous density function and is bounded away from 0, i.e.,  $g(s) > \delta > 0$  for all  $s \in (a, b)$ . If  $0 < \varepsilon \leq 1$ , then*

$$\int_a^b |f'(s)|^2 g(s) ds \leq \gamma \left[ \varepsilon \int_a^b |f''(s)|^2 g(s) ds + \varepsilon^{-1} \int_a^b |f(s)|^2 g(s) ds \right],$$

where  $\gamma$  depends only on  $\delta, a, b$ , and  $\max_{s \in (a, b)} g(s)$ .

Take  $g$  to be the density of  $\theta_0^\top X$  with respect to the Lebesgue measure. By assumption (A5), we have that  $g$  is continuous and bounded away from zero on the bounded set  $D_0$ . Furthermore, let  $f = \hat{m} - m_0$ . By assumption (P1), 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 [16]. Thus by an easy application of the Lemma G.1, we have that

$$\|\hat{m}' \circ \theta_0 - m_0' \circ \theta_0\|^2 \leq \gamma [\varepsilon \|\hat{m}'' \circ \theta_0 - m_0'' \circ \theta_0\|^2 + \varepsilon^{-1} \|\hat{m} \circ \theta_0 - m_0 \circ \theta_0\|^2].$$

By Theorem 3.6, we have that  $J(\hat{m}) = O_p(1)$ . Because  $g$  is bounded away from both zero and infinity, we have that  $\int_{D_0} \hat{m}''(s)^2 g(s) ds \lesssim J(\hat{m})$  and  $\int_{D_0} m_0''(s)^2 g(s) ds \lesssim J(m_0)$ . Fixing  $\varepsilon = \hat{\lambda}_n$ , by Theorem 3.6, we have

$$\|\hat{m}' \circ \theta_0 - m_0' \circ \theta_0\|^2 \leq \gamma [\hat{\lambda}_n (J^2(\hat{m}) + J^2(m_0)) + \hat{\lambda}_n^{-1} O_p(\hat{\lambda}_n^2)] = O_p(\hat{\lambda}_n).$$

## Appendix H: Proofs of results in Appendix A

**Remark H.1** (Quadratic mean differentiability). *If the errors are Gaussian random variables then in the following, we show that the model is quadratic mean differentiable in  $\theta$ . The proof of quadratic mean differentiability for any error distribution satisfying assumption (B3) follows similarly. Under Gaussian error, the density of  $(Y, X)$  is*

$$f_{\theta, m}(y, x) = \exp \left( -\frac{1}{2\sigma^2} (y - m(\theta^\top x))^2 \right) p_X(x).$$

Note that  $\theta \mapsto f_{\theta, m}(y, x)$  is differentiable a.e.  $(y, x)$ . Define

$$\eta(y, x, \theta, m) = \begin{cases} \frac{1}{2} f'_{\theta, m}(y, x) f_{\theta, m}^{-1/2}(y, x), & f_{\theta, m}(y, x) > 0 \text{ and } f'_{\theta, m}(y, x) \text{ exists,} \\ 0 & \text{otherwise,} \end{cases}$$

where  $f'_{\theta, m}(y, x)$  denotes the derivative with respect to  $\theta$ . Hájek [20] proved that the family of distributions is quadratic mean differentiable (q.m.d) at  $\theta_0$  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  $\theta$ . In the following we prove that  $I_{i,j}(\theta)$  is finite and continuous at  $\theta$ . Observe that,

$$\begin{aligned} I_{i,j}(\theta) &= \int_{\mathcal{X} \times \mathbb{R}} \eta_i(y, x, \theta, m) \eta_j(y, x, \theta, m) dP_X dy \\ &= \int_{\mathcal{X} \times \mathbb{R}} (y - m(\theta^\top x))^2 [m'(\theta^\top x)]^2 x_i x_j \exp\left(-\frac{1}{2\sigma^2}(y - m(\theta^\top x))^2\right) dP_X(x) dy \\ &= P_{\theta, m}[(Y - m(\theta^\top X))^2 [m'(\theta^\top X)]^2 X_i X_j] \\ &= P_{\theta, m}[E[(Y - m(\theta^\top X))^2 [m'(\theta^\top X)]^2 X_i X_j | \theta^\top X]] \\ &= P_{\theta, m}[m'(\theta^\top X)^2 E[X_i X_j | \theta^\top X]]. \end{aligned}$$

As both  $m(\cdot)$  and  $E[X_i X_j | \theta^\top X = \cdot]$  are bounded functions, we have that  $I_{i,j}(\theta)$  is finite and continuous at  $\theta_0$ . Thus, the model is differentiable in quadratic mean in  $\theta$ .

### H.1. Proof of Theorem A.1

We will first show that  $\xi_t(u; \theta, \eta, m)$  is a valid submodel. Let us define

$$\varphi_{\theta, \eta, t}(u) := \phi_{\theta, \eta, t}(u + (\theta - \zeta_t(\theta, \eta))^\top k(u)). \quad (\text{H.1})$$

Note that to prove that  $\xi_t(u; \theta, \eta, m)$  is a convex function it is enough to show that  $\nabla_t(\cdot; \theta, \eta, m)$  is an increasing function. First observe that

$$k'(u) = 2h'_{\theta_0}(u) - \frac{m'_0(u)m''_0(u)}{(m''_0(u))^2}h'_{\theta_0}(u) + \frac{m'_0(u)}{m''_0(u)}h''_{\theta_0}(u). \quad (\text{H.2})$$

Hence, by assumptions **(P1)**, **(B1)**, and **(B2)**, we can find  $M^*$  such that  $\|k\|_D^S \leq M^*$ . Thus  $u \mapsto u + (\theta - \zeta_t(\theta, \eta))^\top k(u)$  is a strictly increasing function for  $t$  in neighborhood of zero as  $k$  is a Lipschitz function on  $D$ ; see **(H.2)**. As  $\phi_{\theta, \eta, t}(\cdot)$  is a strictly increasing function for  $t$  sufficiently close to zero. It now follows that  $u \mapsto \varphi_{\theta, \eta, t}(u)$  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

$$\nabla_t(u; \theta, \eta, m) = m' \circ \varphi_{\theta, \eta, t}(u).$$

Thus we have that  $\nabla_t(\cdot; \theta, \eta, m)$  is an increasing function for  $t \in \mathbb{R}$  is close enough to 0. Next we show that  $\xi_t(u; \theta, \eta, m) = m(u)$  when  $t = 0$ . By definition we have

$$\xi_t(s^\top x; \theta, \eta, m) = \int_{s_0}^{s^\top x} m' \circ \varphi_{\theta, \eta, t}(u) dy + (\zeta_t(\theta, \eta) - \theta)^\top [(m'_0(s_0) - m'(s_0))k(s_0) - m'_0(s_0)h_{\theta_0}(s_0)] + m(s_0).$$

We have that  $\varphi_{\theta, \eta, 0}(u) = u$ ,  $\forall u \in D$ . Hence,

$$\xi_0(\zeta_0(\theta, \eta)^\top x; \theta, \eta, m) = \int_{s_0}^{\theta^\top x} m' \circ \varphi_{\theta, \eta, 0}(y) dy + m(s_0) = \int_{s_0}^{\theta^\top x} m'(y) dy + m(s_0) = m(\theta^\top x)$$

and  $\zeta_0(\theta, \eta) = \theta$  for all  $\eta \in S^{d-2}$ . Now we show that  $J^2(\xi_t(\cdot; \theta, \eta, m)) < \infty$ . Observe that

$$\begin{aligned} J^2(\xi_t(\cdot; \theta, \eta, m)) &= \int_D \{\xi_t''(u; \theta, \eta, m)(u)\}^2 du \\ &= \int_{D_r} \left[ \frac{\partial}{\partial u} \nabla_t(u; \theta, \eta, m) \right]^2 du \\ &= \int_D \{m'' \circ \varphi_{\theta, \eta, t}(u) \varphi'_{\theta, \eta, t}(u)\}^2 du \\ &= \int_D \{m''(u)\}^2 \varphi'_{\theta, \eta, t} \circ \varphi_{\theta, \eta, t}^{-1}(u) du \end{aligned}$$

where  $\varphi_{\theta,\eta,t}$  is defined in (H.1) and  $\varphi'_{\theta,\eta,t}(u) := \frac{\partial}{\partial u}\varphi_{\theta,\eta,t}(u)$ . Thus, we have that  $J^2(\xi_t(\cdot; \theta, \eta, m)) < \infty$  whenever  $J(m) < \infty$ .

Next we compute  $\partial \Upsilon_t(u; \theta, \eta, m)/\partial t$  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

$$\begin{aligned} \frac{\partial}{\partial t} \Upsilon_t(u; \theta, \eta, m) &= \frac{\partial}{\partial t} m' \circ \phi_{\theta,\eta,t}(u + (\theta - \zeta_t(\theta, \eta))^\top k(u)) \\ &= m'' \circ \phi_{\theta,\eta,t}(u + (\theta - \zeta_t(\theta, \eta))^\top k(u)) \left[ \dot{\phi}_{\theta,\eta,t}(u + (\theta - \zeta_t(\theta, \eta))^\top k(u)) \right. \\ &\quad \left. - \phi'_{\theta,\eta,t}(u + (\theta - \zeta_t(\theta, \eta))^\top k(u)) \frac{\partial \zeta_t(\theta, \eta)}{\partial t}^\top k(u) \right], \end{aligned}$$

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,

$$\begin{aligned} &\frac{\partial}{\partial t} \xi_t(\zeta_t(\theta, \eta)^\top x; \theta, \eta, m) \\ &= \frac{\partial}{\partial t} \left\{ \int_{s_0}^{\zeta_t(\theta, \eta)^\top x} \Upsilon_t(y; \theta, \eta, m) dy \right\} + \frac{\partial \zeta_t(\theta, \eta)}{\partial t}^\top \left[ (m'_0(s_0) - m'(s_0))k(s_0) - m'_0(s_0)h_{\theta_0}(s_0) \right] \\ &= \Upsilon_t(\zeta_t(\theta, \eta)^\top x; \theta, \eta, m) \frac{\partial \zeta_t(\theta, \eta)}{\partial t}^\top x + \int_{s_0}^{\zeta_t(\theta, \eta)^\top x} \frac{\partial \Upsilon_t(y; \theta, \eta, m)}{\partial t} dy \\ &\quad + \frac{\partial \zeta_t(\theta, \eta)}{\partial t}^\top \left[ (m'_0(s_0) - m'(s_0))k(s_0) - m'_0(s_0)h_{\theta_0}(s_0) \right] \\ &= \frac{\partial \zeta_t(\theta, \eta)}{\partial t}^\top \left[ \Upsilon_t(\zeta_t(\theta, \eta)^\top x; \theta, \eta, m)x + (m'_0(s_0) - m'(s_0))k(s_0) - m'_0(s_0)h_{\theta_0}(s_0) \right] \\ &\quad + \int_{s_0}^{\zeta_t(\theta, \eta)^\top x} m'' \circ \phi_{\theta,\eta,t}(u + (\theta - \zeta_t(\theta, \eta))^\top k(u)) \left[ \dot{\phi}_{\theta,\eta,t}(u + (\theta - \zeta_t(\theta, \eta))^\top k(u)) \right. \\ &\quad \left. - \phi'_{\theta,\eta,t}(u + (\theta - \zeta_t(\theta, \eta))^\top k(u)) \frac{\partial \zeta_t(\theta, \eta)}{\partial t}^\top k(u) \right] du. \end{aligned}$$

The interchange of derivative and the integral is possible by assumptions (P1), (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 (A.3)) and  $\partial \zeta_t(\theta, \eta)/\partial t = -|\eta|^2 t / \sqrt{1 - t^2 |\eta|^2} \theta + H_\theta \eta$ , we get

$$\begin{aligned} &\left. \frac{\partial}{\partial t} \xi_t(\zeta_t(\theta, \eta)^\top x; \theta, \eta, m) \right|_{t=0} \\ &= \eta^\top H_\theta^\top [m'(\theta^\top x) + (m'_0(s_0) - m'(s_0))k(s_0) - m'_0(s_0)h_{\theta_0}(s_0)] \\ &\quad - \eta^\top H_\theta^\top \int_{s_0}^{\theta^\top x} m''(u)k(u) du, \end{aligned}$$

and

$$\begin{aligned} &-\frac{1}{2} \frac{\partial}{\partial t} [y - \xi_t(\zeta_t(\theta, \eta)^\top x; \theta, \eta, m)]^2 \Big|_{t=0} \\ &= (y - m(\theta^\top x)) \eta^\top H_\theta^\top \left[ m'(\theta^\top x) + \int_{s_0}^{\theta^\top x} m''(u)[-k(u)] du \right. \\ &\quad \left. + (m'_0(s_0) - m'(s_0))k(s_0) - m'_0(s_0)h_{\theta_0}(s_0) \right] \\ &= (y - m(\theta^\top x)) \eta^\top H_\theta^\top \left[ m'(\theta^\top x) + \int_{s_0}^{\theta^\top x} m'(u)k'(u) du - m'(\theta^\top x)k(\theta^\top x) \right. \\ &\quad \left. + m'(s_0)k(s_0) + (m'_0(s_0) - m'(s_0))k(s_0) - m'_0(s_0)h_{\theta_0}(s_0) \right] \\ &= \eta^\top \mathfrak{S}_{\theta,m}(x, y). \end{aligned} \tag{H.3}$$

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

$$t \mapsto \frac{1}{n} \sum_{i=1}^n (y_i - \xi_t(\zeta_t(\hat{\theta}, \eta)^\top x; \hat{\theta}, \eta, \hat{m}))^2 + \hat{\lambda}_n^2 \int_D \{\xi_t(\cdot; \hat{\theta}, \eta, \hat{m})''(u)\}^2 du$$

is minimized at  $t = 0$  for every  $\eta \in S^{d-2}$ . Observe that (H.3) and the fact that  $J^2(\xi_t(\cdot; \theta, \eta, m))$  is differentiable imply that the above function is differentiable in  $t$  on a small neighborhood of 0 (which depends on  $\eta$ ). Hence, we have that

$$\mathbb{P}_n \mathfrak{S}_{\hat{\theta}, \hat{m}} - \frac{\hat{\lambda}_n^2}{2} \frac{\partial J^2(\xi_t(\cdot; \hat{\theta}, \eta, \hat{m}))}{\partial t} \Big|_{t=0} = 0. \quad (\text{H.4})$$

Moreover, after some tedious algebra it can be seen that

$$\frac{\partial}{\partial t} J^2(\xi_t(\cdot; \theta, \eta, m)) \Big|_{t=\theta} \leq C \int_{D_\theta} k'(p) \{m''(p)\}^2 dp, \quad (\text{H.5})$$

where  $C$  is constant independent of  $\eta$ ,  $k$  is defined in (4.17). Thus by Theorem 3.5, (4.18), and (H.5), we have

$$\frac{\partial}{\partial t} J^2(\xi_t(\cdot; \hat{\theta}, \eta, \hat{m})) \Big|_{t=0} \leq M^* J(\hat{m}) = O_p(1).$$

Finally, (H.4) and (P2) imply  $\mathbb{P}_n \mathfrak{S}_{\hat{\theta}, \hat{m}} = o_p(n^{-1/2})$ .

Next, we show that  $\mathfrak{S}_{\theta_0, m_0} = \ell_{\theta_0, m_0}$ . By definition, it is enough to show that,

$$\begin{aligned} & m'_0(\theta_0^\top x)(x - k(\theta_0^\top x)) + m'_0(s_0)k(s_0) + \int_{s_0}^{\theta_0^\top x} m'_0(u)k'(u)du - m'_0(s_0)h_{\theta_0}(s_0) = m'_0(\theta_0^\top x)(x - h_{\theta_0}(\theta_0^\top x)) \\ \Leftrightarrow & m'_0(\theta_0^\top x)k(\theta_0^\top x) - m'_0(s_0)k(s_0) - \int_{s_0}^{\theta_0^\top x} m'_0(u)k'(u)du + m'_0(s_0)h_{\theta_0}(s_0) = m'_0(\theta_0^\top x)h_{\theta_0}(\theta_0^\top x) \\ \Leftrightarrow & \int_{s_0}^{\theta_0^\top x} \frac{\partial [m'_0(u)k(u)]}{\partial u} du - \int_{s_0}^{\theta_0^\top x} m'_0(u)k'(u)du + m'_0(s_0)h_{\theta_0}(s_0) = m'_0(\theta_0^\top x)h_{\theta_0}(\theta_0^\top x) \\ \Leftrightarrow & \int_{s_0}^{\theta_0^\top x} m''_0(u)k(u)du + m'_0(s_0)h_{\theta_0}(s_0) = m'_0(\theta_0^\top x)h_{\theta_0}(\theta_0^\top x), \end{aligned} \quad (\text{H.6})$$

which is true by (4.17). 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.

## H.2. Proof of Lemma A.1

From the definitions of  $\mathfrak{S}_{\theta, m}$  and  $\psi_{\theta, m}$ , we have

$$\begin{aligned} & \mathfrak{S}_{\theta, m}(x, y) - \psi_{\theta, m}(x, y) \\ &= \{y - m(\theta^\top x)\} H_\theta^\top \left[ \int_{s_0}^{\theta^\top x} m'(u)k'(u)du - m'(\theta^\top x)k(\theta^\top x) + m'_0(s_0)k(s_0) \right. \\ & \quad \left. - m'_0(s_0)h_{\theta_0}(s_0) + (m'_0 h_{\theta_0})(\theta^\top x) \right]. \end{aligned}$$

Observe that

$$\begin{aligned}
& \int_{s_0}^{\theta^\top x} m'(u)k'(u)du - m'(\theta^\top x)k(\theta^\top x) + m'_0(s_0)k(s_0) - m'_0(s_0)h_{\theta_0}(s_0) + (m'_0 h_{\theta_0})(\theta^\top x) \\
&= \int_{s_0}^{\theta^\top x} m'(u)k'(u)du - \int_{s_0}^{\theta^\top x} m'_0(u)k'(u)du + \int_{s_0}^{\theta^\top x} m'_0(u)k'(u)du - m'(\theta^\top x)k(\theta^\top x) \\
&\quad + m'_0(s_0)k(s_0) - m'_0(s_0)h_{\theta_0}(s_0) + (m'_0 h_{\theta_0})(\theta^\top x) \\
&= \int_{s_0}^{\theta^\top x} \{m'(u) - m'_0(u)\}k'(u)du + \int_{s_0}^{\theta^\top x} m'_0(u)k'(u)du - m'(\theta^\top x)k(\theta^\top x) \\
&\quad + m'_0(s_0)k(s_0) + (m'_0 h_{\theta_0})(\theta^\top x) - (m'_0 h_{\theta_0})(s_0). \tag{H.7}
\end{aligned}$$

We now analyze the terms in the right hand side of the above display. First observe that

$$\begin{aligned}
& \int_{s_0}^{\theta^\top x} m'_0(u)k'(u)du - m'(\theta^\top x)k(\theta^\top x) + m'_0(s_0)k(s_0) \\
&= m'_0(u)k(u)\Big|_{s_0}^{\theta^\top x} - \int_{s_0}^{\theta^\top x} m''_0(u)k(u)du - m'(\theta^\top x)k(\theta^\top x) + m'_0(s_0)k(s_0) \tag{H.8} \\
&= - \int_{s_0}^{\theta^\top x} m''_0(u)k(u)du + (m'_0(\theta^\top x) - m'(\theta^\top x))k(\theta^\top x).
\end{aligned}$$

Finally, by definition (4.17) and integration by parts, we have

$$\int_{s_0}^{\theta^\top x} m''_0(u)k(u)du = \int_{s_0}^{\theta^\top x} [m''_0(u)h_{\theta_0}(u) + m'_0(u)h'_{\theta_0}(u)]du = m'_0(u)h_{\theta_0}(u)\Big|_{s_0}^{\theta^\top x}. \tag{H.9}$$

By substituting (H.8) and (H.9) in (H.7), 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  $\sqrt{n}\mathbb{P}_n[(Y - \hat{m}(\hat{\theta}^\top X))U_{\hat{\theta}, \hat{m}}(X)]$ :

$$\begin{aligned}
& |\sqrt{n}\mathbb{P}_n[(Y - \hat{m}(\hat{\theta}^\top X))U_{\hat{\theta}, \hat{m}}(X)]| \\
&= |\sqrt{n}\mathbb{P}_n[(m_0(\theta_0^\top X) - \hat{m}(\hat{\theta}^\top X))U_{\hat{\theta}, \hat{m}}(X)] + \sqrt{n}\mathbb{P}_n\epsilon U_{\hat{\theta}, \hat{m}}(X)| \\
&\leq |\sqrt{n}\mathbb{P}_n[(m_0 - \hat{m})(\theta_0^\top X)U_{\hat{\theta}, \hat{m}}(X)]| + |\sqrt{n}\mathbb{P}_n[(\hat{m}(\theta_0^\top X) - \hat{m}(\hat{\theta}^\top X))U_{\hat{\theta}, \hat{m}}(X)]| \\
&\quad + |\sqrt{n}\mathbb{P}_n\epsilon U_{\hat{\theta}, \hat{m}}(X)| \\
&\leq |\mathbb{G}_n[(m_0 - \hat{m})(\theta_0^\top X)U_{\hat{\theta}, \hat{m}}(X)]| + |\mathbb{G}_n[(\hat{m}(\theta_0^\top X) - \hat{m}(\hat{\theta}^\top X))U_{\hat{\theta}, \hat{m}}(X)]| \\
&\quad + |\sqrt{n}\mathbb{P}_n\epsilon U_{\hat{\theta}, \hat{m}}(X)| + \sqrt{n}|P_{\theta_0, m_0}[(\hat{m}(\theta_0^\top X) - \hat{m}(\hat{\theta}^\top X))U_{\hat{\theta}, \hat{m}}(X)]| \\
&\quad + \sqrt{n}|P_{\theta_0, m_0}[(m_0 - \hat{m})(\theta_0^\top X)U_{\hat{\theta}, \hat{m}}(X)]|.
\end{aligned}$$

### H.3. Proof of Lemma A.2

We will first show that

$$N(\varepsilon, \mathcal{W}_{M_1, M_2, M_3}^*, \|\cdot\|_\infty) \leq c \exp(c/\varepsilon) \varepsilon^{-4d}, \tag{H.10}$$

where  $c$  depends only on  $M_1, M_2$ , and  $M_3$ . By Theorem 2.4 of [49], we have

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

where  $c$  is a constant depending only on  $M_1, M_2$ , and  $M_3$ . Let us denote the functions in the  $\varepsilon$ -cover by  $l_1, \dots, l_t$ . By Lemma 15 of [32], we have that there exists  $\theta_1, \dots, \theta_s$  for  $s \lesssim \varepsilon^{-4d}$  such that  $\{\theta_i\}_{1 \leq i \leq s}$  form an  $\varepsilon^2$ -cover of  $\Theta \cap B_{\theta_0}(1/2)$  and satisfies (86) of [32] (with  $\varepsilon^2$  instead of  $\varepsilon$ ). 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  $\varepsilon$ -cover is  $l_1$  and the vector nearest to  $\theta$  in the  $\varepsilon^2$  cover of  $\Theta \cap B_{\theta_0}(1/2)$  is  $\theta_1$ , 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

$$V_{\theta,m}(x) := \left[ \int_{s_0}^{\theta^\top x} [m'(u) - m'_0(u)] k'(u) du + (m'_0(\theta^\top x) - m'(\theta^\top x)) k(\theta^\top x) \right].$$

Recall that  $U_{\theta,m} = H_\theta^\top V_{\theta,m}$ . Now for every  $x \in \mathcal{X}$ , observe that

$$\begin{aligned} & |U_{\theta,m}(x) - U_{\theta_1,r_1}(x)| \\ & \leq |U_{\theta,m}(x) - U_{\theta,r_1}(x)| + |(H_\theta^\top - H_{\theta_1}^\top) V_{\theta,r_1}| + |H_{\theta_1}^\top (V_{\theta,r_1}(x) - V_{\theta_1,r_1}(x))| \\ & \leq \left| H_\theta^\top \int_{s_0}^{\theta^\top x} [m'(u) - r'_1(u)] k'(u) du \right| + \left| H_\theta^\top (m' - r'_1)(\theta^\top x) k(\theta^\top x) \right| \\ & \quad + 4M^* M_2(T+1) \sqrt{d-1} \varepsilon^2 + \left| H_{\theta_1}^\top [(r'_1 - m'_0)(\theta^\top x) k(\theta^\top x) - (r'_1 - m'_0)(\theta_1^\top x) k(\theta_1^\top x)] \right| \\ & \quad + \left| H_{\theta_1}^\top \int_{\theta_1^\top x}^{\theta^\top x} [r'_1(u) - m'_0(u)] k'(u) du \right| \\ & \leq 2T(1 + |\theta_0|) M^* \|m' - r'_1\|_\infty + M^* \|m' - r'_1\|_\infty + 4M^* M_2(T+1) \sqrt{d-1} \varepsilon^2 \\ & \quad + \left| (r'_1 - m'_0)(\theta^\top x) k(\theta^\top x) - (r'_1 - m'_0)(\theta_1^\top x) k(\theta^\top x) \right| + 2M_2 M^* T |\theta - \theta_1|. \end{aligned}$$

Furthermore, note that

$$\begin{aligned} & \left| (r'_1 - m'_0)(\theta^\top x) k(\theta^\top x) - (r'_1 - m'_0)(\theta_1^\top x) k(\theta_1^\top x) \right| \\ & \leq \left| (r'_1 - m'_0)(\theta^\top x) k(\theta^\top x) - (r'_1 - m'_0)(\theta_1^\top x) k(\theta^\top x) \right| \\ & \quad + \left| (r'_1 - m'_0)(\theta_1^\top x) [k(\theta^\top x) - k(\theta_1^\top x)] \right| \\ & \leq M^* \left| (r'_1 - m'_0)(\theta^\top x) - (r'_1 - m'_0)(\theta_1^\top x) \right| + 2M_2 M^* T |\theta - \theta_1| \\ & \leq 2M_3 M^* T |\theta - \theta_1|^{1/2} + 2M_2 M^* T |\theta - \theta_1|, \end{aligned}$$

where the last inequality in the previous display follows from Lemma E.1. Combining the above two displays, we have

$$\begin{aligned} |U_{\theta,m}(x) - U_{\theta_1,r_1}(x)| & \leq M^* \|m' - r'_1\|_\infty (4T+1) + 4M^* M_2(T+1) \sqrt{d-1} \varepsilon^2 \\ & \quad + 2M_3 M^* T |\theta - \theta_1|^{1/2} + 2M_2 M^* T |\theta - \theta_1| + 2M_2 M^* T |\theta - \theta_1|. \end{aligned}$$

Thus,  $\{U_{\theta_i, l_j}\}$  form an (constant multiple of)  $\varepsilon$ -cover (with respect to  $\|\cdot\|_{2,\infty}$  norm) of  $\mathcal{W}_{M_1, M_2, M_3}^*$ , and we have (H.10). 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  $|H_\theta^\top x| \leq |x|$  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

$$\begin{aligned}
|U_{\theta,m}(x)| &\leq \left| \int_{s_0}^{\theta^\top x} [m'(u) - m'_0(u)]k'(u)du \right| + \left| (m' - m'_0)(\theta^\top x)k(\theta^\top x) \right| \\
&\leq \left| \int_{s_0}^{\theta_0^\top x} [m'(u) - m'_0(u)]k'(u)du \right| + \left| \int_{\theta_0^\top x}^{\theta^\top x} [m'(u) - m'_0(u)]k'(u)du \right| \\
&\quad + \left| (m' - m'_0)(\theta^\top x)k(\theta^\top x) \right| \\
&\leq \sqrt{d-1}M^*(T\|m - m_0\|_{D_0}^S + 2M_2T|\theta - \theta_0| + 2M_3\sqrt{T}|\theta - \theta_0|^{\frac{1}{2}} + \|m - m_0\|_{D_0}^S) \\
&\leq W_{M_1,M_2,M_3}(n).
\end{aligned}$$

Thus,  $W_{M_1,M_2,M_3}(n)$  satisfies (A.9).

#### H.4. Proof of Lemma A.3

For every  $(\theta, m) \in \mathcal{C}_{M_1,M_2,M_3}(n)$ , note that

$$|(m_0 - m)(\theta_0^\top x)U_{\theta,m}(x)| \leq 2M_1|U_{\theta,m}(x)| \leq 2M_1W_{M_1,M_2,M_3}(n) = D_{M_1,M_2,M_3}(n).$$

Furthermore, we have

$$N(\varepsilon, \{(m_0 - m)(\theta_0^\top \cdot) : m \in \mathcal{C}_{M_1,M_2,M_3}^{m*}\}, \|\cdot\|_\infty) = N(\varepsilon, \mathcal{C}_{M_1,M_2,M_3}^{m*}, \|\cdot\|_\infty) < \exp(c/\sqrt{\varepsilon}),$$

where the inequality follows from Theorem 2.4 of [49] and  $c$  is a constant depending only on  $M_1, M_2$ , and  $M_3$ . By Lemma 9.25 of [31] (for entropy of product of uniformly bounded function classes), and Lemma A.2, we have that

$$N(\varepsilon, \mathcal{D}_{M_1,M_2,M_3}^*, \|\cdot\|_{2,\infty}) \leq c\varepsilon^{-4d} \exp\left(\frac{c}{\sqrt{\varepsilon}} + \frac{c}{\varepsilon}\right).$$

Since,  $N(\varepsilon, \mathcal{D}_{M_1,M_2,M_3}(n), \|\cdot\|_{2,\infty}) \leq N(\varepsilon, \mathcal{D}_{M_1,M_2,M_3}^*, \|\cdot\|_{2,\infty})$  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  $J_{[]}(\gamma, \mathcal{D}_{M_1,M_2,M_3}(n), \|\cdot\|_{2,P_{\theta_0,m_0}}) \lesssim c\gamma^{1/2}$ .

Observe that  $f \in \mathcal{D}_{M_1,M_2,M_3}(n)$  maps  $\chi$  to  $\mathbb{R}^{d-1}$ . For any  $f \in \mathcal{D}_{M_1,M_2,M_3}(n)$ , let  $f_1, \dots, f_{d-1}$  denote each of the real valued components, i.e.,  $f(\cdot) := (f_1(\cdot), \dots, f_{d-1}(\cdot))$ . With this notation, we have

$$\mathbb{P}\left(\sup_{f \in \mathcal{D}_{M_1,M_2,M_3}(n)} |\mathbb{G}_n f| > \delta\right) \leq \sum_{i=1}^{d-1} \mathbb{P}\left(\sup_{f \in \mathcal{D}_{M_1,M_2,M_3}(n)} |\mathbb{G}_n f_i| > \delta/\sqrt{d-1}\right). \quad (\text{H.11})$$

We can bound each term in the summation of (H.11) using the maximal inequality in Corollary 19.35 of [52]. We have

$$\begin{aligned}
&\mathbb{P}\left(\sup_{f \in \mathcal{D}_{M_1,M_2,M_3}(n)} |\mathbb{G}_n f| > \delta\right) \\
&\leq \delta^{-1}\sqrt{d-1} \sum_{i=1}^{d-1} \mathbb{E}\left(\sup_{f \in \mathcal{D}_{M_1,M_2,M_3}(n)} |\mathbb{G}_n f_i|\right) \\
&\leq \delta^{-1}d\sqrt{d-1}J_{[]}(\|D_{M_1,M_2,M_3}(n)\|, \mathcal{D}_{M_1,M_2,M_3}(n), \|\cdot\|_{2,P_{\theta_0,m_0}}) \\
&\lesssim \delta^{-1}\|D_{M_1,M_2,M_3}(n)\|^{1/2} \\
&\lesssim \left[\hat{\lambda}_n^{1/4} + \frac{1}{a_n}\right]^{1/2} \rightarrow 0, \quad \text{as } n \rightarrow \infty,
\end{aligned} \quad (\text{H.12})$$

where we have used (A.10) and the fact that  $D_{M_1,M_2,M_3}^2(n)$  is non-random in the last inequality.

### H.5. Proof of Lemma A.4

First, note that for every  $(\theta, m) \in \mathcal{C}_{M_1, M_2, M_3}(n)$ , we have

$$|(m(\theta_0^\top x) - m(\theta^\top x))U_{\theta, m}(x)| \leq 2M_1 |U_{\theta, m}(x)| \leq D_{M_1, M_2, M_3}(n).$$

Observe that the proof of Lemma A.4 will be complete (by arguments similar to the proof of Lemma A.3) if we can show that

$$\log N(\varepsilon, \mathcal{A}_{M_1, M_2, M_3}^*, \|\cdot\|_{2, \infty}) \leq c \exp\left(\frac{c}{\varepsilon} + \frac{c}{\sqrt{\varepsilon}}\right) \varepsilon^{-4d}, \quad (\text{H.13})$$

where the constant  $c$  depends only on  $M_1, M_2, M_3$ , and  $d$ .

However, arguments similar to the proof of Lemma 8 of [32] 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 [31] and Lemma A.2, we have (H.13).

### H.6. Proof of Lemma A.5

Note that, we have

$$\begin{aligned} & \mathbb{P}(|\sqrt{n}\mathbb{P}_n(\epsilon U_{\hat{\theta}, \hat{m}})| > \delta) \\ & \leq \mathbb{P}\left(\sup_{(\theta, m) \in \mathcal{C}_{M_1, M_2, M_3}(n)} |\sqrt{n}\mathbb{P}_n(\epsilon U_{\theta, m})| > \delta\right) + \mathbb{P}((\hat{\theta}, \hat{m}) \notin \mathcal{C}_{M_1, M_2, M_3}(n)) \\ & \leq \mathbb{P}\left(\sup_{f \in \mathcal{W}_{M_1, M_2, M_3}(n)} |\sqrt{n}\mathbb{P}_n \epsilon f| > \delta\right) + \mathbb{P}((\hat{\theta}, \hat{m}) \notin \mathcal{C}_{M_1, M_2, M_3}(n)) \\ & = \mathbb{P}\left(\sup_{f \in \mathcal{W}_{M_1, M_2, M_3}(n)} |\mathbb{G}_n \epsilon f| > \delta\right) + \mathbb{P}((\hat{\theta}, \hat{m}) \notin \mathcal{C}_{M_1, M_2, M_3}(n)), \end{aligned}$$

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

$$\mathbb{P}\left(\sup_{f \in \mathcal{W}_{M_1, M_2, M_3}(n)} |\mathbb{G}_n \epsilon f| > \delta\right) \rightarrow 0, \quad \text{as } n \rightarrow \infty.$$

By Lemma A.2, 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

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

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 the maximal inequality in Corollary 19.35 of [52] and an argument similar to (H.11) and (H.12), we have

$$\begin{aligned} \mathbb{P}\left(\sup_{f \in \mathcal{W}_{M_1, M_2, M_3}(n)} |\mathbb{G}_n \epsilon f| > \delta\right) & \lesssim \delta^{-1} \sqrt{d-1} \sum_{i=1}^{d-1} \mathbb{E}\left(\sup_{f \in \mathcal{W}_{M_1, M_2, M_3}(n)} |\mathbb{G}_n \epsilon f_i|\right) \\ & \lesssim \delta^{-1} J_{[]} \left(P_{\theta_0, m_0}(|\epsilon|^2 W_{M_1, M_2, M_3}^2(n))^{\frac{1}{2}}, \mathcal{W}_{M_1, M_2, M_3}(n), \|\cdot\|_{2, P_{\theta_0, m_0}}\right) \\ & \lesssim \hat{\lambda}_n^{-1/4} + \frac{1}{a_n} \rightarrow 0 \quad \text{as } n \rightarrow \infty, \end{aligned}$$

where in the first inequality  $f_1, \dots, f_{d-1}$  denote each component of  $f$ . Now, we prove the second and third equations in (A.11). First, note that

$$\begin{aligned} |P_{\theta_0, m_0}[(m_0 - \hat{m})(\theta_0^\top X)U_{\hat{\theta}, \hat{m}}(X)]| &\leq \sqrt{P_{\theta_0, m_0}[(m_0 - \hat{m})^2(\theta_0^\top X)]P_{\theta_0, m_0}|U_{\hat{\theta}, \hat{m}}(X)|^2} \\ &= O_p(\hat{\lambda}_n) \left[ P_{\theta_0, m_0}|U_{\hat{\theta}, \hat{m}}(X)|^2 \right]^{1/2}, \end{aligned} \quad (\text{H.14})$$

where the inequality is an application of the Cauchy-Schwarz inequality and the equality is due to Theorem 3.7. Similarly, using Theorems 3.6, 3.7, and the mean value theorem we have

$$\begin{aligned} |P_{\theta_0, m_0}[(\hat{m}(\theta_0^\top X) - \hat{m}(\hat{\theta}^\top X))U_{\hat{\theta}, \hat{m}}(X)]| &\leq \sqrt{P_{\theta_0, m_0}[\hat{m}(\theta_0^\top X) - \hat{m}(\hat{\theta}^\top X)]^2 P_{\theta_0, m_0}|U_{\hat{\theta}, \hat{m}}(X)|^2} \\ &\leq \|\hat{m}'\|_\infty |\hat{\theta} - \theta_0| T \left[ P_{\theta_0, m_0}|U_{\hat{\theta}, \hat{m}}(X)|^2 \right]^{1/2} \\ &= O_p(\hat{\lambda}_n) \left[ P_{\theta_0, m_0}|U_{\hat{\theta}, \hat{m}}(X)|^2 \right]^{1/2}. \end{aligned} \quad (\text{H.15})$$

Now we find an upper bound for  $P_{\theta_0, m_0}|U_{\hat{\theta}, \hat{m}}(X)|^2$ . Note that

$$\begin{aligned} P_{\theta_0, m_0}|U_{\hat{\theta}, \hat{m}}(X)|^2 &\lesssim P_{\theta_0, m_0}|H_{\hat{\theta}}^\top(\hat{m}' - m'_0)(\hat{\theta}^\top X)k(\hat{\theta}^\top X)|^2 \\ &\quad + P_{\theta_0, m_0} \left| H_{\hat{\theta}}^\top \int_{s_0}^{\hat{\theta}^\top X} [\hat{m}'(u) - m'_0(u)]k'(u)du \right|^2 \\ &\leq M^{*2}(d-1)P_{\theta_0, m_0} \left[ (\hat{m}' - m'_0)(\hat{\theta}^\top X) \right]^2 \\ &\quad + P_{\theta_0, m_0} \left[ \int_{s_0}^{\hat{\theta}^\top X} [\hat{m}'(u) - m'_0(u)]^2 du \int_{s_0}^{\hat{\theta}^\top X} |k'(u)|^2 du \right] \\ &\leq M^{*2}(d-1)P_{\theta_0, m_0} \left[ (\hat{m}' - m'_0)(\hat{\theta}^\top X) \right]^2 \\ &\quad + M^{*2}(d-1)TP_{\theta_0, m_0} \left[ \int_{s_0}^{\hat{\theta}^\top X} [\hat{m}'(u) - m'_0(u)]^2 du \right] \\ &\leq M^{*2}(d-1)P_{\theta_0, m_0} \left[ (\hat{m}' - m'_0)(\hat{\theta}^\top X) \right]^2 \\ &\quad + M^{*2}(d-1)TP_{\theta_0, m_0} \left[ \int_{D_{\hat{\theta}}} [\hat{m}'(u) - m'_0(u)]^2 du \right], \end{aligned} \quad (\text{H.16})$$

where  $M^*$  is defined in (4.18). 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,

$$\int_{D_{\hat{\theta}}} \{\hat{m}'(u) - m'_0(u)\}^2 du \lesssim \|\hat{m}' \circ \hat{\theta} - m'_0 \circ \hat{\theta}\|^2 = O_p(\hat{\lambda}_n).$$

The theorem now follows, as

$$\begin{aligned} |P_{\theta_0, m_0}[(m_0 - \hat{m})(\theta_0^\top X)U_{\hat{\theta}, \hat{m}}(X)]| &= O_p(\hat{\lambda}_n^{3/2}) = O_p(n^{-3/5}), \\ |P_{\theta_0, m_0}[(\hat{m}(\theta_0^\top X) - \hat{m}(\hat{\theta}^\top X))U_{\hat{\theta}, \hat{m}}(X)]| &= O_p(\hat{\lambda}_n^{3/2}) = O_p(n^{-3/5}). \end{aligned}$$

### H.7. Proof of Theorem A.3

We start with some notation. Recall that for any (fixed or random)  $(\theta, m) \in \Theta \times \mathcal{R}$ ,  $P_{\theta, m}$  denotes the joint distribution of  $Y$  and  $X$ , where  $Y = m(\theta^\top X) + \epsilon$  and  $P_X$  denotes the distribution of  $X$ . Now, let  $P_{\theta, m}^{(Y, X)|\theta^\top X}$  denote the joint distribution of  $(Y, X)$  given  $\theta^\top X$ . For any  $(\theta, m) \in \Theta \times \mathcal{R}$  and  $f \in L_2(P_{\theta, m})$ , we have  $P_{\theta, m}[f(X)] = P_X(f(X))$  and

$$\begin{aligned} P_{\theta, m}[(Y - m_0(\theta^\top X))f(X)] &= P_X[P_{\theta, m}^{(Y, X)|\theta^\top X}[f(X)(Y - m_0(\theta^\top X))]] \\ &= P_X[\mathbb{E}(f(X)|\theta^\top X)(m(\theta^\top X) - m_0(\theta^\top X))]. \end{aligned}$$



By above display, we have that

$$\begin{aligned}
P_{\hat{\theta}, m_0} \psi_{\hat{\theta}, \hat{m}} &= H_{\hat{\theta}}^\top P_{\hat{\theta}, m_0} \left[ (Y - \hat{m}(\hat{\theta}^\top X)) [\hat{m}'(\hat{\theta}^\top X) X - (m'_0 h_{\theta_0})(\hat{\theta}^\top X)] \right] \\
&= H_{\hat{\theta}}^\top P_X \left[ (m_0 - \hat{m})(\hat{\theta}^\top X) [\hat{m}'(\hat{\theta}^\top X) E(X|\hat{\theta}^\top X) - (m'_0 h_{\theta_0})(\hat{\theta}^\top X)] \right] \\
&= H_{\hat{\theta}}^\top P_X \left[ (m_0 - \hat{m})(\hat{\theta}^\top X) E(X|\hat{\theta}^\top X) (\hat{m}' - m'_0)(\hat{\theta}^\top X) \right] \\
&\quad + H_{\hat{\theta}}^\top P_X \left[ (m_0 - \hat{m})(\hat{\theta}^\top X) m'_0(\hat{\theta}^\top X) [E(X|\hat{\theta}^\top X) - h_{\theta_0}(\hat{\theta}^\top X)] \right].
\end{aligned} \tag{H.17}$$

Now we will show that each of the terms in (H.17) are  $o_p(n^{-1/2})$ . By (A1) and the Cauchy-Schwarz inequality, for the first term in (H.17) we have

$$\begin{aligned}
&|P_X[(m_0 - \hat{m})(\hat{\theta}^\top X) E(X|\hat{\theta}^\top X) (\hat{m}' - m'_0)(\hat{\theta}^\top X)]| \\
&\leq T \sqrt{P_X[(m_0 - \hat{m})(\hat{\theta}^\top X)^2] P_X[(\hat{m}'(\hat{\theta}^\top X) - m'_0(\hat{\theta}^\top X))^2]} \\
&\lesssim \|m_0 \circ \hat{\theta} - \hat{m} \circ \hat{\theta}\| \|\hat{m}' \circ \hat{\theta} - m'_0 \circ \hat{\theta}\|.
\end{aligned} \tag{H.18}$$

We can bound the two terms on the right side above display as follows. For the first term, note that by Theorems 3.5, 3.6, and 3.7, we have

$$\begin{aligned}
\|m_0 \circ \hat{\theta} - \hat{m} \circ \hat{\theta}\| &\leq \|m_0 \circ \theta_0 - m_0 \circ \hat{\theta}\| + \|\hat{m} \circ \hat{\theta} - m_0 \circ \theta_0\| \\
&\leq T \|m'_0\|_\infty |\theta_0 - \hat{\theta}| + \|\hat{m} \circ \hat{\theta} - m_0 \circ \theta_0\| \\
&= O_p(\hat{\lambda}_n).
\end{aligned} \tag{H.19}$$

For the second term in (H.18), observe that by Lemma E.1 and Theorems 3.7 and 3.8, 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 (H.17) can be bounded as

$$\begin{aligned}
&|P_X[(m_0 - \hat{m})(\hat{\theta}^\top X) m'_0(\hat{\theta}^\top X) (E(X|\hat{\theta}^\top X) - h_{\theta_0}(\hat{\theta}^\top X))]| \\
&\leq \|m'_0\|_\infty \sqrt{P_X[(m_0 - \hat{m})^2 (\hat{\theta}^\top X)] P_X[|h_{\hat{\theta}}(\hat{\theta}^\top X) - h_{\theta_0}(\hat{\theta}^\top X)|^2]} \\
&= \|m'_0\|_\infty \|m_0 \circ \hat{\theta} - \hat{m} \circ \hat{\theta}\| \|h_{\hat{\theta}} \circ \hat{\theta} - h_{\theta_0} \circ \hat{\theta}\|_{2, P_{\theta_0, m_0}} \\
&\leq \|m'_0\|_\infty O_p(\hat{\lambda}_n) \bar{M} |\hat{\theta} - \theta_0| = O_p(\hat{\lambda}_n^2),
\end{aligned} \tag{H.20}$$

where  $\bar{M}$  is defined in (4.2). The last inequality in the above display follows from assumption (B2) and (H.19). The theorem now follows by combining these results.

### H.8. Consistency of $\psi_{\hat{\theta}, \hat{m}}$

The following lemma is used in the proof of Step 5 in Theorem 4.1; also see [32, Section 10.4].

**Lemma H.1.** *If the conditions in Theorem 4.1 hold, then*

$$P_{\theta_0, m_0} |\psi_{\hat{\theta}, \hat{m}} - \psi_{\theta_0, m_0}|^2 = o_p(1), \tag{H.21}$$

$$P_{\hat{\theta}, m_0} |\psi_{\hat{\theta}, \hat{m}}|^2 = O_p(1). \tag{H.22}$$

*Proof.* We first prove (H.21). 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} \left| (y - \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)] \right. \\
&\quad \left. - (y - m_0(\theta_0^\top X)) H_{\theta_0}^\top [m'_0(\theta_0^\top X) X - (m'_0 h_{\theta_0})(\theta_0^\top X)] \right|^2 \\
&= P_{\theta_0, m_0} \left| [(m_0(\theta_0^\top X) - \hat{m}(\hat{\theta}^\top X)) + \epsilon] H_{\hat{\theta}}^\top [\hat{m}'(\hat{\theta}^\top X) X - (m'_0 h_{\theta_0})(\hat{\theta}^\top X)] \right. \\
&\quad \left. - \epsilon H_{\theta_0}^\top [m'_0(\theta_0^\top X) X - (m'_0 h_{\theta_0})(\theta_0^\top X)] \right|^2 \\
&= P_{\theta_0, m_0} \left| [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)] \right|^2 \\
&\quad + P_{\theta_0, m_0} \left| \left[ H_{\hat{\theta}}^\top [\hat{m}'(\hat{\theta}^\top X) X - (m'_0 h_{\theta_0})(\hat{\theta}^\top X)] - H_{\theta_0}^\top [m'_0(\theta_0^\top X) X - (m'_0 h_{\theta_0})(\theta_0^\top X)] \right] \right|^2 \\
&\leq 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 \\
&\quad + P_{\theta_0, m_0} \left| \epsilon H_{\hat{\theta}}^\top [\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)] \right|^2 \\
&\quad + P_{\theta_0, m_0} \left| \epsilon [H_{\hat{\theta}}^\top - H_{\theta_0}^\top] [m'_0(\theta_0^\top X) X - (m'_0 h_{\theta_0})(\theta_0^\top X)] \right|^2 \\
&\leq 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 \\
&\quad + \|\sigma^2(\cdot)\|_\infty P_{\theta_0, m_0} \left| \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) \right|^2 \\
&\quad + 4M_1^2 T^2 \|\sigma^2(\cdot)\|_\infty \|H_{\hat{\theta}} - H_{\theta_0}\|_2^2 \\
&\leq 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 \\
&\quad + 2\|\sigma^2(\cdot)\|_\infty P_{\theta_0, m_0} \left| \hat{m}'(\hat{\theta}^\top X) X - m'_0(\theta_0^\top X) X \right|^2 \\
&\quad + 2\|\sigma^2(\cdot)\|_\infty 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 + 4M_1^2 T^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_1^2 T^2 \|\sigma^2(\cdot)\|_\infty |\hat{\theta} - \theta_0|^2. \tag{H.23}
\end{aligned}$$

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

$$\begin{aligned}
\mathbf{II} &\leq T^2 P_{\theta_0, m_0} \left| \hat{m}'(\hat{\theta}^\top X) - m'_0(\theta_0^\top X) \right|^2 \\
&\leq P_{\theta_0, m_0} |(\hat{m}'(\hat{\theta}^\top X) - \hat{m}'(\theta_0^\top X))|^2 + P_{\theta_0, m_0} |(\hat{m}'(\theta_0^\top X) - m'_0(\theta_0^\top X))|^2 \\
&\leq J^2(\hat{m}) T |\hat{\theta} - \theta_0| + \|\hat{m}' \circ \theta_0 - m'_0 \circ \theta_0\|^2 \\
&= o_p(1).
\end{aligned}$$

Here the last inequality follows from Lemma E.1 and the last equality is due to Theorems 3.7 and 3.8. For  $\mathbf{I}$ , recall that by Theorem 3.5, we have  $\|m_0 \circ \theta_0 - \hat{m} \circ \hat{\theta}\| \xrightarrow{P} 0$ . Thus,

$$\begin{aligned}
\mathbf{I} &= P_{\theta_0, m_0} |(m_0(\theta_0^\top X) - \hat{m}(\hat{\theta}^\top X))(\hat{m}'(\hat{\theta}^\top X) X - (m'_0 h_{\theta_0})(\hat{\theta}^\top X))|^2 \\
&\leq \|m_0 \circ \theta_0 - \hat{m} \circ \hat{\theta}\|^2 (M_2 T + L \|h_{\theta_0}\|_{2, \infty})^2 = o_p(1).
\end{aligned}$$

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 (H.22). Note that

$$\begin{aligned}
& P_{\hat{\theta}, m_0} |\psi_{\hat{\theta}, \hat{m}}|^2 \\
& \leq P_{\hat{\theta}, m_0} \left| (Y - \hat{m}(\hat{\theta}^\top X))^2 [\hat{m}'(\hat{\theta}^\top X)X - m'_0(\hat{\theta}^\top X)h_{\theta_0}(\hat{\theta}^\top X)] \right|^2 \\
& = P_{\hat{\theta}, m_0} \left| [(m_0(\hat{\theta}^\top X) - \hat{m}(\hat{\theta}^\top X)) + \epsilon] [\hat{m}'(\hat{\theta}^\top X)X - (m'_0 h_{\theta_0})(\hat{\theta}^\top X)] \right|^2 \\
& = P_{\hat{\theta}, m_0} \left| [(m_0(\hat{\theta}^\top X) - \hat{m}(\hat{\theta}^\top X))] [\hat{m}'(\hat{\theta}^\top X)X - (m'_0 h_{\theta_0})(\hat{\theta}^\top X)] \right|^2 \\
& \quad + P_{\hat{\theta}, m_0} |\hat{m}'(\hat{\theta}^\top X)X - (m'_0 h_{\theta_0})(\hat{\theta}^\top X)|^2 \\
& \leq (\|m_0\|_\infty^2 + \|\hat{m}\|_\infty^2) P_{\hat{\theta}, m_0} |\hat{m}'(\hat{\theta}^\top X)X - (m'_0 h_{\theta_0})(\hat{\theta}^\top X)|^2 \\
& \quad + P_{\hat{\theta}, m_0} |\hat{m}'(\hat{\theta}^\top X)X - (m'_0 h_{\theta_0})(\hat{\theta}^\top X)|^2 \\
& \leq (\|m_0\|_\infty^2 + \|\hat{m}\|_\infty^2 + 1) P_{\hat{\theta}, m_0} |\hat{m}'(\hat{\theta}^\top X)X - (m'_0 h_{\theta_0})(\hat{\theta}^\top X)|^2.
\end{aligned} \tag{H.24}$$

The result now follows.  $\square$

### H.9. Proof of Theorem A.4

Recall the definition (4.20). Under model (1.1),

$$\begin{aligned}
\psi_{\hat{\theta}, \hat{m}} - \psi_{\theta_0, m_0} &= [\epsilon + m_0(\theta_0^\top x) - \hat{m}(\hat{\theta}^\top x)] H_{\hat{\theta}}^\top [\hat{m}'(\hat{\theta}^\top x)x - (m'_0 h_{\theta_0})(\hat{\theta}^\top x)] \\
&\quad - \epsilon H_{\theta_0}^\top [m'_0(\theta_0^\top x)x - (m'_0 h_{\theta_0})(\theta_0^\top x)] \\
&= \epsilon H_{\hat{\theta}}^\top [\hat{m}'(\hat{\theta}^\top x) - m'_0(\theta_0^\top x)]x + [(m'_0 h_{\theta_0})(\theta_0^\top x) - (m'_0 h_{\theta_0})(\hat{\theta}^\top x)] \\
&\quad + \epsilon(H_{\hat{\theta}}^\top - H_{\theta_0}^\top)[m'_0(\theta_0^\top x)x - (m'_0 h_{\theta_0})(\theta_0^\top x)] \\
&\quad + H_{\hat{\theta}}^\top [m_0(\theta_0^\top x) - \hat{m}(\hat{\theta}^\top x)][\hat{m}'(\hat{\theta}^\top x)x - (m'_0 h_{\theta_0})(\hat{\theta}^\top x)].
\end{aligned} \tag{H.25}$$

For every  $(\theta, m) \in \Theta \times \mathcal{R}$ , define functions  $v_{\theta, m} : \mathcal{X} \rightarrow \mathbb{R}^{d-1}$  and  $\tau_{\theta, m} : \mathcal{X} \rightarrow \mathbb{R}^{d-1}$  as follows:

$$\begin{aligned}
\tau_{\theta, m}(x) &:= H_{\hat{\theta}}^\top \{[m'(\theta^\top x) - m'_0(\theta_0^\top x)]x + [(m'_0 h_{\theta_0})(\theta_0^\top x) - (m'_0 h_{\theta_0})(\theta^\top x)]\} \\
&\quad + (H_{\hat{\theta}}^\top - H_{\theta_0}^\top)[m'_0(\theta_0^\top x)x - (m'_0 h_{\theta_0})(\theta_0^\top x)], \\
v_{\theta, m}(x) &:= H_{\hat{\theta}}^\top [m_0(\theta_0^\top x) - m(\theta^\top x)][m'(\theta^\top x)x - (m'_0 h_{\theta_0})(\theta^\top x)],
\end{aligned} \tag{H.26}$$

and the classes of such functions

$$\begin{aligned}
\Xi_{M_1, M_2, M_3}(n) &= \{\tau_{\theta, m} : (\theta, m) \in \mathcal{C}_{M_1, M_2, M_3}(n)\}, \\
\Upsilon_{M_1, M_2, M_3}(n) &= \{v_{\theta, m} : (\theta, m) \in \mathcal{C}_{M_1, M_2, M_3}(n)\}.
\end{aligned}$$

Observe that, for every fixed  $M_1, M_2$ , and  $M_3$ , we have

$$\begin{aligned}
& \mathbb{P}(|\mathbb{G}_n(\psi_{\hat{\theta}, \hat{m}} - \psi_{\theta_0, m_0})| > \delta) \\
& \leq \mathbb{P}(|\mathbb{G}_n(\epsilon \tau_{\hat{\theta}, \hat{m}} + v_{\hat{\theta}, \hat{m}})| > \delta, (\hat{\theta}, \hat{m}) \in \mathcal{C}_{M_1, M_2, M_3}(n)) + \mathbb{P}((\hat{\theta}, \hat{m}) \notin \mathcal{C}_{M_1, M_2, M_3}(n)) \\
& \leq \mathbb{P}\left(|\mathbb{G}_n(\epsilon \tau_{\hat{\theta}, \hat{m}})| > \frac{\delta}{2}, (\hat{\theta}, \hat{m}) \in \mathcal{C}_{M_1, M_2, M_3}(n)\right) \\
& \quad + \mathbb{P}\left(|\mathbb{G}_n v_{\hat{\theta}, \hat{m}}| > \frac{\delta}{2}, (\hat{\theta}, \hat{m}) \in \mathcal{C}_{M_1, M_2, M_3}(n)\right) + \mathbb{P}((\hat{\theta}, \hat{m}) \notin \mathcal{C}_{M_1, M_2, M_3}(n)) \\
& \leq \mathbb{P}\left(\sup_{f \in \Xi_{M_1, M_2, M_3}(n)} |\mathbb{G}_n \epsilon f| > \frac{\delta}{2}\right) \\
& \quad + \mathbb{P}\left(\sup_{f \in \Upsilon_{M_1, M_2, M_3}(n)} |\mathbb{G}_n f| > \frac{\delta}{2}\right) + \mathbb{P}((\hat{\theta}, \hat{m}) \notin \mathcal{C}_{M_1, M_2, M_3}(n)).
\end{aligned} \tag{H.27}$$

By the discussion following Lemma A.1, we have

$$\mathbb{P}((\hat{\theta}, \hat{m}) \notin \mathcal{C}_{M_1, M_2, M_3}(n)) \rightarrow 0.$$

Hence to prove Theorem A.4, we only need to show that the first two terms in (H.27) are  $o(1)$ . We prove this in Lemmas H.2 and H.3.

**Lemma H.2.** Fix  $M_1, M_2, M_3$ , and  $\delta > 0$ . For  $n \in \mathbb{N}$ , as  $n \rightarrow \infty$ , we have

$$\mathbb{P}\left(\sup_{f \in \Xi_{M_1, M_2, M_3}(n)} |\mathbb{G}_n \epsilon f| > \frac{\delta}{2}\right) \rightarrow 0.$$

*Proof.* The proof of this lemma is similar to the first part of the proof of Lemma A.5. Let us define,

$$\Xi_{M_1, M_2, M_3}^* := \{\tau_{\theta, m} : (\theta, m) \in \mathcal{C}_{M_1, M_2, M_3}^*\}.$$

We will prove that

$$N(\varepsilon, \Xi_{M_1, M_2, M_3}^*, \|\cdot\|_{2, \infty}) \leq c \exp(c/\varepsilon) \varepsilon^{-4d}, \quad (\text{H.28})$$

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 Theorem 2.4 of [49] 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  $\varepsilon$ -cover of  $\{m' : (\cdot, m) \in \mathcal{C}_{M_1, M_2, M_3}^*\}$  by  $l_1, \dots, l_t$ . By Lemma 15 of [32], we have that there exists  $\theta_1, \dots, \theta_s$  for  $s \lesssim \varepsilon^{-4d}$  such that  $\{\theta_i\}_{1 \leq i \leq s}$  form an  $\varepsilon^2$ -cover of  $\Theta \cap B_{\theta_0}(1/2)$  and satisfies (86) of [32] (with  $\varepsilon^2$  instead of  $\varepsilon$ ). 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  $\varepsilon$ -cover is  $l_1$  and the vector nearest to  $\theta$  in the  $\varepsilon^2$ -cover of  $\Theta \cap B_{\theta_0}(1/2)$  is  $\theta_1$  i.e.,

$$\|m' - l_1\|_{\infty} \leq \varepsilon, \quad \|H_{\theta}^{\top} - H_{\theta_1}^{\top}\| \leq \varepsilon^2, \quad \text{and} \quad |\theta - \theta_1| \leq \varepsilon^2. \quad (\text{H.29})$$

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 (H.28), 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

$$\begin{aligned} & |\tau_{\theta, m}(x) - \tau_{\theta_1, r_1}(x)| \\ & \leq |H_{\theta}^{\top} \varrho_{\theta, m}(x) - H_{\theta_1}^{\top} \varrho_{\theta_1, r_1}(x)| + |(H_{\theta}^{\top} - H_{\theta_1}^{\top})[m'_0(\theta_0^{\top} x)x - (m'_0 h_{\theta_0})(\theta_0^{\top} x)]| \\ & \leq |(H_{\theta}^{\top} - H_{\theta_1}^{\top}) \varrho_{\theta, m}(x)| + |H_{\theta_1}^{\top}(\varrho_{\theta, m}(x) - \varrho_{\theta_1, r_1}(x))| + \varepsilon^2 |m'_0(\theta_0^{\top} x)x - (m'_0 h_{\theta_0})(\theta_0^{\top} x)| \\ & \leq \varepsilon^2 |\varrho_{\theta, m}(x)| + |\varrho_{\theta, m}(x) - \varrho_{\theta_1, r_1}(x)| + 2M_2 T \varepsilon^2 \\ & \leq \varepsilon^2 4M_2 T + |\varrho_{\theta, m}(x) - \varrho_{\theta_1, r_1}(x)| + 2M_2 T \varepsilon^2, \end{aligned} \quad (\text{H.30})$$

where the last two inequalities follow from properties of  $H_{\theta}$  (Lemma 1 of [32]), (H.29), and definition of  $\mathcal{C}_{M_1, M_2, M_3}^*$  (see (A.8)). Furthermore, we have

$$\begin{aligned} & |\varrho_{\theta, m}(x) - \varrho_{\theta_1, r_1}(x)| \\ & \leq |(m'(\theta^{\top} x) - r'_1(\theta_1^{\top} x))x| + |((m'_0 h_{\theta_0})(\theta_1^{\top} x) - (m'_0 h_{\theta_0})(\theta^{\top} x))| \\ & \leq |(m'(\theta^{\top} x) - m'(\theta_1^{\top} x))x| + |(m'(\theta_1^{\top} x) - r'_1(\theta_1^{\top} x))x| \\ & \quad + |(m'_0(\theta_1^{\top} x) - m'_0(\theta^{\top} x))h_{\theta_0}(\theta_1^{\top} x)| + |m'_0(\theta^{\top} x)(h_{\theta_0}(\theta_1^{\top} x) - h_{\theta_0}(\theta^{\top} x))| \\ & \leq M_3 T^2 |\theta - \theta_1|^{1/2} + \|m - r_1\|_{\infty} T + \|h_{\theta_0}\|_{\infty} M_3 |\theta - \theta_1|^{1/2} + M_2 \|h'_{\theta_0}\|_{\infty} |\theta - \theta_1| T \\ & \lesssim \varepsilon \end{aligned} \quad (\text{H.31})$$

Thus combining (H.30) and (H.31), we have  $\|\tau_{\theta, m} - \tau_{\theta_1, r_1}\|_{2, \infty} \leq c_1 \varepsilon$ .

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

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

If  $[\bar{h}_1, \bar{h}_2]$  is a bracket for  $\tau_{\theta,m}$ , then  $[\bar{h}_1 \epsilon^+ - \bar{h}_2 \epsilon^-, \bar{h}_2 \epsilon^+ - \bar{h}_1 \epsilon^-]$  is a bracket (coordinate-wise) for  $\epsilon \tau_{\theta,m}$ . Therefore, we have

$$N_{[]}(\varepsilon, \{\epsilon f : f \in \Xi_{M_1,M_2,M_3}^*\}, \|\cdot\|_{2,P_{\theta_0,m_0}}) \lesssim c \exp(c/\varepsilon).$$

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  $x \in \mathcal{X}$  note that,

$$\begin{aligned} |\tau_{\theta,m}(x)| &\leq [|m'(\theta^\top x) - m'(\theta_0^\top x)| + |m'(\theta_0^\top x) - m'_0(\theta_0^\top x)|] |x| \\ &\quad + |m'_0(\theta_0^\top x) h_{\theta_0}(\theta_0^\top x) - m'_0(\theta^\top x) h_{\theta_0}(\theta_0^\top x)| \\ &\quad + |m'_0(\theta^\top x) h_{\theta_0}(\theta_0^\top x) - m'_0(\theta^\top x) h_{\theta_0}(\theta^\top x)| \\ &\quad + |\theta - \theta_0| |m'_0(\theta^\top x) x - (m'_0 h_{\theta_0})(\theta^\top x)| \\ &\leq J(m) |\theta^\top x - \theta_0^\top x|^{1/2} |x| + \|m' - m'_0\|_{D_0}^S |x| \\ &\quad + |h_{\theta_0}(\theta_0^\top x)| J(m_0) |\theta_0^\top x - \theta^\top x|^{1/2} \\ &\quad + |m'_0(\theta^\top x)| |h_{\theta_0}(\theta_0^\top x) - h_{\theta_0}(\theta^\top x)| + |\theta - \theta_0| 2M_2 T \\ &\leq \hat{\lambda}_n^{1/4} (M_3 T^2 + \|h_{\theta_0}\|_{2,\infty} M_3 T + M_2 \|h'_{\theta_0}\|_{2,\infty} T + 2M_2 T) + \frac{1}{a_n} T. \end{aligned}$$

Hence,

$$|\epsilon \tau_{\theta,m}(x)| \leq |\epsilon| \hat{\lambda}_n^{1/4} (M_3 T^2 + \|h_{\theta_0}\|_{2,\infty} M_3 T + M_2 \|h'_{\theta_0}\|_{2,\infty} T + 2M_2 T) + |\epsilon| \frac{1}{a_n} T := W_n$$

Thus using arguments similar to (H.11) and (H.12) and the maximal inequality in Corollary 19.35 of [52] (also see proof of Lemma A.3), we have

$$\begin{aligned} \mathbb{P}\left(\sup_{f \in \Xi_{M_1,M_2,M_3}(n)} |\mathbb{G}_n \epsilon f| > \frac{\delta}{2}\right) &\lesssim 2\delta^{-1} \sqrt{d-1} \sum_{i=1}^{d-1} \mathbb{E}\left(\sup_{f \in \Xi_{M_1,M_2,M_3}(n)} |\mathbb{G}_n \epsilon f_i|\right) \\ &\lesssim \delta^{-1} d \sqrt{d-1} J_{[]} \left(\|W_n\|, \Xi_{M_1,M_2,M_3}(n), \|\cdot\|_{2,P_{\theta_0,m_0}}\right) \\ &\lesssim \left[\hat{\lambda}_n^{1/4} + \frac{1}{a_n}\right]^{1/2} \rightarrow 0 \text{ as } n \rightarrow \infty, \end{aligned}$$

where in the first inequality  $f_1, \dots, f_{d-1}$  denote each component of  $f$ . □

**Lemma H.3.** 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 A.3 and A.4. 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  $x \in \mathcal{X}$ ,

$$\begin{aligned} |v_{\theta,m}(x)| &\leq |m_0(\theta_0^\top x) - m(\theta_0^\top x)| \cdot |m'(\theta^\top x) x - m'_0(\theta^\top x) h_{\theta_0}(\theta^\top x)| \\ &\quad + |m(\theta_0^\top x) - m(\theta^\top x)| \cdot |m'(\theta^\top x) x - m'_0(\theta^\top x) h_{\theta_0}(\theta^\top x)| \\ &\leq \|m_0 - m\|_{D_0}^S |m'(\theta^\top x) x - m'_0(\theta^\top x) h_{\theta_0}(\theta^\top x)| \\ &\quad + \|m'\|_\infty T |\theta - \theta_0| \cdot |m'(\theta^\top x) x - m'_0(\theta^\top x) h_{\theta_0}(\theta^\top x)| \\ &\leq \left[\frac{1}{a_n} + T M_2 \hat{\lambda}_n^{1/2}\right] 2M_2 T \leq C \left[\frac{1}{a_n} + \hat{\lambda}_n^{1/2}\right], \end{aligned}$$

where  $C$  is a constant depending only on  $T, M_1, M_2$ , and  $M_3$ . Let us now define

$$\Upsilon_{M_1, M_2, M_3}^* := \{v_{\theta, m} : (\theta, m) \in \mathcal{C}_{M_1, M_2, M_3}^*\}.$$

Thus using arguments similar to the previous lemma, we have

$$\begin{aligned} \mathbb{P}\left(\sup_{f \in \Upsilon_{M_1, M_2, M_3}(n)} |\mathbb{G}_n f| > \frac{\delta}{2}\right) &\lesssim 2\delta^{-1} \sqrt{d-1} \sum_{i=1}^{d-1} \mathbb{E}\left(\sup_{f \in \Upsilon_{M_1, M_2, M_3}(n)} |\mathbb{G}_n f_i|\right) \\ &\lesssim C\delta^{-1} J_{[\cdot]} \left(C \left[\frac{1}{a_n} + \hat{\lambda}_n^{1/2}\right], \Upsilon_{M_1, M_2, M_3}(n), \|\cdot\|_{2, \infty}\right) \\ &\lesssim C\delta^{-1} J_{[\cdot]} \left(C \left[\frac{1}{a_n} + \hat{\lambda}_n^{1/2}\right], \Upsilon_{M_1, M_2, M_3}^*, \|\cdot\|_{2, \infty}\right), \end{aligned}$$

where  $C$  is a constant depending only on  $M_1, M_2$ , and  $M_3$  and  $f_1, \dots, f_{d-1}$  denote each component of  $f$ . 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 it is enough to show that,  $J_{[\cdot]}(\gamma, \Upsilon_{M_1, M_2, M_3}^*, \|\cdot\|_{2, \infty}) \leq \gamma^{1/2}$ , for all  $\gamma > 0$ , which is implied by

$$N_{[\cdot]}(\varepsilon, \Upsilon_{M_1, M_2, M_3}^*, \|\cdot\|_{2, \infty}) \lesssim \exp\left(\frac{c}{\varepsilon} + \frac{c}{\sqrt{\varepsilon}}\right) \varepsilon^{-5d}, \quad (\text{H.32})$$

where  $c$  is a constant depending only on  $d, M_1, M_2$ , and  $M_3$ . In the following, we show (H.32). Observe that by an argument similar to the proof of Lemma 8 of [32], 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 (H.26)) and Lemma 9.25 of [31] (for the entropy of product of classes of uniformly bounded functions) to prove (H.32), it is enough to show that

$$N(\varepsilon, \{H_\theta^\top V_{\theta, m} : (\theta, m) \in \mathcal{C}_{M_1, M_2, M_3}^*\}, \|\cdot\|_{2, \infty}) \lesssim \varepsilon^{-4d} \exp(c/\sqrt{\varepsilon}). \quad (\text{H.33})$$

We will prove (H.33) by constructing a cover for  $\{H_\theta^\top V_{\theta, m} : (\theta, m) \in \mathcal{C}_{M_1, M_2, M_3}^*\}$ . By Theorem 2.4 of [49], 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  $\varepsilon$ -cover and their anti-derivatives by  $l_1, \dots, l_t$  and  $r_1, \dots, r_t$ , i.e.,  $l_i = r'_i$  for  $1 \leq i \leq t$ . By Lemma 15 of [32], we have that there exists  $\theta_1, \dots, \theta_s$  for  $s \lesssim \varepsilon^{-4d}$  such that  $\{\theta_i\}_{1 \leq i \leq s}$  form an  $\varepsilon^2$ -cover of  $\Theta \cap B_{\theta_0}(1/2)$  and satisfies (86) of [32] (with  $\varepsilon^2$  instead of  $\varepsilon$ ). We now show that  $\{H_{\theta_i} V_{\theta_i, r_j}\}_{1 \leq i \leq s, 1 \leq j \leq t}$  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  $\varepsilon$ -cover is  $l_1$  and the vector nearest to  $\theta$  in the  $\varepsilon^2$  cover of  $\Theta \cap B_{\theta_0}(1/2)$  is  $\theta_1$ , i.e.,

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

Observe that

$$\begin{aligned} |H_\theta^\top V_{\theta, m}(x) - H_{\theta_1}^\top V_{\theta_1, r_1}(x)| &\leq |H_\theta^\top V_{\theta, m}(x) - H_{\theta_1}^\top V_{\theta, m}(x)| + |H_{\theta_1}^\top V_{\theta, m}(x) - H_{\theta_1}^\top V_{\theta_1, r_1}(x)| \\ &\leq \varepsilon^2 |V_{\theta, m}(x)| + |V_{\theta, m}(x) - V_{\theta_1, r_1}(x)|. \end{aligned} \quad (\text{H.34})$$

Furthermore, we have

$$\begin{aligned}
& |V_{\theta,m}(x) - V_{\theta_1,r_1}(x)| \\
& \leq |m'(\theta^\top x) - (m'_0 h_{\theta_0})(\theta^\top x) - r'_1(\theta_1^\top x)x + (m'_0 h_{\theta_0})(\theta_1^\top x)| \\
& \leq T|m'(\theta^\top x) - r'_1(\theta_1^\top x)| + |(m'_0 h_{\theta_0})(\theta^\top x) - (m'_0 h_{\theta_0})(\theta_1^\top x)| \\
& \leq T|m'(\theta^\top x) - m'(\theta_1^\top x)| + T|m'(\theta_1^\top x) - r'_1(\theta_1^\top x)| \\
& \quad + |(m'_0 h_{\theta_0})(\theta^\top x) - (m'_0 h_{\theta_0})(\theta_1^\top x)| \\
& \leq TM_3|\theta^\top x - \theta_1^\top x|^{1/2} + T\varepsilon \\
& \quad + |(m'_0 h_{\theta_0})(\theta^\top x) - (m'_0 h_{\theta_0})(\theta_1^\top x)| \lesssim \varepsilon.
\end{aligned} \tag{H.35}$$

Thus combining (H.34), (H.35), and the fact that  $|V_{\theta,m}(x)| \leq 2TM_2$ , we have

$$\|H_\theta^\top V_{\theta,m} - H_{\theta_1}^\top V_{\theta_1,r_1}\|_{2,\infty} \lesssim \varepsilon. \quad \square$$

## Appendix I: Proof of Results in Appendix B

### I.1. Proof of Theorem B.1

We start by the following definition

$$\varphi_{\theta,\eta,t}(u) := \phi_{\theta,\eta,t}(u + (\theta - \zeta_t(\theta, \eta))^\top k(u)).$$

Note that by (A.3),  $\varphi_{\theta,\eta,t}(\cdot)$  is an increasing function for  $t \in \mathbb{R}$  close to 0. Hence, as  $m'$  is a bounded increasing function,  $\xi_t(\cdot; \theta, \eta, m)$  is a uniformly Lipschitz convex function for  $t$  sufficiently close to 0. Since  $|m'|_\infty$  is bounded by  $L$ , so is  $\nabla_t$ . By definition we have

$$\xi_t(s^\top x; \theta, \eta, m) = \int_{s_0}^{s^\top x} \nabla_t(y; \theta, \eta, m) dy + (\zeta_t(\theta, \eta) - \theta)^\top [(m'(s_0) - m'_0(s_0))k(s_0) + m'_0(s_0)h_{\theta_0}(s_0)] + m(s_0).$$

We have that  $\phi_{\theta,\eta,0}(u + (\theta - \zeta_0(\theta, \eta))^\top k(u)) = u$ ,  $\forall u \in D$ . Hence,

$$\xi_0(\zeta_0(\theta, \eta)^\top x; \theta, \eta, m) = \int_{s_0}^{\theta^\top x} \nabla_0(u; \theta, \eta, m) du + m(s_0) = \int_{s_0}^{\theta^\top x} m' \circ \phi_{\theta,\eta,0}(u) du + m(s_0) = m(\theta^\top x).$$

Observe that,

$$\begin{aligned}
& \frac{\partial}{\partial t} \xi_t(\zeta_t(\theta, \eta)^\top x; \theta, \eta, m) \\
& = \frac{\partial}{\partial t} \left\{ \int_{s_0}^{\zeta_t(\theta, \eta)^\top x} m' \circ \varphi_{\theta,\eta,t}(u) du \right\} + \frac{\partial \zeta_t(\theta, \eta)}{\partial t}^\top \left[ (m'(s_0) - m'_0(s_0))k(s_0) + m'_0(s_0)h_{\theta_0}(s_0) \right] \quad (\text{I.1})
\end{aligned}$$

We next evaluate the first term on the right hand side of the above display. But first, we introduce some notations. Let us define,

$$\phi'_{\theta,\eta,t}(u) := \frac{\partial}{\partial u} \phi_{\theta,\eta,t}(u), \quad \phi''_{\theta,\eta,t}(u) := \frac{\partial}{\partial u} \phi'_{\theta,\eta,t}(u), \quad \dot{\phi}_{\theta,\eta,t}(u) := \frac{\partial}{\partial t} \phi_{\theta,\eta,t}(u),$$

and

$$\varphi'_{\theta,\eta,t}(y) := \frac{\partial \varphi_{\theta,\eta,t}(u)}{\partial u} = \phi'_{\theta,\eta,t}(u + (\theta - \zeta_t(\theta, \eta))^\top k(u))(1 + (\theta - \zeta_t(\theta, \eta))^\top k'(u)).$$

Now, observe that

$$\begin{aligned}
\frac{\partial \varphi_{\theta, \eta, t}(u)}{\partial t} &= \dot{\phi}_{\theta, \eta, t}(u + (\theta - \zeta_t(\theta, \eta))^\top k(u)) - \phi'_{\theta, \eta, t}(u + (\theta - \zeta_t(\theta, \eta))^\top k(u)) \frac{\partial \zeta_t(\theta, \eta)}{\partial t}^\top k(u), \\
\frac{\partial \phi'_{\theta, \eta, t}(u)}{\partial t} &= (1 + (\theta - \zeta_t(\theta, \eta))^\top k'(u)) \left[ \frac{\partial \phi'_{\theta, \eta, t}}{\partial t}(u + (\theta - \zeta_t(\theta, \eta))^\top k(u)) \right. \\
&\quad \left. - \phi''_{\theta, \eta, t}(u + (\theta - \zeta_t(\theta, \eta))^\top k(u)) \frac{\partial \zeta_t(\theta, \eta)}{\partial t}^\top k(u) \right] \\
&\quad - \phi'_{\theta, \eta, t}(u + (\theta - \zeta_t(\theta, \eta))^\top k(u)) \frac{\partial \zeta_t(\theta, \eta)}{\partial t}^\top k'(u), \tag{I.2} \\
\frac{\partial \varphi_{\theta, \eta, t}(\zeta_t(\theta, \eta)^\top x)}{\partial t} &= \dot{\phi}_{\theta, \eta, t}(\zeta_t(\theta, \eta)^\top x + (\theta - \zeta_t(\theta, \eta))^\top k(\zeta_t(\theta, \eta)^\top x)) \\
&\quad + \phi'_{\theta, \eta, t}(\zeta_t(\theta, \eta)^\top x + (\theta - \zeta_t(\theta, \eta))^\top k(\zeta_t(\theta, \eta)^\top x)) \frac{\partial \zeta_t(\theta, \eta)}{\partial t}^\top \\
&\quad \left[ x - k(\zeta_t(\theta, \eta)^\top x) + (\theta - \zeta_t(\theta, \eta))^\top k'(\zeta_t(\theta, \eta)^\top x) x \right].
\end{aligned}$$

Now, we evaluate the first term on the right hand side of (I.1). Note that

$$\begin{aligned}
&\frac{\partial}{\partial t} \left\{ \int_{s_0}^{\zeta_t(\theta, \eta)^\top x} m' \circ \varphi_{\theta, \eta, t}(u) du \right\} \\
&= \frac{\partial}{\partial t} \left\{ \int_{\varphi_{\theta, \eta, t}(s_0)}^{\varphi_{\theta, \eta, t}(\zeta_t(\theta, \eta)^\top x)} \frac{m'(u)}{\varphi'_{\theta, \eta, t} \circ \varphi_{\theta, \eta, t}^{-1}(u)} du \right\} \\
&= \frac{m' \circ \varphi_{\theta, \eta, t}(\zeta_t(\theta, \eta)^\top x)}{\varphi'_{\theta, \eta, t}(\zeta_t(\theta, \eta)^\top x)} \frac{\partial \varphi_{\theta, \eta, t}(\zeta_t(\theta, \eta)^\top x)}{\partial t} - \frac{m' \circ \varphi_{\theta, \eta, t}(s_0)}{\varphi'_{\theta, \eta, t}(s_0)} \frac{\partial \varphi_{\theta, \eta, t}(s_0)}{\partial t} \\
&\quad - \int_{\varphi_{\theta, \eta, t}(s_0)}^{\varphi_{\theta, \eta, t}(\zeta_t(\theta, \eta)^\top x)} \frac{m'(u)}{[\varphi'_{\theta, \eta, t} \circ \varphi_{\theta, \eta, t}^{-1}(u)]^2} \frac{\partial [\varphi'_{\theta, \eta, t} \circ \varphi_{\theta, \eta, t}^{-1}(u)]}{\partial t} du \\
&= \frac{m' \circ \varphi_{\theta, \eta, t}(\zeta_t(\theta, \eta)^\top x)}{\varphi'_{\theta, \eta, t}(\zeta_t(\theta, \eta)^\top x)} \frac{\partial \varphi_{\theta, \eta, t}(\zeta_t(\theta, \eta)^\top x)}{\partial t} - \frac{m' \circ \varphi_{\theta, \eta, t}(s_0)}{\varphi'_{\theta, \eta, t}(s_0)} \frac{\partial \varphi_{\theta, \eta, t}(s_0)}{\partial t} \\
&\quad - \int_{s_0}^{\zeta_t(\theta, \eta)^\top x} \frac{m' \circ \varphi_{\theta, \eta, t}(u)}{[\varphi'_{\theta, \eta, t}(u)]^2} \frac{\partial \varphi'_{\theta, \eta, t}(u)}{\partial t} \varphi'_{\theta, \eta, t}(u) du.
\end{aligned}$$

Using the fact that  $\phi'_{\theta, \eta, t}(u) = 1$  and  $\dot{\phi}_{\theta, \eta, t}(u) = 0$  for all  $u \in D_\theta$  and  $t$  close to 0 (follows from the definition (A.3)) and  $\partial \zeta_t(\theta, \eta)/\partial t = -|\eta|^2 t / \sqrt{1 - t^2 |\eta|^2} \theta + H_\theta \eta$ , i.e.,  $\partial \zeta_t(\theta, \eta)/\partial t|_{t=0} = H_\theta \eta$ , we get from (I.2)

$$\begin{aligned}
&\frac{\partial}{\partial t} \left\{ \int_{s_0}^{\zeta_t(\theta, \eta)^\top x} m' \circ \varphi_{\theta, \eta, t}(u) du \right\} \Big|_{t=0} \\
&= (H_\theta \eta)^\top \left[ m'(\theta^\top x)(x - k(\theta^\top x)) - m'(s_0)[-k(s_0)] - \int_{s_0}^{\theta^\top x} m'(y)[-k'(y)] dy \right].
\end{aligned}$$



We now show that the score function of the sub-model  $\{\zeta_t(\theta, \eta), \xi_t(\cdot; \theta, \eta, m)\}$  is  $\mathfrak{S}_{\theta, m}(x, y)$ , i.e.,

$$\begin{aligned}
& -\frac{1}{2} \frac{\partial}{\partial t} [(y - \xi_t(\zeta_t(\theta, \eta)^\top x; \theta, \eta, m))^2] \Big|_{t=0} \\
&= (y - \xi_t(\zeta_t(\theta, \eta)^\top x; \theta, \eta, m)) \frac{\partial \xi_t(\zeta_t(\theta, \eta)^\top x; \theta, \eta, m)}{\partial t} \Big|_{t=0} \\
&= (y - m(\theta^\top x))(H_\theta \eta)^\top \left[ m'(\theta^\top x)(x - k(\theta^\top x)) - m'(s_0)[-k(s_0)] - \int_{s_0}^{\theta^\top x} m'(y)[-k'(y)] dy \right. \\
&\quad \left. + (m'(s_0) - m'_0(s_0))k(s_0) + m'_0(s_0)h_{\theta_0}(s_0) \right] \\
&= (y - m(\theta^\top x))(H_\theta \eta)^\top \left[ m'(\theta^\top x)x - m'(\theta^\top x)k(\theta^\top x) + \int_{s_0}^{\theta^\top x} m'(y)k'(y) dy \right. \\
&\quad \left. + m'_0(s_0)k(s_0) - m'_0(s_0)h_{\theta_0}(s_0) \right].
\end{aligned}$$

The rest of the proof is similar to the proof of Theorem A.1; see (H.6).

### I.2. Proof of Lemma B.1

Recall that  $U_{\theta, m} : \mathcal{X} \rightarrow \mathbb{R}^{d-1}$  is defined as  $U_{\theta, m}(x) = H_\theta^\top [\int_{s_0}^{\theta^\top x} [m'(u) - m'_0(u)]k'(u)du + (m'_0(\theta^\top x) - m'(\theta^\top x))k(\theta^\top x)]$ ; see (A.7). Observe that  $D$  is a bounded set,  $\sup_{u \in D} (|k(u)| + |k'(u)|) \leq M^*$  and  $\|m'\|_\infty \leq L$ . Hence

$$\begin{aligned}
|U_{\theta, m}(x)| &\leq M^* \int_{s_0}^{\theta^\top x} |m'(u) - m'_0(u)|du + M^* |m'_0(\theta^\top x) - m'(\theta^\top x)| \\
&\leq 2LM^* |\theta^\top x - s_0| + 2M^*L \leq 4LM^*T + 2M^*L := V^*.
\end{aligned}$$

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

$$\begin{aligned}
\mathcal{H}^* &:= \{q : \mathcal{X} \rightarrow \mathbb{R} \mid q(x) = g(\theta^\top x), \theta \in \Theta \text{ and} \\
&\quad g : \mathcal{D} \rightarrow \mathbb{R} \text{ is an increasing function and } \|g\|_\infty \leq S\}.
\end{aligned}$$

The following lemma, proved in Appendix I.3, does this.

**Lemma I.1.** *If  $\sup_{\theta \in \Theta} \|f_{\theta^\top X}\|_D \leq c < \infty$ , where  $f_{\theta^\top X}$  denotes the density of  $\theta^\top X$  with respect to the Lebesgue measure. Then  $\log N_{[]}(\varepsilon, \mathcal{H}^*, L_2(P_{\theta_0, m_0})) \lesssim \varepsilon^{-1}$ .*

Fix  $(\theta, m) \in \mathcal{C}_{M_1}(n)$ . By definition we have that both  $H_\theta^\top k(\cdot)$  and  $H_\theta^\top k'(\cdot)$  are coordinate-wise bounded functions (see (4.17)) 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  $U_{\theta, m}(x) = U_{\theta, m}^{(1)}(x) - U_{\theta, m}^{(2)}(x) + U_{\theta, m}^{(3)}(x)$ , where

$$\begin{aligned}
U_{\theta, m}^{(1)}(x) &:= \int_{s_0}^{\theta^\top x} [m'(u) - m'_0(u)](H_\theta^\top k'(u) + M^* \mathbf{1})du, \\
U_{\theta, m}^{(2)}(x) &:= M^* \mathbf{1} \int_{s_0}^{\theta^\top x} [m'(u) - m'_0(u)]du, \\
U_{\theta, m}^{(3)}(x) &:= (m'_0(\theta^\top x) - m'(\theta^\top x))H_\theta^\top k(\theta^\top x).
\end{aligned}$$

We will find  $c_i \eta$ -brackets (with respect to  $\|\cdot\|_{2, P_{\theta_0, m_0}}$ ) for  $U_{\theta, m}^{(i)}, i = 1, 2$ , 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 I.1 there exists a  $N'_\eta \leq \exp(\eta^{-1})$  such that  $\{(\ell_k, u_k)\}_{1 \leq k \leq N'_\eta}$  form a  $\eta$ -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  $x \in \mathcal{X}$

$$\ell_k(x) \leq m'(\theta^\top x) \leq u_k(x), \quad (\text{I.3})$$

and  $\|u_k - \ell_k\| \leq C\eta$  for some constant  $C$ . Similarly by Lemma 15 of [32], we can find a  $\theta_1, \theta_2, \dots, \theta_{N_\eta}$  with  $N_\eta \leq C\eta^{-2d}$  for some constant  $C$  such that for every  $\theta \in \Theta \cap B_{\theta_0}(1/2)$ , there exists a  $\theta_j$  such that

$$|\theta - \theta_j| \leq \eta/T, \quad \|H_\theta - H_{\theta_j}\|_2 \leq \eta/T, \quad \text{and} \quad |\theta^\top x - \theta_j^\top x| \leq \eta, \quad \forall x \in \mathcal{X}.$$

We first find a  $\|\cdot\|_{2, P_{\theta_0, m_0}}$  bracket for  $U_{\theta, m}^{(3)}$  using Lemma 9.25 of [31]. For this application, we need to find bracketing entropy for the following two classes 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

$$m'_0(\theta_j^\top x - \eta_1) \leq m'_0(\theta^\top x) \leq m'_0(\theta_j^\top x + \eta_1).$$

Thus by (I.3), we have

$$m'_0(\theta_j^\top x - \eta_1) - u_k(x) \leq m'_0(\theta^\top x) - m'(\theta^\top x) \leq m'_0(\theta_j^\top x + \eta_1) - \ell_k(x).$$

The length of the above bracket is given by

$$\begin{aligned} & \|m'_0(\theta_j^\top \cdot + \eta_1) - \ell_k - m'_0(\theta_j^\top \cdot - \eta_1) + u_k\|_{2, P_{\theta_0, m_0}} \\ & \leq [P_{\theta_0, m_0} |m'_0(\theta_j^\top X + \eta_1) - m'_0(\theta_j^\top X - \eta_1)|^2]^{1/2} + \|u_k - \ell_k\| \\ & \leq 2\|m'_0\|_\infty \eta + \eta = (2\|m'_0\|_\infty + 1)\eta. \end{aligned}$$

Thus

$$N_{[]}(\eta, \{m'_0(\theta^\top \cdot) - m'(\theta^\top \cdot) : (\theta, m) \in \mathcal{C}_{M_1}(n)\}, \|\cdot\|) \lesssim \exp(\eta^{-1})\eta^{-2d} \quad (\text{I.4})$$

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

$$H_{\theta_j}^\top k(\theta_j^\top x) - 2\eta M^* \mathbf{1} \leq H_\theta^\top k(\theta^\top x) \leq H_{\theta_j}^\top k(\theta_j^\top x) + 2\eta M^* \mathbf{1},$$

with  $\|\cdot\|_{2, P_{\theta_0, m_0}}$ -length  $4\eta M^* \sqrt{d-1}$ . Thus

$$N_{[]}(\eta, \{H_\theta^\top k(\theta^\top \cdot) : (\theta, m) \in \mathcal{C}_{M_1}(n)\}, \|\cdot\|_{2, P_{\theta_0, m_0}}) \lesssim \exp(\eta^{-1})\eta^{-2d} \quad (\text{I.5})$$

Thus by Lemma 9.25 of [31], (I.4) and (I.5) gives us

$$N_{[]}(\eta, \{[m'_0(\theta^\top \cdot) - m'(\theta^\top \cdot)]H_\theta^\top k(\theta^\top \cdot) : (\theta, m) \in \mathcal{C}_{M_1}(n)\}, \|\cdot\|_{2, P_{\theta_0, m_0}}) \lesssim \exp(\eta^{-1})\eta^{-4d}$$

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 [53], 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  $\eta$ -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  $\theta_j$  satisfies  $|\theta - \theta_j| \leq \eta/T$ , by Lemma 1 of [32] we have

$$|H_\theta^\top k'(u) - H_{\theta_j}^\top k'(u)| \leq |k'(u)|\eta/T \leq M^*\eta/T.$$

This implies

$$H_{\theta_j}^\top k'(u) + M^* \mathbf{1} (1 - \eta/T) \leq H_\theta^\top k'(u) + M^* \mathbf{1} \leq H_{\theta_j}^\top k'(u) + M^* \mathbf{1} (1 + \eta/T).$$

The  $\|\cdot\|_{2, P_{\theta_0, m_0}}$ -length of the above bracket is  $2M^*\eta/T$ . Since  $H_\theta^\top k'(u) + M^* \mathbf{1} \geq 0$  for all  $\theta$  and  $u$ , we can take the brackets to be

$$\{H_{\theta_j}^\top k'(u) + M^* (1 - \eta/T) \mathbf{1}\} \vee 0 \leq H_\theta^\top k'(u) + M^* \mathbf{1} \leq \{H_{\theta_j}^\top k'(u) + M^* (1 + \eta/T) \mathbf{1}\} \wedge (2M^*).$$

From the brackets  $[m_L, m_U]$  of  $m'$ , we get that

$$m_L(u) - m'_0(u) \leq m'(u) - m'_0(u) \leq m_U(u) - m'_0(u).$$

Combining the above two displays and the fact that  $\theta^\top x > s_0$ , we see that for every  $x \in \mathcal{X}$  and  $\theta \in \Theta \cap B_{\theta_0}(1/2)$ ,

$$\begin{aligned} \int_{s_0}^{\theta^\top x} [m_L(u) - m'_0(u)](\{H_{\theta_j}^\top k'(u) + M^*(1 - \eta/T)\mathbf{1}\} \vee 0) du &\preceq U_{\theta, m}^{(1)}(x), \\ U_{\theta, m}^{(1)}(x) &\preceq \int_{s_0}^{\theta^\top x} [m_U(u) - m'_0(u)](\{H_{\theta_j}^\top k'(u) + M^*(1 + \eta/T)\mathbf{1}\} \wedge (2M^*)) du. \end{aligned} \quad (\text{I.6})$$

These bounding functions are not brackets since they depend on  $\theta$  (in the limits of the integral). Since  $m_L, m_U$ , and  $m'$  are 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 (I.6), 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 can be bounded above as follows:

$$\begin{aligned} &\|M_U^{(1)} - M_L^{(1)}\|_{2, P_{\theta_0, m_0}} \\ &\leq 8M^*L\eta\sqrt{d-1} + \left\| \int_{s_0}^{\theta_j^\top \cdot} [m_U(u) - m_L(u)](\{H_{\theta_j}^\top k'(u) + M^*(1 + \eta/T)\mathbf{1}\} \wedge (2M^*)) du \right\|_{2, P_{\theta_0, m_0}} \\ &\quad + \left\| \int_{s_0}^{\theta_j^\top \cdot} [m_U(u) - m'_0(u)] \times \right. \\ &\quad \left. \left[ (\{H_{\theta_j}^\top k'(u) + M^*(1 + \eta/T)\mathbf{1}\} \wedge (2M^*\mathbf{1})) - (\{H_{\theta_j}^\top k'(u) + M^*(1 - \eta/T)\mathbf{1}\} \vee 0) \right] du \right\|_{2, P_{\theta_0, m_0}} \\ &\leq 8M^*L\eta\sqrt{d-1} + \left\| 2M^*\mathbf{1} \int_{s_0}^{\theta_j^\top \cdot} [m_U(u) - m_L(u)] du \right\|_{2, P_{\theta_0, m_0}} \\ &\quad + \left\| 2L \int_{s_0}^{\theta_j^\top \cdot} \left[ (H_{\theta_j}^\top k'(u) + M^*(1 + \eta/T)\mathbf{1}) - (H_{\theta_j}^\top k'(u) + M^*(1 - \eta/T)\mathbf{1}) \right] du \right\|_{2, P_{\theta_0, m_0}} \\ &\leq 8M^*L\eta\sqrt{d-1} + 2M^*\sqrt{d-1} \left( \int_D (m_U(u) - m_L(u))^2 du \right)^{1/2} + 4M^*L\eta\sqrt{d-1}/T \\ &= \sqrt{d-1}(12M^*L\eta + 2M^*\eta). \end{aligned}$$

Thus, we get that  $[M_L^{(1)}, M_U^{(1)}]$  is a  $\sqrt{d-1}(12M^*L + 2M^*)\eta$ -bracket for  $U_{\theta, m}^{(1)}$  with respect to the  $\|\cdot\|_{2, P_{\theta_0, m_0}}$ .

Following very similar arguments, we can show that  $[M_L^{(2)}, M_U^{(2)}]$  forms a bracket for  $U_{\theta, m}^{(2)}$ , where

$$\begin{aligned} M_L^{(2)}(x) &:= \left[ \int_{s_0}^{\theta_j^\top x} [m_L(u) - m'_0(u)] du - 2L\eta \right] \mathbf{1}, \\ M_U^{(2)}(x) &:= \left[ \int_{s_0}^{\theta_j^\top x} [m_U(u) - m'_0(u)] du + 2L\eta \right] \mathbf{1}. \end{aligned}$$

The  $\|\cdot\|_{2, P_{\theta_0, m_0}}$ -length of this bracket is

$$\begin{aligned} \|M_U^{(2)} - M_L^{(2)}\|_{2, P_{\theta_0, m_0}} &\leq 4L\eta\sqrt{d-1} + \sqrt{d-1} \left\| \int_{s_0}^{\theta_j^\top \cdot} (m_U(u) - m_L(u)) du \right\| \\ &\leq 4L\eta\sqrt{d-1} + \eta\sqrt{d-1} = \sqrt{d-1}(4L+1)\eta. \end{aligned}$$

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 (B.1).

Next we show (B.2). Observe that

$$\begin{aligned} \|U_{\theta, m}(x)\|_{2, P_{\theta_0, m_0}}^2 &\leq \left\| \int_{s_0}^{\theta^\top \cdot} [m'(u) - m'_0(u)] k'(u) du \right\|_{2, P_{\theta_0, m_0}}^2 + \left\| (m' - m'_0)(\theta^\top \cdot) k(\theta^\top \cdot) \right\|_{2, P_{\theta_0, m_0}}^2 \\ &\leq \left\| \int_{s_0}^{\theta_0^\top \cdot} [m'(u) - m'_0(u)] k'(u) du \right\|_{2, P_{\theta_0, m_0}}^2 + \left\| \int_{\theta_0^\top \cdot}^{\theta^\top \cdot} [m'(u) - m'_0(u)] k'(u) du \right\|_{2, P_{\theta_0, m_0}}^2 \\ &\quad + \left\| (m' - m'_0)(\theta^\top \cdot) k(\theta^\top \cdot) \right\|_{2, P_{\theta_0, m_0}}^2 \\ &\leq \mathbf{I} + \mathbf{II} + \mathbf{III}. \end{aligned}$$

Observe that

$$\begin{aligned} \mathbf{I} &= \int_{\mathcal{X}} \left| \int_{s_0}^{\theta_0^\top x} [m'(u) - m'_0(u)] k'(u) du \right|^2 dP_{\theta_0, m_0} \\ &\leq \int_{\mathcal{X}} \int_{D_0} [m'(u) - m'_0(u)]^2 |k'(u)|^2 du dP_{\theta_0, m_0} \\ &\leq \|k'\|_{2, \infty}^2 \int_{D_0} [m'(u) - m'_0(u)]^2 du \leq \|k'\|_{2, \infty}^2 n^{-1/5}, \end{aligned}$$

and

$$\begin{aligned} \mathbf{II} &= \left\| \int_{\theta_0^\top \cdot}^{\theta^\top \cdot} [m'(u) - m'_0(u)] k'(u) du \right\|_{2, P_{\theta_0, m_0}}^2 \leq L^2 \|k'\|_{2, \infty}^2 \|(\theta_0 - \theta)^\top \cdot\|^2 \leq L^2 \|k'\|_{2, \infty}^2 T^2 |\theta_0 - \theta|^2, \\ \mathbf{III} &= \left\| (m' - m'_0)(\theta^\top \cdot) k(\theta^\top \cdot) \right\|_{2, P_{\theta_0, m_0}}^2 \leq \|k'\|_{2, \infty}^2 \left\| (m' - m'_0)(\theta^\top \cdot) \right\|^2 = \|k'\|_{2, \infty}^2 n^{-1/5}. \end{aligned}$$

Combining the above two displays, we have

$$\sup_{(\theta, m) \in \mathcal{C}_{M_1}(n)} \|U_{\theta, m}\|_{2, P_{\theta_0, m_0}}^2 \leq 2\|k'\|_{2, \infty}^2 n^{-1/5} + L^2 \|k'\|_{2, \infty}^2 T^2 n^{-1/5} = K_L^2 n^{-1/5}.$$

### I.3. Proof of Lemma I.1

Observe that by Lemma 4.1 of [44] we can get  $\theta_1, \theta_2, \dots, \theta_{N_{\eta_1}}$ , with  $N_{\eta_1} \lesssim \eta_1^{-d}$  such that for every  $\theta \in \Theta$ , there exists a  $j$  satisfying  $|\theta - \theta_j| \leq \eta_1/T$  and

$$|\theta^\top x - \theta_j^\top x| \leq |\theta - \theta_j| \cdot |x| \leq \eta_1 \quad \forall x \in \mathcal{X}.$$

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 of notation, define  $t_j^{(1)}(x) := \theta_j^\top x - \eta_1, t_j^{(2)}(x) := \theta_j^\top x + \eta_1$ , and

$$\mathcal{G}^* := \{g \mid g : \mathcal{D} \rightarrow \mathbb{R} \text{ is a uniformly bounded increasing function and } \|g\|_\infty \leq S\}.$$

Recall that  $\mathbf{m}$  denotes the Lebesgue measure on  $D$ . By Theorem 2.7.5 of [53], we have that  $N_{[]}(\eta_2, \mathcal{G}^*, L_2(\mathbf{m})) \lesssim \exp(\eta_2^{-1})$ , i.e., there exists  $[l_1, u_1], \dots, [l_{M_{\eta_2}}, u_{M_{\eta_2}}]$  with  $l_i \leq u_i, \int_D |u_i(t) - l_i(t)|^2 dt \leq \eta_2^2$  and  $M_{\eta_2} \lesssim \exp(\eta_2^{-1})$  such that for every  $g \in \mathcal{G}^*$ , we can find a  $k \in \{1, \dots, M_{\eta_2}\}$  such that  $l_k \leq g \leq u_k$ . Without loss of generality we can assume that both  $l_i$  and  $u_i$  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  $[l_k, u_k]$  be the  $\eta_2$ -bracket for  $g$ , then for every  $x \in \mathcal{X}$ ,

$$l_k(t_j^{(1)}(x)) \leq l_k(\theta^\top x) \leq g(\theta^\top x) \leq u_k(\theta^\top x) \leq u_k(t_j^{(2)}(x)),$$

where the outer inequalities follow from the fact that both  $l_k$  and  $u_k$  are increasing functions. Proof of Lemma I.1 will be complete if we can show that

$$\{[l_k \circ t_j^{(1)}, u_k \circ t_j^{(2)}] : 1 \leq j \leq N_{\eta_1}, 1 \leq k \leq M_{\eta_2}\},$$

form a  $L_2(P_{\theta_0, m_0})$  bracket for  $\mathcal{H}^*$ . To complete the proof, we now choose  $\eta_1$  (the bracket length for  $\Theta$ ) and  $\eta_2$  (the bracket length for  $\mathcal{G}^*$ ) such that the  $\|\cdot\|$ -length of each bracket of  $\mathcal{H}^*$  is bounded by  $\varepsilon$ . Note that by the triangle inequality, we have

$$\|u_k \circ t_j^{(2)} - l_k \circ t_j^{(1)}\| \leq \|u_k \circ t_j^{(2)} - l_k \circ t_j^{(2)}\| + \|l_k \circ t_j^{(2)} - l_k \circ t_j^{(1)}\|. \quad (\text{I.7})$$

Assuming that the density (with respect to the Lebesgue measure) of  $X^\top \theta$  is uniformly bounded above (by  $C$ ), we get that

$$\|u_k \circ t_j^{(2)} - l_k \circ t_j^{(2)}\|^2 = \int [u_k(r) - l_k(r)]^2 dP_j(r) \leq C \int [u_k(r) - l_k(r)]^2 dr \leq C\eta_2^2.$$

For the second term in (I.7), we first approximate the lower bracket  $l_k$  by a right-continuous 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 [4]. Since  $l_k$  is bounded (by  $S$  say), we can get an increasing step function  $A : D \rightarrow [-S, S]$ , such that  $\int \{l_k(r) - A(r)\}^2 dr \leq \eta_2^2$ . Let  $v_1 < \dots < v_{A_d}$  denote an points of discontinuity of  $A$ . Then for every  $r \in D$ , we can write

$$A(r) = -S + \sum_{i=1}^{A_d} c_i \mathbb{1}_{\{r \geq v_i\}}, \text{ where } c_i > 0 \text{ and } \sum_{i=1}^{A_d} c_i \leq 2S.$$

Using triangle inequality, we get that

$$\begin{aligned} \|l_k \circ t_j^{(2)} - l_k \circ t_j^{(1)}\| &\leq \|l_k \circ t_j^{(2)} - A \circ t_j^{(2)}\| + \|A \circ t_j^{(2)} - A \circ t_j^{(1)}\| + \|A \circ t_j^{(1)} - l_k \circ t_j^{(1)}\| \\ &\leq \sqrt{C}\eta_2 + \|A \circ t_j^{(2)} - A \circ t_j^{(1)}\| + \sqrt{C}\eta_2, \end{aligned}$$

where  $C$  is the (uniform) upper bound on the density of  $X^\top \theta_j$ . Now observe that

$$\begin{aligned} \|A \circ t_j^{(2)} - A \circ t_j^{(1)}\|^2 &= \mathbb{E} \left[ \sum_{i=1}^{A_d} c_i \left( \mathbb{1}_{\{X^\top \theta_j + \eta_1 \geq v_i\}} - \mathbb{1}_{\{X^\top \theta_j \geq v_i\}} \right) \right]^2 \\ &\leq 2S \mathbb{E} \left| \sum_{i=1}^{A_d} c_i \left( \mathbb{1}_{\{X^\top \theta_j + \eta_1 \geq v_i\}} - \mathbb{1}_{\{X^\top \theta_j \geq v_i\}} \right) \right| \\ &\leq 2S \sum_{i=1}^{A_d} c_i \mathbb{P}(X^\top \theta_j - \eta_1 < v_i \leq X^\top \theta_j + \eta_1) \\ &\leq 2S \sum_{i=1}^{A_d} c_i \mathbb{P}(v_i - \eta_1 \leq X^\top \theta_j < v_i + \eta_1) \\ &\leq 2S \sum_{i=1}^{A_d} c_i (2C\eta_1) \leq 8CS^2\eta_1. \end{aligned}$$

Therefore by choosing  $\eta_2 = \varepsilon/(6\sqrt{C})$  and  $\eta_1 = \varepsilon^2/(32CS^2)$ , we have

$$\|u_k \circ t_j^{(2)} - l_k \circ t_j^{(1)}\| \leq 3\sqrt{C}\eta_2 + 2\sqrt{2CS}\sqrt{\eta_1} \leq \varepsilon.$$

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  $\varepsilon$ .

#### I.4. Proof of Lemma B.2

For every  $(\theta, m) \in \mathcal{C}_{M_1}(n)$ , note that

$$\|(m_0 \circ \theta_0 - m \circ \theta_0)U_{\theta, m}\|_{2, P_{\theta_0, m_0}}^2 \leq 4M_1^2 \|U_{\theta, m}\|_{2, P_{\theta_0, m_0}}^2 \leq 4M_1^2 K_L^2 n^{-1/5} = D_{M_1}^2 n^{-1/5}.$$

Furthermore, note that  $\mathcal{D}_{M_1}^*$  is a class of uniformly bounded functions, i.e.,

$$|(m_0 - m)(\theta_0^\top x)U_{\theta, m}(x)| \leq 2M_1 |U_{\theta, m}(x)| \leq 2M_1 V^*.$$

and by Lemma F.6 there exists a constant  $c$  depending only on  $M_1$  and  $L$  such that

$$N(\varepsilon, \{(m_0 \circ \theta_0 - m \circ \theta_0) : m \in \mathcal{D}_{M_1}^*\}, \|\cdot\|_\infty) = N(\varepsilon, \mathcal{D}_{M_1}^*, \|\cdot\|_\infty) \leq c \exp(c/\sqrt{\varepsilon}).$$

By Lemma B.1 and Lemma 9.25 of [31] (for bracketing entropy of product of uniformly bounded function classes), we have

$$N_{[]}(\varepsilon, \mathcal{D}_{M_1}(n), \|\cdot\|_{2, P_{\theta_0, m_0}}) \leq N_{[]}(\varepsilon, \mathcal{D}_{M_1}^*, \|\cdot\|_{2, P_{\theta_0, m_0}}) \leq c\varepsilon^{-2d} \exp\left(\frac{c}{\sqrt{\varepsilon}} + \frac{c}{\varepsilon}\right).$$

It follows that

$$J_{[]}(\gamma, \mathcal{D}_{M_1}(n), \|\cdot\|_{2, P_{\theta_0, m_0}}) \lesssim \gamma^{\frac{1}{2}}.$$

Now using arguments similar to (H.11) and (H.12) and the maximal inequality in Lemma 3.4.2 of [53] (for uniformly bounded function classes), we have

$$\begin{aligned} & \mathbb{P}\left(\sup_{f \in \mathcal{D}_{M_1}(n)} |\mathbb{G}_n f| > \delta\right) \\ & \lesssim 2\delta^{-1} \sqrt{d-1} \sum_{i=1}^{d-1} \mathbb{E}\left(\sup_{f \in \mathcal{D}_{M_1}(n)} |\mathbb{G}_n f_i|\right) \\ & \lesssim \delta^{-1} J_{[]}(\delta, \mathcal{D}_{M_1}(n), \|\cdot\|_{2, P_{\theta_0, m_0}}) \left(1 + \frac{J_{[]}(\delta, \mathcal{D}_{M_1}(n), \|\cdot\|_{2, P_{\theta_0, m_0}})}{D_{M_1}^2 n^{-1/5} \sqrt{n}} 2M_1 V^*\right) \\ & \lesssim \delta^{-1} \left(\sqrt{D_{M_1} n^{-1/20}} + \frac{2M_1 V^* D_{M_1} n^{-1/10}}{D_{M_1}^2 n^{-1/5} \sqrt{n}}\right) \rightarrow 0, \quad \text{as } n \rightarrow \infty, \end{aligned}$$

where in the first inequality  $f_1, \dots, f_{d-1}$  denote each component of  $f$ .

#### I.5. Proof of Lemma B.3

For every  $(\theta, m) \in \mathcal{C}_{M_1}(n)$ , note that

$$\|[m \circ \theta_0 - m \circ \theta]U_{\theta, m}\|_{2, P_{\theta_0, m_0}}^2 \leq 4M_1^2 \|U_{\theta, m}\|_{2, P_{\theta_0, m_0}}^2 \leq 4M_1^2 K_L^2 n^{-1/5} = D_{M_1}^2 n^{-1/5}.$$

By Lemmas F.5 and F.6, 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 B.1 and Lemma 9.25 of [31] (for bracketing entropy of product of uniformly bounded function classes), we have

$$N_{[]}(\varepsilon, \mathcal{A}_{M_1}(n), \|\cdot\|_{2, P_{\theta_0, m_0}}) \leq N_{[]}(\varepsilon, \mathcal{A}_{M_1}^*, \|\cdot\|_{2, P_{\theta_0, m_0}}) \leq c\varepsilon^{-2d} \exp\left(\frac{c}{\sqrt{\varepsilon}} + \frac{c}{\varepsilon}\right).$$

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 B.2.

### I.6. Proof of Lemma B.4

We first prove the first equation of (B.4). Note that, we have

$$\begin{aligned} & \mathbb{P}(|\sqrt{n}\mathbb{P}_n \epsilon U_{\check{\theta}, \check{m}}(X)| > \delta) \\ & \leq \mathbb{P}\left(\sup_{(\theta, m) \in \mathcal{C}_{M_1}(n)} |\sqrt{n}\mathbb{P}_n \epsilon U_{\theta, m}(X)| > \delta\right) + \mathbb{P}((\check{\theta}, \check{m}) \notin \mathcal{C}_{M_1}(n)) \\ & \leq \mathbb{P}\left(\sup_{f \in \mathcal{W}_{M_1}(n)} |\sqrt{n}\mathbb{P}_n \epsilon f| > \delta\right) + \mathbb{P}((\check{\theta}, \check{m}) \notin \mathcal{C}_{M_1}(n)) \\ & = \mathbb{P}\left(\sup_{f \in \mathcal{W}_{M_1}(n)} |\mathbb{G}_n \epsilon f| > \delta\right) + \mathbb{P}((\check{\theta}, \check{m}) \notin \mathcal{C}_{M_1}(n)), \end{aligned}$$

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

$$\mathbb{P}\left(\sup_{f \in \mathcal{W}_{M_1}(n)} |\mathbb{G}_n \epsilon f| > \delta\right) = o(1). \quad (\text{I.8})$$

We will prove (I.8) by applying Lemma F.7 with  $\mathcal{F} = \mathcal{W}_{M_1}(n)$ . Observe that by Lemma B.1, we have

$$\log N_{[]}(\varepsilon, \mathcal{W}_{M_1}(n), \|\cdot\|_{2, P_{\theta_0, m_0}}) \lesssim \varepsilon^{-1}, \quad \sup_{f \in \mathcal{W}_{M_1}^*} \|f\|_{2, \infty} \leq V^*, \quad \text{and} \quad \sup_{f \in \mathcal{W}_{M_1}(n)} \|f\|_{2, P_{\theta_0, m_0}}^2 \leq K_L^2 n^{-1/5}.$$

By Markov inequality, we have

$$\mathbb{P}\left(\sup_{f \in \mathcal{W}_{M_1}(n)} |\mathbb{G}_n \epsilon f| > \delta\right) \lesssim 2\delta^{-1} \sqrt{d-1} \sum_{i=1}^{d-1} \mathbb{E}\left(\sup_{f \in \mathcal{W}_{M_1}(n)} |\mathbb{G}_n \epsilon f_i|\right),$$

where  $f_1, \dots, f_{d-1}$  denote each component of  $f$ . We can bound each term in the summation of the above display by Lemma F.7 (see Appendix F.3.1) with  $\Phi = V^*$ ,  $\kappa = K_L n^{-1/10}$ , and  $\alpha = -1$ . Thus by (F.13), we have

$$\mathbb{E}\left(\sup_{f \in \mathcal{W}_{M_1}(n)} |\mathbb{G}_n \epsilon f_i|\right) \lesssim \sqrt{K_L} n^{-1/20} + \frac{1}{n^{1/10} \sqrt{n}} \rightarrow 0, \quad \text{as } n \rightarrow \infty.$$

We now verify the second and third equations in (B.4). The proofs are similar to the proof of Lemma A.5. Observe that by (H.14), (H.15), and (H.16) (with  $(\check{m}, \check{\theta})$  instead of  $(\hat{m}, \hat{\theta})$ ), we have

$$\begin{aligned} & |P_{\theta_0, m_0}[(m_0 - \check{m})(\theta_0^\top X) U_{\check{\theta}, \check{m}}(X)]| = O_p(n^{-2/5}) \left[ P_{\theta_0, m_0} |U_{\check{\theta}, \check{m}}(X)|^2 \right]^{1/2}, \\ & |P_{\theta_0, m_0}[(\check{m}(\theta_0^\top X) - \check{m}(\check{\theta}^\top X)) U_{\check{\theta}, \check{m}}(X)]| = O_p(n^{-2/5}) \left[ P_{\theta_0, m_0} |U_{\check{\theta}, \check{m}}(X)|^2 \right]^{1/2}. \end{aligned} \quad (\text{I.9})$$

and

$$\begin{aligned} P_{\theta_0, m_0} |U_{\check{\theta}, \check{m}}(X)|^2 & \leq M^{*2} (d-1) P_{\theta_0, m_0} \left[ (\check{m}' - m'_0)(\check{\theta}^\top X) \right]^2 \\ & \quad + M^{*2} (d-1) T P_{\theta_0, m_0} \left[ \int_{D_{\check{\theta}}} [\check{m}'(u) - m'_0(u)]^2 du \right]. \end{aligned} \quad (\text{I.10})$$

Finally by (3.3) of Theorem 3.4, we have that

$$\int_{D_{\check{\theta}}} \{\check{m}'(u) - m'_0(u)\}^2 du \lesssim \|\check{m}' \circ \check{\theta} - m'_0 \circ \check{\theta}\|^2 = O_p(n^{-4/15}).$$

The required result now follows by combining (I.9) and (I.10).

### I.7. Proof of Theorem B.3

Proof of this theorem follows along the lines of the proof of Theorem A.3. By calculations similar to (H.17), (H.18), and (H.20) (with  $(\hat{m}, \hat{\theta})$  replaced by  $(\tilde{m}, \tilde{\theta})$ ), we have that

$$|P_{\tilde{\theta}, m_0} \psi_{\tilde{\theta}, \tilde{m}}| \lesssim \|m_0 \circ \tilde{\theta} - \tilde{m} \circ \tilde{\theta}\| \|\tilde{m}' \circ \tilde{\theta} - m'_0 \circ \tilde{\theta}\| + \|m'_0\|_\infty \|m_0 \circ \tilde{\theta} - \tilde{m} \circ \tilde{\theta}\| \|h_{\tilde{\theta}} \circ \tilde{\theta} - h_{\theta_0} \circ \tilde{\theta}\|_{2, P_{\tilde{\theta}, m_0}}. \quad (\text{I.11})$$

By Theorem 3.4, we have  $\|\tilde{m}' \circ \tilde{\theta} - m'_0 \circ \tilde{\theta}\| = O_p(n^{-2/15})$ . Furthermore, by Theorems 3.1 and 3.3 and assumption (B2), we have

$$\begin{aligned} \|m_0 \circ \tilde{\theta} - \tilde{m} \circ \tilde{\theta}\| &\leq \|\tilde{m} \circ \tilde{\theta} - m_0 \circ \theta_0\| + \|m_0 \circ \theta_0 - m_0 \circ \tilde{\theta}\| \\ &\leq \|\tilde{m} \circ \tilde{\theta} - m_0 \circ \theta_0\| + L_0 T^2 |\theta_0 - \tilde{\theta}| \\ &= O_p(n^{-2/5}) \end{aligned}$$

and  $\|h_{\tilde{\theta}} \circ \tilde{\theta} - h_{\theta_0} \circ \tilde{\theta}\|_{2, P_{\tilde{\theta}, m_0}} \leq \bar{M} |\tilde{\theta} - \theta_0|$ . Thus the first term on the right hand side of (I.11) is  $O_p(n^{-8/15})$  and the second term on the right hand side of (I.11) is  $O_p(n^{-4/5})$ . Thus  $|P_{\tilde{\theta}, m_0} \psi_{\tilde{\theta}, \tilde{m}}| = o_p(n^{-1/2})$ .

### I.8. Consistency of $\psi_{\tilde{\theta}, \tilde{m}}$

**Lemma I.2.** *If the conditions in Theorem 4.2 hold, then*

$$P_{\theta_0, m_0} |\psi_{\tilde{\theta}, \tilde{m}} - \psi_{\theta_0, m_0}|^2 = o_p(1), \quad (\text{I.12})$$

$$P_{\tilde{\theta}, m_0} |\psi_{\tilde{\theta}, \tilde{m}}|^2 = O_p(1). \quad (\text{I.13})$$

*Proof.* Observe that the proof of (I.13) is identical to the proof of (H.22) (with  $(\hat{\theta}, \hat{m})$  replaced by  $(\tilde{\theta}, \tilde{m})$ ); see (H.24).

We now prove (I.12). By assumption (B2) and calculations similar to (H.23), we have

$$P_{\theta_0, m_0} |\psi_{\tilde{\theta}, \tilde{m}} - \psi_{\theta_0, m_0}|^2 \leq \mathbf{I} + 2\|\sigma^2(\cdot)\|_\infty \mathbf{II} + 2\|\sigma^2(\cdot)\|_\infty \mathbf{III} + 4M_1^2 T^2 \|\sigma^2(\cdot)\|_\infty |\tilde{\theta} - \theta_0|^2,$$

where

$$\begin{aligned} \mathbf{I} &:= P_{\theta_0, m_0} \left| [m_0(\theta_0^\top X) - \tilde{m}(\tilde{\theta}^\top X)] [\tilde{m}'(\tilde{\theta}^\top X)X - (m'_0 h_{\theta_0})(\tilde{\theta}^\top X)] \right|^2, \\ \mathbf{II} &:= P_{\theta_0, m_0} \left| \tilde{m}'(\tilde{\theta}^\top X)X - m'_0(\theta_0^\top X)X \right|^2, \\ \mathbf{III} &:= P_{\theta_0, m_0} \left| (m'_0 h_{\theta_0})(\theta_0^\top X) - (m'_0 h_{\theta_0})(\tilde{\theta}^\top X) \right|^2. \end{aligned}$$

It is enough to show that  $\mathbf{I}$ ,  $\mathbf{II}$ , and  $\mathbf{III}$  are  $o_p(1)$ . By Theorems 3.3 and 3.4, we have

$$\begin{aligned} \mathbf{II} &\leq T^2 P_{\theta_0, m_0} \left| \tilde{m}'(\tilde{\theta}^\top X) - m'_0(\theta_0^\top X) \right|^2 \\ &\leq T^2 P_{\theta_0, m_0} \left| \tilde{m}'(\tilde{\theta}^\top X) - m'_0(\tilde{\theta}^\top X) \right|^2 + T^2 P_{\theta_0, m_0} \left| m'_0(\tilde{\theta}^\top X) - m'_0(\theta_0^\top X) \right|^2 = o_p(1). \end{aligned}$$

For  $\mathbf{I}$ , observe that

$$|\tilde{m}'(\tilde{\theta}^\top x)x - m'_0(\tilde{\theta}^\top x)h_{\theta_0}(\tilde{\theta}^\top x)| \leq |\tilde{m}'(\tilde{\theta}^\top x)x| + |m'_0(\tilde{\theta}^\top x)h_{\theta_0}(\tilde{\theta}^\top x)| \leq LT + L\|h_{\theta_0}\|_{2, \infty}.$$

Moreover, by Theorem 3.1, we have  $\|\tilde{m} \circ \tilde{\theta} - m_0 \circ \theta_0\| \xrightarrow{P} 0$ . Thus,

$$\begin{aligned} \mathbf{I} &= P_{\theta_0, m_0} \left| (m_0(\theta_0^\top X) - \tilde{m}(\tilde{\theta}^\top X))(\tilde{m}'(\tilde{\theta}^\top X)X - (m'_0 h_{\theta_0})(\tilde{\theta}^\top X)) \right|^2 \\ &\leq (LT + L\|h_{\theta_0}\|_{2, \infty}) \|m_0 \circ \theta_0 - \tilde{m} \circ \tilde{\theta}\|^2 = o_p(1). \end{aligned}$$

Finally, we have

$$\begin{aligned} \mathbf{III} &= P_{\theta_0, m_0} \left| (m'_0 h_{\theta_0})(\theta_0^\top X) - (m'_0 h_{\theta_0})(\tilde{\theta}^\top X) \right|^2 \\ &\leq P_{\theta_0, m_0} \left[ \|m'_0 h_{\theta_0} + m'_0 h'_{\theta_0}\|_{2, \infty} |(\theta_0 - \tilde{\theta})^\top X| \right]^2 \\ &\leq \|m'_0 h_{\theta_0} + m'_0 h'_{\theta_0}\|_{2, \infty}^2 T^2 |\theta_0 - \tilde{\theta}|^2 = o_p(1). \end{aligned} \quad \square$$



### I.9. Proof of Theorem B.4

Observe that (H.25) and (H.26) imply that

$$\psi_{\check{\theta}, \check{m}} - \psi_{\theta_0, m_0} = \epsilon \tau_{\check{\theta}, \check{m}} + v_{\check{\theta}, \check{m}}.$$

Thus, for every fixed  $M_1$ , we have

$$\begin{aligned} & \mathbb{P}(|\mathbb{G}_n(\psi_{\check{\theta}, \check{m}} - \psi_{\theta_0, m_0})| > \delta) \\ & \leq \mathbb{P}(|\mathbb{G}_n(\epsilon \tau_{\check{\theta}, \check{m}} + v_{\check{\theta}, \check{m}})| > \delta, (\check{\theta}, \check{m}) \in \mathcal{C}_{M_1}(n)) + \mathbb{P}((\check{\theta}, \check{m}) \notin \mathcal{C}_{M_1}(n)) \\ & \leq \mathbb{P}\left(|\mathbb{G}_n(\epsilon \tau_{\check{\theta}, \check{m}})| > \frac{\delta}{2}, (\check{\theta}, \check{m}) \in \mathcal{C}_{M_1}(n)\right) \\ & \quad + \mathbb{P}\left(|\mathbb{G}_n v_{\check{\theta}, \check{m}}| > \frac{\delta}{2}, (\check{\theta}, \check{m}) \in \mathcal{C}_{M_1}(n)\right) + \mathbb{P}((\check{\theta}, \check{m}) \notin \mathcal{C}_{M_1}(n)) \\ & \leq \mathbb{P}\left(\sup_{(\theta, m) \in \mathcal{C}_{M_1}(n)} |\mathbb{G}_n \epsilon \tau_{\theta, m}| > \frac{\delta}{2}\right) + \mathbb{P}\left(\sup_{(\theta, m) \in \mathcal{C}_{M_1}(n)} |\mathbb{G}_n v_{\theta, m}| > \frac{\delta}{2}\right) \\ & \quad + \mathbb{P}((\check{\theta}, \check{m}) \notin \mathcal{C}_{M_1}(n)). \end{aligned} \tag{I.14}$$

Recall that by Theorems 3.1–3.4, we have  $\mathbb{P}((\check{\theta}, \check{m}) \notin \mathcal{C}_{M_1}(n)) = o(1)$ . Thus the proof of Theorem B.4 will be complete if we show that the first two terms in (I.14) are  $o(1)$ . Lemmas I.3 and I.4 do this.

**Lemma I.3.** Fix  $M_1$  and  $\delta > 0$ . For  $n \in \mathbb{N}$ , as  $n \rightarrow \infty$ , we have

$$\mathbb{P}\left(\sup_{(\theta, m) \in \mathcal{C}_{M_1}(n)} |\mathbb{G}_n \epsilon \tau_{\theta, m}| > \frac{\delta}{2}\right) \rightarrow 0.$$

*Proof.* Recall that

$$\begin{aligned} \tau_{\theta, m}(x) &:= H_{\theta}^{\top} \{[m'(\theta^{\top} x) - m'_0(\theta_0^{\top} x)]x + [(m'_0 h_{\theta_0})(\theta_0^{\top} x) - (m'_0 h_{\theta_0})(\theta^{\top} x)]\} \\ &\quad + (H_{\theta}^{\top} - H_{\theta_0}^{\top})[m'_0(\theta_0^{\top} x)x - (m'_0 h_{\theta_0})(\theta_0^{\top} x)], \end{aligned}$$

Let us define,

$$\Xi_{M_1}(n) := \{\tau_{\theta, m} | (\theta, m) \in \mathcal{C}_{M_1}(n)\} \quad \text{and} \quad \Xi_{M_1}^* := \{\tau_{\theta, m} | (\theta, m) \in \mathcal{C}_{M_1}^*\}.$$

We will prove that

$$N(\varepsilon, \Xi_{M_1}^*, \|\cdot\|_{\infty}) \leq c \exp(c/\varepsilon) \varepsilon^{-10d}, \tag{I.15}$$

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 I.1, we have

$$N_{[]}(\varepsilon, \{m'(\theta^{\top} \cdot) | (\theta, m) \in \mathcal{C}_{M_1}^*\}, \|\cdot\|) \lesssim \exp(1/\varepsilon). \tag{I.16}$$

Moreover, by Lemma 15 of [32], we can find a  $\theta_1, \theta_2, \dots, \theta_{N_{\varepsilon}}$  with  $N_{\varepsilon} \lesssim \varepsilon^{-2d}$  such that for every  $\theta \in \Theta \cap B_{\theta_0}(1/2)$ , there exists a  $\theta_j$  such that

$$|\theta - \theta_j| \leq \varepsilon/T, \quad \|H_{\theta} - H_{\theta_j}\|_2 \leq \varepsilon/T, \quad \text{and} \quad |\theta^{\top} x - \theta_j^{\top} x| \leq \varepsilon, \quad \forall x \in \mathcal{X}.$$

Observe that for all  $x \in \mathcal{X}$ , we have  $H_{\theta_j}^{\top} x - \varepsilon \preceq H_{\theta}^{\top} x \preceq H_{\theta_j}^{\top} x + \varepsilon$ . Thus

$$N_{[]}(\varepsilon, \{f : \mathcal{X} \rightarrow \mathbb{R}^d | f(x) = H_{\theta}^{\top} x, \forall x \in \mathcal{X}, \theta \in \Theta \cap B_{\theta_0}(1/2)\}, \|\cdot\|_{2, \infty}) \lesssim \varepsilon^{-2d} \tag{I.17}$$

Similarly as  $|m'_0(\theta^{\top} x) - m'_0(\theta_j^{\top} x)| \leq L_0 \varepsilon$ , we have

$$N_{[]}(\varepsilon, \{m'_0 \circ \theta : \theta \in \Theta \cap B_{\theta_0}(1/2)\}, \|\cdot\|) \lesssim \varepsilon^{-2d} \tag{I.18}$$

Finally observe that

$$\begin{aligned} & |H_{\theta}^{\top} h_{\theta_0}(\theta^{\top} x) - H_{\theta_j}^{\top} h_{\theta_0}(\theta_j^{\top} x)| \\ & \leq |H_{\theta}^{\top} h_{\theta_0}(\theta^{\top} x) - H_{\theta}^{\top} h_{\theta_0}(\theta_j^{\top} x)| + |H_{\theta}^{\top} h_{\theta_0}(\theta_j^{\top} x) - H_{\theta_j}^{\top} h_{\theta_0}(\theta_j^{\top} x)| \\ & \leq |h_{\theta_0}(\theta^{\top} x) - h_{\theta_0}(\theta_j^{\top} x)| + \|H_{\theta}^{\top} - H_{\theta_j}^{\top}\|_2 \|h_{\theta_0}\|_{2, \infty} \\ & \leq \|h'_{\theta_0}\|_{2, \infty} |\theta - \theta_j| T + \|H_{\theta}^{\top} - H_{\theta_j}^{\top}\|_2 \|h_{\theta_0}\|_{2, \infty} \leq \varepsilon (\|h'_{\theta_0}\|_{2, \infty} + \|h_{\theta_0}\|_{2, \infty}/T) \lesssim \varepsilon \end{aligned}$$

and

$$|H_{\theta}^{\top} h_{\theta_0}(\theta_0^{\top} x) - H_{\theta_j}^{\top} h_{\theta_0}(\theta_0^{\top} x)| \leq \|h_{\theta_0}(\theta_0^{\top} x)\|_{2,\infty} \varepsilon / T.$$

Thus we have

$$N_{[]}(\varepsilon, \{f : \chi \rightarrow \mathbb{R}^d | f(x) = H_{\theta}^{\top} h_{\theta_0}(\theta^{\top} x), \theta \in \Theta \cap B_{\theta_0}(1/2)\}, \|\cdot\|_{2,\infty}) \lesssim \varepsilon^{-2d}, \quad (\text{I.19})$$

$$N_{[]}(\varepsilon, \{f : \chi \rightarrow \mathbb{R}^d | f(x) = H_{\theta}^{\top} h_{\theta_0}(\theta_0^{\top} x), \theta \in \Theta \cap B_{\theta_0}(1/2)\}, \|\cdot\|_{2,\infty}) \lesssim \varepsilon^{-2d}. \quad (\text{I.20})$$

Thus by applying Lemma 9.25 of [31] to sums and product of classes of functions in (I.16), (I.17), (I.18), (I.19), and (I.20) we have (I.15).

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  $x \in \chi$  note that,

$$\begin{aligned} |\tau_{\theta,m}(x)| &\leq [|m'(\theta^{\top} x) - m'(\theta_0^{\top} x)| + |m'(\theta_0^{\top} x) - m'_0(\theta_0^{\top} x)|] |x| \\ &\quad + |m'_0(\theta_0^{\top} x) h_{\theta_0}(\theta_0^{\top} x) - m'_0(\theta^{\top} x) h_{\theta_0}(\theta_0^{\top} x)| \\ &\quad + |m'_0(\theta^{\top} x) h_{\theta_0}(\theta_0^{\top} x) - m'_0(\theta^{\top} x) h_{\theta_0}(\theta^{\top} x)| \\ &\quad + |\theta - \theta_0| |m'_0(\theta^{\top} x) x - (m'_0 h_{\theta_0})(\theta^{\top} x)| \\ &\leq L |\theta^{\top} x - \theta_0^{\top} x| |x| + \|m' - m'_0\|_{D_0} |x| \\ &\quad + |h_{\theta_0}(\theta_0^{\top} x)| \|m''_0\|_{\infty} |\theta_0^{\top} x - \theta^{\top} x| \\ &\quad + |m'_0(\theta^{\top} x)| \|h'_{\theta_0}\|_{2,\infty} |\theta_0^{\top} x - \theta^{\top} x| \\ &\quad + |\theta - \theta_0| |m'_0(\theta^{\top} x) x - (m'_0 h_{\theta_0})(\theta^{\top} x)| \\ &\leq |\theta - \theta_0| L T^2 + n^{-1/5} T + T^2 \|m''_0\|_{\infty} T |\theta - \theta_0| + L_0 \|h'_{\theta_0}\|_{2,\infty} T |\theta - \theta_0| + |\theta - \theta_0| L_0 T. \\ &\leq C_{11} n^{-1/10}, \end{aligned}$$

where  $C_{11}$  is constant depending only on  $L, L_0, T, m_0$ , and  $h_{\theta_0}$ . Now observe that,

$$\mathbb{P}\left(\sup_{f \in \Xi_{M_1}(n)} |\mathbb{G}_n \epsilon f| > \frac{\delta}{2}\right) \leq 2\delta^{-1} \sqrt{d-1} \sum_{i=1}^{d-1} \mathbb{E}\left(\sup_{f \in \Xi_{M_1}(n)} |\mathbb{G}_n \epsilon f_i|\right)$$

where  $f_1, \dots, f_{d-1}$  denote each component of  $f$ . We can bound each term in the summation of the above display by Lemma F.7 with  $\Phi = \kappa = C_{11} n^{-1/10}$ , and  $\alpha = -1$ . By (F.13), we have

$$\mathbb{E}\left(\sup_{f \in \Xi_{M_1}(n)} |\mathbb{G}_n \epsilon f_i|\right) \lesssim n^{-1/20} + n^{-4/10} = o(1) \quad \square$$

**Lemma I.4.** Fix  $M_1$  and  $\delta > 0$ . For  $n \in \mathbb{N}$ , we have

$$\mathbb{P}\left(\sup_{(\theta, m) \in \mathcal{C}_{M_1}(n)} |\mathbb{G}_n v_{\theta, m}| > \frac{\delta}{2}\right) = o_p(1).$$

*Proof.* Recall that

$$v_{\theta, m}(x) := [m_0(\theta_0^{\top} x) - m(\theta^{\top} x)] [m'(\theta^{\top} x) H_{\theta}^{\top} x - m'_0(\theta^{\top} x) H_{\theta}^{\top} h_{\theta_0}(\theta^{\top} x)].$$

We will first show that

$$J_{[]}(\nu, \{v_{\theta, m} : (\theta, m) \in \mathcal{C}_{M_1}(n)\}, \|\cdot\|_{2, P_{\theta_0, m_0}}) \lesssim \nu^{1/2} \quad (\text{I.21})$$

By Lemmas F.5 and I.1 and (I.17), (I.18), and (I.19), we have

$$\begin{aligned} N_{[]}(\varepsilon, \{m_0(\theta_0^{\top} \cdot) - m(\theta^{\top} \cdot) | (\theta, m) \in \mathcal{C}_{M_1}^*\}, \|\cdot\|_{\infty}) &\lesssim \exp(1/\sqrt{\varepsilon}), \\ N_{[]}(\varepsilon, \{m'(\theta^{\top} \cdot) | (\theta, m) \in \mathcal{C}_{M_1}^*\}, \|\cdot\|) &\lesssim \exp(1/\varepsilon), \\ N_{[]}(\varepsilon, \{f : \chi \rightarrow \mathbb{R}^d | f(x) = H_{\theta}^{\top} x, \forall x \in \chi, \theta \in \Theta \cap B_{\theta_0}(1/2)\}, \|\cdot\|_{2,\infty}) &\lesssim \varepsilon^{-2d} \\ N_{[]}(\varepsilon, \{m'_0 \circ \theta : \theta \in \Theta \cap B_{\theta_0}(1/2)\}, \|\cdot\|) &\lesssim \varepsilon^{-2d} \\ N_{[]}(\varepsilon, \{f : \chi \rightarrow \mathbb{R}^d | f(x) = H_{\theta}^{\top} h_{\theta_0}(\theta^{\top} x), \theta \in \Theta \cap B_{\theta_0}(1/2)\}, \|\cdot\|_{2,\infty}) &\lesssim \varepsilon^{-2d}. \end{aligned} \quad (\text{I.22})$$

Thus by applying Lemma 9.25 of [31] to sums and product of classes of functions in (I.22), we have

$$N_{[]}(\varepsilon, \{v_{\theta,m} : (\theta, m) \in \mathcal{C}_{M_1}^*\}, \|\cdot\|_{2, P_{\theta_0, m_0}}) \lesssim \exp\left(\frac{1}{\varepsilon} + \frac{1}{\sqrt{\varepsilon}}\right) \varepsilon^{-6d}.$$

Now (I.21) follows from the definition of  $J_{[]}$  by observing that

$$J_{[]}(\nu, \{v_{\theta,m} : (\theta, m) \in \mathcal{C}_{M_1}(n)\}, \|\cdot\|_{2, P_{\theta_0, m_0}}) \leq J_{[]}(\nu, \{v_{\theta,m} : (\theta, m) \in \mathcal{C}_{M_1}^*\}, \|\cdot\|_{2, P_{\theta_0, m_0}}).$$

Now we will find  $\sup_{(\theta, m) \in \mathcal{C}_{M_1}(n)} \|v_{\theta,m}\|_{2, \infty}$ . For every  $x \in \mathcal{X}$  observe that,

$$\begin{aligned} |v_{\theta,m}(x)| &\leq |m_0(\theta_0^\top x) - m(\theta_0^\top x)| \cdot |m'(\theta^\top x)x - m'_0(\theta^\top x)h_{\theta_0}(\theta^\top x)| \\ &\quad + |m(\theta_0^\top x) - m(\theta^\top x)| \cdot |m'(\theta^\top x)x - m'_0(\theta^\top x)h_{\theta_0}(\theta^\top x)| \\ &\leq \|m_0 - m\|_{D_0} |m'(\theta^\top x)x - m'_0(\theta^\top x)h_{\theta_0}(\theta^\top x)| \\ &\quad + L|\theta_0^\top x - \theta^\top x| |m'(\theta^\top x)x - m'_0(\theta^\top x)h_{\theta_0}(\theta^\top x)| \\ &\leq b_n^{-1} 2LT + 2T^2 L^2 M_2 |\theta - \theta_0| \\ &\leq C[b_n^{-1} + n^{-1/10}], \end{aligned}$$

where  $C$  is a constant depending only on  $T, L$ , and  $M_1$ . Thus

$$\sup_{(\theta, m) \in \mathcal{C}_{M_1}(n)} \|v_{\theta,m}\|_{2, P_{\theta_0, m_0}} \leq \sup_{(\theta, m) \in \mathcal{C}_{M_1}(n)} \|v_{\theta,m}\|_{2, \infty} \leq C^2 [b_n^{-1} + n^{-1/10}].$$

Now using arguments similar to (H.11) and (H.12) and the maximal inequality in Lemma 3.4.2 of [53] (for uniformly bounded function classes), we have

$$\begin{aligned} &\mathbb{P}\left(\sup_{(\theta, m) \in \mathcal{C}_{M_1}(n)} |\mathbb{G}_n v_{\theta,m}| > \frac{\delta}{2}\right) \\ &\lesssim 2\delta^{-1} \sqrt{d-1} \sum_{i=1}^{d-1} \mathbb{E}\left(\sup_{(\theta, m) \in \mathcal{C}_{M_1}(n)} |\mathbb{G}_n v_{\theta,m,i}|\right) \\ &\lesssim J_{[]}([b_n^{-1} + n^{-1/10}], \mathcal{W}_{M_1}(n), \|\cdot\|_{2, P_{\theta_0, m_0}}) + \frac{J_{[]}^2([b_n^{-1} + n^{-1/10}], \mathcal{W}_{M_1}(n), \|\cdot\|_{2, P_{\theta_0, m_0}})}{[b_n^{-1} + n^{-1/10}]^2 \sqrt{n}} \\ &\lesssim [b_n^{-1} + n^{-1/10}]^{1/2} + \frac{[b_n^{-1} + n^{-1/10}]}{[b_n^{-1} + n^{-1/10}]^2 \sqrt{n}} \\ &\lesssim [b_n^{-1} + n^{-1/10}]^{1/2} + \frac{1}{b_n^{-1} \sqrt{n} + n^{4/10}} = o(1), \end{aligned}$$

as  $b_n = o(n^{1/2})$ , here in the first inequality  $v_{\theta,m,1}, \dots, v_{\theta,m,d-1}$  denote each component of  $v_{\theta,m}$ .  $\square$

## References

- [1] Agmon, S. (2010). *Lectures on elliptic boundary value problems*. AMS Chelsea Publishing, Providence, RI. Prepared for publication by B. Frank Jones, Jr. with the assistance of George W. Batten, Jr., Revised edition of the 1965 original.
- [2] Aït-Sahalia, Y. and J. Duarte (2003). Nonparametric option pricing under shape restrictions. *J. Econometrics* 116(1-2), 9–47. Frontiers of financial econometrics and financial engineering.
- [3] Balabdaoui, F., C. Durot, and H. Jankowski (2016, October). Least squares estimation in the monotone single index model. *ArXiv e-prints*.
- [4] Bogachev, V. I. (2007). *Measure theory. Vol. I, II*. Springer-Verlag, Berlin.
- [5] Breiman, L. and J. H. Friedman (1985). Estimating optimal transformations for multiple regression and correlation. *Journal of the American statistical Association* 80(391), 580–598.

- [6] Carroll, R. J., J. Fan, I. Gijbels, and M. P. Wand (1997). Generalized partially linear single-index models. *Journal of the American Statistical Association* 92(438), 477–489.
- [7] Chen, C.-H. and K.-C. Li (1998). Can SIR be as popular as multiple linear regression? *Statist. Sinica* 8(2), 289–316.
- [8] Chen, D. and R. J. Plemmons (2010). Nonnegativity constraints in numerical analysis. In *The birth of numerical analysis*, pp. 109–139. World Sci. Publ., Hackensack, NJ.
- [9] Chen, Y. and R. J. Samworth (2014, April). Generalised additive and index models with shape constraints. *ArXiv e-prints*.
- [10] Cheng, G., Y. Zhao, and B. Li (2012). Empirical likelihood inferences for the semiparametric additive isotonic regression. *J. Multivariate Anal.* 112, 172–182.
- [11] Cui, X., W. K. Härdle, and L. Zhu (2011). The EFM approach for single-index models. *Ann. Statist.* 39(3), 1658–1688.
- [12] Delecroix, M., M. Hristache, and V. Patilea (2006). On semiparametric m-estimation in single-index regression. *Journal of Statistical Planning and Inference* 136(3), 730–769.
- [13] Dontchev, A. L., H. Qi, and L. Qi (2003). Quadratic convergence of Newton’s method for convex interpolation and smoothing. *Constr. Approx.* 19(1), 123–143.
- [14] Dümbgen, L., S. Freitag, and G. Jongbloed (2004). Consistency of concave regression with an application to current-status data. *Math. Methods Statist.* 13(1), 69–81.
- [15] Durrett, R. (2010). *Probability: theory and examples* (Fourth ed.). Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge.
- [16] Elfving, T. and L.-E. Andersson (1988). An algorithm for computing constrained smoothing spline functions. *Numer. Math.* 52(5), 583–595.
- [17] Groeneboom, P. and K. Hendrickx (2016, January). Current status linear regression. *ArXiv e-prints arXiv:1601.00202*.
- [18] Guntuboyina, A. and B. Sen (2013, April). Covering numbers for convex functions. *Information Theory, IEEE Transactions on* 59(4), 1957–1965.
- [19] Györfi, L., M. Kohler, A. Krzyżak, and H. Walk (2002). *A distribution-free theory of nonparametric regression*. Springer Series in Statistics. Springer-Verlag, New York.
- [20] Hájek, J. (1972). Local asymptotic minimax and admissibility in estimation. In *Proceedings of the sixth Berkeley symposium on mathematical statistics and probability*, Volume 1, pp. 175–194.
- [21] Härdle, W., P. Hall, and H. Ichimura (1993). Optimal smoothing in single-index models. *Ann. Statist.* 21(1), 157–178.
- [22] Harrison, D. and D. L. Rubinfeld (1978). Hedonic housing prices and the demand for clean air. *Journal of environmental economics and management* 5(1), 81–102.
- [23] Hatcher, A. (2002). *Algebraic topology*. Cambridge University Press, Cambridge.
- [24] Horowitz, J. L. (1998). *Semiparametric methods in econometrics*, Volume 131 of *Lecture Notes in Statistics*. Springer-Verlag, New York.
- [25] Hristache, M., A. Juditsky, and V. Spokoiny (2001). Direct estimation of the index coefficient in a single-index model. *Ann. Statist.* 29(3), 595–623.
- [26] Huang, J. (1996). Efficient estimation for the proportional hazards model with interval censoring. *Ann. Statist.* 24(2), 540–568.
- [27] Huang, J. and J. A. Wellner (1997). Interval censored survival data: a review of recent progress.
- [28] Ichimura, H. (1993). Semiparametric least squares (SLS) and weighted SLS estimation of single-index models. *J. Econometrics* 58(1-2), 71–120.
- [29] Keshavarz, A., Y. Wang, and S. Boyd (2011). Imputing a convex objective function. In *2011 IEEE International Symposium on Intelligent Control*, pp. 613–619. IEEE.
- [30] Klaassen, C. A. J. (1987). Consistent estimation of the influence function of locally asymptotically linear estimators. *Ann. Statist.* 15(4), 1548–1562.
- [31] Kosorok, M. R. (2008). *Introduction to empirical processes and semiparametric inference*. Springer Series in Statistics. Springer, New York.
- [32] Kuchibhotla, A. K. and R. K. Patra (2016a). Efficient estimation in single index models through

- smoothing splines. *arXiv preprint arXiv:1612.00068v2*.
- [33] Kuchibhotla, A. K. and R. K. Patra (2016b). *simest: Single Index Model Estimation with Constraints on Link Function*. R package version 0.6.
  - [34] Lawson, C. L. and R. J. Hanson (1974). *Solving least squares problems*. Prentice-Hall, Inc., Englewood Cliffs, N.J. Prentice-Hall Series in Automatic Computation.
  - [35] Li, K.-C. and N. Duan (1989). Regression analysis under link violation. *Ann. Statist.* 17(3), 1009–1052.
  - [36] Li, Q. and J. S. Racine (2007). *Nonparametric econometrics*. Princeton University Press, Princeton, NJ. Theory and practice.
  - [37] Li, W. and V. Patilea (2015). A new inference approach for single-index models.
  - [38] Ma, Y. and L. Zhu (2013). Doubly robust and efficient estimators for heteroscedastic partially linear single-index models allowing high dimensional covariates. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* 75(2), 305–322.
  - [39] Matzkin, R. L. (1991). Semiparametric estimation of monotone and concave utility functions for polychotomous choice models. *Econometrica* 59(5), 1315–1327.
  - [40] Murphy, S. A. and A. W. van der Vaart (2000). On profile likelihood. *J. Amer. Statist. Assoc.* 95(450), 449–485. With comments and a rejoinder by the authors.
  - [41] Murphy, S. A., A. W. van der Vaart, and J. A. Wellner (1999). Current status regression. *Math. Methods Statist.* 8(3), 407–425.
  - [42] Newey, W. K. (1990). Semiparametric efficiency bounds. *Journal of applied econometrics* 5(2), 99–135.
  - [43] Newey, W. K. and T. M. Stoker (1993). Efficiency of weighted average derivative estimators and index models. *Econometrica* 61(5), 1199–1223.
  - [44] Pollard, D. (1990). *Empirical processes: theory and applications*. NSF-CBMS Regional Conference Series in Probability and Statistics, 2. Institute of Mathematical Statistics, Hayward, CA; American Statistical Association, Alexandria, VA.
  - [45] Powell, J. L., J. H. Stock, and T. M. Stoker (1989). Semiparametric estimation of index coefficients. *Econometrica* 57(6), 1403–1430.
  - [46] Samworth, R. J. and M. Yuan (2012). Independent component analysis via nonparametric maximum likelihood estimation. *Ann. Statist.* 40(6), 2973–3002.
  - [47] Seijo, E. and B. Sen (2011, 06). Nonparametric least squares estimation of a multivariate convex regression function. *The Annals of Statistics* 39(3), 1633–1657.
  - [48] Tsiatis, A. A. (2006). *Semiparametric theory and missing data*. Springer Series in Statistics. Springer, New York.
  - [49] van de Geer, S. (1990). Estimating a regression function. *The Annals of Statistics*, 907–924.
  - [50] van de Geer, S. A. (2000). *Applications of empirical process theory*, Volume 6 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge.
  - [51] van der Vaart, A. (2002). Semiparametric statistics. In *Lectures on probability theory and statistics (Saint-Flour, 1999)*, Volume 1781 of *Lecture Notes in Math.*, pp. 331–457. Springer, Berlin.
  - [52] van der Vaart, A. W. (1998). *Asymptotic statistics*, Volume 3 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge.
  - [53] van der Vaart, A. W. and J. A. Wellner (1996). *Weak convergence and empirical processes*. Springer Series in Statistics. Springer-Verlag, New York. With applications to statistics.
  - [54] Varian, H. R. (1984). The nonparametric approach to production analysis. *Econometrica* 52(3), 579–597.
  - [55] Wahba, G. (1990). *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.
  - [56] Wang, J. and L. Yang (2009). Efficient and fast spline-backfitted kernel smoothing of additive models. *Ann. Inst. Statist. Math.* 61(3), 663–690.

- [57] Wang, J.-L., L. Xue, L. Zhu, and Y. S. Chong (2010). Estimation for a partial-linear single-index model. *Ann. Statist.* 38(1), 246–274.
- [58] Wen, Z. and W. Yin (2013). A feasible method for optimization with orthogonality constraints. *Math. Program.* 142(1-2, Ser. A), 397–434.
- [59] Xia, Y., H. Tong, W. Li, and L.-X. Zhu (2002). An adaptive estimation of dimension reduction space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 64(3), 363–410.
- [60] Yu, K., E. Mammen, and B. U. Park (2011). Semi-parametric regression: efficiency gains from modeling the nonparametric part. *Bernoulli* 17(2), 736–748.