Bernoulli **25**(4B), 2019, 3276–3310
https://doi.org/10.3150/18-BEJ1090

Least squares estimation in the monotone single index model

FADOUA BALABDAOUI[1,2], CÉCILE DUROT[3] and HANNA JANKOWSKI[4]

[1]*Seminar für Statistik, ETH Zürich, 8092, Zürich, Schweiz. E-mail: fadoua.balabdaoui@stat.math.ethz.ch*
[2]*Université Paris-Dauphine, PSL Research University, CNRS, CEREMADE, 75016 Paris, France*
[3]*Modal'x, UPL, Univ Paris Nanterre, F92000 Nanterre, France. E-mail: cecile.durot@gmail.com*
[4]*Department of Mathematics and Statistics, York University, 4700 Keele Street, Toronto ON, M3J 1P3, Canada. E-mail: hkj@yorku.ca*

We study the monotone single index model where a real response variable Y is linked to a d-dimensional covariate X through the relationship $E[Y|X] = \Psi_0(\alpha_0^T X)$, almost surely. Both the ridge function, Ψ_0, and the index parameter, α_0, are unknown and the ridge function is assumed to be monotone. Under some appropriate conditions, we show that the rate of convergence in the L_2-norm for the least squares estimator of the bundled function $\Psi_0(\alpha_0^T \cdot)$ is $n^{1/3}$. A similar result is established for the isolated ridge function, and the index is shown to converge at least at the rate $n^{1/3}$. Since the least squares estimator of the index is computationally intensive, we also consider alternative estimators of the index α_0 from earlier literature. Moreover, we show that if the rate of convergence of such an alternative estimator is at least $n^{1/3}$, then the corresponding least-squares type estimators (obtained via a "plug-in" approach) of both the bundled and isolated ridge functions still converge at the rate $n^{1/3}$.

Keywords: least squares; maximum likelihood; monotone; semi-parametric; shape-constraints; single-index model

1. Introduction

1.1. The generalized linear model and the single index model

The generalized linear model is widely used in econometrics and biometrics as a standard tool in parametric regression analysis, see, *for example*, Dobson and Barnett [9]. It assumes that the observations are n i.i.d. copies of a random pair (X, Y) such that Y is real valued, X is d-dimensional, and

$$E(Y|X) = \Psi_0\bigl(\alpha_0^T X\bigr) \tag{1.1}$$

almost surely with an unknown index $\alpha_0 \in \mathbb{R}^d \setminus \{0\}$ and a monotone ridge function Ψ_0. In the generalized linear model, Ψ_0 is assumed to be known and the conditional density of Y given $X = x$ with respect to a given dominating measure (typically either Lebesgue measure or counting measure) is assumed to be of the form

$$y \mapsto h(y, \phi) \exp\left\{\frac{y\ell(\mu(x)) - B \circ \ell(\mu(x))}{\phi}\right\}, \tag{1.2}$$

1350-7265

where h is the normalizing function, $\mu(x)$ is the mean, $\phi > 0$ is a possibly unknown dispersion parameter, ℓ is a given real valued function with first derivative $\ell' > 0$, and inverse $\ell^{-1} = B'$. The generalized linear model includes very popular parametric regression models but nevertheless, it lacks the flexibility offered by non-parametric approaches.

The single index model extends the generalized linear model in order to gain more flexibility. It is widely used, for instance, in econometrics, as a compromise between restrictive parametric assumptions and a fully non-parametric setting that can suffer from the "curse of dimensionality" in high-dimensional problems, see, for example, Chapter 8 in Li and Racine [21]. It assumes that the conditional expectation of Y depends only on the linear predictor $\alpha_0^T X$. Hence, as in the generalized linear model, we have $E(Y|X) = \Psi_0(\alpha_0^T X)$ almost surely, however, the ridge function Ψ_0 is now unknown. Furthermore, it is no longer assumed that the conditional distribution of Y given X takes the form (1.2), making the model even more flexible.

Standard methods for estimating α_0 and Ψ_0 rely on smoothness assumptions on Ψ_0, and hence involve a smoothing parameter which has to be carefully chosen, see, *for example*, Härdle, Hall and Ichimura [16], Chiou and Müller [5], Hristache, Juditsky and Spokoiny [17] and references therein. Note also that α_0 and Ψ_0 are not identifiable if left unrestricted. To see this, let $\|\alpha_0\|$ denote the Euclidean norm of α_0, and note that $\Psi_0(\alpha_0^T x) = \Phi_0(\beta_0^T x)$ if $\beta_0 = \alpha_0/\|\alpha_0\|$ and $\Phi_0(t) = \Psi_0(\|\alpha_0\| t)$ for all t. Similarly, $\Psi_0(\alpha_0^T x) = \Phi_0(\beta_0^T x)$ if $\beta_0 = -\alpha_0$ and $\Phi_0(t) = \Psi_0(-t)$ for all t. This issue could be resolved by assuming, for example, that $\|\alpha_0\| = 1$ and the first non-null entry of α_0 is positive. Under some additional constraints on Ψ_0 and the distribution of X, the model can be shown to be identifiable, see, for example, Proposition 5.1 below.

1.2. The monotone single index model

In this paper, we assume that the unknown ridge function in the single index model is monotone. This is motivated by the fact that monotonicity appears naturally in various applications, which is one of the reasons behind the popularity of the generalized linear model. Moreover, the monotonicity assumption has a great advantage. Estimators based only on smoothness conditions on the ridge function typically depend on a tuning parameter that has to be chosen by the practitioner. The monotonicity assumption avoids all this by opening the door to non-parametric estimators which are completely data driven, and do not involve any tuning parameters. To be precise, we assume (1.1) where $\alpha_0 \in \mathbb{R}^d \setminus \{0\}$ is such that $\|\alpha_0\| = 1$, and Ψ_0 is assumed to be *non-decreasing*. Note that the assumption made on the direction of monotonicity of the ridge function replaces the assumption that the first non-null entry of α_0 is positive in the identifiability conditions. This can be seen by defining the function $\Phi_0(t) = \Psi_0(-t)$ for $t \in \mathbb{R}$, which is non-increasing if and only if Ψ_0 is non-decreasing. Throughout this paper, we will refer to this model as the monotone single index model, a term that has been used previously in the literature.

The monotone single index model, with the additional assumption that $Y - E(Y|X)$ is independent of X, has been considered by Foster, Taylor and Nan [11], where an estimator for α_0 was proposed based on combining isotonic regression with a smoothing method (which involves a tuning parameter), and also by Kakade et al. [18], where an algorithm for simultaneously estimating the index and the ridge function is provided under the assumption that the ridge function is Lipschitz (the Lipschitz constant is a parameter of the algorithm). This fits in the setting of Han

[15] with (using the notation of that paper) $F(x, u) = f(x) + u$ and $D(t) = t$ with f a monotone function. Han [15] proves consistency of a non-parametric estimator of the index, which does not require a tuning parameter. The monotone single index model is also closely related to the model considered by Chen and Samworth [4], who in contrast to the approach followed here, assume that the conditional distribution of Y given X takes the form (1.2). Chen and Samworth [4] also consider additive index models where, with ℓ as in (1.2), $\ell(E(Y|X))$ can be written as a sum of ridge functions of linear predictors, with each ridge function satisfying a certain shape constraint. The authors show consistency for a slightly modified maximum likelihood estimator obtained by maximizing the likelihood over the closure of the set of all possible parameters.

Current status regression can also be seen as a special case of the monotone single index model. In the current status regression setting, the response $Y \geq 0$ is subjected to interval censoring and is not completely observed. Instead, independent copies of (X, C, Δ) are observed, where $X \in \mathbb{R}^d$ is the predictor, $C \geq 0$ is an observed censoring time independent of Y, and $\Delta = 1_{\{Y \leq C\}}$. Although not observed, Y is assumed to satisfy the linear regression model $Y = \alpha_0^T X + \varepsilon$, where $\alpha_0 \in \mathbb{R}^d$ and ε is independent of (C, X) with unknown distribution function F_0. Let $\tilde{X}$ denote the random vector in $\mathbb{R}^{d+1}$ such that $\tilde{X}^T = (C, X^T)$, $\tilde{\alpha}_0$ the vector in $\mathbb{R}^{d+1}$ such that $\tilde{\alpha}_0^T = (1, -\alpha_0^T)$ and $\tilde{Y} = \Delta$. Then, $E(\tilde{Y}|\tilde{X}) = F_0(\tilde{\alpha}_0^T \tilde{X})$ where F_0 is non-decreasing (since it is a distribution function). Here, the conditional distribution of $\tilde{Y}$ given $\tilde{X}$ is Bernoulli, with $\ell(\mu) = \log(\mu/(1-\mu))$ for $\mu \in (0, 1)$ in (1.2). Note that the particular case where the censoring time $C \equiv 0$ has been widely used in econometrics and is usually referred to as the binary choice model. The maximum likelihood estimator (MLE) of α_0 was proved to be consistent by Cosslett [7], and Murphy, Van der Vaart and Wellner [23] prove that the rate of convergence is $O(n^{1/3})$ in the one-dimensional case (that is, when $d = 1$). The latter also shows that an appropriately penalized MLE is $\sqrt{n}$-consistent, but the considered estimator is difficult to implement. Groeneboom and Hendrickx [14] consider several alternative $\sqrt{n}$-consistent estimators based on a truncated likelihood.

1.3. Contents of the paper

In the monotone single index model, we consider the least squares estimator (LSE) which estimates both the index and the ridge function without the use of a tuning parameter. We give a characterization of the LSE of (Ψ_0, α_0) under the monotonicity constraint using a profile approach. Furthermore, letting $g_0(x) = E(Y|X = x) = \Psi_0(\alpha_0^T x)$, we prove that, under appropriate conditions, the LSE of g_0 converges at an $n^{1/3}$-rate in the L_2-norm. Then, we consider the LSE of α_0 and Ψ_0 separately, and also prove their $n^{1/3}$-consistency. The $n^{1/3}$-rate of convergence obtained for the index may be due to our strategy of proof, as we derive this rate from the $n^{1/3}$-rate of the LSE of g_0. Thus, sharper rates could potentially be obtained using alternative methods. This is however out of the scope of this work.

The least squares estimator of the index α_0 is computationally intensive, so we also consider alternative estimators of the index taken from earlier literature. Among them, the so-called linear estimator, due to Brillinger [3], is especially appealing since it is very easy to implement and converges at the $\sqrt{n}$-rate to the true index under appropriate conditions, see Section 3.2. We then consider "plug-in" estimators of g_0: we first estimate the index using the first pn data points for some fixed $p \in (0,1)$, then plug the obtained estimator $\tilde{\alpha}_n$ in the least squares criterion and finally minimize the criterion based on the remaining $(1-p)n$ data points over the space of monotone ridge functions. See Section 3.1 for details. Combining these two estimators gives an estimator of g_0, and we show in Section 6 that if the rate of convergence of $\widetilde{\alpha}_n$ is sufficiently fast, then the corresponding estimators of g_0 and Ψ_0 converge at the $n^{1/3}$-rate. This means that the practitioner can choose his favorite estimator for the index and using the least-squares approach, obtain $n^{1/3}$-convergent estimators for the bundled and the ridge functions. This shows flexibility of the approach.

The paper is organized as follows. In Section 2, we show existence of the LSE of (Ψ_0, α_0) and give its characterization. Section 3 is devoted to the description of the plug-in approach based on alternative estimators for the index, as well as to the description of such alternative estimators. Our main result is given in Theorem 4.1 in Section 4, where we establish the $n^{1/3}$-convergence rate of the LSE of g_0. In Section 5, we show under some specified assumptions that the LSE of α_0 and Ψ_0 converge separately at the same rate in the Euclidean norm on $\mathbb{R}^d$ and the L_2-norm on the set of real valued functions respectively, provided that we restrict integration to a bounded subset of the domain of Ψ_0. Section 6 studies the rate of convergence of the above-mentioned plug-in estimators. The proof of Theorem 4.1 is given in Section 7. Other proofs are deferred to the supplemental article (Balabdaoui, Durot and Jankowski [1]).

2. Existence and characterization of the least squares estimator

Assume that we observe an i.i.d. sample $(X_1, Y_1), \dots, (X_n, Y_n)$ from (X, Y) such that $E(Y|X) = \Psi_0(\alpha_0^T X)$ almost surely, where both the index α_0 and the monotone ridge function Ψ_0 are unknown. To ensure model identifiability (see Section 1.2), α_0 is assumed to belong to the d-dimensional unit sphere $\mathcal{S}_{d-1}$ and the ridge function Ψ_0 is assumed to be non-decreasing on its domain, which contains the range of the linear predictor $\alpha_0^T X$. For technical reasons, in what follows we will extend all functions outside their actual support by taking the extension to be constant to the left and right of the endpoints of the original support.

The goal is to find the LSE of (Ψ_0, α_0), the minimizer of the least-squares criterion

$$h_n(\Psi, \alpha) = \sum_{i=1}^{n} \{Y_i - \Psi(\alpha^T X_i)\}^2$$

over $\mathcal{M} \times \mathcal{S}_{d-1}$, where $\mathcal{M}$ is the class of all non-decreasing functions on $\mathbb{R}$. Using a profile least-squares approach, we first minimize $\Psi \mapsto h_n(\Psi, \alpha)$ over $\mathcal{M}$ for a fixed α, and then minimize over α. All proofs for Section 2 are given in Section 8 of the supplemental article (Balabdaoui, Durot and Jankowski [1]).

Theorem 2.1. *For any $\alpha \in \mathbb{R}^d$, the minimum of $\Psi \mapsto h_n(\Psi, \alpha)$ over $\mathcal{M}$ is achieved. The minimizer is not unique; it is uniquely defined at the points $\alpha^T X_i$, $i = 1, \dots, n$.*

Next, we search for $\widehat{\alpha}_n$ that minimizes

$$\widehat{h}_n(\alpha) := \min_{\Psi \in \mathcal{M}} h_n(\Psi, \alpha) \tag{2.1}$$

over $\alpha \in \mathcal{S}_{d-1}$. The following proposition shows that the minimum is attained on $\mathcal{S}^X$, the set of all $\alpha \in \mathcal{S}_{d-1}$ which satisfy $\alpha^T X_i \neq \alpha^T X_j$ for all $i \neq j$ such that $X_i \neq X_j$. This will prove very helpful to provide a characterization of the LSE, see Theorem 2.3 below.

Proposition 2.2. *The infimum of $\widehat{h}_n$ over $\mathcal{S}_{d-1}$ is achieved on $\mathcal{S}^X$ and the minimizer is not unique: the set of minimizers contains an open subset of $\mathcal{S}_{d-1}$.*

Combining Theorem 2.1 and Proposition 2.2, we prove existence and non-uniqueness of the LSE. Some notation is needed before giving a precise characterization of the LSEs. The characterization uses the fact that (thanks to Proposition 2.2) one can restrict attention to those $\alpha \in \mathcal{S}^X$ in the minimization process. Let $x_1, \dots, x_m$ denote the distinct values of $X_1, \dots, X_n$, where $m \in \mathbb{N}$ is random. We define

$$\tilde{n}_k = \sum_{i=1}^n \mathbb{I}_{X_i = x_k} \quad \text{and} \quad \tilde{y}_k = \frac{1}{\tilde{n}_k} \sum_{i=1}^n Y_i \mathbb{I}_{X_i = x_k} \tag{2.2}$$

for all $k = 1, \dots, m$. Let $\mathcal{P}^X$ be the set of all permutations (i.e., orderings) π on $\{1, \dots, m\}$ such that there exists an $\alpha \in \mathcal{S}_{d-1}$ that linearly induces π in the sense that

$$\alpha^T x_{\pi(1)} < \cdots < \alpha^T x_{\pi(m)}. \tag{2.3}$$

Note that for each $\alpha \in \mathcal{S}^X$, the $\alpha^T x_k$'s are all different from each other and therefore, there exists a unique permutation π on $\{1, \dots, m\}$ that is linearly induced by α, that is, that satisfies (2.3). Then, for each $\pi \in \mathcal{P}^X$, we denote by $d_1^\pi \le \cdots \le d_m^\pi$ the left derivatives of the greatest convex minorant of the cumulative sum diagram defined by the set of points

$$\left\{(0,0), \left(\sum_{j=1}^k \tilde{n}_{\pi(j)}, \sum_{j=1}^k \tilde{n}_{\pi(j)} \tilde{y}_{\pi(j)}\right), k = 1, \dots, m\right\}.$$

Theorem 2.3. *The infimum of $(\Psi, \alpha) \mapsto h_n(\Psi, \alpha)$ over $\mathcal{M} \times \mathcal{S}_{d-1}$ is achieved. Moreover, if $(\widehat{\Psi}_n, \widehat{\alpha}_n)$ satisfies the following conditions, then it is a minimizer:*

- *$\widehat{\alpha}_n \in \mathcal{S}^X$ linearly induces $\widehat{\pi}_n$ that minimizes $\pi \mapsto \tilde{h}_n(\pi) := \sum_{k=1}^m \tilde{n}_{\pi(k)} (\tilde{y}_{\pi(k)} - d_k^\pi)^2$ over $\mathcal{P}^X$, and*
- *$\widehat{\Psi}_n$ is monotone non-decreasing with $\widehat{\Psi}_n(\widehat{\alpha}_n^T x_{\widehat{\pi}_n(k)}) = d_k^{\widehat{\pi}_n}$.*

To compute a LSE, one can implement the following steps: (1) compute $\tilde{n}_k$ and $\tilde{y}_k$ for all $k = 1, \dots, m$; (2) compute $d_1^\pi, \dots, d_m^\pi$ for all π in the finite set $\mathcal{P}^X$ using, for example, the pool adjacent violators algorithm (Barlow et al. [2], PAVA); (3) compute $\widehat{\pi}_n$ that minimizes $\tilde{h}_n(\pi)$ over the finite set $\mathcal{P}^X$; (4) compute $\widehat{\alpha}_n \in \mathcal{S}^X$ that linearly induces $\widehat{\pi}_n$; (5) compute $\widehat{\Psi}_n \in \mathcal{M}$ such that $\widehat{\Psi}_n(\widehat{\alpha}_n^T x_{\widehat{\pi}_n(k)}) = d_k^{\widehat{\pi}_n}$ for all k (one can consider for simplicity a piecewise constant function).

The difficulty with the above line of implementation is that it requires that the set of all linearly inducible permutations $\mathcal{P}^X$ be computable (steps (2) and (3)). Also, it requires that given a linearly inducible permutation, one can compute an index in $\mathcal{S}^X$ that induces the permutation (step (4)). The cardinality of $\mathcal{P}^X$ is known to be on the order of $m^{2(d-1)}$, see Cover [8], but we are not aware of an efficient algorithm to implement (2)–(4).

Therefore, instead of using inducible permutationss, one could use an alternative optimization algorithm; for example, stochastic search was used in Chen and Samworth [4], Table 4, page 740. When adapted to our setting, the algorithm simplifies as follows: (1) choose the total number N of stochastic searches to perform and set $k = 1$; (2) draw a standard Gaussian vector Z_k in $\mathbb{R}^d$ and compute $\alpha_k = Z_k/\|Z_k\|$; (3) compute the ordered distinct values $t_1 < \cdots < t_L$ of $\alpha_k^T X_i$, $i \in \{1, \dots, n\}$ and also

$$n_l = \sum_{i=1}^{n} \mathbb{I}_{\alpha_k^T X_i = t_l} \quad \text{and} \quad y_l = \frac{1}{n_l} \sum_{i=1}^{n} Y_i \mathbb{I}_{\alpha_k^T X_i = t_l}$$

for all $l = 1, \dots, L$; (4) compute $d_1 \le \cdots \le d_L$, the left derivatives of the greatest convex minorant of the cumulative sum diagram defined by the set of points

$$\left\{(0, 0), \left(\sum_{j=1}^{l} n_j, \sum_{j=1}^{l} n_j y_j\right), l = 1, \dots, L\right\}$$

using the PAVA; (5) compute $A_k := \sum_{l=1}^{L} n_l (y_l - d_l)^2$, set $k := k + 1$, go to (2) if $k \le N$ and to (6) otherwise; (6) compute $\widehat{k}$ that minimizes A_k over $k \in \{1, \dots, N\}$. An approximated value of the LSE $(\widehat{\alpha}_n, \widehat{\Psi}_n)$ is then given by $(\alpha_{\widehat{k}}, \Psi_{\widehat{k}})$, where using the same notation as in (3) and (4) where $k = \widehat{k}$, $\Psi_{\widehat{k}}$ is piecewise constant function such that $\Psi_{\widehat{k}}(t_l) = d_l$ for all $l = 1, \dots, L$. Note that in the algorithm, the variables $Z_1, \dots, Z_N$ are drawn independently from each other.

For completeness, in the supplemental article (Balabdaoui, Durot and Jankowski [1]), we also give an algorithm to compute the LSE exactly for the special case when $d = 2$, see Section 8.4.

3. Alternative estimators

Alternative estimators can be obtained by combining the above least squares approach with an alternative estimator of the index α_0, as detailed in Section 3.1 below. As can be seen from Section 3.1, the main difficulty in computing the LSE in the monotone single index model lies in computing an estimator of the unknown index α_0. Hence, we consider below various estimators of α_0 from earlier literature on single index models with a non-necessarily monotone ridge function. For notational convenience, all the considered estimators are denoted by $\tilde{\alpha}_n$. Among the considered estimators, the linear estimator of Section 3.2 is of particular interest since it is very easy to compute and converges at the $\sqrt{n}$-rate in the monotone single index model, see Theorem 3.1 below.

3.1. Plug-in estimators

First, randomly split the sample into two independent sub-samples of respective sizes n_1 and n_2, where n_1 is the integer part of pn for some fixed $p \in (0, 1)$ and $n_2 = (1 - p)n$. Let $\tilde{\alpha}_n$ denote some appropriate estimator of the true index α_0 using the n_1 data points in the first sub-sample. Next, we consider the "plug-in" estimator $\widetilde{\Psi}_n := \widehat{\Psi}^{\tilde{\alpha}_n}$ of Ψ_0, where for all α, $\widehat{\Psi}^{\alpha}$ is the minimizer of

$$\Psi \mapsto \sum_{i \in I_2} \{Y_i - \Psi(\alpha^T X_i)\}^2 \tag{3.1}$$

over $\Psi \in \mathcal{M}$, where $\{(X_i, Y_i), i \in I_2\}$ are the observations from the second sub-sample. Once $\tilde{\alpha}_n$ is given, the estimator $\widetilde{\Psi}_n$ is easy to compute using again the PAVA. Indeed, it follows from Barlow et al. [2], Theorem 1, that any $\widetilde{\Psi}_n \in \mathcal{M}$ such that $\widetilde{\Psi}_n(Z_k) = d_k$ is a minimizer. Here, $Z_1 < \cdots < Z_m$ denote the ordered distinct values of $\tilde{\alpha}_n^T X_i$, $i \in I_2$, and $d_1 \le \cdots \le d_m$ are the left derivatives of the greatest convex minorant of the cumulative sum diagram defined by the set of points

$$\left\{(0,0), \left(\sum_{i \in I_2} \mathbb{I}_{\tilde{\alpha}_n^T X_i \le Z_k}, \sum_{i \in I_2} Y_i \mathbb{I}_{\tilde{\alpha}_n^T X_i \le Z_k}\right), k = 1, \ldots, m\right\}.$$

Below, we consider several estimators $\tilde{\alpha}_n$ that could be used in this plug-in procedure.

3.2. The linear estimator

The linear estimator goes back to Brillinger [3], who also considered a single index model (1.1) with an unknown, not necessarily monotone ridge function Ψ_0. This estimator is exactly what one would use if the regression model were known to be linear. To be precise, based on observations $(X_1, Y_1), \ldots, (X_n, Y_n)$ where the Y_is take real values whereas the X_is take values in $\mathbb{R}^d$, the linear estimator of α_0 is defined as follows: $\tilde{\alpha}_n = \widehat{\alpha}_n / \|\widehat{\alpha}_n\|$ where here,

$$\widehat{\alpha}_n = \operatorname*{argmin}_{\alpha \in \mathbb{R}^d} \sum_{i=1}^n \left(Y_i - \alpha^T (X_i - \bar{X}_n)\right)^2 \tag{3.2}$$

with $\bar{X}_n = n^{-1} \sum_{i=1}^n X_i$. The linear estimator can therefore be easily computed using standard tools from linear regression. Moreover, it typically converges to the true index α_0 at the square-root rate and is asymptotically Gaussian, even if the linearity assumption is not valid. Typical assumptions required for these results to hold are that the variables $\Psi_0(\alpha_0^T X)$ and $\alpha_0^T X$ are correlated, and that the conditional expectation of X given $\alpha_0^T X$ is a linear function of $\alpha_0^T X$. The latter condition is met under elliptic symmetry of X (which holds in particular if X is Gaussian, see Chmielewski [6]), a condition that has been considered for instance by Li and Duan [20] and Goldstein, Minsker and Wei [13]. It turns out that the condition $\mathrm{Cov}(\Psi_0(\alpha_0^T X), \alpha_0^T X) \neq 0$ is met in our setting where Ψ_0 is monotone and not constant, whence the linear estimator is $\sqrt{n}$-consistent and asymptotically Gaussian. The precise statement is given in the following theorem, which is a close variant to earlier results in the literature on linear estimators. Here, the distribution of X is assumed to be continuous since α_0 is not identifiable under a discrete distribution of X. The assumption on boundedness of Ψ_0 ensures existence of the above covariance. For completeness, the proof is provided in Section 9.1 of the supplemental article (Balabdaoui, Durot and Jankowski [1]).

Theorem 3.1. *Let $(X_1, Y_1), \dots, (X_n, Y_n)$ be an i.i.d. sample from (X, Y) such that $E(Y|X) = \Psi_0(\alpha_0^T X)$ almost surely where $\alpha_0 \in \mathcal{S}_{d-1}$ and Ψ_0 is bounded and non-decreasing such that there exists a nonempty interval $[a, b]$ in the domain of $\alpha_0^T X$ on which Ψ_0 is strictly increasing. Suppose furthermore that X has a continuous elliptically symmetric distribution with finite mean $\mu \in \mathbb{R}^d$ with a positive definite $d \times d$ covariance matrix Σ, and $E(\|X\|^2 Y^2) < \infty$. Then $\sqrt{n}(\tilde{\alpha}_n - \alpha_0)$ converges weakly to a centered d-dimensional Gaussian distribution.*

Note that by definition of $\widehat{\alpha}_n$, we necessarily have that $\widehat{\alpha}_n = \widehat{\Sigma}_n^{-1} n^{-1} \sum_{i=1}^n Y_i(X_i - \bar{X}_n)$, where $\widehat{\Sigma}_n = n^{-1} \sum_{i=1}^n (X_i - \bar{X}_n)(X_i - \bar{X}_n)^T$. If the X_i's were known to be centered with identity covariance matrix, one would merely consider the estimator $n^{-1} \sum_{i=1}^n Y_i X_i$, which is precisely the estimator considered in Section 1.2 of Plan, Vershynin and Yudovina [25]. Under the assumptions that (in our notation) X is standard Gaussian and Y is independent of X conditionally on $\alpha_0^T X$, and $E(\|X\|^2 Y^2) < \infty$ (Plan, Vershynin and Yudovina [25], Proposition 1.1) shows that their estimator is equal to $\lambda \alpha_0 + O_p(\sqrt{d/n})$ in the case $\|\alpha_0\| = 1$, with $\lambda = E(Y \alpha_0^T X)$. As explained in Section 7 of that paper, this result can be generalized to non-standard Gaussian covariates. In fact, if X has a Gaussian distribution with finite mean $\mu \in \mathbb{R}^d$ and covariance matrix Σ such that $\|\mu\| \le K$ and all eigenvalues of Σ belong to $[K_-, K_+]$ for some positive K, K_-, K_+ that do not depend on d, then this result can be extended to prove that our $\widehat{\alpha}_n$ is equal to $\lambda^* \alpha_0 + O_p(\sqrt{d/n})$ uniformly in d, n, with λ^* defined as in Section 9.1 of the supplemental article (Balabdaoui, Durot and Jankowski [1]). This implies that the convergence rate of the linear estimator $\tilde{\alpha}_n$ depends on the dimension like $\sqrt{d/n}$. As mentioned by a referee, this rate is optimal as it is the rate one would obtain if the ridge function was equal to the identity.

3.3. Additional estimators

In the following we discuss other possible ways of index estimation in the present model. A well-known estimator in the monotone single index model is the so-called maximum rank correlation (MRC) estimator; see Han [15]. This estimator is defined as the location of the maximum of $S_n(\alpha)$ over $\alpha \in \mathcal{S}_{d-1}$ where

$$S_n(\alpha) = \binom{n}{2}^{-1} \sum_{1 \le i \ne j \le n} [\mathbb{I}_{\{Y_i > Y_j\}} \mathbb{I}_{\{\alpha^T X_i > \alpha^T X_j\}} + \mathbb{I}_{\{Y_i < Y_j\}} \mathbb{I}_{\{\alpha^T X_i < \alpha^T X_j\}}].$$

Strong consistency of a variant of the MRC estimator is proved in Han [15] under the assumption that (a) the noise $Y - E(Y|X)$ is independent of X, (b) for a given $h \in \{1, \dots, d\}$ the component h of α_0 is in absolute value greater than a given $\eta > 0$, and (c) the distribution of X behaves "nicely". Hence, the variant of the MRC estimator is defined as the location of the maximum of $S_n(\alpha)$ over the set of all α's in $\mathcal{S}_{d-1}$ whose component h is in absolute value greater than η. This implies in particular that one would need to know both h and η, which is quite unrealistic in our opinion.

Building upon the Isotron algorithm of Kalai and Sastry [19], Kakade et al. [18] propose an iterative algorithm in the monotone single index model, called the Slisotron, which finds estimators of the index and monotone ridge function under the additional assumption that the latter is

Lipschitz. More precisely, in the updates of the isotonic estimator, the Slisotron algorithm looks for the least squares monotone estimator which is also Lipschitz. Slisotron produces estimators for both the index and ridge function, and in view of the current discussion we are interested here in the former. Theorem 2 of Kakade et al. [18] shows that the both true and empirical mean squared errors are of order $n^{-1/3}\log n$ with large probability for some appropriate iteration of Slisotron. This indicates that their estimator of the index converges, ignoring the logarithmic factor, at the $n^{1/6}$ rate, which is significantly worse than the cubic rate achieved by our least squares estimator of the index, see Corollary 5.3 below.

Assuming again that the ridge function is Lipschitz, and assuming also that the response variable takes values in [0, 1], Ganti et al. [12] provide an estimation method that applies even for the case of high-dimensional covariates. Their "SILO" method can be viewed as an extension of the linear estimator (described above in Section 3.2) to the high-dimensional case: similar to the linear estimator, the SILO estimator of the index does not take into account the ridge function. Other estimation methods can be found in the compressed sensing literature (some of them are designed for the case of binary response variables), see e.g. Plan and Vershynin [24], Plan, Vershynin and Yudovina [25].

There are several other alternatives which return an estimate of the index in the single index model with a non-necessarily monotone ridge function, and these could also be used here. For example, one could use kernel-based methods, discussed for example, in Härdle, Hall and Ichimura [16], Chiou and Müller [5], Hristache, Juditsky and Spokoiny [17]. Although these methods yield an estimator which is $\sqrt{n}$-consistent, they do rely on smoothing parameters (the bandwidth) for their estimator.

4. Convergence of the LSE for the regression function

We consider the same setting as in Section 2, however, we now also assume that X has a continuous distribution. This means that we observe an i.i.d. sample (X_i, Y_i), $i = 1, \dots, n$ from a pair (X, Y) where, with probability one, all the X_i's are different from each other (hence, in the notation of Section 2, $n = m$ and $\tilde{n}_i = 1$ for all $i = 1, \dots, n$). It is assumed that $E(Y|X) = \Psi_0(\alpha_0^T X)$ almost surely, where both the index α_0 and the monotone ridge function Ψ_0 are unknown, so the regression function is defined by

$$g_0(x) = \Psi_0\big(\alpha_0^T x\big) \tag{4.1}$$

for (almost-) all x in the support of X and its least-squares estimator (LSE) is given by

$$\widehat{g}_n(x) = \widehat{\Psi}_n\big(\widehat{\alpha}_n^T x\big) \tag{4.2}$$

for (almost-) all x in the support of X, where $(\widehat{\Psi}_n, \widehat{\alpha}_n)$ is a LSE of (Ψ_0, α_0) as studied in Section 2. For convenience, we consider below a solution $\widehat{\Psi}_n$ that is left continuous and piecewise constant, with jumps only possible at the points $\widehat{\alpha}_n^T X_i$, for $i = 1, \dots, n$. Hence, $\widehat{\Psi}_n$ is uniquely defined whereas $\widehat{\alpha}_n$ is not unique ($\widehat{\alpha}_n$ denotes an arbitrary minimizer of $\widehat{h}_n$, in the notation of Section 2). In this section, we are interested in the consistency and rate of convergence of $\widehat{g}_n$ in the L_2-sense.

We begin with some notation and assumptions. Let $\mathcal{X}$ be the support of the random vector of covariates X. Let $\mathbb{P}$ be the joint distribution of (X, Y), $\mathbb{P}_x$ the conditional distribution of Y given $X = x$, and $\mathbb{Q}$ the marginal distribution of X. Theorem 4.1 below will be established under the following assumptions.

(A1) $\mathcal{X}$ is a bounded convex set of $\mathbb{R}^d$,
(A2) there exists a constant $K_0 > 0$ such that $|g_0(x)| \le K_0$ for all $x \in \mathcal{X}$,
(A3) there exist constants $a_0 > 0$ and $M > 0$ such that for all integers $s \ge 2$ and $x \in \mathcal{X}$

$$\int |y|^s \, d\mathbb{P}_x(y) \le a_0 s! M^{s-2}, \tag{4.3}$$

(A4) there exist $\overline{q} > 0$ and $\underline{q} > 0$ such that with respect to the Lebesgue measure, for all $\alpha \in \mathcal{S}_{d-1}$, the variable $\alpha^T X$ has a density that is bounded from above by $\overline{q}$ and bounded from below by $\underline{q}$ on its support.

Assumption (A1) ensures that for all $\alpha \in \mathcal{S}_{d-1}$, the set $\{\alpha^T x, x \in \mathcal{X}\}$, which is the support of the linear predictor $\alpha^T X$ corresponding to α, is convex (*i.e.* is an interval). Hence, we consider functions of the form $\Psi(\alpha^T X)$ where $\alpha \in \mathcal{S}_{d-1}$ and Ψ is a non-decreasing function on its interval of support $\{\alpha^T x, x \in \mathcal{X}\}$. It turns out that for Theorem 4.1 below, it is sufficient to assume, instead of Assumption (A1), that the support of X is connected, since the continuous image of a connected set is connected, and a connected subset of the real line is necessarily an interval. However, convexity of the support of X is also used in Proposition 5.1 below so to alleviate the exposition, we consider the same assumption (A1) in Theorem 4.1 than in Proposition 5.1.

Assumption (A3) is clearly satisfied if Y is a bounded random variable. It is also satisfied if the conditional distribution of Y given X belongs to an exponential family, see Proposition 9.2 in Appendix 9.8 for more details. The assumption ensures that conditionally on $X = x$, the response variable Y is uniformly integrable in x. It also ensures that $\max_i |Y_i|$ is of maximal order $\log n$, see (7.11) below, which in turn ensures that $\widehat{g}_n$ also is of maximal order (in sup-norm) $\log n$, see (7.2).

Assumption (A4) makes the distribution of $\alpha^T X$, with $\alpha \in \mathcal{S}_{d-1}$, equivalent to the Lebesgue measure on its support.

The following theorem proves the $n^{1/3}$-rate of convergence of the bundled estimator $\widehat{g}_n$ under the above assumptions, that is, under the assumption of a continuous design distribution $\mathbb{Q}$. The case of a discrete distribution with finite support will be considered in a separate paper. In this case, a $\sqrt{n}$-rate of convergence can be proved. We conjecture that in the case when some of the components of X are continuous, and the other ones are discrete, the rate of convergence is still $n^{1/3}$. Another case where the $\sqrt{n}$-rate of convergence emerges (up to a $\log n$-factor) is when the true ridge function is constant. This case also will be studied elsewhere.

Theorem 4.1. *With g_0 and $\widehat{g}_n$ defined by* (4.1) *and* (4.2) *respectively, where $\widehat{\Psi}_n$ is the same piecewise constant function described above, and under assumptions* (A1)–(A4) *we have*

$$\left(\int_{\mathcal{X}} \left(\widehat{g}_n(x) - g_0(x)\right)^2 d\mathbb{Q}(x)\right)^{1/2} = O_p\left(n^{-1/3}\right). \tag{4.4}$$

Remark 4.2.

- If instead of assuming (A4) we only assume that there exists a $\overline{q} > 0$ such that with respect to the Lebesgue measure and for all $\alpha \in \mathcal{S}_{d-1}$, the variable $\alpha^T X$ has a density bounded above by $\overline{q}$, then we obtain a rate of convergence $n^{-1/3}(\log n)^{5/3}$ instead of $n^{-1/3}$, see Section 7.1 below for details.
- The convergence rate obtained above depends on the dimension d. A closer look at the proof reveals that this dependence takes the form of $O_p(d(1+\sqrt{\overline{q}}R)n^{-1/3}(\log n)^{5/3})$, see Theorem 7.3 below. Note that the constant R may hide a dependence on d since in the case where $\mathcal{X}$ is the ℓ_∞-unit ball in $\mathbb{R}^d$ we have $R = \sqrt{d}$.
- Suppose that we relax assumptions (A1) and (A4). That is, instead of assuming (A4), we only assume here that there exists $\overline{q} > 0$ such that with respect to the Lebesgue measure, for all $\alpha \in \mathcal{S}_{d-1}$, the variable $\alpha^T X$ has a density that is bounded from above by $\overline{q}$. Moreover, instead of assuming that X has a bounded support, we assume that X has a sub-Gaussian distribution. This means that there exists $\sigma^2 > 0$ such that for all vectors $u \in \mathcal{S}_{d-1}$, and all $t \in \mathbb{R}$, with $T = u^T X$ we have

$$P\big(T - E(T) > t\big) \le \exp\big(-t^2/\big(2\sigma^2\big)\big) \quad \text{and} \quad P\big(T - E(T) < -t\big) \le \exp\big(-t^2/\big(2\sigma^2\big)\big).$$

  Then, for all $\varepsilon > 0$ there exists $A > 0$ such that with probability larger than $1-\varepsilon$ we have

$$\int_{\mathcal{X}} \big(\widehat{g}_n(x) - g_0(x)\big)^2 d\mathbb{Q}(x) \le A\sqrt{\overline{q}} d^{5/4}\big(\log(n \vee d)\big)^{1/4} n^{-1/3}(\log n)^{5/3}, \tag{4.5}$$

  see Section 9.2 in the supplemental article (Balabdaoui, Durot and Jankowski [1]) for details. If d does not depend on n, then this yields a rate of convergence $n^{-1/3}(\log n)^{23/12}$.

5. Convergence of the separated LSE estimators

We now derive from Theorem 4.1 convergence of $\widehat{\alpha}_n$ to α_0 and $\widehat{\Psi}_n$ to Ψ_0. Moreover, we are interested in the rate of convergence of the two estimators. Convergence can happen only under uniqueness of the limit so first we prove identifiability of Ψ_0 and α_0 under appropriate conditions.

5.1. Identifiability of the separated parameters

Let (X, Y) be a pair of random variables, where X takes values in $\mathbb{R}^d$ and Y is an integrable real valued random variable such that (1.1) holds for some $\alpha_0 \in \mathcal{S}_{d-1}$ and $\Psi_0 \in \mathcal{M}$. Identifiability of the parameter (α_0, Ψ_0) means here that if we can find β in $\mathcal{S}_{d-1}$, and h in $\mathcal{M}$ such that $\Psi_0(\alpha_0^T X) = h(\beta^T X)$ a.s. then $\beta = \alpha_0$ and $h = \Psi_0$ on $\mathcal{C}_{\alpha_0} = \mathcal{C}_\beta$, where for all $\alpha \in \mathcal{S}_{d-1}$ we set $\mathcal{C}_\alpha = \{\alpha^T x, x \in \mathcal{X}\}$ with $\mathcal{X}$ being the support of X. Although identifiability can be derived from Lin and Kulasekera [22] when assuming that Ψ_0 is non-constant and continuous, for completeness we state below identifiability under a slightly less restrictive assumption, namely left- (or right-) continuity instead of continuity. A proof can be found in Section 9.3 in the supplemental article (Balabdaoui, Durot and Jankowski [1]). Since $\mathcal{X}$ is convex, it follows that $\mathcal{C}_\alpha$ is an interval for any α. Moreover, recall that monotone functions on an interval can be extended to monotone functions on $\mathbb{R}$.

Proposition 5.1. *Assume that $\mathcal{X}$ is convex with at least one interior point. Assume also that X has a density with respect to Lebesgue measure which is strictly positive on $\mathcal{X}$, and that* (1.1) *holds for some $\alpha_0 \in \mathcal{S}_{d-1}$ and $\Psi_0 \in \mathcal{M}$ that is not constant on $\mathcal{C}_{\alpha_0}$, and either left- or right-continuous on $\mathcal{C}_{\alpha_0}$ with no discontinuity point at the boundaries of $\mathcal{C}_{\alpha_0}$. Then, (Ψ_0, α_0) is uniquely defined.*

5.2. Convergence of the separated estimators

We begin by establishing consistency of $(\widehat{\alpha}_n, \widehat{\Psi}_n)$ where $\widehat{\Psi}_n$ denotes the left-continuous LSE of Ψ_0 extended to $\mathbb{R}$ and $\widehat{\alpha}_n$ is a minimizer of $\widehat{h}_n$ defined in (2.1), see Section 9.4 in the supplemental article (Balabdaoui, Durot and Jankowski [1]) for a proof.

Theorem 5.2. *Assume that assumptions* (A1)–(A3) *are satisfied and that there exists a $\overline{q} > 0$ such that for all $\alpha \in \mathcal{S}_{d-1}$, with respect to the Lebesgue measure, the variable $\alpha^T X$ has a density that is bounded above by $\overline{q}$. Assume, moreover, that Ψ_0 is non-constant and left-continuous with no discontinuity points at the boundaries of $\mathcal{C}_{\alpha_0}$, and that $\mathcal{X}$ has at least one interior point. Assume also that X has a density with respect to Lebesgue measure which is strictly positive on $\mathcal{X}$.*

1. *We then have $\widehat{\alpha}_n = \alpha_0 + o_p(1)$, and for all fixed continuity points t of Ψ_0 in the interior of $\mathcal{C}_{\alpha_0}$, $\Psi_n(t)$ converges in probability to $\Psi_0(t)$ as $n \to \infty$.*
2. *If, moreover, Ψ_0 is continuous, then*

$$\sup_{t\in I}\bigl|\widehat{\Psi}_n(t) - \Psi_0(t)\bigr| = o_p(1) \tag{5.1}$$

for all compact intervals $I \subset \mathbb{R}$ such that $K_- < \Psi_0(t) < K_+$ for all $t \in I$. Here, K_+ and K_- denote the largest and smallest values of Ψ_0 on $\mathcal{C}_{\alpha_0}$.

Next, we establish rates of convergence for $\widehat{\alpha}_n$ and $\widehat{\Psi}_n$. To show that both $\widehat{\alpha}_n$ and $\widehat{\Psi}_n$ inherit the $n^{1/3}$ rate of convergence from the joint convergence established for the full estimator $\widehat{g}_n(\cdot) = \widehat{\Psi}_n(\widehat{\alpha}_n^T \cdot)$, some additional assumptions are needed.

(A5) There exists an interior point $z_0 \in \mathcal{C}_{\alpha_0}$ such that Ψ_0 is continuously differentiable in the neighborhood of z_0, with $\Psi_0'(z_0) > 0$.
(A6) The density of X, q, is continuous on $\mathcal{X}$.

Let $\underline{c} = \inf \mathcal{C}_{\alpha_0}$ and $\overline{c} = \sup \mathcal{C}_{\alpha_0}$. Our main result here is the following. It is proved in Section 9.4 in the supplemental article (Balabdaoui, Durot and Jankowski [1]).

Corollary 5.3. *Assume that Ψ_0 is non-constant and left-continuous with no discontinuity points at the boundaries of $\mathcal{C}_{\alpha_0}$, that $\mathcal{X}$ has at least one interior point, and that* (A1)–(A6) *hold. Then,*

$\|\widehat{\alpha}_n - \alpha_0\| = O_p(n^{-1/3})$. *If moreover,* Ψ_0 *has a derivative bounded from above on* $\mathcal{C}_{\alpha_0}$, *then*

$$\left(\int_{\underline{c}+v_n}^{\bar{c}-v_n} \big(\widehat{\Psi}_n(t) - \Psi_0(t)\big)^2\, dt\right)^{1/2} = O_p\big(n^{-1/3}\big) \tag{5.2}$$

for all sequences v_n *such that* $n^{1/3} v_n \to \infty$ *and* $\underline{c} + v_n \le \bar{c} - v_n$.

Remark 5.4. The above result holds under Assumption (A1) on the support of the covariate X. The result can be made stronger under additional regularity conditions on the support $\mathcal{X}$. For example, when $\mathcal{X}$ is a ball in $\mathbb{R}^d$ centered at the origin and of radius r then the above result holds with $v_n = 0$. Indeed, in this setting the support of the linear predictor $\alpha^T X$, for any α, is $[-r, r]$. Therefore, in the proof, $\mathcal{C}_{\widehat{\alpha}_n} = [-r, r]$ and hence $v_n = 0$, $\bar{c} = r$ and $\underline{c} = -r$ in inequality (9.16) of the supplemental article (Balabdaoui, Durot and Jankowski [1]). Notably, Kakade et al. [18] consider this choice of $\mathcal{X}$ with $r = 1$.

The $n^{1/3}$-rate obtained in Corollary 5.3 for convergence of the LSE $\widehat{\alpha}_n$ towards the truth raises the question whether this convergence actually occurs at a faster rate, for example $n^{1/2}$. In order to investigate this question, we have performed simulations for $d = 2$ and two different monotone single index models: the first one is a Gaussian model where $Y \sim \mathcal{N}((\alpha_0^T X)^3, 1)$, whereas the second one is a logistic regression model where $Y \sim \text{Bin}(10, \exp(\alpha_0^T X)(1 + \exp(\alpha_0^T X))^{-1})$. In both settings, the two-dimensional covariate $X \sim \mathcal{U}[0,1] \times \mathcal{U}[0,1]$ and $\alpha_0 = (\cos(\theta_0), \sin(\theta_0))^T$ with $\theta_0 \in \{\pi/4, \pi/3, \pi/2\}$. From each of these monotone single index models we have drawn 100 times n i.i.d. pairs (X_i, Y_i) and computed the LSE $\widehat{\alpha}_n$ for $n \in \{10^2, 10^3, 10^4, 10^5\}$. Based on these 100 replications we computed the empirical estimates for the covariance matrix of $n^{1/3}(\widehat{\alpha}_n - \alpha_0)$ and $n^{1/2}(\widehat{\alpha}_n - \alpha_0)$. The main idea behind is that the correct rate of convergence should yield estimates that are more or less stable for large n. Our simulation results for the Gaussian and logistic model are reported in Table 1 and Table 2, respectively. For the settings we have chosen, the variances $\widehat{\sigma}_{11}^2$ and $\widehat{\sigma}_{22}^2$ as well as the absolute value of the covariance $|\widehat{c}_{12}|$ seem to increase with the sample size n if $n^{1/2}$ is the stipulated rate of convergence. This picture is completely reversed for the rate $n^{1/3}$. This first investigation can only allow us to conclude (for the chosen models, true monotone link functions and indices) that the convergence of our LSE occurs at a rate that is faster than $n^{1/3}$ and slower than $n^{1/2}$.

Upon request of one referee, we have performed additional simulations with equally spaced values of n on the logarithmic scale and computed for the same settings as above the average value of the square of the L_2-norm of the estimation error; that ish, $\|\widehat{\alpha}_n - \alpha_0\|_2^2$. For a given sample size n we denote by m_n this average. For this new set of simulations, we have increased the number of replications from 100 to 500. The plots of the $\log m_n$ versus $\log n$, shown in the supplemental article (Balabdaoui, Durot and Jankowski [1]), are unfortunately less conclusive than our results in Table 1 and Table 2. The non-linear trend of the plots indicates that the rate of convergence is not of the form n^{ν} for some $\nu > 0$. If one conjectures that this convergence rate is rather of the form $n^{\nu}(\log n)^{-\gamma}$ for some $\gamma > 0$, then regressing $\log m_n$ on the "predictors" $\log n$ and $\log(\log n)$ does not give meaningful outputs neither: in some cases it was found that

Table 1. Values of the empirical covariances matrices for the case $d = 2$ of $n^{1/2}(\widehat{\alpha}_n - \alpha_0)$ and $n^{1/3}(\widehat{\alpha}_n - \alpha_0)$ with entries $\widehat{\sigma}^2_{11}, \widehat{\sigma}^2_{22}, \widehat{c}_{12} = \widehat{c}_{21}$. The sample size $n \in \{10^2, 10^3, 10^4, 10^5\}$ and $\alpha_0 = (\cos(\theta_0), \sin(\theta_0))^T$ with $\theta_0 \in \{\pi/4, \pi/3, \pi/2\}$. The obtained estimates were computed based on 100 replications and the model $Y \sim \mathcal{N}((\alpha_0^T X)^3, 1)$ and $X \sim \mathcal{U}[0,1] \times \mathcal{U}[0,1]$

θ_0	Rate	n	$\widehat{\sigma}^2_{11}$	$\widehat{\sigma}^2_{22}$	$\widehat{c}_{12}$
$\pi/4$	$n^{1/2}$	100	1.180	1.167	−1.162
		1000	2.260	2.247	−2.247
		10 000	4.415	4.340	−4.374
		100 000	6.663	6.604	−6.633
	$n^{1/3}$	100	0.254	0.251	−0.250
		1000	0.226	0.225	−0.225
		10 000	0.205	0.201	−0.203
		100 000	0.143	0.142	−0.143
$\pi/3$	$n^{1/2}$	100	3.143	1.110	−1.836
		1000	5.404	1.700	−3.011
		10 000	7.078	2.418	−4.133
		100 000	7.565	2.513	−4.360
	$n^{1/3}$	100	0.677	0.239	−0.395
		1000	0.540	0.170	−0.301
		10000	0.328	0.112	−0.192
		100000	0.163	0.054	−0.094
$\pi/2$	$n^{1/2}$	100	6.740	0.132	0.066
		1000	11.633	0.051	−0.047
		10 000	12.110	0.014	−0.195
		100 000	10.904	<1e−03	−0.011
	$n^{1/3}$	100	1.452	0.028	0.014
		1000	1.163	0.005	−0.005
		10 000	0.562	<1e−03	−0.009
		100 000	0.235	<1e−03	<1e−03

the estimator of ν is smaller than -1, which is of course unrealistic. We believe that simulations for much larger sample sizes are needed to obtain a much better picture.

Proving the exact rate of convergence of $\widehat{\alpha}_n$ is an interesting question but goes beyond the scope of this work. We believe that in establishing this exact rate, under suitable assumptions, it is necessary to overcome the difficulty of non-smoothness of $\widehat{\Psi}_n$, the monotone estimator of the true link function Ψ_0 and the fact that $\widehat{\alpha}_n$ and $\widehat{\Psi}_n$ are intertwined. Consequently, a useful device such as Taylor expansion cannot be used. Also, when this $\widehat{\Psi}_n$ converges at the cubic rate (as it is the case under our assumptions), it is not immediate how to show that $\widehat{\alpha}_n$ converges at a faster rate as both Ψ_n and $\widehat{\alpha}_n$ depend on each other. We intend to investigate these questions in a future work.

Table 2. Values of the empirical covariances matrices for the case $d = 2$ of $n^{1/2}(\widehat{\alpha}_n - \alpha_0)$ and $n^{1/3}(\widehat{\alpha}_n - \alpha_0)$ with entries $\widehat{\sigma}^2_{11}, \widehat{\sigma}^2_{22}, \widehat{c}_{12} = \widehat{c}_{21}$. The sample size $n \in \{10^2, 10^3, 10^4, 10^5\}$ and $\alpha_0 = (\cos(\theta_0), \sin(\theta_0))^T$ with $\theta_0 \in \{\pi/4, \pi/3, \pi/2\}$. The obtained estimates were computed based on 100 replications and the model $Y \sim \text{Bin}(10, \exp(\alpha_0^T X)/(1 + \exp(\alpha_0^T X))$ and $X \sim \mathcal{U}[0, 1] \times \mathcal{U}[0, 1]$

θ_0	Rate	n	$\widehat{\sigma}^2_{11}$	$\widehat{\sigma}^2_{22}$	$\widehat{c}_{12}$
$\pi/4$	$n^{1/2}$	100	1.291	1.330	−1.300
		1000	2.981	2.977	−2.970
		10 000	5.718	5.600	−5.654
		100 000	7.140	7.180	−7.159
	$n^{1/3}$	100	0.278	0.287	0.280
		1000	0.298	0.298	−0.297
		10 000	0.265	0.260	−0.262
		100 000	0.154	0.155	−0.154
$\pi/3$	$n^{1/2}$	100	3.399	1.083	−1.884
		1000	5.250	1.788	−3.047
		10 000	9.769	3.307	−5.675
		100 000	13.059	4.420	−7.596
	$n^{1/3}$	100	0.732	0.233	−0.406
		1000	0.525	0.179	−0.305
		10000	0.453	0.153	−0.263
		100000	0.281	0.095	−0.164
$\pi/2$	$n^{1/2}$	100	6.689	0.127	−0.045
		1000	7.583	0.030	−0.125
		10 000	9.220	0.004	−0.015
		100 000	13.354	<1e−03	0.008
	$n^{1/3}$	100	1.441	0.027	−0.010
		1000	0.758	0.003	−0.012
		10 000	0.428	<1e−03	<1e−03
		100 000	0.288	<1e−03	<1e−03

6. Convergence of alternative estimators

We now consider convergence of plug-in estimators of Section 3.1: we randomly split the sample into two independent sub-samples of respective sizes n_1 and n_2, where n_1 is the integer part of pn for some fixed $p \in (0, 1)$ and $n_2 = (1 - p)n$, we compute an index estimator $\tilde{\alpha}_n$ based on the first sub-sample, and then compute $\tilde{\Psi}_n$, the minimizer of (3.1) over $\Psi \in \mathcal{M}$, where $\alpha = \tilde{\alpha}_n$ and $\{(X_i, Y_i), i \in I_2\}$ are the observations from the second sub-sample. Note that arguing conditionally on the first sub-sample, $\tilde{\alpha}_n$ can be considered as non-random when studying the limiting behavior of $\tilde{\Psi}_n$. In the sequel, we set $\tilde{g}_n(x) = \tilde{\Psi}_n(\tilde{\alpha}_n^T x)$ for all $x \in \mathbb{R}^d$. We prove below that, provided that $\tilde{\alpha}_n$ converges at the $n^{1/3}$-rate (which is the case of the linear estimator of Section 3.2 and some estimators from Section 3.3 under appropriate assumptions), $\tilde{g}_n$ also converges at the same rate. The complete proof is given in Section 9.7 of the supplemental article (Balabdaoui, Durot and Jankowski [1]). Below, we implicitly assume that α_0 is identifiable.

Theorem 6.1. *Assume* (A1)–(A4). *Assume, moreover, that* Ψ_0 *is non-constant and Lipschitz continuous, and that* $\tilde{\alpha}_n = \alpha_0 + O_p(n^{-1/3})$.

With $\tilde{g}_n$ *as above we then have*

$$\left(\int_{\mathcal{X}} \big(\tilde{g}_n(x) - g_0(x)\big)^2 \, dx\right)^{1/2} = O_p\big(n^{-1/3}\big). \tag{6.1}$$

Furthermore, if (A1)–(A6) *hold, and* Ψ_0 *has a first derivative that is bounded from above on* $\mathcal{C}_{\alpha_0}$, *then*

$$\left(\int_{\underline{c}+v_n}^{\overline{c}-v_n} \big(\tilde{\Psi}_n(t) - \Psi_0(t)\big)^2 \, dt\right)^{1/2} = O_p\big(n^{-1/3}\big) \tag{6.2}$$

for all sequences v_n *such that* $n^{1/3}v_n \to \infty$ *and* $\underline{c} + v_n \le \overline{c} - v_n$.

Remark 6.2. Similarly to Remark 5.4, the result can be made stronger under additional regularity conditions on the support $\mathcal{X}$. Moreover, similar to Remark 4.2, if instead of assuming that X has a bounded support we assume that it has a sub-Gaussian distribution, then the rate of convergence is only inflated by the factor $(\log n)^{5/3}$.

7. Proof of Theorem 4.1

As the proof of Theorem 4.1 is quite long and technical, we first give the main ideas of the proof of this theorem in Section 7.1 below. Here, we give two preparatory lemmas and an intermediate rate theorem (Theorem 7.3). The latter compares to Theorem 4.1 but with an additional $\log n$ term in the rate of convergence. The proof of Theorem 7.3 requires entropy results that are described in Section 7.2 and proved in subsequent subsections. The proof of Theorem 4.1 is finally completed in Section 7.9.

7.1. The main steps of the proof of Theorem 4.1

By definition of the LSE, $\widehat{g}_n$ maximizes the criterion

$$\mathbb{M}_n g := \frac{1}{n}\sum_{i=1}^{n}\left\{Y_i g(X_i) - \frac{g(X_i)^2}{2}\right\} \tag{7.1}$$

over the set of all functions g of the form $g(x) = \Psi(\alpha^T x)$, $x \in \mathcal{X}$ with $\alpha \in \mathcal{S}_{d-1}$ and $\Psi \in \mathcal{M}$. It would have been easier to prove Theorem 4.1 using standard results from empirical process theory if the LSE were known to be bounded in probability by some constant which is independent of n. Unfortunately, we do not know whether this holds true. Instead, the following lemma can be established (see Section 7.3 for a proof).

Lemma 7.1. *We have*

$$\min_{1\le k\le n} Y_k \le \widehat{g}_n(x) \le \max_{1\le k\le n} Y_k \tag{7.2}$$

for all $x \in \mathcal{X}$. Moreover, under assumptions (A2) *and* (A3) *we have*

$$\sup_{x\in\mathcal{X}} \left|\widehat{g}_n(x)\right| \le \max_{1\le k\le n} |Y_k| = O_p(\log n).$$

Note that under the more restrictive assumption that Y is a bounded random variable (so that $\max_k |Y_k|$ is bounded), we obtain that $\widehat{g}_n$ is also bounded. This is the case for instance, in the current status model which, as explained in the introduction, is a special case of the model we consider with $Y \in \{0, 1\}$. For this reason, the arguments developed in Groeneboom and Hendrickx [14] in the current status model cannot be directly adapted to our setting.

Now it follows from Lemma 7.1 that, with arbitrarily large probability, $\widehat{g}_n$ maximizes $\mathbb{M}_n g$ over the set of all functions g that are bounded in absolute value by $C \log n$ for some appropriately chosen $C > 0$, and take the form $g(x) = \Psi(\alpha^T x)$, $x \in \mathcal{X}$ with $(\alpha, \Psi) \in \mathcal{S}_{d-1} \times \mathcal{M}$. Denote by $\mathbb{P}_n$ the empirical distribution corresponding to $(X_1, Y_1), \dots, (X_n, Y_n)$, and let $\widehat{f}_n(x, y) = y\widehat{g}_n(x) - \widehat{g}_n^2(x)/2$ for $x \in \mathcal{X}$ and $y \in \mathbb{R}$. Since $\mathbb{M}_n g = \mathbb{P}_n f$ with

$$f(x, y) = yg(x) - g^2(x)/2 \tag{7.3}$$

for all $x \in \mathcal{X}$, $y \in \mathbb{R}$, this means that, with arbitrarily large probability, $\widehat{f}_n$ maximizes $\mathbb{P}_n f$ over the set of all functions f of the form (7.3) for some function g that is bounded in absolute value by $C \log n$ and takes the form $g(x) = \Psi(\alpha^T x)$, $x \in \mathcal{X}$ with $(\alpha, \Psi) \in \mathcal{S}_{d-1} \times \mathcal{M}$. Hence, classical arguments for maximizers of the empirical process over a class of functions (where g can be assumed to be bounded by $C \log n$) can be used to compute the rate of convergence of the estimator. This requires bounds for the entropy of the class of functions f of the form (7.3) together with a basic inequality that makes the connection between the mean of $\mathbb{M}_n g - \mathbb{M}_n g_0$ and a distance between g and g_0. The entropy bounds are given in Section 7.2 below whereas the basic inequality is given in the following lemma, which is proved in Section 7.4. For each bounded function $g : \mathcal{X} \to \mathbb{R}$, we define $\mathbb{Q}g = \int g\,d\mathbb{Q}$ and $\mathbb{M}g = \mathbb{P}f$ where f is given by (7.3) and $\mathbb{P}f = \int f\,d\mathbb{P}$, which means that $\mathbb{M}g$ is the expected value of $\mathbb{M}_n g$:

$$\mathbb{M}g = \int_{\mathcal{X}\times\mathbb{R}} \left\{ yg(x) - \frac{g^2(x)}{2} \right\} d\mathbb{P}(x, y). \tag{7.4}$$

Lemma 7.2. *Let $g : \mathcal{X} \to \mathbb{R}$ with $\mathbb{Q}g^2 < \infty$. Then, $\mathbb{M}g - \mathbb{M}g_0 \le -D^2(g, g_0)/2$ where*

$$D(g, g_0) = \left(\int_{\mathcal{X}} \big(g(x) - g_0(x)\big)^2 \, d\mathbb{Q}(x) \right)^{1/2}. \tag{7.5}$$

If classical arguments for maximizers over a class of functions are used based on the previous basic inequality and Lemma 7.1 (which allows to restrict attention to functions g that are bounded by $C \log n$, for some large $C > 0$ that does not depend on n), the obtained rate of convergence would be inflated by a logarithmic factor.

Theorem 7.3. *Assume that assumptions* (A1)–(A3) *are satisfied and that there exists a constant* $\overline{q} > 0$ *such that for all* $\alpha \in \mathcal{S}_{d-1}$, *with respect to the Lebesgue measure, the variable* $\alpha^T X$ *has a density which is bounded by* $\overline{q}$. *Then,* $D(\widehat{g}_n, g_0) = O_p(n^{-1/3}(\log n)^{5/3})$. *More precisely, for all* $\varepsilon > 0$, *there exists* $A > 0$ *that depends only on* ε, a_0 *and* M *such that*

$$P\big(D(\widehat{g}_n, g_0) > Ad(1 + \sqrt{\overline{q}R})n^{-1/3}(\log n)^{5/3}\big) \leq \varepsilon.$$

It may seem superfluous to add Theorem 7.3. However, the obtained unrefined rate of convergence will be used to get rid of the additional logarithmic factor. To explain how this works, set $v = Cn^{-1/3}(\log n)^2$ and $K = C \log n$ for some constant $C > 0$ to be chosen appropriately. Lemma 7.1 and Theorem 7.3 are used to show that with a probability that can be made arbitrarily large by choice of C, the LSE $\widehat{g}_n$ is bounded by K while $D(\widehat{g}_n, g_0)$ is smaller that v. Hence, with arbitrarily large probability, the LSE maximizes $\mathbb{M}_n g$ over the set $\mathcal{G}_{Kv}$ of all functions g of the form $g(x) = \Psi(\alpha^T x)$ with $\alpha \in \mathcal{S}_{d-1}$ and $\Psi \in \mathcal{M}$ such that $|g(x)| \leq K$ for all $x \in \mathcal{X}$ and

$$D(g, g_0) \leq v. \tag{7.6}$$

Hence, although optimization cannot be restricted to a set of functions that are uniformly bounded in n, we can work with a class of functions that are bounded in the $L_2(\mathbb{Q})$-norm. The merit of the latter is that under (A2) and equivalence of $\mathbb{Q}$ with the Lebesgue measure, a function $g \in \mathcal{G}_{Kv}$ can be shown to exceed $2K_0$ only on a subset of $\mathcal{X}$ with Lebesgue measure of maximal order $(v/K_0)^2$. The fact that considered functions g are bounded by $2K_0$ except on such a small region will balance out the large values that g might have on the same region, and this will prove to be very advantageous in computing the final entropy of the original class of functions. To estimate this entropy, each function $g(.) = \Psi(\alpha^T .)$ will be decomposed as follows:

$$g = (g - \bar{g}) + \bar{g}, \tag{7.7}$$

where $\bar{g}$ is the truncated version of g defined by

$$\bar{g}(x) = \begin{cases} g(x) & \text{if } \left|g(x)\right| \leq 2K_0, \\ 2K_0 & \text{if } g(x) > 2K_0, \\ -2K_0 & \text{if } g(x) < -2K_0. \end{cases} \tag{7.8}$$

The set of all possible $\bar{g}$ forms now a class of bounded functions on which standard arguments from empirical processes theory apply. On the other hand, the differences $g - \bar{g}$ form a set of functions whose supremum norm increases with n and which, by the discussion above, take the value zero except on regions of a very small size. Those two classes of functions will be treated with different arguments. The assumption that $\alpha^T X$ has a density bounded from above will be used to compute entropy bounds for the former class (see Lemma 7.6 and the preceding comment) whereas the assumption of a density bounded away from zero will be used to compute entropy bounds for the latter class (see Lemma 7.7 and the preceding comment). Below are some entropy results required in the proof of Theorems 7.3 and 4.1. The complete proofs of the theorems are given in Sections 7.8 and 7.9 whereas the proofs of the lemmas are given in Sections 7.3 to 7.7.

7.2. Entropy results

We begin with some notation. For any class of functions $\mathcal{F}$ equipped with a norm $\|\cdot\|$, and $\varepsilon > 0$, we denote by $H_B(\varepsilon, \mathcal{F}, \|\cdot\|)$ the corresponding bracketing entropy:

$$H_B(\varepsilon, \mathcal{F}, \|\cdot\|) = \log N_B(\varepsilon, \mathcal{F}, \|\cdot\|),$$

where $N_B(\varepsilon, \mathcal{F}, \|\cdot\|) = N$ is the smallest number of pairs of functions $(f_1^U, f_1^L), \dots, (f_N^U, f_N^L)$ such that all $\|f_j^L - f_j^U\| \le \varepsilon$ and for each $f \in \mathcal{F}$, there exists a $j \in \{1, \dots, N\}$ such that $f_j^L \le f \le f_j^U$. Moreover, assuming that X has a bounded support $\mathcal{X}$, we set $R = \sup_{x\in\mathcal{X}} \|x\|$ where $\|\cdot\|$ denotes the Euclidean norm in $\mathbb{R}^d$. We then have

$$|\alpha^T x| \le R \qquad \text{for all } x \in \mathcal{X} \tag{7.9}$$

for all $\alpha \in \mathcal{S}_{d-1}$, using the Cauchy–Schwarz inequality. In the rest of the paper, we will use the following notation

- $\|\cdot\|_{\mathbb{P}}$ and $\|\cdot\|_{\mathbb{Q}}$ are the L_2-norms corresponding to respectively $\mathbb{P}$ and $\mathbb{Q}$: $\|f\|_{\mathbb{P}}^2 = \int_{\mathcal{X}\times\mathbb{R}} f^2(x, y)\, d\mathbb{P}(x, y)$ and $\|g\|_{\mathbb{Q}}^2 = \int_{\mathcal{X}} g^2(x)\, d\mathbb{Q}(x)$ for all $f : \mathcal{X} \times \mathbb{R} \to \mathbb{R}$ and $g : \mathcal{X} \to \mathbb{R}$,
- $\mathcal{M}_K$ is the class of all nondecreasing functions on $\mathbb{R}$ that are bounded in absolute value by K,
- $\mathcal{G}_K$ is the class of functions $g(x) = \Psi(\alpha^T x)$, $x \in \mathcal{X}$ where $\alpha \in \mathcal{S}_{d-1}$, $\Psi \in \mathcal{M}_K$,
- $\mathcal{F}_K$ is the class of functions f of the form (7.3), $x \in \mathcal{X}$, $y \in \mathbb{R}$, and $g \in \mathcal{G}_K$,
- $\mathcal{G}_{Kv}$ is the class of functions $g \in \mathcal{G}_K$ satisfying the condition (7.6),
- $\mathcal{F}_{Kv}$ is the class of functions f of the form (7.3), $x \in \mathcal{X}$, $y \in \mathbb{R}$, and $g \in \mathcal{G}_{Kv}$,
- $\overline{\mathcal{G}}_{Kv}$ is the class of functions $g - \bar{g}$, where $g \in \mathcal{G}_{Kv}$ and $\bar{g}$ is given in (7.8),
- $\overline{\mathcal{F}}_{Kv}$ is the class of functions $f - \bar{f}$, where f takes the form (7.3) for some $g \in \mathcal{G}_{Kv}$ and $\bar{f}(x, y) = y\bar{g}(x) - \bar{g}^2(x)/2$, $x \in \mathcal{X}$, $y \in \mathbb{R}$.

Our starting point is the following result, which follows from Theorem 2.7.5 in van der Vaart and Wellner [26].

Lemma 7.4. *There exists a universal constant $A > 0$ such that*

$$H_B(\varepsilon, \mathcal{M}_K, \|\cdot\|_Q) \le \frac{AK}{\varepsilon}$$

for all $\varepsilon > 0$, $K > 0$, and all probability measures Q on $\mathbb{R}$, where $\|\cdot\|_Q$ is the L_2-norm corresponding to Q: $\|\Psi\|_Q^2 = \int \Psi^2\, dQ$ for all $\Psi : \mathbb{R} \to \mathbb{R}$.

The next result, which follows from Lemma 22 of Feige and Schechtman [10], gives a bound on the minimal number of subsets with diameter at most ε, say, into which $\mathcal{S}_{d-1}$ can be divided. Here, the diameter of some given subset $\mathcal{A} \subset \mathcal{S}_{d-1}$, is given by $\sup_{(x,y)\in\mathcal{A}^2} \|x-y\|$.

Lemma 7.5. *Fix $\varepsilon \in (0, \pi/2)$, and let $N(\varepsilon, \mathcal{S}_{d-1})$ be the number of subsets of equal size with diameter at most ε into which $\mathcal{S}_{d-1}$ can be partitioned. Then, there exists a universal constant $A > 0$, such that $N(\varepsilon, \mathcal{S}_{d-1}) \le (A/\varepsilon)^d$.*

In what follows, we assume that the assumptions (A1)–(A3) are satisfied. The next step is to use the results above to construct ε-brackets for the classes $\mathcal{G}_K$, $\mathcal{F}_K$, $\overline{\mathcal{G}}_{Kv}$ and $\overline{\mathcal{F}}_{Kv}$. We begin with the classes $\mathcal{G}_K$ and $\mathcal{F}_K$. In the next lemma, we assume that $\alpha^T X$ has a bounded density on a bounded support for all α. This assumptions is used in the proof to show that with Q the distribution of $\alpha^T X$ with arbitrary $\alpha \in \mathcal{S}_{d-1}$, and $\Psi \in \mathcal{M}_K$, there exists a constant C such that

$$\int_{\mathbb{R}} \big(\Psi(t+u) - \Psi(t-u)\big)\, dQ(t) \le Cu.$$

Lemma 7.6. *Assume that the assumptions* (A1)–(A3) *are satisfied and that there exists $\overline{q} > 0$ such that for all $\alpha \in \mathcal{S}_{d-1}$, with respect to the Lebesgue measure, the variable $\alpha^T X$ has a density that is bounded by $\overline{q}$. Let $K > \varepsilon > 0$. There exists a universal constant $A_1 > 0$ such that*

$$H_B\big(\varepsilon, \mathcal{G}_K, \|\cdot\|_{\mathbb{Q}}\big) \le \frac{A_1 K d(1+\sqrt{\overline{q}R})}{\varepsilon}.$$

Moreover, if $K > 1$ then there exists $A_2 > 0$ depending only on a_0 such that

$$H_B\big(\varepsilon, \mathcal{F}_K, \|\cdot\|_{\mathbb{P}}\big) \le \frac{A_2 K^2 d(1+\sqrt{\overline{q}R})}{\varepsilon}.$$

The next lemma will be used to control the differences $g(X) - \overline{g}(X) = h(\alpha^T X)$, $g \in \mathcal{G}_{Kv}$. Here, we assume that for all α, the variable $\alpha^T X$ has a density that is bounded away from zero on its support. The assumption is used to show that under (7.6), $h = 0$ except on a set whose Lebesgue measure is at most $\underline{q}^{-1} K_0^{-2} v^2$, see (7.21) below. Since the distribution of $\alpha^T X$ is also assumed to have a bounded density with respect to the Lebesgue measure, this implies that the probability that $h(\alpha^T X) \neq 0$ is of maximal order v^2, for all α, leading to a sharp bound for $\bar{\mathcal{G}}_{Kv}$ and then $\bar{\mathcal{F}}_{Kv}$.

Lemma 7.7. *Assume that the assumptions* (A1)–(A4) *are satisfied. Let $\varepsilon > 0$ and $v > 0$. There exists a constant $A_1 > 0$ depending only on K_0, $\overline{q}$, $\underline{q}$ and R such that*

$$H_B\big(\varepsilon, \bar{\mathcal{G}}_{Kv}, \|\cdot\|_{\mathbb{Q}}\big) \le \frac{A_1 K v}{\varepsilon} + d\log\left(\frac{A_1 K^2}{\varepsilon^2}\right)$$

for all $K > \varepsilon$ *such that* $Kv > \varepsilon K_0\sqrt{2R\underline{q}}$. *Moreover, there exists* $A_1 > 0$ *depending only on* K_0, $\overline{q}$, $\underline{q}$, R *and* a_0 *such that for all* $K > 1 \vee \varepsilon$ *such that* $K^2 v > \varepsilon K_0\sqrt{2R\underline{q}}$, *we have*

$$H_B\big(\varepsilon, \bar{\mathcal{F}}_{Kv}, \|\cdot\|_{\mathbb{P}}\big) \le \frac{A_1 K^2 v}{\varepsilon} + d\log\left(\frac{A_1 K^4}{\varepsilon^2}\right).$$

The next lemma will be needed to give entropy bounds in the Bernstein norm. We recall that the Bernstein norm of some function f with respect to $\mathbb{P}$ is defined by

$$\|f\|_{B,\mathbb{P}} = \big(2\mathbb{P}\big(e^{|f|} - 1 - |f|\big)\big)^{1/2}.$$

Although not technically a norm, it is typically referred to as such in the literature (van der Vaart and Wellner [26], page 324), and we do not stray from this in what follows. With $\tilde{C} > 0$ a constant appropriately chosen, Lemma 7.8 below derives from Lemmas 7.6 and 7.7 upper bounds on the bracketing number of the classes

$$\widetilde{\mathcal{F}}_{Kv} := \big\{(f - f_0)\tilde{C}^{-1}, f \in \mathcal{F}_{Kv}\big\}, \tag{7.10}$$

where $f_0(x, y) = yg_0(x) - g_0^2(x)/2$ and

$$\tilde{\bar{\mathcal{F}}}_{Kv} := \big\{f\tilde{C}^{-1}, f \in \bar{\mathcal{F}}_{Kv}\big\}$$

with respect to the Bernstein norm. This will enable us to use Lemma 3.4.3 of van der Vaart and Wellner [26] which does not require the class of functions of interest to be bounded.

Lemma 7.8. *Assume that the assumptions* (A1)–(A3) *are satisfied and that there exists* $\overline{q} > 0$ *such that for all* $\alpha \in \mathcal{S}_{d-1}$, *with respect to the Lebesgue measure, the variable* $\alpha^T X$ *has a density that is bounded by* $\overline{q}$. *Let* $\varepsilon > 0$ *and* $v > 0$. *Let* M *be the same constant from the moment condition* (4.3) *of Assumption* (A3). *Let* $\tilde{C} = 4MK^2$ *such that* $K \geq (2K_0) \vee 2$. *Then, there exist constants* $A_1 > 0$ *and* $A_2 > 0$ *that depend on* a_0 *and* M *only such that*

$$H_B\big(\varepsilon, \widetilde{\mathcal{F}}_{Kv}, \|\cdot\|_{B,\mathbb{P}}\big) \le \frac{A_1 d(1 + \sqrt{\overline{q}}R)}{\varepsilon} \quad \textit{and} \quad \|\tilde{f}\|_{B,\mathbb{P}} \le A_2 v$$

for all $\tilde{f} \in \widetilde{\mathcal{F}}_{Kv}$. *If moreover, the assumption* (A4) *is fulfilled, then there exist constants* $A_1 > 0$ *and* $A_2 > 0$ *depending only on* K_0, R, a_0, M, $\bar{q}$ *and* $\underline{q}$ *such that*

$$H_B\big(\varepsilon, \tilde{\bar{\mathcal{F}}}_{Kv}, \|\cdot\|_{B,\mathbb{P}}\big) \le \frac{A_1 v}{\varepsilon} + d\log\left(\frac{A_1}{\varepsilon^2}\right) \quad \textit{and} \quad \|\tilde{f}\|_{B,\mathbb{P}} \le A_2 v$$

for all $\tilde{f} \in \tilde{\bar{\mathcal{F}}}_{Kv}$, *provided that* $K^2 v > A_2\varepsilon$ *and* $K > \varepsilon$.

Note that the condition $\varepsilon < 2$ guarantees that $K > \varepsilon$ since $K \geq 2$. Also, we point out that the constants A_1 and A_2 may not be the same ones as in Lemma 7.7 but we can always increase their respective values so that they are suitable for Lemma 7.8.

7.3. Proof of Lemma 7.1

For a fixed $\alpha \in \mathcal{S}_{d-1}$, let $\widehat{\Psi}_n^\alpha$ be a minimizer of $h_n(\Psi, \alpha)$ over $\Psi \in \mathcal{M}$. It follows from Theorem 1 in Barlow et al. [2] that $\widetilde{\Psi}_n(Z_k) = d_k$ for $k = 1, \ldots, m$, where $Z_1 < \cdots < Z_m$ are the ordered distinct values of $\alpha^T X_1, \ldots, \alpha^T X_n$, and $d_1 \le \cdots \le d_m$ are the left derivatives of the greatest convex minorant of the cumulative sum diagram defined by the set of points

$$\left\{(0,0), \left(\sum_{i=1}^n \mathbb{I}_{\alpha^T X_i \le Z_k}, \sum_{i=1}^n Y_i \mathbb{I}_{\alpha^T X_i \le Z_k}\right), k = 1, \ldots, m\right\}.$$

Hence, we have

$$\min_{1\le k\le m} \frac{\sum_{i=1}^n Y_i \mathbb{I}_{\alpha^T X_i \le Z_k}}{\sum_{i=1}^n \mathbb{I}_{\alpha^T X_i \le Z_k}} \le \widehat{\Psi}_n^\alpha\big(\alpha^T X_j\big) \le \max_{0\le k\le m-1} \frac{\sum_{i=1}^n Y_i \mathbb{I}_{\alpha^T X_i > Z_k}}{\sum_{i=1}^n \mathbb{I}_{\alpha^T X_i > Z_k}},$$

with $Z_0 = -\infty$, for all $j = 1, \ldots, n$. Therefore, $\min_{1\le i\le n} Y_i \le \widehat{\Psi}_n^\alpha(\alpha^T X_j) \le \max_{1\le i\le n} Y_i$ for all $\alpha \in \mathcal{S}_{d-1}$ and $j = 1, \ldots, n$. The inequalities in (7.2) follow since by definition, $\widehat{g}_n(x)$ takes the form $\widehat{\Psi}_n^\alpha(\alpha^T x)$ for all x, where α is replaced by the LSE $\widehat{\alpha}_n \in \mathcal{S}_{d-1}$.

Now we prove that under the assumptions (A2) and (A3),

$$\max_{1\le i\le n} |Y_i| = O_p(\log n). \tag{7.11}$$

For an integer $s \ge 2$, it follows from convexity of the function $z \mapsto |z|^s$ on $\mathbb{R}$ that

$$\begin{aligned} E\big[\big|Y - g_0(x)\big|^s | X = x\big] &\le 2^{s-1}\big(E\big[|Y|^s | X = x\big] + \big|g_0(x)\big|^s\big) \\ &\le 2^{s-1}\big(s! a_0 M^{s-2} + K_0^s\big) \le s! b_0\big(M'\big)^{s-2} \end{aligned}$$

with $b_0 = 2(a_0 + K_0^2)$ and $M' = 2(M \vee K_0)$. Now, using Lemma 2.2.11 of van der Vaart and Wellner [26] with $n = 1$ and $\tilde{Y} = Y - g_0(x)$ and after integrating out with respect to $d\mathbb{Q}$, we obtain

$$P\big(\big|Y - g_0(X)\big| > t\big) \le 2\exp\left(\frac{-t^2}{2(2b_0 + M't)}\right)$$

for all $t > 0$. Hence, with $t = C\log n$ such that $K_0 < C\log(n)/2$ we have that

$$\begin{aligned} P\left(\max_{1\le i\le n} |Y_i| > C\log n\right) &\le \sum_{i=1}^n P\big(\big|Y_i - g_0(X_i)\big| > C\log(n)/2\big) \\ &\le 2n\exp\left(\frac{-C^2\log n}{4(4b_0(\log n)^{-1} + M'C)}\right) \end{aligned}$$

which converges to 0 as $n \to \infty$ provided that C is sufficiently large so that $8M'C < C^2$. Lemma 7.1 follows.

7.4. Proof of Lemma 7.2

Since $E[Y|X=x]=g_0(x)$ we have

$$\mathbb{M}g - \mathbb{M}g_0 = \int_{\mathcal{X}} \left\{ g_0(x)\big(g(x)-g_0(x)\big) - \frac{g(x)^2}{2} + \frac{g_0(x)^2}{2} \right\} d\mathbb{Q}(x)$$
$$= -\frac{1}{2}\int_{\mathcal{X}} \big(g(x)-g_0(x)\big)^2\, d\mathbb{Q}(x),$$

which proves Lemma 7.2.

7.5. Proof of Lemma 7.6

Let $\varepsilon_\alpha = \varepsilon^2 K^{-2} \in (0,1)$. By Lemma 7.5, $\mathcal{S}_{d-1}$ can be covered by N neighborhoods with diameter at most ε_α where $N \le (A\varepsilon_\alpha^{-1})^d$ with $A>0$ a universal constant. Let $\{\alpha_1,\dots,\alpha_N\}$ denote elements of each of these neighborhoods. Now, consider an arbitrary $g \in \mathcal{G}_K$. Then, $g(x)=\Psi(\alpha^T x)$, $x \in \mathcal{X}$, for some $\Psi \in \mathcal{M}_K$ and $\alpha \in \mathcal{S}_{d-1}$. We can find $i \in \{1,\dots,N\}$ such that $\|\alpha-\alpha_i\| \le \varepsilon_\alpha$. Then, using the monotonicity of Ψ together with the Cauchy–Schwarz inequality we can write for all $x \in \mathcal{X}$ that

$$g(x) = \Psi\big(\alpha_i^T x + (\alpha-\alpha_i)^T x\big) \le \Psi\big(\alpha_i^T x + \varepsilon_\alpha R\big)$$

and $g(x) \ge \Psi(\alpha_i^T x - \varepsilon_\alpha R)$. Then, Lemma 7.4 implies that with $N' = \exp(AK\varepsilon^{-1})$, at the cost of increasing A, we can find brackets $[\Psi_j^L, \Psi_j^U]$ covering the class of functions $\mathcal{M}_K$ such that

$$\int_{\mathbb{R}} \big(\Psi_j^U(t)-\Psi_j^L(t)\big)^2\, dQ_i^-(t) \le \varepsilon^2 \quad \text{and} \quad \int_{\mathbb{R}} \big(\Psi_j^U(t)-\Psi_j^L(t)\big)^2\, dQ_i^+(t) \le \varepsilon^2$$

for $j=1,\dots,N'$, where Q_i^- and Q_i^+ respectively, denote the distribution of $\alpha_i^T X - \varepsilon_\alpha R$ and $\alpha_i^T X + \varepsilon_\alpha R$. Now returning to g, and using that $\Psi \in \mathcal{M}_K$, we can see that

$$\Psi_j^L\big(\alpha_i^T x - \varepsilon_\alpha R\big) \le g(x) \le \Psi_j^U\big(\alpha_i^T x + \varepsilon_\alpha R\big) \tag{7.12}$$

for some $j=1,\dots,N'$ and all $x \in \mathcal{X}$. We will show that there exists $B>0$ depending only on $\overline{q}$, R such that the new bracket $[\Psi_j^L(\alpha_i^T x - \varepsilon_\alpha R), \Psi_j^U(\alpha_i^T x + \varepsilon_\alpha R)]$, $x \in \mathcal{X}$ satisfies

$$\left(\int_{\mathcal{X}} \big(\Psi_j^U\big(\alpha_i^T x + \varepsilon_\alpha R\big) - \Psi_j^L\big(\alpha_i^T x - \varepsilon_\alpha R\big)\big)^2\, d\mathbb{Q}(x)\right)^{1/2} \le B\varepsilon. \tag{7.13}$$

It follows from the Minkowski inequality that the left-hand side in (7.13) is at most

$$\begin{aligned}&\left(\int_{\mathcal{X}}\left(\Psi\left(\alpha_i^T x-\varepsilon_\alpha R\right)-\Psi_j^L\left(\alpha_i^T x-\varepsilon_\alpha R\right)\right)^2 d\mathbb{Q}(x)\right)^{1/2}\\&\quad+\left(\int_{\mathcal{X}}\left(\Psi_j^U\left(\alpha_i^T x+\varepsilon_\alpha R\right)-\Psi\left(\alpha_i^T x+\varepsilon_\alpha R\right)\right)^2 d\mathbb{Q}(x)\right)^{1/2}\\&\quad+\left(\int_{\mathcal{X}}\left(\Psi\left(\alpha_i^T x+\varepsilon_\alpha R\right)-\Psi\left(\alpha_i^T x-\varepsilon_\alpha R\right)\right)^2 d\mathbb{Q}(x)\right)^{1/2}. \qquad (7.14)\end{aligned}$$

We have

$$\begin{aligned}\int_{\mathcal{X}}\left(\Psi\left(\alpha_i^T x-\varepsilon_\alpha R\right)-\Psi_j^L\left(\alpha_i^T x-\varepsilon_\alpha R\right)\right)^2 d\mathbb{Q}(x)&=\int_{\mathbb{R}}\left(\Psi(t)-\Psi_j^L(t)\right)^2 dQ_i^-(t)\\&\le\varepsilon^2\end{aligned}$$

and a similar bound is found for the square of the second integral in (7.14). Hence, the left-hand side in (7.13) is less than or equal to

$$2\varepsilon+\left(\int_{\mathbb{R}}\left(\Psi(t+\varepsilon_\alpha R)-\Psi(t-\varepsilon_\alpha R)\right)^2 dQ_i(t)\right)^{1/2},$$

where Q_i is the distribution of $\alpha_i^T X$. By monotonicity of Ψ and the fact that it is bounded in absolute value by K, we can write

$$\begin{aligned}\int_{\mathbb{R}}\left(\Psi(t+\varepsilon_\alpha R)-\Psi(t-\varepsilon_\alpha R)\right)^2 dQ_i(t)&\le 2K\int_{\mathbb{R}}\left(\Psi(t+\varepsilon_\alpha R)-\Psi(t-\varepsilon_\alpha R)\right) dQ_i(t)\\&\le 2K\overline{q}\int_{-R}^{R}\left(\Psi(t+\varepsilon_\alpha R)-\Psi(t-\varepsilon_\alpha R)\right) dt,\end{aligned}$$

with $\overline{q}$ an upper bound of the density of Q_i, that is supported on $[-R, R]$, with respect to the Lebesgue measure. This is at most

$$2K\overline{q}\left(\int_{R-\varepsilon_\alpha R}^{R+\varepsilon_\alpha R}\Psi(t)\,dt-\int_{-R-\varepsilon_\alpha R}^{-R+\varepsilon_\alpha R}\Psi(t)\,dt\right)\le 8\overline{q}R\varepsilon^2,$$

using that $\varepsilon_\alpha=\varepsilon^2/K^2$. Hence,

$$\left(\int_{\mathcal{X}}\left(\Psi\left(\alpha_i^T x+\varepsilon_\alpha R\right)-\Psi\left(\alpha_i^T x-\varepsilon_\alpha R\right)\right)^2 d\mathbb{Q}(x)\right)^{1/2}\le(8\overline{q}R)^{1/2}\varepsilon. \qquad (7.15)$$

Combining the inequalities above, we get the claimed inequality in (7.13) with $B = 2 + (8\overline{q}R)^{1/2}$. It follows that

$$\begin{aligned} H_B\big(B\varepsilon, \mathcal{G}_K, \|\cdot\|_{\mathbb{Q}}\big) &\le \log N + \log N' \\ &\le d\log\big(AK^2\varepsilon^{-2}\big) + AK\varepsilon^{-1} \\ &\le K\varepsilon^{-1}\big(dA^{1/2} + A\big) \end{aligned} \tag{7.16}$$

since $\log x \le \sqrt{x}$ for all $x > 0$. The first assertion of Lemma 7.6 follows.

To prove the second assertion, we need to build brackets for the class of functions $(x, y) \mapsto yg(x)$, $x \in \mathcal{X}$, $y \in \mathbb{R}$, and then for the class of functions g^2, with $g \in \mathcal{G}_K$. In the following, we denote the former class by $\mathcal{G}_1$ and the latter by $\mathcal{G}_2$ with which we begin. Note that $g^2(x) = s(x) = \Psi^2(\alpha^T x) = h(\alpha^T x)$ for some function h that is either monotone non-decreasing, monotone non-increasing or U-shaped depending on the sign of Ψ. Hence, the function h can be always decomposed into the difference of two monotone functions that are bounded by K^2. If $K^2 > \varepsilon$ (which holds for all $\varepsilon > 0$ and $K > \varepsilon$ such that $K > 1$), we can use similar arguments as above to conclude that there exists a universal constant $B_0 > 0$ such that

$$H_B\big(\varepsilon, \mathcal{G}_2, \|\cdot\|_{\mathbb{Q}}\big) \le \frac{B_0 K^2 d(1 + \sqrt{\overline{q}R})}{\varepsilon}. \tag{7.17}$$

Using the fact that any element $s \in \mathcal{G}_2$ satisfies

$$\int_{\mathbb{R}}\int_{\mathcal{X}} s^2(x)\, d\mathbb{P}(x, y) = \int_{\mathcal{X}} s^2(x)\, d\mathbb{Q}(x),$$

it follows that

$$H_B\big(\varepsilon, \mathcal{G}_2, \|\cdot\|_{\mathbb{P}}\big) \le \frac{B_0 K^2 d(1 + \sqrt{\overline{q}R})}{\varepsilon}.$$

Now we turn to $\mathcal{G}_1$. With $N = N_B(\varepsilon, \mathcal{G}_K, \|\cdot\|_{\mathbb{Q}})$, we will denote by $\{(g_i^L, g_i^U), i \in \{1, \dots, N\}\}$ a cover of ε-brackets for $\mathcal{G}_K$. For all $i = 1, \dots, N$, define

$$k_i^U(x, y) = \begin{cases} yg_i^U(x) & \text{if } y \ge 0, \\ yg_i^L(x) & \text{if } y \le 0, \end{cases} \qquad k_i^L(x, y) = \begin{cases} yg_i^L(x) & \text{if } y \ge 0, \\ yg_i^U(x) & \text{if } y \le 0. \end{cases} \tag{7.18}$$

Now, take $g \in \mathcal{G}_K$ and let $i \in \{1, \dots, N\}$ such that $g_i^L \le g \le g_i^U$. Then, we have $k_i^L(x, y) \le yg(x) \le k_i^U(x, y)$ so that $\{(k_i^L, k_i^U), i \in \{1, \dots, N\}\}$ form a bracketing cover for $\mathcal{G}_1$. We will now compute its size. We have that

$$\begin{aligned} \int_{\mathbb{R}\times\mathcal{X}} \big(k_i^U(x, y) - k_i^L(x, y)\big)^2\, d\mathbb{P}(x, y) &= \int_{\mathbb{R}\times\mathcal{X}} y^2 \times \big(g_i^U(x) - g_i^L(x)\big)^2\, d\mathbb{P}(x, y) \\ &\le 2a_0 \int_{\mathcal{X}} \big(g_i^U(x) - g_i^L(x)\big)^2\, d\mathbb{Q}(x) \\ &\le 2a_0\varepsilon^2, \end{aligned}$$

where a_0 is taken from (4.3). Hence,

$$H_B\big(\sqrt{2a_0}\varepsilon, \mathcal{G}_1, \|\cdot\|_{\mathbb{P}}\big) \le H_B\big(\varepsilon, \mathcal{G}_K, \|\cdot\|_{\mathbb{Q}}\big) \le \frac{A_1 K d(1+\sqrt{\overline{q}}R)}{\varepsilon}, \tag{7.19}$$

using the first assertion of the lemma. But for all $\varepsilon > 0$, we have

$$H_B\big(\varepsilon, \mathcal{F}_K, \|\cdot\|_{\mathbb{P}}\big) \le H_B\big(\varepsilon/2, \mathcal{G}_1, \|\cdot\|_{\mathbb{P}}\big) + H_B\big(\varepsilon, \mathcal{G}_2, \|\cdot\|_{\mathbb{P}}\big) \tag{7.20}$$

and hence we obtain the second assertion of the lemma, which completes the proof.

7.6. Proof of Lemma 7.7

Let $g \in \mathcal{G}_{Kv}$ and $\Psi \in \mathcal{M}_K$, $\alpha \in \mathcal{S}_{d-1}$ such that $g(x) = \Psi(\alpha^T x)$. We shall show below that $|\Psi(t)| \le 2K_0$ except on a region of small size. To do that, we shall use the condition (7.6) together with the triangle inequality to get

$$\int_{\mathcal{X}} \mathbb{1}_{\{|\Psi(\alpha^T x)|>2K_0\}}\, d\mathbb{Q}(x) \le \int_{\mathcal{X}} \mathbb{1}_{\{|\Psi(\alpha^T x)-\Psi_0(\alpha_0^T x)|>K_0\}}\, d\mathbb{Q}(x) \le \int_{\mathcal{X}} \left(\frac{\Psi(\alpha^T x) - g_0(x)}{K_0}\right)^2 d\mathbb{Q}(x) \le v^2 K_0^{-2}.$$

With a, b the boundaries of the interval $\{\alpha^T x, x \in \mathcal{X}\}$ and Q_α the distribution of $\alpha^T X$ we have

$$\int_{\mathcal{X}} \mathbb{1}_{\{|\Psi(\alpha^T x)|>2K_0\}}\, d\mathbb{Q}(x) = \int_a^b \mathbb{1}_{\{|\Psi(t)|>2K_0\}}\, dQ_\alpha(t) \ge \underline{q} \int_a^b \mathbb{1}_{\{|\Psi(t)|>2K_0\}}\, dt.$$

Combining the two preceding displays, we conclude that

$$\int_a^b \mathbb{1}_{\{|\Psi(t)|>2K_0\}}\, dt \le D_2 v^2, \tag{7.21}$$

where $D_2 = \underline{q}^{-1} K_0^{-2}$. By monotonicity of Ψ, this means that $|\Psi(t)| \le 2K_0$ for all t in the interval $[a + D_2 v^2, b - D_2 v^2]$. Now, from (7.8), $g(x) - \bar{g}(x)$ takes the form of $h(\alpha^T x)$ where $h \in \mathcal{M}_K$ is such that

$$h(t) = \begin{cases} 0 & \text{if } \big|\Psi(t)\big| \le 2K_0, \\ \Psi(t) + 2K_0 & \text{if } \Psi(t) < -2K_0, \\ \Psi(t) - 2K_0 & \text{if } \Psi(t) > 2K_0. \end{cases}$$

Hence, for $t \in [a, b]$ we can have $h(t) \neq 0$ only for $t \in [a, a + D_2 v^2] \cup [b - D_2 v^2, b]$. Consider $\{\alpha_1, \dots, \alpha_N\}$ the grid providing a ε_α-cover for $\mathcal{S}_{d-1}$ with $\varepsilon_\alpha = \varepsilon^2/K^2$ and $N \leq (A\varepsilon_\alpha^{-1})^d$, see Lemma 7.5. Similar to the proof of Lemma 7.6, with α_i such that $\|\alpha - \alpha_i\| \leq \varepsilon_\alpha$, we then have

$$h\big(\alpha_i^T x - \varepsilon_\alpha R\big) \leq g(x) - \bar{g}(x) \leq h\big(\alpha_i^T x + \varepsilon_\alpha R\big) \tag{7.22}$$

for all $x \in \mathcal{X}$, where h is considered on the interval $[a_i, b_i]$ where

$$a_i = \inf\big\{\alpha_i^T x - \varepsilon_\alpha R, x \in \mathcal{X}\big\} \quad \text{and} \quad b_i = \sup\big\{\alpha_i^T x + \varepsilon_\alpha R, x \in \mathcal{X}\big\}.$$

Note that the support of $\alpha_i^T X$, $\alpha_i^T X - \varepsilon_\alpha R$ and $\alpha_i^T X + \varepsilon_\alpha R$ are all included in $[a_i, b_i]$, and we have $|a - a_i| \leq 2\varepsilon_\alpha R$ and $|b - b_i| \leq 2\varepsilon_\alpha R$. From what precedes, for $t \in [a_i, b_i]$ we can have $h(t) \neq 0$ only for $t \in \mathcal{I}_{i,1} \cup \mathcal{I}_{i,2}$ where

$$\mathcal{I}_{i,1} = \big[a_i, a_i + D_2 v^2 + 2\varepsilon_\alpha R\big] \quad \text{and} \quad \mathcal{I}_{i,2} = \big[b_i - D_2 v^2 - 2\varepsilon_\alpha R, b_i\big]$$

have length at most $2D_2 v^2$ under the assumption that $Kv > \varepsilon K_0\sqrt{2R\underline{q}}$. Hence, we only need to construct brackets for the class of monotone functions on $[a_i, b_i]$ that are bounded by K and constant equal to zero outside $\mathcal{I}_{i,1} \cup \mathcal{I}_{i,2}$. This can be done by using Lemma 7.4 with Q denoting the uniform distribution on $\mathcal{I}_{i,1} \cup \mathcal{I}_{i,2}$: it follows from that lemma that with $N_i \leq \exp(2A\sqrt{D_2}Kv/\varepsilon)$, we can find brackets (h_j^L, h_j^U), $j = 1, \dots, N_i$ such that every function in the class belongs to $[h_j^L, h_j^U]$ for some j, and

$$\begin{aligned}\int_{\mathbb{R}} \big(h_j^L(t) - h_j^U(t)\big)^2 \, dt &\leq 4D_2 v^2 \int_{\mathcal{I}_{i,1} \cup \mathcal{I}_{i,2}} \big(h_j^L(t) - h_j^U(t)\big)^2 \, dQ(t) \\ &\leq \varepsilon^2\end{aligned} \tag{7.23}$$

for all j. Note that we have omitted writing the dependence on i for the functions in the brackets.

Let $j \in \{1, \dots, N_i\}$ such that $h_j^L \leq h \leq h_j^U$ on $[a_i, b_i]$. By (7.22), we have

$$b^L(x) \equiv h_j^L\big(\alpha_i^T x - \varepsilon_\alpha R\big) \leq g(x) - \bar{g}(x) \leq h_j^U\big(\alpha_i^T x + \varepsilon_\alpha R\big) \equiv b^U(x)$$

for all $x \in \mathcal{X}$ and it remains to compute the size of the obtained brackets. By the Minkowski inequality, with Q_i the distribution of $\alpha_i^T X$ we have

$$\begin{aligned}\|b^U - b^L\|_{\mathbb{Q}} &= \left(\int_{\mathbb{R}} \big(h_j^U(t + \varepsilon_\alpha R) - h_j^L(t - \varepsilon_\alpha R)\big)^2 \, dQ_i(t)\right)^{1/2} \\ &\leq \left(\overline{q} \int_{a_i + \varepsilon_\alpha R}^{b_i - \varepsilon_\alpha R} \big(h_j^U(t + \varepsilon_\alpha R) - h_j^L(t - \varepsilon_\alpha R)\big)^2 \, dt\right)^{1/2} \\ &\leq \left(\overline{q} \int_{a_i + \varepsilon_\alpha R}^{b_i - \varepsilon_\alpha R} \big(h_j^U(t + \varepsilon_\alpha R) - h_j^U(t - \varepsilon_\alpha R)\big)^2 \, dt\right)^{1/2} \\ &\quad + \left(\overline{q} \int_{a_i + \varepsilon_\alpha R}^{b_i - \varepsilon_\alpha R} \big(h_j^U(t - \varepsilon_\alpha R) - h_j^L(t - \varepsilon_\alpha R)\big)^2 \, dt\right)^{1/2}.\end{aligned}$$

Since h_j^L and h_j^U can be chosen monotone on $[a_i, b_i]$ and bounded in absolute value by K we conclude from (7.23) that

$$\|b^U - b^L\|_{\mathbb{Q}} \le \left(2K\overline{q}\int_{a_i}^{b_i - 2\varepsilon_\alpha R}\left(h_j^U(t + 2\varepsilon_\alpha R) - h_j^U(t)\right)dt\right)^{1/2} + \sqrt{\overline{q}}\varepsilon$$
$$\le \left(8\varepsilon_\alpha R K^2 \overline{q}\right)^{1/2} + \sqrt{\overline{q}}\varepsilon.$$

Since by definition, $\varepsilon_\alpha = \varepsilon^2/K^2$, this means that we have $C\varepsilon$-brackets where C depends on $\overline{q}$ and R only. Hence,

$$H_B\left(C\varepsilon, \bar{\mathcal{G}}_{Kv}, \|\cdot\|_{\mathbb{Q}}\right) \le \log(N_i) + \log(N) \le \frac{2A\sqrt{D_2}Kv}{\varepsilon} + d\log\frac{AK^2}{\varepsilon^2}$$

and the first assertion of the lemma follows.

To prove the second assertion, recall that $(f - \bar{f})(x, y) = y(g(x) - \bar{g}(x)) - \frac{1}{2}(g^2(x) - \bar{g}^2(x))$. We need then to build brackets for the functions $(x, y) \mapsto y(g(x) - \bar{g}(x))$, and those for the functions $x \mapsto g^2(x) - \bar{g}^2(x)$. The construction of the brackets goes along the same line as the construction of the brackets for the classes $\mathcal{G}_1$ and $\mathcal{G}_2$ in the proof of Lemma 7.6 above, where for the latter class, we use the fact that all functions in the class take the form $h(\alpha^T x)$ where h is either monotone or U-shaped, and vanishes when $|g(\alpha^T x)| \le 2K_0$, which implies that the function vanishes except on at most two intervals of maximal length $2D_2 v^2$.

Hence, we can find $A_1 > 0$ and $A_2 > 0$ depending on K_0, $\overline{q}$, $\underline{q}$, R and a_0 such that

$$H_B\left(\varepsilon, \bar{\mathcal{F}}_{Kv}, \|\cdot\|_{\mathbb{P}}\right) \le \frac{A_1 K^2 v}{\varepsilon} + d\log\frac{A_1 K^4}{\varepsilon^2}$$

for $K > 1 \vee \varepsilon$ such that $K^2 v > A_2\varepsilon$. This completes the proof of Lemma 7.7.

7.7. Proof of Lemma 7.8

We start by noting that entropy bound on the class $\widetilde{\mathcal{F}}_{Kv}$ is smaller than the entropy bound for the class $\widetilde{\mathcal{F}}_K$ as a consequence of inclusion of the former class in the latter. We will now show that the upper bound with respect to the Bernstein norm for the class $\widetilde{\mathcal{F}}_{Kv}$ is of the claimed order. Let $N_1 = N_B(\varepsilon, \mathcal{G}_K, \|\cdot\|_{\mathbb{Q}})$ and $N_2 = N_B(\varepsilon, \mathcal{G}_2, \|\cdot\|_{\mathbb{Q}})$, where $\mathcal{G}_2$ is the class of functions $\{x \mapsto g^2(x), g \in \mathcal{G}_K\}$. Consider brackets $[g_j^L, g_j^U]$, $j = 1, \ldots, N_1$ covering $\mathcal{G}_K$ and $[s_i^L, s_i^U]$, $i = 1, \ldots, N_2$ covering $\mathcal{G}_2$. Note that g_i^L and g_i^U can be always taken to be bounded by K, because otherwise we can take instead $g_i^L \vee (-K)$ and $g_i^U \wedge K$. The same thing holds for s_i^L and s_i^U which can be taken to be bounded by K^2. Let $f \in \mathcal{F}_{Kv}$ and $\tilde{C} > 0$ a fixed constant to be chosen later. Then, there exists $(i, j) \in \{1, \ldots, N_1\} \times \{1, \ldots, N_2\}$ such that $f_{i,j}^L \le f \le f_{i,j}^U$ and

$$f_{i,j}^L(x, y) = \begin{cases} y g_j^L(x) - \dfrac{1}{2}s_i^U(x) & \text{if } y \ge 0, \\ y g_j^U(x) - \dfrac{1}{2}s_i^U(x) & \text{if } y < 0, \end{cases}$$

$$f^U_{i,j}(x,y)=\begin{cases} yg^U_j(x)-\dfrac{1}{2}s^L_i(x) & \text{if } y\geq 0,\\ yg^L_j(x)-\dfrac{1}{2}s^L_i(x) & \text{if } y<0.\end{cases}$$

Now note that for a given function h such that h^k is $\mathbb{P}$-integrable for all $k\geq 2$, we can write $\|h\|^2_{B,\mathbb{P}}=2\sum_{k=2}^{\infty}(\int|h|^k\,d\mathbb{P})/k!$. Hence, by convexity of $x\mapsto x^k$ for $k\geq 2$, we have

$$\begin{aligned}&\left\|\frac{f^U_{i,j}-f^L_{i,j}}{\tilde C}\right\|^2_{B,\mathbb{P}}\\ &\quad\leq 2\sum_{k=2}^{\infty}\frac{2^{k-1}}{k!\tilde C^k}\int_{\mathcal{X}\times\mathbb{R}}\left\{|y|^k\left|g^U_j(x)-g^L_j(x)\right|^k+\frac{1}{2^k}\left|s^U_i(x)-s^L_i(x)\right|^k\right\}d\mathbb{P}(x,y).\end{aligned}$$

Integrating first with respect to y using the assumption (A3) yields that the right-hand side is at most

$$\begin{aligned}&2\sum_{k=2}^{\infty}\frac{2^{k-1}}{k!\tilde C^k}\left\{a_0M^{k-2}k!\times(2K)^{k-2}\int_{\mathcal{X}}\left|g^U_j-g^L_j\right|^2d\mathbb{Q}+\frac{(2K^2)^{(k-2)}}{2^k}\int_{\mathcal{X}}\left|s^U_j-s^L_j\right|^2d\mathbb{Q}\right\}\\ &\quad\leq 2\sum_{k=2}^{\infty}\frac{2^{k-1}}{k!\tilde C^k}\left\{a_0M^{k-2}k!\times(2K)^{k-2}\varepsilon^2+\frac{(2K^2)^{(k-2)}}{2^k}\varepsilon^2\right\}.\end{aligned}$$

Hence,

$$\left\|(f^U_{i,j}-f^L_{i,j})\tilde C^{-1}\right\|^2_{B,\mathbb{P}}\leq\frac{2}{\tilde C^2}\left(2a_0\sum_{k=2}^{\infty}\left(\frac{4MK}{\tilde C}\right)^{k-2}+\frac{1}{4}\sum_{k=2}^{\infty}\left(\frac{2K^2}{\tilde C}\right)^{k-2}\frac{1}{(k-2)!}\right)\varepsilon^2,$$

using the fact that $k!\geq 2(k-2)!$ for all $k\geq 2$. We conclude that

$$\left\|(f^U_{i,j}-f^L_{i,j})\tilde C^{-1}\right\|^2_{B,\mathbb{P}}\leq\frac{2}{\tilde C^2}(2a_0\vee 1/4)\left(\frac{1}{1-4MK/\tilde C}+e^{2K^2/\tilde C}\right)\varepsilon^2.$$

Since $K\geq 2$, the choice $\tilde C=4MK^2$ yields

$$\begin{aligned}\left\|(f^U_{i,j}-f^L_{i,j})\tilde C^{-1}\right\|^2_{B,\mathbb{P}}&\leq\frac{2}{\tilde C^2}(2a_0\vee 1/4)\left(\frac{K}{K-1}+e^{(2M)^{-1}}\right)\varepsilon^2\\ &\leq B^2K^{-4}\varepsilon^2,\end{aligned}\tag{7.24}$$

where B depends on a_0 and M only. Using (7.17) and the first assertion of Lemma 7.6, this means that there exists a universal constant A_2 such that

$$H_B\left(\frac{B\varepsilon}{K^2},\widetilde{\mathcal{F}}_{Kv},\|\cdot\|_{B,\mathbb{P}}\right)\leq\log N_1+\log N_2\leq\frac{A_2K^2d(1+\sqrt{q}R)}{\varepsilon}.\tag{7.25}$$

This in turn implies that

$$H_B\big(\varepsilon, \widetilde{\mathcal{F}}_{Kv}, \|\cdot\|_{B,\mathbb{P}}\big) \le \frac{A_2 B K^2 d(1+\sqrt{q}R)}{\varepsilon K^2} = \frac{A_1 d(1+\sqrt{q}R)}{\varepsilon}$$

as claimed in the statement of the lemma.

To show the second claim, we will use again the series expansion of the Bernstein norm. Similar as above, using that g_0 is bounded by $K_0 \le K$, and that for arbitrary $\tilde{f} = (f - f_0)\tilde{C}^{-1} \in \widetilde{\mathcal{F}}_{Kv}$, the corresponding $g \in \mathcal{G}_{Kv}$ satisfies (7.6), we obtain

$$\begin{aligned}
\|\tilde{f}\|^2_{B,\mathbb{P}} &\le \sum_{k=2}^{\infty} \frac{2^k}{k!\tilde{C}^k}\bigg\{a_0 M^{k-2} k!(2K)^{k-2}\int_{\mathcal{X}} (g-g_0)^2\, d\mathbb{Q} \\
&\quad + \frac{(2K^2)^{(k-2)}}{2^k}\int_{\mathcal{X}} \big(g^2 - g_0^2\big)^2\, d\mathbb{Q}\bigg\} \\
&\le \sum_{k=2}^{\infty} \frac{2^k}{k!\tilde{C}^k}\bigg\{a_0 M^{k-2} k!(2K)^{k-2}\int_{\mathcal{X}} (g-g_0)^2\, d\mathbb{Q} \\
&\quad + \frac{2(2K^2)^{k-1}}{2^k}\int_{\mathcal{X}} (g-g_0)^2\, d\mathbb{Q}\bigg\} \\
&\le \bigg(\frac{4a_0}{\tilde{C}^2}\sum_{k=2}^{\infty}\bigg(\frac{4MK}{\tilde{C}}\bigg)^{k-2} + \frac{(4K)^2}{\tilde{C}^2}\sum_{k=2}^{\infty}\bigg(\frac{4K^2}{\tilde{C}}\bigg)^{k-2}\frac{1}{k!}\bigg)v^2 \\
&\le \bigg(\frac{a_0}{2M^2K^4} + \frac{1}{4M^2K^2}e^{1/M}\bigg)v^2, \qquad (7.26)
\end{aligned}$$

using that $K \ge 2$. The second claim follows.

Using the same arguments as above in combination with the entropy bound for $\bar{\mathcal{F}}_{Kv}$ obtained in Lemma 7.7 we can show that

$$H_B\big(\varepsilon, \tilde{\bar{\mathcal{F}}}_{Kv}, \|\cdot\|_{B,\mathbb{P}}\big) \le \frac{A_1 v}{\varepsilon} + d\log\bigg(\frac{A_1}{\varepsilon^2}\bigg)$$

at the cost of increasing the constant A_1. To show the second assertion for the elements of $\tilde{\bar{\mathcal{F}}}_{Kv}$, we can use again the same arguments as for the class $\widetilde{\mathcal{F}}_{Kv}$. Indeed, the condition $K \ge 2K_0 \vee 2$ implies that $\max(|g|, |\bar{g}|) \le K$ since $|\bar{g}| \le 2K_0$. Moreover, with $\tilde{C} = 4MK^2$ we get for any element $\tilde{f} \in \tilde{\bar{\mathcal{F}}}_{Kv}$ that

$$\begin{aligned}
\|\tilde{f}\|^2_{B,\mathbb{P}} &\le \bigg(\frac{a_0}{32M^2} + \frac{1}{16M^2}e^{1/M}\bigg)\int_{\mathcal{X}} \big(g(x) - \bar{g}(x)\big)^2\, d\mathbb{Q}(x) \\
&\le 4\bigg(\frac{a_0}{32M^2} + \frac{1}{16M^2}e^{1/M}\bigg)v^2, \qquad (7.27)
\end{aligned}$$

where in the last line we used the fact that by convexity of $x \mapsto x^2$,

$$\int_{\mathcal{X}} \big(g(x) - \bar{g}(x)\big)^2 \, d\mathbb{Q}(x) \leq 2 \int_{\mathcal{X}} \big(g(x) - g_0(x)\big)^2 \, d\mathbb{Q}(x) + 2 \int_{\mathcal{X}} \big(\bar{g}(x) - g_0(x)\big)^2 \, d\mathbb{Q}(x),$$
$$\leq 4 \int_{\mathcal{X}} \big(g(x) - g_0(x)\big)^2 \, d\mathbb{Q}(x), \tag{7.28}$$

as $(\bar{g} - g_0)^2 \leq (g - g_0)^2$ by definition of $\bar{g}$. This completes the proof of Lemma 7.8.

7.8. Proof of Theorem 7.3

In the sequel, we assume that the assumptions of the theorem hold. Below, we give a uniform bound for the centered process $\mathbb{M}_n - \mathbb{M}$, with $\mathbb{M}_n$ and $\mathbb{M}$ as in (7.1) and (7.4) respectively. In the sequel, the notation $\lesssim$ means "is bounded up to an absolute constant". Moreover, the capital E denotes outer expectation in cases when we consider expectation of a random variable which we have not proved to be measurable.

Proposition 7.9. *Let $K > 2 \vee (2K_0)$. Then, for all $v \in (0, 2K]$, there exists $A > 0$ that depends only on a_0 and M such that*

$$\sqrt{n} E\Big[\sup_{g \in \mathcal{G}_{Kv}} \big|(\mathbb{M}_n - \mathbb{M})g - (\mathbb{M}_n - \mathbb{M})g_0\big|\Big] \leq Ad(1 + \sqrt{q}R)\phi_n(v), \tag{7.29}$$

where $\phi_n(v) = v^{1/2} K^{5/2} (1 + K^{1/2} v^{-3/2} n^{-1/2})$.

Proof. Define for $\eta > 0$ and fixed $v \in (0, 2K]$

$$J(\eta) = \int_0^\eta \sqrt{1 + H_B\big(\varepsilon, \widetilde{\mathcal{F}}_{Kv}, \|\cdot\|_{B,\mathbb{P}}\big)} \, d\varepsilon,$$

where we recall that $\|\cdot\|_{B,\mathbb{P}}$ is the Bernstein norm, and $\widetilde{\mathcal{F}}_{Kv}$ is defined in (7.10) with $\tilde{C} = 4MK^2$. By Lemma 7.8, there exists a constant $A_2 > 0$ depending only on a_0 and M such that $\|\tilde{f}\|_{B,\mathbb{P}} \leq A_2 v$ for all $\tilde{f} \in \widetilde{\mathcal{F}}_{Kv}$. It follows from Lemma 3.4.3 of van der Vaart and Wellner [26] (using the notation of that book) that

$$E\big[\|\mathbb{G}_n\|_{\widetilde{\mathcal{F}}_{Kv}}\big] \lesssim J(A_2 v)\bigg(1 + \frac{J(A_2 v)}{A_2^2 v^2 \sqrt{n}}\bigg),$$

where by Lemma 7.8 and the inequality $\sqrt{u+v} \leq \sqrt{u} + \sqrt{v}$ for $u, v \geq 0$ we have that

$$J(\eta) \leq \int_0^\eta \sqrt{1 + \frac{A_1 d(1 + \sqrt{q}R)}{\varepsilon}} \, d\varepsilon \leq \eta + 2\big(A_1 d(1 + \sqrt{q}R)\big)^{1/2} \eta^{1/2}$$

for all $\eta > 0$. Note that $v \in (0, 2K]$ implies that $v \le v^{1/2}2^{1/2}K^{1/2}$, and hence

$$J(A_2 v) \le \left(A_2 2^{1/2} K^{1/2} + 2\left(A_1 d(1+\sqrt{\overline{q}}R)\right)^{1/2} A_2^{1/2}\right) v^{1/2} \le A_3\left(d(1+\sqrt{\overline{q}}R)\right)^{1/2} K^{1/2} v^{1/2}$$

using that $K \ge 1$, where A_3 depends only on a_0 and M. Hence, by definition of $\widetilde{\mathcal{F}}_{Kv}$, which has the same entropy has the class $\mathcal{F}_{Kv} - f_0 = \{f - f_0, f \in \mathcal{F}_{Kv}\}$, we obtain

$$\begin{aligned}\sqrt{n} E\Big[\sup_{g \in \mathcal{G}_{Kv}} \big|(\mathbb{M}_n - \mathbb{M})(g) - (\mathbb{M}_n - \mathbb{M})(g_0)\big|\Big] &= E\big[\|\mathbb{G}_n\|_{\mathcal{F}_{Kv} - f_0}\big] \\ &= 4MK^2 E\big[\|\mathbb{G}_n\|_{\widetilde{\mathcal{F}}_{Kv} - f_0}\big] \lesssim \phi_n(v)\end{aligned}$$

which completes the proof of Proposition 7.9. □

Now we are ready to give the proof of Theorem 7.3.

Proof of Theorem 7.3. In the sequel, we consider $K = C \log n$ for some $C > 0$ that does not depend on n, and $v \in (0, 2K]$. It follows from Proposition 7.9 that for all n sufficiently large, we have (7.29) where

$$\phi_n(v) = (\log n)^{5/2}\sqrt{v}\left(1 + (\log n)^{1/2} v^{-3/2} n^{-1/2}\right)$$

and A depends only on a_0, M and C. Since with D taken from (7.5), we have $D(g, g_0) \le \|g\|_\infty + \|g_0\|_\infty \le 2K$ for sufficiently large n and all $g \in \mathcal{G}_K$, the above inequality holds for all $v > 0$. Furthermore, $\widehat{g}_n$ maximizes $\mathbb{M}_n g$ over the set of all functions g of the form $g(x) = \Psi(\alpha^T x)$, $x \in \mathcal{X}$ with $\alpha \in \mathcal{S}_{d-1}$ and Ψ a non-decreasing function on $\mathbb{R}$, and it follows from Lemma 7.1 that with arbitrarily large probablity by choice of C, $\widehat{g}_n$ maximizes $\mathbb{M}_n g$ over the restricted set $\mathcal{G}_K$. Hence, we can use Lemma 7.2 and Proposition 7.9 above, together with Theorem 3.2.5 in van der Vaart and Wellner [26] with $\alpha = 1/2$ and $r_n \sim n^{1/3}(\log n)^{-5/3}$, to conclude that $D(\widehat{g}_n, g_0) = O_p^*(n^{-1/3}(\log n)^{5/3})$, which completes the proof of Theorem 7.3. □

7.9. Proof of Theorem 4.1

Assuming that (A1)–(A4) hold, we give a second uniform bound for $\mathbb{M}_n - \mathbb{M}$. The bound is sharper than the one obtained in Proposition 7.9 for the case where all functions g in the considered class of functions satisfy (7.6) for some $v \le (\log n)^2 n^{-1/3}$. As before, the notation $\lesssim$ means "is bounded up to an absolute constant".

Proposition 7.10. *Let $K = C \log n$ for some fixed $C > 0$, $v \in (0, (\log n)^2 n^{-1/3}]$ and $\phi_n(v) = v^{1/2}(1 + v^{-3/2} n^{-1/2})$. Then for n large enough we have that*

$$\sqrt{n} E\Big[\sup_{g \in \mathcal{G}_{Kv}} \big|(\mathbb{M}_n - \mathbb{M})g - (\mathbb{M}_n - \mathbb{M})g_0\big|\Big] \le A\phi_n(v),$$

where A depends only on R, a_0, M, $\overline{q}$, $\underline{q}$ and K_0.

Proof. Assume n large enough so that $K \geq (2K_0) \vee 2$ and use (7.7) to write that the expectation on the left-hand side of the previous display is bounded above by

$$\sqrt{n}E\Big[\sup_{g\in\mathcal{G}_{Kv}}\big|(\mathbb{M}_n-\mathbb{M})g-(\mathbb{M}_n-\mathbb{M})\overline{g}\big|\Big]+\sqrt{n}E\Big[\sup_{g\in\mathcal{G}_{Kv}}\big|(\mathbb{M}_n-\mathbb{M})\overline{g}-(\mathbb{M}_n-\mathbb{M})g_0\big|\Big].$$

Hence, in the notation of van der Vaart and Wellner [26]

$$\sqrt{n}E\Big[\sup_{g\in\mathcal{G}_{Kv}}\big|(\mathbb{M}_n-\mathbb{M})g-(\mathbb{M}_n-\mathbb{M})g_0\big|\Big]\leq E\big[\|\mathbb{G}_n\|_{\mathcal{F}_1}\big]+E\big[\|\mathbb{G}_n\|_{\mathcal{F}_2}\big], \tag{7.30}$$

where $\mathcal{F}_1=\bar{\mathcal{F}}_{Kv}$ and $\mathcal{F}_2$ is the class of functions $\bar{f}-f_0$ such that $\bar{f}\in\mathcal{F}_{(2K_0)v}$. To give a bound for the first term on the right-hand side, consider $v'=(\log n)^2n^{-1/6}\gg\sqrt{v}$. It follows from Lemma 7.8 that for all $\varepsilon\in(0,A_2^{-1}v]$, we have

$$H_B\big(\varepsilon,\tilde{\bar{\mathcal{F}}}_{Kv'},\|\cdot\|_{B,\mathbb{P}}\big)\leq\frac{A_1v'}{\varepsilon}+d\log\left(\frac{A_1}{\varepsilon^2}\right)\leq A_1(1+d)\frac{v'}{\varepsilon} \tag{7.31}$$

provided that n is sufficiently large, where we used the fact that $\log(x)\leq\sqrt{x}$ for all $x>0$ for the second inequality. Since the class $\widetilde{\mathcal{F}}_1:=\tilde{\bar{\mathcal{F}}}_{Kv}$ is included in $\tilde{\bar{\mathcal{F}}}_{Kv'}$, its ε-bracketing entropy can be also bounded above by (7.31) for all $\varepsilon\in(0,A_2^{-1}v]$. Using again the inequality $\sqrt{x+y}\leq\sqrt{x}+\sqrt{y}$ for all $x,y\geq0$, we can write

$$\begin{aligned}J_1(A_2v)&:=\int_0^{A_2v}\sqrt{1+H_B(\varepsilon,\widetilde{\mathcal{F}}_1,\|\cdot\|_{B,\mathbb{P}})}\,d\varepsilon\\&\leq A_2v+2\big(A_1(1+d)v'\big)^{1/2}(A_2v)^{1/2}\leq A_3\big(v'v\big)^{1/2}\end{aligned}$$

using that $v<v'$ and $K>1$, where $A_3>0$ is a constant depending on a_0, M and d. Lemma 7.8 implies that $\|\tilde{f}\|_{B,\mathbb{P}}\leq A_2v$ for all $\tilde{f}\in\widetilde{\mathcal{F}}_1$. Invoking Lemma 3.4.3 of van der Vaart and Wellner [26] allows us to write that

$$E\big[\|\mathbb{G}_n\|_{\widetilde{\mathcal{F}}_1}\big]\lesssim J_1(A_2v)\left(1+\frac{J_1(A_2v)}{A_2^2v^2\sqrt{n}}\right)\leq A_3\big(v'v\big)^{1/2}\left(1+\frac{(v')^{1/2}}{v^{3/2}\sqrt{n}}\right)$$

at the cost of increasing A_3. Now, using the definition of $\widetilde{\mathcal{F}}_1$, we have that

$$E\big[\|\mathbb{G}_n\|_{\mathcal{F}_1}\big]=4MK^2E\big[\|\mathbb{G}_n\|_{\widetilde{\mathcal{F}}_1}\big]\leq A_3v^{1/2}\left(1+\frac{1}{v^{3/2}\sqrt{n}}\right)$$

at the cost of increasing A_3. This gives a bound for the first term on the right-hand side of (7.30).

To deal with the second term, we apply Lemma 7.8 to the class $\widetilde{\mathcal{F}}_2=\widetilde{\mathcal{F}}_{(2K_0)v}$ with $K=2K_0$. Here, $\tilde{C}=4MK_0^2$ is independent of n, and $J_2(A_2v)\leq A_3v^{1/2}$ for some $A_3>0$ that does not depend on n, where J_2 is defined in the same manner as J_1 with $\widetilde{\mathcal{F}}_1$ replaced by $\widetilde{\mathcal{F}}_2$. By Lemma 3.4.3 of van der Vaart and Wellner [26], we have

$$E\big[\|\mathbb{G}_n\|_{\mathcal{F}_2}\big]=16MK_0^2E\big[\|\mathbb{G}_n\|_{\widetilde{\mathcal{F}}_2}\big]\leq A_3v^{1/2}\left(1+\frac{1}{v^{3/2}\sqrt{n}}\right)$$

at the cost of increasing A_3. Combining the calculations developed for both classes together with (7.30) gives the claimed form of the entropy bound. □

Proof of Theorem 4.1. Theorem 7.3 implies that with a probability that can be made arbitrarily large, the LSE $\widehat{g}_n$ belongs to $\mathcal{G}_{Kv}$ with $K = C\log n$ and $v = (\log n)^2 n^{-1/3}$ for some $C > 0$ that does not depend on n. The result follows now from Theorem 3.2.5 of van der Vaart and Wellner [26] with $\alpha = 1/2$ and $r_n \sim n^{1/3}$. □

Acknowledgements

Part of this work was done while the third author was at the University Paris Nanterre as an invited visiting researcher. She thanks Paris Nanterre for the support. The contribution of the second author has been conducted as part of the project Labex MME-DII (ANR11-LBX-0023-01).

Supplementary Material

Supplement to "Least squares estimation in the monotone single index model" (DOI: 10.3150/18-BEJ1090SUPP; .pdf). We provide additional proofs, we give an algorithm to compute the LSE exactly for the special case when $d = 2$, we give properties of exponential families, and we provide additional simulations for Section 5.

References

[1] Balabdaoui, F., Durot, C. and Jankowski, H. (2019). Supplement to "Least squares estimation in the monotone single index model." DOI:10.3150/18-BEJ1090SUPP.

[2] Barlow, R.E., Bartholomew, D.J., Bremner, J.M. and Brunk, H.D. (1972). *Statistical Inference Under Order Restrictions. The Theory and Application of Isotonic Regression*. London–Sydney: Wiley. Wiley Series in Probability and Mathematical Statistics. MR0326887

[3] Brillinger, D.R. (1983). A generalized linear model with "Gaussian" regressor variables. In *A Festschrift for Erich L. Lehmann*. *Wadsworth Statist./Probab. Ser.* 97–114. Belmont, CA: Wadsworth. MR0689741

[4] Chen, Y. and Samworth, R.J. (2016). Generalized additive and index models with shape constraints. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **78** 729–754. MR3534348

[5] Chiou, J.-M. and Müller, H.-G. (2004). Quasi-likelihood regression with multiple indices and smooth link and variance functions. *Scand. J. Stat.* **31** 367–386. MR2087831

[6] Chmielewski, M.A. (1981). Elliptically symmetric distributions: A review and bibliography. *Int. Stat. Rev.* **49** 67–74. MR0623010

[7] Cosslett, S.R. (1983). Distribution-free maximum likelihood estimator of the binary choice model. *Econometrica* **51** 765–782. MR0712369

[8] Cover, T.M. (1967). The number of linearly inducible orderings of points in d-space. *SIAM J. Appl. Math.* **15** 434–439. MR0211892

[9] Dobson, A.J. and Barnett, A.G. (2008). *An Introduction to Generalized Linear Models*, 3rd ed. *Texts in Statistical Science Series*. Boca Raton, FL: CRC Press. MR2459739

[10] Feige, U. and Schechtman, G. (2002). On the optimality of the random hyperplane rounding technique for MAX CUT. *Random Structures Algorithms* **20** 403–440. Probabilistic methods in combinatorial optimization. MR1900615

[11] Foster, J.C., Taylor, J.M.G. and Nan, B. (2013). Variable selection in monotone single-index models via the adaptive LASSO. *Stat. Med.* **32** 3944–3954. MR3102450

[12] Ganti, R., Rao, N., Willett, R.M. and Nowak, R. (2015). Learning single index models in high dimensions. Preprint. Available at arXiv:1506.08910.

[13] Goldstein, L., Minsker, S. and Wei, X. (2018). Structured signal recovery from non-linear and heavy-tailed measurements. *IEEE Trans. Inform. Theory* **64** 5513–5530. MR3832320

[14] Groeneboom, P. and Hendrickx, K. (2018). Current status linear regression. *Ann. Statist.* **46** 1415–1444. MR3819105

[15] Han, A.K. (1987). Nonparametric analysis of a generalized regression model. The maximum rank correlation estimator. *J. Econometrics* **35** 303–316. MR0903188

[16] Härdle, W., Hall, P. and Ichimura, H. (1993). Optimal smoothing in single-index models. *Ann. Statist.* **21** 157–178. MR1212171

[17] Hristache, M., Juditsky, A. and Spokoiny, V. (2001). Direct estimation of the index coefficient in a single-index model. *Ann. Statist.* **29** 595–623. MR1865333

[18] Kakade, S.M., Kanade, V., Shamir, O. and Kalai, A. (2011). Efficient learning of generalized linear and single index models with isotonic regression. In *Advances in Neural Information Processing Systems* 927–935.

[19] Kalai, A. and Sastry, R. (2009). The isotron algorithm: High-dimensional isotonic regression. In *Proceedings of the* 22*nd Annual Conference on Learning Theory* (*COLT*).

[20] Li, K.-C. and Duan, N. (1989). Regression analysis under link violation. *Ann. Statist.* **17** 1009–1052. MR1015136

[21] Li, Q. and Racine, J.S. (2007). *Nonparametric Econometrics*: *Theory and Practice*. Princeton, NJ: Princeton Univ. Press. MR2283034

[22] Lin, W. and Kulasekera, K.B. (2007). Identifiability of single-index models and additive-index models. *Biometrika* **94** 496–501. MR2380574

[23] Murphy, S.A., van der Vaart, A.W. and Wellner, J.A. (1999). Current status regression. *Math. Methods Statist.* **8** 407–425. MR1735473

[24] Plan, Y. and Vershynin, R. (2013). Robust 1-bit compressed sensing and sparse logistic regression: A convex programming approach. *IEEE Trans. Inform. Theory* **59** 482–494. MR3008160

[25] Plan, Y., Vershynin, R. and Yudovina, E. (2017). High-dimensional estimation with geometric constraints. *Inf. Inference* **6** 1–40. MR3636866

[26] van der Vaart, A.W. and Wellner, J.A. (1996). *Weak Convergence and Empirical Processes*: *With Applications to Statistics*. *Springer Series in Statistics*. New York: Springer. MR1385671

Received July 2017 and revised August 2018