Electronic Journal of Statistics
Vol. 9 (2015) 2528–2641
ISSN: 1935-7524
DOI: 10.1214/15-EJS1079

Finite sample behavior of a sieve profile estimator in the single index model

Andreas Andresen[*]

Weierstrass-Institute,
Mohrenstr. 39, 10117 Berlin, Germany
e-mail: andresen@wias-berlin.de

Abstract: We apply the results of Andresen et. al. (2014) on finite sample properties of sieve M-estimators and Andresen et. al. (2015) on the convergence of an alternating maximization procedure to analyse a sieve profile maximization estimator in the single index model with linear index function. The link function is approximated with C^3-Daubechies-wavelets with compact support. We derive results like Wilks phenomenon and Fisher Theorem in a finite sample setup even when the model is miss-specified. Furthermore we show that an alternating maximization procedure converges to the global maximizer and we assess the performance of Friedman's projection pursuit procedure. The approach is based on showing that the conditions of Andresen et. al. (2014) and (2015) can be satisfied under a set of mild regularity and moment conditions on the link function, the regressors and the additive noise. The results allow to construct non-asymptotic confidence sets and to derive asymptotic bounds for the estimator as corollaries.

MSC 2010 subject classifications: Primary 62F10; secondary 62J12, 62F25, 62H12.
Keywords and phrases: Profile estimator, sieve, projection pursuit procedure, alternating maximization, alternating minimization, single index.

Received April 2015.

Contents

[*] The author is supported by Research Units 1735 "Structural Inference in Statistics: Adaptation and Efficiency".

1. Finding the most interesting directions of a data set

Assume observations $(Y_i, \mathbf{X}_i) \in \mathbb{R} \times \mathbb{R}^p$ with $p \in \mathbb{N}$

$$Y_i = g(\mathbf{X}_i) + \varepsilon_i,\ i = 1, \dots, n, \tag{1.1}$$

where $g : \mathbb{R}^p \to \mathbb{R}$ is some continuous function, $\varepsilon_i \in \mathbb{R}$ are additive centered errors independent of the random regressors $(\mathbf{X}_i)$. Consider the task of estimating

$$\mathbb{E}[\boldsymbol{Y}|\mathbf{X}] = g(\mathbf{X}).$$

Statistical theory for nonparametric models shows that even for moderate $p \in \mathbb{N}$ the accuracy of estimating $g(\mathbf{X})$ increases very slow in the sample size $n \in \mathbb{N}$ as the rates are lower bounded by $n^{-\alpha/(2\alpha+p)}$ – with $\alpha > 0$ quantifying the smoothness of $g : \mathbb{R}^p \to \mathbb{R}$ – as was for instance noted in [20]. [7] propose to use a projection pursuit approach to circumvent this problem in situations where

$$g(\mathbf{X}) \approx \sum_{l=1}^{M} f_{(l)}(\mathbf{X}^\top \boldsymbol{\theta}^*_{(l)}), \tag{1.2}$$

for a set of functions $f_{(l)} : \mathbb{R} \to \mathbb{R}$, vectors $\boldsymbol{\theta}^*_{(l)} \in S_1^{p,+} := \{\boldsymbol{\theta} \in \mathbb{R}^p : \|\boldsymbol{\theta}\| = 1,\, \theta_1 > 0\} \subset \mathbb{R}^p$ and some $M \in \mathbb{N}$. As each nonparametric estimation task is uni-variate, better performance can be expected in comparison to a full nonparametric regression as long as $M, p \in \mathbb{N}$ are not very large. But of course (1.2) is a structural assumption whose usefulness depends on the size of $M \in \mathbb{N}$ and $p \in \mathbb{N}$. For small $M \in \mathbb{N}$ and $p \in \mathbb{N}$ one can get important gains but the assumption (1.2) becomes rather restrictive. On the other hand, for large $M \in \mathbb{N}$ and large $p \in \mathbb{N}$ the assumption (1.2) becomes true for any smooth function. This can be seen as follows. Assume that one observes $(Y_i, \boldsymbol{Z}_i)$ for a given vector of regressors $\boldsymbol{Z} \in \mathbb{R}^{p_1}$ and that the aim is to estimate $g^\circ(\boldsymbol{Z}) = \mathbb{E}[Y|\boldsymbol{Z}]$. We can define for some $D \in \mathbb{N}$ an extended vector of regressors $\mathbf{X} \in \mathbb{R}^{p_1 + \sum_{d=2}^{D+1} p_1^d - p_1}$ via

$$\mathbf{X} \overset{\text{def}}{=} (Z_1, \ldots, Z_{p_1}, Z_1 Z_2, Z_1 Z_3, \ldots, Z_{p_1 - 1} Z_{p_1}, Z_1 Z_1 Z_2, \ldots, Z_{p_1 - 1} Z_{p_1}^D).$$

For large $D \in \mathbb{N}$ this means that (1.2) demands that $g^\circ(\boldsymbol{Z}) = g(\mathbf{X})$ can be well approximated by polynomials of maximal degree $D + 1 \in \mathbb{N}$, which of course is the case for smooth functions. See [11] and [13] for a more sophisticated approach of showing that smooth functions g can be well approximated as in (1.2). [7] suggest to estimate the pairs $(f_{(l)}, \boldsymbol{\theta}^*_{(l)})$ iteratively. The first task is to estimate

$$\boldsymbol{\theta}^*_{(1)} \overset{\text{def}}{=} \operatorname*{argmin}_{\boldsymbol{\theta} \in S_1^{p,+}} \mathbb{E}\left[\left(g(\mathbf{X}) - \mathbb{E}[g(\mathbf{X})|\mathbf{X}^\top \boldsymbol{\theta}]\right)^2\right]. \tag{1.3}$$

Given an estimator $\widetilde{\boldsymbol{\theta}}_{(1)} \in S_1^{p,+}$ one can determine an estimator $\widehat{f}_{(1)}$ for $f_{(1)}$ and generate a new sample via

$$Y_{i(1)} \overset{\text{def}}{=} Y_i - \widehat{f}_{(1)}(\mathbf{X}_i^\top \widetilde{\boldsymbol{\theta}}_{(1)}).$$

Using this new data set $(Y_{i(1)})_{i=1,\ldots,n}$ one can estimate $\boldsymbol{\theta}^*_{(2)}$ and $f_{(2)}$ as in the first step and again generate a new data set $(Y_{i(2)})_{i=1,\ldots,n}$. These steps are repeated $M - 1 \in \mathbb{N}$ times if $M \in \mathbb{N}$ was fixed or known in the beginning, otherwise until a certain level of variability in the data is explained by the obtained sum

$$\sum_{l=1}^{M} \widehat{f}_{(l)}(\mathbf{X}_i^\top \widetilde{\boldsymbol{\theta}}_{(l)}),$$

which then serves as an estimator for $\mathbb{E}[Y|\mathbf{X}]$. This way of estimating the conditional expectation is called Projection Pursuit Procedure (PPP), cf. [7].

In this work we will mainly focus on the task (1.3). It has been observed in [9] that the estimation of $\boldsymbol{\theta}^*_{(1)}$ – from now on denoted simply by $\boldsymbol{\theta}^*$ – can be attained with root-n rate even though the full model is nonparametric.

In the particular case that $M = 1$, i.e. that

$$g(\mathbf{X}) = f(\mathbf{X}^\top \boldsymbol{\theta}^*), \tag{1.4}$$

for some $f : \mathbb{R} \to \mathbb{R}$ and $\boldsymbol{\theta}^* \in S_1^{p,+} \subset \mathbb{R}^p$, the estimation problem (1.3) becomes the task to estimate the linear response vector in a semiparametric single-index model (see [12]). The single-index model supposes that the observations satisfy with two functions $f : \mathbb{R} \to \mathbb{R}$ and $h : \mathbb{R}^p \to \mathbb{R}$ and with errors $(\varepsilon_i) \in \mathbb{R}$

$$Y_i = f(h(\mathbf{X}_i)) + \varepsilon_i, \; i = 1, \ldots, n.$$

Usually it is assumed that the index function h is known up to some parameter $\boldsymbol{\theta} \in \mathbb{R}^p$ such that one writes $h(\boldsymbol{\theta}, \boldsymbol{x})$. In our setting $h(\boldsymbol{\theta}, \boldsymbol{x}) = \boldsymbol{\theta}^\top \boldsymbol{x}$. [23] compares the asymptotic distributions of two different prominent estimation procedures for $\boldsymbol{\theta}^*$. The first is the average derivative estimation introduced by [17] and refined by [10] and is based on the fact that if (1.4) is correct

$$\mathbb{E}\left[\frac{d}{d\mathbf{X}} g(\mathbf{X})\right] = \mathbb{E}\left[f'(\mathbf{X}^\top \boldsymbol{\theta}^*)\right] \boldsymbol{\theta}^*,$$

which suggests to estimate $\boldsymbol{\theta}^*$ via an estimate of $\mathbb{E}\left[f'(\mathbf{X}^\top \boldsymbol{\theta}^*)\right]$. The second one is the minimal conditional variance estimation by [24] which is inspired by [8] and aims at directly solving (1.3) via a local linear approximation of $\mathbb{E}[y|\mathbf{X}^\top \boldsymbol{\theta}]$. Further results are the asymptotic efficiency of a semiparametric maximum-likelihood estimator shown by [5] for particular examples and in [8] the right choice of the bandwidth for the nonparametric estimation of the link function.

In this work we want to use a different approach to carry out the first step (1.3) that allows to apply the results of [2] and [3]. For this purpose denote

$$\mathbb{E}[g(\mathbf{X})|\mathbf{X}^\top \boldsymbol{\theta}^*] = f(\mathbf{X}^\top \boldsymbol{\theta}^*). \tag{1.5}$$

Assume that $f|_{[-s_\mathbf{X}, s_\mathbf{X}]} \in L^2([-s_\mathbf{X}, s_\mathbf{X}]) = \overline{\operatorname{span}}\{(\boldsymbol{e}_k)_{k\in\mathbb{N}}\}$ for the set of Daubechies wavelet basis functions $(\boldsymbol{e}_k)_{k\in\mathbb{N}} \subset L^2([-s_\mathbf{X}, s_\mathbf{X}])$ that we present in Section 3.1. For some $m \geq 1$ and $\boldsymbol{\eta} \in \mathbb{R}^m$ denote

$$\boldsymbol{f}_{\boldsymbol{\eta}} \overset{\text{def}}{=} \sum_{k=0}^{m} \eta_k \boldsymbol{e}_k,$$

with properly selected coefficients $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_m)^\top \in \mathbb{R}^m$. Further assume that $\mathbb{P}(\mathbf{X}_1 \in B_{s_\mathbf{X}}(0)) \approx 1$ for some $s_\mathbf{X} > 0$, where $B_\mathtt{r}(\boldsymbol{z})$ denotes the euclidean ball of radius $\mathtt{r} > 0$ around $\boldsymbol{z}$. Set $\boldsymbol{\upsilon} \overset{\text{def}}{=} (\boldsymbol{\theta}, \boldsymbol{\eta}) \in \mathbb{R}^{p+m}$, define $p^* = p + m$ and $\Pi_{\boldsymbol{\theta}}, \Pi_{\boldsymbol{\eta}}$ as the orthogonal projections on the $\boldsymbol{\theta}$- or $\boldsymbol{\eta}$-component of $\boldsymbol{\upsilon}$ respectively. We assume that p is fixed and can be treated as a constant. This means that $p^* = p + m = O(m)$. Our aim is to analyze for $m \in \mathbb{N}$ the properties of the estimator

$$\widetilde{\boldsymbol{\theta}}_m \overset{\text{def}}{=} \Pi_{\boldsymbol{\theta}} \widetilde{\boldsymbol{\upsilon}}_m \overset{\text{def}}{=} \Pi_{\boldsymbol{\theta}} \underset{(\boldsymbol{\theta},\boldsymbol{\eta})\in\Upsilon_m}{\operatorname{argmax}} \mathcal{L}_m(\boldsymbol{\theta}, \boldsymbol{\eta}), \tag{1.6}$$

where

$$\begin{aligned} \mathcal{L}_m(\boldsymbol{\theta}, \boldsymbol{\eta}) &= \sum_{\{i\colon \|\mathbf{X}_i\| \leq s_\mathbf{X}\}} \ell_{i,m}(\boldsymbol{\theta}, \boldsymbol{\eta}) \qquad (1.7)\\ &= -\sum_{\{i\colon \|\mathbf{X}_i\| \leq s_\mathbf{X}\}} \|Y_i - \boldsymbol{f}_{\boldsymbol{\eta}}(\mathbf{X}_i^\top \boldsymbol{\theta})\|^2/2. \end{aligned}$$

The set $\varUpsilon_m$ is defined as

$$\varUpsilon_m \overset{\text{def}}{=} \{(\boldsymbol{\theta},\boldsymbol{\eta}) \subset S_1^{p,+} \times \mathbb{R}^m,\ \|\boldsymbol{\eta}\| \le \mathtt{r}^\circ\},$$

with some $\mathtt{r}^\circ < \infty$ defined in (5.2). Note that this is exactly the type of estimator presented in Section 2.7 of [2]. In [12] a very similar estimator is analyzed based on a "leave one out" kernel estimation of $\mathbb{E}[Y_i|\mathbf{X}_i^\top\boldsymbol{\theta}]$ instead of using $\boldsymbol{f}_{\boldsymbol{\eta}}(\mathbf{X}_i^\top\boldsymbol{\theta})$. Ichimura shows $\sqrt{n}$-consistency and asymptotic normality of his proposed estimator.

Remark 1.1. The radius $\mathtt{r}^\circ$ is needed to control the large deviations of the full maximizer $\widetilde{\boldsymbol{\upsilon}}_m$. We ensure that the estimator $\widetilde{\boldsymbol{\upsilon}}_m$ does not lie on the boundary in Lemma 5.2.

Remark 1.2. Our approach relies on an invertible operator $\nabla^2\mathbb{E}\mathcal{L}(\boldsymbol{\upsilon}^*)$. One way to achieve this it to ensure that the density of $\mathbf{X}^\top\boldsymbol{\theta}$ is strictly greater zero on the whole interval $[-s_{\mathbf{X}}, s_{\mathbf{X}}]$ for any $\boldsymbol{\theta} \in S_+^1$. To avoid undesirable boundary effects – i.e. that the density vanishes near $s_{\mathbf{X}}, -s_{\mathbf{X}}$ (see Remark A.5) – we do not use all available data: We only consider realizations $(Y_i, \mathbf{X}_i)$ for which $\|\mathbf{X}_i\| \le s_{\mathbf{X}}$ but in Section 2.1 we assume in condition $(\mathbf{Cond}_{\mathbf{X}})$ that there is positive probability that $\mathbf{X} \in B_{s_{\mathbf{X}}+c_B}(0)\backslash B_{s_{\mathbf{X}}}(0)$ for some $c_B > 0$. We assume that the proportion of ignored data is small such that we can neglect this in the following and pretend that we can use the full data set.

For an appropriate sequence $m(n) \to \infty$ the estimator $\widetilde{\boldsymbol{\theta}}_m$ in (1.6) is supposed to approach

$$\boldsymbol{\theta}^* \overset{\text{def}}{=} \varPi_{\boldsymbol{\theta}}(\boldsymbol{\theta}^*,\boldsymbol{\eta}^*) \overset{\text{def}}{=} \varPi_{\boldsymbol{\theta}}\boldsymbol{\upsilon}^* \overset{\text{def}}{=} \varPi_{\boldsymbol{\theta}} \operatorname*{argmax}_{(\boldsymbol{\theta},\boldsymbol{\eta})\in\varUpsilon} \mathbb{E}\mathcal{L}(\boldsymbol{\theta},\boldsymbol{\eta}), \tag{1.8}$$

where $\varUpsilon = S_1^{p,+} \times l^2$ and for $(\boldsymbol{\theta},\boldsymbol{\eta}) \in \varUpsilon$

$$\mathcal{L}(\boldsymbol{\theta},\boldsymbol{\eta}) \overset{\text{def}}{=} -\sum_{\{i:\,\|\mathbf{X}_i\|\le s_{\mathbf{X}}\}} \left\| Y_i - \sum_{k=1}^{\infty} \eta_k \boldsymbol{e}_k(\mathbf{X}_i^\top\boldsymbol{\theta}) \right\|^2 /2.$$

Remark 1.3. To understand the motivation of this functional note that for any $\boldsymbol{\theta} \in S_1^{p,+}$ the sequence

$$\boldsymbol{\eta}_{\boldsymbol{\theta}}^* \overset{\text{def}}{=} \varPi_{\boldsymbol{\eta}} \operatorname*{argmax}_{\substack{\boldsymbol{\upsilon}\in\mathbb{R}^p\times l^2 \\ \varPi_{\boldsymbol{\theta}}\boldsymbol{\upsilon}=\boldsymbol{\theta}}} \mathbb{E}\mathcal{L}(\boldsymbol{\upsilon}),$$

solves by first order criteria of maximality for any $A \in \mathcal{F}(\mathbf{X}^\top\boldsymbol{\theta})$ – where $\mathcal{F}(\mathbf{X}^\top\boldsymbol{\theta})$ denotes the sigma algebra associated to the law of $\mathbf{X}^\top\boldsymbol{\theta}$ – the equation

$$\mathbb{E}\left[\left(g(\mathbf{X}) - \boldsymbol{f}_{\boldsymbol{\eta}_{\boldsymbol{\theta}}^*}(\mathbf{X}^\top\boldsymbol{\theta})\right) 1_A\right] = 0.$$

This means that with equivalence in $L^2(\mathbb{P}^{\mathbf{X}})$

$$\boldsymbol{f}_{\boldsymbol{\eta}^*_{\boldsymbol{\theta}}}(\mathbf{X}^\top\boldsymbol{\theta}) = \mathbb{E}[g(\mathbf{X})|\mathbf{X}^\top\boldsymbol{\theta}], \tag{1.9}$$

such that the target (1.8) indeed coincides with the most informative direction in (1.3).

Remark 1.4. Note that there is a model bias and an approximation bias of the form

$$\begin{aligned} \text{“}\textit{model bias}\text{”} &= \min_{\boldsymbol{\upsilon}\in\varUpsilon} \mathbb{E}\|g(\mathbf{X}) - \boldsymbol{f}_{\boldsymbol{\eta}}(\mathbf{X}^\top\boldsymbol{\theta})\|^2, \\ \text{“}\textit{approximation bias}\text{”} &= \min_{\boldsymbol{\upsilon}\in\varUpsilon_m} \mathbb{E}\|\boldsymbol{f}_{\boldsymbol{\eta}}(\mathbf{X}^\top\boldsymbol{\theta}) - \boldsymbol{f}_{\boldsymbol{\eta}^*}(\mathbf{X}^\top\boldsymbol{\theta}^*)\|^2, \end{aligned} \tag{1.10}$$

which both have to be accounted for.

As pointed out we will analyze the properties of the estimator $\widetilde{\boldsymbol{\theta}}_m$ in (1.6) using the results of [2] and [3]. It turns out that this is possible with a series of conditions on the additive noise $\varepsilon_i \in \mathbb{R}$, the function $g : \mathbb{R}^p \to \mathbb{R}$ and on the random design $\mathbf{X} \in \mathbb{R}^p$. In particular the choice of the basis is independent of the model. Due to the support structure of compactly supported wavelets – see Section 3.1 – we still manage to control the sieve bias in (1.10). Even though we assume what is necessary to apply the results of [2] and [3], the calculations necessary to check the conditions still remain rather tedious and lengthy. We present most steps in full detail, which at some points leads to repetitions of very similar arguments. Also the regression setup leads to some peculiarities that we elaborate on in Section 3.2. The treatment of these issues involves bounds for the spectral norm or random matrices from [22]. It is worthy to point out here that a fixed design setting would not resolve these issues either as one for instance would still have to deal with convergence issues of the operator

$$\sum_{i=1}^n \nabla\mathcal{L}(\mathbf{X}_i, Y_i, \boldsymbol{\upsilon})\nabla\mathcal{L}(\mathbf{X}_i, Y_i, \boldsymbol{\upsilon})^\top \in \mathbb{R}^{p^*\times p^*}.$$

There is another peculiarity to the results we present in this work. A naive approach to satisfy the important condition $(\mathcal{L}\mathtt{r})$ from Section 4.1 would include a bound for

$$\sup_{\boldsymbol{\upsilon}\in\varUpsilon_m} |\mathbb{E}[\mathcal{L}(\boldsymbol{\upsilon},\boldsymbol{\upsilon}^*)|(\mathbf{X}_i)_{i=1,\ldots,n}] - \mathbb{E}\mathcal{L}(\boldsymbol{\upsilon},\boldsymbol{\upsilon}^*)|\,. \tag{1.11}$$

But as $\mathcal{L}$ is quadratic and $\varUpsilon_m \subset \mathbb{R}^{p^*}$ can be quite large this becomes hard to achieve with nice bounds. We circumvent this problem using an idea of [15]. Mendelson's crucial insight is that to obtain

$$\inf_{\boldsymbol{\upsilon}\in\{\|\boldsymbol{\upsilon}-\boldsymbol{\upsilon}^*\|_\circ>\mathtt{r}\}} \mathbb{E}[\mathcal{L}(\boldsymbol{\upsilon},\boldsymbol{\upsilon}^*)|(\mathbf{X}_i)_{i=1,\ldots,n}] \geq \mathtt{br}^2,$$

with some norm $\|\cdot\|_\circ$, one only has to ensure that

$$\inf_{\boldsymbol{\upsilon}\in\{\|\boldsymbol{\upsilon}-\boldsymbol{\upsilon}^*\|_\circ>\mathtt{r}\}} \mathbb{P}\left(\|Y_i-\boldsymbol{f}_{\boldsymbol{\eta}^*}(\mathbf{X}_i^\top\boldsymbol{\theta}^*)\|^2-\|Y_i-\boldsymbol{f}_{\boldsymbol{\eta}}(\mathbf{X}_i^\top\boldsymbol{\theta})\|^2\geq \mathtt{br}^2/n\right)>0,$$

We follow this route in the proof of Lemma 5.3. But we only apply this idea in the case that $\mathtt{C}_{bias}=0$. In the general case we derive a bound for (1.11) to avoid too lengthy derivations. The price is an additional $\log(n)$-factor in the sufficient full dimension i.e. we need ${p^*}^3\log(n)=o(\sqrt{n})$ instead of ${p^*}^3=o(\sqrt{n})$ to apply Theorem 4.3.

1.1. Finite sample Wilks and Fisher Theorems

Before we present our main results we want to explain what type of results we aim at and how they can be interpreted. Hopefully this will ease the understanding and will make some of the apparently cumbersome notation more intelligible.

Usually in asymptotic treatments of semiparametric M-estimators like $\widetilde{\boldsymbol{\theta}}_m$ in (1.6) the aim is to derive statements of the kind

$$\begin{aligned}
\sqrt{n}(\widetilde{\boldsymbol{\theta}}_m-\boldsymbol{\theta}^*)-\breve{\boldsymbol{\xi}}_m &= o_{\mathbb{P}}(1), && (1.12)\\
\breve{L}_m(\widetilde{\boldsymbol{\theta}}_m,\boldsymbol{\theta}^*)-\|\breve{d}_m\breve{\boldsymbol{\xi}}_m\|^2 &= o_{\mathbb{P}}(1), && (1.13)\\
\breve{\boldsymbol{\xi}}_m &\xrightarrow{w} \mathcal{N}(0,\breve{d}^{-2}\breve{v}^2\breve{d}^{-2}),
\end{aligned}$$

where we use the shorthand notation

$$\breve{L}_m(\boldsymbol{\theta},\boldsymbol{\theta}^\circ)\stackrel{\text{def}}{=}\max_{\boldsymbol{\eta}}\mathcal{L}_m(\boldsymbol{\theta},\boldsymbol{\eta})-\max_{\boldsymbol{\eta}}\mathcal{L}_m(\boldsymbol{\theta}^\circ,\boldsymbol{\eta}).$$

The random variable $\breve{\boldsymbol{\xi}}_m\in\mathbb{R}^p$ is called semiparametric score. Below we will briefly explain its derivation along with the explanation of the matrices $\breve{v}^2,\breve{d}^2$, $\breve{d}_m^2\in\mathbb{R}^{p\times p}$. But before, we sketch how (1.12) and (1.13) can be used for the construction of asymptotic confidence sets that yield statistical tests. Given the matrices $\breve{v}^2,\breve{d}^2$ the construction works as follows. Let $q_\alpha^2>0$ be an α-level quantile of a $\chi_p^2(\breve{d}^{-2}\breve{v}^2\breve{d}^{-2})$-distribution. Set

$$\mathcal{E}(q_\alpha)=\left\{\boldsymbol{\theta}:\sqrt{n}\|(\widetilde{\boldsymbol{\theta}}_m-\boldsymbol{\theta})\|\leq q_\alpha\right\}; \qquad (1.14)$$

then one can use (1.12) to show

$$\mathbb{P}\left\{\boldsymbol{\theta}^*\notin\mathcal{E}\left(q_\alpha\right)\right\}=\mathbb{P}\left\{\sqrt{n}\|(\widetilde{\boldsymbol{\theta}}_m-\boldsymbol{\theta}^*)\|\geq q_\alpha\right\}\to 1-\alpha.$$

Similarly one can exploit (1.13).

Now we explain the definition of $\breve{v}^2$, $\breve{d}^2$ and $\breve{d}^2_m$. Remember that $\mathcal{L}(\boldsymbol{\upsilon}) = \sum_i \ell_i(\boldsymbol{\upsilon})$ by the definition in (1.7). Consider the Fréchet-derivatives $\nabla\ell(\boldsymbol{\upsilon}^*) \in \mathbb{R}^p \times l^2$ and

$$-\nabla^2\mathbb{E}\mathcal{L}(\boldsymbol{\upsilon}) = -n\nabla^2\mathbb{E}\ell(\boldsymbol{\upsilon}) : \mathbb{R}^p \times l^2 \to \mathbb{R}^p \times l^2.$$

Then $\breve{v}^2 = \breve{v}^2(\boldsymbol{\upsilon}^*)$ and $\breve{d}^2 = \breve{d}^2(\boldsymbol{\upsilon}^*)$ where

$$\begin{aligned} \breve{v}^{-2}(\boldsymbol{\upsilon}) &\overset{\text{def}}{=} \varPi_{\boldsymbol{\theta}} \operatorname{Cov}(\nabla\ell(\boldsymbol{\upsilon}))^{-1} \varPi_{\boldsymbol{\theta}}^{\top}, \\ \breve{d}^{-2}(\boldsymbol{\upsilon}) &\overset{\text{def}}{=} -\varPi_{\boldsymbol{\theta}} \left(\nabla^2\mathbb{E}\ell(\boldsymbol{\upsilon})\right)^{-1} \varPi_{\boldsymbol{\theta}}^{\top}, \end{aligned} \tag{1.15}$$

where $\varPi_{\boldsymbol{\theta}}$ is the orthogonal projection onto the $\boldsymbol{\theta}$-component in $\mathbb{R}^p$ and $\varPi_{\boldsymbol{\theta}}^{\top}$ is its dual operator.

Remark 1.5. Note that these two matrices coincide if the functional $\mathcal{L}$ was the complete loglikelihood of the observations and that then $\breve{d}^2$ would equal the covariance of the efficient influence function (see [14] for more details).

For the definition of the semiparametric score $\breve{\boldsymbol{\xi}}_m \in \mathbb{R}^p$ and of $\breve{d}_m^{-2}(\boldsymbol{\upsilon})$ consider

$$d_m^2(\boldsymbol{\upsilon}) = \begin{pmatrix} d^2(\boldsymbol{\upsilon}) & a_m(\boldsymbol{\upsilon}) \\ a_m^{\top}(\boldsymbol{\upsilon}) & h_m^2(\boldsymbol{\upsilon}) \end{pmatrix} \overset{\text{def}}{=} -\varPi_m \nabla^2\mathbb{E}\ell(\boldsymbol{\upsilon})\varPi_m^{\top} \in \mathbb{R}^{p^*\times p^*}, \tag{1.16}$$

$$\breve{d}_m^{-2}(\boldsymbol{\upsilon}) \overset{\text{def}}{=} -\varPi_{\boldsymbol{\theta}} \left(d_m(\boldsymbol{\upsilon})\right)^{-2} \varPi_{\boldsymbol{\theta}}^{\top}. \tag{1.17}$$

where $\varPi_m : \mathbb{R}^p \times l^2 \to \mathbb{R}^p \times \mathbb{R}^m = \mathbb{R}^{p^*}$ is the canonical projection from l^2 onto $\mathbb{R}^{p^*}$. Further consider the possibly biased target $\boldsymbol{\upsilon}_m^* \in \mathbb{R}^{p^*}$

$$\boldsymbol{\upsilon}_m^* = (\boldsymbol{\theta}_m^*, \boldsymbol{\eta}_m^*) = \underset{\boldsymbol{\upsilon}\in\varUpsilon_m^*}{\operatorname{argmax}}\, \mathbb{E}[\mathcal{L}_m(\boldsymbol{\upsilon})], \tag{1.18}$$

where

$$\varUpsilon_m^* \overset{\text{def}}{=} \{(\boldsymbol{\theta},\boldsymbol{\eta}) \subset S_1^{p,+} \times \mathbb{R}^m\}.$$

Then

$$\begin{aligned} \breve{\boldsymbol{\xi}}_m &\overset{\text{def}}{=} \frac{1}{\sqrt{n}}(1-\mathbb{E}_\epsilon)\varPi_p d_m^{-2}(\boldsymbol{\upsilon}_m^*)\nabla\mathcal{L}_m(\boldsymbol{\upsilon}_m^*) \\ &= \frac{1}{\sqrt{n}}(1-\mathbb{E}_\epsilon)\breve{d}_m^{-2}(\boldsymbol{\upsilon}_m^*)\left\{\nabla_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\upsilon}_m^*) - a_m h_m^{-2}(\boldsymbol{\upsilon}_m^*)\varPi_m\nabla_{\boldsymbol{\eta}}\mathcal{L}(\boldsymbol{\upsilon}_m^*)\right\}, \end{aligned} \tag{1.19}$$

where $\mathbb{E}_\varepsilon$ denotes the expectation operator of the law of $(\varepsilon_i)_{i=1,\dots,n}$ given $(\mathbf{X}_i)_{i=1,\dots,n}$. This random variable is related to the efficient influence function in semiparametric estimation and it plays the role that the usual score $\nabla\mathcal{L}(\boldsymbol{\upsilon}^*)$ plays in the setting of parametric M-estimation.

In Section A.2 we calculate $\nabla\mathcal{L}(\boldsymbol{\upsilon})$, $\mathbb{E}[\nabla\mathcal{L}(\boldsymbol{\upsilon}^*)\nabla\mathcal{L}(\boldsymbol{\upsilon}^*)^\top]$ and $\nabla^2\mathbb{E}\mathcal{L}(\boldsymbol{\upsilon}^*)$, which could be used for a plug-in approach to estimate $\breve{d}^{-2}\breve{v}^2\breve{d}^{-2}$. But we do not present closed form expressions for $\breve{\boldsymbol{\xi}}_m$, $\breve{d}^2$, $\breve{v}^2$ or $\breve{d}_m^2$. One reason – besides the fact that it would not be clear how to do that – is that these objects depend on the true parameter $\boldsymbol{\upsilon}^*$ in (1.8). In fact to make use of such results as in (1.12) and (1.13) in practice one needs to somehow assess the distribution of $\breve{\boldsymbol{\xi}}_m$, which could be done via an estimation of $\breve{d}^{-2}\breve{v}^2\breve{d}^{-2}$ and a Gaussian approximation or via some bootstrap scheme. On this level the statements are merely a theoretical justification of such inference procedures.

In this work we restrict ourselves to derive finite sample bounds for the terms on the right-hand sides of (1.12) and (1.13). To be more precise we derive statements of the following kind. With probability greater than $1-\mathtt{C}e^{-\mathtt{x}}$

$$\left\|\sqrt{n}(\widetilde{\boldsymbol{\theta}}_m-\boldsymbol{\theta}^*)-\breve{\boldsymbol{\xi}}_m\right\| \le \tau(\mathtt{x},p^*,n), \tag{1.20}$$

$$\left|\breve{L}_m(\widetilde{\boldsymbol{\theta}}_m,\boldsymbol{\theta}^*)-\|\breve{d}_m\breve{\boldsymbol{\xi}}_m\|^2\right| = \tau(\mathtt{x},p^*,n), \tag{1.21}$$

with some small value $\tau(\mathtt{x},p^*)>0$. Using the scheme in (1.14) the bounds (1.20) and (1.21) allow the construction of (conservative) finite sample "confidence sets". Assume that (approximate) quantiles q_α for $\|\breve{\boldsymbol{\xi}}\|$ are available, i.e. that with some small $\epsilon>0$ and any $\alpha\in[0,1]$

$$\mathbb{P}(\|\breve{\boldsymbol{\xi}}_m\|\le q_\alpha)\in(\alpha-\epsilon,\alpha+\epsilon),$$

then with some generic constant $\mathtt{C}>0$ (see Remark 2.13 of [2])

$$\alpha+\epsilon+\mathtt{C}e^{-\mathtt{x}} \le \mathbb{P}\left\{\boldsymbol{\theta}^*\in\mathcal{E}(q_\alpha+\tau(\mathtt{x},p^*)\right\},$$
$$\mathbb{P}\left\{\boldsymbol{\theta}^*\in\mathcal{E}(q_\alpha-\tau(\mathtt{x},p^*)\right\} \le \alpha-\epsilon-\mathtt{C}e^{-\mathtt{x}}.$$

The important achievement is that one can now make approximate confidence statements even in the finite sample case, without ignoring "hopefully small enough" terms. As remarked above such approximate quantiles could be attained via an plug-in-estimation of $\breve{d}^{-2}\breve{v}^2\breve{d}^{-2}$ combined with a Gaussian approximation or a bootstrap. Those steps are beyond the scope of this work, in which we merely serve the first step for such an analysis namely the bounds (1.20) and (1.21), which allow to correct for "non-quadratacity" of the functional $\mathcal{L}$.

To derive such a bound $\tau(\mathtt{x},p^*,n)>0$, in [2] the problem is split into two parts. The first part is to derive a bound $\tau_s(\mathtt{x})$ such that with probability greater than $1-\mathtt{C}e^{-\mathtt{x}}$

$$\left\|\sqrt{n}(\widetilde{\boldsymbol{\theta}}_m-\boldsymbol{\theta}_m^*)-\breve{\boldsymbol{\xi}}_m\right\| \le \tau_s(\mathtt{x},p^*,n),$$
$$\left|\breve{L}_m(\widetilde{\boldsymbol{\theta}}_m,\boldsymbol{\theta}_m^*)-\|\breve{d}\breve{\boldsymbol{\xi}}_m\|^2\right| \le \tau_s(\mathtt{x},p^*,n).$$

The corresponding result in this paper is Proposition 2.1. The second part consists in bounding with some $\alpha(m) > 0$

$$\left\|\sqrt{n}(\boldsymbol{\theta}^* - \boldsymbol{\theta}^*_m)\right\| \leq \alpha(m), \quad \left|\breve{L}_m(\boldsymbol{\theta}^*, \boldsymbol{\theta}^*_m)\right| \leq \alpha(m),$$

which yields $\tau(\mathtt{x}, p^*, n) \leq \tau_s(\mathtt{x}, p^*, n) + \alpha(m)$ in Proposition 2.2.

Remark 1.6. We will see that $(\boldsymbol{\upsilon}^*_m, 0) \in l^2$ lies close to the true point $\boldsymbol{\upsilon}^* \in l^2$ but we will not proof that it is unique. We neither proof nor use uniqueness of the profile ME either. In the following we will denote by $\boldsymbol{\upsilon}^*_m$ the set of maximizers and we will always make statements about $\widetilde{\boldsymbol{\theta}}_m \in \mathbb{R}^p$, whereby we mean any element of the set of maximizers of the profiled functional. Non-uniqueness is not a problem, as the concentration on the local set $\varUpsilon_\circ$ is ensured via Theorem 4.2.

Remark 1.7. Note that we maximize over different sets when defining $\widetilde{\boldsymbol{\upsilon}}_m$ and $\boldsymbol{\upsilon}^*_m$. To control the large deviations and avoid boundary effects we have to ensure that with overwhelming probability $\widetilde{\boldsymbol{\upsilon}}_m \subset \operatorname{int}\{\varUpsilon_m\} \subset \varUpsilon^*_m$. We do this with Lemma 5.2, which tells us that we may set $\mathtt{r}^\circ \leq \mathtt{C}\sqrt{m}$ with some constant $\mathtt{C} \in \mathbb{R}$. This lemma also ensures that the alternating sequence $(\widetilde{\boldsymbol{\theta}}^{(k)}, \widetilde{\boldsymbol{\eta}}^{(k(+1))})_{k\in\mathbb{N}}$ from Section 2.3 lies in $S_1^{p,+} \times B^m_{\mathtt{r}^\circ}(0)$.

2. Main results

2.1. Assumptions

To apply the technique presented in [2] and [3] we need a list of assumptions. We denote this list by $(\mathcal{A})$. We start with conditions on the regressors $\mathbf{X} \in \mathbb{R}^p$:

$(\mathbf{Cond_X})$

- The random variables $(\mathbf{X}_i)_{i=1,\dots,n} \subset \mathbb{R}^p$ are i.i.d with distribution denoted by $\mathbb{P}^{\mathbf{X}}$ and independent of $(\varepsilon_i)_{i=1,\dots,n} \subset \mathbb{R}$.
- The measure $\mathbb{P}^{\mathbf{X}}$ is absolutely continuous with respect to the Lebesgue measure. The Lebesgue density $p_{\mathbf{X}}$ of $\mathbb{P}^{\mathbf{X}}$ is Lipschitz continuous on $B_{s_{\mathbf{X}}}(0) \subset \mathbb{R}^p$ and satisfies $p_{\mathbf{X}} > 0$ on $B_{s_{\mathbf{X}}+c_B}(0)$ for some $c_B > 0$.
- For any pair $\boldsymbol{\theta} \in S_1^{+,p}$ with $\boldsymbol{\theta} \perp \boldsymbol{\theta}^*$ we have almost surely

$$\operatorname{Var}\left(\mathbf{X}^\top\boldsymbol{\theta}\middle|\mathbf{X}^\top\boldsymbol{\theta}^*\right) > 0.$$

  Furthermore all pairs $\boldsymbol{\theta}, \boldsymbol{\theta}^\circ \in S_1^{+,p}$ with $\boldsymbol{\theta} \perp \boldsymbol{\theta}^\circ$ satisfy $\left\|\frac{p_{\boldsymbol{\theta}^\circ,\boldsymbol{\theta}}}{p_{\boldsymbol{\theta}}}\right\|_\infty < \infty$ with $p_{\boldsymbol{\theta}^\circ,\boldsymbol{\theta}} : \mathbb{R}^2 \to \mathbb{R}_+$ denoting the density of $(\mathbf{X}^\top\boldsymbol{\theta}^\circ, \mathbf{X}^\top\boldsymbol{\theta}) \in \mathbb{R}^2$.

Remark 2.1. $\operatorname{Var}\left(\mathbf{X}^\top\boldsymbol{\theta}^\circ\middle|\mathbf{X}^\top\boldsymbol{\theta}^*\right) = 0$ would mean that $\mathbf{X}^\top\boldsymbol{\theta}^\circ = a(\mathbf{X}^\top\boldsymbol{\theta}^*)$ for some measurable function $a : \mathbb{R} \to \mathbb{R}$. But then we would have for any $(\alpha, \beta) \in \mathbb{R}^2$ with $\alpha^2 + \beta^2 = 1$ that

$$f(\mathbf{X}^\top(\alpha\boldsymbol{\theta}^* + \beta\boldsymbol{\theta}^\circ)) = f(\alpha\mathbf{X}^\top\boldsymbol{\theta}^* + \beta a(\mathbf{X}^\top\boldsymbol{\theta}^*)) \overset{\text{def}}{=} f^\circ_{\alpha,\beta}(\mathbf{X}^\top\boldsymbol{\theta}^*),$$

such that the problem would no longer be identifiable. We bound $p_{\mathbf{X}} > 0$ on $B_{s_{\mathbf{X}}+c_B}(0)$ to ensure identifiability, also see Remark A.5.

Remark 2.2. We assume that the support of $\mathbb{P}^{\mathbf{X}}$ contains 0 without loss of generality. If that was not the case one could modify the sample as follows. Let $\boldsymbol{x}_0$ be an inner point of the support of $\mathbb{P}^{\mathbf{X}}$. Generate a new sample $(\mathbf{X}'_i)_{i=1,\ldots,n} = (\mathbf{X}_i - \boldsymbol{x}_0)_{i=1,\ldots,n}$ and assume $(\mathbf{Cond}_{\mathbf{X}})$ for this new sample instead.

Of course we need some regularity of the link function $f \in \{f : [-s_{\mathbf{X}}, s_{\mathbf{X}}] \mapsto \mathbb{R}\}$ in (1.5):

$(\mathbf{Cond}_f)$ For some $\boldsymbol{\eta}^* \in B_{\mathbf{r}^\circ}(0) \subset l^2 \overset{\text{def}}{=} \{(u_k)_{k\in\mathbb{N}} : \sum_{k=1}^{\infty} u_k^2 < \infty\}$

$$f = \mathbb{E}[g(\mathbf{X})|\mathbf{X}^\top\boldsymbol{\theta}^* = \cdot] = \boldsymbol{f}_{\boldsymbol{\eta}^*} = \sum_{k=1}^{\infty} \eta_k^* \boldsymbol{e}_k, \tag{2.1}$$

where $\|\boldsymbol{f}'_{\boldsymbol{\eta}^*}\|_\infty < \infty$ and $\|\boldsymbol{f}''_{\boldsymbol{\eta}^*}\|_\infty < \infty$ and where with some $\alpha > 2$

$$\sum_{k=0}^{\infty} k^{2\alpha} \eta_k^{*2} < \infty. \tag{2.2}$$

Remark 2.3. In the case that the data is not from the model (1.4) but from the model in (1.1) the implications of this condition to the function $g : \mathbb{R}^p \to \mathbb{R}$ become somewhat unclear. One way of ensuring that it is satisfied is to assume that for every $\boldsymbol{\theta} \in S_1^{p,+}$ and any $\boldsymbol{x} \in B_{s_{\mathbf{X}}}(0) \cap \boldsymbol{\theta}^\perp$ the function

$$f_{\boldsymbol{\theta},\boldsymbol{x}} : \mathbb{R} \to \mathbb{R}, \quad t \mapsto g(t + \boldsymbol{\theta}^\top \boldsymbol{x}),$$

satisfies (2.1) with some $\boldsymbol{\eta}(\boldsymbol{\theta}, \boldsymbol{x})$ and $\alpha(\boldsymbol{\theta}, \boldsymbol{x}) > 2 + \epsilon$, where $\epsilon > 0$ is independent of $\boldsymbol{x}$. More precisely set for any $\boldsymbol{\theta} \in S_1^{p,+}$

$$f_{\boldsymbol{\theta}}(t) \overset{\text{def}}{=} \mathbb{E}[Y_i|\mathbf{X}^\top\boldsymbol{\theta} = t] = \int_{B_{s_{\mathbf{X}}}(0)\cap\boldsymbol{\theta}^\perp} f_{\boldsymbol{\theta},\boldsymbol{x}}(t) p_{\mathbf{X}|\mathbf{X}^\top\boldsymbol{\theta}=t}(\boldsymbol{x}) d\boldsymbol{x},$$

where $p_{\mathbf{X}|\mathbf{X}^\top\boldsymbol{\theta}=t}(\boldsymbol{x})$ is the conditional density of $\mathbf{X}|\mathbf{X}^\top\boldsymbol{\theta} = t$. Due to the smoothness assumption on $f_{\boldsymbol{\theta},\boldsymbol{x}}(t)$ the function $f_{\boldsymbol{\theta}}(t)$ satisfies (2.1) as well with some $\boldsymbol{\eta}(\boldsymbol{\theta})$ and $\alpha(\boldsymbol{\theta}) \geq \inf_{\boldsymbol{x}\in B_{s_{\mathbf{X}}}(0)\cap\boldsymbol{\theta}^\perp}\{\alpha(\boldsymbol{\theta}, \boldsymbol{x})\} > 2$. We proof this in Section A.

To control the large deviations of $\widetilde{\boldsymbol{v}}_m \in \mathbb{R}^{p^*}$ we use the following assumption:

$(\mathbf{Cond}_{\mathbf{X}\boldsymbol{\theta}^*})$ On some interval $[t_0 - h, t_0 + h] \subseteq [-s_{\mathbf{X}}, s_{\mathbf{X}}]$ with $h > 0$ it holds true that

$$|\boldsymbol{f}'_{\boldsymbol{\eta}^*}(t)| > 0.$$

Remark 2.4. A condition of this kind is necessary to ensure identifiability. Otherwise the function $g : \mathbb{R}^p \to \mathbb{R}$ would be $\mathbb{P}^{\mathbf{X}}$-almost surely constant. But for a constant function $\boldsymbol{\theta}^* \in \mathbb{R}^p$ in (1.3) is not defined.

To be able to apply the finite sample device we need constraints on the moments of the additive noise:

(**Cond**$_\varepsilon$) The errors $(\varepsilon_i) \in \mathbb{R}$ are i.i.d. with $\mathbb{E}[\varepsilon_i] = 0$, $\mathrm{Cov}(\varepsilon_i) = \sigma^2$ and satisfy for all $|\mu| \leq \widetilde{g}$ for some $\widetilde{g} > 0$ and some $\widetilde{\nu} > 0$

$$\log \mathbb{E}[\exp\{\mu\varepsilon_1\}] \leq \widetilde{\nu}^2\mu^2/2.$$

Remark 2.5. Note that our assumptions in terms of moments and smoothness are quite common in this model. For instance [8] assume that the density $p_{\mathbf{X}}$ of the regressors $(\mathbf{X}_i)$ is twice continuously differentiable, that f has two bounded derivatives and that the errors (ε_i) are centered with bounded polynomial moments of arbitrary degree.

Unfortunately these conditions do not facilitate an easy proof of our desired results in the case that the data is not from the model (1.4). To control the large deviations of $\widetilde{\boldsymbol{\upsilon}}_m$ and for identifiability we impose some more "esoteric" conditions on the interplay of the function $g : \mathbb{R}^p \to \mathbb{R}$ and the measure $\mathbb{P}^{\mathbf{X}}$.

(model bias) Assume that

$$\|\mathbb{E}[g(\mathbf{X})|\mathbf{X}^\top\boldsymbol{\theta}^*] - g(\mathbf{X})\| = \|\boldsymbol{f}_{\boldsymbol{\eta}^*}(\mathbf{X}^\top\boldsymbol{\theta}^*) - g(\mathbf{X})\| \leq \mathtt{C}_{bias},$$

for some constant $\mathtt{C}_{bias} \geq 0$. Furthermore we need if $\mathtt{C}_{bias} > 0$ that there exists an open ball $B_{\mathtt{r}_{\boldsymbol{\theta}}}(\boldsymbol{\theta}^*) \subset \mathbb{R}^p$ around $\boldsymbol{\theta}^*$ and a constant $\mathtt{b}_\theta > 0$ such that for $\boldsymbol{\theta} \notin B_{\mathtt{r}_{\boldsymbol{\theta}}}(\boldsymbol{\theta}^*)$

$$-\mathbb{E}\left[\left(g(\mathbf{X}) - \mathbb{E}[g(\mathbf{X})|\mathbf{X}^\top\boldsymbol{\theta}]\right)^2\right] + \mathbb{E}\left[\left(g(\mathbf{X}) - \mathbb{E}[g(\mathbf{X})|\mathbf{X}^\top\boldsymbol{\theta}^*]\right)^2\right] \leq -\mathtt{b}_\theta,$$

and such that on $B_{\mathtt{r}_{\boldsymbol{\theta}}}(\boldsymbol{\theta}^*) \subset \mathbb{R}^{p-1}$ the second derivative exists and satisfies with some $\mathtt{C}_\theta > 0$

$$\nabla^2_{\boldsymbol{\theta}}\mathbb{E}\left[\left(g(\mathbf{X}) - \mathbb{E}[g(\mathbf{X})|\mathbf{X}^\top\boldsymbol{\theta}]\right)^2\right] \geq \mathtt{b}_\theta > 0.$$

Remark 2.6. The conditions (model bias) are of course rather peculiar and not a very accurate characterization of the class of functions that allow the application of our approach. As this paper – even with these conditions – is still very technical we do not elaborate on this issue further. We only point out that this condition is a kind of quantification of how salient the direction $\boldsymbol{\theta}^* \in \mathbb{R}^p$ in (1.3) is.

2.2. Properties of the Wavelet Sieve profile M-estimator

This section presents the application of the results of [2] to the estimator $\widetilde{\boldsymbol{\theta}}_m$ in (1.6). Unfortunately a presentation of the results in full detail would involve constants that are characterized by formulas that would cover many pages. This is why in this work we restrict ourselves to the mere presentation of an upper bound for the critical dimension. This means that we do not specify the size of the appearing constants even though this would be crucial in a true finite sample approach. To further simplify the presentation we bound with some constants $\mathtt{C}, c > 0$

$$1 - 12\mathrm{e}^{-\mathtt{x}} - \exp\left\{-m^3\mathtt{x}\right\} - \exp\left\{-nc_{(\boldsymbol{Q})}/4\right\} \geq 1 - \mathtt{C}\mathrm{e}^{-\mathtt{x}-nc}.$$

Also – in the proofs as well – the same symbol $\mathtt{C}$ can stand for different values, that do not depend on $p^*, m, n, \mathtt{x}$. We use this convention to make the presentation less cumbersome and hope the reader appreciates this despite the loss of rigor.

Remark 2.7. The constant $c_{(\boldsymbol{Q})} > 0$ is derived in the proof of Lemma A.20 and does not depend on $\mathtt{x}$, n, p^*.

2.2.1. The single index case

In this part we only consider the properties of the estimator $\widetilde{\boldsymbol{\theta}}_m$ in (1.6) under the assumption that the model (1.4) is correct. We get the following result by applying Theorem 4.3:

Proposition 2.1. *Assume* $(\mathcal{A})$ *and that the model* (1.4) *is correct. Suppose that* $m^{-(2\alpha+1)}n \to 0$ *and that* $p^{*4}/n \to 0$*. If* $n \in \mathbb{N}$ *is large enough, it holds with probability greater than* $1 - \mathtt{C}\mathrm{e}^{-\mathtt{x}-nc}$

$$\left\|\sqrt{n}(\widetilde{\boldsymbol{\theta}}_m - \boldsymbol{\theta}^*_m) - \breve{\boldsymbol{\xi}}_m\right\| \leq \mathtt{C}\frac{p^{*5/2} + \mathtt{x}}{\sqrt{n}},$$

$$\left|2\breve{L}(\widetilde{\boldsymbol{\theta}}_m, \boldsymbol{\theta}^*_m) - \|\breve{d}_m\breve{\boldsymbol{\xi}}_m(\boldsymbol{\upsilon}^*_m)\|^2\right| \leq \mathtt{C}\left(\sqrt{p+\mathtt{x}} + \frac{p^{*5/2} + \mathtt{x}}{\sqrt{n}}\right)\frac{p^{*5/2} + \mathtt{x}}{\sqrt{n}},$$

where $\breve{d}^2_m$ *is defined in* (1.17) *and* $\breve{\boldsymbol{\xi}}_m$ *in* (1.19)*.*

Remark 2.8. The constraint that n is large enough means, that it should exceed a constant, that depends on characteristics of the model, the basis, on $s_{\mathbf{X}}$, etc. but also on the full dimension p^*. The necessary size of $n \in \mathbb{N}$ is determined by the speed with which $p^{*4}/n \to 0$ and $m^{-2\alpha-1}n \to 0$. In the proof of Proposition 2.1 we impose conditions on $n \in \mathbb{N}$ of the kind

$$p^{*2}/\sqrt{n} \leq \mathtt{C}_1^{-1}, \quad m^{-2\alpha-1}n \leq \mathtt{C}_2^{-1},$$

for certain constants $\mathtt{C}_1, \mathtt{C}_2 > 0$. But note that the approximation error on the right hand sides of the theorem are determined by the size of p^{*5}/n, which has to be small. Consequently for accurate results one needs $p^* = o(n^{1/5})$ if p is assumed to be constant.

So far we only addressed the behavior of the sieve profile ME with respect to the possibly biased target $\boldsymbol{\theta}_m^* \in \mathbb{R}^p$ and with a weighting matrix $\breve{d}_m$ that depends on the dimension $m \in \mathbb{N}$ of the nuisance parameter $\boldsymbol{\eta} \in \mathbb{R}^m$. Addressing the bias we get the following result.

Proposition 2.2. *Assume* $(\mathcal{A})$ *and that the model* (1.4) *is correct. Suppose that* $m^{-(2\alpha+1)}n \to 0$ *and that* $p^{*4}/n \to 0$. *If* $n \in \mathbb{N}$ *is large enough it holds with probability greater than* $1 - \mathtt{C}\mathrm{e}^{-\mathtt{x}-nc}$

$$\left\|\sqrt{n}\big(\widetilde{\boldsymbol{\theta}}_m - \boldsymbol{\theta}^*\big) - \breve{\boldsymbol{\xi}}_m(\boldsymbol{\upsilon}_m^*)\right\| \le \mathtt{C}\left(\frac{p^{*5/2} + \mathtt{x}}{\sqrt{n}} + \sqrt{n}m^{-(\alpha+1/2)}\right),$$

$$\left|2\breve{L}(\widetilde{\boldsymbol{\theta}}_m, \boldsymbol{\theta}^*) - \|\breve{d}_m\breve{\boldsymbol{\xi}}_m(\boldsymbol{\upsilon}_m^*)\|^2\right| \le \mathtt{C}\left(\frac{p^{*5/2} + \mathtt{x}}{\sqrt{n}} + \sqrt{n}m^{-(\alpha+1/2)}\right) \cdot \left(\sqrt{p + \mathtt{x}} + \frac{p^{*5/2} + \mathtt{x}}{\sqrt{n}}\right).$$

Further if $p^{*5}/n \to 0$ *we find as* $n \to \infty$ *– and with* $\breve{d}^2$ *defined in* (1.15) *–*

$$\sqrt{n}\big(\widetilde{\boldsymbol{\theta}}_m - \boldsymbol{\theta}^*\big) \xrightarrow{w} \mathcal{N}(0, \sigma^2\breve{d}^{-2}), \qquad 2\breve{L}(\widetilde{\boldsymbol{\theta}}_m, \boldsymbol{\theta}^*) \xrightarrow{w} \sigma^2\chi_p^2.$$

Remark 2.9. The constraints $m^{-(2\alpha+1)}n \to 0$ and $p^{*5/2}/\sqrt{n} \to 0$ exclude the case $\alpha \le 2$. But note that if $0 < \alpha - 2 = \epsilon$ and $m \ge n^{1/5-\delta}$ with $\delta > 2\epsilon/(25+5\epsilon)$ we get

$$m^{-2\alpha-1}n^1 \le n^{-(1+2\varepsilon_\alpha/5)+\delta(2\alpha+1)+1} = n^{-2\epsilon_\alpha/5+\delta(5+2\epsilon)} \to 0,$$

such that $n = o(m^{2\alpha+1})$ and $p^* = o(n^{1/5})$. Also note that the choice $m = n^{1/(2\alpha+1)}$ is the optimal choice for m – for known $\boldsymbol{\theta}^* \in \mathbb{R}^m$ – in the given setting as a consequence of the bias variance decomposition in nonparametric series estimation; see [16]. It leads to the optimal rate for the mean squared error in the estimation of $f_{\boldsymbol{\eta}^*}$, i.e. $n^{\alpha/(2\alpha+1)}$.

Remark 2.10. It can be shown that if the model (1.4) is correct the matrix $\sigma^2\breve{d}^{-2}$ is the lower bound for the variance of regular estimators of $\boldsymbol{\theta}^* \in \mathbb{R}^m$ if $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ and $\mathbf{X}$ is uniformly distributed on $\mathcal{B}_{s_\mathbf{X}} \subset \mathbb{R}^p$.

2.2.2. The general case

Assume now that the model (1.4) could be wrong. In this case the results for the estimator in (1.6) slightly change. This mainly is a result of the fact that in this case

$$\|g(\mathbf{X}) - \mathbb{E}[g(\mathbf{X})|\mathbf{X}^\top\boldsymbol{\theta}^*]\| > 0.$$

This leads to a more complicated form of the information operator $-\nabla^2\mathbb{E}\mathcal{L}(\boldsymbol{\upsilon}^*)$, which is harder to bound. It is worth mentioning that theses terms would disappear if for all $\boldsymbol{\theta}\perp\boldsymbol{\theta}^*$ the random variable $\mathbf{X}^\top\boldsymbol{\theta}$ was independent of $\mathbf{X}^\top\boldsymbol{\theta}^*$.

Proposition 2.3. *Assume* $(\mathcal{A})$. *Suppose that* $p^{*6}\log(n)/n \to 0$ *and that* $m^{-2(\alpha-1)}n \to 0$. *If* $n\in\mathbb{N}$ *is large enough, it holds with probability greater than* $1-\mathtt{C}e^{-\mathtt{x}-nc}$

$$\left\|\sqrt{n}\big(\widetilde{\boldsymbol{\theta}}_m-\boldsymbol{\theta}^*_m\big)-\breve{\boldsymbol{\xi}}_m(\boldsymbol{\upsilon}^*_m)\right\| \le \mathtt{C}\frac{p^{*7/2}+\mathtt{x}}{\sqrt{n}},$$

$$\left|2\breve{L}(\widetilde{\boldsymbol{\theta}}_m,\boldsymbol{\theta}^*_m)-\|\breve{d}_m\breve{\boldsymbol{\xi}}_m(\boldsymbol{\upsilon}^*_m)\|^2\right| \le \mathtt{C}\left(\sqrt{p+\mathtt{x}}+\frac{p^{*7/2}+\mathtt{x}}{\sqrt{n}}\right)\frac{p^{*7/2}+\mathtt{x}}{\sqrt{n}}.$$

For the unbiased target $\boldsymbol{\theta}^*$ we get the following result.

Proposition 2.4. *Assume* $(\mathcal{A})$. *Suppose that* $p^{*6}\log(n)/n \to 0$ *and that* $m^{-2(\alpha-1)}n \to 0$. *If* $n\in\mathbb{N}$ *is large enough it holds with probability greater than* $1-\mathtt{C}e^{-\mathtt{x}-nc}$

$$\left\|\sqrt{n}\big(\widetilde{\boldsymbol{\theta}}_m-\boldsymbol{\theta}^*\big)-\breve{\boldsymbol{\xi}}_m(\boldsymbol{\upsilon}^*_m)\right\| \le \mathtt{C}\left(\frac{p^{*7/2}+\mathtt{x}}{\sqrt{n}}+\sqrt{n}m^{-(\alpha-1)}\right),$$

$$\left|2\breve{L}(\widetilde{\boldsymbol{\theta}}_m,\boldsymbol{\theta}^*)-\|\breve{d}_m\breve{\boldsymbol{\xi}}_m(\boldsymbol{\upsilon}^*_m)\|^2\right| \le \mathtt{C}\left(\frac{p^{*7/2}+\mathtt{x}}{\sqrt{n}}+\sqrt{n}m^{-(\alpha-1)}\right) \cdot\left(\sqrt{p+\mathtt{x}}+\frac{p^{*7/2}+\mathtt{x}}{\sqrt{n}}\right).$$

Remark 2.11. Note that we do not show any weak convergence statements for the general case. The approach of [1] is not applicable – at least not with the arguments we use in Lemma A.6 for the case that the model in (1.4) is correct. Also note that to control the approximation bias the necessary smoothness of $\mathbb{E}[g(\mathbf{X})|\mathbf{X}^\top\boldsymbol{\theta}^*=\cdot]=\boldsymbol{f}_{\boldsymbol{\eta}^*}(\cdot):\mathbb{R}\to\mathbb{R}$ measured in $\alpha>0$ in (2.2) increases from $\alpha>2$ to $\alpha>14/3$ to ensure that $\alpha(m)\to 0$.

2.3. A way to calculate the profile estimator

In this section we briefly sketch how to actually calculate $\widetilde{\boldsymbol{\upsilon}}\in\mathbb{R}^{p^*}$ in practice. We assume that the full dimension $p+m\in\mathbb{N}$ is finite and thus suppress the index$\cdot_m$. For this note that the maximization problem

$$\widetilde{\boldsymbol{\upsilon}}=\operatorname*{argmax}_{\Upsilon}\sum_{i=1}^n(Y_i-\boldsymbol{f}_{\boldsymbol{\eta}}(\boldsymbol{\theta}^\top\mathbf{X}_i))^2/2,$$

is not convex and thus computationally involved. We propose to obtain the maximizer via the alternation maximization procedure as it is analyzed in [3]. This sequential algorithm – which is a variant of the EM algorithm – works as follows: Start with some initial guess $\widetilde{\boldsymbol{\upsilon}}^{(0)} = (\widetilde{\boldsymbol{\theta}}^{(0)}, \widetilde{\boldsymbol{\eta}}^{(0)}) \in \varUpsilon$. Then calculate for $k \in \mathbb{N}$ iteratively

$$\widetilde{\boldsymbol{\upsilon}}^{(k,k+1)} \stackrel{\text{def}}{=} (\widetilde{\boldsymbol{\theta}}^{(k)}, \widetilde{\boldsymbol{\eta}}^{(k+1)}) = \left(\widetilde{\boldsymbol{\theta}}^{(k)}, \operatorname*{argmax}_{\boldsymbol{\eta}} \mathcal{L}(\widetilde{\boldsymbol{\theta}}^{(k)}, \boldsymbol{\eta})\right),$$

$$\widetilde{\boldsymbol{\upsilon}}^{(k,k)} \stackrel{\text{def}}{=} (\widetilde{\boldsymbol{\theta}}^{(k)}, \widetilde{\boldsymbol{\eta}}^{(k)}) = \left(\operatorname*{argmax}_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}, \widetilde{\boldsymbol{\eta}}^{(k)}), \widetilde{\boldsymbol{\eta}}^{(k)}\right).$$

In the following we write $\widetilde{\boldsymbol{\upsilon}}^{(k,k(+1))}$ in statements that are true for both $\widetilde{\boldsymbol{\upsilon}}^{(k,k+1)}$ and $\widetilde{\boldsymbol{\upsilon}}^{(k,k)}$. For the initial guess we propose a simple grid search. For this generate a uniform grid $G_N \stackrel{\text{def}}{=} (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_N) \subset S_1^+$ and define

$$\widetilde{\boldsymbol{\upsilon}}^{(0)} \stackrel{\text{def}}{=} \operatorname*{argmax}_{\substack{(\boldsymbol{\theta},\boldsymbol{\eta}) \in \varUpsilon \\ \boldsymbol{\theta} \in G_N}} \mathcal{L}(\boldsymbol{\upsilon}). \tag{2.3}$$

Note that given the grid the above maximizer is easily obtained. Simply calculate

$$\begin{aligned} \widetilde{\boldsymbol{\eta}}_l^{(0)} &\stackrel{\text{def}}{=} \operatorname*{argmax}_{l=1,\ldots,N} \mathcal{L}(\boldsymbol{\theta}_l, \boldsymbol{\eta}) \\ &= \left(\frac{1}{n}\sum_{i=1}^n \boldsymbol{e}\boldsymbol{e}^\top(\mathbf{X}_i^\top \boldsymbol{\theta}_l)\right)^{-1} \frac{1}{n}\sum_{i=1}^n Y_i \boldsymbol{e}^\top(\mathbf{X}_i^\top \boldsymbol{\theta}_l) \in \mathbb{R}^m, \end{aligned} \tag{2.4}$$

where by abuse of notation $\boldsymbol{e} = (\boldsymbol{e}_1, \ldots, \boldsymbol{e}_m) \in \mathbb{R}^m$. Observe that

$$\widetilde{\boldsymbol{\upsilon}}^{(0)} = \operatorname*{argmax}_{l=1,\ldots,N} \mathcal{L}(\boldsymbol{\theta}_l, \widetilde{\boldsymbol{\eta}}_l^{(0)}).$$

Define the fineness of the grid via $\tau \stackrel{\text{def}}{=} \sup_{\boldsymbol{\theta},\boldsymbol{\theta}^\circ \in G_N} \|\boldsymbol{\theta} - \boldsymbol{\theta}^\circ\|$. To assess the properties of the alternating procedure we apply Theorem 4.6 and Theorem 2.4 of [3]. Before we present the results we need to introduce the constant ρ, that plays a central role in this part. It is defined – with the blocks of the information matrix in (1.16) – as

$$\|d^{-1} a h^{-1}\|^2 = \rho.$$

With Lemma 5.3 we know that $0 \le \rho < 1$. To ease the presentation we again distinguish the case that the model (1.4) is correct from the general one.

2.3.1. *The single index case*

Theorem 4.6 yields:

Proposition 2.5. *Assume* $(\mathcal{A})$ *and that the model* (1.4) *is correct. Suppose that* $m^{-(2\alpha+1)}n \to 0$ *and that* ${p^*}^4/n \to 0$*. Set* $\tau = o({p^*}^{-3/2})$*. With the initial guess given by Equation* (2.3) *and for* $n \in \mathbb{N}$ *large enough the alternating sequence satisfies with probability greater than* $1 - \mathtt{C}e^{-\mathtt{x}-nc}$

$$\left\|\sqrt{n}(\widetilde{\boldsymbol{\theta}}^{(k)} - \boldsymbol{\theta}^*) - \breve{\boldsymbol{\xi}}\right\| \le \mathtt{C}\frac{{p^*}^{5/2} + \rho^{2k}R_0^2(\mathtt{x}) + \mathtt{x}}{\sqrt{n}}, \tag{2.5}$$

$$\begin{aligned}\left|2\breve{L}(\widetilde{\boldsymbol{\theta}}^{(k)}, \boldsymbol{\theta}^*) - \|\breve{d}\breve{\boldsymbol{\xi}}\|^2\right| \le \mathtt{C}&\frac{{p^*}^{5/2} + \rho^{2k}R_0^2(\mathtt{x}) + \mathtt{x}}{\sqrt{n}} \\ &\cdot\left(\sqrt{p+\mathtt{x}} + \frac{{p^*}^{5/2} + \rho^{2k}R_0^2(\mathtt{x}) + \mathtt{x}}{\sqrt{n}}\right),\end{aligned} \tag{2.6}$$

where

$$R_0(\mathtt{x}) \le \mathtt{C}{p^*}^{3/4}\sqrt{p^* + \mathtt{x} + n\tau^2 + \sqrt{n}\tau\sqrt{\mathtt{x}}}.$$

Remark 2.12. The constraint $\tau = o({p^*}^{-3/2})$ implies that for the calculation of the initial guess the vector $\widetilde{\boldsymbol{\eta}}^{(0)}_{(l)}$ in (2.4) and the functional $\mathcal{L}(\cdot)$ have to be evaluated $N = {p^*}^{3(p-1)/2}$ times. This means – since $m^5 = o(n)$ is necessary for the right-hand sides in (2.6) and (2.5) to vanish – that we need an accuracy of the first guess of order $o(n^{-3/10})$, while the accuracy of the output of the alternating procedure is of order $n^{-1/2}$. In the general case we need – see below – an accuracy of the first guess of order $o(n^{-9/26})$ because $\tau = o(m^{-9/4})$ and $m^{13/2} = o(n)$. Although this difference does not seem large the number of grid points necessary for $n^{-1/2}$-accuracy of the grid search is by a factor $n^{(p-1)/5}$ or $n^{2(p-1)/13}$ larger than those for a sufficient initial guess.

The above Theorem tells us that as far as statistical inference is concerned the estimator $\widetilde{\boldsymbol{\theta}}^{(k)}$ and the profile ME $\widetilde{\boldsymbol{\theta}}$ are interchangeable as soon as $\rho^k \mathrm{R}_0$ is of the order of p^*. If not the statistical properties but mere convergence of the sequence $\widetilde{\boldsymbol{\upsilon}}^{(k,k(+1))} \to \widetilde{\boldsymbol{\upsilon}}$ is desired we can prove the following result using Theorem 2.4 of [3].

Proposition 2.6. *Assume* $(\mathcal{A})$ *and that the model* (1.4) *is correct. Suppose that* $m^{-(2\alpha+1)}n \to 0$ *and that* ${p^*}^4/n \to 0$*. Set* $\tau = o({p^*}^{-3/2})$*. Let* $\mathtt{x} > 0$ *be chosen such that*

$$\mathtt{x} \le \frac{1}{2}\left(\widetilde{\nu}^2 n\widetilde{\mathtt{g}}^2 - \log(p^*)\right) \wedge p^*.$$

Then with the initial guess given by Equation (2.3) and for $n \in \mathbb{N}$ large enough

$$\mathbb{P}\left(\bigcap_{k\in\mathbb{N}}\left\{\sqrt{n}\left\|\widetilde{\boldsymbol{\upsilon}}-\widetilde{\boldsymbol{\upsilon}}^{(k,k(+1))}\right\|\le \mathtt{r}_k^*\right\}\right)\ge 1-\mathtt{C}\mathrm{e}^{-\mathtt{x}-nc},$$

where with some constant $\mathtt{C}_0>0$ and $\kappa(\mathtt{x})\le \mathtt{C}{p^*}^{3/2}\sqrt{p^*+\mathtt{x}}$

$$\mathtt{r}_k^* \le \begin{cases}\rho^k\frac{\sqrt{n}}{\sqrt{n}-\kappa(\mathtt{x})k}R_0, & \kappa(\mathtt{x})k\le\sqrt{n},\\ \rho^{\frac{k}{\log(k)}\log\left(\frac{\sqrt{n}(1-\rho)}{\kappa(\mathtt{x})}\right)}c_kR_0, & \textit{otherwise,}\end{cases}$$

with some sequence $(c_k)\in\mathbb{N}$, where $0<c_k\to 2$ and with

$$R_0 \le \mathtt{C}\sqrt{p^*+\mathtt{x}+n\tau^2+\sqrt{n}\tau\sqrt{\mathtt{x}}}.$$

Remark 2.13. In a nutshell this tells us that up to a $1/\log(k)$-factor we can ensure linear convergence of the sequence $(\widetilde{\boldsymbol{\theta}}^{(k)},\widetilde{\boldsymbol{\eta}}^{(k)})$ to the global maximizer $\widetilde{\boldsymbol{\upsilon}}$.

Remark 2.14. Note that the constraint on the size of the dimension $p^*\in\mathbb{N}$ for accurate results is weaker in Proposition 2.6 than in Proposition 2.5 because there are no "right-hand sides" and thus $m^4=o(n)$ is sufficient.

2.3.2. The general case

Again the results become worse in the general case:

Proposition 2.7. *Assume $(\mathcal{A})$. Suppose that $m^6\log(n)/n\to 0$ and that $m^{-2(\alpha-1)}n\to 0$. Furthermore let $\mathtt{x}\le 2\widetilde{\nu}^2\widetilde{\mathtt{g}}^2(1+\mathtt{C}_{bias})n$ set $\tau=o(m^{-11/4})$. With the initial guess given by Equation (2.3) the alternating sequence satisfies with probability greater than $1-\mathtt{C}\mathrm{e}^{-\mathtt{x}-nc}$*

$$\begin{aligned}\left\|\sqrt{n}(\widetilde{\boldsymbol{\theta}}^{(k)}-\boldsymbol{\theta}^*)-\breve{\boldsymbol{\xi}}\right\| &\le \mathtt{C}\frac{{p^*}^{7/2}+\rho^{2k}R_0^2(\mathtt{x})+\mathtt{x}}{\sqrt{n}},\\ \left|2\breve{L}(\widetilde{\boldsymbol{\theta}}^{(k)},\boldsymbol{\theta}^*)-\|\breve{d}\breve{\boldsymbol{\xi}}\|^2\right| &\le \mathtt{C}\frac{{p^*}^{7/2}+\rho^{2k}R_0^2(\mathtt{x})+\mathtt{x}}{\sqrt{n}}\\ &\qquad\cdot\left(\sqrt{p+\mathtt{x}}+\frac{{p^*}^{7/2}+\rho^{2k}R_0^2(\mathtt{x})+\mathtt{x}}{\sqrt{n}}\right),\end{aligned}$$

where

$$R_0(\mathtt{x})\le \mathtt{C}{p^*}^{5/4}\sqrt{p^*\log(n)+\mathtt{x}+\sqrt{m}n\tau^2+\sqrt{n}\tau\sqrt{\mathtt{x}}}.$$

Proposition 2.8. *Assume* $(\mathcal{A})$. *Suppose* $m^6 \log(n) = o(n)$ *and assume that* $m^{-2(\alpha-1)} n \to 0$. *Set* $\tau = o(m^{-11/4})$ *and let* $\mathtt{x} > 0$ *be chosen such that*

$$\mathtt{x} \le \frac{1}{2}\left(\widetilde{\nu}^2 n \widetilde{\mathtt{g}}^2 - \log(p^*)\right) \wedge p^*.$$

Then

$$\mathbb{P}\left(\bigcap_{k\in\mathbb{N}} \left\{\sqrt{n}\left\|\widetilde{\boldsymbol{\upsilon}} - \widetilde{\boldsymbol{\upsilon}}^{(k,k(+1))}\right\| \le \mathtt{r}_k^*\right\}\right) \ge 1 - \mathtt{C}\mathrm{e}^{-2\mathtt{x}-nc},$$

where with some constant $\mathtt{C}_0$ *and* $\kappa(\mathtt{x}) \le \mathtt{C}{p^*}^{5/2}\sqrt{p^* + \mathtt{x}}$

$$\mathtt{r}_k^* \le \begin{cases} \rho^k \frac{\sqrt{n}}{\sqrt{n} - \kappa(\mathtt{x})k} R_0, & \kappa(\mathtt{x})k \le \sqrt{n}, \\ \rho^{\frac{k}{\log(k)} \log\left(\frac{\sqrt{n}(1-\rho)}{\kappa(\mathtt{x})}\right)} c_k R_0, & \textit{otherwise}, \end{cases}$$

with some sequence $(c_k) \in \mathbb{N}$, *where* $0 < c_k \to 2$ *and with*

$$R_0 \le \mathtt{C}\sqrt{p^* \log(n) + \mathtt{x} + \sqrt{m} n \tau^2 + \sqrt{n}\tau\sqrt{\mathtt{x}}}.$$

2.4. Performance of Projection Pursuit Procedure

In this section we want to briefly assess the performance of the Projection Pursuit procedure of [7] as we explained it in the introduction (i.e. in 1). We assume that the iteration $k \in \mathbb{N}$ in the alternation maximization procedure is large enough so that we can pretend that one can directly access the maximizer $\widetilde{\boldsymbol{\upsilon}}$. Also we assume that the number of iterations $M \in \mathbb{N}$ is fixed. Further we again suppress $\cdot_m$ to ease notation. In the previous sections we already established that for observations of the kind

$$Y_i = g(\mathbf{X}_i) + \varepsilon_i, \; i = 1, \ldots, n,$$

the estimator in (1.6) satisfies

$$\begin{aligned} &\left|\mathbb{E}[Y|\mathbf{X}^\top \boldsymbol{\theta}^*_{(1)}] - \boldsymbol{f}_{\widetilde{\boldsymbol{\eta}}_{(1)}}(\mathbf{X}^\top \widetilde{\boldsymbol{\theta}}_{(1)})\right| \\ &\qquad \le \mathtt{C}\left(\mathtt{r}^* + \alpha(m) + \diamondsuit(\mathtt{x}) + \|\mathcal{D}_{(1)}{}^{-1}\nabla\mathcal{L}_{(1)}(\boldsymbol{\upsilon}^*_{(1)})\|\right)/\sqrt{n}, \end{aligned} \tag{2.7}$$

with high probability. But in each step a new data set is generated, i.e. given $Y_{i(l)}, \widetilde{\boldsymbol{\upsilon}}_{(l)}$ we generate

$$Y_{i(l+1)} \overset{\text{def}}{=} Y_{i(l)} - \boldsymbol{f}_{\widetilde{\boldsymbol{\eta}}_{(l)}}(\mathbf{X}_i^\top \widetilde{\boldsymbol{\theta}}_{(l)}) = g_{(l+1)}(\mathbf{X}_i) + \varepsilon_i + \tau_{i(l)},$$

where

$$g_{(l)}(\mathbf{X}_i) \approx \sum_{s=l}^{M} \boldsymbol{f}_{\boldsymbol{\eta}^*_{(s)}}(\mathbf{X}_i^\top \boldsymbol{\theta}^*_{(s)}),$$

$$\tau_{i(l)} = \sum_{s=1}^{l} \boldsymbol{f}_{\boldsymbol{\eta}^*_{(s)}}(\mathbf{X}_i^\top \boldsymbol{\theta}^*_{(s)}) - \boldsymbol{f}_{\widetilde{\boldsymbol{\eta}}_{(s)}}(\mathbf{X}_i^\top \widetilde{\boldsymbol{\theta}}_{(s)}).$$

The errors $\tau_{i(l)}$ are not i.i.d. and not necessarily centered such that we can not directly apply the results from above for $l > 1$. But a slight modification serves a remedy. For this remember that the central tool for Theorems of the type of 4.3 is to bound with probability $1 - \mathrm{e}^{-\mathtt{x}}$

$$\sup_{\boldsymbol{\upsilon} \in \Upsilon_\circ(\mathtt{r}_0)} \left\| \mathcal{D}^{-1}\left(\nabla \mathcal{L}(\boldsymbol{\upsilon}) - \nabla \mathcal{L}(\boldsymbol{\upsilon}^*)\right) + \mathcal{D}(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^*) \right\| \le \diamondsuit(\mathtt{r}_0, \mathtt{x}),$$

and to show that $\mathbb{P}(\widetilde{\boldsymbol{\upsilon}}, \widetilde{\boldsymbol{\upsilon}}_{\boldsymbol{\theta}^*} \in \Upsilon_\circ(\mathtt{r}_0)) \ge 1 - \mathrm{e}^{-\mathtt{x}}$. So we decompose

$$\begin{aligned}
&\mathcal{L}_{(l)}(\boldsymbol{\upsilon}, Y_{i(l)}) \\
&\quad = -\sum_{i=1}^{n} \left(g_{(l)}(\mathbf{X}_i) + \varepsilon_i - \boldsymbol{f}_{\boldsymbol{\eta}}(\mathbf{X}_i^\top \boldsymbol{\theta})\right)^2 \\
&\qquad - \sum_{i=1}^{n} {\tau_{i(l-1)}}^2 + 2\sum_{i=1}^{n} \tau_{i(l-1)} \left(\boldsymbol{f}_{\boldsymbol{\eta}}(\mathbf{X}_i^\top \boldsymbol{\theta}) - \boldsymbol{f}_{\boldsymbol{\eta}^*_{(l)}}(\mathbf{X}_i^\top \boldsymbol{\theta}^*_{(l)})\right) \\
&\quad \overset{\mathrm{def}}{=} \mathcal{L}_{\varepsilon(l)}(\boldsymbol{\upsilon}, Y_{i(l)}) + \mathcal{L}_{(l)_\tau}(\boldsymbol{\upsilon}, Y_{i(l)}),
\end{aligned}$$

and define

$$\begin{aligned}
\boldsymbol{\upsilon}^*_{m(l)} &\overset{\mathrm{def}}{=} \operatorname*{argmax}_{\boldsymbol{\upsilon} \in \Upsilon_m} \mathbb{E}\mathcal{L}_{\varepsilon(l)}(\boldsymbol{\upsilon}), \\
{\mathcal{D}_{m(l)}}^2 &\overset{\mathrm{def}}{=} \nabla^2 \mathbb{E}[\mathcal{L}_{\varepsilon(l)}(\boldsymbol{\upsilon}^*_{m(l)})], \\
\boldsymbol{\zeta}_{\varepsilon(l)}(\boldsymbol{\upsilon}) &\overset{\mathrm{def}}{=} \mathcal{L}_{\varepsilon(l)}(\boldsymbol{\upsilon}) - \mathbb{E}\mathcal{L}_{\varepsilon(l)}(\boldsymbol{\upsilon}).
\end{aligned}$$

We assume that the condition (model bias) holds for every function $g_{(l)}$. With Remark 2.3, Lemma 5.3 and Lemma A.6 this means that the conditions of Section 4.1 and 4.3 are met for $(\mathcal{L}_{\varepsilon(l)}, \Upsilon_m, \mathcal{D}_{m(l)})$ with high probability for every $l = 1, \ldots, M$. It remains to show that for each $l \in \mathbb{N}$ and $m \in \mathbb{N}$ large enough the contribution of $\tau_{i(l)}$ remains insignificant. We do this in the proof of the following Proposition.

Proposition 2.9. *Assume that $M = O(p^*)$ and that the conditions $(\mathcal{A})$ hold for every $l = 1 \ldots, M$. Assume further $\frac{{p^*}^3 \log(n) M + \mathtt{x}}{\sqrt{n}} \to 0$ and assume that*

$m^{-2(\alpha-1)}n \to 0$. For $n \in \mathbb{N}$ large enough and with probability greater than $1 - \mathtt{C}M\mathrm{e}^{-\mathtt{x}-nc}$ we have

$$\sup_{\boldsymbol{x}\in B_{s_{\mathbf{X}}}(0)} \left| \sum_{l=1}^{M} \boldsymbol{f}_{\boldsymbol{\eta}^*_{(l)}}(\boldsymbol{x}^\top \boldsymbol{\theta}^*_{(l)}) - \boldsymbol{f}_{\widetilde{\boldsymbol{\eta}}_{(l)}}(\boldsymbol{x}^\top \widetilde{\boldsymbol{\theta}}_{(l)}) \right| \le \mathtt{C}M\sqrt{m}\left(\frac{p^{*7/2} + \mathtt{x}}{n} + \frac{\sqrt{p^* + \mathtt{x}}}{\sqrt{n}} \right).$$

Remark 2.15. Denoting the bias

$$b(M) \stackrel{\mathrm{def}}{=} \left\| g - \sum_{l=1}^{M} \boldsymbol{f}_{\boldsymbol{\eta}^*_{(l)}}(\cdot^\top \boldsymbol{\theta}^*_{(l)}) \right\|_\infty ,$$

Proposition 2.9 implies if $\mathtt{x} \le p^* = o(n^{1/6})$ that

$$\sup_{\boldsymbol{x}\in B_{s_{\mathbf{X}}}(0)} \left| g(\boldsymbol{x}) - \sum_{l=1}^{M} \boldsymbol{f}_{\widetilde{\boldsymbol{\eta}}_{(l)}}(\boldsymbol{x}^\top \widetilde{\boldsymbol{\theta}}_{(l)}) \right| \le \mathtt{C}Mo(n^{-1/3}) + b(M).$$

Depending on the speed with which $b(M)$ decays in M the resulting rate can be substantially faster than $n^{-\alpha/(2\alpha+p)}$.

3. Technical peculiarities

Before we explain in more detail how the above statements can be derived based on the theory presented in [2] and [3], we address two technical issues that arise with the regression setup with random design and due to the peculiarities of the sieve approach.

3.1. Choice of basis

To control the approximation bias of the sieve estimator $\widetilde{\boldsymbol{\theta}}_m \in \mathbb{R}^p$ with the approach from [1] we can not use any basis $(\boldsymbol{e}_k)_{k\in\mathbb{N}}$ in $L^2([-s_{\mathbf{X}}, s_{\mathbf{X}}])$. We need to show in the proof of Lemma A.6 that the following terms vanish as $m \to \infty$

$$\int_{\mathbb{R}} \boldsymbol{e}_{m+k}(x)\boldsymbol{e}_{m+l}(x)p_{\mathbf{X}^\top\boldsymbol{\theta}^*}(x)dx;\ l,k \in \mathbb{N}, \tag{3.1}$$

where $p_{\mathbf{X}^\top\boldsymbol{\theta}^*}$ denotes the density of $\mathbf{X}^\top\boldsymbol{\theta}^* \in \mathbb{R}$. But it is not clear whether terms as in (3.1) vanish for any basis of $L^2([-s_{\mathbf{X}}, s_{\mathbf{X}}])$. Of course – following [18] – we could assume that the basis is orthogonal in the inner product induced by the Hessian $\nabla^2\mathbb{E}\mathcal{L}(\boldsymbol{\upsilon}^*)$. But for this one would need to know the true parameter $\boldsymbol{\theta}^* \in \mathbb{R}^p$ and the density $p_{\mathbf{X}} : \mathbb{R}^p \to \mathbb{R}$ in advance. We want to avoid such assumptions and also the tedious calculations resulting from using an estimator of $\boldsymbol{\theta}^*$ plugged into an estimator of $p_{\mathbf{X}^\top\cdot}$ for the construction of a suitable basis. As it turns out an orthonormal wavelet basis is suitable for our purpose. For high indexes $k \in \mathbb{N}$ the support of each wavelet $\boldsymbol{e}_k$ is contained in a small interval on which the density $p_{\mathbf{X}^\top \boldsymbol{\theta}^*}$ can be well approximated by a constant. Due to orthogonality and shrinking supports of the basis the term in (3.1) can be shown to diminish sufficiently fast for a Lipschitz continuous density $p_{\mathbf{X}^\top \boldsymbol{\theta}^*}$ (see Lemma A.6). The trouble is that our approach relies on smoothness of the basis elements. Consequently we need a smooth orthogonal wavelet basis on an interval. Thanks to [4] (Theorem 4.4) there can indeed be obtained a basis $(\boldsymbol{e}_k))$ for $L^2([-s_{\mathbf{X}}, s_{\mathbf{X}}])$, that is contained in $C^3(\mathbb{R})$ and satisfies for any $l, k \in \mathbb{N}$ with $k = 2^{j_k} + j_k 17 - 1 + r_k \in \mathbb{N}$ and $r_k \in \{0, \ldots, 2^{j_k} + 2 * 17 - 1\}$

$$\langle \boldsymbol{e}_l, \boldsymbol{e}_k \rangle_{L^2(\mathbb{R})} = \delta_{l,k}, \quad |\mathrm{supp}(\boldsymbol{e}_k)| \le 2^{-j_k} 17 s_{\mathbf{X}}.$$

This basis has another useful property that will come in handy in the proof of Lemma A.6: For any $k \in \mathbb{N}$ with $k = 2^{j_k} + j_k 17 - 1 + r_k \in \mathbb{N}$ it holds

$$\begin{aligned} &\left| \left\{ l = 2^{j_l} + j_l 17 + r_l \middle| r_l \in \{0, \ldots, 2^{j_l} + 16\}, \, \mathrm{supp}(\boldsymbol{e}_k) \cap \mathrm{supp}(\boldsymbol{e}_l) \neq \emptyset \right\} \right| \\ &\quad \le \ulcorner 2^{(j_l - j_k)} 17 \urcorner. \end{aligned} \tag{3.2}$$

In words this means that the number of nonempty intersections of the supports of $\boldsymbol{e}_k$ and $\boldsymbol{e}_l$ can be controlled well. For nearly all basis functions $\boldsymbol{e}_l$ with $l \ge k$ we have

$$\int_{\mathbb{R}} \boldsymbol{e}_k(x) \boldsymbol{e}_l(x) p_{\mathbf{X}^\top \boldsymbol{\theta}^*}(x) dx = 0.$$

This will allow to satisfy the conditions $(\boldsymbol{\kappa})$ and $(\boldsymbol{\upsilon\kappa})$ – which allows to bound the sieve bias (1.10) as is shown in [1] – in Lemma A.6.

3.2. Implications of Regression setup

Due to the regression set up there are some particularities to the analysis that we have to point out here. The definition of $\boldsymbol{\upsilon}_m^* \in \varUpsilon$ reads

$$\boldsymbol{\upsilon}_m^* \overset{\text{def}}{=} \operatorname*{argmax}_{\boldsymbol{\upsilon} \in \varUpsilon_m} \mathbb{E} \mathcal{L}_m(\boldsymbol{\upsilon}),$$

where $\mathbb{E}$ denotes the expectation operator with respect to the joint measure of $(\mathbf{X}, \varepsilon) \in \mathbb{R}^p \times \mathbb{R}$, similarly $\mathcal{D}^2(\boldsymbol{\upsilon})$ is also based on the full expectation $\mathbb{E}$. But in Lemma 5.3 we show the conditions $(\mathcal{E}\mathcal{D}_0)$, $(\mathcal{E}\mathbf{r})$ and $(\mathcal{E}\mathcal{D}_1)$ for the random variables

$$\nabla(1 - \mathbb{E}_\varepsilon)\mathcal{L}_m(\boldsymbol{\upsilon}) \in \mathbb{R}^{p+m},$$

i.e. we use only the expectation with respect to the noise $(\varepsilon_i)_{i=1,\ldots,n}$ but conditional on $(\mathbf{X}_i)_{i=1,\ldots,n}$. This leads to rather weak conditions on the errors (ε_i).

Especially the conditions $(\mathcal{E}\mathtt{r})$ and $(\mathcal{E}\mathcal{D}_1)$ would otherwise become quite restrictive. But on the other hand this means that a list of additional steps become necessary to apply the theory of [2] and [3]. As becomes evident from the proof of Theorem 2.2 of [2], we have to bound the term

$$\sup_{\boldsymbol{\upsilon}\in\Upsilon_\circ(\mathtt{r})} \|\mathcal{D}_m^{-1}\nabla(\mathbb{E}-\mathbb{E}_\varepsilon)[\mathcal{L}_m(\boldsymbol{\upsilon}_m^*)-\mathcal{L}_m(\boldsymbol{\upsilon})]\|,$$

and ad the obtained bound to the error term $\breve{\diamondsuit}(\mathtt{r}_0,\mathtt{x})$. The matrix $\mathcal{D}_m$ is defined in (1.16). Also the probability of the desired bound has to be subtracted from the probability under which the event in Theorem 2.2 of [2] is valid. See Section 4.2 for more details. The following lemma serves this bound.

Lemma 3.1. *With some constant* $\mathtt{C}>0$

$$\mathbb{P}\Bigg(\bigcap_{\mathtt{r}\le R_0}\Bigg\{\sup_{\boldsymbol{\upsilon}\in\Upsilon}\|\mathcal{D}_m^{-1}\nabla(\mathbb{E}-\mathbb{E}_\varepsilon)[\mathcal{L}_m(\boldsymbol{\upsilon}_m^*)-\mathcal{L}_m(\boldsymbol{\upsilon})]\| \\ \ge \mathtt{C}\mathtt{r}\sqrt{\mathtt{x}+p^*\log(p^*)}/\sqrt{n}\Bigg\}\Bigg)\le \mathrm{e}^{-\mathtt{x}}.$$

Remark 3.1. We will see that the error term

$$\mathtt{C}\mathtt{r}\sqrt{\mathtt{x}+p^*\log(p^*)}/\sqrt{n},$$

is of smaller order than the bounds that we will derive for $\breve{\diamondsuit}(\mathtt{r},\mathtt{x})$ in the subsequent analysis, namely bounds of the order $p^{*5/2}/\sqrt{n}$. Consequently we neglect it in the following and let a constant $\mathtt{C}>0$ account for its contribution in the formulation of our results.

Furthermore in the derivation of the conditions $(\mathcal{E}\mathcal{D}_0)$, $(\mathcal{E}\mathtt{r})$ and $(\mathcal{E}\mathcal{D}_1)$ we obtain bounds for $\nu_1,\nu_0,\nu_\mathtt{r}$ that involve terms of the kind

$$\|\mathbb{E}\left[\boldsymbol{S}_n\right]-\boldsymbol{S}_n\|, \qquad \boldsymbol{S}_n=\frac{1}{n}\sum_{i=1}^n \boldsymbol{M}(\mathbf{X}_i), \qquad \boldsymbol{M}(\mathbf{X}_i)\in\mathbb{R}^{p^*\times p^*}.$$

This leads to concentration bounds for sums of i.i.d. random matrices which can be handled with the results of [22]. We do this in Section A.8.3. Again the set on which Theorem 2.2 of [2] occurs has to be intersected with the set on which the matrix deviation bounds are valid. Another implication is that when proving condition $(\mathcal{L}\mathtt{r})$ we have to consider $\mathbb{E}_\epsilon\mathcal{L}(\boldsymbol{\upsilon},\boldsymbol{\upsilon}_m^*)$ instead of $\mathbb{E}\mathcal{L}(\boldsymbol{\upsilon},\boldsymbol{\upsilon}_m^*)$, which makes the proof quite involved and again makes the restriction to a set of high probability necessary. This is why in Proposition 2.1 the probability of the desired results can only be bounded from bellow by $1-12\mathrm{e}^{-\mathtt{x}}-\mathtt{C}\mathrm{e}^{-nc-p^*\mathtt{x}}$ instead of $1-5\mathrm{e}^{-\mathtt{x}}$ as in Proposition 2.4 of [2].

4. Synopsis of the finite sample theory for M-Estimators

In this section we briefly summarize the results of [2] and [3] and thereby adapt them to the regression setting of the given model. The presentation mimics that of those papers and thus is rather abstract. Readers familiar with those papers can skip this section.

4.1. Conditions

This section collects the conditions that underlie the results of [2] and [3]. They are taken from [2] but are adapted to our setting. This means in particular that the expectation operator in the moment conditions is $\mathbb{E}_\varepsilon$ and not the full one. [2] assume that the function $\mathcal{L}(\boldsymbol{\upsilon})\colon \mathbb{R}^{p^*} \to \mathbb{R}$ is sufficiently smooth in $\boldsymbol{\upsilon} \in \mathbb{R}^{p^*}$, $\nabla\mathcal{L}(\boldsymbol{\upsilon}) \in \mathbb{R}^{p^*}$ stands for the gradient and $\nabla^2\mathbb{E}\mathcal{L}(\boldsymbol{\upsilon}) \in \mathbb{R}^{p^*\times p^*}$ for the Hessian of the expectation $\mathbb{E}\mathcal{L} : \mathbb{R}^{p^*} \to \mathbb{R}$ at $\boldsymbol{\upsilon} \in \mathbb{R}^{p^*}$. By smooth enough we mean that all appearing derivatives exist and that we can interchange $\nabla\mathbb{E}\mathcal{L}(\boldsymbol{\upsilon}) = \mathbb{E}\nabla\mathcal{L}(\boldsymbol{\upsilon})$ on $\varUpsilon_\circ(\mathtt{r}_0)$, where $\mathtt{r}_0 > 0$ is defined in equation (4.2) and $\varUpsilon_\circ(\mathtt{r})$ in equation (4.1). This clearly is the case for the given problem of this work.

With $\boldsymbol{\upsilon}^*_m \in \mathbb{R}^{p^*}$ from (1.18) define $\mathcal{D}_m \overset{\text{def}}{=} nd_m(\boldsymbol{\upsilon}^*_m)$ with $d_m(\boldsymbol{\upsilon})$ in (1.16) and $\mathcal{V}^2_m = n\operatorname{Cov}(\nabla\ell_m(\boldsymbol{\upsilon}^*_m)) \in \mathbb{R}^{p^*\times p^*}$. Using the matrix $\mathcal{D}^2_m$ we define the local set $\varUpsilon_\circ(\mathtt{r}) \subset \varUpsilon_m \subseteq \mathbb{R}^{p^*}$ with some $\mathtt{r} \ge 0$:

$$\varUpsilon_\circ(\mathtt{r}) \overset{\text{def}}{=} \bigl\{\boldsymbol{\upsilon} = (\boldsymbol{\theta},\boldsymbol{\eta}) \in \varUpsilon_m \colon \|\mathcal{D}_m(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^*_m)\| \le \mathtt{r}\bigr\}. \tag{4.1}$$

We introduce $\widetilde{\boldsymbol{\upsilon}}_{\boldsymbol{\theta}^*_m} \in \varUpsilon$, which maximizes $\mathcal{L}_m(\boldsymbol{\upsilon})$ subject to $\varPi_0\boldsymbol{\upsilon} = \boldsymbol{\theta}^*_m$:

$$\widetilde{\boldsymbol{\upsilon}}_{\boldsymbol{\theta}^*_m} \overset{\text{def}}{=} (\boldsymbol{\theta}^*_m, \widetilde{\boldsymbol{\eta}}_{\boldsymbol{\theta}^*_m}) \overset{\text{def}}{=} \underset{\substack{\boldsymbol{\upsilon}\in\varTheta \\ \varPi_0\boldsymbol{\upsilon}=\boldsymbol{\theta}^*_m}}{\operatorname{argmax}}\, \mathcal{L}(\boldsymbol{\upsilon}),$$

and define the radius $\mathtt{r}_0 > 0$

$$\mathtt{r}_0(\mathtt{x}) \overset{\text{def}}{=} \inf_{\mathtt{r}>0}\bigl\{\mathbb{P}(\widetilde{\boldsymbol{\upsilon}}, \widetilde{\boldsymbol{\upsilon}}_{\boldsymbol{\theta}^*_m} \in \varUpsilon_\circ(\mathtt{r})) \ge 1 - \mathrm{e}^{-\mathtt{x}}\bigr\}, \tag{4.2}$$

which is set to infinity if $\widetilde{\boldsymbol{\upsilon}} = \emptyset$ or $\widetilde{\boldsymbol{\upsilon}}_{\boldsymbol{\theta}^*} = \emptyset$. Under the conditions $(\mathcal{L}\mathbf{r})$ and $(\mathcal{E}\mathbf{r})$ Theorem 2.3 of [2] states that $\mathtt{r}_0 = \mathtt{r}_0(\mathtt{x}) \approx \mathtt{C}\sqrt{\mathtt{x} + p^*} > 0$. Further introduce the projected gradient and the covariance of the projected score

$$\breve{\nabla}_{\boldsymbol{\theta}} = \nabla_{\boldsymbol{\theta}} - \mathrm{A}_m\mathcal{H}^{-2}_m\nabla_{\boldsymbol{\eta}}, \qquad \breve{V}^2 = n\operatorname{Cov}(\breve{\nabla}_{\boldsymbol{\theta}}\ell(\boldsymbol{\upsilon}^\circ)).$$

Finally we define

$$\breve{\boldsymbol{\xi}} \overset{\text{def}}{=} \frac{1}{\sqrt{n}}\breve{d}^{-2}\breve{\nabla}\mathcal{L}(\boldsymbol{\upsilon}^*), \qquad \breve{\boldsymbol{\xi}}_m \overset{\text{def}}{=} \frac{1}{\sqrt{n}}\breve{d}^{-2}_m\breve{\nabla}\mathcal{L}_m(\boldsymbol{\upsilon}^*_m).$$

A sufficient list of conditions

The following three conditions ensure that $\mathcal{D}_m^2$ is not degenerated and further quantify the smoothness properties on $\varUpsilon_{\circ}(\mathtt{r})$ of the expected value $\mathbb{E}\mathcal{L}(\boldsymbol{\upsilon})$ and of the stochastic component $\zeta_m(\boldsymbol{\upsilon}) = \mathcal{L}_m(\boldsymbol{\upsilon}) - \mathbb{E}_{\varepsilon}\mathcal{L}_m(\boldsymbol{\upsilon})$.

Represent

$$\mathcal{D}_m^2(\boldsymbol{\upsilon}) = \begin{pmatrix} \mathrm{D}^2(\boldsymbol{\upsilon}) & \mathrm{A}_m(\boldsymbol{\upsilon}) \\ \mathrm{A}_m^{\top}(\boldsymbol{\upsilon}) & \mathcal{H}_m^2(\boldsymbol{\upsilon}) \end{pmatrix}.$$

$(\boldsymbol{\mathcal{I}})$ It holds for some $\rho < 1$

$$\|\mathrm{D}^{-1}\mathrm{A}_m^{\top}\mathcal{H}_m^{-1}\| \le \rho.$$

$(\breve{\boldsymbol{\mathcal{L}}}_{\mathbf{0}})$ For each $\mathtt{r} \le 4\mathtt{r}_0$, there is a constant $\delta(\mathtt{r})$ such that it holds on the set $\varUpsilon_{\circ}(\mathtt{r})$:

$$\|\mathrm{D}^{-1}\mathrm{D}^2(\boldsymbol{\upsilon})\mathrm{D}^{-1} - I_p\| \le \breve{\delta}(\mathtt{r}),$$
$$\|\mathrm{D}^{-1}(\mathrm{A}_m(\boldsymbol{\upsilon}) - \mathrm{A}_m)\mathcal{H}_m^{-1}\| \le \breve{\delta}(\mathtt{r})$$
$$\left\|\mathrm{D}^{-1}\mathrm{A}_m\mathcal{H}_m^{-1}\left(I_m - \mathcal{H}_m^{-1}\mathcal{H}_m^2(\boldsymbol{\upsilon})\mathcal{H}_m^{-1}\right)\right\| \le \breve{\delta}(\mathtt{r}).$$

$(\breve{\boldsymbol{\mathcal{E}}}\boldsymbol{\mathcal{D}}_{\mathbf{1}})$ $\zeta(\boldsymbol{\upsilon}) \to \zeta(\boldsymbol{\upsilon}')$ as $\boldsymbol{\upsilon} \to \boldsymbol{\upsilon}'$. Further for all $0 < \mathtt{r} < 4\mathtt{r}_0$, there exists a constant $\omega \le 1/2$ such that for all $|\mu| \le \breve{\mathfrak{g}}$ and $\boldsymbol{\upsilon}, \boldsymbol{\upsilon}' \in \varUpsilon_{\circ}(\mathtt{r})$

$$\sup_{\boldsymbol{\upsilon},\boldsymbol{\upsilon}' \in \varUpsilon_{\circ}(\mathtt{r})} \sup_{\|\boldsymbol{\gamma}\| \le 1} \log \mathbb{E}_{\varepsilon} \exp\left\{ \frac{\mu}{\breve{\omega}} \frac{\boldsymbol{\gamma}^{\top}\frac{1}{\sqrt{n}}\breve{d}^{-1}\{\breve{\nabla}_{\boldsymbol{\theta}}\boldsymbol{\zeta}(\boldsymbol{\upsilon}) - \breve{\nabla}_{\boldsymbol{\theta}}\boldsymbol{\zeta}(\boldsymbol{\upsilon}')\}}{\|\mathcal{D}_m(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}')\|} \right\} \le \frac{\breve{\nu}_1^2\mu^2}{2}.$$

$(\breve{\boldsymbol{\mathcal{E}}}\boldsymbol{\mathcal{D}}_{\mathbf{0}})$ There exist a matrix $\breve{V}^2 \in \mathbb{R}^{p \times p}$, constants $\nu_0 > 0$ and $\breve{\mathfrak{g}} > 0$ such that for all $|\mu| \le \breve{\mathfrak{g}}$

$$\sup_{\boldsymbol{\gamma} \in \mathbb{R}^p} \log \mathbb{E}_{\varepsilon} \exp\left\{ \mu \frac{\langle \breve{\nabla}_{\boldsymbol{\theta}}\zeta(\boldsymbol{\upsilon}^{\circ}), \boldsymbol{\gamma}\rangle}{\|\breve{V}\boldsymbol{\gamma}\|} \right\} \le \frac{\breve{\nu}_0^2\mu^2}{2}.$$

Remark 4.1. Please see [2] for a discussion and explanation of the above conditions.

Stronger conditions for the full model

In many situations the following, stronger conditions, are easier to verify and allow to derive more accurate results:

$(\boldsymbol{\mathcal{L}}_{\mathbf{0}})$ For each $\mathtt{r} \le \mathtt{r}_0$, there is a constant $\delta(\mathtt{r})$ such that it holds on the set $\varUpsilon_{\circ}(\mathtt{r})$:

$$\left\|\mathcal{D}_m^{-1}\left\{\nabla^2\mathbb{E}\mathcal{L}_m(\boldsymbol{\upsilon})\right\}\mathcal{D}_m^{-1} - \mathbb{I}_{p^*}\right\| \le \delta(\mathtt{r}).$$

($\boldsymbol{\mathcal{E}}\mathbf{\mathcal{D}_1}$) There exists a constant $\omega \le 1/2$, such that for all $|\mu| \le \mathtt{g}$ and all $0 < \mathtt{r} < \mathtt{r}_0$

$$\sup_{\boldsymbol{\upsilon},\boldsymbol{\upsilon}' \in \varUpsilon_{\circ}(\mathtt{r})} \sup_{\|\boldsymbol{\gamma}\|=1} \log \mathbb{E}_{\varepsilon} \exp\left\{ \frac{\mu\, \boldsymbol{\gamma}^{\top} \mathcal{D}_m^{-1} \{\nabla \boldsymbol{\zeta}(\boldsymbol{\upsilon}) - \nabla \boldsymbol{\zeta}(\boldsymbol{\upsilon}')\}}{\omega\, \|\mathcal{D}_m(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}')\|} \right\} \le \frac{\nu_1^2 \mu^2}{2}.$$

($\boldsymbol{\mathcal{E}}\mathbf{\mathcal{D}_0}$) There exist a matrix $\mathcal{V}_0^2 \in \mathbb{R}^{p^* \times p^*}$, constants $\nu_0 > 0$ and $\mathtt{g} > 0$ such that for all $|\mu| \le \mathtt{g}$

$$\sup_{\boldsymbol{\gamma} \in \mathbb{R}^{p^*}} \log \mathbb{E}_{\varepsilon} \exp\left\{ \mu \frac{\langle \nabla \zeta(\boldsymbol{\upsilon}^{\circ}), \boldsymbol{\gamma} \rangle}{\|\mathcal{V}\boldsymbol{\gamma}\|} \right\} \le \frac{\nu_0^2 \mu^2}{2}.$$

The following lemma shows, that these conditions imply the weaker ones from above:

Lemma 4.1 (Lemma 2.1 of [2])**.** *Assume* $(\mathcal{I})$*. Then* $(\mathcal{E}\mathcal{D}_1)$ *implies* $(\breve{\mathcal{E}}\mathcal{D}_1)$, $(\mathcal{L}_0)$ *implies* $(\breve{\mathcal{L}}_0)$*, and* $(\mathcal{E}\mathcal{D}_0)$ *implies* $(\breve{\mathcal{E}}\mathcal{D}_0)$ *with*

$$\breve{\mathtt{g}} = \frac{\sqrt{1-\rho^2}}{(1+\rho)\sqrt{1+\rho^2}}\mathtt{g},\ \breve{\nu}_i = \frac{(1+\rho)\sqrt{1+\rho^2}}{\sqrt{1-\rho^2}}\nu_i,\ \breve{\delta}(\mathtt{r}) = \delta(\mathtt{r}),\quad \textit{and } \breve{\omega} = \omega.$$

Conditions to ensure concentration of the ME

Finally we present two conditions that allow a specific approach to determine the radius $\mathtt{r}_0(\mathtt{x}) > 0$ from (4.2). These conditions have to be satisfied on the whole set $\varUpsilon \subseteq \mathbb{R}^{p^*}$.

($\boldsymbol{\mathcal{L}}\mathtt{r}$) For any $\mathtt{r} > \mathtt{r}_0$ there exists a value $\mathtt{b}(\mathtt{r}) > 0$, such that

$$\frac{-\mathbb{E}\mathcal{L}(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^{\circ})}{\|\mathcal{D}_m(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^{\circ})\|^2} \ge \mathtt{b}(\mathtt{r}), \qquad \boldsymbol{\upsilon} \in \varUpsilon_{\circ}(\mathtt{r}).$$

($\boldsymbol{\mathcal{E}}\mathtt{r}$) For any $\mathtt{r} \ge \mathtt{r}_0$ there exists a constant $\mathtt{g}(\mathtt{r}) > 0$ such that

$$\sup_{\boldsymbol{\upsilon} \in \varUpsilon_{\circ}(\mathtt{r})} \sup_{\mu \le \mathtt{g}(\mathtt{r})} \sup_{\boldsymbol{\gamma} \in \mathbb{R}^{p^*}} \log \mathbb{E}_{\varepsilon} \exp\left\{ \mu \frac{\langle \nabla \zeta(\boldsymbol{\upsilon}), \boldsymbol{\gamma} \rangle}{\|\mathcal{D}_m \boldsymbol{\gamma}\|} \right\} \le \frac{\nu_{\mathtt{r}}^2 \mu^2}{2}.$$

4.2. General results for profile M-estimators

[2] define for some $\mathtt{x}, \mathtt{r} > 0$ the semiparametric spread

$$\breve{\diamondsuit}(\mathtt{r}, \mathtt{x}) \overset{\text{def}}{=} 4\left(\frac{4}{(1-\rho^2)^2} \breve{\delta}(4\mathtt{r}) + 6\nu_1 \breve{\omega} \mathfrak{z}_1(\mathtt{x}, 2p^* + 2p) \right) \mathtt{r}. \tag{4.3}$$

Remark 4.2. The constant $\mathfrak{z}_1(\mathtt{x}, \cdot)$ is of order of $\sqrt{\mathtt{x} + \cdot}$. For a precise definition see Appendix C of [2].

[2] present the following three results, that we adapted for the regression setup. They can be proved in exactly the same way:

Theorem 4.2 ([2], Theorem 2.3)**.** *Suppose that on some set* $\boldsymbol{\mathcal{N}}(\mathtt{x}) \subset \varOmega$ *the condition*$(\mathcal{E}\mathtt{r})$ *and* $(\mathcal{L}\mathtt{r})$ *with* $\mathtt{b}(\mathtt{r}) \equiv \mathtt{b}$ *is met. Further define the following random set*

$$\varUpsilon(K) \overset{\text{def}}{=} \{\boldsymbol{\upsilon} \in \varUpsilon : \mathcal{L}(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^*) \geq -K\}.$$

If for a fixed $\mathtt{r}_0$ *and any* $\mathtt{r} \geq \mathtt{r}_0$*, the following conditions are fulfilled:*

$$\begin{aligned} 6\nu_{\mathtt{r}} \sqrt{\mathtt{x} + 2p^* + \frac{\mathtt{b}}{9\nu_{\mathtt{r}}^2} K} &\leq \mathtt{r}\mathtt{b}, \\ 1 + \sqrt{\mathtt{x} + 2p^*} &\leq 3\nu_{\mathtt{r}}^2 \mathtt{g}(\mathtt{r})/\mathtt{b}, \end{aligned} \tag{4.4}$$

then

$$\mathbb{P}(\varUpsilon(K) \subseteq \varUpsilon_\circ(\mathtt{r}_0)) \geq 1 - \mathrm{e}^{-\mathtt{x}} - \mathbb{P}(\boldsymbol{\mathcal{N}}(\mathtt{x})^c).$$

Theorem 4.3 (Theorem 2.2 of [2])**.** *Assume* $(\breve{\mathcal{L}}_0)$ *and* $(\mathcal{I})$*. Further assume that on some set* $\boldsymbol{\mathcal{N}}(\mathtt{x}) \subset \varOmega$ *the condition* $(\breve{\mathcal{E}}\mathcal{D}_1)$ *is met. Further assume that on* $\boldsymbol{\mathcal{N}}(\mathtt{x}) \subset \varOmega$ *the sets of maximizers* $\widetilde{\boldsymbol{\upsilon}}$, $\widetilde{\boldsymbol{\upsilon}}_{\boldsymbol{\theta}^*}$ *are not empty and that it contains with some* $\tau(\cdot) \in \mathbb{R}$ *the set*

$$\left\{ \sup_{\boldsymbol{\upsilon} \in \varUpsilon_\circ(\mathtt{r}_0)} \|\nabla(\mathbb{E} - \mathbb{E}_\varepsilon)[\mathcal{L}(\boldsymbol{\upsilon}_m^*) - \mathcal{L}(\boldsymbol{\upsilon})]\| \leq \tau(\mathtt{r}_0) \right\} \cap \{\widetilde{\boldsymbol{\upsilon}},\, \widetilde{\boldsymbol{\upsilon}}_{\boldsymbol{\theta}^*} \in \varUpsilon_\circ(\mathtt{r}_0)\}.$$

Then it holds on a set of probability greater $1 - \mathrm{e}^{-\mathtt{x}} - \mathbb{P}(\boldsymbol{\mathcal{N}}(\mathtt{x})^c)$

$$\begin{aligned} \left\| \sqrt{n}\breve{d}(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) - \breve{d}\breve{\boldsymbol{\xi}} \right\| &\leq \breve{\diamondsuit}(\mathtt{r}_0, \mathtt{x}) + \tau(\mathtt{r}_0), \\ \left| 2\breve{L}(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}^*) - \|\breve{d}\breve{\boldsymbol{\xi}}\|^2 \right| &\leq 5\left(\|\breve{\boldsymbol{\xi}}\| + \breve{\diamondsuit}(\mathtt{r}_0, \mathtt{x}) + \tau(\mathtt{r}_0) \right) \left(\breve{\diamondsuit}(\mathtt{r}_0, \mathtt{x}) + \tau(\mathtt{r}_0) \right), \end{aligned}$$

where the spread $\breve{\diamondsuit}(\mathtt{r}_0, \mathtt{x})$ *is defined in* (4.3) *and where* $\mathtt{r}_0 > 0$ *is defined in* (4.2)*.*

Proposition 4.4 (Proposition 2.4 of [2])**.** *Assume the conditions of Theorem 4.3 and additionally assume* $(\mathcal{L}_0)$ *and that* $(\mathcal{E}\mathcal{D}_1)$ *and* $(\mathcal{E}\mathcal{D}_0)$ *are met on* $\boldsymbol{\mathcal{N}}(\mathtt{x})$*. Then the results of Theorem 4.3 hold with* $\mathtt{r}_1 \leq \mathtt{r}_0$ *instead of* $\mathtt{r}_0$ *and with probability greater* $1 - 4\mathrm{e}^{-\mathtt{x}} - \mathbb{P}(\boldsymbol{\mathcal{N}}(\mathtt{x})^c)$ *where*

$$\mathtt{r}_1 \leq \mathfrak{z}(\mathtt{x}, \mathbb{B}) + \diamondsuit_Q(R_0, \mathtt{x}) \wedge \mathtt{r}_0(\mathtt{x}),$$

where $\diamondsuit_Q(\mathtt{r}, \mathtt{x})$ *is of similar order as* $\diamondsuit(\mathtt{r}, \mathtt{x})$ *defined in* (4.3)*. Further if there is some* $\epsilon > 0$ *such that* $\delta(\mathtt{r})/\mathtt{r} \vee 6\nu_1\omega \leq \epsilon$ *for all* $\mathtt{r} \leq \mathtt{r}_0$ *and with* $6\epsilon\mathtt{r}_0(\mathtt{x}) < c$

and $6\epsilon\mathtt{r}_0(\mathtt{x}) < 1$ *then* $\mathtt{r}_0$ *can be replaces with* $\mathtt{r}_0^*$ *which is bounded by*

$$\mathtt{r}_0^* \le \mathfrak{z}(\mathtt{x},\mathbb{B}) + \epsilon\mathfrak{z}_Q(\mathtt{x},4p^*)^2 + \epsilon^2\frac{18}{1-c}\mathfrak{z}_\epsilon(\mathtt{x}).$$

Remark 4.3. The constant $\mathfrak{z}(\mathtt{x},\mathbb{B})$ is of order of $\sqrt{\mathtt{x}+p^*}$. For a precise definition see Appendix A of [2]. Similarly the constant $\mathfrak{z}_Q(\mathtt{x},4p^*)$ is of order of $\sqrt{\mathtt{x}+p^*}$. For a precise definition see the supplement of [19].

Remark 4.4. This is a slightly refined version of Proposition 2.4 of [2], that can be derived using arguments that are similar to those underlying Theorem 2.4 of [3].

4.3. A way to bound the sieve bias

Theorem 4.3 involves two kinds of bias once it is applied to the sieve estimator $\widetilde{\boldsymbol{\theta}}_m$: one that concerns the difference $\boldsymbol{\theta}_m^* - \boldsymbol{\theta}^* \in \mathbb{R}^p$ and the other the difference between $\breve{d}_m(\boldsymbol{\upsilon}_m^*) \in \mathbb{R}^{p\times p}$ and $\breve{d}(\boldsymbol{\upsilon}^*) \in \mathbb{R}^{p\times p}$ where [2] combined with [1] present the following conditions to control these biases:

Represent $\boldsymbol{\upsilon} = (\mathit{\Pi}_{p^*}\boldsymbol{\upsilon}, \boldsymbol{\kappa}) \in \mathbb{R}^{p^*} \times l^2$ and

$$-\nabla^2\mathbb{E}\mathcal{L}(\boldsymbol{\upsilon}) = \begin{pmatrix} \mathcal{D}_m^2(\boldsymbol{\upsilon}) & A_{\boldsymbol{\upsilon}_m\boldsymbol{\kappa}}(\boldsymbol{\upsilon}) \\ A_{\boldsymbol{\upsilon}\boldsymbol{\kappa}}(\boldsymbol{\upsilon})^\top & \mathcal{H}^2(\kappa\kappa) \end{pmatrix} : \mathbb{R}^{p^*} \times l^2 \to \mathbb{R}^{p^*} \times l^2.$$

$(\boldsymbol{\kappa})$ The vector $\boldsymbol{\kappa}^* \overset{\text{def}}{=} (Id_{l^2} - \mathit{\Pi}_{p^*})\boldsymbol{\upsilon}^* \in l^2$ satisfies $\|\mathcal{H}_{\boldsymbol{\kappa}\boldsymbol{\kappa}}\boldsymbol{\kappa}^*\|^2 \le \mathtt{C}_{\boldsymbol{\kappa}^*} m$ for some $\mathtt{C}_{\boldsymbol{\kappa}^*} > 0$ and with $\alpha(m) \to 0$

$$\|\mathcal{D}_m^{-1} A_{\boldsymbol{\kappa}\boldsymbol{\upsilon}_m}^\top \boldsymbol{\kappa}^*\| \le \widehat{\alpha}(m).$$

Further for any $\lambda \in [0,1]$ with some $\tau(m) \to 0$

$$\|\mathcal{D}_m^{-1}\left(\nabla_{\boldsymbol{\upsilon}_m\boldsymbol{\kappa}}\mathbb{E}\mathcal{L}(\mathit{\Pi}_{p^*}\boldsymbol{\upsilon}^*, \lambda\boldsymbol{\kappa}^*) - A_{\boldsymbol{\kappa}\boldsymbol{\upsilon}_m}^\top\right)\boldsymbol{\kappa}^*\| \le \tau(m),$$
$$\left|{\boldsymbol{\kappa}^*}^\top(\mathcal{H}_{\boldsymbol{\kappa}\boldsymbol{\kappa}} - \nabla_{\boldsymbol{\kappa}\boldsymbol{\kappa}}\mathbb{E}\mathcal{L}\left((\mathit{\Pi}_{p^*}\boldsymbol{\upsilon}^*, \lambda\boldsymbol{\kappa}^*)\right)\boldsymbol{\kappa}^*\right| \le \mathtt{C}_{\boldsymbol{\kappa}^*} m.$$

$(\boldsymbol{\upsilon}\boldsymbol{\kappa})$ Assume that with some $\beta(m) \to 0$

$$\|\mathcal{H}_{\boldsymbol{\kappa}\boldsymbol{\kappa}}^{-1} A_{\boldsymbol{\kappa}\boldsymbol{\upsilon}_m}^\top \mathcal{D}_m^{-1}\| \le \beta(m).$$

$(\boldsymbol{\mathcal{L}}\mathbf{r}_\infty)$ For any $\mathtt{r} > \mathtt{r}_0$ there exists a value $\mathtt{b}(\mathtt{r}) > 0$, such that

$$\frac{-\mathbb{E}\mathcal{L}(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^*)}{\|\mathcal{D}(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^*)\|^2} \ge \mathtt{b}(\mathtt{r}).$$

(bias″) As $m \to \infty$ with $\| \cdot \|$ denoting the spectral norm

$$\|\breve{d}_m^{-1}(\boldsymbol{\upsilon}_m^*)\breve{\mathfrak{v}}_m^2(\boldsymbol{\upsilon}_m^*)\breve{d}_m^{-1}(\boldsymbol{\upsilon}_m^*) - \breve{d}^{-1}\breve{\mathfrak{v}}^2\breve{d}^{-1}\| \to 0,$$

where $\breve{\mathfrak{v}}_m^2 = \dfrac{1}{n}\operatorname{Cov}(\breve{\nabla}(\boldsymbol{\upsilon}_m^*))$.

For some $\mathtt{r} > 0$ define the set

$$\varUpsilon_{\circ,m}(\mathtt{r}) \overset{\text{def}}{=} \{\boldsymbol{\upsilon} \in \mathbb{R}^{p^*},\ \|\mathcal{D}_m(\boldsymbol{\upsilon}_m^*)(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}_m^*)\|\}.$$

Theorem 4.5 (Corollary 2.8, 2.10 and Theorem 2.9 of [2]; Theorem 2.1 of [1])**.** *Let the condition* $(\mathcal{L}\mathtt{r}_\infty)$ *with* $\mathtt{b}(\mathtt{r}) \equiv \mathtt{b} > 0$, $(\boldsymbol{\kappa})$ *and condition* $(\mathcal{I})$ *from Section 4.1 be satisfied for both* $\mathcal{D}_m(\boldsymbol{\upsilon}^*)$ *and* $\mathcal{D}_m(\boldsymbol{\upsilon}_m^*)$ *and for* $\mathbb{E}\mathcal{L} : l^2 \to \mathbb{R}$. *Set* $\mathtt{r}^{*2} = 4\mathtt{C}_{\boldsymbol{\kappa}^*}^2 m/\mathtt{b}$. *Assume that on some set* $\boldsymbol{\mathcal{N}}(\mathtt{x}) \subset \varOmega$ *and for some* $m_0 \in \mathbb{N}$ *and all* $m \geq m_0$ *the conditions* $(\breve{\mathcal{E}}\mathcal{D}_0)$, $(\breve{\mathcal{E}}\mathcal{D}_1)$ *and* $(\breve{\mathcal{L}}_0)$ *from Section 4.1 are satisfied for all* $m \geq m_0$ *for some* $m_0 \in \mathbb{N}$ *and with* $\mathcal{D}_0^2 = \nabla_{p+m}^2\mathbb{E}\mathcal{L}_m(\boldsymbol{\upsilon}_m^*) \in \mathbb{R}^{p^* \times p^*}$, $\mathcal{V}_0^2 = \operatorname{Cov}[\nabla_{p+m}\mathcal{L}_m(\boldsymbol{\upsilon}_m^*)] \in \mathbb{R}^{p^* \times p^*}$ *and* $\boldsymbol{\upsilon}^\circ = \boldsymbol{\upsilon}_m^* \in \mathbb{R}^{p^*}$ *and for any* $\mathtt{r} \leq \mathtt{r}^* \vee \mathtt{r}_0^\circ$. *Further assume that on* $\boldsymbol{\mathcal{N}}(\mathtt{x}) \subset \varOmega$ *the sets of maximizers* $\widetilde{\boldsymbol{\upsilon}}$, $\widetilde{\boldsymbol{\upsilon}}_{\boldsymbol{\theta}^*}$ *are not empty and that it contains with some* $\tau(\cdot) \in \mathbb{R}$ *the set*

$$\left\{\sup_{\boldsymbol{\upsilon}\in\varUpsilon_\circ(\mathtt{r}_0)} \|\nabla(\mathbb{E} - \mathbb{E}_\varepsilon)[\mathcal{L}(\boldsymbol{\upsilon}_m^*) - \mathcal{L}(\boldsymbol{\upsilon})]\| \leq \tau(\mathtt{r}_0^\circ)\right\}$$
$$\cap\{\widetilde{\boldsymbol{\upsilon}}_m, \widetilde{\boldsymbol{\upsilon}}_{\boldsymbol{\theta}_m^*,m}, \widetilde{\boldsymbol{\upsilon}}_{\boldsymbol{\theta}^*,m} \in \varUpsilon_{0,m}(\mathtt{r}_0^\circ)\}.$$

Then it holds for any $m \geq m_0$ *with probability greater* $1 - \mathrm{e}^{-\mathtt{x}_n} - \mathbb{P}(\boldsymbol{\mathcal{N}}(\mathtt{x})^c)$

$$\begin{aligned}
\left\|\sqrt{n}\big(\widetilde{\boldsymbol{\theta}}_m - \boldsymbol{\theta}^*\big) - \breve{d}_m\breve{\boldsymbol{\xi}}_m(\boldsymbol{\upsilon}_m^*)\right\| &\leq\ \breve{\diamondsuit}(\mathtt{r}_0^\circ, \mathtt{x}) + \tau(\mathtt{r}_0) + \alpha(m),\\
\left|2\breve{L}(\widetilde{\boldsymbol{\theta}}_m, \boldsymbol{\theta}^*) - \|\breve{d}_m\breve{\boldsymbol{\xi}}_m(\boldsymbol{\upsilon}_m^*)\|^2\right| &\leq 5\left(\breve{\diamondsuit}(\mathtt{r}_0^\circ, \mathtt{x}) + \tau(\mathtt{r}_0) + \alpha(m)\right)\\
&\qquad \cdot\left(\|\breve{d}_m\breve{\boldsymbol{\xi}}_m(\boldsymbol{\upsilon}_m^*)\| + \breve{\diamondsuit}(\mathtt{r}_0^\circ, \mathtt{x}) + \tau(\mathtt{r}_0) + \alpha(m)\right),
\end{aligned}$$

where

$$\alpha(m) = \sqrt{\frac{1+\rho^2}{1-\rho^2}}\left(\alpha(m) + \tau(m) + 2\breve{\delta}(2\mathtt{r}^*)\mathtt{r}^*\right).$$

If further the condition $(\boldsymbol{\upsilon\kappa})$ *and* (*bias″*) *are fulfilled and if for any* $\mathtt{r} > 0$

$$\breve{\delta}(\mathtt{r}^*) \to 0, \quad \breve{\delta}_n(\mathtt{r}) \to 0, \quad \mathtt{r}_0(\mathtt{x}) < \infty,$$
$$\mathbb{P}(\boldsymbol{\mathcal{N}}(\mathtt{x})^c) \to 0, \quad \textit{as } \mathtt{x} \to \infty,$$

there is a sequence $m_n \to \infty$ such that as $n \to \infty$

$$\begin{aligned} \sqrt{n}\breve{d}\bigl(\widetilde{\boldsymbol{\theta}}_m - \boldsymbol{\theta}^*\bigr) &\xrightarrow{w} \mathcal{N}(0, \breve{d}^{-1}\breve{v}^2\breve{d}^{-1}), \\ 2\breve{L}(\widetilde{\boldsymbol{\theta}}_m, \boldsymbol{\theta}^*) \quad &\xrightarrow{w} \mathcal{L}(\|\breve{d}\breve{\boldsymbol{\xi}}_\infty\|^2),\ \breve{\boldsymbol{\xi}}_\infty \sim \mathcal{N}(0, \breve{d}^{-2}\breve{v}^2\breve{d}^{-2}). \end{aligned}$$

Remark 4.5. With remark 2.26 of [2] the radius $\mathtt{r}_0^\circ$ which is defined via

$$\mathtt{r}_0^\circ \overset{\text{def}}{=} \inf\left\{\mathtt{r} > 0 / \mathbb{P}\left(\left\{\widetilde{\boldsymbol{\upsilon}}_m, \widetilde{\boldsymbol{\upsilon}}_{\boldsymbol{\theta}^*_m, m}, \widetilde{\boldsymbol{\upsilon}}_{\boldsymbol{\theta}^*, m} \in \Upsilon_{0,m}(\mathtt{r})\right\}\right) > 1 - \mathrm{e}^{-\mathtt{x}}\right\},$$

is of similar order as $\mathtt{r}_0 > 0$ which satisfies

$$\mathtt{r}_0 \overset{\text{def}}{=} \inf\left\{\mathtt{r} > 0 / \mathbb{P}\left(\left\{\widetilde{\boldsymbol{\upsilon}}_m, \widetilde{\boldsymbol{\upsilon}}_{\boldsymbol{\theta}^*_m, m} \in \Upsilon_{0,m}(\mathtt{r})\right\}\right) > 1 - \mathrm{e}^{-\mathtt{x}}\right\}.$$

The later can be determined using the arguments we present in Section 5.2, using Theorem 4.2.

4.4. Convergence results for the alternating procedure

To derive convergence statements for the alternating procedure sketched in Section 2.3 [3] present the following list of conditions that the initial guess (2.3) has to satisfy to allow applying their results.

($\mathbf{A}_1$) With probability greater $1 - \beta_{(\mathbf{A})}(\mathtt{x})$ the initial guess satisfies $\mathcal{L}(\widetilde{\boldsymbol{\upsilon}}_0, \boldsymbol{\upsilon}^*) \geq -\mathrm{K}_0(\mathtt{x})$ for some $\mathrm{K}_0(\mathtt{x}) \geq 0$.

($\mathbf{A}_2$) The conditions $(\breve{\mathcal{E}}\mathcal{D}_1)$, $(\breve{\mathcal{L}}_0)$, $(\mathcal{E}\mathcal{D}_1)$ and $(\mathcal{L}_0)$ from Section 4.1 hold for all $\mathtt{r} \leq \mathrm{R}_0(\mathtt{x}, \mathrm{K}_0)$ where

$$\mathrm{R}_0(\mathtt{x}) \leq \mathtt{C}\sqrt{\mathtt{x} + p^* + \mathrm{K}_0}. \tag{4.5}$$

($\mathbf{A}_3$) There is some $\epsilon > 0$ such that $\tau(\mathtt{r})/\mathtt{r} \vee \delta(\mathtt{r})/\mathtt{r} \vee 12\nu_1\omega \leq \epsilon$ for all $\mathtt{r} \leq \mathrm{R}_0$. Further ($\mathbf{A}_3$)] $\mathrm{K}_0 \in \mathbb{R}$ and $\epsilon > 0$ are small enough to ensure

$$\epsilon\mathtt{C}(\rho)\mathrm{R}_0 < 1,$$

with

$$\mathtt{C}(\rho) \overset{\text{def}}{=} \frac{16\sqrt{2}(1 + \sqrt{\rho})}{(1 - \rho)(1 - \sqrt{\rho})}.$$

($\mathbf{B}_1$) Assume for all $\mathtt{r} \geq \frac{6\nu_0}{\mathtt{b}}\sqrt{\mathtt{x} + 4p^*}$

$$1 + \sqrt{\mathtt{x} + 4p^*} \leq \frac{3\nu_{\mathtt{r}}^2}{\mathtt{b}}\mathtt{g}(\mathtt{r}).$$

Theorem 4.6 (Theorem 2.2 of [3]). *Assume that the conditions* $(\mathcal{L}_0)$ *and* $(\breve{\mathcal{L}}_0)$ *are met. Assume that on some set* $\boldsymbol{\mathcal{N}}(\mathtt{x}) \subset \Omega$ *the conditions* $(\mathcal{E}\mathcal{D}_0)$,$(\mathcal{E}\mathcal{D}_1)$, $(\mathcal{L}_{\mathtt{r}})$, $(\breve{\mathcal{E}}\mathcal{D}_1)$ *and* $(\mathcal{E}\mathtt{r})$ *of Section 4.1 are met with a constant* $\mathtt{b}(\mathtt{r}) \equiv \mathtt{b}$ *and where* $\mathcal{V}_0^2 = \operatorname{Cov}\left(\nabla\mathcal{L}(\boldsymbol{\upsilon}^*)\right)$, $\mathcal{D}_0^2 = -\nabla^2\mathbb{E}\mathcal{L}(\boldsymbol{\upsilon}^*)$ *and where* $\boldsymbol{\upsilon}^\circ = \boldsymbol{\upsilon}^*$. *Further assume that on* $\boldsymbol{\mathcal{N}}(\mathtt{x}) \subset \Omega$ *the sets* $(\widetilde{\boldsymbol{\upsilon}}^{(k,k(+1))})$ *are not empty and that it contains the set*

$$\bigcap_{\mathtt{r} \le R_0} \left\{ \sup_{\boldsymbol{\upsilon} \in \Upsilon_\circ(\mathtt{r})} \|\nabla(\mathbb{E} - \mathbb{E}_\varepsilon)[\mathcal{L}(\boldsymbol{\upsilon}_m^*) - \mathcal{L}(\boldsymbol{\upsilon})]\| \le \tau(\mathtt{r}) \right\}$$

$$\cap \{(\widetilde{\boldsymbol{\upsilon}}^{(k,k(+1))}) \subset \Upsilon_{0,m}(R_0)\}.$$

Further assume (B_1) *and that the initial guess satisfies* (A_1) *and* (A_2). *Then it holds with probability greater* $1 - 8\mathrm{e}^{-\mathtt{x}} - \beta_{(\mathbf{A})} - \mathbb{P}(\boldsymbol{\mathcal{N}}(\mathtt{x})^c)$ *for all* $k \in \mathbb{N}$

$$\left\| \breve{D}\left(\widetilde{\boldsymbol{\theta}}_k - \boldsymbol{\theta}^*\right) - \breve{\boldsymbol{\xi}} \right\| \le \breve{\Diamond}_Q(\mathtt{r}_k, \mathtt{x}) + \tau(\mathtt{r}_k),$$

$$\left|\max_{\boldsymbol{\eta}} \mathcal{L}(\widetilde{\boldsymbol{\theta}}_k, \boldsymbol{\eta}) - \max_{\boldsymbol{\eta}} \mathcal{L}(\boldsymbol{\theta}^*, \boldsymbol{\eta}) - \|\breve{\boldsymbol{\xi}}\|^2/2\right| \le 5\left(\|\breve{\boldsymbol{\xi}}\| + \breve{\Diamond}_Q(\mathtt{r}_k, \mathtt{x}) + \tau(\mathtt{r}_k) \right)$$
$$(\breve{\Diamond}_Q(\mathtt{r}_k, \mathtt{x}) + \tau(\mathtt{r}_k)),$$

where

$$\mathtt{r}_k \le \mathtt{C}\left(\sqrt{p^* + \mathtt{x}} + \rho^k R_0 \right),$$

with a constant $\mathtt{C}$ *that depends on* $\rho < 1$ *and* $1 - \mathtt{C}(\rho)\epsilon R_0 > 0$. $\breve{\Diamond}_Q(\mathtt{r}, \mathtt{x})$ *is of similar order as* $\breve{\Diamond}(\mathtt{r}, \mathtt{x})$ *defined in* (4.3)

[3] also present a result that shows under which conditions the sequence of estimators $(\widetilde{\boldsymbol{\upsilon}}^{(k,k(+1))})$ actually converges to the maximizer $\widetilde{\boldsymbol{\upsilon}}$. For this result consider the following condition.

$(\boldsymbol{\mathcal{E}\mathcal{D}_2})$ There exists a constant $\omega \le 1/2$, such that for all $|\mu| \le \mathtt{g}$ and all $0 < \mathtt{r} < \mathtt{r}_0$

$$\sup_{\boldsymbol{\upsilon}, \boldsymbol{\upsilon}' \in \Upsilon_\circ(\mathtt{r})} \sup_{\|\boldsymbol{\gamma}_1\| = 1} \sup_{\|\boldsymbol{\gamma}_2\| = 1} \log \mathbb{E} \exp\left\{ \frac{\mu \boldsymbol{\gamma}_1^\top \mathcal{D}^{-1}\{\nabla^2\boldsymbol{\zeta}(\boldsymbol{\upsilon}) - \nabla^2\boldsymbol{\zeta}(\boldsymbol{\upsilon}')\}\boldsymbol{\gamma}_2}{\omega_2 \|\mathcal{D}(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}')\|} \right\}$$
$$\le \frac{\nu_2^2 \mu^2}{2}.$$

Define $\mathfrak{z}(\mathtt{x}, \nabla^2\mathcal{L}(\boldsymbol{\upsilon}^*))$ via

$$\mathbb{P}\left\{ \|\mathcal{D}^{-1}\nabla^2\mathcal{L}(\boldsymbol{\upsilon}^*)\| \ge \mathfrak{z}\left(\mathtt{x}, \nabla^2\mathcal{L}(\boldsymbol{\upsilon}^*)\right) \right\} \le \mathrm{e}^{-\mathtt{x}},$$

and $\kappa(\mathtt{x}, \mathrm{R}_0)$ as

$$\kappa(\mathtt{x}, \mathrm{R}_0) \overset{\text{def}}{=} \frac{2\sqrt{2}(1+\sqrt{\rho})}{\sqrt{1-\rho}}\Big[\delta(\mathrm{R}_0) + 9\omega_2\nu_2\|\mathcal{D}^{-1}\|\mathfrak{z}_1(\mathtt{x}, 6p^*)\mathrm{R}_0 \\ + \|\mathcal{D}^{-1}\|\mathfrak{z}\left(\mathtt{x}, \nabla^2\mathcal{L}(\boldsymbol{\upsilon}^*)\right)\Big].$$

Theorem 4.7 (Theorem 2.4 of [3])**.** *Assume that the condition* $(\mathcal{L}_0)$ *is met. Assume that on some set* $\boldsymbol{\mathcal{N}}(\mathtt{x}) \subset \Omega$ *the conditions* $(\mathcal{E}\mathcal{D}_0)$,$(\mathcal{E}\mathcal{D}_1)$, $(\mathcal{L}_{\mathtt{r}})$ *and* $(\mathcal{E}\mathtt{r})$ *of Section 4.1 are met with a constant* $\mathtt{b}(\mathtt{r}) \equiv \mathtt{b}$ *and where* $\mathcal{V}_0^2 = \operatorname{Cov}\big(\nabla\mathcal{L}(\boldsymbol{\upsilon}^*)\big)$, $\mathcal{D}_0^2 = -\nabla^2\mathbb{E}\mathcal{L}(\boldsymbol{\upsilon}^*)$ *and where* $\boldsymbol{\upsilon}^\circ = \boldsymbol{\upsilon}^*$. *Furthermore, assume that on* $\boldsymbol{\mathcal{N}}(\mathtt{x}) \subset \Omega$ *the sets* $(\widetilde{\boldsymbol{\upsilon}}^{(k,k(+1))})$ *are not empty and that it contains the set*

$$\bigcap_{\mathtt{r}\le R_0}\left\{\sup_{\boldsymbol{\upsilon}\in\Upsilon_\circ(\mathtt{r})}\|\nabla(\mathbb{E}-\mathbb{E}_\varepsilon)[\mathcal{L}(\boldsymbol{\upsilon}_m^*)-\mathcal{L}(\boldsymbol{\upsilon})]\| \le \tau(\mathtt{r})\right\} \\ \cap\{(\widetilde{\boldsymbol{\upsilon}}^{(k,k(+1))}) \subset \Upsilon_{0,m}(R_0)\}.$$

Suppose (B_1) *and that the initial guess satisfies* (A_1) *and* (A_2). *Assume that* $\kappa(\mathtt{x}, R_0) < (1-\rho)$. *Then*

$$\mathbb{P}\left(\bigcap_{k\in\mathbb{N}}\left\{\left\|\mathcal{D}(\widetilde{\boldsymbol{\upsilon}}^{(k,k(+1))} - \boldsymbol{\upsilon}^*)\right\| \le \mathtt{r}_k^*\right\}\right) \ge 1 - 3\mathrm{e}^{-\mathtt{x}} - \beta_{(\mathbf{A})},$$

where

$$\mathtt{r}_k^* \le \begin{cases}\rho^k\frac{4\sqrt{2}}{1-\kappa(\mathtt{x},R_0)k}R_0, & \kappa(\mathtt{x}, R_0)k \le 1,\\ \rho^{\frac{k}{\log(k)}\log\left(\frac{1-\rho}{\kappa(\mathtt{x},R_0)}\right)}c_kR_0, & \textit{otherwise},\end{cases}$$

with some sequence $(c_k) \in \mathbb{N}$, *where* $0 < c_k \to 2$.

5. Application of the finite sample theory

We will now apply the results presented in the previous section to our problem. First we will show that the conditions $(\mathcal{E}\mathcal{D}_0)$, $(\mathcal{E}\mathcal{D}_1)$, $(\mathcal{L}_0)$, $(\mathcal{I})$, of Section 4.1 can be satisfied under the assumptions $(\mathcal{A})$. These imply – by Lemma 4.1 - $(\breve{\mathcal{L}}_0)$, $(\breve{\mathcal{E}}\mathcal{D}_1)$ and $(\breve{\mathcal{E}}\mathcal{D}_0)$ from Section 4.1, necessary for Theorem 2.2 of [2]. Further we will show that the conditions $(\mathcal{E}\mathtt{r})$ and $(\mathcal{L}\mathtt{r})$ from 4.1 are met. This will allow to determine $\mathtt{r}_0 > 0$ and ensure that the sets of maximizers $\widetilde{\boldsymbol{\upsilon}}_m$, $\widetilde{\boldsymbol{\upsilon}}_{m\boldsymbol{\theta}^*}$ are not empty. The subsequent analysis will then serve to determine the necessary size of $n \in \mathbb{N}$ that allows to obtain good bounds for $\breve{\diamondsuit}(\mathtt{r}_0, \mathtt{x}) \in \mathbb{R}$. Concerning the alternation procedure we will show that the initial guess from (2.3) and the values of $\delta(\mathtt{r}),\omega$ from $(\breve{\mathcal{L}}_0)$, $(\breve{\mathcal{E}}\mathcal{D}_1)$ allow to apply the Theorems 4.6 and 4.7. Finally we present the proof of Proposition 2.9. Unfortunately this section is rather technical and is not entirely intelligible without results from Section A.

5.1. Conditions satisfied

In this section we show that the conditions of Section 4.1 are satisfied. First we derive an a priori bound for the distance between the target $\boldsymbol{\upsilon}_m^* \in \mathbb{R}^p \times \mathbb{R}^m$ and the true parameter $\boldsymbol{\upsilon}^* \in \mathbb{R}^p \times l^2$

Lemma 5.1. *Assume* $(\mathcal{A})$ *then there is a constant* $\mathtt{C} > 0$ *such that we get* $\|\mathcal{D}_m(\boldsymbol{\upsilon}_m^* - \boldsymbol{\upsilon}^*)\| \le \mathtt{r}^*$ *with*

$$\mathtt{r}^* = \mathtt{C}\sqrt{n}m^{-(1+2\alpha)/2}\sqrt{m}. \tag{5.1}$$

The next step is to determine a radius $\mathtt{r}^\circ$ that ensures that $\widetilde{\boldsymbol{\upsilon}} \in S_1^{p,+} \times B_{\mathtt{r}^\circ}(0)$ with large probability.

Lemma 5.2. *Define*

$$\widetilde{\boldsymbol{\eta}}_{m,\boldsymbol{\theta}}^{(\infty)} \overset{\text{def}}{=} \operatorname*{argmax}_{\boldsymbol{\eta}\in\mathbb{R}^m} \mathcal{L}_m(\boldsymbol{\theta},\boldsymbol{\eta}),$$

then with some constant $\mathtt{C} \in \mathbb{R}$

$$\mathbb{P}\left(\sup_{\boldsymbol{\theta}\in S_1^{p,+}} \left\|\widetilde{\boldsymbol{\eta}}_{m,\boldsymbol{\theta}}^{(\infty)}\right\| \ge \mathtt{C}\sqrt{p^*\log(p^*)+\mathtt{x}}\right) \le \mathrm{e}^{-\mathtt{x}}.$$

Remark 5.1. This Lemma also ensures that the alternating sequence $(\widetilde{\boldsymbol{\theta}}^{(k)}, \widetilde{\boldsymbol{\eta}}^{(k(+1))})$ introduced in Section 2.3 lies in $S_1^{p,+} \times B_{\mathtt{r}^\circ}^m(0)$, with

$$\mathtt{r}^\circ \overset{\text{def}}{=} \mathtt{C}\sqrt{p^*\log(p^*)+\mathtt{x}}. \tag{5.2}$$

Let $c_{\mathcal{D}} > 0$ denote the smallest eigenvalue $c_{\mathcal{D}} \overset{\text{def}}{=} \lambda_{\min}(\mathcal{D}_m)/\sqrt{n}$ which as shown in Lemma A.8 is bounded away from 0. This also means that

$$\Upsilon_m \subseteq \Upsilon_\circ(\sqrt{n}\mathtt{r}^\circ/c_{\mathcal{D}}) \overset{\text{def}}{=} \left\{\boldsymbol{\upsilon} \in \Upsilon :\, \|\mathcal{D}_m(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}_m^*)\| \le \sqrt{n}\mathtt{r}^\circ/c_{\mathcal{D}}\right\}.$$

Now we show that the general conditions of Section 4.1 are met under the assumptions $(\mathcal{A})$. For this we point out again that due to the random design regression approach we define the random component of $\mathcal{L}$ via $\mathcal{L} - \mathbb{E}_\varepsilon\mathcal{L}$ where $\mathbb{E}_\varepsilon$ denotes the expectation operator of the law of $(\varepsilon_i)_{i=1,\dots,n}$ given $(\mathbf{X}_i)_{i=1,\dots,n}$. This facilitates the proof of the conditions $(\mathcal{E}\mathcal{D}_0)$, $(\mathcal{E}\mathcal{D}_1)$ and $(\mathcal{E}\mathtt{r})$ but leads to additional randomness, in the sense that the claim of the following lemma is only true with a certain high probability.

Lemma 5.3. *Assume the conditions* $(\mathcal{A})$*. Then with* $\boldsymbol{\upsilon}^{\circ} = \boldsymbol{\upsilon}_m^* \in \mathbb{R}^{p^*}$ *and*

$$\mathcal{V}_0^2 = \operatorname{Cov}\left(\nabla \mathcal{L}_m(\boldsymbol{\upsilon}_m^*)\right), \quad \mathcal{D}_0^2 = -\nabla^2 \mathbb{E}\mathcal{L}_m(\boldsymbol{\upsilon}_m^*),$$

and $\mathtt{x} \le m$ *we get the conditions of section 4.1 on the set*

$$\left\{ \sup_{\boldsymbol{\theta} \in S_1^{p,+}} \left\| \widetilde{\boldsymbol{\eta}}_{m,\boldsymbol{\theta}}^{(\infty)} \right\| \le \mathtt{C}\sqrt{p^* \log(p^*) + \mathtt{x}} \right\}.$$

More precisely we get $(\mathcal{I})$ *and:*

$(\mathcal{E}\mathfrak{D}_0)$ *with probability greater than* $1 - \mathrm{e}^{-\mathtt{x}}$ *and with*

$$\mathtt{g} = \sqrt{\frac{n}{\mathtt{C}m}}\widetilde{g}, \qquad \nu_m^2 = 2\widetilde{\nu}^2,$$

$(\mathcal{E}\mathtt{r})$ *with probability greater than* $1 - \mathrm{e}^{-\mathtt{x}}$ *and with*

$$\begin{aligned} \mathtt{g}(\mathtt{r}) \;&=\; \sqrt{n}\widetilde{g}\mathtt{C}^{-1}\left(\sqrt{m} + m^{3/2}\mathtt{r}/\sqrt{n}\right)^{-1}, \\ \nu_{\mathtt{r},m}^2 \;&=\; \widetilde{\nu}^2\Big(1 + \mathtt{C}\left(m^{3/2} + \mathtt{r}m^2/\sqrt{n}\right)\mathtt{r}/\sqrt{n}\Big) \\ &\qquad +\mathtt{C}\left(m + m^3\mathtt{r}^2/n\right)\Big(\mathtt{x} + \log(2m)\Big)^{1/2}/\sqrt{n}\Big). \end{aligned} \tag{5.3}$$

$(\mathcal{E}\mathfrak{D}_1)$ *on* $\Upsilon_{\circ}(\mathtt{r})$ *for all* $\mathtt{r} > 0$ *with* $\mathtt{r}m^2/\sqrt{n} \le 1$ *with probability greater than* $1 - \mathrm{e}^{-\mathtt{x}}$ *and with*

$$\mathtt{g} \ge \frac{\sqrt{n}}{\mathtt{r}m^{3/2}C}, \qquad \omega \overset{\text{def}}{=} \frac{2}{\sqrt{n}}, \qquad \nu_{1,m}^2 = \widetilde{\nu}^2\mathtt{C}m^2.$$

$(\mathcal{L}_0)$ *is satisfied for all* $\mathtt{r} > 0$ *with* $\mathtt{r}m^{3/2}/\sqrt{n} \le 1$ *and where*

$$\delta(\mathtt{r}) = \frac{\mathtt{C}\left\{m^{3/2} + \mathtt{C}_{bias}m^{5/2}\right\}\mathtt{r}}{\sqrt{n}}.$$

$(\mathcal{L}\mathtt{r})$ *if* $\mathtt{C}_{bias} = 0$ *and for* $n \in \mathbb{N}$ *large enough with* $\mathtt{b} = c_{(\mathcal{L}\mathtt{r})} > 0$ *as soon as*

$$\mathtt{r}^2 \ge \mathtt{C}\mathtt{r}^{*2}/\mathtt{b} \vee m, \tag{5.4}$$

and with probability greater than $1 - \exp\left\{-m^3\mathtt{x}\right\} - \exp\left\{-nc_{(\boldsymbol{Q})}\right\}$*, for some* $c_{(\boldsymbol{Q})} > 0$*. In the case that* $\mathtt{C}_{bias} \ne 0$ *we get for*

$$\mathtt{r}^2 \ge \sqrt{\mathtt{x} + \mathtt{C}p^*[\log(p^*) + \log(n)]}/\mathtt{b}_{\mathbb{E}} \vee 2\mathtt{r}^{*2},$$

that with some $\mathtt{b}_{bias} > 0$ independent of $n, m, \mathtt{x}, \mathtt{r}$ and with probability greater than $1 - \mathrm{e}^{-\mathtt{x}}$

$$-\mathbb{E}_\epsilon \mathcal{L}_m(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}_m^*) \ge \mathtt{b}_{bias}\mathtt{r}^2.$$

Finally we apply Lemma 4.1 to obtain the conditions $(\breve{\mathcal{L}}_0)$, $(\breve{\mathcal{E}}\mathcal{D}_1)$ and $(\breve{\mathcal{E}}\mathcal{D}_0)$.

Remark 5.2. The condition $\mathtt{r}m^2/\sqrt{n} \le 1$ needed for $(\boldsymbol{\mathcal{E}\mathcal{D}_1})$ can be relaxed to read $\mathtt{r}m^{3/2}/\sqrt{n} \le 1$ if one increases $\nu_{1,m}^2 = \widetilde{\nu}^2\mathtt{C}m^3$. This does not change the bounds for $\diamondsuit(\mathtt{r},\mathtt{x})$, as $\delta(\mathtt{r})$ will still be of the same order as $\omega\nu_{1,m}$. With this correction the conditions apply for all $\mathtt{r} \le \mathrm{R}_0$, where R_0 is the deviation bound for the elements of the alternation procedure started in $\widetilde{\boldsymbol{\upsilon}}_0$ in (2.3), as we explain in Remark 5.6.

Remark 5.3. We do not show the conditions $(\breve{\mathcal{L}}_0)$, $(\breve{\mathcal{E}}\mathcal{D}_1)$ and $(\breve{\mathcal{E}}\mathcal{D}_0)$ directly. To benefit from the weaker conditions we would need entry-wise bounds for the operator $A\mathcal{H}_m^{-2}$ for better bounds in the proof of condition $(\breve{\mathcal{L}}_0)$. As this work is very long and technical without this sophistication we postpone this improvement to future work.

5.2. Large deviations

Next we determine the necessary size of the radius $\mathtt{r}_0(\mathtt{x})$ defined in (4.2). We want to use Theorem 4.2. We have with Lemma 5.2 combined with Lemma A.16 that condition $(\boldsymbol{\mathcal{E}\mathtt{r}})$ is met with probability $1 - 2\mathrm{e}^{-\mathtt{x}}$ and with (setting $\mathtt{r} = \mathtt{C}\sqrt{n}\sqrt{p^*\log(p^*)}$ in (5.3))

$$\mathtt{g}(\mathtt{r}) = \sqrt{n}\widetilde{g}\mathtt{C}^{-1}\left(\sqrt{m} + m^2\log(p^*)\right)^{-1}, \qquad \nu_{\mathtt{r},m}^2 \le \widetilde{\nu}^2\mathtt{C}m^3\log(p^*)^2.$$

Furthermore due to $\mathtt{r}^* \le \mathtt{C}\sqrt{p^*}$ and for moderate $\mathtt{x} > 0$ we find if

$$\mathtt{r}^2 \ge \begin{cases} \mathtt{C}p^*, & \text{if } \mathtt{C}_{bias} = 0, \\ \mathtt{C}p^*\log(n) & \text{if } \mathtt{C}_{bias} > 0. \end{cases}$$

that with some $\mathtt{b} > 0$

$$\mathbb{P}\left(-\mathbb{E}_\epsilon\mathcal{L}_m(\boldsymbol{\upsilon},\boldsymbol{\upsilon}_m^*) \ge \mathtt{b}\mathtt{r}^2\right) \ge 1 - \mathrm{e}^{-\mathtt{x}} - \exp\left\{-m^3\mathtt{x}\right\} - \exp\left\{-nc_{(\boldsymbol{Q})}/4\right\}.$$

Note that the second condition (4.4) of Theorem 4.2 is automatically satisfied in our setting for $n \in \mathbb{N}$ large enough. Finally we only have to ensure that $\mathtt{r}_0 > 0$ is large enough to satisfy (5.4), then Theorem 4.2 yields the following corollary.

Corollary 5.4. *Consider the set*

$$\mathcal{A} \overset{\text{def}}{=} \{(\boldsymbol{\mathcal{E}\mathtt{r}}) \textit{ and } (\boldsymbol{\mathcal{L}_\mathtt{r}}) \textit{ are met}\} \cap \left\{\sup_{\boldsymbol{\theta}\in S_1^{p,+}} \left\|\widetilde{\boldsymbol{\eta}}_{m,\boldsymbol{\theta}}^{(\infty)}\right\| \le \mathtt{C}\sqrt{p^*\log(p^*)+\mathtt{x}}\right\},$$

Then it holds that

$$\mathbb{P}\left(\mathcal{A} \cap \left\{ \sup_{\boldsymbol{\upsilon} \in \Upsilon_m \setminus \Upsilon_\circ(\mathtt{r}_0^\circ)} \mathcal{L}(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}_m^*) < 0 \right\}\right) \geq 1 - \mathrm{e}^{-\mathtt{x}} - \mathbb{P}(\mathcal{A}^c),$$

where

$$\mathtt{r}_0^\circ \overset{\text{def}}{=} \begin{cases} \mathtt{C} m^{3/2}\sqrt{\mathtt{x} + p^*} & \text{if } \mathtt{C}_{bias} = 0, \\ \mathtt{C}\left(\sqrt{p^* \log(n)} \vee m^{3/2}\sqrt{\mathtt{x} + p^*}\right) & \text{if } \mathtt{C}_{bias} > 0. \end{cases}$$

Repeating the same steps from above gives that on the set

$$\{(\boldsymbol{\mathcal{E}}\mathbf{r}) \text{ and } (\boldsymbol{\mathcal{L}}_{\mathtt{r}}) \text{ are met}\} \cap \left\{ \sup_{\boldsymbol{\theta} \in S_1^{p,+}} \left\|\widetilde{\boldsymbol{\eta}}_{m,\boldsymbol{\theta}}^{(\infty)}\right\| \leq \mathtt{C}\sqrt{p^* \log(p^*) + \mathtt{x}} \right\}$$

$$\cap \left\{ \sup_{\boldsymbol{\upsilon} \in \Upsilon_m \setminus \Upsilon_\circ(\mathtt{r}_0^\circ)} \mathcal{L}(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}_m^*) < 0 \right\}.$$

condition $(\boldsymbol{\mathcal{E}}\mathbf{r})$ is actually met on $\Upsilon_\circ(\mathtt{r}_0^\circ)$ with

$$\mathtt{g}(\mathtt{r}) = \sqrt{n}\widetilde{g}(\mathtt{C}\sqrt{m})^{-1}, \quad \nu_m^2 \leq \mathtt{C}\widetilde{\nu}^2 m,$$

if ${p^*}^5(1 + \mathtt{C}_{bias}\log(n))/n \to 0$. This gives

Corollary 5.5. *Consider the set*

$$\mathcal{B} \overset{\text{def}}{=} \{(\boldsymbol{\mathcal{E}}\mathbf{r}) \textit{ and } (\boldsymbol{\mathcal{L}}_{\mathtt{r}}) \textit{ are met}\} \cap \left\{ \sup_{\boldsymbol{\theta} \in S_1^{p,+}} \left\|\widetilde{\boldsymbol{\eta}}_{m,\boldsymbol{\theta}}^{(\infty)}\right\| \leq \mathtt{C}\sqrt{p^* \log(p^*) + \mathtt{x}} \right\}$$

$$\cap \left\{ \sup_{\boldsymbol{\upsilon} \in \Upsilon_m \setminus \Upsilon_\circ(\mathtt{r}_0^\circ)} \mathcal{L}(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}_m^*) < 0 \right\}.$$

Then it holds that

$$\mathbb{P}\left(\mathcal{B} \cap \left\{ \sup_{\boldsymbol{\upsilon} \in \Upsilon_m \setminus \Upsilon_\circ(\mathtt{r}_0)} \mathcal{L}(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}_m^*) < 0 \right\}\right) \geq 1 - 2\mathrm{e}^{-\mathtt{x}} - \mathbb{P}(\mathcal{A}^c),$$

where

$$\mathtt{r}_0 \leq \begin{cases} \mathtt{C}\sqrt{\mathtt{x} + p^*} & \text{if } \mathtt{C}_{bias} = 0, \\ \mathtt{C}\sqrt{\mathtt{x} + p^* \log(n)} & \text{if } \mathtt{C}_{bias} > 0. \end{cases} \tag{5.5}$$

Remark 5.4. If only ${p^*}^4(1 + \mathtt{C}_{bias}\log(n))/n \to 0$ we would get $\nu_m^2 \leq \mathtt{C}\widetilde{\nu}^2 m^2$ and would have to iterate the above argument once more.

5.3. Proof of finite sample Wilks and Fisher expansion

Combining Lemma 5.3 and Corollary 5.5 we obtain the following bound if $\mathtt{C}_{bias} = 0$ and ${p^*}^4/n \to 0$ and if $n \in \mathbb{N}$ is large enough:

$$\breve{\diamondsuit}(\mathtt{r}_0, \mathtt{x}) \leq \mathtt{C}\frac{{p^*}^{5/2} + \mathtt{x}}{\sqrt{n}}.$$

With these results Proposition 2.1 is merely a corollary of Theorem 4.3 and of Lemma 4.1. More precisely define the set

$$\begin{aligned}\boldsymbol{\mathcal{N}}(\mathtt{x}) \overset{\text{def}}{=} & \left\{ \sup_{\boldsymbol{\theta} \in S_1^{p,+}} \left\| \widetilde{\boldsymbol{\eta}}^{(\infty)}_{m,\boldsymbol{\theta}} \right\| \leq \mathtt{C}\sqrt{p^* \log(p^*) + \mathtt{x}} \right\} \\ & \cap \left\{ \sup_{\boldsymbol{\upsilon} \in \Upsilon_m \setminus \Upsilon_\circ(\mathtt{r}_0^\circ)} \mathcal{L}(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^*_m) < 0 \right\} \\ & \cap \{ \widetilde{\boldsymbol{\upsilon}}_m, \widetilde{\boldsymbol{\upsilon}}_{\boldsymbol{\theta}^*_m, m} \in \{ \| \mathcal{D}_m(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^*_m) \| \leq \mathtt{r}_0 \} \} \\ & \cap \left\{ \sup_{\boldsymbol{\upsilon} \in \Upsilon_\circ(\mathtt{r}_0)} \| \nabla (\mathbb{E} - \mathbb{E}_\varepsilon)[\mathcal{L}(\boldsymbol{\upsilon}^*_m) - \mathcal{L}(\boldsymbol{\upsilon})] \| \leq \mathtt{C}(\mathtt{x} + p^*)^2 \mathtt{r}_0 / \sqrt{n} \right\} \\ & \cap \{ \text{The conditions of Section 4.1 are met for } (\mathcal{L}, \Upsilon_m, \mathcal{D}) \} .\end{aligned}$$

It is of Probability greater $1 - 7\mathrm{e}^{-\mathtt{x}} - \exp\left\{-m^3 \mathtt{x}\right\} - \exp\left\{-n c_{\boldsymbol{(Q)}}/4\right\}$. Finally with the results of Appendix A of [2] on the deviation behavior of quadratic forms we can bound with some constant

$$\mathbb{P}\left(\| \frac{1}{\sqrt{n}} \breve{d}^{-1} \breve{\nabla} \| \leq \mathfrak{z}(\mathtt{x}, \breve{\mathbb{B}}) \right) \geq 1 - 2\mathrm{e}^{-\mathtt{x}}, \quad \mathfrak{z}(\mathtt{x}, \breve{\mathbb{B}}) \leq \sigma \mathtt{C}\sqrt{p^* + \mathtt{x}}.$$

So we get the claim with Theorem 4.3.

For the case that $\mathtt{C}_{bias} > 0$ we want to apply Proposition 4.4. For this define

$$\epsilon \overset{\text{def}}{=} 6\nu_1 \omega \vee \delta(\mathtt{r})/\mathtt{r} \leq \mathtt{C}_\diamond m^{5/2}/\sqrt{n}.$$

Then $\mathtt{r}_0 > 0$ in (5.5) satisfies by assumption

$$6\epsilon \mathtt{r}_0 \to 0.$$

since $m^3 \log(n)/\sqrt{n} \to 0$. Consequently Proposition 4.4 applies with $\boldsymbol{\mathcal{N}}(\mathtt{x})$ from above, which yields the claim of Proposition 2.3 with an error term

$$\mathtt{C}_\diamond (1 + \mathtt{C}_{bias}) \frac{\mathtt{x} + {p^*}^{5/2} {\mathtt{r}_0^*}^2}{\sqrt{n}},$$

where

$$\mathtt{r}_0^* \leq \mathtt{C}\sqrt{p^* + \mathtt{x}}.$$

5.4. Bounding the sieve bias

We prove this claim via showing that the conditions of and Theorem 4.5 are met. For this we need to show that the conditions $(\mathcal{L}\mathtt{r}_{\infty})$ and $(\boldsymbol{\kappa})$ are met. But exactly this is done in Lemma A.6. So we simply have to plug in our estimates.

Finally we determine an admissible rate for $m(n) \in \mathbb{N}$ which ensures that the error terms vanish. We exemplify this for the case $\mathtt{C}_{bias} = 0$. We can show that

$$\breve{\diamondsuit}(\mathtt{r}_0^{\circ}, \mathtt{x}) \le \mathtt{C}(p^* + \mathtt{x})^{5/2}/\sqrt{n}.$$

If $p^{*5/2}/\sqrt{n} \to 0$, we can get that $2(\|\breve{\mathrm{D}}^{-1}\breve{\nabla}\| + \mathtt{r}_p^*(\mathtt{x}_n))\diamondsuit(\mathtt{r}_2, \mathtt{x}_n) \xrightarrow{\mathbb{P}} 0$ by choosing a sequence $\mathtt{x}_n > 0$, that increases slow enough. If $\sqrt{n}m^{-\alpha-1/2} \to 0$ we get the desired result. Clearly such a sequence exists and in this case $\mathbb{P}(\Omega(\mathtt{x}_n)) \to 1$.

For the the weak convergence statements we focus on the case $\mathtt{C}_{bias} = 0$ and use Theorem 4.5. As $\delta(\mathtt{r}), \omega \to 0$ and $\mathtt{r}_0(\mathtt{x}) < \infty$ we further only have to prove condition $\boldsymbol{(bias')}$ which means that we have to bound

$$\|I_{p^*} - \breve{d}_m^{-1}(\boldsymbol{\upsilon}^*)\breve{d}(\boldsymbol{\upsilon}^*)\breve{d}_m^{-1}(\boldsymbol{\upsilon}^*)\| \text{ and } \|I_{p^*} - \breve{d}_m^{-1}(\boldsymbol{\upsilon}_m^*)\breve{d}_m(\boldsymbol{\upsilon}^*)\breve{d}_m^{-1}(\boldsymbol{\upsilon}_m^*)\|.$$

With $(\boldsymbol{\upsilon\kappa})$ – as proven in Lemma A.6 – we can apply Lemma A.4 of [1] to find

$$\|I - \breve{d}_m^{-1}\breve{d}\breve{d}_m^{-1}\| \le \sqrt{\frac{1+\rho^2+m^{-1}}{1-\rho^2}\frac{\mathtt{C}_1^2 m^{-1}}{c_{\mathfrak{D}}^2 - \mathtt{C}_1^2 m^{-1}}} \to 0,$$

and with Lemma A.5 of [1]

$$\begin{aligned}&\|I - \breve{d}_m(\boldsymbol{\upsilon}_m^*)^{-1}\breve{d}_m(\boldsymbol{\upsilon}^*)^2\breve{d}_m(\boldsymbol{\upsilon}_m^*)^{-1}\|\\ &\quad\le \frac{\sqrt{\rho}\left(2+\sqrt{1-\breve{\delta}(\mathtt{r}^*)}\right)+1+\breve{\delta}(\mathtt{r}^*)}{(1-\sqrt{\rho})^2}\breve{\delta}(\mathtt{r}^*) \to 0.\end{aligned}$$

Furthermore we need to satisfy $\boldsymbol{(bias'')}$, which in our setting becomes

$\boldsymbol{(bias'')}$ The i.i.d. random variables $Y_i(m) \in \mathbb{R}^p$ satisfy $\mathrm{Cov}(Y_i(m)) \to 0$ where

$$\begin{aligned}Y_i(m) \overset{\text{def}}{=} \breve{d}_m^{-1}&\{\nabla_{\boldsymbol{\theta}}(\ell_i(\boldsymbol{\upsilon}_m^*) - \ell_i(\boldsymbol{\upsilon}^*))\\ &\qquad -\mathrm{A}_m\mathcal{H}_m^{-2}\nabla_{(\eta_1,\dots,\eta_m)}(\ell_i(\boldsymbol{\upsilon}_m^*) - \ell_i(\boldsymbol{\upsilon}^*))\}.\end{aligned}$$

which is done with Lemma A.26. This completes the proof after plugging in the bounds.

5.5. Convergence of the alternating procedure

Here we want to explain in more detail how the Propositions 2.5, 2.7, 2.6 and 2.8 can be derived.

We want to use Theorem 4.6. For this it remains to check the conditions $(\mathbf{A}_1)$, $(\mathbf{A}_2)$ and $(\mathbf{A}_3)$ from Section 4.4 for the initial guess defined in (2.3).

Remark 5.5. Condition $(\mathbf{B}_1)$ is met in our case as we pointed out in Section 5.2.

We can prove the following lemma:

Lemma 5.6. *It holds for* $\mathtt{x} \le \mathtt{C}\widetilde{\nu}^2\widetilde{\mathtt{g}}^2 n$ *that*

$$\mathbb{P}\left(\mathcal{L}_m(\widetilde{\boldsymbol{\upsilon}}^{(0)}, \boldsymbol{\upsilon}^*_m) \le -\mathtt{C}\left\{(1+\mathtt{C}_{bias}\sqrt{m})n\tau^2 + (1+\mathtt{C}_{bias})\sqrt{\mathtt{x}}\tau\sqrt{n}\right\}\right) \le 2\mathrm{e}^{-\mathtt{x}}.$$

If $\mathtt{C}_{bias} = 0$ *set* $\tau = o(p^{*-3/2})$ *and* $m^4 = o(n)$. *If* $\mathtt{C}_{bias} > 0$ *set* $\tau = o(m^{-9/4})$ *and* $m^6 = o(n)$. *Then the initial radius* $R_0 > 0$ *in* (4.5) *satisfies* $\epsilon R_0 \to 0$ *such that the conditions* $(\mathbf{A}_1)$,$(\mathbf{A}_2)$ *and* $(\mathbf{A}_3)$ *are satisfied for* $n \in \mathbb{N}$ *large enough (as in Lemma 5.3).*

Together with Theorem 4.6 this implies Proposition 2.5 as we can bound

$$\breve{\diamondsuit}_Q(\mathtt{r},\mathtt{x}) \le \mathtt{C}_\diamond \frac{\mathtt{x} + p^{*3/2}\mathtt{r}^2 + \mathtt{C}_{bias}p^{*2}\mathtt{r}^2}{\sqrt{n}}.$$

Remark 5.6. $\epsilon \mathrm{R}_0 \to 0$ implies $\mathrm{R}_0 m^{3/2}/\sqrt{n} \to 0$. As pointed out in Remark 5.2 this means that the conditions from Section 4.1 can be satisfied on $\varUpsilon_\circ(\mathrm{R}_0)$.

For Proposition 2.6 we apply Theorem 4.7. It remains to show condition $(\mathcal{E}\mathcal{D}_2)$ and to bound $\mathfrak{z}(\mathtt{x}, \nabla^2\mathcal{L}(\boldsymbol{\upsilon}^*))$ which is defined via

$$\mathbb{P}\left\{\|\mathcal{D}^{-1}\nabla^2\mathcal{L}(\boldsymbol{\upsilon}^*)\| \ge \mathfrak{z}\left(\mathtt{x}, \nabla^2\mathcal{L}(\boldsymbol{\upsilon}^*)\right)\right\} \le \mathrm{e}^{-\mathtt{x}}.$$

We derive a bound for $\mathfrak{z}(\mathtt{x}, \nabla^2\mathcal{L}(\boldsymbol{\upsilon}^*))$ in Lemma A.31 which is based on Corollary 3.7 of [22], as is proposed in Remark 2.17 of [3]. The claim of Proposition 2.6 is shown with the following Lemma.

Lemma 5.7. *Assume* $(\mathcal{A})$. *Assume further that* $p^{*4}/n \to 0$ *and* $\tau = o(p^{*-3/2})$ *if* $\mathtt{C}_{bias} = 0$ *and* $p^{*6}/n \to 0$ *and* $\tau = o(p^{*-9/4})$ *if* $\mathtt{C}_{bias} > 0$. *Let* $\mathtt{x} > 0$ *be chosen such that*

$$\mathtt{x} \le \frac{1}{2}\left(\widetilde{\nu}^2 n\widetilde{\mathtt{g}}^2 - \log(p^*)\right).$$

then the conditions $(\mathcal{E}\mathcal{D}_2)$, $(\mathcal{L}_0)$, $(\mathcal{L}_{\mathtt{r}})$ *and* $(\mathcal{E}\mathtt{r})$ *are met and* $\kappa(\mathtt{x}, R_0) \to 0$ *with* $n \to \infty$.

Remark 5.7. The bound for $\mathtt{x}$ comes from Lemma A.31 but also from the definition of $\mathfrak{z}_1(\mathtt{x}, \cdot)$ and ensures that $\mathfrak{z}_1(\mathtt{x}, 3p^*) = O(\sqrt{\mathtt{x} + p^*})$.

5.6. Convergence of PPP

Define the set

$$\boldsymbol{\mathcal{M}}_M(\mathtt{x}) \overset{\text{def}}{=} \left\{ \sup_{\boldsymbol{\theta} \in S_1^{p,+}} \left\| \widetilde{\boldsymbol{\eta}}_{m,\boldsymbol{\theta}}^{(\infty)} \right\| \le \mathtt{C}(\mathtt{x})\sqrt{m} \right\} \cap \bigcap_{l=1}^{M} \boldsymbol{\mathcal{N}}_l(\mathtt{x}),$$

where

$$\begin{aligned}
\boldsymbol{\mathcal{N}}_l(\mathtt{x}) \overset{\text{def}}{=} & \left\{ \sup_{\boldsymbol{\upsilon} \in \varUpsilon_m \setminus \varUpsilon_\circ(\mathtt{r}_0^\circ)} \mathcal{L}(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}_m^*) < 0 \right\} \\
& \cap \{ \widetilde{\boldsymbol{\upsilon}}_{m(l)}, \widetilde{\boldsymbol{\upsilon}}_{m\boldsymbol{\theta}^*_{(l)}(l)}, \widetilde{\boldsymbol{\upsilon}}_{m\boldsymbol{\eta}^*_{(l)}(l)} \in \{ \| \mathcal{D}_{(l)}(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^*_{m(l)}) \| \le \mathtt{r}_0 \} \} \\
& \cap \bigcap_{\mathtt{r} \le \mathtt{r}_0} \left\{ \sup_{\boldsymbol{\upsilon} \in \varUpsilon_\circ(\mathtt{r})} \left\| \mathcal{D}_{(l)}{}^{-1} \left(\nabla \zeta_{\varepsilon(l)}(\boldsymbol{\upsilon}) - \nabla \zeta_{\varepsilon(l)}(\boldsymbol{\upsilon}^*_{m(l)}) \right) \right\| - 2\mathtt{r}^2 \le \mathtt{C}\omega\nu_1(\mathtt{x} + p^*) \right\} \\
& \cap \left\{ \| \mathcal{D}_{(l)}{}^{-1} \nabla \mathcal{L}_{\varepsilon(l)}(\boldsymbol{\upsilon}^*_{m(l)}) \| \le \mathtt{C}\sqrt{\mathtt{x} + p^*} \right\} \\
& \cap \left\{ \sup_{\boldsymbol{\upsilon} \in \varUpsilon_{\circ(l)}(\mathtt{r}^\infty)} \| \nabla(\mathbb{E} - \mathbb{E}_\varepsilon)[\mathcal{L}_{\varepsilon(l)}(\boldsymbol{\upsilon}^*_{m(l)}) - \mathcal{L}_{\varepsilon(l)}(\boldsymbol{\upsilon})] \| \le \mathtt{C}(\mathtt{x} + p^*)^2 \mathtt{r}^\infty / \sqrt{n} \right\} \\
& \cap \left\{ \text{The conditions of Section 4.1 are met for } (\mathcal{L}_{\varepsilon(l)}, \varUpsilon_m, \mathcal{D}_{(l)}) \right\},
\end{aligned}$$

where $\mathtt{r}_0 = \mathtt{C}(p^* + \mathtt{x})M$, $\mathtt{r}_0^\circ = \mathtt{C}[p^{*3/2}\sqrt{p^* + \mathtt{x}} \vee (p^* + \mathtt{x})M]$ and where

$$\mathtt{r}^\infty(\mathtt{x}) = \mathtt{C}\sqrt{p^* + \mathtt{x}}.$$

Remark 5.8. For $M = 1$ this is the set on which Proposition 2.3 applies.

Lemma 5.8. *We have on the set $\boldsymbol{\mathcal{M}}_M(\mathtt{x})$ if ${p^*}^5/n < l$*

$$\tau_{i(l)} \le \mathtt{C} l \sqrt{m} \left(\frac{{p^*}^{7/2} + \mathtt{x}}{n} + \frac{\sqrt{p^* + \mathtt{x}}}{\sqrt{n}} \right). \tag{5.6}$$

Proof. We obtain with Proposition 4.4 that if

$$(\delta(\mathtt{r})/\mathtt{r} + 6\nu_1\omega)\,\mathtt{r}_0 < 1, \qquad \text{and } (\delta(\mathtt{r})/\mathtt{r} + 6\nu_1\omega)\,\mathtt{C}\sqrt{\mathtt{x} + p^*} < 1,$$

that then

$$\boldsymbol{\mathcal{M}}_M(\mathtt{x}) \subset \{ \widetilde{\boldsymbol{\upsilon}}_{m(l)}, \widetilde{\boldsymbol{\upsilon}}_{m\boldsymbol{\theta}^*_{(l)}(l)} \subset \varUpsilon_\circ(\mathtt{r}^\infty) \},$$

where

$$\mathtt{r}^\infty(\mathtt{x}) \le \mathtt{C}\sqrt{p^* + \mathtt{x}}.$$

But by assumption

$$(\delta(\mathtt{r})/\mathtt{r}+6\nu_1\omega)\,\mathtt{C}\sqrt{\mathtt{x}+p^*}\;\le\;\mathtt{C}\frac{{p^*}^{5/2}+\mathtt{x}}{\sqrt{n}}\to 0,$$

$$(\delta(\mathtt{r})/\mathtt{r}+6\nu_1\omega)\,\mathtt{r}_0(\mathtt{x})\;\le\;\mathtt{C}\frac{{p^*}^{3}\log(n)M+\mathtt{x}}{\sqrt{n}}\to 0.$$

Consequently we can restrict our selves to the set $\varUpsilon_\circ(\mathtt{r}^\infty)$. We show the claim via induction. For this note that with (2.7) we already showed the claim for $l=1$. Assume that the claim is already shown for $0<l-1<M$. Remember that

$$\varsigma_{i,m}(\boldsymbol{\upsilon})\overset{\text{def}}{=}\Big(\boldsymbol{f}_{\boldsymbol{\eta}}'(\mathbf{X}_i^\top\boldsymbol{\theta})\nabla\varPhi(\boldsymbol{\theta})^\top\mathbf{X}_i,\boldsymbol{e}(\mathbf{X}_i^\top\boldsymbol{\theta})\Big)\in\mathbb{R}^{p+m}.$$

We find with the same arguments as in the proof of Proposition 4.4 and using Lemma A.11 that on the set $\boldsymbol{\mathcal{M}}_M$ (we suppress $\cdot_{(l)}$)

$$\begin{aligned}
&\sup_{\boldsymbol{\upsilon}\in\varUpsilon_\circ(\mathtt{r}^\infty)}\left\|\mathcal{D}^{-1}\left(\nabla\mathcal{L}(\boldsymbol{\upsilon})-\nabla\mathcal{L}(\boldsymbol{\upsilon}_m^*)\right)+\mathcal{D}(\boldsymbol{\upsilon}-\boldsymbol{\upsilon}_m^*)\right\|\\
&\quad\le\sup_{\boldsymbol{\upsilon}\in\varUpsilon_\circ(\mathtt{r}^\infty)}\left\|\mathcal{D}^{-1}\left(\nabla\mathcal{L}_\varepsilon(\boldsymbol{\upsilon})-\nabla\mathcal{L}_\varepsilon(\boldsymbol{\upsilon}_m^*)\right)+\mathcal{D}(\boldsymbol{\upsilon}-\boldsymbol{\upsilon}_m^*)\right\|\\
&\qquad+\sup_{\boldsymbol{\upsilon}\in\varUpsilon_\circ(\mathtt{r}^\infty)}\left\|\mathcal{D}^{-1}\left(\nabla\mathcal{L}_\tau(\boldsymbol{\upsilon})-\nabla\mathcal{L}_\tau(\boldsymbol{\upsilon}_m^*)\right)\right\|\\
&\quad\le\diamondsuit_Q(\mathtt{r}^\infty,\mathtt{x})+\sup_{\boldsymbol{\upsilon}\in\varUpsilon_\circ(\mathtt{r}^\infty)}\frac{2}{c_{\mathcal{D}}\sqrt{n}}\sum_{i=1}^n\tau_i(l-1)\left\|\varsigma_{i,m}(\boldsymbol{\upsilon})-\varsigma_{i,m}(\boldsymbol{\upsilon}_m^*)\right\|\\
&\quad\le\diamondsuit_Q(\mathtt{r}^\infty,\mathtt{x})+\frac{\mathtt{C}m^{3/2}\mathtt{r}^\infty}{c_{\mathcal{D}}^2}\max_i|\tau_i(l-1)|,
\end{aligned}$$

We find

$$\begin{aligned}
&\left\|\mathcal{D}_{(l)}(\widetilde{\boldsymbol{\upsilon}}_{m\,(l)}-\boldsymbol{\upsilon}_{m\,(l)}^*)\right\|\\
&\quad\le\left\|\mathcal{D}_{(l)}{}^{-1}\nabla\mathcal{L}_{\varepsilon\,(l)}(\boldsymbol{\upsilon}_{m\,(l)}^*)\right\|+\mathtt{C}\frac{{p^*}^{7/2}+\mathtt{x}}{\sqrt{n}}+\mathtt{C}{p^*}^2\max_i|\tau_i(l-1)|\\
&\quad\le\mathtt{C}\left(\sqrt{p^*+\mathtt{x}}+\mathtt{C}\frac{{p^*}^{7/2}+\mathtt{x}}{\sqrt{n}}+{p^*}^2\max_i|\tau_i(l-1)|\right).
\end{aligned}$$

It remains to address the bias $\|\mathcal{D}_{(l)}(\boldsymbol{\upsilon}_{m\,(l)}^*-\boldsymbol{\upsilon}^*{}_{(l)})\|$. Using that the assumptions $(\mathcal{A})$ hold for all $(g_{(l)})_{l=1,\dots,M}$ we can bound as in Lemma A.7

$$\mathbb{E}\mathcal{L}_{\varepsilon\,(l)}(\boldsymbol{\upsilon},\boldsymbol{\upsilon}^*{}_{(l)})\;\le\;-\mathtt{b}\mathtt{r}^2,$$

where $\mathtt{r} = \|\mathcal{D}_{(l)}(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^*{}_{(l)})\|$. With Lemma A.2 of [1] this gives

$$\|\mathcal{D}_{(l)}(\boldsymbol{\upsilon}^*_{m\,(l)} - \boldsymbol{\upsilon}^*{}_{(l)})\|^2 \le \mathtt{r}^{*2},$$

where we point out that $\mathtt{r}^* \le \mathtt{C}\sqrt{n}m^{-\alpha}$ in (5.1) is a uniform upper bound for all $l \le M$. We derived that on the set $\boldsymbol{\mathcal{M}}_M$ using that $\mathtt{r}^* \le \mathtt{C}\sqrt{p^* + \mathtt{x}}$

$$\begin{aligned}
&\big\|\mathcal{D}_{(l)}(\widetilde{\boldsymbol{\upsilon}}_{m\,(l)} - \boldsymbol{\upsilon}^*{}_{(l)})\big\| \\
&\quad\le\ \mathtt{C}\left(\sqrt{p^* + \mathtt{x}} + \mathtt{C}\frac{p^{*\,7/2} + \mathtt{x}}{\sqrt{n}} + p^{*\,2}\max_i |\tau_i(l-1)|\right) \\
&\quad\overset{\text{def}}{=}\ \mathtt{C}\boldsymbol{T}_{(l-1)}.
\end{aligned} \tag{5.7}$$

Finally we bound

$$\begin{aligned}
\Big|\boldsymbol{f}_{\boldsymbol{\eta}^*_{(l)}}(\mathbf{X}_i^\top \boldsymbol{\theta}^*_{(l)}) - \boldsymbol{f}_{\widetilde{\boldsymbol{\eta}}_{(l)}}(\mathbf{X}_i^\top \widetilde{\boldsymbol{\theta}}_{(l)})\Big| &\le \Big|\boldsymbol{f}_{\boldsymbol{\eta}^*_{(l)} - \widetilde{\boldsymbol{\eta}}_{(l)}}(\mathbf{X}^\top \widetilde{\boldsymbol{\theta}})\Big| \\
&\quad + \Big|\boldsymbol{f}_{\boldsymbol{\eta}^*_{(l)}}(\mathbf{X}^\top \boldsymbol{\theta}^*_{(l)}) - \boldsymbol{f}_{\boldsymbol{\eta}^*_{(l)}}(\mathbf{X}^\top \widetilde{\boldsymbol{\theta}}_{(l)})\Big|.
\end{aligned}$$

We estimate separately using (5.7)

$$\begin{aligned}
\Big|\boldsymbol{f}_{\boldsymbol{\eta}^*_{(l)} - \widetilde{\boldsymbol{\eta}}_{(l)}}(\mathbf{X}^\top \widetilde{\boldsymbol{\theta}}_{(l)})\Big| &\le \|\|\mathcal{H}_m^{-1}\boldsymbol{e}\|_{\mathbb{R}^m}\|_\infty \mathtt{C}\boldsymbol{T}_{(l-1)} \\
&\le \mathtt{C}\sqrt{m}\boldsymbol{T}_{(l-1)}/\sqrt{n}.
\end{aligned}$$

Furthermore we find with (5.7)

$$\Big|\boldsymbol{f}_{\boldsymbol{\eta}^*_{(l)}}(\mathbf{X}^\top \boldsymbol{\theta}^*_{(l)}) - \boldsymbol{f}_{\boldsymbol{\eta}^*_{(l)}}(\mathbf{X}^\top \widetilde{\boldsymbol{\theta}})_{(l)}\Big| \le \mathtt{C}s_{\mathbf{X}}\|\boldsymbol{f}'_{\boldsymbol{\eta}^*_{(l)}}\|\boldsymbol{T}_{(l-1)}/\sqrt{n}.$$

Consequently

$$\begin{aligned}
\big|\tau_{i\,(l)}\big| &= \left|\sum_{s=1}^{l} \boldsymbol{f}_{\boldsymbol{\eta}^*_{(s)}}(\mathbf{X}_i^\top \boldsymbol{\theta}^*_{(s)}) - \boldsymbol{f}_{\widetilde{\boldsymbol{\eta}}_{(s)}}(\mathbf{X}_i^\top \widetilde{\boldsymbol{\theta}}_{(s)})\right| \\
&\le \mathtt{C}l\sqrt{m}\left(\frac{p^{*\,7/2} + \mathtt{x}}{n} + \frac{\sqrt{p^* + \mathtt{x}}}{\sqrt{n}}\right) + \mathtt{C}\sum_{s=1}^{l} \frac{p^{*\,5/2}}{\sqrt{n}} \max_i |\tau_i(s-1)|.
\end{aligned}$$

Denote

$$a \overset{\text{def}}{=} \mathtt{C}\sqrt{m}\left(\frac{p^{*\,7/2} + \mathtt{x}}{n} + \frac{\sqrt{p^* + \mathtt{x}}}{\sqrt{n}}\right), \qquad b \overset{\text{def}}{=} \frac{p^{*\,5/2}}{\sqrt{n}}.$$

Furthermore define

$$S_{k(l)} \overset{\text{def}}{=} \sum_{s=1}^{l} S_{k-1(s-1)}, \qquad S_{0(l)} = l.$$

Then we can write

$$\left|\tau_{i(l)}\right| \le a \sum_{k=0}^{l-1} b^k S_{k(l)},$$

which gives with the crude bound $S_{k(l)} \le l \sum_{s=0}^{k} l^s = l\frac{l^{k+1}-1}{l-1} \le 2l^{k+1}$ that

$$\left|\tau_{i(l)}\right| \le 2la \sum_{k=0}^{l-1} b^k l^k \le \mathtt{C} la,$$

if $b < l \le M$. This gives the claim. □

To complete this section we show that the set $\boldsymbol{\mathcal{M}}_M$ is of large probability as long as $M \in \mathbb{N}$ is not too big.

Lemma 5.9. *We have*

$$\mathbb{P}(\boldsymbol{\mathcal{M}}_M) \ge 1 - \mathrm{e}^{-\mathtt{x}} - M\left(12\mathrm{e}^{-\mathtt{x}} + \exp\left\{-m^3\mathtt{x}\right\} + \exp\left\{-nc_{(\boldsymbol{Q})}/4\right\}\right)$$

Proof. With Lemma 5.2 we find

$$\mathbb{P}\left(\sup_{\boldsymbol{\theta}\in S_1^{p,+}} \left\|\widetilde{\boldsymbol{\eta}}_{m,\boldsymbol{\theta}}^{(\infty)}\right\| \ge \mathtt{C}(\mathtt{x})\sqrt{m}\right) \le \mathrm{e}^{-\mathtt{x}}.$$

Due to the assumptions we find with Lemma 5.3 that

$$\begin{aligned}&\mathbb{P}\left(\text{The conditions of Section 4.1 are met for } (\mathcal{L}_{\varepsilon(l)}, \varUpsilon_m, \mathcal{D}_{(l)})\right)\\ &\quad\ge 1 - 4\mathrm{e}^{-\mathtt{x}} - \exp\left\{-m^3\mathtt{x}\right\} - \exp\left\{-nc_{(\boldsymbol{Q})}/4\right\}.\end{aligned}$$

On that set we find as in the proof of Proposition 4.4 for $\mathtt{C} > 0$ large enough

$$\begin{aligned}\mathbb{P}\Bigg(\bigcap_{\mathtt{r}\le\mathtt{r}_0}\Bigg\{&\sup_{\boldsymbol{\upsilon}\in\varUpsilon_\circ(\mathtt{r})}\left\|\mathcal{D}_{(l)}{}^{-1}\left(\nabla\zeta_{\varepsilon(l)}(\boldsymbol{\upsilon}) - \nabla\zeta_{\varepsilon(l)}(\boldsymbol{\upsilon}_{m(l)}^*)\right)\right\| - 2\mathtt{r}^2\\ &\le \mathtt{C}\omega\nu_1(\mathtt{x}+p^*)\Bigg\}\Bigg) \ge 1 - \mathrm{e}^{-\mathtt{x}}.\end{aligned}$$

and

$$\mathbb{P}\left(\left\|\mathcal{D}_{(l)}{}^{-1}\nabla\zeta_{\varepsilon(l)}(\boldsymbol{\upsilon}_{m(l)}^*)\right\| \ge \mathtt{C}\sqrt{\mathtt{x}+p^*}\right) \ge 1 - 2\mathrm{e}^{-\mathtt{x}}.$$

Furthermore by Lemma 3.1 We have that

$$\mathbb{P}\left(\sup_{\boldsymbol{\upsilon}\in\Upsilon_{\circ(l)}(\mathtt{r})}\|\nabla(\mathbb{E}-\mathbb{E}_{\varepsilon})[\mathcal{L}_{\varepsilon(l)}(\boldsymbol{\upsilon}^{*}_{m(l)})-\mathcal{L}_{\varepsilon(l)}(\boldsymbol{\upsilon})]\geq\mathtt{C}(\mathtt{x}+p^{*})^{2}\mathtt{r}/\sqrt{n}\right)$$
$$\leq 2\mathrm{e}^{-\mathtt{x}}.$$

For the large deviation bound we proceed as follows. Note that

$$\mathcal{L}_{(l)}(\boldsymbol{\upsilon},\boldsymbol{\upsilon}^{*}_{m(l)},Y_{i(l)}) = \mathcal{L}_{\varepsilon(l)}(\boldsymbol{\upsilon},\boldsymbol{\upsilon}^{*}_{m(l)},Y_{i(l)}) +2\sum_{i=1}^{n}\tau_i(l-1)\left(\boldsymbol{f}_{\boldsymbol{\eta}}(\mathbf{X}_i^{\top}\boldsymbol{\theta})-\boldsymbol{f}_{\boldsymbol{\eta}^{*}_{m(l)}}(\mathbf{X}_i^{\top}\boldsymbol{\theta}^{*}_{m(l)})\right).$$

We can bound

$$\sum_{i=1}^{n}\tau_i(l-1)\left(\boldsymbol{f}_{\boldsymbol{\eta}}(\mathbf{X}_i^{\top}\boldsymbol{\theta})-\boldsymbol{f}_{\boldsymbol{\eta}^{*}_{m(l)}}(\mathbf{X}_i^{\top}\boldsymbol{\theta}^{*}_{m(l)})\right) \leq \mathtt{C}\max_i|\tau_i(l-1)|\sqrt{n}\sqrt{m}\mathtt{r}.$$

As the conditions $(\mathcal{A})$ are satisfied for all $l=1,\ldots,M$ we can establish as in Lemma 5.3 for $\mathtt{r}^2\geq\mathtt{C}_{\mathtt{b}}p^{*}\log(n)$

$$-\mathbb{E}_{\varepsilon}\sum_{i=1}^{n}\left(g_{(l)}(\mathbf{X}_i)+\varepsilon_i-\boldsymbol{f}_{\boldsymbol{\eta}}(\mathbf{X}_i^{\top}\boldsymbol{\theta})\right)^2 -\left(g_{(l)}(\mathbf{X}_i)+\varepsilon_i-\boldsymbol{f}_{\boldsymbol{\eta}^{*}_{m(l)}}(\mathbf{X}_i^{\top}\boldsymbol{\theta}^{*}_{m(l)})\right)^2\leq-\mathtt{b}_{(l)}\mathtt{r}^2.$$

Together this implies for $\mathtt{r}\geq\mathtt{C}_{\mathtt{b}}p^{*}$

$$\mathbb{E}_{\varepsilon}\mathcal{L}_{(l)}(\boldsymbol{\upsilon},\boldsymbol{\upsilon}^{*}_{m(l)},Y_{i(l)})\leq-\mathtt{b}_{(l)}\mathtt{r}^2+\mathtt{C}\boldsymbol{B}_{(l-1)}\sqrt{n}\sqrt{m}\mathtt{r}.$$

This gives for $\mathtt{r}\geq\mathtt{C}\boldsymbol{B}_{(l-1)}\sqrt{n}\sqrt{m}$ and $\mathtt{C}>0$ large enough

$$\mathbb{E}_{\varepsilon}\mathcal{L}(\boldsymbol{\upsilon},\boldsymbol{\upsilon}^{*}_{m(l)},Y_{i(l)})\leq-\mathtt{b}_{(l)}\mathtt{r}^2/2.$$

Plugging in (5.6) the lower bound becomes

$$\mathtt{r}_{0(l)}\geq\mathtt{C}\sqrt{p^{*}+\mathtt{x}}\left(1+l\sqrt{m}\frac{{p^{*}}^{7/2}+\mathtt{x}}{\sqrt{n}}\right)=\mathtt{C}'M(p^{*}+\mathtt{x}).$$

For the remaining part we proceed as in Section 5.2. This gives the claim. □

Appendix A: Technical proofs

In the following all the technical steps necessary to prove the Lemmas of Section 5 are presented. But first we cite an important result that will be used in our arguments, namely the bounded difference inequality:

Theorem A.1 (Bounded differences inequality)**.** *Let a function $f : \mathcal{X}^n \to \mathbb{R}$ satisfy for any $\mathbf{X}_1, \dots, \mathbf{X}_n, \mathbf{X}_i' \in \mathcal{X}$*

$$|f(\mathbf{X}_1, \dots, \mathbf{X}_{i-1}, \mathbf{X}_i, \mathbf{X}_{i+1}, \dots, \mathbf{X}_n) - f(\mathbf{X}_1, \dots, \mathbf{X}_{i-1}, \mathbf{X}_i', \mathbf{X}_{i+1}, \dots, \mathbf{X}_n)| \leq c_i.$$

Then for any vector of independent random variables $\mathbf{X} \in \mathcal{X}^n$

$$\mathbb{P}\left(f(\mathbf{X}) - \mathbb{E}f(\mathbf{X}) \geq t\right) \leq \mathrm{e}^{-\frac{2t^2}{\sum_{i=1}^n c_i^2}},$$

$$\mathbb{P}\left(f(\mathbf{X}) - \mathbb{E}f(\mathbf{X}) \leq -t\right) \leq \mathrm{e}^{-\frac{2t^2}{\sum_{i=1}^n c_i^2}}.$$

Furthermore we will use the basic chaining device as it was introduced by [6] (see Section 2 of [21] for a more concise description). As we use the idea several times, we summarize the central step in the following Lemma

Lemma A.2. *Let $\{\mathcal{Y}(\boldsymbol{\upsilon}) - \mathcal{Y}(\boldsymbol{\upsilon}^*),\, \boldsymbol{\upsilon} \in \varUpsilon\}$ be a family of random variables index by a set $\varUpsilon$ that is contained in a normed space $(\mathcal{X}, \|\cdot\|)$. Define $\varUpsilon_0 = \{\boldsymbol{\upsilon}^*\}$ and with some $\mathtt{r} > 0$ the sequence $\mathtt{r}_k = 2^{-k}\mathtt{r}$ and the sequence of sets $\varUpsilon_k$ each with minimal cardinality such that*

$$\varUpsilon \subset \bigcup_{\boldsymbol{\upsilon} \in \varUpsilon_k} B_{\mathtt{r}_k}(\boldsymbol{\upsilon}), \qquad B_{\mathtt{r}}(\boldsymbol{\upsilon}) \overset{\text{def}}{=} \{\boldsymbol{\upsilon}^\circ \in \varUpsilon,\, \|\boldsymbol{\upsilon}^\circ - \boldsymbol{\upsilon}\| \leq \mathtt{r}\}.$$

Then for any $\mathfrak{z} > 0$

$$\begin{aligned}
&\mathbb{P}\left(\sup_{\boldsymbol{\upsilon} \in \varUpsilon} |\mathcal{Y}(\boldsymbol{\upsilon}) - \mathcal{Y}(\boldsymbol{\upsilon}^*)| \geq \mathfrak{z}\right) \\
&\quad \leq \sum_{k=1}^{\infty} |\varUpsilon_k| \sup_{\boldsymbol{\upsilon}^\circ \in \varUpsilon_k} \mathbb{P}\left(\inf_{\boldsymbol{\upsilon} \in \varUpsilon_{k-1}} |\mathcal{Y}(\boldsymbol{\upsilon}) - \mathcal{Y}(\boldsymbol{\upsilon}^\circ)| \geq 2^{-(k-1)/2}(1 - 1/\sqrt{2})\mathfrak{z}\right).
\end{aligned}$$

Proof. We simply use the definition and estimate

$$\begin{aligned}
&\mathbb{P}\left(\sup_{\boldsymbol{\upsilon} \in \varUpsilon} |\mathcal{Y}(\boldsymbol{\upsilon}) - \mathcal{Y}(\boldsymbol{\upsilon}^*)| \geq \mathfrak{z}\right) \\
&\quad \leq \mathbb{P}\left(\sum_{k=1}^{\infty} \sup_{\boldsymbol{\upsilon}_k \in \varUpsilon_k} \inf_{\boldsymbol{\upsilon}_{k-1} \varUpsilon_{k-1}} |\mathcal{Y}(\boldsymbol{\upsilon}_k) - \mathcal{Y}(\boldsymbol{\upsilon}_{k-1})| \geq \mathfrak{z}\right) \\
&\quad \leq \sum_{k=1}^{\infty} \mathbb{P}\left(\sup_{\boldsymbol{\upsilon}_k \in \varUpsilon_k} \inf_{\boldsymbol{\upsilon}_{k-1} \in \varUpsilon_{k-1}} |\mathcal{Y}(\boldsymbol{\upsilon}_k) - \mathcal{Y}(\boldsymbol{\upsilon}_{k-1})| \geq 2^{-(k-1)/2}(1 - 1/\sqrt{2})\mathfrak{z}\right)
\end{aligned}$$

$$\leq \sum_{k=1}^{\infty} |\Upsilon_k| \sup_{\boldsymbol{\upsilon}_k \in \Upsilon_k} \mathbb{P}\left(\inf_{\boldsymbol{\upsilon}_{k-1}\in\Upsilon_{k-1}} |\mathcal{Y}(\boldsymbol{\upsilon}_k) - \mathcal{Y}(\boldsymbol{\upsilon}_{k-1})| \geq 2^{-(k-1)/2}(1-1/\sqrt{2})\mathfrak{z} \right),$$

where we used that $\sum_{k=1}^{\infty} 2^{-(k-1)/2} \leq 1/(1-1/\sqrt{2})$. □

A.1. Proof of Remark 2.3

Proof. This can be seen as follows. First with Fubini's Theorem we find

$$\begin{aligned}
\boldsymbol{\eta}_k(\boldsymbol{\theta}) &\overset{\text{def}}{=} \int_{[-s_{\mathbf{X}},s_{\mathbf{X}}]} f_{\boldsymbol{\theta}}(t)\boldsymbol{e}_k(t)dt \\
&= \int_{[-s_{\mathbf{X}},s_{\mathbf{X}}]} \int_{B_{s_{\mathbf{X}}}(0)\cap\boldsymbol{\theta}^{\perp}} f_{\boldsymbol{\theta},\boldsymbol{x}}(t)\boldsymbol{e}_k(t) p_{\mathbf{X}|\mathbf{X}^{\top}\boldsymbol{\theta}=t}(\boldsymbol{x})d\boldsymbol{x}dt \\
&= \int_{B_{s_{\mathbf{X}}}(0)\cap\boldsymbol{\theta}^{\perp}} \left(\int_{[-s_{\mathbf{X}},s_{\mathbf{X}}]} f_{\boldsymbol{\theta},\boldsymbol{x}}(t)\boldsymbol{e}_k(t)dt \right) p_{\mathbf{X}|\mathbf{X}^{\top}\boldsymbol{\theta}=t}(\boldsymbol{x})d\boldsymbol{x}, \\
&= \int_{B_{s_{\mathbf{X}}}(0)\cap\boldsymbol{\theta}^{\perp}} \boldsymbol{\eta}_k(\boldsymbol{\theta},\boldsymbol{x}) p_{\mathbf{X}|\mathbf{X}^{\top}\boldsymbol{\theta}=t}(\boldsymbol{x})d\boldsymbol{x}.
\end{aligned}$$

Note that the application of Fubini's theorem is justified since by assumtion

$$|f_{\boldsymbol{\theta},\boldsymbol{x}}(t)\boldsymbol{e}_k(t)p_{\mathbf{X}|\mathbf{X}^{\top}\boldsymbol{\theta}=t}(\boldsymbol{x})| < \infty.$$

Furthermore with Jensen's inequality and exchanging the order integration and summation as the lim sup is finite we find

$$\begin{aligned}
\sum_{k=0}^{\infty} k^{2\alpha(\boldsymbol{\theta})}\boldsymbol{\eta}_k^2(\boldsymbol{\theta})^2 &= \sum_{k=0}^{\infty} k^{2\alpha(\boldsymbol{\theta})} \left(\int_{B_{s_{\mathbf{X}}}(0)\cap\boldsymbol{\theta}^{\perp}} \boldsymbol{\eta}_k(\boldsymbol{\theta},\boldsymbol{x}) p_{\mathbf{X}|\mathbf{X}^{\top}\boldsymbol{\theta}=t}(\boldsymbol{x})d\boldsymbol{x} \right)^2 \\
&\leq \sum_{k=0}^{\infty} \int_{B_{s_{\mathbf{X}}}(0)\cap\boldsymbol{\theta}^{\perp}} k^{2\alpha(\boldsymbol{\theta})}\boldsymbol{\eta}_k(\boldsymbol{\theta},\boldsymbol{x})^2 p_{\mathbf{X}|\mathbf{X}^{\top}\boldsymbol{\theta}=t}(\boldsymbol{x})d\boldsymbol{x} \\
&\leq \int_{B_{s_{\mathbf{X}}}(0)\cap\boldsymbol{\theta}^{\perp}} \left(\sum_{k=0}^{\infty} k^{2\alpha(\boldsymbol{\theta},\boldsymbol{x})}\boldsymbol{\eta}_k(\boldsymbol{\theta},\boldsymbol{x})^2 \right) p_{\mathbf{X}|\mathbf{X}^{\top}\boldsymbol{\theta}=t}(\boldsymbol{x})d\boldsymbol{x} \\
&< \infty,
\end{aligned}$$

where we used in the second to last step that $\alpha(\boldsymbol{\theta}) \leq \alpha(\boldsymbol{\theta},\boldsymbol{x})$. □

A.2. Calculating the elements

First we calculate the relevant objects in this setting. For this we have to emphasize one subtlety about this analysis. As the parameter $\boldsymbol{\theta} \in \mathbb{R}^p$ lies in $S_1^{p,+} \subset \mathbb{R}^p$

a more appropriate parameter set is $W_S \overset{\text{def}}{=} [0,\pi] \times [-\pi/2,\pi/2] \times [-\pi/2,\pi/2] \times \cdots \times [-\pi/2,\pi/2] \subset \mathbb{R}^{p-1}$. This gives, parametrising the half sphere $S_1^{p,+} \subset \mathbb{R}^p$ via the standard spherical coordinates

$$\varPhi : [0,\pi] \times [-\pi/2,\pi/2] \times [-\pi/2,\pi/2] \times \cdots \times [-\pi/2,\pi/2] \subset \mathbb{R}^{p-1} \to S_1^{p,+},$$

that our actual likelihood functional is defined on $W_S \times \mathbb{R}^m$ as

$$\mathcal{L}_m(\boldsymbol{\theta},\boldsymbol{\eta}) = \sum_{i=1}^{n} \|Y_i - \boldsymbol{f}_{\boldsymbol{\eta}}(\mathbf{X}_i^\top \varPhi(\boldsymbol{\theta}))\|^2/2,$$

where with abuse of notation we denote the preimage of an element of the sphere by the same symbol. Fix any element of the set of maximizers $\boldsymbol{\upsilon}_m^*$ for some $m \in \mathbb{N}$.

First we calculate

$$\zeta(\boldsymbol{\upsilon},\boldsymbol{\upsilon}^*) := \mathcal{L}_m(\boldsymbol{\upsilon},\boldsymbol{\upsilon}^*) - \mathbb{E}_\varepsilon \mathcal{L}_m(\boldsymbol{\upsilon},\boldsymbol{\upsilon}^*) = -\sum_{i=1}^{n} \varepsilon_i \Big(g(\mathbf{X}_i) - \boldsymbol{f}_{\boldsymbol{\eta}}(\mathbf{X}_i^\top \varPhi(\boldsymbol{\theta})) \Big).$$

This gives that with $\nabla_{p^*} = (\nabla_{\theta_1},\ldots,\nabla_{\theta_{p-1}},\nabla_{\eta_1},\ldots,\nabla_{\eta_m})$ and $\boldsymbol{\varepsilon} = (\varepsilon_1,\ldots,\epsilon)$

$$\begin{aligned}\nabla_{p^*}\zeta(\boldsymbol{\upsilon}) &= \sum_{i=1}^{n} \Big(\boldsymbol{f}_{\boldsymbol{\eta}}'(\mathbf{X}_i^\top\boldsymbol{\theta})\nabla\varPhi(\boldsymbol{\theta})^\top \mathbf{X}_i, \boldsymbol{e}(\mathbf{X}_i^\top\boldsymbol{\theta}) \Big)\varepsilon_i \\ &\overset{\text{def}}{=} \sum_{i=1}^{n} \varsigma_{i,m}(\boldsymbol{\upsilon})\varepsilon_i \overset{\text{def}}{=} W_m(\boldsymbol{\upsilon})\boldsymbol{\varepsilon}.\end{aligned}$$

where with $\boldsymbol{e} = (\boldsymbol{e}_1,\ldots,\boldsymbol{e}_m)$

$$W_m(\boldsymbol{\upsilon}) = \begin{pmatrix} \boldsymbol{f}_{\boldsymbol{\eta}}'(\mathbf{X}_1^\top\boldsymbol{\theta})\nabla\varPhi(\boldsymbol{\theta})^\top\mathbf{X}_1 & \ldots & \boldsymbol{f}_{\boldsymbol{\eta}}'(\mathbf{X}_n^\top\boldsymbol{\theta})\nabla\varPhi(\boldsymbol{\theta})^\top\mathbf{X}_n \\ \boldsymbol{e}(\mathbf{X}_1^\top\boldsymbol{\theta}) & \ldots & \boldsymbol{e}(\mathbf{X}_n^\top\boldsymbol{\theta}) \end{pmatrix}.$$

As we use this notation in the following, we repeat the definition

$$\varsigma_{i,m}(\boldsymbol{\upsilon}) \overset{\text{def}}{=} \Big(\boldsymbol{f}_{\boldsymbol{\eta}}'(\mathbf{X}_i^\top\boldsymbol{\theta})\nabla\varPhi(\boldsymbol{\theta})^\top\mathbf{X}_i, \boldsymbol{e}(\mathbf{X}_i^\top\boldsymbol{\theta}) \Big) \in \mathbb{R}^{p^*}. \tag{A.1}$$

By assumption the ε_i are i.i.d. with covariance $\sigma^2 > 0$ and the design points $(\mathbf{X}_i)$ are i.i.d. as well. We set

$$\begin{aligned}\mathcal{V}_m^2 &\overset{\text{def}}{=} \sigma^2 \mathbb{E} W_m(\boldsymbol{\upsilon}^*) W_m(\boldsymbol{\upsilon}^*)^\top \\ &= n\sigma^2 \begin{pmatrix} d_{\boldsymbol{\theta}}^2(\boldsymbol{\upsilon}^*) & a_m(\boldsymbol{\upsilon}^*) \\ a_m^\top(\boldsymbol{\upsilon}^*) & h_m^2(\boldsymbol{\upsilon}^*) \end{pmatrix} \overset{\text{def}}{=} n\sigma^2 d_m^2 \in \mathbb{R}^{(p-1+m)\times(p-1+m)}.\end{aligned}$$

where with $\mathbb{E}[\cdot]$ denoting the expectation under the measure $\mathbb{P}^{\mathbf{X}_1}$

$$\begin{aligned}
d^2_{\boldsymbol{\theta}}(\boldsymbol{\upsilon}) &= \mathbb{E}\Big[\boldsymbol{f}'_{\boldsymbol{\eta}}(\mathbf{X}_1^\top\boldsymbol{\theta})^2 \nabla\Phi(\boldsymbol{\theta})^\top \mathbf{X}_1 \mathbf{X}_1{}^\top \nabla\Phi(\boldsymbol{\theta})\Big],\\
h^2_m(\boldsymbol{\upsilon}) &= \mathbb{E}\Big[\boldsymbol{e}\boldsymbol{e}^\top(\mathbf{X}_1^\top\boldsymbol{\theta})\Big],\\
a_m(\boldsymbol{\upsilon}) &= \mathbb{E}\Big[\boldsymbol{f}'_{\boldsymbol{\eta}}(\mathbf{X}_1^\top\boldsymbol{\theta})\nabla\Phi(\boldsymbol{\theta})^\top \mathbf{X}_1 \boldsymbol{e}^\top(\mathbf{X}_1^\top\boldsymbol{\theta})\Big].
\end{aligned}$$

Furthermore we get because of the quadratic functional and sufficient smoothness of the basis $(\boldsymbol{e}_i)$ for any $\boldsymbol{\upsilon} \in \mathbb{R}^{p^*-1}$

$$\begin{aligned}
\mathcal{D}^2_m(\boldsymbol{\upsilon}) &\overset{\text{def}}{=} -\nabla^2_{p^*}\mathbb{E}[\mathcal{L}_m(\boldsymbol{\upsilon})] = n d^2_m(\boldsymbol{\upsilon}) + n r^2_m(\boldsymbol{\upsilon}),\\
d^2_m &= \begin{pmatrix} d^2_{\boldsymbol{\theta}}(\boldsymbol{\upsilon}) & a_m(\boldsymbol{\upsilon}) \\ a_m^\top(\boldsymbol{\upsilon}) & h^2_m(\boldsymbol{\upsilon}) \end{pmatrix},\\
r^2_m(\boldsymbol{\upsilon}) &= \mathbb{E}\left[\big[\boldsymbol{f}_{\boldsymbol{\eta}}(\mathbf{X}^\top\boldsymbol{\theta}) - g(\mathbf{X})\big]\begin{pmatrix} v^2_{\boldsymbol{\theta}}(\boldsymbol{\upsilon}) & b_m(\boldsymbol{\upsilon}) \\ b_m^\top(\boldsymbol{\upsilon}) & 0 \end{pmatrix}\right],\\
v^2_{\boldsymbol{\theta}}(\boldsymbol{\upsilon}) &= 2\boldsymbol{f}''_{\boldsymbol{\eta}}(\mathbf{X}^\top\boldsymbol{\theta})\nabla\Phi_{\boldsymbol{\theta}}^\top \mathbf{X}\mathbf{X}^\top\nabla\Phi_{\boldsymbol{\theta}} + |\boldsymbol{f}'_{\boldsymbol{\eta}}(\mathbf{X}^\top\boldsymbol{\theta})|^2 \mathbf{X}^\top \nabla^2\Phi_{\boldsymbol{\theta}}^\top[\mathbf{X},\cdot,\cdot],\\
b_m(\boldsymbol{\upsilon}) &= \nabla\Phi_{\boldsymbol{\theta}}\mathbf{X}^\top \boldsymbol{e}'^\top(\mathbf{X}^\top\boldsymbol{\theta}).
\end{aligned}$$

For the analysis of the sieve bias we also define the corresponding full operator $\mathcal{D}^2 \in L(l^2, \{(x_k)_{k\in\mathbb{N}}, x \in \mathbb{R}\})$

$$\mathcal{D}^2(\boldsymbol{\upsilon}) = n d^2(\boldsymbol{\upsilon}) + \mathbb{E}\left[\big[\boldsymbol{f}_{\boldsymbol{\eta}}(\mathbf{X}^\top\boldsymbol{\theta}) - g(\mathbf{X})\big]\begin{pmatrix} v^2_{\boldsymbol{\theta}}(\boldsymbol{\upsilon}) & b_\infty^\top(\boldsymbol{\upsilon}) \\ b_\infty^\top(\boldsymbol{\upsilon}) & 0 \end{pmatrix}\right],$$

where with the obvious adaptations

$$d^2(\boldsymbol{\upsilon}) = \begin{pmatrix} d^2_{\boldsymbol{\theta}}(\boldsymbol{\upsilon}) & a_\infty(\boldsymbol{\upsilon}) \\ a_\infty^\top(\boldsymbol{\upsilon}) & h^2_\infty(\boldsymbol{\upsilon}) \end{pmatrix}.$$

Remark A.1. If $\mathbf{X}^\top\boldsymbol{\theta}^*$ was independent to $\mathbf{X}^\top\boldsymbol{\theta}^\circ$ for any $\boldsymbol{\theta}^\circ \in {\boldsymbol{\theta}^*}^\perp$, we would have $b_m(\boldsymbol{\upsilon}^*) = 0$ for $m \in \mathbb{N}\cup\{\infty\}$ by the definition of $\boldsymbol{f}_{\boldsymbol{\eta}^*}(\mathbf{X}^\top\boldsymbol{\theta}^*) \overset{\text{def}}{=} \mathbb{E}[g(\mathbf{X})|\mathbf{X}^\top\boldsymbol{\theta}^*]$.

Furthermore we calculate – with $\varsigma_{i,m}$ from (A.1) –

$$\nabla^2\zeta(\boldsymbol{\upsilon}) = \sum_{i=1}^n \nabla\varsigma_{i,m}(\boldsymbol{\upsilon}),$$

where

$$\begin{aligned}
\mathit{\Pi}_{\boldsymbol{\theta}}\nabla_{\boldsymbol{\theta}}\varsigma_{i,m}(\boldsymbol{\upsilon}) &= \boldsymbol{f}_{\boldsymbol{\eta}}''(\mathbf{X}_i^\top\boldsymbol{\theta})\nabla\mathit{\Phi}(\boldsymbol{\theta})^\top\mathbf{X}_i\mathbf{X}_i^\top\nabla\mathit{\Phi}(\boldsymbol{\theta})\\
&\quad+\boldsymbol{f}_{\boldsymbol{\eta}}'(\mathbf{X}_i^\top\boldsymbol{\theta})\mathbf{X}_i\nabla^2\mathit{\Phi}(\boldsymbol{\theta}^\top\mathbf{X}_i)[\mathbf{X}_i,\cdot,\cdot],\\
\mathit{\Pi}_{\boldsymbol{\eta}}\nabla_{\boldsymbol{\theta}}\varsigma_{i,m}(\boldsymbol{\upsilon}) &= \boldsymbol{e}'(\boldsymbol{\theta}^\top\mathbf{X}_i)\mathbf{X}_i^\top\nabla\mathit{\Phi}(\boldsymbol{\theta}),\\
\nabla_{\boldsymbol{\eta}}\varsigma_{i,m}(\boldsymbol{\upsilon}) &= 0.
\end{aligned}$$

A.3. Preliminary calculations

By assumption the density of $p_{\mathbf{X}}:\mathbb{R}^p\mapsto\mathbb{R}$ is Lipshitz continuous. Denote by $L_{p_{\mathbf{X}}}$ its Lipshitz constant. Define

$$I_k\overset{\text{def}}{=}\operatorname{supp}(\boldsymbol{e}_k)\subset[-s_{\mathbf{X}},s_{\mathbf{X}}].$$

Lemma A.3. *We have for all $k,l\in\mathbb{N}$*

$$\begin{aligned}
&|\mathbb{E}[\boldsymbol{e}_k\boldsymbol{e}_l(\mathbf{X}^\top\boldsymbol{\theta})]|\\
&\quad\le 17s_{\mathbf{X}}^{p+1}L_{p_{\mathbf{X}}}\|\psi\|_\infty2^{-j_l-1}2^{(j_k-j_l)/2}1_{\{I_l\cap I_k\neq\emptyset\}}(k,l),\ \textit{for } l\ge k, &(\text{A.2})\\
&|\mathbb{E}[(\mathbf{X}^\top\boldsymbol{\theta})\boldsymbol{e}_k'\boldsymbol{e}_l(\mathbf{X}^\top\boldsymbol{\theta}^\circ)]|\\
&\quad\le 17\frac{\sqrt{p+2}}{2}\pi\|\psi'\|_\infty s_{\mathbf{X}}^2\|p_{\mathbf{X}}\|_\infty2^{3j_k/2}2^{-(j_l\vee j_k)/2}1_{\{I_l\cap I_k\neq\emptyset\}}(k,l), &(\text{A.3})\\
&|\mathbb{E}[\boldsymbol{e}_l'\boldsymbol{e}_k'(\mathbf{X}^\top\boldsymbol{\theta})]|\\
&\quad\le 17s_{\mathbf{X}}\|\psi'\|_\infty\|p_{\mathbf{X}}\|_\infty2^{3(j_l+j_k)/2-(j_l\vee j_k)}1_{\{I_k\cap I_l\neq\emptyset\}}(k,l), &(\text{A.4})\\
&\mathbb{E}\left[(\boldsymbol{e}_k(\mathbf{X}^\top\boldsymbol{\theta})-\boldsymbol{e}_k(\mathbf{X}^\top\boldsymbol{\theta}'))(\boldsymbol{e}_l(\mathbf{X}^\top\boldsymbol{\theta})-\boldsymbol{e}_l(\mathbf{X}^\top\boldsymbol{\theta}'))\right]\\
&\quad\le \mathtt{C}\|\boldsymbol{\theta}-\boldsymbol{\theta}'\|^22^{j_k}2^{j_l}\|\psi'\|_\infty^2s_{\mathbf{X}}^417^21_{\{I_k\cap I_l\neq\emptyset\}}, &(\text{A.5})\\
&\mathbb{E}\left[(\boldsymbol{e}_k'(\mathbf{X}^\top\boldsymbol{\theta})-\boldsymbol{e}_k'(\mathbf{X}^\top\boldsymbol{\theta}'))(\boldsymbol{e}_l'(\mathbf{X}^\top\boldsymbol{\theta})-\boldsymbol{e}_l'(\mathbf{X}^\top\boldsymbol{\theta}'))\right]\\
&\quad\le \mathtt{C}\|\boldsymbol{\theta}-\boldsymbol{\theta}'\|^22^{2j_k}2^{2j_l}\|\psi''\|_\infty^2s_{\mathbf{X}}^417^21_{\{I_k\cap I_l\neq\emptyset\}}, &(\text{A.6})\\
&\mathbb{E}\left[\left(\boldsymbol{e}_l(\mathbf{X}^\top\boldsymbol{\theta})-\boldsymbol{e}_l(\mathbf{X}^\top\boldsymbol{\theta}')\right)\boldsymbol{e}_k(\mathbf{X}^\top\boldsymbol{\theta})\right]\le\mathtt{C}\|\boldsymbol{\theta}-\boldsymbol{\theta}'\|2^{j_l/2}2^{(j_k\wedge j_l)/2}. &(\text{A.7})
\end{aligned}$$

Proof. Observe that if the density of $p_{\mathbf{X}}:\mathbb{R}^p\mapsto\mathbb{R}$ is Lipshitz continuous with Lipshitz constant $L_{p_{\mathbf{X}}}$ and its support contained in a ball of radius $s_{\mathbf{X}}>0$ then the density $p_{\mathbf{X}^\top\boldsymbol{\theta}^*}:\mathbb{R}\mapsto\mathbb{R}$ of $\mathbf{X}^\top\boldsymbol{\theta}\in\mathbb{R}$ is Lipshitz continuous with Lipshitz constant $L_{p_{\mathbf{X}^\top\boldsymbol{\theta}}}\le s_{\mathbf{X}}^pL_{p_{\mathbf{X}}}$. Furthermore for $k,l\in\mathbb{N}$

$$\mathbb{E}[\boldsymbol{e}_k\boldsymbol{e}_l(\mathbf{X}^\top\boldsymbol{\theta})]=\int_{[-s_{\mathbf{X}},s_{\mathbf{X}}]}\boldsymbol{e}_k(x)\boldsymbol{e}_l(x)p_{\mathbf{X}^\top\boldsymbol{\theta}}(x)dx.$$

Denote by $I_k \subset \mathbb{R}$ the support of $\boldsymbol{e}_k(x)$. We write

$$\begin{aligned}
\mathbb{E}[\boldsymbol{e}_k\boldsymbol{e}_l(\mathbf{X}^\top\boldsymbol{\theta})] &= \int_{I_l} \boldsymbol{e}_k(x)\boldsymbol{e}_l(x)p_{\mathbf{X}^\top\boldsymbol{\theta}}(x)dx \\
&= \int_{I_l} \boldsymbol{e}_k(x)\boldsymbol{e}_l(x)p_{\mathbf{X}^\top\boldsymbol{\theta}}(x_0)dx 1_{\{I_l\cap I_k\neq\emptyset\}}(k,l) \\
&\quad + \int_{I_l} \boldsymbol{e}_k(x)\boldsymbol{e}_l(x)\Big(p_{\mathbf{X}^\top\boldsymbol{\theta}}(x) - p_{\mathbf{X}^\top\boldsymbol{\theta}}(x_0)\Big)dx 1_{\{I_l\cap I_k\neq\emptyset\}}(k,l),
\end{aligned}$$

where $x_0 \in I_l$ is the center of the support of $\boldsymbol{e}_l(x)$, which is of length $2^{-j_l}17s_{\mathbf{X}}$ for $l = 2^{j_l}+17j_l-1+r_l \in \mathbb{N}$. Because of orthogonality the first summand on the right-hand side is equal to zero. For the second summand we use the Lipshitz continuity and Cauchy-Schwarz to estimate

$$\begin{aligned}
&|\int_{I_l} \boldsymbol{e}_k(x)\boldsymbol{e}_l(x)\Big(p_{\mathbf{X}^\top\boldsymbol{\theta}}(x) - p_{\mathbf{X}^\top\boldsymbol{\theta}}(x_0)\Big)dx|1_{\{I_l\cap I_k\neq\emptyset\}}(k,l) \\
&\quad\leq s_{\mathbf{X}}^p L_{p_{\mathbf{X}}} 2^{-j_l-1}\int_{I_l} |\boldsymbol{e}_k(x)||\boldsymbol{e}_l(x)|dx 1_{\{I_l\cap I_k\neq\emptyset\}}(k,l) \\
&\quad\leq s_{\mathbf{X}}^p L_{p_{\mathbf{X}}} 2^{-j_l-1}\left(\int_{I_l} \boldsymbol{e}_l(x)^2dx\int_{I_l}\boldsymbol{e}_k(x)^2dx\right)^{1/2} 1_{\{I_l\cap I_k\neq\emptyset\}}(k,l) \\
&\quad\leq s_{\mathbf{X}}^p L_{p_{\mathbf{X}}} 2^{-j_l-1}\left(\int_{I_l} \boldsymbol{e}_k(x)^2dx\right)^{1/2} 1_{\{I_l\cap I_k\neq\emptyset\}}(k,l) \\
&\quad\leq 17s_{\mathbf{X}}^{p+1} L_{p_{\mathbf{X}}}\|\psi\|_\infty 2^{-j_l-1}2^{j_k/2-j_l/2}1_{\{I_l\cap I_k\neq\emptyset\}}(k,l),
\end{aligned}$$

where we used that the $(\boldsymbol{e}_k)$ form an orthonormal basis, that $\|\boldsymbol{e}_k\|_\infty \leq 2^{j_k/2}\|\psi\|_\infty$ and that I_l is of length $2^{-j_l}17s_{\mathbf{X}}$. This gives (A.2). Using that for any $\boldsymbol{\theta} \in W_S$ it holds true that $\|\nabla\Phi(\boldsymbol{\theta}^*)\boldsymbol{\theta}\| \leq \frac{\sqrt{p+2}}{2}\pi$ we estimate similarly to before

$$\begin{aligned}
&|\mathbb{E}[(\mathbf{X}^\top\boldsymbol{\theta})\boldsymbol{e}_k'\boldsymbol{e}_l(\mathbf{X}^\top\boldsymbol{\theta}^\circ)]| \\
&\quad\leq \frac{\sqrt{p+2}}{2}\pi s_{\mathbf{X}}^2\mathbb{E}[|\boldsymbol{e}_k'\boldsymbol{e}_l(\mathbf{X}^\top\boldsymbol{\theta}^\circ)|] \\
&\quad\leq \frac{\sqrt{p+2}}{2}\pi s_{\mathbf{X}}^2\int_{I_l}\boldsymbol{e}_k'(x)\boldsymbol{e}_l(x)p_{\mathbf{X}^\top\boldsymbol{\theta}^\circ}(x)dx \\
&\quad\leq \frac{\sqrt{p+2}}{2}\pi s_{\mathbf{X}}^2\|p_{\mathbf{X}^\top\boldsymbol{\theta}}\|_\infty\left(\int_{I_l}\boldsymbol{e}_k'(x)^2dx\right)^{1/2}\left(\int_{I_l}\boldsymbol{e}_l(x)^2dx\right)^{1/2} \\
&\quad\leq 17\frac{\sqrt{p+2}}{2}\pi\|\psi'\|_\infty s_{\mathbf{X}}^2\|p_{\mathbf{X}^\top\boldsymbol{\theta}}\|_\infty 2^{3j_k/2}2^{-(j_l\vee j_k)/2}1_{\{I_l\cap I_k\neq\emptyset\}}(k,l).
\end{aligned}$$

The bound (A.4) follows with exactly the same calculations. To show (A.5) we calculate with $M_k \overset{\mathrm{def}}{=} \{(x,y) \in \mathbb{R}^2,\, x \in I_k\} \cup \{(x,y) \in \mathbb{R}^2,\, x+y \in I_k\}$ and with $p_{\boldsymbol{\theta},(\boldsymbol{\theta}^\circ-\boldsymbol{\theta})} : \mathbb{R}^2 \to \mathbb{R}_+$ denoting the density of $(\mathbf{X}^\top\boldsymbol{\theta}, \mathbf{X}^\top(\boldsymbol{\theta}^\circ - \boldsymbol{\theta})) \in \mathbb{R}^2$

$$\begin{aligned}
&\mathbb{E}\left[(\boldsymbol{e}_k(\mathbf{X}^\top\boldsymbol{\theta}) - \boldsymbol{e}_k(\mathbf{X}^\top\boldsymbol{\theta}^\circ))(\boldsymbol{e}_l(\mathbf{X}^\top\boldsymbol{\theta}) - \boldsymbol{e}_l(\mathbf{X}^\top\boldsymbol{\theta}^\circ))\right] \\
&\quad = 1_{\{I_k \cap I_l \neq \emptyset\}} \int_{M_k} (\boldsymbol{e}_k(x) - \boldsymbol{e}_k(x+y))(\boldsymbol{e}_l(x) - \boldsymbol{e}_l(x+y)) p_{\boldsymbol{\theta},(\boldsymbol{\theta}^\circ-\boldsymbol{\theta})}(x,y)d(x,y) \\
&\quad \le 1_{\{I_k \cap I_l \neq \emptyset\}} \left(\int_{M_k} (\boldsymbol{e}_k(x) - \boldsymbol{e}_k(x+y))^2 p_{\boldsymbol{\theta},(\boldsymbol{\theta}^\circ-\boldsymbol{\theta})}(x,y)d(x,y)\right)^{1/2} \\
&\qquad \left(\int_{M_l} (\boldsymbol{e}_l(x) - \boldsymbol{e}_l(x+y))^2 p_{\boldsymbol{\theta},(\boldsymbol{\theta}^\circ-\boldsymbol{\theta})}(x,y)d(x,y)\right)^{1/2}.
\end{aligned}$$

We estimate separately

$$\begin{aligned}
&\int_{M_k} (\boldsymbol{e}_k(x) - \boldsymbol{e}_k(x+y))^2 p_{\boldsymbol{\theta},(\boldsymbol{\theta}^\circ-\boldsymbol{\theta})}(x,y)d(x,y) \\
&\qquad \le 2^{3j_k}\|\psi''\|_\infty^2 \int_{M_k} y^2 p_{\boldsymbol{\theta},(\boldsymbol{\theta}^\circ-\boldsymbol{\theta})}(x,y)d(x,y),
\end{aligned}$$

Note that $p_{\boldsymbol{\theta},(\boldsymbol{\theta}^\circ-\boldsymbol{\theta})}(x,y) > 0$ only for $|y| \le \|\boldsymbol{\theta} - \boldsymbol{\theta}^\circ\|(s_{\mathbf{X}} + h)$, where we suppress h in the following such that

$$\begin{aligned}
&\int_{M_k} (\boldsymbol{e}_k(x) - \boldsymbol{e}_k(x+y))^2 p_{\boldsymbol{\theta},(\boldsymbol{\theta}^\circ-\boldsymbol{\theta})}(x,y)d(x,y) \\
&\quad \le \|\boldsymbol{\theta} - \boldsymbol{\theta}^\circ\|^2 2^{3j_k}\|\psi''\|_\infty^2 s_{\mathbf{X}}^2 \\
&\qquad \left(\int_{\mathbb{R}}\int_{I_k - x} p_{\boldsymbol{\theta},(\boldsymbol{\theta}^\circ-\boldsymbol{\theta})}(x,y)dydx + \int_{I_k}\int_{\mathbb{R}} p_{\boldsymbol{\theta},(\boldsymbol{\theta}^\circ-\boldsymbol{\theta})}(x,y)dydx\right) \\
&\quad \le \|\boldsymbol{\theta} - \boldsymbol{\theta}^\circ\|^2 2^{3j_k}\|\psi''\|_\infty^2 s_{\mathbf{X}}^2 \\
&\qquad \left(\int_{\mathbb{R}} \mathbb{P}\left\{(\boldsymbol{\theta}^\circ - \boldsymbol{\theta})^\top\mathbf{X} \in I_k - x | \boldsymbol{\theta}^\top\mathbf{X} = x\right\} p_{\boldsymbol{\theta}}(x)dx + \int_{I_k} p_{\boldsymbol{\theta}}(x)dx\right).
\end{aligned}$$

represent $\boldsymbol{\theta}^\circ = \alpha\boldsymbol{\theta} + \beta\boldsymbol{\theta}'$ where $\boldsymbol{\theta}' \perp \boldsymbol{\theta}$ with $\|\boldsymbol{\theta}^\circ\| = 1$. Then we find with condition $(\mathbf{Cond}_{\mathbf{X}})$

$$\begin{aligned}
&\mathbb{P}\left\{(\boldsymbol{\theta}^\circ - \boldsymbol{\theta})^\top\mathbf{X} \in I_k - x | \boldsymbol{\theta}^\top\mathbf{X} = x\right\} = \mathbb{P}\left\{\boldsymbol{\theta}'^\top\mathbf{X} \in \frac{1}{\beta}(I_k - (1-\alpha)x) | \boldsymbol{\theta}^\top\mathbf{X} = x\right\} \\
&\quad \le \left\|\frac{p_{\boldsymbol{\theta}',\boldsymbol{\theta}}}{p_{\boldsymbol{\theta}}}\right\|_\infty \lambda\left\{\frac{1}{\beta}(I_k - (1-\alpha)x)\right\} \le \mathtt{C}2^{-j_k}/\|\boldsymbol{\theta} - \boldsymbol{\theta}^\circ\|.
\end{aligned}$$

With the bound $p_{\boldsymbol{\theta}}(x) \le \mathtt{C}$ we find (since $\|\boldsymbol{\theta} - \boldsymbol{\theta}^{\circ}\| < \sqrt{2}$)

$$\begin{aligned}\int_{M_k} &(\boldsymbol{e}_k(x) - \boldsymbol{e}_k(x+y))^2 p_{\boldsymbol{\theta},(\boldsymbol{\theta}^{\circ}-\boldsymbol{\theta})}(x,y)d(x,y)\\ &\le \mathtt{C}\|\boldsymbol{\theta} - \boldsymbol{\theta}^{\circ}\| 2^{2j_k}\|\psi''\|_{\infty}^2 s_{\mathbf{X}}^4 17^2,\end{aligned}$$

which yields (A.5). With the same calculations we can show (A.6). with $M_{l,k} \overset{\text{def}}{=} \{(x,y) \in I_k \times \mathbb{R},\, x \in I_l \cap I_k\} \cup \{(x,y) \in I_k \times \mathbb{R},\, x+y \in I_l\}$

$$\begin{aligned}&\mathbb{E}\left[\left(\boldsymbol{e}_l(\mathbf{X}^\top\boldsymbol{\theta}) - \boldsymbol{e}_l(\mathbf{X}^\top\boldsymbol{\theta}_m^*)\right)\boldsymbol{e}_k(\mathbf{X}^\top\boldsymbol{\theta})\right]\\ &\quad\le \left(\int_{M_{l,k}} \left(\boldsymbol{e}_l(x) - \boldsymbol{e}_l(x+y)\right)^2 p_{\boldsymbol{\theta},(\boldsymbol{\theta}_m^*-\boldsymbol{\theta})}(x,y)d(x,y)\right)^{1/2}\\ &\quad\left(\int_{M_{l,k}} \boldsymbol{e}_k^2(x)p_{\boldsymbol{\theta},(\boldsymbol{\theta}_m^*-\boldsymbol{\theta})}(x,y)d(x,y)\right)^{1/2}.\end{aligned}$$

We have by (A.5)

$$\begin{aligned}\int_{M_{l,k}} &\left(\boldsymbol{e}_l(x) - \boldsymbol{e}_l(x+y)\right)^2 p_{\boldsymbol{\theta},(\boldsymbol{\theta}_m^*-\boldsymbol{\theta})}(x,y)d(x,y)\\ &\le 2^{2j_l}\|\boldsymbol{\theta} - \boldsymbol{\theta}_m^*\|^2\|\psi'\|^2 s_{\mathbf{X}}^4 17^2\mathtt{C}.\end{aligned}$$

As above we can bound

$$\begin{aligned}&\int_{M_{l,k}} \boldsymbol{e}_k^2(x)p_{\boldsymbol{\theta},(\boldsymbol{\theta}'-\boldsymbol{\theta})}(x,y)d(x,y)\\ &\quad= \int_{\mathbb{R}} \boldsymbol{e}_k^2(x)\int_{I_l - x} p_{\boldsymbol{\theta},(\boldsymbol{\theta}'-\boldsymbol{\theta})}(x,y)d(x,y)\\ &\quad\quad+ \int_{I_l\cap I_k} \boldsymbol{e}_k^2(x)\int_{\mathbb{R}} p_{\boldsymbol{\theta},(\boldsymbol{\theta}'-\boldsymbol{\theta})}(x,y)d(x,y)\\ &\quad\le \int_{\mathbb{R}} \boldsymbol{e}_k^2(x)\mathbb{P}\left\{(\boldsymbol{\theta}'-\boldsymbol{\theta})^\top\mathbf{X} \in (I_l - x)\middle|\, \boldsymbol{\theta}^\top\mathbf{X} = x\right\}p_{\boldsymbol{\theta}}(x)d(x)\\ &\quad\quad+ \int_{I_l\cap I_k} \boldsymbol{e}_k^2(x)p_{\boldsymbol{\theta}}(x)d(x)\\ &\quad\le \frac{\mathtt{C}}{\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|}2^{-j_l}\mathtt{C} + 2^{-j_l}2^{(j_k\wedge j_l)}\|\psi\|_{\infty}^2.\end{aligned}$$

□

Lemma A.4. *For any* $(\boldsymbol{\theta}, \boldsymbol{\eta}) \in \mathbb{R}^{p+m}$

$$
\begin{aligned}
\|\boldsymbol{e}(\boldsymbol{x})\| &\leq \mathtt{C}\|\psi\|_\infty \sqrt{m}, && \text{(A.8)}\\
\left|\boldsymbol{f}_{\boldsymbol{\eta}}(\boldsymbol{x})\right| &\leq \mathtt{C}\|\psi\|_\infty \sqrt{m}\|\boldsymbol{\eta}\|, &&\\
\|\boldsymbol{e}'(\boldsymbol{x})\| &\leq \sqrt{17}\|\psi'\| m^{3/2}. && \text{(A.9)}
\end{aligned}
$$

Proof. Clearly $\left|\boldsymbol{f}_{\boldsymbol{\eta}}(\boldsymbol{x})\right| \leq \|\boldsymbol{\eta}\|\|\boldsymbol{e}(\mathbf{X}_i^\top \boldsymbol{\theta}_m^*)\|$. Because of the wavelet structure and the choice $m = 2^{j_m} + j_m 17 - 1$ we have for each $j = 0, \ldots, j_m - 1$ that

$$
\begin{aligned}
&|M(j)| && \text{(A.10)}\\
&\overset{\text{def}}{=} \left|\left\{k \in \{(2^j + j17 - 1, \ldots, 2^{j+1} + (j+1)17 - 2\} : |\boldsymbol{e}_k(\boldsymbol{x})| \neq 0\right\}\right| \leq 17.
\end{aligned}
$$

This implies

$$
\begin{aligned}
\|\boldsymbol{e}(\boldsymbol{x})\| &= \left(\sum_{k=0}^{m-1} |\boldsymbol{e}_k(\boldsymbol{x})|^2\right)^{1/2} = \left(\sum_{j=0}^{j_m-1} \sum_{k \in M(j)} |\boldsymbol{e}_k(\boldsymbol{x})|^2\right)^{1/2}\\
&\leq \sqrt{17}\|\psi\|_\infty \left(\sum_{j=0}^{j_m-1} 2^j\right)^{1/2} = \sqrt{17}\|\psi\|_\infty 2^{j_m/2} \leq \sqrt{17}\|\psi\|_\infty \sqrt{m}.
\end{aligned}
$$

The proof of (A.9) works analogously. □

A.4. Lower bound for the information operator

Lemma A.5. *Under* $(\mathbf{Cond}_{\mathbf{X},\boldsymbol{e}})$, $(\mathbf{Cond}_{\mathbf{X}\boldsymbol{\theta}^*})$ *and (model bias) we find for all* $m \in \mathbb{N} \cup \{\infty\}$ *that* $\mathcal{D}_m(\boldsymbol{\upsilon}^*) \geq c_{\mathcal{D}}*$ *with some constant* $c_{\mathcal{D}}* > 0$.

Remark A.2. The constant $c_{\mathcal{D}}* > 0$ is specified – to some extend – in the proof.

Proof. We represent for any $\boldsymbol{\gamma} \in \mathbb{R}^{p^*}$ with $\|\boldsymbol{\gamma}\| = 1$

$$
\begin{aligned}
&\boldsymbol{\gamma}^\top \mathcal{D}_m \boldsymbol{\gamma}\\
&\quad = n \lim_{h \to 0} \frac{1}{h^2}\Bigg(\mathbb{E}\left[\left(g(\mathbf{X}) - \sum_{k=1}^{m} (\eta_k^* + h\gamma_{p+k}) \boldsymbol{e}_k(\mathbf{X}^\top(\boldsymbol{\theta}^* + h\mathit{\Pi}_{\boldsymbol{\theta}}\boldsymbol{\gamma}))\right)^2\right]\\
&\qquad - \mathbb{E}\left[(g(\mathbf{X}) - \mathbb{E}[g(\mathbf{X})|\mathbf{X}^\top\boldsymbol{\theta}^*])^2\right]\Bigg).
\end{aligned}
$$

Using the properties of conditional expectation we can write

$$
\begin{aligned}
&\mathbb{E}\left[\left(g(\mathbf{X})-\sum_{k=1}^{m}(\eta_k^*+h\gamma_{p+k})\boldsymbol{e}_k(\mathbf{X}^\top(\boldsymbol{\theta}^*+h\varPi_{\boldsymbol{\theta}}\boldsymbol{\gamma}))\right)^2\right]\\
&\quad=\mathbb{E}\Bigg[\bigg(\mathbb{E}[g(\mathbf{X})|\mathbf{X}^\top(\boldsymbol{\theta}^*+\varPi_{\boldsymbol{\theta}}\boldsymbol{\gamma})]\\
&\qquad\qquad\qquad-\sum_{k=1}^{m}(\eta_k^*+h\gamma_{p+k})\boldsymbol{e}_k(\mathbf{X}^\top(\boldsymbol{\theta}^*+h\varPi_{\boldsymbol{\theta}}\boldsymbol{\gamma}))\bigg)^2\Bigg]\\
&\qquad+\mathbb{E}\left[(g(\mathbf{X})-\mathbb{E}[g(\mathbf{X})|\mathbf{X}^\top(\boldsymbol{\theta}^*+h\varPi_{\boldsymbol{\theta}}\boldsymbol{\gamma})])^2\right]
\end{aligned}
$$

Using assumption (model bias) we find

$$
\begin{aligned}
&\boldsymbol{\gamma}^\top\mathcal{D}_m\boldsymbol{\gamma}\geq n\mathtt{b}_\theta\|\varPi_{\boldsymbol{\theta}}\boldsymbol{\gamma}\|^2\\
&\qquad+n\lim_{h\to 0}\frac{1}{h^2}\mathbb{E}\bigg(\mathbb{E}[g(\mathbf{X})|\mathbf{X}^\top(\boldsymbol{\theta}^*+h\varPi_{\boldsymbol{\theta}}\boldsymbol{\gamma})]\\
&\qquad\qquad\qquad-\sum_{k=1}^{m}(\eta_k^*+h\gamma_{p+k})\boldsymbol{e}_k(\mathbf{X}^\top(\boldsymbol{\theta}^*+h\varPi_{\boldsymbol{\theta}}\boldsymbol{\gamma}))\bigg)^2.
\end{aligned}
$$

In case that $\|\varPi_{\boldsymbol{\theta}}\boldsymbol{\gamma}\|^2\geq\tau^2>0$ with some $\tau>0$ this implies $\mathcal{D}_m\geq\mathtt{b}_\theta\tau^2$. Assume $\|\varPi_{\boldsymbol{\theta}}\boldsymbol{\gamma}\|^2\leq\tau^2$. Using the smoothness of the density $p_{\mathbf{X}}$ and of g we find with some constant

$$
\left|\mathbb{E}[g(\mathbf{X})|\mathbf{X}^\top(\boldsymbol{\theta}^*+h\varPi_{\boldsymbol{\theta}}\boldsymbol{\gamma})]-\mathbb{E}[g(\mathbf{X})|\mathbf{X}^\top\boldsymbol{\theta}^*]\right|\leq\mathtt{C}\|\varPi_{\boldsymbol{\theta}}\boldsymbol{\gamma}\|\leq n\mathtt{C}h\tau.
$$

Furthermore we show in Lemma A.24 that with some $\mathtt{b}^*>0$ and $Q>0$

$$
\inf_{\boldsymbol{\upsilon}\in\varUpsilon_m}\mathbb{P}\left(\left|\mathbb{E}[g(\mathbf{X})|\mathbf{X}^\top\boldsymbol{\theta}^*]-\sum_{k=1}^{m}(\eta_k)\boldsymbol{e}_k(\mathbf{X}^\top\boldsymbol{\theta})\right|\geq\mathtt{b}^*\|\boldsymbol{\upsilon}-\boldsymbol{\upsilon}^*\|\right)\geq Q>0.
$$

Remark A.3. A close look at the proof of Lemma A.24 reveals that the claim can be shown with $\|\boldsymbol{\upsilon}-\boldsymbol{\upsilon}^*\|$ instead of $\|\mathcal{D}(\boldsymbol{\upsilon}-\boldsymbol{\upsilon}^*)\|$ on the right-hand side with the same arguments.

Consequently

$$
\begin{aligned}
&\mathbb{E}\left[\left(\mathbb{E}[g(\mathbf{X})|\mathbf{X}^\top(\boldsymbol{\theta}^*+h\varPi_{\boldsymbol{\theta}}\boldsymbol{\gamma})]-\sum_{k=1}^{m}(\eta_k^*+h\gamma_{p+k})\boldsymbol{e}_k(\mathbf{X}^\top(\boldsymbol{\theta}^*+h\varPi_{\boldsymbol{\theta}}\boldsymbol{\gamma}))\right)^2\right]\\
&\quad\geq Qh^2(\mathtt{b}^*-\mathtt{C}\tau)^2.
\end{aligned}
$$

Setting $\tau\leq\mathtt{b}^*/(2\mathtt{C})$ gives the claim. □

A.5. Regularity

We represent the full parameter $\boldsymbol{\upsilon} \in \mathbb{R}^\infty$ in the form

$$\boldsymbol{\upsilon} = (\boldsymbol{\theta}, \boldsymbol{f}) = (\Pi_{p^*}\boldsymbol{\upsilon}, \boldsymbol{\kappa}) = (\boldsymbol{\theta}, \Pi_m\boldsymbol{\eta}, \boldsymbol{\kappa}) \in \mathbb{R}^{p+m} \times l^2.$$

where $\boldsymbol{\kappa} = (\eta_{m+1}, \ldots)^\top$ stands for the remaining components of the expansion (2.1). Consider the following block representations of of the full *information operator*: Remember the representation the full operator $\mathcal{D} \in L(l^2, \{(x_k)_{k\in\mathbb{N}}, x \in \mathbb{R}\})$ in block form

$$\mathcal{D}^2(\boldsymbol{\upsilon}^*) = \begin{pmatrix} \mathrm{D}^2 & A \\ A & \mathcal{H}^2 \end{pmatrix} = \begin{pmatrix} \mathcal{D}_m^2 & A_{\boldsymbol{\upsilon\kappa}} \\ A_{\boldsymbol{\upsilon\kappa}} & \mathcal{H}_{\boldsymbol{\kappa\kappa}}^2 \end{pmatrix} = \begin{pmatrix} \mathrm{D}^2 & A_m & A_{\boldsymbol{\theta\kappa}} \\ A_m & \mathcal{H}_m & A_{\boldsymbol{\eta\kappa}} \\ A_{\boldsymbol{\eta\kappa}} & A_{\boldsymbol{\theta\kappa}} & \mathcal{H}_{\boldsymbol{\kappa\kappa}}^2 \end{pmatrix}.$$

where $A_{\boldsymbol{\upsilon\kappa}}$ is a – possibly unbounded – operator from l^2 to $\mathbb{R}^{p+m}$.

Lemma A.6. *Assume that the density $p_{\mathbf{X}} : \mathbb{R}^p \to \mathbb{R}$ is Lipschitz continuous and that the $\mathbf{X} \in \mathbb{R}$ are bounded by some constant $s_{\mathbf{X}} > 0$. Then using our orthogonal and sufficiently smooth wavelet basis we get for any $\lambda \in [0,1]$*

$$\begin{aligned} \|\mathcal{H}_m^{1/2}\boldsymbol{\kappa}^*\|^2 \; &< \; \mathtt{C}nm^{-2\alpha}, \\ \alpha(m) \overset{\text{def}}{=} \; &\|\mathcal{D}_m^{-1}A_{\boldsymbol{\upsilon\kappa}}\boldsymbol{\kappa}^*\| \le \mathtt{C}\sqrt{n}\left(m^{-(\alpha+1/2)} + \mathtt{C}_{bias}m^{-(\alpha-1)}\right), \\ \tau(m) \overset{\text{def}}{=} \; &\|\mathcal{D}_m^{-1}\nabla_{\boldsymbol{\upsilon\kappa}}\mathbb{E}[\mathcal{L}\big((\Pi_{p^*}\boldsymbol{\upsilon}^*, \lambda\boldsymbol{\kappa}^*) - A_{\boldsymbol{\upsilon\kappa}}\big)]\boldsymbol{\kappa}^*\| \le \mathtt{C}m^{-2\alpha+1/2}\sqrt{n}, \\ 0 \; &= \; \left|\boldsymbol{\kappa}^{*\top}(\mathcal{H}_m - \nabla_{\boldsymbol{\kappa\kappa}}\mathbb{E}\mathcal{L}(\Pi_{p^*}\boldsymbol{\upsilon}^*, \lambda\boldsymbol{\kappa}^*))\boldsymbol{\kappa}^*\right|, \end{aligned}$$

if $\mathtt{C}_{bias} = 0$ one can bound with some $\mathtt{C} > 0$

$$\beta(m) \overset{\text{def}}{=} \|\mathcal{D}_m^{-1}A_{\boldsymbol{\upsilon\kappa}}\mathcal{H}_m^{-1}\| \le \mathtt{C}m^{-1/2}.$$

Furthermore we find that

$$\|D^2\| \le \mathtt{C}np$$

Proof. Throughout this proof we assume that $m > 0$ is large enough to ensure for all $k, l > m$ that $I_k \cap I_l \subseteq [-s_{\mathbf{X}} - c_B, s_{\mathbf{X}} + c_B]$. We have that

$$\|\mathcal{D}_m^{-1}A_{\boldsymbol{\upsilon\kappa}}\boldsymbol{\kappa}^*\| \le \|\mathcal{D}_m^{-1}\|\|A_{\boldsymbol{\upsilon\kappa}}\boldsymbol{\kappa}^*\|.$$

Due to Lemma A.5

$$\|\mathcal{D}_m^{-1}\| \le \frac{1}{c_{\mathcal{D}}\sqrt{n}}.$$

And we have by definition that for any $\boldsymbol{\upsilon} = (\boldsymbol{\theta}, \boldsymbol{\eta}) \in W_S \times \mathbb{R}^m$

$$\frac{1}{n}|\boldsymbol{\upsilon}^\top A_{\boldsymbol{\upsilon}\boldsymbol{\kappa}}\boldsymbol{\kappa}^*| \leq \frac{1}{n}|\boldsymbol{\theta} A_{\boldsymbol{\theta}\boldsymbol{\kappa}}\boldsymbol{\kappa}^*| + \frac{1}{n}|\boldsymbol{\eta} A_{\boldsymbol{\eta}\boldsymbol{\kappa}}\boldsymbol{\kappa}^*|.$$

We first analyze the second summand

$$\frac{1}{n}\boldsymbol{\eta} A_{\boldsymbol{\eta}\boldsymbol{\kappa}}\boldsymbol{\kappa}^* \;=\; \sum_{l=m+1}^{\infty} \eta_l^* \sum_{k=1}^{m} \eta_k \mathbb{E}[\boldsymbol{e}_k \boldsymbol{e}_l (\mathbf{X}^\top \boldsymbol{\theta}^*)].$$

We use (A.2) from Lemma A.3 to find

$$\frac{1}{n}\boldsymbol{\eta} A_{\boldsymbol{\eta}\boldsymbol{\kappa}}\boldsymbol{\kappa}^*| \;\leq\; 17 s_{\mathbf{X}}^{p+1} L_{p\mathbf{X}} \|\psi\|_\infty \sum_{l=m+1}^{\infty}\sum_{k=1}^{m} |\eta_l^*||\eta_k| 2^{-j_l-1} 2^{j_k/2 - j_l/2} 1_{\{I_l \cap I_k \neq \emptyset\}}(k,l).$$

Note that for each $j_k = 0, \ldots, j_m$ there exists at most 17 $r_k(l) \in \{0, \ldots, 2^{j_k}+16\}$ with $I_l \cap I_k \neq \emptyset$. Remember that $m = 2^{j_m} + j_m 17 - 1$ and note that $2^{j_m} \leq m$. This implies using the Cauchy-Schwarz inequality and that $\|\boldsymbol{\eta}\| = 1$

$$\begin{aligned}
|\frac{1}{n}\boldsymbol{\eta} A_{\boldsymbol{\eta}\boldsymbol{\kappa}}\boldsymbol{\kappa}^*| \;&\leq\; 17 s_{\mathbf{X}}^{p+1} L_{p\mathbf{X}} \|\psi^2\|_\infty \\
&\qquad \sum_{l=m+1}^{\infty}\sum_{k=1}^{m} |\eta_l^*||\eta_k| 2^{-j_l-1} 2^{j_k/2 - j_l/2} 1_{\{I_l \cap I_k \neq \emptyset\}}(k,l) \\
&\leq\; 17 s_{\mathbf{X}}^{p+1} L_{p\mathbf{X}} \|\psi^2\|_\infty \sum_{l=m+1}^{\infty} |\eta_l^*| 2^{-3j_l/2} \left(\sum_{k=1}^{m} 2^{j_k} 1_{\{I_l \cap I_k \neq \emptyset\}}(k,l) \right)^{1/2} \\
&\leq\; 17\sqrt{17} s_{\mathbf{X}}^{p+1} L_{p\mathbf{X}} \|\psi^2\|_\infty \sum_{l=m+1}^{\infty} |\eta_l^*| 2^{-3j_l/2} \left(\sum_{j_k=0}^{j_m-1} 2^{j_k} \right)^{1/2} \\
&\leq\; 17^{3/2} s_{\mathbf{X}}^{p+1} L_{p\mathbf{X}} \|\psi^2\|_\infty \sqrt{m} \left(\sum_{l=m+1}^{\infty} |\eta_l^*|^2 \right)^{1/2} \left(\sum_{l=m}^{\infty} 2^{-3j_l} \right)^{1/2}.
\end{aligned}$$

By assumption $\mathbf{Cond}_{\boldsymbol{\upsilon}^*}$

$$\left(\sum_{l=m+1}^{\infty} |\eta_l^*|^2 \right)^{1/2} \leq m^{-\alpha} \left(\sum_{l=m+1}^{\infty} l^{2\alpha} |\eta_l^*|^2 \right)^{1/2} \leq m^{-\alpha} \mathtt{C}_{\|\boldsymbol{f}^*\|}.$$

Since $m = 2^{j_m} + j_m 17 - 1$ and $l = 2^{j_l} + j_l 17 - 1 + r_l$ with $r_l \in \{0, \ldots, 2^{j_l} + 16\}$

$$\begin{aligned}
\left(\sum_{l=m+1}^{\infty} 2^{-3j_l} \right)^{1/2} \;&=\; \left(\sum_{j_l = j_m}^{\infty} C(m) 2^{j_l} 2^{-3j_l} \right)^{1/2} \\
&=\; C(m)^{1/2} 2^{-j_m} 2 \leq \sqrt{2} C(m)^{3/2} m^{-1},
\end{aligned}$$

with

$$C(m) = \frac{2^{j_m} + j_m 17 - 1}{2^{j_m}} \le 34.$$

Consequently

$$|\frac{1}{n}\boldsymbol{\eta} A_{\boldsymbol{\eta}\boldsymbol{\kappa}}\boldsymbol{\kappa}^*| \le \sqrt{2}17^3 \mathtt{C}_{\|\boldsymbol{f}^*\|} s_{\mathbf{X}}^{p+1} L_{p_{\mathbf{X}}} \|\psi^2\|_\infty m^{-\alpha-1/2}.$$

For the second summand we remind the reader that

$$\begin{aligned} A_{\boldsymbol{\theta}\boldsymbol{\kappa}} &= n a_{\boldsymbol{\theta}\boldsymbol{\kappa}}, \\ a_{\boldsymbol{\theta}\boldsymbol{\kappa}} &= \mathbb{E}[\boldsymbol{f}'_{\boldsymbol{\eta}^*}(\mathbf{X}^\top\boldsymbol{\theta}^*)\nabla\Phi_{\boldsymbol{\theta}^*}^\top \mathbf{X}(\boldsymbol{e}_{m+1}(\mathbf{X}^\top\boldsymbol{\theta}^*), \ldots)], \end{aligned}$$

Similarly to the first summand we get by the dominated convergence theorem

$$\boldsymbol{\theta} a_{\boldsymbol{\theta}\boldsymbol{\kappa}}\boldsymbol{\kappa}^* = \sum_{k=1}^{\infty}\sum_{l=m+1}^{\infty} \eta_k^*\eta_l^* \mathbb{E}[(\mathbf{X}^\top\nabla\Phi(\boldsymbol{\theta}^*)\boldsymbol{\theta})\boldsymbol{e}'_k\boldsymbol{e}_l(\mathbf{X}^\top\boldsymbol{\theta}^*)]1_{\{I_l\cap I_k\neq\emptyset\}}(k,l).$$

To justify the exchange of summation and expectation note that for each $l \in \mathbb{N}$

$$\begin{aligned} &\mathbb{E}[|(\mathbf{X}^\top\nabla\Phi(\boldsymbol{\theta}^*)\boldsymbol{\theta})\boldsymbol{e}_l f'_{\boldsymbol{\eta}^*}(\mathbf{X}^\top\boldsymbol{\theta}^*)|] \\ &\quad\le \|\nabla\Phi(\boldsymbol{\theta}^*)\boldsymbol{\theta}\| s_{\mathbf{X}} 2^{j_l/2}\mathbb{E}[|f'_{\boldsymbol{\eta}^*}(\mathbf{X}^\top\boldsymbol{\theta}^*)|] \\ &\quad\le \|\nabla\Phi(\boldsymbol{\theta}^*)\boldsymbol{\theta}\| s_{\mathbf{X}} 2^{j_l/2}\mathbb{E}\left[\left|\sum_{k=1}^{\infty}\eta_k^*\boldsymbol{e}'_k(\mathbf{X}^\top\boldsymbol{\theta}^*)\right|\right] \\ &\quad\le \|\nabla\Phi(\boldsymbol{\theta}^*)\boldsymbol{\theta}\| s_{\mathbf{X}} 2^{j_l/2}\left(\sum_{k=1}^{\infty} l^{2\alpha}\eta_k^{*2}\right)^{1/2}\left(\sum_{k=1}^{\infty} l^{-2\alpha}2^{3j_k}\|\psi'\|^2\right)^{1/2} \\ &\quad\le \|\nabla\Phi(\boldsymbol{\theta}^*)\boldsymbol{\theta}\| s_{\mathbf{X}} \mathtt{C}_{\|\boldsymbol{f}^*\|}\|\psi'\|_\infty 2^{j_l/2}\left(\frac{17}{2}\sum_{j=0}^{\infty} l^{-2\alpha}2^{4j}\right)^{1/2} < \infty. \end{aligned}$$

The exchange of the order of summation is justified by the subsequent bounds and again the dominated convergence theorem. We again use Lemma A.3 to find with (A.3) and with similar arguments to those from above

$$\begin{aligned} |\boldsymbol{\theta} a_{\boldsymbol{\theta}\boldsymbol{\kappa}}\boldsymbol{\kappa}^*| &\le 17\frac{\sqrt{p+2}}{2}\pi\|\psi'\|_\infty s_{\mathbf{X}}^2\|p_{\mathbf{X}}\|_\infty \\ &\qquad \sum_{k=1}^{\infty}\eta_k^* 2^{3j_k/2}\sum_{l=m+1}^{\infty}\eta_l^* 2^{-(j_l\vee j_k)/2}1_{\{I_l\cap I_k\neq\emptyset\}}(k,l) \end{aligned}$$

$$\leq 17\frac{\sqrt{p+2}}{2}\pi\|\psi'\|_\infty s_{\mathbf{X}}^2\|p_{\mathbf{X}}\|_\infty\sum_{k=1}^{\infty}\eta_k^* k^{3/2}\left(\sum_{l=m+1}^{\infty}l^{2\alpha}\eta_l^{*2}\right)^{1/2}$$
$$\left(\sum_{j_l=j_m+1}^{\infty}\sum_{r_l=0}^{2^{j_l}+16}2^{-2\alpha j_l}2^{-(j_l\vee j_k)}1_{\{I_l\cap I_k\neq\emptyset\}}(k,l)\right)^{1/2}.$$

We have due to (3.2) that

$$\sum_{j_l=j_m+1}^{\infty}\sum_{r_l=0}^{2^{j_l}+16}2^{-2\alpha j_l}2^{-(j_l\vee j_k)}1_{\{I_l\cap I_k\neq\emptyset\}}(k,l)$$
$$= \sum_{j_l=j_m+1}^{\infty}2^{-2\alpha j_l}2^{-(j_l\vee j_k)}\sum_{r_l=0}^{2^{j_l}+16}1_{\{I_l\cap I_k\neq\emptyset\}}(k,l)$$
$$= \sum_{j_l=j_m+1}^{\infty}2^{-2\alpha j_l}2^{-(j_l\vee j_k)}$$
$$\left|\left\{l=2^{j_l}+17j_l-1+r_l\middle| r_l\in\{0,\ldots,2^{j_l}+16\},\, I_l\cap I_k\neq\emptyset\right\}\right|$$
$$= \sum_{j_l=j_m+1}^{\infty}2^{-(2\alpha+1)j_l}2^{-(j_k-j_l)_+}\ulcorner 2^{(j_l-j_k)}17\urcorner$$
$$\leq 2^{-(2\alpha+1)j_m}18 \leq 17m^{-(2\alpha+1)}18.$$

Which gives

$$|\boldsymbol{\theta}a_{\boldsymbol{\theta}\boldsymbol{\kappa}}\boldsymbol{\kappa}^*| \leq 17^{3/2}\sqrt{18}\frac{\sqrt{p+2}}{2}\pi\|\psi'\|_\infty s_{\mathbf{X}}^2\|p_{\mathbf{X}}\|_\infty \mathtt{C}_{\|\boldsymbol{f}^*\|}m^{-\alpha-1/2}$$
$$\left(\sum_{k=1}^{\infty}\eta_k^{*2}k^{2\alpha}\right)^{1/2}\left(\sum_{k=1}^{\infty}k^{-(2\alpha-3)}\right)^{1/2}$$
$$\leq 17^{3/2}\sqrt{18}\frac{\sqrt{p+2}}{2}\pi\|\psi'\|_\infty s_{\mathbf{X}}^2\|p_{\mathbf{X}}\|_\infty \mathtt{C}_{\|\boldsymbol{f}^*\|}^2$$
$$\sqrt{(2\alpha-3)/(2\alpha-4)}m^{-(\alpha+1/2)},$$

since $\alpha > 2$ such that $\sum_{k=1}^{\infty}k^{-(2\alpha-3)} < (2\alpha-3)/(2\alpha-4)$.

Furthermore with $\boldsymbol{\theta}^\circ = \nabla\mathit{\Phi}_{\boldsymbol{\theta}^*}\boldsymbol{\theta} \in \boldsymbol{\theta}^{*\perp}$

$$|\boldsymbol{\theta}b_{\boldsymbol{\theta}\boldsymbol{\kappa}}\boldsymbol{\kappa}^*| = \left|\mathbb{E}\left[(\boldsymbol{f}_{\boldsymbol{\eta}^*}(\mathbf{X}^\top\boldsymbol{\theta}^*)-g(\mathbf{X}))\mathbf{X}^\top\boldsymbol{\theta}^\circ\sum_{k=m}^{\infty}\eta_k^*\boldsymbol{e}_k'(\mathbf{X}^\top\boldsymbol{\theta}^*)\right]\right|$$
$$\leq \mathtt{C}_{bias}\frac{\sqrt{p+2}}{2}\pi s_{\mathbf{X}}\mathbb{E}\left[\left|\boldsymbol{f}_{\boldsymbol{\kappa}^*}'(\mathbf{X}^\top\boldsymbol{\theta}^*)\right|\right].$$

We bound

$$\begin{aligned}\mathbb{E}\left[\left|\boldsymbol{f}'_{\boldsymbol{\kappa}^*}(\mathbf{X}^\top\boldsymbol{\theta}^*)\right|\right] &\le \sqrt{17}\|\psi'\|_\infty\left(\sum_{k=m}^\infty \eta_k^{*2}k^{2\alpha}\right)\left(\sum_{j=j_m+1}^\infty 2^{-(2\alpha-3)j}\right)\\ &\le \mathtt{C}(m)\mathtt{C}_{\|\boldsymbol{\eta}^*\|}\sqrt{17}\|\psi'\|_\infty m^{-(\alpha-3/2)}<\infty.\end{aligned}$$

We can exchange summation and expectation to find

$$\mathbb{E}\left[\left|\boldsymbol{f}'_{\boldsymbol{\kappa}^*}(\mathbf{X}^\top\boldsymbol{\theta}^*)\right|\right]=\sum_{k=m}^\infty \eta_k^*\mathbb{E}[\left|\boldsymbol{e}'_k(\mathbf{X}^\top\boldsymbol{\theta}^*)\right|].$$

We estimate

$$\begin{aligned}\mathbb{E}\left[\left|\boldsymbol{e}'_k(\mathbf{X}^\top\boldsymbol{\theta}^*)\right|\right] &= \int_{\mathbb{R}}|\boldsymbol{e}'_k(x)p_{\mathbf{X}^\top\boldsymbol{\theta}^*}(x)|\,dx\\ &\le \left(\int_{I_k}\boldsymbol{e}'_k(x)^2dx\right)^{1/2}\left(\int_{I_k}p^2_{\mathbf{X}^\top\boldsymbol{\theta}^*}(x)dx\right)^{1/2}\\ &\le \|\psi'\|_\infty\mathtt{C}_d2^{j_k/2}.\end{aligned}$$

Such that

$$\begin{aligned}\mathbb{E}\left[\left|\boldsymbol{f}'_{\boldsymbol{\kappa}^*}(\mathbf{X}^\top\boldsymbol{\theta}^*)\right|\right] &\le \mathtt{C}(m)\mathtt{C}_{\|\boldsymbol{\eta}^*\|}\mathtt{C}_d\|\psi'\|_\infty\sum_{k=m}^\infty 2^{j_k/2}\eta_k^*\\ &\le \mathtt{C}(m)\mathtt{C}_{\|\boldsymbol{\eta}^*\|}\mathtt{C}_d\|\psi'\|_\infty\left(\sum_{j=j_m+1}^\infty 2^{-2(\alpha-1)j}\right)^{1/2}\\ &\le \mathtt{C}(m)\mathtt{C}_{\|\boldsymbol{\eta}^*\|}\mathtt{C}_d\|\psi'\|_\infty m^{-(\alpha-1)}.\end{aligned}$$

Collecting both summands

$$\|\mathcal{D}_m^{-1}A_{\boldsymbol{\upsilon}\boldsymbol{\kappa}}\boldsymbol{\kappa}^*\|\le\mathtt{C}\left(\sqrt{n}m^{-(\alpha+1/2)}+\mathtt{C}_{bias}m^{-(\alpha-1)}\right).$$

with some $\mathtt{C}>0$. The same arguments give for the case $\mathtt{C}_{bias}=0$

$$\begin{aligned}\|\mathcal{D}_m^{-1}A_{\boldsymbol{\upsilon}\boldsymbol{\kappa}}\mathcal{H}_m^{-1}\| &\le \frac{1}{c_{\mathcal{D}}^2}\left(\sup_{\|\boldsymbol{\theta}\|=1,\,\|\boldsymbol{\kappa}\|_{l^2}=1}\frac{1}{n}|\boldsymbol{\theta}A_{\boldsymbol{\theta}\boldsymbol{\kappa}}\boldsymbol{\kappa}|+\sup_{\|\boldsymbol{\eta}\|=1,\,\|\boldsymbol{\kappa}\|_{l^2}=1}\frac{1}{n}|\boldsymbol{\eta}A_{\boldsymbol{\eta}\boldsymbol{\kappa}}\boldsymbol{\kappa}|\right)\\ &\le \frac{\mathtt{C}_1}{c_{\mathcal{D}}^2}2m^{-1/2}.\end{aligned}$$

Remark A.4. In case $\mathsf{C}_{bias} > 0$ we do not manage to get a bound for $\boldsymbol{\theta} b_{\boldsymbol{\theta}\boldsymbol{\kappa}}\boldsymbol{\kappa}$ for general $\boldsymbol{\kappa} \in l^2$. How to get a bound for $\beta(m)$ in this setting remains unclear.

We bound using the dominated convergence theorem (applicable due to similar bounds as above)

$$\|\mathcal{H}_m \boldsymbol{\kappa}^*\|^2 \leq n \sum_{k=m+1}^{\infty} {\eta_k^*}^2 \|p_{\mathbf{X}^\top \boldsymbol{\theta}^*}\|_\infty + 2n \left| \sum_{l>k} \eta_l^* \eta_k^* \mathbb{E}[\boldsymbol{e}_k \boldsymbol{e}_l(\mathbf{X}^\top \boldsymbol{\theta}^*)] \right| .$$

As above we find

$$|\mathbb{E}[\boldsymbol{e}_k \boldsymbol{e}_l(\mathbf{X}^\top \boldsymbol{\theta}^*)]| \leq 17 s_{\mathbf{X}}^{p+1} L_{p_{\mathbf{X}}} \|\psi\|_\infty 2^{-3j_l/2-1} 2^{j_k/2} 1_{\{I_l \cap I_k \neq \emptyset\}}(k,l).$$

We estimate

$$\begin{aligned}
&\sum_{l>k>m} \eta_l^* \eta_k^* \mathbb{E}[\boldsymbol{e}_k \boldsymbol{e}_l(\mathbf{X}^\top \boldsymbol{\theta}^*)] \\
&\quad \leq 17 s_{\mathbf{X}}^{p+1} L_{p_{\mathbf{X}}} \|\psi\|_\infty \sum_{l>k} \eta_l^* \eta_k^* 2^{-3j_l/2-1} 2^{j_k/2} 1_{\{I_l \cap I_k \neq \emptyset\}}(k,l) \\
&\quad \leq 17 s_{\mathbf{X}}^{p+1} L_{p_{\mathbf{X}}} \|\psi\|_\infty \sum_{k=1}^{\infty} \eta_k^* 2^{j_k/2} \sum_{l=k+1}^{\infty} \eta_l^* 2^{-3j_l/2-1} 1_{\{I_l \cap I_k \neq \emptyset\}}(k,l) \\
&\quad \leq 17 s_{\mathbf{X}}^{p+1} L_{p_{\mathbf{X}}} \|\psi\|_\infty \sum_{k=1}^{\infty} \eta_k^* 2^{j_k/2} \left(\sum_{l=k+1}^{\infty} {\eta_l^*}^2 l^{2\alpha} \right)^{1/2} \\
&\qquad \left(\sum_{l=k+1}^{\infty} l^{-2\alpha} 2^{-3j_l} 1_{\{I_l \cap I_k \neq \emptyset\}}(k,l) \right)^{1/2} .
\end{aligned}$$

We continue using that $l \geq 2^{j_l}$

$$\begin{aligned}
&\sum_{l=k+1}^{\infty} l^{-2\alpha} 2^{-3j_l} 1_{\{I_l \cap I_k \neq \emptyset\}}(k,l) \\
&\quad \leq \sum_{j=j_{k+1}}^{\infty} 2^{-(3+2\alpha)j} \\
&\qquad\quad |\{l = 2^j + j17 - 1, \dots, 2^{j+1} + (j+1)17 - 1 - 1 :\, I_l \cap I_k \neq \emptyset\}| \\
&\quad \leq \sum_{j=j_{k+1}}^{\infty} 2^{-(3+2\alpha)j} \ulcorner 2^{j-j_k} 17 \urcorner \leq 2^{-(3+2\alpha)j_k} 36.
\end{aligned}$$

Plugging this in we find

$$
\begin{aligned}
&\sum_{l>k>m} \eta_l^* \eta_k^* \mathbb{E}[\boldsymbol{e}_k \boldsymbol{e}_l (\mathbf{X}^\top \boldsymbol{\theta}^*)] \\
&\quad \le 17\sqrt{36} s_{\mathbf{X}}^{p+1} L_{p\mathbf{X}} \|\psi\|_\infty \sum_{k=m+1}^{\infty} \eta_k^* 2^{-(2+2\alpha)j_k/2} \mathtt{C}_{\|\boldsymbol{f}^*\|} \\
&\quad \le 17\sqrt{36} s_{\mathbf{X}}^{p+1} L_{p\mathbf{X}} \|\psi\|_\infty \mathtt{C}_{\|\boldsymbol{f}^*\|} \left(\sum_{k=m+1}^{\infty} \eta_k^{*2} k^{2\alpha} \right)^{1/2} \left(\sum_{k=m+1}^{\infty} k^{-2\alpha} 2^{-(2+2\alpha)j_k} \right)^{1/2} \\
&\quad \le 17\sqrt{36} s_{\mathbf{X}}^{p+1} L_{p\mathbf{X}} \|\psi\|_\infty \mathtt{C}_{\|\boldsymbol{f}^*\|}^2 \left(\sum_{k=m+1}^{\infty} 2^{-(2+4\alpha)j_k} \right)^{1/2} \\
&\quad \le 17\sqrt{36} s_{\mathbf{X}}^{p+1} L_{p\mathbf{X}} \|\psi\|_\infty \mathtt{C}_{\|\boldsymbol{f}^*\|}^2 \left(\sum_{j=j_m}^{\infty} 2^{-(1+4\alpha)j_k} \right)^{1/2} \\
&\quad \le 17\sqrt{36} s_{\mathbf{X}}^{p+1} L_{p\mathbf{X}} \|\psi\|_\infty \mathtt{C}_{\|\boldsymbol{f}^*\|}^2 2^{-(1+4\alpha)(j_m)/2} \left(\sum_{j=0}^{\infty} 2^{-(1+4\alpha)j_k} \right)^{1/2}.
\end{aligned}
$$

From which we obtain

$$
\begin{aligned}
\|\mathcal{H}_m \boldsymbol{\eta}_2^*\|^2 &= n \sum_{k=m+1}^{\infty} \eta_k^{*2} \|p_{\mathbf{X}^\top \boldsymbol{\theta}^*}\|_\infty + 2 s_{\mathbf{X}}^{p} L_{p\mathbf{X}} \|\psi\|_\infty \mathtt{C}_{\|\boldsymbol{f}^*\|}^2 n 2^{-(1+4\alpha)(j_m+1)/2} \\
&\le \|p_{\mathbf{X}^\top \boldsymbol{\theta}^*}\|_\infty n m^{-(1+2\alpha)} m \left(\sum_{k=m+1}^{\infty} \eta_k^{*2} k^{2\alpha} \right) \\
&\quad + 17^2 \sqrt{36} s_{\mathbf{X}}^{p+1} L_{p\mathbf{X}} \|\psi\|_\infty \mathtt{C}_{\|\boldsymbol{f}^*\|}^2 n m^{-(1/2+2\alpha)} \\
&\le \Big(17 \|p_{\mathbf{X}^\top \boldsymbol{\theta}^*}\|_\infty \mathtt{C}_{\|\boldsymbol{f}^*\|} + 17^2 \sqrt{36} s_{\mathbf{X}}^{p+1} L_{p\mathbf{X}} \|\psi\|_\infty \mathtt{C}_{\|\boldsymbol{f}^*\|}^2 \Big) n m^{-(1+2\alpha)} m.
\end{aligned}
$$

Next we show

$$
\|\mathcal{D}_m^{-1} \left(\nabla_{\boldsymbol{\upsilon}\boldsymbol{\kappa}} \mathbb{E}[\mathcal{L}\big((\mathit{\Pi}_{p^*} \boldsymbol{\upsilon}^*, \lambda \boldsymbol{\kappa}^*)\big)] - A_{\boldsymbol{\upsilon}\boldsymbol{\kappa}} \right) \boldsymbol{\kappa}^* \| \le \tau(m).
$$

For this note that

$$
\begin{aligned}
&\left(\nabla_{\boldsymbol{\upsilon}\boldsymbol{\kappa}} \mathbb{E}[\mathcal{L}\big((\mathit{\Pi}_{p^*} \boldsymbol{\upsilon}^*, \lambda \boldsymbol{\kappa}^*)\big)] - A_{\boldsymbol{\upsilon}\boldsymbol{\kappa}} \right) \boldsymbol{\kappa}^* \\
&\quad = n \begin{pmatrix} \mathbb{E}[\boldsymbol{f}'_{(0,\lambda\boldsymbol{\kappa}^*)} \mathbf{X} \boldsymbol{f}_{(0,\boldsymbol{\kappa}^*)} (\mathbf{X}^\top \boldsymbol{\theta}^*)] \\ \mathbb{E}[\boldsymbol{e} \boldsymbol{f}_{(0,\boldsymbol{\kappa}^*)} (\mathbf{X}^\top \boldsymbol{\theta}^*)] \end{pmatrix} \\
&\qquad + n \begin{pmatrix} \mathbb{E}[\boldsymbol{f}'_{(0,\boldsymbol{\kappa}^*)} \mathbf{X} \boldsymbol{f}_{(0,\lambda\boldsymbol{\kappa}^*)} (\mathbf{X}^\top \boldsymbol{\theta}^*)] \\ 0 \end{pmatrix}.
\end{aligned}
$$

We infer

$$
\begin{aligned}
&\|\mathcal{D}_m^{-1}\left(\nabla_{\boldsymbol{\upsilon}\boldsymbol{\kappa}}\mathbb{E}[\mathcal{L}\big((\mathit{\Pi}_{p^*}\boldsymbol{\upsilon}^*, \lambda\boldsymbol{\kappa}^*)\big)] - A_{\boldsymbol{\upsilon}\boldsymbol{\kappa}}\right)\boldsymbol{\kappa}^*\| \\
&\quad\le n\mathbb{E}[\boldsymbol{f}^2_{(0,\boldsymbol{\kappa}^*)}(\mathbf{X}^\top\boldsymbol{\theta}^*)]^{1/2}\Bigg(\mathbb{E}\left[\left\|\mathcal{D}_m^{-1}\begin{pmatrix} \boldsymbol{f}'_{(0,\lambda\boldsymbol{\kappa}^*)}\mathbf{X} \\ \boldsymbol{e} \end{pmatrix}\right\|^2\right]^{1/2} \\
&\qquad+\mathbb{E}\left[\left\|\mathcal{D}_m^{-1}\begin{pmatrix} \boldsymbol{f}'_{(0,\boldsymbol{\kappa}^*)}\mathbf{X} \\ 0 \end{pmatrix}\right\|^2\right]^{1/2}\Bigg) \\
&\quad\le \frac{\sqrt{n}}{c_{\mathcal{D}}}\left(s_{\mathbf{X}}\left\{\mathbb{E}[{\boldsymbol{f}'_{(0,\lambda\boldsymbol{\kappa}^*)}}^2]^{1/2} + \mathbb{E}[{\boldsymbol{f}'_{(0,\boldsymbol{\kappa}^*)}}^2]^{1/2}\right\} + \|p_{\mathbf{X}^\top\boldsymbol{\theta}}\|^{1/2}17^{1/4}\sqrt{m}\right) \\
&\qquad\mathbb{E}[\boldsymbol{f}^2_{(0,\boldsymbol{\kappa}^*)}(\mathbf{X}^\top\boldsymbol{\theta}^*)]^{1/2}.
\end{aligned}
$$

We estimate separately using the same bounds as before to apply the dominated convergence theorem to exchange summation and expectation. We bound as above using (A.4)

$$
\begin{aligned}
&\mathbb{E}[{\boldsymbol{f}'_{(0,\lambda\boldsymbol{\kappa}^*)}}^2] \\
&\quad= \lambda\sum_{k,l=m+1}^{\infty}\eta^*_k\eta^*_l\mathbb{E}[\boldsymbol{e}'_l\boldsymbol{e}'_k(\mathbf{X}^\top\boldsymbol{\theta}^*)] \\
&\quad\le 17 s_{\mathbf{X}}\|\psi'\|_\infty\|p_{\mathbf{X}}\|_\infty\sum_{k,l=m+1}^{\infty}\eta^*_k\eta^*_l 2^{3(j_l+j_k)/2-(j_l\vee j_k)}1_{\{I_k\cap I_l\neq\emptyset\}}(k,l) \\
&\quad\le 17 s_{\mathbf{X}}\|\psi'\|_\infty\|p_{\mathbf{X}}\|_\infty\sum_{k=m+1}^{\infty}\eta^*_k 2^{3j_k/2}\left(\sum_{l=m+1}^{\infty}l^{2\alpha}{f^*_l}^2\right)^{1/2} \\
&\qquad\left(\sum_{l=m+1}^{\infty}2^{(3-2\alpha)2j_l-2(j_l\vee j_k)}1_{\{I_k\cap I_l\neq\emptyset\}}(k,l)\right)^{1/2}.
\end{aligned}
$$

Observe

$$
\begin{aligned}
&\sum_{l=m+1}^{\infty}2^{(3-2\alpha)j_l-2(j_l\vee j_k)}1_{\{I_k\cap I_l\neq\emptyset\}}(k,l) \\
&\quad=\sum_{j=j_m+1}^{\infty}2^{(3-2\alpha)j-2(j\vee j_k)} \\
&\qquad\left|\left\{l = j12 + 2^j + r_l \,\middle|\, r_l\in\{0,\dots,2^j+11\},\ I_l\cap I_k\neq\emptyset\right\}\right| \\
&\quad=\sum_{j=j_m+1}^{\infty}2^{(3-2\alpha)j-2(j\vee j_k)}\ulcorner 2^{(j-j_k)}17\urcorner
\end{aligned}
$$

$$\leq 18 \sum_{j=j_m+1}^{\infty} 2^{(2-2\alpha)j} = 17^3 18 m^{-2\alpha+2}.$$

Such that again using the Cauchy-Schwarz inequality for any $\lambda \in [0,1]$

$$\begin{aligned} \mathbb{E}[\boldsymbol{f}'_{(0,\lambda\boldsymbol{\kappa}^*)}{}^2] &\leq 17^{5/2}\sqrt{18} s_{\mathbf{X}} \|\psi'\|_\infty \|p_{\mathbf{X}}\|_\infty \mathtt{C}_{\|\boldsymbol{f}^*\|} m^{-\alpha+1} \sum_{k=m+1}^{\infty} \eta_k^* 2^{3j_k/2} \\ &\leq 17^3 \sqrt{18} s_{\mathbf{X}} \|\psi'\|_\infty \|p_{\mathbf{X}}\|_\infty \mathtt{C}^2_{\|\boldsymbol{f}^*\|} m^{-2\alpha+3}. \end{aligned}$$

Furthermore

$$\begin{aligned} &\mathbb{E}[\boldsymbol{f}^2_{(0,\boldsymbol{\kappa}^*)}(\mathbf{X}^\top \boldsymbol{\theta}^*)] = \sum_{k,l=m+1}^{\infty} \eta_k^* \eta_l^* \mathbb{E}[\boldsymbol{e}_k \boldsymbol{e}_l(\mathbf{X}^\top \boldsymbol{\theta}^*)] \\ &\quad\leq 17 s_{\mathbf{X}}^{p+1} L_{p_{\mathbf{X}}} \|\psi\|_\infty \sum_{k,l=m+1}^{\infty} \eta_k^* \eta_l^* 2^{-3(j_l \vee j_k)/2 + (j_l \wedge j_k)/2} 1_{\{I_l \cap I_k \neq \emptyset\}}(k,l) \\ &\quad= 17 s_{\mathbf{X}}^{p+1} L_{p_{\mathbf{X}}} \|\psi\|_\infty \sum_{k=m+1}^{\infty} \eta_k^* 2^{-j_k} \sum_{l=m+1}^{\infty} \eta_l^* 1_{\{I_l \cap I_k \neq \emptyset\}}(k,l) \\ &\quad\leq 17 s_{\mathbf{X}}^{p+1} L_{p_{\mathbf{X}}} \|\psi\|_\infty \sum_{k=m+1}^{\infty} \eta_k^* 2^{-j_k} \mathtt{C}_{\|\boldsymbol{f}^*\|} \left(\sum_{j=j_m+1}^{\infty} 2^{-2\alpha j} 18 \right)^{1/2} \\ &\quad\leq 17\sqrt{18} 17^{1/2} s_{\mathbf{X}}^{p+1} L_{p_{\mathbf{X}}} \|\psi\|_\infty \mathtt{C}_{\|\boldsymbol{f}^*\|} m^{-\alpha} \sum_{k=m+1}^{\infty} \eta_k^* 2^{-j_k} \\ &\quad\leq 17\sqrt{36} 17^2 s_{\mathbf{X}}^{p+1} L_{p_{\mathbf{X}}} \|\psi\|_\infty \mathtt{C}^2_{\|\boldsymbol{f}^*\|} m^{-2\alpha}. \end{aligned}$$

Together this implies

$$\begin{aligned} &\|\mathcal{D}_m^{-1} \left(\nabla_{\boldsymbol{\upsilon}\boldsymbol{\kappa}} \mathbb{E}[\mathcal{L}\left((\mathit{\Pi}_{p^*} \boldsymbol{\upsilon}^*, \lambda\boldsymbol{\kappa}^*) \right)] - A_{\boldsymbol{\upsilon}\boldsymbol{\kappa}} \right) \boldsymbol{\kappa}^* \| \\ &\quad \leq \frac{1}{c_{\mathcal{D}}} \left(2 s_{\mathbf{X}} \left\{ 17^3 \sqrt{18} s_{\mathbf{X}} \|\psi'\|_\infty \|p_{\mathbf{X}}\|_\infty \mathtt{C}^2_{\|\boldsymbol{f}^*\|} \right\}^{1/2} + \|p_{\mathbf{X}^\top \boldsymbol{\theta}}\|^{1/2} 17^{1/4} \right) \\ &\quad\quad \sqrt{\sqrt{36} 17^3 s_{\mathbf{X}}^{p+1} L_{p_{\mathbf{X}}} \|\psi\|_\infty \mathtt{C}^2_{\|\boldsymbol{f}^*\|}} m^{-2\alpha+1/2)} \sqrt{n} \\ &\quad \leq \mathtt{C}_1 m^{-2\alpha+1/2} \sqrt{n}. \end{aligned}$$

Clearly

$$\left| {\boldsymbol{\kappa}^*}^\top (\mathcal{H}_m - \nabla_{\boldsymbol{\kappa}\boldsymbol{\kappa}} \mathbb{E}\mathcal{L}(\mathit{\Pi}_{p^*} \boldsymbol{\upsilon}^*, \lambda\boldsymbol{\kappa}^*)) \boldsymbol{\kappa}^* \right| = 0.$$

To see this simply note that for any $\boldsymbol{f} \in \mathcal{S}$ and any $\boldsymbol{\kappa} \in \mathcal{S}$

$$\boldsymbol{\kappa}^\top \nabla_{\boldsymbol{\kappa}\boldsymbol{\kappa}} \mathbb{E}\mathcal{L}(\boldsymbol{\theta}^*, \boldsymbol{f}) \boldsymbol{\kappa} = \mathbb{E}[\boldsymbol{f}^2_{(0,\boldsymbol{\kappa})}(\mathbf{X}^\top \boldsymbol{\theta}^*)] = \boldsymbol{\kappa}^\top \mathcal{H}_m \boldsymbol{\kappa}.$$

Furthermore we find that

$$\begin{aligned}\boldsymbol{\theta}^\top d^2_{\boldsymbol{\theta}}(\boldsymbol{\upsilon}^*)\boldsymbol{\theta} &= \mathbb{E}[f_f'(\mathbf{X}^\top\boldsymbol{\theta}^*)^2(\mathbf{X}^\top\nabla\varPhi(\boldsymbol{\theta}^*)\boldsymbol{\theta})^2]\\ &\le \|\boldsymbol{f}'_{\boldsymbol{\eta}^*}\|^2_\infty s^2_{\mathbf{X}}\|\nabla\varPhi(\boldsymbol{\theta}^*)\|\\ &\le \frac{p+2}{4}\mathsf{C}_{\|f\|}\|\psi'\|^2_\infty s^2_{\mathbf{X}}\pi^2,\end{aligned}$$

and

$$\begin{aligned}\boldsymbol{\theta}^\top v^2_{\boldsymbol{\theta}}(\boldsymbol{\upsilon}^*)\boldsymbol{\theta} &= \mathbb{E}\left[\left(\boldsymbol{f}_{\boldsymbol{\eta}^*}(\mathbf{X}^\top\boldsymbol{\theta}^*) - g(\mathbf{X})\right)\left((\mathbf{X}^\top\boldsymbol{\theta})^2\boldsymbol{f}''_{\boldsymbol{\eta}^*}(\mathbf{X}^\top\boldsymbol{\theta}^*)\right]\right.\\ &\qquad \left.+|\boldsymbol{f}'_{\boldsymbol{\eta}^*}(\mathbf{X}^\top\boldsymbol{\theta}^*)|^2\mathbf{X}^\top\nabla^2\varphi^\top_{\boldsymbol{\theta}^*}[\mathbf{X},\boldsymbol{\theta},\boldsymbol{\theta}]\right)\Big]\\ &\le \mathsf{C}_{bias}\left(s^2_{\mathbf{X}}\mathsf{C}_{\|\boldsymbol{f}''_{\boldsymbol{\eta}^*}\|_\infty} + 34\|\psi'\|^2_\infty C^2_{\|\boldsymbol{\eta}^*\|}s^2_{\mathbf{X}}\|\nabla^2\varphi_{\boldsymbol{\theta}^*}\|_\infty\right).\end{aligned}$$

This completes the proof. □

A.6. Proof or Lemma 5.1

We proof the claim via validating condition $(\mathcal{L}\mathtt{r}_\infty)$ from Section 4.3. For this condition we can use the full expectation $\mathbb{E}$ instead of $\mathbb{E}_\varepsilon$:

Lemma A.7. *Assume* $(\mathcal{A})$. *Then there exists a constant* $\mathtt{b} > 0$ *such that*

$$\mathbb{E}\mathcal{L}(\boldsymbol{\upsilon},\boldsymbol{\upsilon}^*) \le -\mathtt{b}\mathtt{r}^2.$$

Proof. As in Lemma A.8 we can make the decomposition

$$\begin{aligned}&\mathbb{E}\mathcal{L}(\boldsymbol{\upsilon},\boldsymbol{\upsilon}^*)\\ &\quad= -n\mathbb{E}\left[\left(g(\mathbf{X}) - \mathbb{E}[g(\mathbf{X})|\mathbf{X}^\top\boldsymbol{\theta}]\right)^2\right] - n\mathbb{E}\left[\left(g(\mathbf{X}) - \mathbb{E}[g(\mathbf{X})|\mathbf{X}^\top\boldsymbol{\theta}^*]\right)^2\right]\\ &\qquad -n\mathbb{E}\left[\left(\mathbb{E}[g(\mathbf{X})|\mathbf{X}^\top\boldsymbol{\theta}] - \sum_{k=1}^n \eta_k\boldsymbol{e}_k(\mathbf{X}^\top\boldsymbol{\theta})\right)^2\right].\end{aligned}$$

We find with condition (model bias) for all $\boldsymbol{\upsilon} = (\boldsymbol{\theta},\boldsymbol{\eta}) \in \varUpsilon_m$

$$\begin{aligned}&-n\mathbb{E}\left[\left(g(\mathbf{X}) - \mathbb{E}[g(\mathbf{X})|\mathbf{X}^\top\boldsymbol{\theta}]\right)^2\right] + n\mathbb{E}\left[\left(g(\mathbf{X}) - \mathbb{E}[g(\mathbf{X})|\mathbf{X}^\top\boldsymbol{\theta}^*]\right)^2\right]\\ &\qquad\le \begin{cases}-n\mathtt{b}_\theta, & \|\mathrm{D}(\boldsymbol{\theta}-\boldsymbol{\theta}^*)\| \ge \sqrt{n}\mathtt{r}_{\boldsymbol{\theta}}/c_{\mathfrak{D}}\\ -\mathtt{b}_\theta\|\mathrm{D}(\boldsymbol{\theta}-\boldsymbol{\theta}^*)\|^2, & \textit{otherwise.}\end{cases}\end{aligned}$$

As $\|\mathrm{D}(\boldsymbol{\theta}-\boldsymbol{\theta}^*)\|^2 \le n\frac{p+2}{2}\mathsf{C}_{\|f\|}\|\psi'\|^2_\infty s^2_{\mathbf{X}}\pi^2$ we find

$$\mathbb{E}\mathcal{L}(\boldsymbol{\upsilon},\boldsymbol{\upsilon}^*) \le -\mathtt{b}''_\theta\|\mathrm{D}(\boldsymbol{\theta}-\boldsymbol{\theta}^*)\|^2, \quad \mathtt{b}''_\theta = \mathtt{b}_\theta\min\left\{1, \frac{1}{\frac{p+2}{2}\mathsf{C}_{\|f\|}\|\psi'\|^2_\infty s^2_{\mathbf{X}}\pi^2}\right\}.$$

We study two cases first assume that $\|\mathrm{D}(\boldsymbol{\theta}-\boldsymbol{\theta}^*)\|^2 \geq \tau^2\mathtt{r}^2$ for some $\tau > 0$, then we get

$$-\mathbb{E}\mathcal{L}(\boldsymbol{\upsilon},\boldsymbol{\upsilon}^*) \geq \tau^2 \mathtt{b}_\theta' \mathtt{r}^2.$$

Otherwise – if $\|\mathrm{D}(\boldsymbol{\theta}-\boldsymbol{\theta}^*)\|^2 \leq \tau^2\mathtt{r}^2$ – we have as in the proof of Lemma A.5

$$\mathbb{E}\left[\left(\mathbb{E}[g(\mathbf{X})|\mathbf{X}^\top\boldsymbol{\theta}] - \sum_{k=1}^n \eta_k \boldsymbol{e}_k(\mathbf{X}^\top\boldsymbol{\theta})\right)^2\right] \geq Q(\mathtt{b}^* - \mathtt{C}\tau)^2\mathtt{r}^2.$$

Choosing $\tau > 0$ small enough gives the claim. □

The claim of Lemma 5.1 now is a direct consequence of Lemma A.2 of [1].

A.7. Proof of Lemma 5.2

Remark A.5. We assume that the density of the regressors satisfies $p_{\mathbf{X}} \geq c_{p_{\mathbf{X}}} > 0$ on $B_{s_{\mathbf{X}}+c_B}(0)$. This implies that for any $\boldsymbol{\theta} \in \mathbb{R}^p$ the density of $\mathbf{X}^\top\boldsymbol{\theta}$ is also bounded away from zero on $[-s_{\mathbf{X}}, s_{\mathbf{X}}]$ by $\lambda(B_{c_B}^{p-1})c_{p_{\mathbf{X}}}$ where $\lambda(B_{\mathtt{r}}^{p-1})$ denotes the Lebesgue measure of the $p-1$ dimensional ball of radius $\mathtt{r} > 0$ on $\mathbb{R}^{p-1}$. As we use a orthonormal wavelet basis on $L^2([-s_{\boldsymbol{x}}, s_{\boldsymbol{x}}])$ this gives

$$\begin{aligned}\lambda_{min}(\mathcal{H}^2(\boldsymbol{\upsilon})) &= \inf_{\boldsymbol{\eta}\in l^2} \mathbb{E}[\boldsymbol{f}_{\boldsymbol{\eta}}(\mathbf{X}^\top\boldsymbol{\theta})^2]/\|\boldsymbol{\eta}\|^2 \\ &\geq \lambda(B_{c_B}^{p-1})c_{p_{\mathbf{X}}} \int_{[-s_{\mathbf{X}},s_{\mathbf{X}}]} \boldsymbol{f}_{\boldsymbol{\eta}}(x)^2 dx/\|\boldsymbol{\eta}\|^2 = \lambda(B_{c_B}^{p-1})c_{p_{\mathbf{X}}}.\end{aligned}$$

Proof. Take any $\boldsymbol{\theta} \in S_1^{p,+}$. Then we have due to the quadratic structure of the problem and using the usual bounds for $\|\boldsymbol{e}\| \leq \mathtt{C}\sqrt{m}$

$$\begin{aligned}\left\|\widetilde{\boldsymbol{\eta}}_{m,\boldsymbol{\theta}}^{(\infty)}\right\| &\overset{\text{def}}{=} \left\|\operatorname*{argmax}_{\boldsymbol{\eta}\in\mathbb{R}^m} \mathcal{L}_m(\boldsymbol{\theta},\boldsymbol{\eta})\right\| \\ &= \left\|\left(\frac{1}{n}\sum_{i=1}^n \boldsymbol{e}\boldsymbol{e}^\top(\mathbf{X}_i^\top\boldsymbol{\theta})\right)^{-1} \frac{1}{n}\sum_{i=1}^n (g(\mathbf{X}_i)+\varepsilon_i)\boldsymbol{e}(\mathbf{X}_i^\top\boldsymbol{\theta})\right\| \\ &\leq \left(\|g\|_\infty \mathtt{C}\sqrt{m} + \left\|\frac{1}{n}\sum_{i=1}^n \varepsilon_i \boldsymbol{e}(\mathbf{X}_i^\top\boldsymbol{\theta})\right\|\right) \\ &\qquad \left\|\left(\frac{1}{n}\sum_{i=1}^n \boldsymbol{e}\boldsymbol{e}^\top(\mathbf{X}_i^\top\boldsymbol{\theta})\right)^{-1}\right\|. \end{aligned} \tag{A.11}$$

We want to bound the above right-hand side. For this we bound

$$\begin{aligned}\mathbb{P}\left(\left\|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i \boldsymbol{e}(\mathbf{X}_i^\top\boldsymbol{\theta})\right\|\geq t\right) &= \mathbb{P}\left(\sup_{\substack{\boldsymbol{\eta}\in\mathbb{R}^m\\ \|\boldsymbol{\eta}\|=1}}\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\sum_{k=1}^{m}\eta_k\boldsymbol{e}_k(\mathbf{X}_i^\top\boldsymbol{\theta})\geq t\right)\\ &\leq \mathbb{P}\left(\sup_{\boldsymbol{\eta}\in B_1(0)}\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\boldsymbol{f}_{\boldsymbol{\eta}}(\mathbf{X}_i^\top\boldsymbol{\theta})\geq t\right)\end{aligned}$$

We want to apply Corollary 2.2 of the supplement of [19] with

$$\mathcal{U}(\boldsymbol{\eta}) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\varepsilon_i\boldsymbol{f}_{\boldsymbol{\eta}}(\mathbf{X}_i^\top\boldsymbol{\theta}),\ \boldsymbol{\upsilon}^* = 0\in\mathbb{R}^m.$$

For this we have to show that

$$\log\mathbb{E}\exp\left\{\lambda\frac{\mathcal{U}(\boldsymbol{\upsilon})-\mathcal{U}(\boldsymbol{\upsilon}^\circ)}{d(\boldsymbol{\upsilon},\boldsymbol{\upsilon}^\circ)}\right\}\leq \nu^2\lambda^2/2,$$

with $d(\boldsymbol{\eta},\boldsymbol{\eta}^\circ) = \|\boldsymbol{\eta}-\boldsymbol{\eta}^\circ\|_{\mathbb{R}^m}$. This is indeed the case since by Lemma A.4 for any pair $\boldsymbol{\eta},\boldsymbol{\eta}^\circ\in B_1(0)$

$$\left|\boldsymbol{f}_{\boldsymbol{\eta}-\boldsymbol{\eta}^\circ}(\mathbf{X}_i^\top\boldsymbol{\theta})\right|\leq \mathtt{C}\|\psi\|_\infty\sqrt{m}\|\boldsymbol{\eta}-\boldsymbol{\eta}^\circ\|.$$

Using ($\mathbf{Cond}_\varepsilon$), the independence of (ε_i) and $(\mathbf{X}_i)$ we find for

$$\lambda\leq\frac{\sqrt{n}}{\mathtt{C}\sqrt{m}}\widetilde{\mathtt{g}},$$

and any pair $\boldsymbol{\eta},\boldsymbol{\eta}^\circ\in B_1(0)$

$$\begin{aligned}&\log\mathbb{E}\exp\left\{\lambda\frac{\mathcal{U}(\boldsymbol{\upsilon})-\mathcal{U}(\boldsymbol{\upsilon}^\circ)}{d(\boldsymbol{\upsilon},\boldsymbol{\upsilon}^\circ)}\right\}\\ &\quad= \log\mathbb{E}\exp\left\{\lambda\frac{1}{\sqrt{n}\|\boldsymbol{\eta}-\boldsymbol{\eta}^\circ\|}\sum_{i=1}^{n}\varepsilon_i\boldsymbol{f}_{\boldsymbol{\eta}-\boldsymbol{\eta}^\circ}(\mathbf{X}_i^\top\boldsymbol{\theta})\right\}\\ &\quad\leq\sum_{i=1}^{n}\log\mathbb{E}\exp\left\{\frac{\lambda}{\sqrt{n}\|\boldsymbol{\eta}-\boldsymbol{\eta}^\circ\|}\varepsilon_i\boldsymbol{f}_{\boldsymbol{\eta}-\boldsymbol{\eta}^\circ}(\mathbf{X}_i^\top\boldsymbol{\theta})\right\}\\ &\quad\leq\sum_{i=1}^{n}\log\mathbb{E}\left[\exp\left\{\frac{\widetilde{\nu}^2\lambda^2}{n}\frac{1}{\|\boldsymbol{\eta}-\boldsymbol{\eta}^\circ\|^2}\boldsymbol{f}^2_{\boldsymbol{\eta}-\boldsymbol{\eta}^\circ}(\mathbf{X}_i^\top\boldsymbol{\theta})\right\}\right]\\ &\quad\leq \mathtt{C}^2 m\widetilde{\nu}^2\lambda^2/2.\end{aligned}$$

This implies with Corollary 2.2 of the supplement of [19]

$$\mathbb{P}\left(\left\|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i \boldsymbol{e}(\mathbf{X}_i^\top\boldsymbol{\theta})\right\| \ge \mathtt{C}\widetilde{\nu}\sqrt{m}\sqrt{\mathtt{x}+2m}/\sqrt{n}\right) \le \mathrm{e}^{-\mathtt{x}}.$$

Two bound the norm of the inverse of the matrix in (A.11) we denote

$$\boldsymbol{M}_n(\boldsymbol{\theta}) \overset{\text{def}}{=} \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{e}\boldsymbol{e}^\top(\mathbf{X}_i^\top\boldsymbol{\theta}).$$

Note that with Remark A.5

$$\mathbb{E}\left[\boldsymbol{M}_n(\boldsymbol{\theta})\right] \ge \lambda(B_h^{p-1})c_{p\mathbf{X}}c_{\boldsymbol{e}},$$

while

$$\sup_{\boldsymbol{\theta}\in S_1^p}\left\|\boldsymbol{M}_n(\boldsymbol{\theta})-\mathbb{E}\left[\boldsymbol{M}_n(\boldsymbol{\theta})\right]\right\| = \sup_{(\boldsymbol{\theta},\boldsymbol{\eta})\in S_1^p\times S_1^m}\left|(P_n-\mathbb{P})\boldsymbol{f}_{\boldsymbol{\eta}}^2(\mathbf{X}^\top\boldsymbol{\theta})\right|.$$

We bound

$$\begin{aligned}
&\mathbb{P}\left(\sup_{(\boldsymbol{\theta},\boldsymbol{\eta})\in S_1^p\times S_1^m}\left|(P_n-\mathbb{P})\boldsymbol{f}_{\boldsymbol{\eta}}^2(\mathbf{X}^\top\boldsymbol{\theta})\right| \ge t+s\right)\\
&\quad\le \mathbb{P}\left(\left|(P_n-\mathbb{P})\boldsymbol{f}_{\boldsymbol{\eta}^*}^2(\mathbf{X}^\top\boldsymbol{\theta}^*)\right| \ge s\right)\\
&\qquad+\mathbb{P}\left(\sup_{(\boldsymbol{\theta},\boldsymbol{\eta})\in S_1^p\times S_1^m}\left|(P_n-\mathbb{P})\left[\boldsymbol{f}_{\boldsymbol{\eta}}^2(\mathbf{X}^\top\boldsymbol{\theta})-\boldsymbol{f}_{\boldsymbol{\eta}^*}^2(\mathbf{X}^\top\boldsymbol{\theta}^*)\right]\right| \ge t\right).
\end{aligned}$$

For the first term we can use the bounded differences inequality (Theorem A.1) to find

$$\mathbb{P}\left(\left|(P_n-\mathbb{P})\boldsymbol{f}_{\boldsymbol{\eta}^*}^2(\mathbf{X}_i^\top\boldsymbol{\theta}^*)^2\right| \ge \|f_{\boldsymbol{\eta}^*}\|_\infty^2\sqrt{\mathtt{x}}/\sqrt{n}\right) \le \mathrm{e}^{-\mathtt{x}}.$$

For the second summand we define $\boldsymbol{\zeta}_{\mathbf{X}}(\boldsymbol{\upsilon}) \overset{\text{def}}{=} (P_n-\mathbb{P})\boldsymbol{f}_{\boldsymbol{\eta}}(\mathbf{X}_i^\top\boldsymbol{\theta})^2$. We use the chaining method, i.e. Lemma A.2. Define $\Upsilon_0=\{\boldsymbol{\upsilon}^*\}$ and with a sequence $\mathtt{r}_k = 2^{-k}\mathtt{r}$ with $\mathtt{r}$ to be specified later the sequence of sets Υ_k each with minimal cardinality such that

$$S_1^p\times S_1^m \subset \bigcup_{\boldsymbol{\upsilon}\in\Upsilon_k}B_{\mathtt{r}_k}(\boldsymbol{\upsilon}),\qquad B_{\mathtt{r}}(\boldsymbol{\upsilon}) \overset{\text{def}}{=} \{\boldsymbol{\upsilon}^\circ\in S_1^p\times S_1^m,\ \|\boldsymbol{\upsilon}^\circ-\boldsymbol{\upsilon}\|\le\mathtt{r}\}.$$

We can estimate with any $\boldsymbol{\upsilon}'\in B_{\mathtt{r}_k,\mathcal{D}}(\boldsymbol{\upsilon})$

$$\inf_{\Upsilon_{k-1,m}}\left|\boldsymbol{\zeta}_{\mathbf{X}}(\boldsymbol{\upsilon})-\boldsymbol{\zeta}_{\mathbf{X}}(\boldsymbol{\upsilon}^\circ)\right| = \left|(P_n-\mathbb{P})\left\{\boldsymbol{f}_{\boldsymbol{\eta}}(\mathbf{X}_i^\top\boldsymbol{\theta})^2-\boldsymbol{f}_{\boldsymbol{\eta}'}(\mathbf{X}_i^\top\boldsymbol{\theta}')^2\right\}\right|$$

We estimate for an application of the bounded differences inequality

$$\begin{aligned}&\left|\left\{\boldsymbol{f}_{\boldsymbol{\eta}}(\mathbf{X}_i^\top\boldsymbol{\theta})^2-\boldsymbol{f}_{\boldsymbol{\eta}'}(\mathbf{X}_i^\top\boldsymbol{\theta}')^2\right\}\right|\\ &\quad\le\left|\left\{\boldsymbol{f}_{\boldsymbol{\eta}}(\mathbf{X}_i^\top\boldsymbol{\theta})-\boldsymbol{f}_{\boldsymbol{\eta}'}(\mathbf{X}_i^\top\boldsymbol{\theta}')\right\}\left\{\boldsymbol{f}_{\boldsymbol{\eta}}(\mathbf{X}_i^\top\boldsymbol{\theta})+\boldsymbol{f}_{\boldsymbol{\eta}'}(\mathbf{X}_i^\top\boldsymbol{\theta}')\right\}\right|\\ &\quad\le\left(\|\boldsymbol{f}_{\boldsymbol{\eta}}\|_\infty+\|\boldsymbol{f}_{\boldsymbol{\eta}'}\|_\infty\right)\left(\|\boldsymbol{f}_{\boldsymbol{\eta}-\boldsymbol{\eta}'}\|_\infty+\|\boldsymbol{f}'_{\boldsymbol{\eta}}\|_\infty\|\boldsymbol{\theta}-\boldsymbol{\theta}'\|\right).\end{aligned}$$

We have as $\|\eta\|=1$ with Lemma A.4

$$\begin{aligned}\|\boldsymbol{f}_{\boldsymbol{\eta}}\|_\infty&\le\|\boldsymbol{\eta}\|\sup_{x\in[-s_{\mathbf{X}},s_{\mathbf{X}}]}\left(\sum_{k=1}^m\boldsymbol{e}_k^2(x)^2\right)^{1/2}\le\sqrt{17}\|\psi\|\sqrt{m},\\ \|\boldsymbol{f}'_{\boldsymbol{\eta}}\|_\infty&\le\|\boldsymbol{\eta}\|\sup_{x\in[-s_{\mathbf{X}},s_{\mathbf{X}}]}\left(\sum_{k=1}^m\boldsymbol{e}_k'^2(x)^2\right)^{1/2}\le\sqrt{17}\|\psi'\|m^{3/2}.\end{aligned}$$

Consequently

$$\left|\left\{\boldsymbol{f}_{\boldsymbol{\eta}}(\mathbf{X}_i^\top\boldsymbol{\theta})^2-\boldsymbol{f}_{\boldsymbol{\eta}'}(\mathbf{X}_i^\top\boldsymbol{\theta}')^2\right\}\right|\le\mathtt{C}_{\boldsymbol{\zeta}}m^{3/2}\mathtt{r}_k.$$

This yields with the bounded difference inequality

$$\mathbb{P}\left(\inf_{\varUpsilon_{k-1,m}}|\boldsymbol{\zeta}_{\mathbf{X}}(\boldsymbol{\upsilon}_k)-\boldsymbol{\zeta}_{\mathbf{X}}(\boldsymbol{\upsilon}_{k-1})|\ge s\mathtt{C}_{\boldsymbol{\zeta}}m^{3/2}\mathtt{r}_k/\sqrt{n}\right)\le\mathrm{e}^{-s^2}.$$

Now we can define $\mathtt{r}\overset{\text{def}}{=}\frac{(1-1/\sqrt{2})}{\mathtt{C}_{\boldsymbol{\zeta}}m^{3/2}}$. Then

$$\mathbb{P}\left(\inf_{\varUpsilon_{k-1,m}}|\boldsymbol{\zeta}_{\mathbf{X}}(\boldsymbol{\upsilon}_k)-\boldsymbol{\zeta}_{\mathbf{X}}(\boldsymbol{\upsilon}_{k-1})|\ge\frac{2^{-(k-1)}(1-1/\sqrt{2})s}{\sqrt{n}}\right)\le\mathrm{e}^{-s^2}.\quad\text{(A.12)}$$

Set

$$\begin{aligned}s&=\sqrt{\mathtt{x}+\log(2)+p^*[1+\log(2)+\log(\mathtt{C}_{\boldsymbol{\zeta}}m^{3/2})-\log(1-1/\sqrt{2})]}/\sqrt{n}\\ &\le\mathtt{C}\sqrt{\mathtt{x}+p^*\log(p^*)}/\sqrt{n},\end{aligned}$$

and plug it into (A.12), then we find with Lemma A.2

$$\begin{aligned}&\mathbb{P}\left(\sup_{\boldsymbol{\theta}\in S_1^p}\|\boldsymbol{M}_n(\boldsymbol{\theta})-\mathbb{E}\left[\boldsymbol{M}_n(\boldsymbol{\theta})\right]\|\ge\mathtt{C}\sqrt{\mathtt{x}+p^*\log(p^*)}/\sqrt{n}\right)\\ &\quad\le\mathbb{P}\left(\sup_{\boldsymbol{\upsilon}\in\varUpsilon_m}\boldsymbol{\zeta}_{\mathbf{X}}(\boldsymbol{\upsilon})-\boldsymbol{\zeta}_{\mathbf{X}}(\boldsymbol{\upsilon}^*)\ge\mathtt{C}\sqrt{\mathtt{x}+p^*\log(p^*)}/\sqrt{n}\right)\\ &\quad\le\sum_{k=1}^\infty\exp\Big\{p^*[1+\log(2)k+\log(\mathtt{C}_{\boldsymbol{\zeta}}m^{3/2})-\log(1-1/\sqrt{2})]\end{aligned}$$

$$- 2^{k-1}\left[\mathtt{x} + \log(2) + p^*[1 + \log(2) + \log(\mathtt{C}_{\boldsymbol{\zeta}} m^{3/2}) - \log(1 - 1/\sqrt{2})]\right]\Bigg\}$$
$$\le \mathrm{e}^{-\mathtt{x}}.$$

Together this implies because $p^* \log(p^*)/\sqrt{n} \to 0$

$$\mathbb{P}\left(\sup_{\boldsymbol{\theta}\in S_1^{p,+}} \left\|\widetilde{\boldsymbol{\eta}}^{(\infty)}_{m,\boldsymbol{\theta}}\right\| \ge \mathtt{C}\sqrt{p^*\log(p^*) + \mathtt{x}}\right) \le 3\mathrm{e}^{-\mathtt{x}}.$$

Adding $\log(3)$ to $\mathtt{x}$ in the above inequality and adapting the constant gives the claim with a probability bound $\mathrm{e}^{-\mathtt{x}}$. □

A.8. Proof of Lemma 5.3

Before we prove the claims we need a series of auxiliary lemmas.

A.8.1. $\mathcal{D}_m(\boldsymbol{\upsilon}^*_m)$ is boundedly invertible

Lemma A.8. *Under $(\mathcal{A})$ we have that*

$$\mathcal{D}_m(\boldsymbol{\upsilon}^*_m)^2 \ge c_{\mathcal{D}}^2 \ge {c^*_{\mathcal{D}}}^2 \Big/ \left(1 - \frac{\mathtt{C}^*_{(\boldsymbol{\mathcal{L}}_0)}\left\{m^{3/2} + \mathtt{C}_{bias} m^{5/2}\right\}\mathtt{r}^*}{c^*_{\mathcal{D}}\sqrt{n}}\right),$$

where $c^*_{\mathcal{D}} > 0$ is defined in Lemma A.5 and is independent of m, n and where $\mathtt{r}^* > 0$ is defined in (5.1).

Remark A.6. By the definition of $\mathtt{r}^* > 0$ in (5.1) it is clear that $c_{\mathcal{D}} \approx c^*_{\mathcal{D}}$, once $(m^2 + \mathtt{C}_{bias} m^3)/\sqrt{n} \to 0$.

To prove this claim, note that using Lemma A.5 we can prove the following result. It is proved very similarly to Lemma A.18:

Lemma A.9. *We have for any $\boldsymbol{\upsilon} \in \{\boldsymbol{\upsilon} \in \varUpsilon_m : \|\mathcal{D}_m(\boldsymbol{\upsilon}^*)(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^*)\| \le \mathtt{r}\}$ and with some constant $\mathtt{C}^*_{(\boldsymbol{\mathcal{L}}_0)} > 0$*

$$\|I - \mathcal{D}_m^{-1}(\boldsymbol{\upsilon}^*_m)\mathcal{D}_m^2(\boldsymbol{\upsilon}^*)\mathcal{D}_m^{-1}(\boldsymbol{\upsilon}^*_m)\| \le \frac{\mathtt{C}^*_{(\boldsymbol{\mathcal{L}}_0)}\left\{m^{3/2} + \mathtt{C}_{bias} m^{5/2}\right\}\mathtt{r}}{c^*_{\mathcal{D}}\sqrt{n}}.$$

We obtain the claim of Lemma A.8 because

$$\mathcal{D}_m^2(\boldsymbol{\upsilon}^*_m) - \mathcal{D}_m(\boldsymbol{\upsilon}^*_m)\left\{I - \mathcal{D}_m^{-1}(\boldsymbol{\upsilon}^*_m)\mathcal{D}_m^2(\boldsymbol{\upsilon}^*)\mathcal{D}_m^{-1}(\boldsymbol{\upsilon}^*_m\right\} = \mathcal{D}_m^2(\boldsymbol{\upsilon}^*),$$

such that using Lemma 5.1 and Lemma A.5

$$\left(1 + \frac{\mathtt{C}^*_{(\boldsymbol{\mathcal{L}}_0)}\left\{m^{3/2} + \mathtt{C}_{bias} m^{5/2}\right\}\mathtt{r}^*}{c^*_{\mathcal{D}}\sqrt{n}}\right)\mathcal{D}_m^2(\boldsymbol{\upsilon}^*_m) \ge \mathcal{D}_m^2(\boldsymbol{\upsilon}^*) \ge c^*_{\mathcal{D}}.$$

A.8.2. Some bounds for the score

Lemma A.10. *We have*

$$
\begin{aligned}
|\boldsymbol{f}'_{\boldsymbol{\eta}^*_m}(\boldsymbol{x})| &\le \mathtt{C},\\
|\boldsymbol{f}_{\boldsymbol{\eta}}(\mathbf{X}^\top\boldsymbol{\theta}) - \boldsymbol{f}_{\boldsymbol{\eta}^\circ}(\mathbf{X}^\top\boldsymbol{\theta}^\circ)| &\le \mathtt{C}\frac{\|\mathcal{D}(\boldsymbol{\upsilon}-\boldsymbol{\upsilon}^\circ)\|\sqrt{m}}{\sqrt{n}}\\
&\quad +\mathtt{C}\left(\frac{\|\mathcal{D}(\boldsymbol{\upsilon}^\circ-\boldsymbol{\upsilon}^*)\|m^2}{\sqrt{n}}+1\right). \qquad \text{(A.13)}
\end{aligned}
$$

Proof. Using assumption $(\mathbf{Cond}_{\boldsymbol{\eta}^*})$, that $|M(j)| \le 17$ (in (A.10)) and $k = (2^{j_k}-1)17 + r_k$ with $r_k \in \{0,\dots,2^{j_k}+16\}$ and $j_k \in \mathbb{N}_0$ we find as $\alpha > 2$

$$
\begin{aligned}
|\boldsymbol{f}'_{\boldsymbol{\eta}^*_m}(\boldsymbol{x})| &\le \sum_{j=0}^{j_m-1}\sum_{k\in M(j)} |\eta^*_{mk}||\boldsymbol{e}'_k(\boldsymbol{x})|\\
&\le \sqrt{17}\|\psi'\|_\infty \left(\sum_{j=0}^{j_m-1}\sum_{k\in M(j)} |\eta^*_{mk}|^2 2^{4j}\right)^{1/2}\left(\sum_{j=0}^{j_m-1} 2^{-4j}2^{3j}\right)^{1/2}\\
&\le \sqrt{17}\|\psi'\|_\infty \left(\sum_{k=0}^{m-1} |\eta^*_{mk}|^2 k^4\right)^{1/2}\left(\sum_{j=0}^{j_m-1} 2^{-j}\right)^{1/2}\\
&\le \sqrt{34}\|\psi'\|_\infty C_{\|\boldsymbol{\eta}^*_m\|},
\end{aligned}
$$

where with Lemma 5.1 and $m \in \mathbb{N}$ large enough $(m^5/n \to 0$ and $\mathtt{r}^* \cong m)$

$$
\begin{aligned}
C_{\|\boldsymbol{\eta}^*_m\|} &\le \left(\sum_{k=1}^{m-1} |\eta^*_{mk}|^2 k^4\right)^{1/2}\\
&\le \left(\sum_{k=0}^{m-1} |\eta^*{}_k|^2 k^4\right)^{1/2} + \left(\sum_{k=0}^{m-1} |\eta^*_{mk} - \eta^*{}_k|^2 k^4\right)^{1/2}\\
&\le \mathtt{C} + m^2\|(\boldsymbol{\eta}^*_m - \mathit{\Pi}_m\boldsymbol{\eta}^*)\|\\
&\le \mathtt{C} + \frac{m^2\mathtt{r}^*}{\sqrt{n}c_{\mathcal{D}}} \le \mathtt{C},
\end{aligned}
$$

For the second claim we bound (A.13) to bound

$$
\begin{aligned}
|\boldsymbol{f}_{\boldsymbol{\eta}}(\mathbf{X}^\top\boldsymbol{\theta}) - \boldsymbol{f}_{\boldsymbol{\eta}^\circ}(\mathbf{X}^\top\boldsymbol{\theta}^\circ)| &\le |\boldsymbol{f}_{\boldsymbol{\eta}-\boldsymbol{\eta}^\circ}(\mathbf{X}^\top\boldsymbol{\theta})| + |\boldsymbol{f}_{\boldsymbol{\eta}^\circ}(\mathbf{X}^\top\boldsymbol{\theta}) - \boldsymbol{f}_{\boldsymbol{\eta}^\circ}(\mathbf{X}^\top\boldsymbol{\theta}^\circ)|\\
&\le \frac{\mathtt{r}\sqrt{m}}{c_{\mathcal{D}}\sqrt{n}} + s_{\mathbf{X}}\|\boldsymbol{f}'_{\boldsymbol{\eta}^\circ}\|_\infty \mathtt{r}/\sqrt{n}.
\end{aligned}
$$

It remains to bound using that $m^5/n \to 0$ and that $\mathtt{r}^* \le \mathtt{C}\sqrt{m}$

$$\begin{aligned} |\boldsymbol{f}'_{\boldsymbol{\eta}^\circ}| &\le \sqrt{17}\left(\sum_{k=1}^{m} \boldsymbol{\eta}^\circ 2^4\right)^{1/2}\left(\sum_{j=1}^{j_m} 2^{(3-4)j}\right)^{1/2} \\ &\le \mathtt{C}\left(\frac{\|\mathcal{D}(\boldsymbol{\upsilon}^\circ - \boldsymbol{\upsilon}^*)\| m^2}{\sqrt{n}} + 1\right). \end{aligned}$$

□

Lemma A.11. *We have with $\varsigma_{i,m}$ from (A.1)*

$$\|\varsigma_{i,m}(\boldsymbol{\upsilon}^*_m)\| \le \mathtt{C}\sqrt{m},$$

and for any $\boldsymbol{\upsilon}, \boldsymbol{\upsilon}' \in \varUpsilon_\circ(\mathtt{r})$ with $\mathtt{r} \le \mathtt{C}\sqrt{m}(1 + \mathtt{C}_{bias}\log(n))$

$$\|\varsigma_{i,m}(\boldsymbol{\upsilon}) - \varsigma_{i,m}(\boldsymbol{\upsilon}')\| \le \mathtt{C} m^{3/2} \frac{\|\mathcal{D}_m(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}')\|}{\sqrt{n} c_{\mathcal{D}}}.$$

Proof. Note

$$\begin{aligned} \|\varsigma_{i,m}(\boldsymbol{\upsilon}^*_m)\| &= \|(\boldsymbol{f}'_{\boldsymbol{\eta}^*_m}(\mathbf{X}_i^\top \boldsymbol{\theta}^*_m) \nabla \varPhi^\top_{\varphi_{\boldsymbol{\theta}^*_m}} \mathbf{X}_i, \boldsymbol{e}(\mathbf{X}_i^\top \boldsymbol{\theta}^*_m))\| \\ &\le \|\boldsymbol{f}'_{\boldsymbol{\eta}^*_m}(\mathbf{X}_i^\top \boldsymbol{\theta}^*_m)\| \|\mathbf{X}_i\| + \|\boldsymbol{e}(\mathbf{X}_i^\top \boldsymbol{\theta}^*_m)\|. \end{aligned}$$

Such that with (A.8) and Lemma A.10

$$\|\varsigma_{i,m}(\boldsymbol{\upsilon}^*_m)\| \le (\mathtt{C}+1)\sqrt{34} s_{\mathbf{X}} \|\psi'\|_\infty + \sqrt{17}\|\psi\|_\infty \sqrt{m}. \tag{A.14}$$

For the second claim we use that for each $j = 1, \ldots, j_m - 1$

$$\begin{aligned} |N(j)| \overset{\text{def}}{=} \Big|\Big\{ \quad & k \in \{(2^j - 1)17, \ldots, 2^{j+1} + (j+1)17 - 1 - 1\} : \\ & |\boldsymbol{e}_k(\mathbf{X}_i^\top \boldsymbol{\theta}') - \boldsymbol{e}_k(\mathbf{X}_i^\top \boldsymbol{\theta})| \vee |\boldsymbol{e}'_k(\mathbf{X}_i^\top \boldsymbol{\theta}') - \boldsymbol{e}'_k(\mathbf{X}_i^\top \boldsymbol{\theta})| > 0 \Big\}\Big| \le 34. \end{aligned} \tag{A.15}$$

Furthermore we always have that

$$|\boldsymbol{e}'_k(\mathbf{X}_i^\top \boldsymbol{\theta}') - \boldsymbol{e}'_k(\mathbf{X}_i^\top \boldsymbol{\theta})| \le 2^{j_k 5/2} \|\psi''\|_\infty s_{\mathbf{X}} \|\boldsymbol{\theta} - \boldsymbol{\theta}'\|.$$

This implies again using that $\alpha > 2$ that $\frac{\mathtt{r}m}{n} \to 0$ for $\mathtt{r}^2 \le \mathtt{C}m$ and with $N(j) \subset \mathbb{N}$ from (A.15)

$$|\boldsymbol{f}'_{\boldsymbol{\eta}}(\boldsymbol{\theta}^\top \mathbf{X}_i) - \boldsymbol{f}'_{\boldsymbol{\eta}}(\mathbf{X}_i^\top \boldsymbol{\theta}')| \le \sqrt{34}(\mathtt{C}+1) m^{3/2} \|\boldsymbol{\theta} - \boldsymbol{\theta}'\| \|\psi''\|_\infty s_{\mathbf{X}}, \tag{A.16}$$

and with the same arguments

$$\begin{aligned}\|\boldsymbol{e}(\mathbf{X}_i^\top\boldsymbol{\theta})-\boldsymbol{e}(\mathbf{X}_i^\top\boldsymbol{\theta}')\| &\le \left(\sum_{k=1}^{m}|\boldsymbol{e}_k(\mathbf{X}_i^\top\boldsymbol{\theta})-\boldsymbol{e}_k(\mathbf{X}_i^\top\boldsymbol{\theta}')|^2\right)^{1/2} \\ &\le \sqrt{34}\left(\sum_{j=0}^{j_m-1}2^{3j}\right)^{1/2}\|\boldsymbol{\theta}-\boldsymbol{\theta}'\|\|\psi'\|_\infty s_{\mathbf{X}} \\ &\le \sqrt{34}m^{3/2}\|\boldsymbol{\theta}-\boldsymbol{\theta}'\|\|\psi'\|_\infty s_{\mathbf{X}},\end{aligned}\tag{A.17}$$

and

$$\begin{aligned}\|f'_{\boldsymbol{\eta}-\boldsymbol{\eta}'}(\boldsymbol{\theta}^\top\mathbf{X}_i)\nabla\varphi_{\boldsymbol{\theta}}^\top\mathbf{X}_i\| &\le s_{\mathbf{X}}\sum_{k=1}^{m}|\eta_k-\eta'_{m,k}||\boldsymbol{e}'_k(\boldsymbol{\theta}^\top\mathbf{X}_i)| \\ &\le \sqrt{34}\|\boldsymbol{\eta}-\boldsymbol{\eta}'\|s_{\mathbf{X}}\|\psi'\|_\infty\left(\sum_{j=0}^{j_m-1}2^{3j}\right)^{1/2} \\ &\le \sqrt{34}\|\boldsymbol{\eta}-\boldsymbol{\eta}'\|s_{\mathbf{X}}\|\psi'\|_\infty m^{3/2}.\end{aligned}\tag{A.18}$$

Finally similar to (A.14) we have with $M(j)\subset\mathbb{N}$ from (A.10)

$$|\boldsymbol{f}'_{\boldsymbol{\eta}}(\mathbf{X}_i^\top\boldsymbol{\theta})| \le \sqrt{34}\|\psi'\|_\infty(C_{\|\boldsymbol{\eta}^*\|}+1),$$

such that

$$\begin{aligned}&\|\boldsymbol{f}'_{\boldsymbol{\eta}'}(\mathbf{X}_i^\top\boldsymbol{\theta}')(\nabla\varphi_{\boldsymbol{\theta}}^\top-\nabla\Phi_{\boldsymbol{\theta}'}^\top)\mathbf{X}_i\| \\ &\quad\le \|\psi'\|_\infty(C_{\|\boldsymbol{\eta}^*\|}+1)\sqrt{34}L_{\nabla\Phi.}\|\boldsymbol{\theta}-\boldsymbol{\theta}'\|s_{\mathbf{X}}.\end{aligned}\tag{A.19}$$

We get combining (A.19), (A.16), (A.18) and (A.17)

$$\|\varsigma_{i,m}(\boldsymbol{\upsilon})-\varsigma_{i,m}(\boldsymbol{\upsilon}')\| \le \mathtt{C}m^{3/2}\frac{\|\mathcal{D}_m(\boldsymbol{\upsilon}-\boldsymbol{\upsilon}')\|}{\sqrt{n}c_{\mathcal{D}}},$$

where we used Lemma A.8 to find that

$$\begin{aligned}\|\boldsymbol{\theta}-\boldsymbol{\theta}'\|\vee\|\boldsymbol{\eta}-\boldsymbol{\eta}'\| &\le \sqrt{\|\boldsymbol{\theta}-\boldsymbol{\theta}'\|^2+\|\boldsymbol{\eta}-\boldsymbol{\eta}'\|^2}\le\|\boldsymbol{\upsilon}-\boldsymbol{\upsilon}'\| \\ &\le \frac{\|\mathcal{D}_m(\boldsymbol{\upsilon}-\boldsymbol{\upsilon}')\|}{\sqrt{n}c_{\mathcal{D}}}.\end{aligned}$$

□

A.8.3. Crude deviation bounds for sums of random matrices

The next auxiliary Lemma relies on a non-commutative Bernstein inequality; see Theorem 1.4 of [22].

Lemma A.12. *Suppose that $\boldsymbol{h}_i \in \mathbb{R}^{p_1}$ are iid random vectors, where $p \in \mathbb{N}$. Define*

$$\mathbf{S}_n^* := \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{h}_i \boldsymbol{h}_i^\top - \mathbb{E}[\boldsymbol{h}_1 \boldsymbol{h}_1^\top],$$

and $B^2 := \mathbb{E}[\|\boldsymbol{h}_1\|^4]$. Assume that $\|\boldsymbol{h}_{i,m}\boldsymbol{h}_{i,m}^\top\| = \|\mathbf{M}_i\| \le U \in \mathbb{R}$ then it holds

$$\mathbb{P}\big(\|\mathbf{S}_n^*\| > n^{-1}t\big) \le 2p_1 \exp\Big\{-\frac{t^2}{4nB^2 + 2Ut/3}\Big\}$$

Proof. This lemma is an immediate consequence of the non-commutative Bernstein inequality (Theorem 1.4 in [22]). We only have to note that

$$\sum_{i=1}^{n} \mathbb{E}[\mathbf{M}_i^2] \le 2n\mathbb{E}[\|\boldsymbol{h}_1\|^4] = 2nB^2. \qquad \square$$

Lemma A.13. *We have with $\mathtt{x} \le 9n/2 - \log(2m)$ that*

$$\mathbb{P}\left(\|\mathbf{S}_n\| \ge \mathtt{C}\sqrt{8}m\Big(\mathtt{x} + \log(2m)\Big)^{1/2}/\sqrt{n}\right) \le \mathrm{e}^{-\mathtt{x}}.$$

where with $\varsigma_{i,m}$ from (A.1)

$$\mathbf{S}_n \;=\; \frac{1}{n}\sum_{i=1}^{n} \varsigma_{i,m}(\boldsymbol{\upsilon}_m^*)\varsigma_{i,m}(\boldsymbol{\upsilon}_m^*)^\top - \frac{1}{n}\mathcal{V}_m^2(\boldsymbol{\upsilon}_m^*).$$

Proof. We want to employ lemma A.12. We estimate using Lemma A.11

$$\|\varsigma_{i,m}(\boldsymbol{\upsilon}_m^*)\varsigma_{i,m}(\boldsymbol{\upsilon}_m^*)^\top\| \le \mathtt{C}m,$$

such that $\|\varsigma_{i,m}\varsigma_{i,m}^\top\| =: \|\mathbf{M}_i\| \le \mathtt{C}m$. Furthermore

$$\mathbb{E}[\|\varsigma_{i,m}(\boldsymbol{\upsilon}_m^*)\|^4] \le \mathtt{C}^2 m^2.$$

Plugging these bounds into lemma A.12 we get

$$\mathbb{P}(\|\mathbf{S}_n\| \ge n^{-1}t) \le 2m \exp\Big\{-\frac{t^2}{4n\mathtt{C}^2m^2 + 2\mathtt{C}mt/3}\Big\}.$$

Setting $t = \mathtt{C}\sqrt{8n}m\Big(\mathtt{x} + \log(2m)\Big)^{1/2}$ and $\mathtt{x} \le 9n/2 - \log(2m)$ this gives

$$\mathbb{P}\left(\|\mathbf{S}_n\| \ge \mathtt{C}\sqrt{8}m\Big(\mathtt{x} + \log(2m)\Big)^{1/2}/\sqrt{n}\right) \le \mathrm{e}^{-\mathtt{x}}. \qquad \square$$

Lemma A.14. *We have with* $\mathtt{x} \le 9n/2 - \log(2m)$ *that*

$$\mathbb{P}\left(\|\mathbf{S}_n\| \ge \sqrt{8}\mathtt{C}(p^* + \mathtt{x})^4 \Big(\mathtt{x} + \log(2m)\Big)^{1/2} / \sqrt{n} \right) \le \mathrm{e}^{-\mathtt{x}}.$$

where with $\varsigma_{i,m}$ *from* (A.1)

$$\mathbf{S}_n(\boldsymbol{\upsilon}) = \frac{1}{n}\sum_{i=1}^{n} \varsigma_{i,m}(\boldsymbol{\upsilon})\varsigma_{i,m}(\boldsymbol{\upsilon})^\top - \frac{1}{n}\mathcal{V}_m^2(\boldsymbol{\upsilon}).$$

Proof. We want to employ lemma A.12. We estimate using Lemma A.11 and that $\mathtt{r}^\circ \le \mathtt{C}\sqrt{p^* + \mathtt{x}}$

$$\begin{aligned} \|\varsigma_{i,m}(\boldsymbol{\upsilon})\varsigma_{i,m}(\boldsymbol{\upsilon})^\top\| &\le 3\|\varsigma_{i,m}(\boldsymbol{\upsilon}_m^*)\|^2 + 3\|\varsigma_{i,m}(\boldsymbol{\upsilon}_m^*) - \varsigma_{i,m}(\boldsymbol{\upsilon})\|^2 \\ &\le \mathtt{C}m + \mathtt{C}\frac{\|\mathcal{D}_m(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}_m^*)\|^2 m^3}{n} \\ &\le \mathtt{C}m + \mathtt{C}m^3\mathtt{r}^2/n. \end{aligned}$$

such that $\|\varsigma_{i,m}\varsigma_{i,m}^\top\| =: \|\mathbf{M}_i\| \le \mathtt{C}m$. Furthermore

$$\mathbb{E}[\|\varsigma_{i,m}(\boldsymbol{\upsilon}_m^*)\|^4] \le \mathtt{C}^2(m^2 + m^6\mathtt{r}^4/n^2).$$

Plugging these bounds into lemma A.12 we get

$$\mathbb{P}(\|\mathbf{S}_n\| \ge n^{-1}t) \le 2m\exp\Big\{ -\frac{t^2}{4n\mathtt{C}^2(m^2 + m^6\mathtt{r}^4/n^2) + 2\mathtt{C}\,(m + m^3\mathtt{r}^2/n)\,t/3} \Big\}.$$

Setting $t = \sqrt{8n}\mathtt{C}\,(m + m^3\mathtt{r}^2/n)\,\Big(\mathtt{x} + \log(2m)\Big)^{1/2}/n^2$ and $\mathtt{x} \le 9n/2 - \log(2m)$ this implies

$$\mathbb{P}\left(\|\mathbf{S}_n\| \ge \sqrt{8}\mathtt{C}\,(m + m^3\mathtt{r}^2/n)\,\Big(\mathtt{x} + \log(2m)\Big)^{1/2}/\sqrt{n} \right) \le \mathrm{e}^{-\mathtt{x}}. \qquad \square$$

Lemma A.15. *We have with*

$$t = \mathtt{C}\|\mathcal{D}_m(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}')\|^2 m^3 \Big(\mathtt{x} + \log(2m)\Big)^{1/2} \sqrt{5/n},$$

and $\mathtt{x} \le 9n/2 - \log(2m)$

$$\mathbb{P}(\|\mathbf{S}_n\| \ge n^{-1}t) \le \mathrm{e}^{-\mathtt{x}},$$

where with $\boldsymbol{\upsilon} \in \Upsilon_\circ(\mathtt{r})$ and with $\varsigma_{i,m}$ from (A.1)

$$\begin{aligned}\mathbf{S}_n &= \frac{1}{n}\sum_{i=1}^{n}(\varsigma_{i,m}(\boldsymbol{\upsilon}') - \varsigma_{i,m}(\boldsymbol{\upsilon}))(\varsigma_{i,m}(\boldsymbol{\upsilon}') - \varsigma_{i,m}(\boldsymbol{\upsilon}))^\top \\ &\quad -\mathbb{E}(\varsigma_{i,m}(\boldsymbol{\upsilon}') - \varsigma_{i,m}(\boldsymbol{\upsilon}))(\varsigma_{i,m}(\boldsymbol{\upsilon}') - \varsigma_{i,m}(\boldsymbol{\upsilon}))^\top.\end{aligned}$$

Proof. We estimate using Lemma A.11

$$\begin{aligned}&\|(\varsigma_{i,m}(\boldsymbol{\upsilon}') - \varsigma_{i,m}(\boldsymbol{\upsilon}))(\varsigma_{i,m}(\boldsymbol{\upsilon}') - \varsigma_{i,m}(\boldsymbol{\upsilon}))^\top\| \\ &\quad \leq \|\varsigma_{i,m}(\boldsymbol{\upsilon}') - \varsigma_{i,m}(\boldsymbol{\upsilon})\|^2 \\ &\quad \leq \mathtt{C}\frac{\|\mathcal{D}_m(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}')\|^2 m^3}{n}.\end{aligned}$$

With the same estimates we obtain

$$\mathbb{E}[\|\varsigma_{i,m}(\boldsymbol{\upsilon}') - \varsigma_{i,m}(\boldsymbol{\upsilon})\|^4] \leq \mathtt{C}m^6\frac{\|\mathcal{D}_m(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}')\|^4}{n^2}.$$

Plugging these bounds into Lemma A.12 we get with $d(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}') \overset{\mathrm{def}}{=} \|\mathcal{D}_m(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}')\|$

$$\begin{aligned}&\mathbb{P}(\|\mathbf{S}_n\| \geq n^{-1}t) \\ &\quad \leq 2m\exp\Bigl\{-\frac{t^2}{4d(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}')^4\mathtt{C}n^{-1}m^6 + 2d(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}')^2\mathtt{C}m^3n^{-1}t/3}\Bigr\}.\end{aligned}$$

Setting $t = \mathtt{C}\|\mathcal{D}_m(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}')\|^2\sqrt{8/n}m^3\Bigl(\mathtt{x} + \log(2m)\Bigr)^{1/2}$ and $\mathtt{x} \leq 9n/2 - \log(2m)$ this yields

$$\mathbb{P}(\|\mathbf{S}_n\| \geq n^{-1}t) \leq \mathrm{e}^{-\mathtt{x}}. \qquad \square$$

A.8.4. Conditions $(\mathcal{I})$

Proof. This follows from $\mathcal{D} \geq c_{\mathcal{D}}Id$ with Lemma B.5 of [2] where

$$\rho^2 \leq 1 - \frac{nc_{\mathcal{D}}}{\lambda_{\max}\mathrm{D} \wedge \lambda_{\max}\mathcal{H}} \leq 1 - \frac{c_{\mathcal{D}}}{\mathtt{C}p}.$$

where we used Lemma A.6 to bound $\lambda_{\max}\mathrm{D} \leq \mathtt{C}p$ in the last step. Finally we have $c_{\mathcal{D}} > 0$ as shown in Lemma A.8. □

A.8.5. Conditions $(\mathfrak{E}\mathfrak{D}_0)$, $(\mathfrak{E}\mathtt{r})$ and $(\mathfrak{E}\mathfrak{D}_{1,m})$

Lemma A.16. *With probability greater than $1-3\mathrm{e}^{-\mathtt{x}}$ we have $(\mathfrak{E}\mathfrak{D}_0)$ with*

$$g = \frac{\sqrt{n}}{\mathtt{C}(1+\sqrt{m})}, \qquad \nu_m^2 = 2\widetilde{\nu}^2\sigma^2,$$

and $(\mathfrak{E}\mathtt{r})$ with

$$\begin{aligned}
\mathtt{g}(\mathtt{r}) &= \sqrt{n}c_{\mathfrak{D}}\widetilde{g}\mathtt{C}\left(\sqrt{m}+m^{3/2}\mathtt{r}/\sqrt{n}\right)^{-1},\\
\nu_{\mathtt{r},m}^2 &= \widetilde{\nu}^2\Big(1+\mathtt{C}\left(m^{3/2}+\mathtt{r}m^2/\sqrt{n}\right)\mathtt{r}/\sqrt{n}\\
&\qquad +\mathtt{C}\left(m+m^3\mathtt{r}^2/n\right)\Big(\mathtt{x}+\log(2m)\Big)^{1/2}/\sqrt{n}\Big).
\end{aligned}$$

where $\mathtt{C}_{(\mathfrak{E}\mathtt{r})}>0$ is independent of $n, m, \mathtt{x}, \mathtt{r}$.

Proof. Lemma A.8 gives with $\widetilde{\boldsymbol{\gamma}} = \mathcal{V}_m^{1/2}\gamma/\|\mathcal{V}_m^{1/2}\gamma\|$

$$\frac{\langle\nabla\zeta(\boldsymbol{\upsilon}_m^*),\gamma\rangle_{\mathbb{R}^{p^*}}}{\|\mathcal{V}_m\gamma\|} = \langle\widetilde{\gamma}^\top\mathcal{V}_m^{-1}A(\boldsymbol{\upsilon}_m^*),\varepsilon\rangle_{\mathbb{R}^n}.$$

Consequently – using Lemma A.11 – we get with $\mu \le \dfrac{\sqrt{n}}{\mathtt{C}(1+\sqrt{m})}$, with $\varsigma_{i,m}$ from (A.1) and assumption $(\mathbf{Cond}_\varepsilon)$

$$\begin{aligned}
&\sup_{\gamma\in\mathbb{R}^{p^*}}\log\mathbb{E}_\varepsilon\exp\left\{\mu\frac{\langle\nabla\zeta(\boldsymbol{\upsilon}_m^*),\gamma\rangle}{\|\mathcal{V}_m(\boldsymbol{\upsilon}_m^*)\gamma\|}\right\}\\
&\quad\le\sum_{i=1}^n\sup_{\gamma\in\mathbb{R}^{p^*},\ \|\widetilde{\gamma}\|=1}\log\mathbb{E}\exp\left\{\mu\langle\widetilde{\gamma},\mathcal{V}_m^{-1}(\boldsymbol{\upsilon}_m^*)\varsigma_{i,m}(\boldsymbol{\upsilon}_m^*)\rangle\varepsilon_i\right\}\\
&\quad\le\widetilde{\nu}^2\mu^2\widetilde{\gamma}^\top\mathcal{V}_m^{-1}(\boldsymbol{\upsilon}_m^*)\left(\sum_{i=1}^n\varsigma_{i,m}(\boldsymbol{\upsilon}_m^*)\varsigma_{i,m}(\boldsymbol{\upsilon}_m^*)^\top\right)\mathcal{V}_m^{-1}(\boldsymbol{\upsilon}^*)\widetilde{\gamma}\\
&\quad=\widetilde{\nu}^2\mu^2+\widetilde{\nu}^2\mu^2\widetilde{\gamma}^\top\mathcal{V}_m^{-1}(\boldsymbol{\upsilon}_m^*)n\mathbf{S}_n\mathcal{V}_m^{-1}(\boldsymbol{\upsilon}_m^*)\widetilde{\gamma}\\
&\quad\le\widetilde{\nu}^2\mu^2+\widetilde{\nu}^2\mu^2\kappa_n,
\end{aligned}\tag{A.20}$$

where

$$\begin{aligned}
\kappa_n &= \widetilde{\gamma}^\top\left(n^{-1}\mathcal{V}_m\right)^{-1/2}\mathbf{S}_n\left(n^{-1}\mathcal{V}_m\right)^{-1/2}\widetilde{\gamma},\\
\mathbf{S}_n &= \frac{1}{n}\sum_{i=1}^n\varsigma_{i,m}(\boldsymbol{\upsilon}_m^*)\varsigma_{i,m}(\boldsymbol{\upsilon}_m^*)^\top-\frac{1}{n}\mathcal{V}_m(\boldsymbol{\upsilon}_m^*).
\end{aligned}$$

With Lemma A.13 we infer that if $\mathtt{x} \leq 9n/2 - \log(2m)$

$$\mathbb{P}\left(\|\mathbf{S}_n\| \geq C_M\sqrt{8}m\Big(\mathtt{x}+\log(2m)\Big)^{1/2}/\sqrt{n}\right) \leq \mathrm{e}^{-\mathtt{x}}.$$

Consequently with probability greater than $1-\mathrm{e}^{-\mathtt{x}}$ we find that for $n \in \mathbb{N}$ large enough

$$\kappa_n \leq \frac{\mathtt{C}_M\sqrt{8}m\Big(\mathtt{x}+\log(2m)\Big)^{1/2}}{\sqrt{n}\sigma^2 c_{\mathcal{D}}^2} \leq 1.$$

Thus we get $\boldsymbol{(\mathcal{E}\mathfrak{D}_0)}$ with probability greater than $1-\mathrm{e}^{-\mathtt{x}}$ and

$$g = \frac{\sqrt{n}}{\mathtt{C}(1+\sqrt{m})}, \quad \nu_m^2 = 2\widetilde{\nu}^2.$$

Concerning $\boldsymbol{(\mathcal{E}\mathtt{r})}$ we bound using the same arguments as in the proof of Lemma A.18

$$\begin{aligned}\|\mathcal{V}_m(\boldsymbol{\upsilon})^{-1}\mathcal{V}_m(\boldsymbol{\upsilon}_m^*)\|^2 &\leq 1+\|I-\mathcal{V}_m(\boldsymbol{\upsilon})^{-1}\mathcal{V}_m(\boldsymbol{\upsilon}_m^*)^2\mathcal{V}_m(\boldsymbol{\upsilon})^{-1}\| \\ &\leq 1+\mathtt{C}\left(m^{3/2}+\mathtt{r}m^2/\sqrt{n}\right)\mathtt{r}/\sqrt{n}.\end{aligned}$$

Thus we get with the arguments from above $\boldsymbol{(\mathcal{E}\mathtt{r})}$ using Lemma A.14 with probability greater than $1-\mathrm{e}^{-\mathtt{x}}$ and

$$\begin{aligned}\mathtt{g}(\mathtt{r}) &= \sqrt{n}c_{\mathcal{D}}\widetilde{g}\mathtt{C}\left(\sqrt{m}+m^{3/2}\mathtt{r}/\sqrt{n}\right)^{-1}, \\ \nu_{\mathtt{r},m}^2 &= \widetilde{\nu}^2\Big(1+\mathtt{C}\left(m^{3/2}+\mathtt{r}m^2/\sqrt{n}\right)\mathtt{r}/\sqrt{n} \\ &\quad +\mathtt{C}\left(m+m^3\mathtt{r}^2/n\right)\Big(\mathtt{x}+\log(2m)\Big)^{1/2}/\sqrt{n}\Big).\end{aligned}$$

□

Lemma A.17. *With probability greater than* $1-\mathrm{e}^{-\mathtt{x}}$ *we have* $\boldsymbol{(\mathcal{E}\mathfrak{D}_1)}$ *with*

$$\mathtt{g} \overset{\text{def}}{=} \frac{\sqrt{n}}{\mathtt{Cr}m^{3/2}}, \quad \omega \overset{\text{def}}{=} \frac{2}{\sqrt{n}c_{\mathcal{D}}}, \quad \nu_{1,m}^2 = \widetilde{\nu}^2\mathtt{C}m^2.$$

Proof. We get with Lemma A.11, with Lemma A.8 and with $\varsigma_{i,m}$ from (A.1)

$$\|\mathcal{D}_m^{-1}(\varsigma_{i,m}(\boldsymbol{\upsilon})-\varsigma_{i,m}(\boldsymbol{\upsilon}'))\| \leq \mathtt{C}\frac{2m^{3/2}}{nc_{\mathcal{D}}^2}\|\mathcal{D}_m(\boldsymbol{\upsilon}-\boldsymbol{\upsilon}')\|,$$

We get with,

$$\begin{aligned}\mu &\leq \mathtt{g} \overset{\text{def}}{=} \frac{\sqrt{n}}{\mathtt{Cr}m^{3/2}} \\ \omega &\overset{\text{def}}{=} \frac{2}{\sqrt{n}c_{\mathcal{D}}},\end{aligned}$$

and the same calculations as in (A.20) with some $\boldsymbol{\upsilon}, \boldsymbol{\upsilon}' \in \Upsilon_\circ(\mathtt{r})$, $\boldsymbol{\gamma} \in \mathbb{R}^{p^*}$ and $\|\boldsymbol{\gamma}\| = 1$

$$
\begin{aligned}
&\log \mathbb{E}_\varepsilon[\exp\left\{\mu \frac{\boldsymbol{\gamma}^\top \mathcal{D}_m^{-1}\left(\nabla\zeta(\boldsymbol{\upsilon}) - \nabla\zeta(\boldsymbol{\upsilon}')\right)}{\omega\|\mathcal{D}_m(\boldsymbol{\upsilon}-\boldsymbol{\upsilon}')\|}\right\}] \\
&\quad \leq \sum_{i=1}^n \log \mathbb{E}_\varepsilon[\exp\left\{\mu\varepsilon_i \frac{\gamma^\top \mathcal{D}_m^{-1}(\varsigma_{i,m}(\boldsymbol{\upsilon}) - \varsigma_{i,m}(\boldsymbol{\upsilon}'))}{\omega\|\mathcal{D}_m(\boldsymbol{\upsilon}-\boldsymbol{\upsilon}')\|}\right\}] \\
&\quad \leq \frac{\mu^2\widetilde{\nu}^2}{2}(\omega\|\mathcal{D}_m(\boldsymbol{\upsilon}-\boldsymbol{\upsilon}')\|)^{-2} \\
&\qquad n\gamma^\top \mathcal{D}_m^{-1}\left(\sum_{i=1}^n (\varsigma_{i,m}(\boldsymbol{\upsilon}') - \varsigma_{i,m}(\boldsymbol{\upsilon}))(\varsigma_{i,m}(\boldsymbol{\upsilon}') - \varsigma_{i,m}(\boldsymbol{\upsilon}))^\top\right)\mathcal{D}_m^{-1}\gamma^\top.
\end{aligned}
$$

We estimate

$$
\begin{aligned}
&\widetilde{\gamma}^\top \mathcal{D}_m^{-1}\left(\sum_{i=1}^n (\varsigma_{i,m}(\boldsymbol{\upsilon}') - \varsigma_{i,m}(\boldsymbol{\upsilon}))(\varsigma_{i,m}(\boldsymbol{\upsilon}') - \varsigma_{i,m}(\boldsymbol{\upsilon}))^\top\right)\mathcal{D}_m^{-1}\widetilde{\gamma}^\top \\
&\quad \leq \|\mathcal{D}_m^{-1} n\mathbb{E}\left[(\varsigma_{i,m}(\boldsymbol{\upsilon}') - \varsigma_{i,m}(\boldsymbol{\upsilon}))(\varsigma_{i,m}(\boldsymbol{\upsilon}') - \varsigma_{i,m}(\boldsymbol{\upsilon}))^\top\right]\mathcal{D}_m^{-1}\| + \kappa_n \\
&\quad \leq \mathbb{E}\|\left(n^{-1/2}\mathcal{D}_m\right)^{-1}(\varsigma_{i,m}(\boldsymbol{\upsilon}') - \varsigma_{i,m}(\boldsymbol{\upsilon}))\|^2 + \kappa_n,
\end{aligned}
$$

where

$$
\begin{aligned}
\kappa_n &= \|\left(n^{-1/2}\mathcal{D}_m\right)^{-1}\mathbf{S}_n\left(n^{-1/2}\mathcal{D}_m\right)^{-1}\|, \\
\mathbf{S}_n &= \frac{1}{n}\sum_{i=1}^n (\varsigma_{i,m}(\boldsymbol{\upsilon}') - \varsigma_{i,m}(\boldsymbol{\upsilon}))(\varsigma_{i,m}(\boldsymbol{\upsilon}') - \varsigma_{i,m}(\boldsymbol{\upsilon}))^\top \\
&\quad -\mathbb{E}(\varsigma_{i,m}(\boldsymbol{\upsilon}') - \varsigma_{i,m}(\boldsymbol{\upsilon}))(\varsigma_{i,m}(\boldsymbol{\upsilon}') - \varsigma_{i,m}(\boldsymbol{\upsilon}))^\top.
\end{aligned}
$$

To control $\kappa_n > 0$ we apply Lemma A.15 and we infer that with $t = C_M^2\|\mathcal{D}_m(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}')\|^2\sqrt{5/n}m^3\left(\mathtt{x} + \log(2m)\right)^{1/2}$ and $\mathtt{x} \leq 9n/2 - \log(2m)$ the set $\{\|\mathbf{S}_n\| \leq n^{-1}t\}$ is of dominating probability and on this set we find

$$
\begin{aligned}
\kappa_n &\leq \frac{C_M^2\|\mathcal{D}_m(\boldsymbol{\upsilon}-\boldsymbol{\upsilon}')\|^2\left(\mathtt{x}+\log(2m)\right)^{1/2}m^3\sqrt{5/n}}{nc_{\mathcal{D}}^2} \\
&\leq \omega^2\|\mathcal{D}_m(\boldsymbol{\upsilon}-\boldsymbol{\upsilon}')\|^2\frac{C_M^2\sqrt{5}m^3\left(\mathtt{x}+\log(2m)\right)^{1/2}}{\sqrt{n}}.
\end{aligned}
$$

For $\mathtt{r} \le \mathtt{r}_0 \le \mathtt{C}_{\mathtt{r}}\sqrt{p^* + \mathtt{x}}$ this gives because $m^{5/2}/\sqrt{n} \to 0$

$$\kappa_n \le \mathtt{C}_\kappa \sqrt{\Big(\mathtt{x} + \log(2m)\Big)p^*}.$$

We calculate with some $(\boldsymbol{\theta}^\circ, \boldsymbol{\eta}^\circ) \overset{\mathrm{def}}{=} (\frac{1}{\sqrt{n}}\mathcal{D}_m)^{-1}\boldsymbol{\gamma}$

$$\begin{aligned} & n\boldsymbol{\gamma}^\top \mathcal{D}_m^{-1}\mathbb{E}\left[(\varsigma_{1,m}(\boldsymbol{\upsilon}') - \varsigma_{1,m}(\boldsymbol{\upsilon}))(\varsigma_{1,m}(\boldsymbol{\upsilon}') - \varsigma_{1,m}(\boldsymbol{\upsilon}))^\top\right]\mathcal{D}_m^{-1}\boldsymbol{\gamma}^\top \\ &\quad = \mathbb{E}\left[\left\{\left[\boldsymbol{f}'_{\boldsymbol{\eta}}(\mathbf{X}^\top\boldsymbol{\theta}) - \boldsymbol{f}'_{\boldsymbol{\eta}'}(\mathbf{X}^\top\boldsymbol{\theta}')\right](\mathbf{X}^\top\boldsymbol{\theta}^\circ)^2 + \boldsymbol{f}_{\boldsymbol{\eta}^\circ}(\mathbf{X}^\top\boldsymbol{\theta}) - \boldsymbol{f}_{\boldsymbol{\eta}^\circ}(\mathbf{X}^\top\boldsymbol{\theta}')\right\}^2\right] \\ &\quad \le 2\mathbb{E}\left[\left\{\left[\boldsymbol{f}'_{\boldsymbol{\eta}}(\mathbf{X}^\top\boldsymbol{\theta}) - \boldsymbol{f}'_{\boldsymbol{\eta}'}(\mathbf{X}^\top\boldsymbol{\theta}')\right](\mathbf{X}^\top\boldsymbol{\theta}^\circ)^2\right\}^2\right] \\ &\qquad + 2\mathbb{E}\left[\left\{\boldsymbol{f}_{\boldsymbol{\eta}^\circ}(\mathbf{X}^\top\boldsymbol{\theta}) - \boldsymbol{f}_{\boldsymbol{\eta}^\circ}(\mathbf{X}^\top\boldsymbol{\theta}')\right\}^2\right]. \end{aligned}$$

We estimate separately

$$\begin{aligned} & \mathbb{E}\left[\left\{\left[\boldsymbol{f}'_{\boldsymbol{\eta}}(\mathbf{X}^\top\boldsymbol{\theta}) - \boldsymbol{f}'_{\boldsymbol{\eta}'}(\mathbf{X}^\top\boldsymbol{\theta}')\right](\mathbf{X}^\top\boldsymbol{\theta}^\circ)^2\right\}^2\right] \\ &\qquad \le 2s_{\mathbf{X}}^4\left(\mathbb{E}\left[\left\{\boldsymbol{f}'_{\boldsymbol{\eta}-\boldsymbol{\eta}'}(\mathbf{X}^\top\boldsymbol{\theta})\right\}^2\right] + \mathbb{E}\left[\left\{\boldsymbol{f}'_{\boldsymbol{\eta}'}(\mathbf{X}^\top\boldsymbol{\theta}) - \boldsymbol{f}'_{\boldsymbol{\eta}'}(\mathbf{X}^\top\boldsymbol{\theta}')\right\}^2\right]\right). \end{aligned}$$

We again estimate separately denoting $\boldsymbol{\gamma} = (\boldsymbol{\eta} - \boldsymbol{\eta}')/\|\boldsymbol{\eta} - \boldsymbol{\eta}'\|$

$$\mathbb{E}\left[\left\{\boldsymbol{f}'_{\boldsymbol{\eta}-\boldsymbol{\eta}'}(\mathbf{X}^\top\boldsymbol{\theta})\right\}^2\right] = \|\boldsymbol{\eta} - \boldsymbol{\eta}'\|^2 2\sum_{k=1}^m\sum_{l=k}^m (1 - 1_{k=l}/2)\gamma_k\gamma_l\mathbb{E}[\boldsymbol{e}'_l\boldsymbol{e}'_l(\mathbf{X}^\top\boldsymbol{\theta})],$$

We have with $l = 2^{j_l} + 17j_l - 1 + r_l \in \mathbb{N}$ and $k = (2^{j_k} - 1)17 + r_k \in \mathbb{N}$ using (A.4)

$$\mathbb{E}[\boldsymbol{e}'_k\boldsymbol{e}'_l(\mathbf{X}^\top\boldsymbol{\theta})] \le 17\mathtt{C}2^{j_k}\|\psi'\|_\infty^2 2^{j_l}1_{I_k\cap I_l\neq 0}. \tag{A.21}$$

This implies

$$\begin{aligned} & \frac{1}{\|\boldsymbol{\eta} - \boldsymbol{\eta}'\|^2}\mathbb{E}\left[\left\{\boldsymbol{f}'_{\boldsymbol{\eta}-\boldsymbol{\eta}'}(\mathbf{X}^\top\boldsymbol{\theta})\right\}^2\right] \\ &\quad = \sum_{k=1}^m\sum_{l=k}^m (1 - 1_{k=l}/2)\gamma_l\gamma_k\mathbb{E}[\boldsymbol{e}'_l\boldsymbol{e}'_l(\mathbf{X}^\top\boldsymbol{\theta})] \\ &\quad \le 17\mathtt{C}\|\psi'\|_\infty\sum_{k=0}^m\gamma_k 2^{j_k}\left(\sum_{j=j_k}^{j_m}\sum_{r=0}^{2^j+17-1} 2^{2j_l}1_{I_k\cap I_l\neq 0}((2^j - 1)17 + r, k)\right)^{1/2} \\ &\quad \le 17\mathtt{C}\|\psi'\|_\infty\sum_{k=0}^m\gamma_k 2^{j_k}\left(\sum_{j=j_k}^{j_m} 2^{2j_l}\ulcorner 2^{(j_l-j_k)}17\urcorner\right)^{1/2} \end{aligned}$$

$$
\begin{aligned}
&\le \sqrt{18}17\mathtt{C}\|\psi'\|_\infty \sum_{k=0}^{m} \gamma_k 2^{j_k/2}\left(\sum_{j=j_k}^{j_m} 2^{3j_l}\right)^{1/2} \\
&\le 18^2\mathtt{C}m^2. \qquad\qquad (\text{A.22})
\end{aligned}
$$

Furthermore

$$
\begin{aligned}
\mathbb{E}\left[\left\{\boldsymbol{f}'_{\boldsymbol{\eta}'}(\mathbf{X}^\top\boldsymbol{\theta}) - \boldsymbol{f}'_{\boldsymbol{\eta}'}(\mathbf{X}^\top\boldsymbol{\theta}')\right\}^2\right] &= 2\sum_{k=1}^{m}\sum_{l=k}^{m}(1-1_{k=l}/2)\eta'_k\eta'_l \\
&\quad \mathbb{E}\left[(\boldsymbol{e}'_k(\mathbf{X}^\top\boldsymbol{\theta}) - \boldsymbol{e}'_k(\mathbf{X}^\top\boldsymbol{\theta}'))(\boldsymbol{e}'_l(\mathbf{X}^\top\boldsymbol{\theta}) - \boldsymbol{e}'_l(\mathbf{X}^\top\boldsymbol{\theta}'))\right].
\end{aligned}
$$

With (A.6) this gives

$$
\begin{aligned}
&\mathbb{E}\left[\left\{\boldsymbol{f}'_{\boldsymbol{\eta}'}(\mathbf{X}^\top\boldsymbol{\theta}) - \boldsymbol{f}'_{\boldsymbol{\eta}'}(\mathbf{X}^\top\boldsymbol{\theta}')\right\}^2\right] \\
&\quad\le \|\boldsymbol{\theta}-\boldsymbol{\theta}'\|^2\mathtt{C}\|\psi''\|_\infty^2 s_{\mathbf{X}}^4 17^2 2\sum_{k=1}^{m}\eta'_k 2^{2j_k}\sum_{l=k}^{m}\eta'_l 2^{2j_l}1_{\{I_k\cap I_l\neq\emptyset\}} \\
&\quad\le \|\boldsymbol{\theta}-\boldsymbol{\theta}'\|^2\mathtt{C}\|\psi''\|_\infty^2 s_{\mathbf{X}}^4 17^2 2 \\
&\qquad \sum_{k=1}^{m}\eta'_k 2^{2j_k}\left(\sum_{l=k}^{m}{\eta'_l}^2 k^4\right)^{1/2}\left(\sum_{l=k}^{m}1_{\{I_k\cap I_l\neq\emptyset\}}\right)^{1/2}.
\end{aligned}
$$

As always

$$
\mathtt{r} \le \mathtt{C}\sqrt{p^*}(1+\mathtt{C}_{bias}\log(n)),
$$

implies $\mathtt{r}m^2/\sqrt{n}\to 0$ such that

$$
\begin{aligned}
\left(\sum_{l=k}^{m}{\eta'_l}^2 k^4\right)^{1/2} &\le \left(\sum_{l=k}^{m}{\eta^*_{ml}}^2 k^4\right)^{1/2} + \left(\sum_{l=k}^{m}|\eta'_l-\eta^*_{ml}|^2 k^4\right)^{1/2} \\
&\le 2(1-\mathtt{C}_{\|\boldsymbol{\eta}^*\|}),
\end{aligned}
$$

which gives using (3.2)

$$
\begin{aligned}
&\mathbb{E}\left[\left\{\boldsymbol{f}'_{\boldsymbol{\eta}'}(\mathbf{X}^\top\boldsymbol{\theta}) - \boldsymbol{f}'_{\boldsymbol{\eta}'}(\mathbf{X}^\top\boldsymbol{\theta}')\right\}^2\right] \\
&\quad\le \|\boldsymbol{\theta}-\boldsymbol{\theta}'\|^2(1-\mathtt{C}_{\|\boldsymbol{\eta}^*\|})\mathtt{C}\|\psi''\|_\infty^2 s_{\mathbf{X}}^4 17^{5/2}4m\sum_{k=1}^{m}\eta'_k 2^{3j_k/2}.
\end{aligned}
$$

Repeating the same arguments gives

$$
\mathbb{E}\left[\left\{\boldsymbol{f}'_{\boldsymbol{\eta}'}(\mathbf{X}^\top\boldsymbol{\theta}) - \boldsymbol{f}'_{\boldsymbol{\eta}'}(\mathbf{X}^\top\boldsymbol{\theta}')\right\}^2\right] \le \|\boldsymbol{\theta}-\boldsymbol{\theta}'\|^2(1-\mathtt{C}_{\|\boldsymbol{\eta}^*\|})\mathtt{C}\|\psi''\|_\infty^2 s_{\mathbf{X}}^4 17^3 4m^{3/2},
$$

such that

$$\mathbb{E}\left[\left\{\left[\boldsymbol{f}'_{\boldsymbol{\eta}}(\mathbf{X}^\top\boldsymbol{\theta}) - \boldsymbol{f}'_{\boldsymbol{\eta}'}(\mathbf{X}^\top\boldsymbol{\theta}')\right](\mathbf{X}^\top\boldsymbol{\theta}^\circ)^2\right\}^2\right] \le \mathtt{C} m^2\|\boldsymbol{\upsilon} - \boldsymbol{\upsilon}'\|^2. \tag{A.23}$$

Finally we can estimate

$$\begin{aligned}
&\mathbb{E}\left[\left\{\boldsymbol{f}_{\boldsymbol{\eta}^\circ}(\mathbf{X}^\top\boldsymbol{\theta}) - \boldsymbol{f}_{\boldsymbol{\eta}^\circ}(\mathbf{X}^\top\boldsymbol{\theta}')\right\}^2\right]\\
&\quad = 2\sum_{k=1}^m\sum_{l=k}^m(1 - 1_{k=l}/2)\eta_k^\circ\eta_l^\circ\\
&\qquad\qquad \mathbb{E}\left[(\boldsymbol{e}_k(\mathbf{X}^\top\boldsymbol{\theta}) - \boldsymbol{e}_k(\mathbf{X}^\top\boldsymbol{\theta}'))(\boldsymbol{e}_l(\mathbf{X}^\top\boldsymbol{\theta}) - \boldsymbol{e}_l(\mathbf{X}^\top\boldsymbol{\theta}'))\right].
\end{aligned}$$

Using (A.5), that $\|\boldsymbol{\eta}^\circ\| \le 1/c_{\mathcal{D}}$ and very similar arguments as before we find

$$\mathbb{E}\left[\left\{\boldsymbol{f}_{\boldsymbol{\eta}^\circ}(\mathbf{X}^\top\boldsymbol{\theta}) - \boldsymbol{f}_{\boldsymbol{\eta}^\circ}(\mathbf{X}^\top\boldsymbol{\theta}')\right\}^2\right] \le \|\boldsymbol{\theta} - \boldsymbol{\theta}'\|^2\mathtt{C}\|\psi'\|_\infty^2 s_{\mathbf{X}}^4 17^{5/2} 4m^2\frac{1}{c_{\mathcal{D}}^2}. \tag{A.24}$$

Putting these bounds together gives

$$\begin{aligned}
&n\boldsymbol{\gamma}^\top\mathcal{D}_m^{-1}\mathbb{E}\left[(\varsigma_{1,m}(\boldsymbol{\upsilon}') - \varsigma_{1,m}(\boldsymbol{\upsilon}))(\varsigma_{1,m}(\boldsymbol{\upsilon}') - \varsigma_{1,m}(\boldsymbol{\upsilon}))^\top\right]\mathcal{D}_m^{-1}\boldsymbol{\gamma}^\top\\
&\quad \le \mathtt{C}^2_{\boldsymbol{(\mathcal{E}\mathfrak{D}_1)}} m^2\|\mathcal{D}_m(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}')\|^2\omega^2.
\end{aligned}$$

This yields $\boldsymbol{(\mathcal{E}\mathfrak{D}_1)}$ with

$$\nu_{1,m}^2 = \widetilde{\nu}^2\mathtt{C}^2_{\boldsymbol{(\mathcal{E}\mathfrak{D}_1)}} m^2.$$

□

A.8.6. Condition $\boldsymbol{(\mathcal{L}_0)}$

Lemma A.18. *The condition* $\boldsymbol{(\mathcal{L}_0)}$ *is satisfied where*

$$\delta(\mathtt{r}) = \frac{\mathtt{C}\left\{m^{3/2} + \mathtt{C}_{bias}m^{5/2}\right\}\mathtt{r}}{c_{\mathcal{D}}\sqrt{n}}.$$

Proof. We will show that $\frac{1}{n}\|\mathcal{D}_m^2(\boldsymbol{\upsilon}) - \mathcal{D}_m^2(\boldsymbol{\upsilon}_m^*)\| \le c_{\mathcal{D}}^2\delta(\mathtt{r})$, which will give the claim due to

$$\|I_{p^*} - \mathcal{D}_m^{-1}\nabla_{p^*}^2\mathbb{E}[\mathcal{L}(\boldsymbol{\upsilon})]\mathcal{D}_m^{-1}\| \le \frac{1}{nc_{\mathcal{D}}^2}\|\mathcal{D}_m^2(\boldsymbol{\upsilon}) - \mathcal{D}_m^2(\boldsymbol{\upsilon}_m^*)\|.$$

We represent

$$-\nabla_{p^*}^2\mathbb{E}[\mathcal{L}_m(\boldsymbol{\upsilon})] \overset{\text{def}}{=} \mathcal{D}_m^2(\boldsymbol{\upsilon}) = nd_m^2(\boldsymbol{\upsilon}) + nr_m^2(\boldsymbol{\upsilon}),$$

$$\begin{aligned}
nd_m^2(\boldsymbol{\upsilon}) &= n\left(\begin{array}{cc} d_{\boldsymbol{\theta}}^2(\boldsymbol{\upsilon}) & a_m(\boldsymbol{\upsilon}) \\ a_m^\top(\boldsymbol{\upsilon}) & h_m^2(\boldsymbol{\upsilon}) \end{array}\right) \overset{\text{def}}{=} \left(\begin{array}{cc} \mathrm{D}(\boldsymbol{\upsilon})^2 & A_m^\top(\boldsymbol{\upsilon}) \\ A_m(\boldsymbol{\upsilon}) & \mathrm{H}_m^2(\boldsymbol{\upsilon}) \end{array}\right), \\
r_m^2(\boldsymbol{\upsilon}) &= \mathbb{E}\left[\left(\boldsymbol{f}_{\boldsymbol{\eta}}(\mathbf{X}^\top\boldsymbol{\theta}) - g(\mathbf{X})\right)\left(\begin{array}{cc} v_{\boldsymbol{\theta}}^2(\boldsymbol{\upsilon}) & b_m(\boldsymbol{\upsilon}) \\ b_m^\top(\boldsymbol{\upsilon}) & 0 \end{array}\right)\right], \\
v_{\boldsymbol{\theta}}^2(\boldsymbol{\upsilon}) &= 2\boldsymbol{f}_{\boldsymbol{\eta}}''(\mathbf{X}^\top\boldsymbol{\theta})\nabla\Phi_{\boldsymbol{\theta}}^\top\mathbf{X}(\mathbf{X})^\top\nabla\Phi_{\boldsymbol{\theta}} \\
&\quad + |\boldsymbol{f}_{\boldsymbol{\eta}}'(\mathbf{X}^\top\boldsymbol{\theta})|^2\mathbf{X}^\top\nabla^2\varphi_{\boldsymbol{\theta}}^\top[\mathbf{X},\cdot,\cdot], \\
b_m(\boldsymbol{\upsilon}) &= \nabla\Phi_{\boldsymbol{\theta}}\mathbf{X}^\top\boldsymbol{e}'^\top(\mathbf{X}^\top\boldsymbol{\theta}),
\end{aligned}$$

such that

$$\begin{aligned}
\frac{1}{n}\|\mathcal{D}_m^2(\boldsymbol{\upsilon}) - \mathcal{D}_m^2(\boldsymbol{\upsilon}_m^*)\| \leq \frac{1}{n}\Big(&\|\mathrm{D}^2(\boldsymbol{\upsilon}) - \mathrm{D}^2(\boldsymbol{\upsilon}_m^*)\| + 2\|A_m(\boldsymbol{\upsilon}) - A_m(\boldsymbol{\upsilon}_m^*)\| \\
&+\|\mathrm{H}_m^2(\boldsymbol{\upsilon}) - \mathrm{H}_m^2(\boldsymbol{\upsilon}_m^*)\| + \|r_m^2(\boldsymbol{\upsilon}) - r_m^2(\boldsymbol{\upsilon}_m^*)\|\Big),
\end{aligned}$$

so that we can calculate separately

$$\begin{aligned}
\frac{1}{n}\|\mathrm{D}^2(\boldsymbol{\upsilon}) - \mathrm{D}^2(\boldsymbol{\upsilon}_m^*)\| \leq \mathbb{E}[\|\mathbf{X}\|^2\Big\{&|((\boldsymbol{f}_{\boldsymbol{\eta}}')^2 - (\boldsymbol{f}_{\boldsymbol{\eta}_m^*}')^2)(\mathbf{X}^\top\boldsymbol{\theta})| \\
&+|(\boldsymbol{f}_{\boldsymbol{\eta}_m^*}')^2(\mathbf{X}^\top\boldsymbol{\theta}) - (\boldsymbol{f}_{\boldsymbol{\eta}_m^*}')^2(\mathbf{X}^\top\boldsymbol{\theta}_m^*)| \\
&+2|(\boldsymbol{f}_{\boldsymbol{\eta}_m^*}')^2(\mathbf{X}^\top\boldsymbol{\theta}_m^*)|\|\nabla\Phi(\boldsymbol{\theta})^\top\mathbf{X} - \nabla\Phi(\boldsymbol{\theta}_m^*)^\top\mathbf{X}\|\Big\}].
\end{aligned}$$

Using Lemma A.10 we find

$$|(\boldsymbol{f}_{\boldsymbol{\eta}_m^*}')^2(\mathbf{X}^\top\boldsymbol{\theta}_m^*)|\|\nabla\Phi(\mathbf{X}^\top\boldsymbol{\theta}) - \nabla\Phi(\mathbf{X}^\top\boldsymbol{\theta}_m^*)\| \leq \|\psi'\|_\infty(\mathtt{C}+1)\sqrt{2}L_{\nabla\Phi}\|\boldsymbol{\theta} - \boldsymbol{\theta}_m^*\|.$$

Furthermore we have with $M(j) \subset \{1,\ldots,m\}$ in (A.10)

$$\begin{aligned}
\mathbb{E}|(\boldsymbol{f}_{\boldsymbol{\eta}}' - \boldsymbol{f}_{\boldsymbol{\eta}_m^*}')(\mathbf{X}^\top\boldsymbol{\theta})| &\leq \sum_{k=1}^{m}|\eta_k - \eta_{mk}^*|\mathbb{E}|\boldsymbol{e}_k'(\mathbf{X}^\top\boldsymbol{\theta})| \qquad \text{(A.25)} \\
&\leq \mathtt{C}_{p_{\mathbf{X}\boldsymbol{\theta}}}\|\psi'\|\|\boldsymbol{\eta} - \boldsymbol{\eta}_m^*\|\left(\sum_{k=1}^{j_m}2^{j_k}|M(j)|\right)^{1/2} \\
&\leq \mathtt{C}\|\boldsymbol{\eta} - \boldsymbol{\eta}_m^*\|\|\psi'\|_\infty m.
\end{aligned}$$

This implies using (A.14), (A.25) and (A.22)

$$\begin{aligned}
&\mathbb{E}\left[|((\boldsymbol{f}_{\boldsymbol{\eta}}')^2 - (\boldsymbol{f}_{\boldsymbol{\eta}_m^*}')^2)(\mathbf{X}^\top\boldsymbol{\theta})|\right] \\
&\quad\leq \mathbb{E}\left[|\boldsymbol{f}_{\boldsymbol{\eta}}'(\mathbf{X}^\top\boldsymbol{\theta})| + |\boldsymbol{f}_{\boldsymbol{\eta}_m^*}'(\mathbf{X}^\top\boldsymbol{\theta})|)|(\boldsymbol{f}_{\boldsymbol{\eta}}' - \boldsymbol{f}_{\boldsymbol{\eta}_m^*}')(\mathbf{X}^\top\boldsymbol{\theta})|\right] \\
&\quad\leq \mathbb{E}\left[(|(\boldsymbol{f}_{\boldsymbol{\eta}}' - \boldsymbol{f}_{\boldsymbol{\eta}_m^*}')(\mathbf{X}^\top\boldsymbol{\theta})| + 2|\boldsymbol{f}_{\boldsymbol{\eta}_m^*}'(\mathbf{X}^\top\boldsymbol{\theta})|)|(\boldsymbol{f}_{\boldsymbol{\eta}}' - \boldsymbol{f}_{\boldsymbol{\eta}_m^*}')(\mathbf{X}^\top\boldsymbol{\theta})|\right]
\end{aligned}$$

$$
\begin{aligned}
&\leq \mathbb{E}\left[|(\boldsymbol{f}'_{\boldsymbol{\eta}} - \boldsymbol{f}'_{\boldsymbol{\eta}^*_m})(\mathbf{X}^\top\boldsymbol{\theta})|^2\right] + 2\|\psi'\|_\infty(\mathtt{C}+1)\sqrt{2}\mathbb{E}[|(\boldsymbol{f}'_{\boldsymbol{\eta}} - \boldsymbol{f}'_{\boldsymbol{\eta}^*_m})(\mathbf{X}^\top\boldsymbol{\theta})|] \\
&\leq \|\psi'\|_\infty\left(2\|\psi'\|_\infty(\mathtt{C}+1)\sqrt{2} + \mathtt{C}\frac{\mathtt{r}m}{\sqrt{n}}\right) m\|\boldsymbol{\eta} - \boldsymbol{\eta}^*_m\| = \mathtt{C}m\|\boldsymbol{\eta} - \boldsymbol{\eta}^*_m\|,
\end{aligned}
$$

where we used $\frac{\mathtt{r}m}{\sqrt{n}} \to 0$ for $\mathtt{r}^2 \leq \mathtt{C}m$. Finally we derive with (A.16) and (A.14)

$$
\begin{aligned}
&\|(\boldsymbol{f}'_{\boldsymbol{\eta}^*_m})^2(\mathbf{X}^\top\boldsymbol{\theta}) - (\boldsymbol{f}'_{\boldsymbol{\eta}^*_m})^2(\mathbf{X}^\top\boldsymbol{\theta}^*_m)\| \\
&\quad\leq (\|\boldsymbol{f}'_{\boldsymbol{\eta}^*_m}(\mathbf{X}^\top\boldsymbol{\theta}^*_m)\| + \|\boldsymbol{f}'_{\boldsymbol{\eta}^*_m}(\mathbf{X}^\top\boldsymbol{\theta})\|)\|\boldsymbol{f}'_{\boldsymbol{\eta}^*_m}(\mathbf{X}^\top\boldsymbol{\theta}) - \boldsymbol{f}'_{\boldsymbol{\eta}^*_m}(\mathbf{X}^\top\boldsymbol{\theta}^*_m)\| \\
&\quad\leq 4\sqrt{2}\|\psi'\|_\infty(\mathtt{C}+1)^2\sqrt{m}\|\psi''\|_\infty s_{\mathbf{X}}\|\boldsymbol{\theta} - \boldsymbol{\theta}^*_m\|.
\end{aligned}
$$

Collecting everything yields with some constant $\mathtt{C} > 0$

$$
\frac{1}{n}\|\mathrm{D}^2(\boldsymbol{\upsilon}) - \mathrm{D}^2(\boldsymbol{\upsilon}^*_m)\| \leq \mathtt{C}m\|\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^*_m\|.
$$

Furthermore

$$
\begin{aligned}
&\frac{1}{n}\|\mathrm{H}^2_m(\boldsymbol{\upsilon}) - \mathrm{H}^2_m(\boldsymbol{\upsilon}^*_m)\| \\
&\quad= \sup_{\substack{\boldsymbol{\gamma}\in\mathbb{R}^m \\ \|\boldsymbol{\gamma}\|=1}} \sum_{k,l=1}^m \gamma_k\gamma_l\left(\mathbb{E}[\boldsymbol{e}_k\boldsymbol{e}_l(\mathbf{X}^\top\boldsymbol{\theta})] - \mathbb{E}[\boldsymbol{e}_k\boldsymbol{e}_l(\mathbf{X}^\top\boldsymbol{\theta}^*_m)]\right)1_{I_l\cap I_k\neq\emptyset} \\
&\quad\leq 2\sup_{\substack{\boldsymbol{\gamma}\in\mathbb{R}^m \\ \|\boldsymbol{\gamma}\|=1}} \sum_{k=1}^m\sum_{l=k}^m \gamma_k\gamma_l\mathbb{E}\left[\left(\boldsymbol{e}_k(\mathbf{X}^\top\boldsymbol{\theta}) - \boldsymbol{e}_k(\mathbf{X}^\top\boldsymbol{\theta}^*_m)\right)\boldsymbol{e}_l(\mathbf{X}^\top\boldsymbol{\theta})\right]1_{I_l\cap I_k\neq\emptyset} \\
&\qquad+2\sup_{\substack{\boldsymbol{\gamma}\in\mathbb{R}^m \\ \|\boldsymbol{\gamma}\|=1}} \sum_{k=1}^m\sum_{l=k}^m \gamma_k\gamma_l\mathbb{E}\left[\boldsymbol{e}_k(\mathbf{X}^\top\boldsymbol{\theta}^*_m)\left(\boldsymbol{e}_l(\mathbf{X}^\top\boldsymbol{\theta}) - \boldsymbol{e}_l(\mathbf{X}^\top\boldsymbol{\theta}^*_m)\right)\right]1_{I_l\cap I_k\neq\emptyset}.
\end{aligned}
$$

Using (A.7) and (3.2) this gives

$$
\begin{aligned}
&\sum_{k=1}^m\sum_{l=k}^m \gamma_k\gamma_l\mathbb{E}\left[\left(\boldsymbol{e}_k(\mathbf{X}^\top\boldsymbol{\theta}) - \boldsymbol{e}_k(\mathbf{X}^\top\boldsymbol{\theta}^*_m)\right)\boldsymbol{e}_l(\mathbf{X}^\top\boldsymbol{\theta})\right]1_{I_l\cap I_k\neq\emptyset} \\
&\quad\leq \|\boldsymbol{\theta} - \boldsymbol{\theta}^*_m\|\|\psi'\|s^2_{\mathbf{X}}17\mathtt{C}\sum_{k=1}^m\gamma_k 2^{j_k}\sum_{l=k}^m\gamma_l 1_{I_l\cap I_k\neq\emptyset} \\
&\quad\leq \|\boldsymbol{\theta} - \boldsymbol{\theta}^*_m\|\|\psi'\|s^2_{\mathbf{X}}17^{3/2}\mathtt{C}m,
\end{aligned}
$$

and with similar arguments

$$\sum_{k=1}^{m}\sum_{l=k}^{m}\gamma_k\gamma_l\mathbb{E}\left[\left(\boldsymbol{e}_l(\mathbf{X}^\top\boldsymbol{\theta})-\boldsymbol{e}_l(\mathbf{X}^\top\boldsymbol{\theta}_m^*)\right)\boldsymbol{e}_k(\mathbf{X}^\top\boldsymbol{\theta})\right]1_{I_l\cap I_k\neq\emptyset}$$
$$\leq \|\boldsymbol{\theta}-\boldsymbol{\theta}_m^*\|\|\psi'\|s_{\mathbf{X}}^2 17^{3/2}(\mathtt{C}+\|\psi\|_\infty)m.$$

Consequently with some constant $\mathtt{C}\in\mathbb{R}$

$$\frac{1}{n}\|\mathrm{H}_m^2(\boldsymbol{\upsilon})-\mathrm{H}_m^2(\boldsymbol{\upsilon}_m^*)\|\leq\frac{\mathtt{C}m}{\sqrt{n}c_{\mathcal{D}}}\|\mathcal{D}(\boldsymbol{\upsilon}-\boldsymbol{\upsilon}_m^*)\|. \tag{A.26}$$

Again with some constant $\mathtt{C}>0$

$$\frac{1}{n}\|A_m(\boldsymbol{\upsilon})-A_m(\boldsymbol{\upsilon}_m^*)\|\leq\mathtt{C}\Bigg(\mathbb{E}\Big[\Big\|\boldsymbol{f}'_{\boldsymbol{\eta}}(\mathbf{X}_1^\top\boldsymbol{\theta})-\boldsymbol{f}'_{\boldsymbol{\eta}_m^*}(\mathbf{X}_1^\top\boldsymbol{\theta}_m^*)\Big\|^2\Big]^{1/2}$$
$$+\mathbb{E}\Big[\|\nabla\varPhi(\boldsymbol{\theta})-\nabla\varPhi(\boldsymbol{\theta}_m^*)\|^2\Big]^{1/2}$$
$$+\mathbb{E}\Big[\|\boldsymbol{e}(\mathbf{X}_1^\top\boldsymbol{\theta})-\boldsymbol{e}(\mathbf{X}_1^\top\boldsymbol{\theta}_m^*)\|^2\Big]^{1/2}\Bigg).$$

Note that using (A.24)

$$\mathbb{E}\Big[\|\boldsymbol{e}(\mathbf{X}_1^\top\boldsymbol{\theta})-\boldsymbol{e}(\mathbf{X}_1^\top\boldsymbol{\theta}_m^*)\|^2\Big]$$
$$\leq\sup_{\substack{\boldsymbol{\eta}^\circ\in\mathbb{R}^m\\ \|\boldsymbol{\eta}^\circ\|=1}}\mathbb{E}\left[\left\{\boldsymbol{f}_{\boldsymbol{\eta}^\circ}(\mathbf{X}^\top\boldsymbol{\theta})-\boldsymbol{f}_{\boldsymbol{\eta}^\circ}(\mathbf{X}^\top\boldsymbol{\theta}_m^*)\right\}^2\right]$$
$$\leq\|\boldsymbol{\theta}-\boldsymbol{\theta}_m^*\|^2\mathtt{C}\|\psi'\|_\infty^2 s_{\mathbf{X}}^4 17^{5/2}4m^2.$$

Using (A.23) this yields

$$\frac{1}{n}\|A_m(\boldsymbol{\upsilon})-A_m(\boldsymbol{\upsilon}_m^*)\|\leq\mathtt{C}m\|\boldsymbol{\upsilon}-\boldsymbol{\upsilon}'\|.$$

Finally we estimate the fourth term.

$$\|r_m^2(\boldsymbol{\upsilon})-r_m^2(\boldsymbol{\upsilon}_m^*)\|\leq\mathbb{E}[|\boldsymbol{f}_{\boldsymbol{\eta}}(\mathbf{X}^\top\boldsymbol{\theta})-\boldsymbol{f}_{\boldsymbol{\eta}_m^*}(\mathbf{X}^\top\boldsymbol{\theta}_m^*)|\|\widetilde{\mathcal{V}}_m^2(\boldsymbol{\upsilon})\|] \tag{A.27}$$
$$+\mathbb{E}[|\boldsymbol{f}_{\boldsymbol{\eta}_m^*}(\mathbf{X}^\top\boldsymbol{\theta}_m^*)-g(\mathbf{X})|\|\widetilde{\mathcal{V}}_m^2(\boldsymbol{\upsilon}_m^*)-\widetilde{\mathcal{V}}_m^2(\boldsymbol{\upsilon})\|].$$

We estimate separately

$$\mathbb{E}[|\boldsymbol{f}_{\boldsymbol{\eta}}(\mathbf{X}^\top\boldsymbol{\theta})-\boldsymbol{f}_{\boldsymbol{\eta}_m^*}(\mathbf{X}^\top\boldsymbol{\theta}_m^*)|\|(\widetilde{\mathcal{V}}_m^2)(\boldsymbol{\upsilon})\|]$$
$$\leq\mathbb{E}[|\boldsymbol{f}_{\boldsymbol{\eta}}(\mathbf{X}^\top\boldsymbol{\theta})-\boldsymbol{f}_{\boldsymbol{\eta}_m^*}(\mathbf{X}^\top\boldsymbol{\theta}_m^*)|\|v_{\boldsymbol{\theta}}^2(\boldsymbol{\upsilon})\|]$$
$$+\mathbb{E}[|\boldsymbol{f}_{\boldsymbol{\eta}}(\mathbf{X}^\top\boldsymbol{\theta})-\boldsymbol{f}_{\boldsymbol{\eta}_m^*}(\mathbf{X}^\top\boldsymbol{\theta}_m^*)|\|b_m(\boldsymbol{\upsilon})\|]$$

To bound the first term, first note that again using the wavelet structure

$$\begin{aligned}
|\boldsymbol{f}''_{\boldsymbol{\eta}}(\mathbf{X}^\top\boldsymbol{\theta})| &\le |\boldsymbol{f}''_{\boldsymbol{\eta}-\boldsymbol{\eta}^*_m}(\mathbf{X}^\top\boldsymbol{\theta})| + |\boldsymbol{f}''_{\boldsymbol{\eta}^*_m}(\mathbf{X}^\top\boldsymbol{\theta})| \\
&\le \sqrt{34}\|\psi''\|_\infty \left(\sum_{k=0}^{m}(\eta_k - \eta^*_{mk})^2\right)^{1/2} \left(\sum_{j=0}^{j_m-1} 2^{5j}\right)^{1/2} + \mathtt{C}_{\|\boldsymbol{f}''_{\boldsymbol{\eta}^*_m}\|_\infty} \\
&\le \sqrt{34}\|\psi''\|_\infty \|\mathcal{D}_m(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^*_m)\| \frac{m^{5/2}}{c_{\mathcal{D}}\sqrt{n}} + \mathtt{C}_{\|\boldsymbol{f}''_{\boldsymbol{\eta}^*_m}\|_\infty},
\end{aligned}$$

which can be treated as a constant as $m^5/n \to 0$. Furthermore using (A.14) we have for any $\varphi \in \mathbb{R}^{p-1}$ with $\|\varphi\| = 1$

$$\||\boldsymbol{f}'_{\boldsymbol{\eta}}(\mathbf{X}^\top\boldsymbol{\theta})|^2 \nabla^2 \Phi^\top_{\boldsymbol{\theta}}[\mathbf{X}, \varphi, \cdot]\|_{\mathbb{R}^p} \le 34\|\psi'\|^2_\infty C^2_{\|\boldsymbol{\eta}^*_m\|} s^2_{\mathbf{X}} \|\nabla^2 \Phi_{\boldsymbol{\theta}^*_m}\|_\infty.$$

To control $\mathbb{E}\|b_m(\boldsymbol{\upsilon})\|^2$ we use (A.21) to bound

$$\begin{aligned}
\mathbb{E}\|b_m(\boldsymbol{\upsilon})\|^2 &\le s^2_{\mathbf{X}} \sum_{k=1}^{m} \mathbb{E}\boldsymbol{e}'_k(\mathbf{X}^\top\boldsymbol{\theta})^2 \\
&\le s^2_{\mathbf{X}} 17^2 \mathtt{C}^2 \|\psi'\|^2_\infty \sum_{k=1}^{m} 2^{2j_k} \\
&\le s^2_{\mathbf{X}} 17^2 \mathtt{C}^2 \|\psi'\|^2_\infty m^3. \qquad \text{(A.28)}
\end{aligned}$$

This implies for the first summand in (A.27) with constants $\mathtt{C}, \mathtt{C}' > 0$ large enough

$$\begin{aligned}
&\mathbb{E}[|\boldsymbol{f}_{\boldsymbol{\eta}}(\mathbf{X}^\top\boldsymbol{\theta}) - \boldsymbol{f}_{\boldsymbol{\eta}^*_m}(\mathbf{X}^\top\boldsymbol{\theta}^*_m)|\|\widetilde{\mathcal{V}}^2_m(\boldsymbol{\upsilon})\|] \\
&\quad\le \Big(\mathbb{E}[|\boldsymbol{f}_{\boldsymbol{\eta}}(\mathbf{X}^\top\boldsymbol{\theta}) - \boldsymbol{f}_{\boldsymbol{\eta}^*_m}(\mathbf{X}^\top\boldsymbol{\theta})|^2]^{1/2} \\
&\qquad + \mathbb{E}[|\boldsymbol{f}_{\boldsymbol{\eta}^*_m}(\mathbf{X}^\top\boldsymbol{\theta}) - \boldsymbol{f}_{\boldsymbol{\eta}^*_m}(\mathbf{X}^\top\boldsymbol{\theta}^*_m)|^2]^{1/2}\Big) \mathtt{C}m^{3/2} \\
&\quad\le \mathtt{C}m^{3/2}\|\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^*_m\| + \mathtt{C}m^{3/2}\mathbb{E}[|\boldsymbol{f}_{\boldsymbol{\eta}}(\mathbf{X}^\top\boldsymbol{\theta}) - \boldsymbol{f}_{\boldsymbol{\eta}^*_m}(\mathbf{X}^\top\boldsymbol{\theta})|^2]^{1/2}.
\end{aligned}$$

We estimate using (A.26), $\mathtt{r}m^{3/2}/\sqrt{n} \to 0$ for $\mathtt{r} \le \mathtt{r}_0$ and constants $\mathtt{C}, \mathtt{C}' > 0$ large enough

$$\begin{aligned}
&\mathbb{E}[|\boldsymbol{f}_{\boldsymbol{\eta}}(\mathbf{X}^\top\boldsymbol{\theta}) - \boldsymbol{f}_{\boldsymbol{\eta}^*_m}(\mathbf{X}^\top\boldsymbol{\theta})|^2]^{1/2} = \frac{1}{\sqrt{n}}\|\mathrm{H}_m(\boldsymbol{\upsilon})(\boldsymbol{\eta} - \boldsymbol{\eta}^*_m)\| \\
&\quad\le \frac{1}{\sqrt{n}}\|\mathrm{H}^2_m(\boldsymbol{\upsilon}) - \mathrm{H}^2_m(\boldsymbol{\upsilon}^*_m)\|^{1/2}\|(\boldsymbol{\eta} - \boldsymbol{\eta}^*_m)\| + \frac{1}{\sqrt{n}}\|\mathrm{H}_m(\boldsymbol{\upsilon}^*_m)(\boldsymbol{\eta} - \boldsymbol{\eta}^*_m)\| \\
&\quad\le \left(\frac{1}{\sqrt{n}}\|\mathrm{H}^2_m(\boldsymbol{\upsilon}) - \mathrm{H}^2_m(\boldsymbol{\upsilon}^*_m)\|^{1/2}\frac{1}{\sqrt{n}c_{\mathcal{D}}} + \frac{1}{\sqrt{n}}\right)\|\mathrm{H}_m(\boldsymbol{\upsilon}^*_m)(\boldsymbol{\eta} - \boldsymbol{\eta}^*_m)\|
\end{aligned}$$

$$
\begin{aligned}
&\le \left\{ \left(\mathtt{r} m^{3/2}/\sqrt{n} \right)^{1/2} + 1 \right\} \frac{\mathtt{C}}{\sqrt{n} c_{\mathcal{D}}} \|\mathcal{D}_m(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^*_m)\| \\
&\le \frac{\mathtt{C}'}{\sqrt{n} c_{\mathcal{D}}} \|\mathcal{D}_m(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^*_m)\|.
\end{aligned}
$$

We also find

$$
\begin{aligned}
&|\boldsymbol{f}_{\boldsymbol{\eta}^*_m}(\mathbf{X}^\top \boldsymbol{\theta}) - \boldsymbol{f}_{\boldsymbol{\eta}^*_m}(\mathbf{X}^\top \boldsymbol{\theta}^*_m)| \\
&\quad \le \left(\sum_{k=1}^m (\boldsymbol{\eta}^*_m)_k^2 k^{2\alpha} \right)^{1/2} \left(\sum_{k=1}^m |\boldsymbol{e}'_k(\mathbf{X}^\top \boldsymbol{\theta}^*_m)|^2 k^{-2\alpha} \right)^{1/2} L_{\nabla \Phi} \|\mathbf{X}\| \|\boldsymbol{\theta} - \boldsymbol{\theta}^*_m\| \\
&\quad \le 2\sqrt{34} C_{\|\boldsymbol{\eta}^*_m\|} \sqrt{2} L_{\nabla \Phi} s_{\mathbf{X}} \|\psi'\|_\infty \|\boldsymbol{\theta} - \boldsymbol{\theta}^*_m\|.
\end{aligned}
$$

Consequently

$$
\mathbb{E}[|\boldsymbol{f}_{\boldsymbol{\eta}}(\mathbf{X}^\top \boldsymbol{\theta}) - \boldsymbol{f}_{\boldsymbol{\eta}^*_m}(\mathbf{X}^\top \boldsymbol{\theta}^*_m)| \|\widetilde{\mathcal{V}}^2_m(\boldsymbol{\upsilon})\|] \le \frac{\mathtt{C} m^{3/2}}{\sqrt{n} c_{\mathcal{D}}} \|\mathcal{D}_m(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^*_m)\|.
$$

Furthermore using that $|\boldsymbol{f}_{\boldsymbol{\eta}^*_m}(\mathbf{X}^\top \boldsymbol{\theta}^*_m) - g(\mathbf{X})| \le \mathtt{C}_{bias}$

$$
\begin{aligned}
&\mathbb{E}[|\boldsymbol{f}_{\boldsymbol{\eta}^*_m}(\mathbf{X}^\top \boldsymbol{\theta}^*_m) - g(\mathbf{X})| \|\widetilde{\mathcal{V}}^2_m(\boldsymbol{\upsilon}^*_m) - \widetilde{\mathcal{V}}^2_m(\boldsymbol{\upsilon})\|] \\
&\quad \le \mathtt{C}_{bias} \left(\mathbb{E}[\|v^2_{\boldsymbol{\theta}}(\boldsymbol{\upsilon}^*_m) - v^2_{\boldsymbol{\theta}}(\boldsymbol{\upsilon})\|] + 2\mathbb{E}[\|b_m(\boldsymbol{\upsilon}^*_m) - b_m(\boldsymbol{\upsilon})\|] \right).
\end{aligned}
$$

For this we estimate with some constants $\mathtt{C}_i$ that only depend on $\|\nabla^2 \Phi_{\boldsymbol{\theta}^*_m}\|, s_{\mathbf{X}}$, $\mathtt{C}_{\|\boldsymbol{f}'_{\boldsymbol{\eta}^*_m}\|_\infty}, \mathtt{C}_{\|\boldsymbol{f}''_{\boldsymbol{\eta}^*_m}\|_\infty}$, etc.

$$
\begin{aligned}
&\|v^2_{\boldsymbol{\theta}}(\boldsymbol{\upsilon}^*_m) - v^2_{\boldsymbol{\theta}}(\boldsymbol{\upsilon})\| \\
&\quad \le \|2\boldsymbol{f}''_{\boldsymbol{\eta}}(\mathbf{X}^\top \boldsymbol{\theta}) \nabla \Phi_{\boldsymbol{\theta}}^\top \mathbf{X}(\mathbf{X})^\top \nabla \Phi_{\boldsymbol{\theta}} - 2\boldsymbol{f}''_{\boldsymbol{\eta}^*_m}(\mathbf{X}^\top \boldsymbol{\theta}^*_m) \nabla \Phi_{\boldsymbol{\theta}^*_m}^\top \mathbf{X}(\mathbf{X})^\top \nabla \Phi_{\boldsymbol{\theta}^*_m}\| \\
&\qquad + \| |\boldsymbol{f}'_{\boldsymbol{\eta}^*_m}(\mathbf{X}^\top \boldsymbol{\theta}^*_m)|^2 \mathbf{X}^\top \nabla^2 \Phi_{\boldsymbol{\theta}^*_m}^\top [\mathbf{X}, \cdot, \cdot] - |\boldsymbol{f}'_{\boldsymbol{\eta}}(\mathbf{X}^\top \boldsymbol{\theta})|^2 \mathbf{X}^\top \nabla^2 \varphi_{\boldsymbol{\theta}}^\top [\mathbf{X}, \cdot, \cdot]\| \\
&\quad \le \mathtt{C}_1 \left| |\boldsymbol{f}'_{\boldsymbol{\eta}^*_m}(\mathbf{X}^\top \boldsymbol{\theta}^*_m)|^2 - |\boldsymbol{f}'_{\boldsymbol{\eta}}(\mathbf{X}^\top \boldsymbol{\theta})|^2 \right| + \mathtt{C}_2 |\boldsymbol{f}''_{\boldsymbol{\eta}^*_m}(\mathbf{X}^\top \boldsymbol{\theta}^*_m) - \boldsymbol{f}''_{\boldsymbol{\eta}}(\mathbf{X}^\top \boldsymbol{\theta})| \\
&\qquad + \mathtt{C}_3 \|\boldsymbol{\theta} - \boldsymbol{\theta}^*_m\|.
\end{aligned}
$$

With the same arguments as those used for the bound of $\frac{1}{n} \|\mathrm{D}^2(\boldsymbol{\upsilon}) - \mathrm{D}^2(\boldsymbol{\upsilon}^*_m)\|$

$$
\mathbb{E} \left| |\boldsymbol{f}'_{\boldsymbol{\eta}^*_m}(\mathbf{X}^\top \boldsymbol{\theta}^*_m)|^2 - |\boldsymbol{f}'_{\boldsymbol{\eta}}(\mathbf{X}^\top \boldsymbol{\theta})|^2 \right| \le \mathtt{C} m \|\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^*_m\|.
$$

Furthermore

$$
\begin{aligned}
|\boldsymbol{f}''_{\boldsymbol{\eta}^*_m}(\mathbf{X}^\top \boldsymbol{\theta}^*_m) - \boldsymbol{f}''_{\boldsymbol{\eta}}(\mathbf{X}^\top \boldsymbol{\theta})| &\le |\boldsymbol{f}''_{\boldsymbol{\eta}^*_m}(\mathbf{X}^\top \boldsymbol{\theta}^*_m) - \boldsymbol{f}''_{\boldsymbol{\eta}^*_m}(\mathbf{X}^\top \boldsymbol{\theta})| \\
&\quad + |\boldsymbol{f}''_{\boldsymbol{\eta}^*_m}(\mathbf{X}^\top \boldsymbol{\theta}) - \boldsymbol{f}''_{\boldsymbol{\eta}}(\mathbf{X}^\top \boldsymbol{\theta})|
\end{aligned}
$$

Using (A.15) we estimate

$$|\boldsymbol{f}''_{\boldsymbol{\eta}^*_m}(\mathbf{X}^\top\boldsymbol{\theta}^*_m) - \boldsymbol{f}''_{\boldsymbol{\eta}^*_m}(\mathbf{X}^\top\boldsymbol{\theta})| \le \mathtt{C} m^{3/2}\|\boldsymbol{\theta} - \boldsymbol{\theta}^*_m\|,$$

and

$$|\boldsymbol{f}''_{\boldsymbol{\eta}^*_m}(\mathbf{X}^\top\boldsymbol{\theta}) - \boldsymbol{f}''_{\boldsymbol{\eta}}(\mathbf{X}^\top\boldsymbol{\theta})| \le \sqrt{17}\|\boldsymbol{\eta} - \boldsymbol{\eta}^*\| \left(\sum_{j=1}^{j_m} 2^{5j)}\right)^{1/2} \le \mathtt{C} m^{5/2}\|\boldsymbol{\theta} - \boldsymbol{\theta}^*_m\|.$$

Furthermore

$$\begin{aligned}\mathbb{E}[\|b_m(\boldsymbol{\upsilon}^*_m) - b_m(\boldsymbol{\upsilon})\|] \le\ & \mathtt{C}\mathbb{E}\|\boldsymbol{e}'(\mathbf{X}^\top\boldsymbol{\theta}) - \boldsymbol{e}'(\mathbf{X}^\top\boldsymbol{\theta}^*_m)\| \\ & + \mathtt{C}\mathbb{E}[\|\boldsymbol{e}'(\mathbf{X}^\top\boldsymbol{\theta}^*_m)\|^2]^{1/2}\|\boldsymbol{\theta} - \boldsymbol{\theta}^*_m\|.\end{aligned}$$

By (A.28) we have

$$\mathbb{E}[\|\boldsymbol{e}'(\mathbf{X}^\top\boldsymbol{\theta}^*_m)\|^2]^{1/2} \le 17\mathtt{C}\|\psi'\|_\infty m^{3/2}.$$

Furthermore

$$\begin{aligned}\mathbb{E}\|\boldsymbol{e}'(\mathbf{X}^\top\boldsymbol{\theta}) - \boldsymbol{e}'(\mathbf{X}^\top\boldsymbol{\theta}^*_m)\| &\le \mathbb{E}\left[\|\boldsymbol{e}'(\mathbf{X}^\top\boldsymbol{\theta}) - \boldsymbol{e}'(\mathbf{X}^\top\boldsymbol{\theta}^*_m)\|^2\right]^{1/2} \\ &= \left(\sum_{k=1}^m \mathbb{E}\left[(\boldsymbol{e}'_k(\mathbf{X}^\top\boldsymbol{\theta}) - \boldsymbol{e}'_k(\mathbf{X}^\top\boldsymbol{\theta}^*_m))^2\right]\right)^{1/2}.\end{aligned}$$

With (A.6) we find

$$\begin{aligned}\mathbb{E}\|\boldsymbol{e}'(\mathbf{X}^\top\boldsymbol{\theta}) - \boldsymbol{e}'(\mathbf{X}^\top\boldsymbol{\theta}^*_m)\| &\le \|\psi''\|_\infty s^2_{\mathbf{X}} 17\|\boldsymbol{\theta} - \boldsymbol{\theta}'\| \left(\sum_{k=1}^m 2^{4j_k}\right)^{1/2} \\ &\le |\psi''\|_\infty s^2_{\mathbf{X}} 17\|\boldsymbol{\theta} - \boldsymbol{\theta}'\| m^{5/2}.\end{aligned}$$

Together this gives

$$\begin{aligned}&\mathbb{E}[|\boldsymbol{f}_{\boldsymbol{\eta}^*_m}(\mathbf{X}^\top\boldsymbol{\theta}^*_m) - g(\mathbf{X})|\|\widetilde{\mathcal{V}}^2_m(\boldsymbol{\upsilon}^*_m) - \widetilde{\mathcal{V}}^2_m(\boldsymbol{\upsilon})\|] \\ &\le \mathtt{C} m^{3/2}\|\boldsymbol{\upsilon} - \mathit{\Pi}_{p^*}\boldsymbol{\upsilon}^*\| + \mathtt{C}_{bias}\mathtt{C} m^{5/2}.\end{aligned}$$

Collecting everything we find

$$\frac{1}{n}\|\mathcal{D}^2_m(\boldsymbol{\upsilon}) - \mathcal{D}^2_m(\boldsymbol{\upsilon}^*_m)\| \le \frac{\mathtt{C}}{\sqrt{n} c_{\mathcal{D}}}\left\{m^{3/2} + \mathtt{C}_{bias} m^{5/2}\right\}\|\mathcal{D}_m(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^*_m)\|.$$

Such that

$$\delta(\mathtt{r}) = \frac{\mathtt{C}_{(\mathcal{L}_0)}\left\{m^{3/2} + \mathtt{C}_{bias} m^{5/2}\right\}\mathtt{r}}{c_{\mathcal{D}}\sqrt{n}}. \qquad \square$$

A.8.7. Condition ($\mathcal{L}\mathtt{r}$)

Before we start with the actual proof we cite the following important result that will be used in our arguments.

The next result is a variant of Theorem 4.3 of [15] and is the key tool of this subsection.

Theorem A.19. *Let for a sequence of independent* $\mathbf{X}_i \in \mathcal{X}$ *for some space* $\mathcal{X}$

$$F(\boldsymbol{\upsilon}) = \sum_{i=1}^{n} f_i(\boldsymbol{\upsilon}, \mathbf{X}_i) - e, \, \boldsymbol{\upsilon} \in \varUpsilon \subset \mathbb{R}^{p^*}$$

and assume that with $\mathtt{r} > \mathtt{r}_{\boldsymbol{Q}} > 0$, $\varUpsilon_\circ(\mathtt{r}) \subset \varUpsilon$ *and* $\chi_{\mathtt{b}} : [0, 2\mathtt{b}] \to \mathbb{R}$ *defined in* (A.29)

$$\mathbb{E}\left[\sup_{\boldsymbol{\upsilon} \in \varUpsilon_\circ(\mathtt{r})^c} (P_n - \mathbb{P})\chi_{\mathtt{b}}(\boldsymbol{\upsilon})\right] \le C_\chi, \qquad \mathbb{P}(e > C_e) \le \tau_e,$$

$$\boldsymbol{Q}(\mathtt{b}) \overset{\text{def}}{=} \inf_{\boldsymbol{\upsilon} \in \varUpsilon_\circ(\mathtt{r})^c} \mathbb{P}\left(f_i(\boldsymbol{\upsilon}, \mathbf{X}_i) \ge \mathtt{b}\mathtt{r}^2/n\right) > 0.$$

Choose

$$0 < \lambda \le \left(\boldsymbol{Q}(2\mathtt{b}) - 2/n - 2C_\chi\right)/4.$$

Then for $\mathtt{r}^2 \ge C_e/(\lambda \mathtt{b}) \vee \mathtt{r}_{\boldsymbol{Q}}^2$

$$\mathbb{P}\left(\inf_{\boldsymbol{\upsilon} \in \varUpsilon_\circ(\mathtt{r})^c} F(\boldsymbol{\upsilon}) \le \lambda \mathtt{b}\mathtt{r}^2\right) \le \exp\left\{-n\boldsymbol{Q}(2\mathtt{b})^2/4\right\} + \tau_e$$

The auxiliary function is defined as

$$\overline{\chi}_u(t) = \begin{cases} 0 & t \le u; \\ t/u - 1 & t \in [u, 2u]; \\ 1 & t \ge 2u; \end{cases} \quad \chi_{\mathtt{b}}(\boldsymbol{\upsilon})_i \overset{\text{def}}{=} \overline{\chi}_{\mathtt{b}}(f_i(\boldsymbol{\upsilon})). \tag{A.29}$$

Remark A.7. The proof is nearly the same as that of Theorem 4.3 of [15]. The set $\varUpsilon_\circ(\mathtt{r})^c \subset \mathbb{R}^{p^*}$ is neither star shaped, nor convex but one can still use the same arguments.

Now we can start with the verification of ($\mathcal{L}\mathtt{r}$). We point out that in this Section we will distinguish $\boldsymbol{\theta} \in S_1^{p,+}$ and $\varphi_{\boldsymbol{\theta}} \in W_S$ with $\varPhi(\varphi_{\boldsymbol{\theta}}) = \boldsymbol{\theta}$ from each other. The result is summarized in the following Lemma:

Lemma A.20. *Assume the conditions $(\mathcal{A})$. Then for $n \in \mathbb{N}$ large enough there exist $c_{(\boldsymbol{Q})}, c_{(\mathcal{L}\mathtt{r})}, \mathtt{C} > 0$ such that with probability $1 - \exp\left\{-m^3\mathtt{x}\right\} - \exp\left\{-nc_{(\boldsymbol{Q})}/4\right\}$*

$$-\inf_{\boldsymbol{\upsilon}\in\varUpsilon_\circ(\mathtt{r})^c} \mathbb{E}[\mathcal{L}(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}_m^*)] > c_{(\mathcal{L}\mathtt{r})}\mathtt{r}^2/2,$$

as soon as $\mathtt{r}^2 \geq \mathtt{C}(m + \mathtt{x})$.

Proof. We will proof this claim using Theorem A.19. First note that we have with expectation taken conditioned on $(\mathbf{X}) = (\mathbf{X}_i)_{i=1,\ldots,n} \subset \mathbb{R}^p$ and using (1.9)

$$\begin{aligned}
&-\mathbb{E}_\varepsilon[\mathcal{L}(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}_m^*)] = -\mathbb{E}[\mathcal{L}(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}_m^*)|(\mathbf{X})] \\
&\quad = \sum_{i=1}^n \left[|\boldsymbol{f}_{\boldsymbol{\eta}}(\mathbf{X}_i^\top\boldsymbol{\theta}) - \boldsymbol{f}_{\boldsymbol{\eta}^*}(\mathbf{X}^\top\boldsymbol{\theta}^*)|^2 - |\boldsymbol{f}_{\boldsymbol{\eta}_m^*}(\mathbf{X}_i^\top\boldsymbol{\theta}_m^*) - \boldsymbol{f}_{\boldsymbol{\eta}^*}(\mathbf{X}^\top\boldsymbol{\theta}^*)|^2\right] \\
&\quad \geq \sum_{i=1}^n \left[|\boldsymbol{f}_{\boldsymbol{\eta}}(\mathbf{X}_i^\top\boldsymbol{\theta}) - \boldsymbol{f}_{\boldsymbol{\eta}^*}(\mathbf{X}_i^\top\boldsymbol{\theta}^*)|^2\right] - n\mathbb{E}[|\boldsymbol{f}_{\boldsymbol{\eta}_m^*}(\mathbf{X}_i^\top\boldsymbol{\theta}_m^*) - \boldsymbol{f}_{\boldsymbol{\eta}^*}(\mathbf{X}_i^\top\boldsymbol{\theta}^*)|^2] \\
&\qquad -n\left|(\mathrm{P}_n - \mathbb{P})|\boldsymbol{f}_{\boldsymbol{\eta}_m^*}(\mathbf{X}^\top\boldsymbol{\theta}_m^*) - \boldsymbol{f}_{\boldsymbol{\eta}^*}(\mathbf{X}_i^\top\boldsymbol{\theta}^*)|^2\right|.
\end{aligned}$$

We define

$$\begin{aligned}
e \stackrel{\text{def}}{=} & n\mathbb{E}[|\boldsymbol{f}_{\boldsymbol{\eta}_m^*}(\mathbf{X}_i^\top\boldsymbol{\theta}_m^*) - \boldsymbol{f}_{\boldsymbol{\eta}^*}(\mathbf{X}^\top\boldsymbol{\theta}^*)|^2] \\
& +n\left|(\mathrm{P}_n - \mathbb{P})|\boldsymbol{f}_{\boldsymbol{\eta}_m^*}(\mathbf{X}^\top\boldsymbol{\theta}_m^*) - \boldsymbol{f}_{\boldsymbol{\eta}^*}(\mathbf{X}^\top\boldsymbol{\theta}^*)|^2\right|,
\end{aligned}$$

such that

$$-\mathbb{E}[\mathcal{L}(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}_m^*)|(\mathbf{X})] \geq \sum_{i=1}^n \left(\boldsymbol{f}_{\boldsymbol{\eta}}(\mathbf{X}_i^\top\boldsymbol{\theta}) - \boldsymbol{f}_{\boldsymbol{\eta}^*}(\mathbf{X}_i^\top\boldsymbol{\theta}^*)\right)^2 - e,$$

This hints that Theorem A.19 gives the desired result. Consider the following list of assumptions:

(1) With some $\mathtt{C} > 0$

$$n\mathbb{E}[|\boldsymbol{f}_{\boldsymbol{\eta}_m^*}(\mathbf{X}^\top\boldsymbol{\theta}_m^*) - \boldsymbol{f}_{\boldsymbol{\eta}^*}(\mathbf{X}^\top\boldsymbol{\theta}^*)|^2] \leq 3(2 + \mathtt{C})\mathtt{r}^{*2},$$

(2) With probability $1 - \exp\left\{-m^3\mathtt{x}\right\}$ and a constant $\mathtt{C}_{\sum} > 0$

$$n\left|(\mathrm{P}_n - \mathbb{P})|\boldsymbol{f}_{\boldsymbol{\eta}_m^*}(\mathbf{X}^\top\boldsymbol{\theta}_m^*) - \boldsymbol{f}_{\boldsymbol{\eta}^*}(\mathbf{X}^\top\boldsymbol{\theta}^*)|^2]\right| \leq \mathtt{C}_{\sum},$$

(3) For some $\mathtt{b} > 0$ and for $n \in \mathbb{N}$ large enough and $\mathtt{r} > \sqrt{m}$

$$\begin{aligned}&\boldsymbol{Q}(2\mathtt{b})\\ &\quad \stackrel{\text{def}}{=} \inf_{(\boldsymbol{\theta},\boldsymbol{\eta})\in\Upsilon_\circ(\mathtt{r})^c} \mathbb{P}\left[\left(\boldsymbol{f}_{\boldsymbol{\eta}}(\mathbf{X}_i^\top\boldsymbol{\theta}) - \boldsymbol{f}_{\boldsymbol{\eta}^*}(\mathbf{X}_i^\top\boldsymbol{\theta}^*)\right)^2 \ge \mathtt{b}\mathtt{r}^2/n\right] > 0,\end{aligned} \tag{A.30}$$

This means that in terms of Theorem A.19 under assumptions (1), (2) and (3) we have $C_e \le 3(2+\mathtt{C})\mathtt{r}^{*2} + \mathtt{C}_{\sum}$ and $\tau_e \le \exp\left\{-m^3\mathtt{x}\right\}$. We prove assumptions (1), (2) and (3) in Lemmas A.22, A.23 and A.24, which will give that $C_e \le \mathtt{C}_m + 3(2+\mathtt{C})\mathtt{r}^{*2}$ with probability greater than $1 - \mathrm{e}^{-m^3\mathtt{x}}$ and that $\boldsymbol{Q}(\mathtt{b}) > 0$ for a certain choice of $\mathtt{b} > 0$ small enough and for $\mathtt{r} \ge \mathtt{C}\sqrt{m}$ with some constant $\mathtt{C}$. Lemma A.21 completes the proof. □

Lemma A.21. *Under the assumptions (1), (2) and (3) we get*

$$\inf_{\boldsymbol{\upsilon}\in\Upsilon_\circ(\mathtt{r})^c} -\mathbb{E}[\mathcal{L}(\boldsymbol{\upsilon},\boldsymbol{\upsilon}_m^*)|(\mathbf{X})] \ge \lambda\mathtt{b}\mathtt{r}^2$$

with probability greater than $1 - \exp\left\{-m^3\mathtt{x}\right\} - \exp\left\{-n\boldsymbol{Q}(2\mathtt{b})^2/4\right\}$ *for*

$$\mathtt{r}^2 \ge (3(2+\mathtt{C})\mathtt{r}^{*2} + \mathtt{C}_{\sum})/(\lambda\mathtt{b}) \vee \mathtt{C}m,$$

if

$$0 < \lambda \stackrel{\text{def}}{=} \left(\boldsymbol{Q}(2\mathtt{b}) - 2/n + \mathtt{C}\sqrt{\frac{\log(n)p^*}{n}}\right)/4,$$

for a constant $\mathtt{C} > 0$ *which is a function of* $\|\psi\|_\infty, \|\psi\|_\infty, s_{\mathbf{X}}$.

Proof. This is a direct consequence of Theorem A.19. It remains to bound using the proof of Theorem 8.15 of [14]

$$\begin{aligned}\mathbb{E}\left[\sup_{\boldsymbol{\upsilon}\in\Upsilon_\circ(\mathtt{r})^c}(\mathrm{P}_n - \mathbb{P})\chi_{\mathtt{b}}(\boldsymbol{\upsilon})\right] &\le \mathbb{E}\left[\sup_{\boldsymbol{\upsilon}\in\Upsilon}(\mathrm{P}_n - \mathbb{P})\chi_{\mathtt{b}}(\boldsymbol{\upsilon})\right]\\ &\le 2\mathtt{C}^*\mathbb{E}\left[\sqrt{\frac{6\{1+\log N(\delta,\mathcal{F},L_1(\mathrm{P}_n))\}}{n}}\right] + \delta,\end{aligned} \tag{A.31}$$

where $N(\delta,\mathcal{F},L_1(\mathrm{P}_n))$ denotes the δ-ball covering number of $\mathcal{F} \stackrel{\text{def}}{=} \{\chi_{\mathtt{b}}(\boldsymbol{\upsilon}) : \boldsymbol{\upsilon} \in \Upsilon\}$ with respect to the norm

$$\|h\|_{L_1(\mathrm{P}_n)} = \mathrm{P}_n|h(\mathbf{X})| = \frac{1}{n}\sum_{i=1}^n |h(\mathbf{X}_i)|.$$

The universal constant $\mathtt{C}^* > 0$ comes from Lemma 8.2 of [14] ($\mathtt{C}^* = K(\exp(x^2) - 1)$). The function $\chi_{\mathtt{b}} : \varUpsilon_\circ \to \mathbb{R}$ is defined via

$$\overline{\chi}_u(t) = \begin{cases} 0 & t \le u; \\ t/u - 1 & t \in [u, 2u]; \\ 1 & t \ge 2u; \end{cases} \qquad \chi_{\mathtt{b}}(\boldsymbol{\upsilon})_i \overset{\text{def}}{=} \overline{\chi}_{\mathtt{b}}(|\boldsymbol{f}_{\boldsymbol{\eta}}(\mathbf{X}_i^\top \boldsymbol{\theta}) - \boldsymbol{f}_{\boldsymbol{\eta}^*}(\mathbf{X}_i^\top \boldsymbol{\theta}^*)|^2).$$

We want to bound the right-hand side of (A.31). For this note that

$$\log N(\delta, \mathcal{F}, L_1(\mathrm{P}_n)) \le \log N(\delta/(L(\mathrm{P}_n) \vee 1), \varUpsilon, \|\cdot\|_2),$$

where

$$L(\mathrm{P}_n) = \sup_{\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^\circ \in \varUpsilon} \frac{\|\chi_{\mathtt{b}}(\boldsymbol{\upsilon}) - \chi_{\mathtt{b}}(\boldsymbol{\upsilon}^\circ)\|_{L_1(\mathrm{P}_n)}}{\|\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^\circ\|_2}.$$

We estimate using that we have $\operatorname{diam}(\varUpsilon_m) < \mathtt{C}\sqrt{m}$

$$\begin{aligned}
&|\chi_{\mathtt{b}}(\boldsymbol{\upsilon})_i - \chi_{\mathtt{b}}(\boldsymbol{\upsilon}^\circ)_i| \\
&\quad \le |\boldsymbol{f}_{\boldsymbol{\eta}}(\mathbf{X}_i^\top \boldsymbol{\theta}) - \boldsymbol{f}_{\boldsymbol{\eta}^\circ}(\mathbf{X}_i^\top \boldsymbol{\theta}^\circ)|^2 \\
&\qquad + 2|(\boldsymbol{f}_{\boldsymbol{\eta}}(\mathbf{X}_i^\top \boldsymbol{\theta}) - \boldsymbol{f}_{\boldsymbol{\eta}^\circ}(\mathbf{X}_i^\top \boldsymbol{\theta}^\circ))(\boldsymbol{f}_{\boldsymbol{\eta}}(\mathbf{X}_i^\top \boldsymbol{\theta}) - \boldsymbol{f}_{\boldsymbol{\eta}^*}(\mathbf{X}^\top \boldsymbol{\theta}^*))| \\
&\quad \le 2|\boldsymbol{f}_{\boldsymbol{\eta}-\boldsymbol{\eta}^\circ}(\mathbf{X}_i^\top \boldsymbol{\theta})|^2 + 2|\boldsymbol{f}_{\boldsymbol{\eta}^\circ}(\mathbf{X}_i^\top \boldsymbol{\theta}) - \boldsymbol{f}_{\boldsymbol{\eta}^\circ}(\mathbf{X}_i^\top \boldsymbol{\theta}^\circ)|^2 \\
&\qquad + \sqrt{2|\boldsymbol{f}_{\boldsymbol{\eta}-\boldsymbol{\eta}^\circ}(\mathbf{X}_i^\top \boldsymbol{\theta})|^2 + 2|\boldsymbol{f}_{\boldsymbol{\eta}^\circ}(\mathbf{X}_i^\top \boldsymbol{\theta}) - \boldsymbol{f}_{\boldsymbol{\eta}^\circ}(\mathbf{X}_i^\top \boldsymbol{\theta}^\circ)|^2} \\
&\qquad |\boldsymbol{f}_{\boldsymbol{\eta}}(\mathbf{X}_i^\top \boldsymbol{\theta}) - \boldsymbol{f}_{\boldsymbol{\eta}^*}(\mathbf{X}^\top \boldsymbol{\theta}^*)| \\
&\quad \le 2\|\boldsymbol{\eta} - \boldsymbol{\eta}^\circ\|^2 m \|\psi\|_\infty^2 + 2\|\boldsymbol{\theta} - \boldsymbol{\theta}^\circ\|^2 s_{\mathbf{X}^2} m^3 \|\psi'\|_\infty^2 \|\boldsymbol{\eta}^\circ\|^2 \\
&\qquad + \sqrt{2\|\boldsymbol{\eta} - \boldsymbol{\eta}^\circ\|^2 m \|\psi\|_\infty^2 + 2\|\boldsymbol{\theta} - \boldsymbol{\theta}^\circ\|^2 s_{\mathbf{X}^2} m^3 \|\psi'\|_\infty^2 \|\boldsymbol{\eta}^\circ\|^2} \\
&\qquad \sqrt{m}\|\psi\|_\infty (\|\boldsymbol{\eta}\| + \|\boldsymbol{\eta}^*\|) \\
&\quad \le \mathtt{C}_1 m^3 \|\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^\circ\| + \mathtt{C}_2 m^4 \|\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^\circ\|^2.
\end{aligned}$$

But note that by the triangular inequality we also have $|\chi_{\mathtt{b}}(\boldsymbol{\upsilon})_i - \chi_{\mathtt{b}}(\boldsymbol{\upsilon}^\circ)_i| \le 2$. This gives

$$\begin{aligned}
\sup_{\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^\circ} \frac{\|\chi_{\mathtt{b}}(\boldsymbol{\upsilon}) - \chi_{\mathtt{b}}(\boldsymbol{\upsilon}^\circ)\|_{L_1(\mathrm{P}_n)}}{\|\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^\circ\|_2} &\le \sup_{\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^\circ} \left(\frac{2}{\|\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^\circ\|_2} \wedge \mathtt{C}_1 m^3 + \mathtt{C}_2 m^4 \|\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^\circ\|_2 \right) \\
&= \mathtt{C}_3 m^3.
\end{aligned}$$

We infer setting $\delta = \sqrt{p^*/n}$

$$
\begin{aligned}
&\sqrt{\frac{6\{1+\log N(\delta,\mathcal{F},L_1(\mathrm{P}_n))\}}{n}}+\delta\\
&\qquad\le \sqrt{\frac{6\{1+\log N(\delta/(L(\mathrm{P}_n)\vee 1),\varUpsilon,\|\cdot\|_2)\}}{n}}+\delta\\
&\qquad\le \sqrt{\frac{6\{1+\log(\mathtt{C}m^3)+\log(1/\delta)p^*\}}{n}}+\delta\\
&\qquad\le \mathtt{C}_1\sqrt{\frac{\log(p^*)+\log(n/p^*)p^*/2}{n}}+\sqrt{p^*/n}\\
&\qquad\le \mathtt{C}_2\sqrt{\frac{\log(n)p^*}{n}}.
\end{aligned}
$$

The claim follows with Theorem A.19. □

It remains to prove the assumptions (1), (2) and (3) which we do in the following three lemmas.

Lemma A.22. *We have for some* $\mathtt{C}>0$

$$
n\mathbb{E}[\|\boldsymbol{f}_{\boldsymbol{\eta}^*_m}(\mathbf{X}^\top\boldsymbol{\theta}^*_m)-\boldsymbol{f}_{\boldsymbol{\eta}^*}(\mathbf{X}^\top\boldsymbol{\theta}^*)\|^2]\le 3(2+\mathtt{C})\mathtt{r}^{*2}.
$$

Proof. We find with the Taylor expansion, Lemma A.2 of [1] (which is applicable because it only needs $(\boldsymbol{\mathcal{L}}\mathbf{r})$ for the full model and with center $\boldsymbol{\upsilon}^*\in\varUpsilon$) and Lemma A.6 with some $\boldsymbol{\theta}^\circ\in\mathbf{Conv}(\boldsymbol{\theta}^*_m,\boldsymbol{\theta}^*)$

$$
\begin{aligned}
&n\mathbb{E}[\|\boldsymbol{f}_{\boldsymbol{\eta}^*_m}(\mathbf{X}^\top\boldsymbol{\theta}^*_m)-\boldsymbol{f}_{\boldsymbol{\eta}^*}(\mathbf{X}^\top\boldsymbol{\theta}^*)\|^2]\\
&\quad\le 3n\left(\mathbb{E}[\|\boldsymbol{f}_{\boldsymbol{\eta}^*}(\mathbf{X}^\top\boldsymbol{\theta}^*_m)-\boldsymbol{f}_{\boldsymbol{\eta}^*}(\mathbf{X}^\top\boldsymbol{\theta}^*)\|^2]+\mathbb{E}[\|\boldsymbol{f}_{\boldsymbol{\eta}^*_m-\boldsymbol{\eta}^*}(\mathbf{X}^\top\boldsymbol{\theta}^*_m)\|^2]\right)\\
&\quad\le 3\left(\|\mathrm{D}(\boldsymbol{\theta}^\circ)(\boldsymbol{\theta}^*_m-\boldsymbol{\theta}^*)\|^2+\|\mathcal{H}(\boldsymbol{\upsilon}^*_m)(\boldsymbol{\eta}^*_m-\boldsymbol{f}^*)\|^2\right)\\
&\quad\le 3\Big((1+\|I-\mathrm{D}^{-1/2}n\mathrm{D}(\boldsymbol{\xi})\mathrm{D}^{-1/2}\|)\|\mathrm{D}(\boldsymbol{\theta}^*_m-\boldsymbol{\theta}^*)\|^2\\
&\qquad+(1+\|I-\mathcal{H}^{-1}n\widetilde{\mathcal{H}}(\boldsymbol{\upsilon}^*_m)\mathcal{H}^{-1}\|)\|\mathcal{H}(\boldsymbol{\eta}^*_m-\boldsymbol{f}^*)\|^2\Big)\\
&\quad\le 3\left[2+\|I-\mathrm{D}^{-1/2}n\mathrm{D}(\boldsymbol{\theta}^\circ)\mathrm{D}^{-1/2}\|+\|I-\mathcal{H}^{-1}n\mathcal{H}(\boldsymbol{\upsilon}^*_m)\mathcal{H}^{-1}\|\right]\\
&\qquad\|\mathcal{D}(\boldsymbol{\upsilon}^*_m-\boldsymbol{\upsilon}^*)\|^2\\
&\quad\le 3(2+\mathtt{C})\mathtt{r}^{*2}.
\end{aligned}
$$

□

Lemma A.23. *We have for a constant* $\mathtt{C} > 0$ *that only depends on* $\|\psi\|_\infty$, $\|\psi'\|_\infty$ *and* $s_{\mathbf{X}^2}$ *that*

$$\mathbb{P}\left(n\left|(P_n-\mathbb{P})|\boldsymbol{f}_{\boldsymbol{\eta}^*_m}(\mathbf{X}^\top\boldsymbol{\theta}^*_m)-\boldsymbol{f}_{\boldsymbol{\eta}^*}(\mathbf{X}^\top\boldsymbol{\theta}^*)|^2\right|\geq \mathtt{C}\sqrt{\mathtt{x}}\right)\leq \exp\left\{-m^3\mathtt{x}\right\}.$$

Proof. We want to use the finite difference inequality. As above define

$$f:\bigotimes_{i=1}^n\mathbb{R}^p\to\mathbb{R},\quad f(\mathbf{X}_1,\ldots,\mathbf{X}_n)\overset{\mathrm{def}}{=}\mathrm{P}_n|\boldsymbol{f}_{\boldsymbol{\eta}^*_m}(\mathbf{X}^\top\boldsymbol{\theta}^*_m)-\boldsymbol{f}_{\boldsymbol{\eta}^*}(\mathbf{X}^\top\boldsymbol{\theta}^*)|^2,$$

and note that for any $i=1,\ldots,n$ and any alternative realization $\mathbf{X}_i'\in\mathbb{R}$

$$\begin{aligned}&n|f(\mathbf{X}_1,\ldots,\mathbf{X}_{i-1},\mathbf{X}_i,\mathbf{X}_{i+1},\ldots,\mathbf{X}_n)-f(\mathbf{X}_1,\ldots,\mathbf{X}_{i-1},\mathbf{X}_i',\mathbf{X}_{i+1},\ldots,\mathbf{X}_n)|\\&\quad\leq|\boldsymbol{f}_{\boldsymbol{\eta}^*_m}(\mathbf{X}_i^\top\boldsymbol{\theta}^*_m)-\boldsymbol{f}_{\boldsymbol{\eta}^*}(\mathbf{X}_i^\top\boldsymbol{\theta}^*)|^2+|\boldsymbol{f}_{\boldsymbol{\eta}^*_m}({\mathbf{X}'}_i^{\top}\boldsymbol{\theta}^*_m)-g({\mathbf{X}'}_i)|^2.\end{aligned}$$

We have

$$\begin{aligned}|\boldsymbol{f}_{\boldsymbol{\eta}^*_m}(\mathbf{X}^\top\boldsymbol{\theta}^*_m)-\boldsymbol{f}_{\boldsymbol{\eta}^*}(\mathbf{X}^\top\boldsymbol{\theta}^*)|^2\leq{}&3|\boldsymbol{f}_{\boldsymbol{\eta}^*-\boldsymbol{\eta}^*_m}(\mathbf{X}_i^\top\boldsymbol{\theta})|^2\\&+3|\boldsymbol{f}_{\boldsymbol{\eta}^*}(\mathbf{X}_i^\top\boldsymbol{\theta}^*)-\boldsymbol{f}_{\boldsymbol{\eta}^*}(\mathbf{X}_i^\top\boldsymbol{\theta}^*_m)|^2.\end{aligned}$$

As in Lemma A.11 there are constants $\mathtt{C},\mathtt{C}'$ such that

$$\begin{aligned}|\boldsymbol{f}_{\boldsymbol{\eta}^*-\boldsymbol{\eta}^*_m}(\mathbf{X}_i^\top\boldsymbol{\theta}^*_m)|^2&\leq 3\left|\sum_{k=1}^m(\eta^*_k-\eta^*_{k,m})\boldsymbol{e}_k(\mathbf{X}_i^\top\boldsymbol{\theta}^*_m)\right|^2+3\left|\sum_{k=m+1}^\infty\eta^*_k\boldsymbol{e}_k(\mathbf{X}_i^\top\boldsymbol{\theta}^*_m)\right|^2\\&\leq 3\left|\sum_{k=1}^m\boldsymbol{e}_k^2(\mathbf{X}_i^\top\boldsymbol{\theta}^*_m)\right|\|\mathit{\Pi}_m\boldsymbol{\eta}^*-\boldsymbol{\eta}^*_m\|^2+\mathtt{C}(\boldsymbol{\kappa}^*)\\&\leq\mathtt{C}'\left|\sum_{j=0}^{j_m}2^j\right|\|\mathit{\Pi}_m\boldsymbol{\eta}^*-\boldsymbol{\eta}^*_m\|^2+\mathtt{C}(\boldsymbol{\kappa}^*)\\&\leq\mathtt{C}m\|\mathit{\Pi}_m\boldsymbol{\eta}^*-\boldsymbol{\eta}^*_m\|^2+\mathtt{C}(\boldsymbol{\kappa}^*),\end{aligned}$$

where $\mathtt{C}(\boldsymbol{\kappa}^*)\leq\mathtt{C}m^{-2\alpha+1}$. Furthermore again as in Lemma A.11 there are constants $\mathtt{C},\mathtt{C}'$ such that

$$\begin{aligned}|\boldsymbol{f}_{\boldsymbol{\eta}^*}(\mathbf{X}_i^\top\boldsymbol{\theta}^*)-\boldsymbol{f}_{\boldsymbol{\eta}^*}(\mathbf{X}_i^\top\boldsymbol{\theta}^*_m)|^2&\leq\left|\sum_{k=1}^m\eta^*_k\left(\boldsymbol{e}_k(\mathbf{X}_i^\top\boldsymbol{\theta}^*)-\boldsymbol{e}_k(\mathbf{X}_i^\top\boldsymbol{\theta}^*_m)\right)\right|^2\\&\leq\mathtt{C}'\left|\sum_{j=0}^{j_m}2^{3j-2\alpha}\right|\|\boldsymbol{\theta}^*-\boldsymbol{\theta}^*_m\|^2\\&\leq\mathtt{C}\|\boldsymbol{\theta}^*-\boldsymbol{\theta}^*_m\|^2.\end{aligned}$$

This implies with Lemma 5.1 and constants $\mathtt{C}_1, \mathtt{C}_2 > 0$

$$|\boldsymbol{f}_{\boldsymbol{\eta}^*_m}(\mathbf{X}^\top \boldsymbol{\theta}^*_m) - \boldsymbol{f}_{\boldsymbol{\eta}^*}(\mathbf{X}^\top \boldsymbol{\theta}^*)|^2 \le \mathtt{C}_1 \left(\frac{m}{n c_{\mathcal{D}}^2} \mathtt{r}^{*2} + m^{-2\alpha+1} \right) \le \mathtt{C} m^{-3}.$$

Note that $\mathtt{r}^{*2} m/n \to 0$. This gives with the bounded difference inequality (Theorem A.1) that

$$\mathbb{P}\left(n \left| (\mathrm{P}_n - \mathbb{P}) |\boldsymbol{f}_{\boldsymbol{\eta}^*_m}(\mathbf{X}^\top \boldsymbol{\theta}^*_m) - \boldsymbol{f}_{\boldsymbol{\eta}^*}(\mathbf{X}^\top \boldsymbol{\theta}^*)|^2 \right| \ge t \mathtt{C} m^{-3} \right) \le \exp\left\{ -t^2 \right\}.$$

From this we infer with $t = m^3 \sqrt{\mathtt{x}} \to \infty$

$$\mathbb{P}\left(n \left| (\mathrm{P}_n - \mathbb{P}) |\boldsymbol{f}_{\boldsymbol{\eta}^*_m}(\mathbf{X}^\top \boldsymbol{\theta}^*_m) - \boldsymbol{f}_{\boldsymbol{\eta}^*}(\mathbf{X}^\top \boldsymbol{\theta}^*)|^2 \right| \ge \mathtt{C}_2 \sqrt{\mathtt{x}} \right) \le \exp\left\{ -m^3 \mathtt{x} \right\}. \quad \square$$

For a set $A \subset \mathbb{R}^p$ we denote by $\lambda(A) \in \mathbb{R}_+$ its Lebesgue measure and define

$$\lambda_{\boldsymbol{e}} \overset{\text{def}}{=} \sup \left\{ \lambda > 0 : \inf_{\substack{\boldsymbol{v} \in \mathbb{R}^m, \|\boldsymbol{v}\|=1 \\ \boldsymbol{\theta} \in S_1^{p,+}}} \mathbb{P}\left(|\langle \boldsymbol{v}, \boldsymbol{e}(\mathbf{X}^\top \boldsymbol{\theta}) \rangle| > \lambda \right) > 3/4 \right\}. \tag{A.32}$$

Remark A.8. $\lambda_{\boldsymbol{e}} \ge \mathbb{R}$ in (A.32) is strictly greater 0 because the basis functions are linearly independent and we assumed the distribution of the regressors $\mathbf{X}$ to be absolutely continuous with respect to the Lebesgue measure.

Lemma A.24. *Denote the cylinder*

$$C_{\rho,x,y}(x_0, y_0) \overset{\text{def}}{=} \{(x, y, z) \in \mathbb{R}^2 \times \mathbb{R}^{p-2};\ (x - x_0)^2 + (y - y_0)^2 \le \rho^2\}.$$

There is a point $(x_0, y_0) \in \mathbb{R}^2$ *such that* $\boldsymbol{Q}(2\mathtt{b})$ *in* (A.30) *satisfies*

$$\boldsymbol{Q}(2\mathtt{b}) + 3\mathrm{e}^{-\mathtt{x}} \ge \frac{1}{2} \wedge c_{p_{\mathbf{X}}} \lambda \Big(B_h(0) \cap C_{h,x,y}(0) \cap B_{s_{\mathbf{X}}}(x_0, y_0, 0)$$
$$\cap \left\{ (x, y) \in \mathbb{R}^2 : \operatorname{sign}(y_0) y \ge \operatorname{sign}(y_0) h/2 \right\} \Big),$$

for $\tau = \lambda_{\boldsymbol{e}} / (8 \boldsymbol{L}_{\boldsymbol{\eta}^*} s_{\mathbf{X}})$ *and*

$$2\mathtt{b} = (1 - \rho^2) \left(\frac{\lambda_{\boldsymbol{e}}^2 c_{\mathcal{D}}^2}{32} \wedge \frac{\tau c_{f'_{\boldsymbol{\eta}^*}}^2 h^2}{4 p \pi^2 s_{\mathbf{X}}^2 \|p_{\mathbf{X}}\|_\infty^2 C_{\|\boldsymbol{\eta}^*\|}} \right),$$

and for

$$\mathtt{r} \ge \sqrt{m} \frac{4 \mathtt{C}_{\boldsymbol{\kappa}}}{\lambda_{\boldsymbol{e}} \sqrt{(1 - \rho)}}.$$

Remark A.9. The constants $h, c_{\boldsymbol{f}'_{\boldsymbol{\eta}^*}} > 0$ are from assumption $(\mathbf{Cond}_{\mathbf{X}\boldsymbol{\theta}^*})$.

Proof. We have to prove

$$\inf_{\boldsymbol{\upsilon}\in\Upsilon_\circ(\mathtt{r})^c} \mathbb{P}\left[\left(\boldsymbol{f}_{\boldsymbol{\eta}}(\mathbf{X}_i^\top\boldsymbol{\theta}) - \boldsymbol{f}_{\boldsymbol{\eta}^*}(\mathbf{X}_i^\top\boldsymbol{\theta}^*)\right)^2 \geq \frac{\mathtt{br}^2}{n}\right] > 0. \tag{A.33}$$

We carry out the proof in two steps.

1. Before we determine $\mathtt{b} > 0$ that allows to prove (A.33) note that

$$\begin{aligned}&\|\mathcal{D}_m(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^*_m)\| - \|\mathcal{D}_m(\varPi_{p^*}\boldsymbol{\upsilon}^* - \boldsymbol{\upsilon}^*_m)\| \leq \|\mathcal{D}_m(\boldsymbol{\upsilon} - \varPi_{p^*}\boldsymbol{\upsilon}^*)\|\\ &\quad \leq \|\mathcal{D}_m(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^*_m)\| + \|\mathcal{D}_m(\varPi_{p^*}\boldsymbol{\upsilon}^* - \boldsymbol{\upsilon}^*_m)\|.\end{aligned}$$

Slightly modifying Lemma A.3 of [1] with $\boldsymbol{\theta} = \boldsymbol{\upsilon}$ gives

$$\|\mathcal{D}_m(\varPi_{p^*}\boldsymbol{\upsilon}^* - \boldsymbol{\upsilon}^*_m)\| \leq \Big(\alpha(m) + \tau(m) + 2\delta(2\mathtt{r}^*)\mathtt{r}^*\Big) \overset{\text{def}}{=} \mathtt{r}^*_\epsilon(m),$$

where due to Lemma A.6 and the definition of $\mathtt{r}^* > 0$ in Lemma 5.1

$$\mathtt{r}^* \leq \mathtt{C}\sqrt{m},\ \alpha(m) = \mathtt{C}\left(m^{-\alpha-1/2} + \mathtt{C}_{bias}m^{-(\alpha-1)}\right)\sqrt{n},\ \tau(m) \leq \mathtt{C}m^{-2\alpha+1/2}\sqrt{n}.$$

With arguments as above we find that $\mathtt{r}^*_\epsilon(m) > 0$ is neglect-ably small for $n \in \mathbb{N}$ large enough. We have with some small $\epsilon > 0$

$$\begin{aligned}(1-\epsilon)\|\mathcal{D}_m(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^*_m)\|^2 &\leq \|\mathcal{D}_m(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^*)\|^2\\ &\leq (1+\epsilon)\|\mathcal{D}_m(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^*_m)\|^2.\end{aligned} \tag{A.34}$$

Assume that $n \in \mathbb{N}$ is large enough to ensure that $\epsilon < 1/2$. Then we find for $\boldsymbol{\upsilon} \in \Upsilon_\circ(\mathtt{r})^c$ and with Lemma B.5 of [2] and (A.34) that

$$\|\mathrm{D}(\varphi_{\boldsymbol{\theta}} - \varphi_{\boldsymbol{\theta}^*})\|^2 + \|\mathrm{H}_m(\boldsymbol{\eta} - \boldsymbol{\eta}^*)\|^2 \geq (1-\rho)\|\mathcal{D}_m(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^*)\|^2 \geq (1-\rho)\mathtt{r}^2/2.$$

2. Now we show (A.30). We treat two cases for $(\varphi_{\boldsymbol{\theta}}, \boldsymbol{\eta}) \in \mathbb{R}^{p-1} \times \mathbb{R}^m$ separately. The first case is that $\|\mathrm{D}(\varphi_{\boldsymbol{\theta}} - \varphi_{\boldsymbol{\theta}^*})\|^2 \leq \frac{1}{4}(1-\rho)\mathtt{r}^2$. In this situation we can use the smoothness of $\boldsymbol{f}_{\boldsymbol{\eta}^*_m}$ and $\boldsymbol{f}_{\boldsymbol{\eta}^*}$ to determine $\mathtt{b} > 0$. In the second case we use the geometric structure of

$$\left(\boldsymbol{f}_{\boldsymbol{\eta}}(\mathbf{X}^\top\boldsymbol{\theta}) - \boldsymbol{f}_{\boldsymbol{\eta}^*}(\mathbf{X}^\top\boldsymbol{\theta}^*)\right)^2 > 0,$$

to obtain a good lower bound.

Case 1: $\|\mathrm{D}(\varphi_{\boldsymbol{\theta}} - \varphi_{\boldsymbol{\theta}^*})\|^2 \le \frac{1}{2}\tau \mathtt{r}^2$. In this case we simply calculate and find

$$\begin{aligned}
&|\boldsymbol{f}_{\boldsymbol{\eta}}(\mathbf{X}^\top\boldsymbol{\theta}) - \boldsymbol{f}_{\boldsymbol{\eta}^*}(\mathbf{X}^\top\boldsymbol{\theta}^*)|^2 \\
&\quad \ge |\boldsymbol{f}_{\boldsymbol{\eta}}(\mathbf{X}^\top\boldsymbol{\theta}) - \boldsymbol{f}_{\boldsymbol{\eta}^*}(\mathbf{X}^\top\boldsymbol{\theta})|^2 \\
&\qquad -2|\boldsymbol{f}_{\boldsymbol{\eta}}(\mathbf{X}^\top\boldsymbol{\theta}) - \boldsymbol{f}_{\boldsymbol{\eta}^*}(\mathbf{X}^\top\boldsymbol{\theta})||\boldsymbol{f}_{\boldsymbol{\eta}^*}(\mathbf{X}^\top\boldsymbol{\theta}) - \boldsymbol{f}_{\boldsymbol{\eta}^*}(\mathbf{X}^\top\boldsymbol{\theta}^*)| \\
&\quad \ge |\boldsymbol{f}_{\boldsymbol{\eta}}(\mathbf{X}^\top\boldsymbol{\theta}) - \boldsymbol{f}_{\boldsymbol{\eta}^*}(\mathbf{X}^\top\boldsymbol{\theta})|^2 - 2|\boldsymbol{f}_{\boldsymbol{\eta}}(\mathbf{X}^\top\boldsymbol{\theta}) - \boldsymbol{f}_{\boldsymbol{\eta}^*}(\mathbf{X}^\top\boldsymbol{\theta})|\boldsymbol{L}_{\boldsymbol{\eta}^*} s_{\mathbf{X}}\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|.
\end{aligned}$$

Now

$$|\boldsymbol{f}_{\boldsymbol{\eta}}(\mathbf{X}^\top\boldsymbol{\theta}) - \boldsymbol{f}_{\boldsymbol{\eta}^*}(\mathbf{X}^\top\boldsymbol{\theta})| \ge |\boldsymbol{f}_{\boldsymbol{\eta}-\boldsymbol{\eta}^*}(\mathbf{X}^\top\boldsymbol{\theta})| - |\boldsymbol{f}_{(0,\boldsymbol{\kappa}^*)}(\mathbf{X}^\top\boldsymbol{\theta})|.$$

We find with probability greater than 3/4

$$\begin{aligned}
|\boldsymbol{f}_{\boldsymbol{\eta}-\boldsymbol{\eta}^*}(\mathbf{X}^\top\boldsymbol{\theta})| &= |\langle \boldsymbol{\eta} - \boldsymbol{\eta}^*, \boldsymbol{e}(\mathbf{X}^\top\boldsymbol{\theta})\rangle| \\
&\ge \|\mathrm{H}_m(\boldsymbol{\eta} - \boldsymbol{\eta}^*)\|\lambda_{\boldsymbol{e}} \\
&\ge \mathtt{r}\lambda_{\boldsymbol{e}}\frac{1}{2}\sqrt{(1-\rho^2)},
\end{aligned}$$

where

$$\lambda_{\boldsymbol{e}} \overset{\text{def}}{=} \sup\left\{\lambda > 0 : \inf_{\substack{\boldsymbol{\eta}\in\mathbb{R}^m, \|\boldsymbol{\eta}\|=1 \\ \boldsymbol{\theta}\in S_1^{p,+}}} \mathbb{P}\left(|\langle \boldsymbol{\eta}, \mathrm{H}_m^{-1}\boldsymbol{e}(\mathbf{X}^\top\boldsymbol{\theta})\rangle| > \lambda\right) > 3/4\right\},$$

which is larger 0 because the basis functions are linearly independent and we assumed the distribution of the regressors $\mathbf{X}$ to be absolutely continuous to the Lebesgue measure. Remember that by Lemma A.6

$$\begin{aligned}
\|\mathcal{H}_m^{1/2}\boldsymbol{\kappa}^*\|^2 &< \left(17\|p_{\mathbf{X}^\top\boldsymbol{\theta}^*}\|_\infty \mathtt{C}_{\|\boldsymbol{f}^*\|} + 17^2\sqrt{36} s_{\mathbf{X}}^{p+1} L_{p_{\mathbf{X}}}\|\psi\|_\infty \mathtt{C}_{\|\boldsymbol{f}^*\|}^2\right) n m^{-2\alpha} \\
&\overset{\text{def}}{=} \mathtt{C}_{\boldsymbol{\kappa}}^2 m.
\end{aligned}$$

We use the Markov inequality to obtain

$$\mathbb{P}\left(|\boldsymbol{f}_{(0,\boldsymbol{\kappa}^*)}(\mathbf{X}^\top\boldsymbol{\theta})|^2 \ge 4\mathtt{C}_{\boldsymbol{\kappa}}\frac{m}{n}\right) \le \frac{\|\mathcal{H}_m^{1/2}\boldsymbol{\kappa}^*\|^2}{4\mathtt{C}_{\boldsymbol{\kappa}}^2 m} \le 1/4.$$

This implies that with probability greater than $1/2 = 3/4 - 1/4$

$$\begin{aligned}
|\boldsymbol{f}_{\boldsymbol{\eta}}(\mathbf{X}^\top\boldsymbol{\theta}) - \boldsymbol{f}_{\boldsymbol{\eta}^*}(\mathbf{X}^\top\boldsymbol{\theta})| &\ge \mathtt{r}\lambda_{\boldsymbol{e}}\frac{1}{2}\sqrt{(1-\rho^2)} - 4\mathtt{C}_{\boldsymbol{\kappa}}\sqrt{\frac{m}{n}} \\
&\ge \frac{\sqrt{(1-\rho^2)}\lambda_{\boldsymbol{e}}}{4\sqrt{n}}\mathtt{r},
\end{aligned}$$

for

$$\mathtt{r} \geq \sqrt{m}\frac{4\mathtt{C}_{\boldsymbol{\kappa}}}{\lambda_{\boldsymbol{e}}\sqrt{(1-\rho^2)}}.$$

We still have to account for the summand $\boldsymbol{L}_{\boldsymbol{\eta}^*} s_{\mathbf{X}}\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|$ via

$$\boldsymbol{L}_{\boldsymbol{\eta}^*} s_{\mathbf{X}}\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\| \leq \frac{\boldsymbol{L}_{\boldsymbol{\eta}^*} s_{\mathbf{X}}\sqrt{\tau(1-\rho^2)}}{2c_{\mathfrak{D}}\sqrt{n}}\mathtt{r}.$$

This gives for the choice of $\tau = \lambda_{\boldsymbol{e}} c_{\mathfrak{D}}/(8\boldsymbol{L}_{\boldsymbol{\eta}^*} s_{\mathbf{X}})$

$$\begin{aligned}
&|\boldsymbol{f}_{\boldsymbol{\eta}}(\mathbf{X}^\top\boldsymbol{\theta}) - \boldsymbol{f}_{\boldsymbol{\eta}^*}(\mathbf{X}^\top\boldsymbol{\theta})| - 2\boldsymbol{L}_{\boldsymbol{\eta}^*} s_{\mathbf{X}}\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\| \\
&\geq \left(\frac{\lambda_{\boldsymbol{e}}}{4} - \frac{\boldsymbol{L}_{\boldsymbol{\eta}^*} s_{\mathbf{X}}\sqrt{\tau}}{c_{\mathfrak{D}}}\right)\frac{\sqrt{(1-\rho^2)}}{\sqrt{n}}\mathtt{r} \\
&= \frac{\lambda_{\boldsymbol{e}} c_{\mathfrak{D}}\sqrt{(1-\rho^2)}}{8\sqrt{n}}\mathtt{r}.
\end{aligned}$$

We obtain in case 1 that $\boldsymbol{Q}(2\mathtt{b}) \geq 1/2$ for

$$2\mathtt{b}/n \overset{\text{def}}{=} \frac{(1-\rho^2)\lambda_{\boldsymbol{e}}^2 c_{\mathfrak{D}}^2}{32n}.$$

Case 2: $\frac{1}{2}\tau(1-\rho)\mathtt{r}^2 \leq \|\mathrm{D}(\varphi_{\boldsymbol{\theta}} - \varphi_{\boldsymbol{\theta}^*})\|^2 \leq \sqrt{2}\lambda_{\max}\mathrm{D}^2$.
Take some $f:\mathbb{R}\to\mathbb{R}$ with $f' > c$ and some $(\alpha,\beta)\in\mathbb{R}^2$ with $\alpha^2+\beta^2 = 1$. Furthermore take any $g:\mathbb{R}\to\mathbb{R}$. We are interested in determining

$$\begin{aligned}
V(\tau) &\overset{\text{def}}{=} \inf_{\substack{f\in C^1(\mathbb{R}),\, f'>c,\\ g:\mathbb{R}\to\mathbb{R}}} \lambda\left(\mathcal{A}(\tau)\right), \\
\mathcal{A}(\tau) &\overset{\text{def}}{=} \left\{(x,y,z)\in\mathbb{R}^2\times\mathbb{R}^{p-2};\, |f(\alpha x+\beta y) - g(x)| > \tau\right\} \\
&\quad \cap C_{\rho,x,y}(0) \cap B_{s_{\mathbf{X}}}(x_0,y_0,0) \subset \mathbb{R}^2\times\mathbb{R}^{p-2}, \\
C_{\rho,x,y}(x_0,y_0) &\overset{\text{def}}{=} \{(x,y,z)\in\mathbb{R}^2\times\mathbb{R}^{p-2};\, (x-x_0)^2 + (y-y_0)^2 \leq \rho^2\},
\end{aligned}$$

where for a set $A\subset\mathbb{R}^p$ we denote by $\lambda(A)\in\mathbb{R}_+$ its Lebesgue measure. For this observe

$$f(\alpha x+\beta y) - g(x) \begin{cases} \geq c\beta y + f(\alpha x) - g(x) & \beta > 0, \\ \leq c\beta y + f(\alpha x) - g(x) & \beta \leq 0. \end{cases}$$

Consequently for fixed $x\in[-\rho,\rho]$ we have $|f(\alpha x+\beta y) - g(x)| > \rho\beta c/2$ on the set

$$\{y\in[-\sqrt{\rho^2-x^2}, \sqrt{\rho^2-x^2}] : |c\beta y + f(\alpha x) - g(x)| > \rho\beta c/2\},$$

which always is of a length greater $\lambda([-\sqrt{\rho^2-x^2},\sqrt{\rho^2-x^2}]\backslash[-\rho/2,\rho/2])$. Addressing the way a centered cylinder intersects with a shifted ball this gives that

$$\begin{aligned}V(\rho\beta c/2) \geq\ & \lambda\left(C_{\rho,x,y}(0)\cap B_{s_{\mathbf{X}}}(x_0,y_0,0)\right.\\ & \cap\{(x,y,z)\in\mathbb{R}^2\times\mathbb{R}^{p-2};\\ & \left.(x,y)\in\mathbb{R}^2: -\operatorname{sign}(y_0)y\geq-\operatorname{sign}(y_0)\rho/2\}\right)\\ \geq\ & \lambda(B_{\rho/4}(0))>0, \end{aligned} \tag{A.35}$$

for the ball $B_{h/4}(0)\subset\mathbb{R}^p$. Now we can prove the claim. For any $(\boldsymbol{\theta},\boldsymbol{\eta})=\boldsymbol{\upsilon}\in\Upsilon$, with $\|\boldsymbol{\theta}\|=1$, we can represent $\boldsymbol{\theta}^*=\alpha\boldsymbol{\theta}+\beta\boldsymbol{\theta}^\circ$ with some $\boldsymbol{\theta}^\circ\in\boldsymbol{\theta}^\perp$ with $\|\boldsymbol{\theta}^\circ\|=1$ and $\alpha^2+\beta^2=1$. By assumption $(\mathbf{Cond}_{\mathbf{X}\boldsymbol{\theta}^*})$ for any $(\boldsymbol{\theta},\boldsymbol{\eta})=\boldsymbol{\upsilon}\in\Upsilon$, there exist constants $c_{f'}$, $c_{p_{\mathbf{X}}}, h>0$ and a value $(x_0,y_0)\in\{x^2+y^2\leq s_{\mathbf{X}}\}\subset\mathbb{R}^2$ such that for $(x,y)\in\{(x-x_0)^2+(y-y_0)^2\leq h^2\}$ we have $|f'_{\boldsymbol{\eta}^*}(x)|>c_{f'}$ and $p_{\mathbf{X}}\geq c_{p_{\mathbf{X}}}$. We can estimate using (A.35)

$$\begin{aligned}&\mathbb{P}\Big\{\big(\boldsymbol{f}_{\boldsymbol{\eta}^*}(\mathbf{X}^\top\boldsymbol{\theta}^*)-\boldsymbol{f}_{\boldsymbol{\eta}}(\mathbf{X}^\top\boldsymbol{\theta})\big)^2\geq c_{f'}^2h^2\beta^2/4\Big\}\\ &\quad\geq \inf_{\substack{f\in C^1(\mathbb{R}),\,f'>0,\\ g:\mathbb{R}\to\mathbb{R}}}\mathbb{P}\Big(\{\mathbf{X}\in B_{s_{\mathbf{X}}}(0)\}\cap\{\mathbf{X}\in C_{h,x,y}(x_0,y_0)\}\\ &\qquad\qquad\cap\{|f(\alpha x+\beta y)-g(x)|\geq c_{f'}h\beta/2\}\Big)\\ &\quad\geq c_{p_{\mathbf{X}}}\inf_{\substack{f\in C^1(\mathbb{R}),\,f'>0,\\ g:\mathbb{R}\to\mathbb{R}}}\lambda\Big(B_{s_{\mathbf{X}}}(-x_0,-y_0,0)\cap C_{h,x,y}(0)\\ &\qquad\qquad\cap\{|f(\alpha x+\beta y)-g(x)|\geq c_{f'}h\beta/2\}\Big)\\ &\quad= c_{p_{\mathbf{X}}}V(h\beta c_{f'}/2)\geq\lambda(B_{h/4}(0))>0.\end{aligned}$$

We need to express $\beta>0$ in terms of $\mathtt{r}>0$. We can use elementary geometry to obtain

$$\beta=\sin\left(2\arcsin\left(\frac{\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|}{2}\right)\right).$$

Using that $\sin(2\alpha)=2\sin(\alpha)\cos(\alpha)$ this yields

$$\beta=\cos\left(\arcsin\left(\frac{\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|}{2}\right)\right)\frac{\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|}{2}.$$

Now as $\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|^2 \le 2$ we get

$$\beta \ge \cos\left(\arcsin\left(\frac{1}{\sqrt{2}}\right)\right)\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\| = \frac{\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|}{\sqrt{2}}.$$

Furthermore for any $\varphi_{\boldsymbol{\theta}}, \varphi_{\boldsymbol{\theta}} \in W_S$ we have with (A.34) that

$$\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|^2 \ge \frac{2}{p\pi^2}\|\varphi_{\boldsymbol{\theta}}-\varphi_{\boldsymbol{\theta}^*}\|^2 \ge \frac{2}{p\pi^2\|\mathrm{D}^2\|}\|\mathrm{D}(\varphi_{\boldsymbol{\theta}}-\varphi_{\boldsymbol{\theta}^*})\|^2 \ge \frac{\tau}{p\pi^2\|\mathrm{D}^2\|}\mathtt{r}^2.$$

With Lemma A.5 this implies

$$\beta^2 \ge \frac{\tau}{2p\pi^2 s_{\mathbf{X}}^2\|f_{\mathbf{X}}\|_\infty^2 C_{\|\boldsymbol{f}^*\|}}\mathtt{r}^2/n.$$

Combined this yields that with

$$2\mathtt{b}/n \overset{\mathrm{def}}{=} \frac{\tau c_{f'}^2 h^2}{4np\pi^2 s_{\mathbf{X}}^2\|p_{\mathbf{X}}\|_\infty^2 C_{\|\boldsymbol{\eta}^*\|}},$$

it holds

$$\begin{aligned}&\mathbb{P}\Big\{\left(\boldsymbol{f}_{\boldsymbol{\eta}}(\mathbf{X}^\top\boldsymbol{\theta})-\boldsymbol{f}_{\boldsymbol{\eta}^*}(\mathbf{X}^\top\boldsymbol{\theta}^*)\right)^2 \ge 9\mathtt{b}\mathtt{r}^2/n\Big\}\\ &\qquad \ge c_{p_{\mathbf{X}}}\lambda(B_1^{p-2})\lambda\left(B_h(0)\cap\{(x,y)\in\mathbb{R}^2 : |y|\le h/2\}\right).\end{aligned}$$

This gives the claim. □

A.8.8. Proof of Condition $(\boldsymbol{\mathcal{L}}\mathtt{r})$ *with modeling bias*

We show the following Lemma

Lemma A.25. *We have with some* $\mathtt{C}>0$ *and with* $\mathtt{r}^\circ>0$ *from* (5.2) *that*

$$\mathbb{P}\left(\sup_{\boldsymbol{\upsilon}\in\Upsilon_\circ(\sqrt{n}\mathtt{r}^\circ)}|\mathbb{E}_\epsilon\mathcal{L}(\boldsymbol{\upsilon},\boldsymbol{\upsilon}^*)-\mathbb{E}\mathcal{L}(\boldsymbol{\upsilon},\boldsymbol{\upsilon}^*)| \ge \sqrt{\mathtt{x}+p^*[\mathtt{C}\log(p^*)+\log(\mathtt{r})]}\right) \le \mathrm{e}^{-\mathtt{x}}.$$

Proof. We bound

$$\begin{aligned}&\sup_{\boldsymbol{\upsilon}\in\Upsilon_\circ(\sqrt{n}\mathtt{r}^\circ)}|\mathbb{E}_\epsilon\mathcal{L}(\boldsymbol{\upsilon},\boldsymbol{\upsilon}^*)-\mathbb{E}\mathcal{L}(\boldsymbol{\upsilon},\boldsymbol{\upsilon}^*)|\\ &\quad\le n\sup_{\boldsymbol{\upsilon}\in\Upsilon_\circ(\sqrt{n}\mathtt{r}^\circ)}\left|(P_n-\mathbb{P})\Big\{\left(g(\mathbf{X}_i)-\boldsymbol{f}_{\boldsymbol{\eta}^*}(\mathbf{X}_i^\top\boldsymbol{\theta}^*)\right)^2-\left(g(\mathbf{X}_i)-\boldsymbol{f}_{\boldsymbol{\eta}}(\mathbf{X}_i^\top\boldsymbol{\theta})\right)^2\Big\}\right|\\ &\quad\le n\sup_{\boldsymbol{\upsilon}\in\Upsilon_\circ(\sqrt{n}\mathtt{r}^\circ)}\left|(P_n-\mathbb{P})\left\{\boldsymbol{f}_{\boldsymbol{\eta}}(\mathbf{X}_i^\top\boldsymbol{\theta})-\boldsymbol{f}_{\boldsymbol{\eta}^*}(\mathbf{X}_i^\top\boldsymbol{\theta}^*)\right\}^2\right|\\ &\qquad+n\mathtt{C}_{bias}\sup_{\boldsymbol{\upsilon}\in\Upsilon_\circ(\sqrt{n}\mathtt{r}^\circ)}\left|(P_n-\mathbb{P})\left|\boldsymbol{f}_{\boldsymbol{\eta}}(\mathbf{X}_i^\top\boldsymbol{\theta})-\boldsymbol{f}_{\boldsymbol{\eta}^*}(\mathbf{X}_i^\top\boldsymbol{\theta}^*)\right|\right|.\end{aligned}$$

Furthermore

$$\left\{\boldsymbol{f}_{\boldsymbol{\eta}}(\mathbf{X}_i^\top\boldsymbol{\theta}) - \boldsymbol{f}_{\boldsymbol{\eta}^*}(\mathbf{X}_i^\top\boldsymbol{\theta}^*)\right\}^2 \le \left|\boldsymbol{f}_{\boldsymbol{\eta}}(\mathbf{X}_i^\top\boldsymbol{\theta}) - \boldsymbol{f}_{\boldsymbol{\eta}^*}(\mathbf{X}_i^\top\boldsymbol{\theta}^*)\right| \\ \left(\|\boldsymbol{f}_{\boldsymbol{\eta}^*}\|_\infty + \|\boldsymbol{f}_{\boldsymbol{\eta}_m^*}\|_\infty + \mathtt{C}\mathtt{r}\sqrt{m}/\sqrt{n}\right).$$

Thus we have

$$\begin{aligned}&\sup_{\boldsymbol{\upsilon}\in\Upsilon_\circ(\sqrt{n}\mathtt{r}^\circ)} |\mathbb{E}_\epsilon\mathcal{L}(\boldsymbol{\upsilon},\boldsymbol{\upsilon}^*) - \mathbb{E}\mathcal{L}(\boldsymbol{\upsilon},\boldsymbol{\upsilon}^*)| \\ &\quad\le n\left(\mathtt{C}_{bias} + \|\boldsymbol{f}_{\boldsymbol{\eta}^*}\|_\infty + \|\boldsymbol{f}_{\boldsymbol{\eta}_m^*}\|_\infty + \mathtt{C}\mathtt{r}^\circ\sqrt{m}\right) \\ &\qquad\sup_{\boldsymbol{\upsilon}\in\Upsilon_\circ(\sqrt{n}\mathtt{r}^\circ)} \left|(P_n-\mathbb{P})\left|\boldsymbol{f}_{\boldsymbol{\eta}}(\mathbf{X}_i^\top\boldsymbol{\theta}) - \boldsymbol{f}_{\boldsymbol{\eta}^*}(\mathbf{X}_i^\top\boldsymbol{\theta}^*)\right|\right|.\end{aligned}$$

Define $\boldsymbol{\zeta}_{\mathbf{X}}(\boldsymbol{\upsilon}) \overset{\text{def}}{=} (P_n-\mathbb{P})|\boldsymbol{f}_{\boldsymbol{\eta}}(\mathbf{X}_i^\top\boldsymbol{\theta}) - \boldsymbol{f}_{\boldsymbol{\eta}^*}(\mathbf{X}_i^\top\boldsymbol{\theta}^*)|$. Then we find using that $\mathtt{r}^\circ \le \mathtt{C}\sqrt{p^*\log(p^*)+\mathtt{x}}$

$$\sup_{\boldsymbol{\upsilon}\in\Upsilon_\circ(\sqrt{n}\mathtt{r}^\circ)} |\mathbb{E}_\epsilon\mathcal{L}(\boldsymbol{\upsilon},\boldsymbol{\upsilon}^*) - \mathbb{E}\mathcal{L}(\boldsymbol{\upsilon},\boldsymbol{\upsilon}^*)| \le n\mathtt{C}m^{3/2}\sup_{\boldsymbol{\upsilon}\in\Upsilon_\circ(\sqrt{n}\mathtt{r}^\circ)} |\boldsymbol{\zeta}_{\mathbf{X}}(\boldsymbol{\upsilon}) - \boldsymbol{\zeta}_{\mathbf{X}}(\boldsymbol{\upsilon}^*)|.$$

We want to use Lemma A.2. Define $\Upsilon_0 = \{\boldsymbol{\upsilon}^*\}$ and with $\mathtt{r}_k = 2^{-k}\mathtt{r}$ with $\mathtt{r}>0$ to be specified later the sequence of sets Υ_k each with minimal cardinality such that

$$\Upsilon_m \subset \bigcup_{\boldsymbol{\upsilon}\in\Upsilon_k} B_{\mathtt{r}_k}(\boldsymbol{\upsilon}), \quad B_{\mathtt{r}}(\boldsymbol{\upsilon}) \overset{\text{def}}{=} \{\boldsymbol{\upsilon}^\circ\in\Upsilon_m,\, \|\mathcal{D}(\boldsymbol{\upsilon}^\circ-\boldsymbol{\upsilon})\| \le \mathtt{r}\}.$$

We estimate for an application of the bounded differences inequality

$$\left|\left\{\boldsymbol{f}_{\boldsymbol{\eta}}(\mathbf{X}_i^\top\boldsymbol{\theta}) - \boldsymbol{f}_{\boldsymbol{\eta}'}(\mathbf{X}_i^\top\boldsymbol{\theta}')\right\}\right| \le \|\boldsymbol{f}_{\boldsymbol{\eta}-\boldsymbol{\eta}'}\|_\infty + \|\boldsymbol{f}'_{\boldsymbol{\eta}}\|_\infty\|\boldsymbol{\theta}-\boldsymbol{\theta}'\|.$$

We have

$$\begin{aligned}\|\boldsymbol{f}_{\boldsymbol{\eta}}\|_\infty &\le \|\boldsymbol{\eta}\|\sup_{x\in[-s_{\mathbf{X}},s_{\mathbf{X}}]}\left(\sum_{k=1}^m \boldsymbol{e}_k^2(x)^2\right)^{1/2} \le \sqrt{17}\|\psi\|\sqrt{m}\mathtt{r}/\sqrt{n}, \\ \|\boldsymbol{f}'_{\boldsymbol{\eta}-\boldsymbol{\eta}'}\|_\infty &\le \|\boldsymbol{\eta}-\boldsymbol{\eta}'\|\sup_{x\in[-s_{\mathbf{X}},s_{\mathbf{X}}]}\left(\sum_{k=1}^m \boldsymbol{e}_k'^2(x)^2\right)^{1/2} \\ &\le \sqrt{17}\|\psi'\|m^{3/2}\|\boldsymbol{\eta}-\boldsymbol{\eta}'\|.\end{aligned}$$

Consequently again using that $\mathtt{r}^\circ \le \mathtt{C}\sqrt{p^*\log(p^*)+\mathtt{x}}$

$$\left|\left\{\boldsymbol{f}_{\boldsymbol{\eta}}(\mathbf{X}_i^\top\boldsymbol{\theta}) - \boldsymbol{f}_{\boldsymbol{\eta}'}(\mathbf{X}_i^\top\boldsymbol{\theta}')\right\}\right| \le \mathtt{C}_{\boldsymbol{\zeta}}m^{3/2}\|\boldsymbol{\upsilon}-\boldsymbol{\upsilon}'\|.$$

This implies with the bounded difference inequality for any $\boldsymbol{\upsilon}_k \in \varUpsilon_k$

$$\mathbb{P}\left(n \inf_{\varUpsilon_{k-1}} |\boldsymbol{\zeta}_{\mathbf{X}}(\boldsymbol{\upsilon}_k) - \boldsymbol{\zeta}_{\mathbf{X}}(\boldsymbol{\upsilon}_{k-1})| \ge t\mathtt{C}_{\boldsymbol{\zeta}} m^{3/2}\frac{\mathtt{r}_{k-1}}{c_{\mathcal{D}}}\right) \le \mathrm{e}^{-t^2}.$$

Define $\mathtt{r} \overset{\mathrm{def}}{=} \frac{(1-1/\sqrt{2})}{m^3}$ then we find

$$\mathbb{P}\left(n \inf_{\varUpsilon_{k-1}} |\boldsymbol{\zeta}_{\mathbf{X}}(\boldsymbol{\upsilon}_k) - \boldsymbol{\zeta}_{\mathbf{X}}(\boldsymbol{\upsilon}_{k-1})| \ge \mathtt{C} m^{-3/2} t 2^{-(k-1)}(1-1/\sqrt{2})\right) \le \mathrm{e}^{-t^2},$$

$$|\varUpsilon_k| \le \exp\left\{\left(\log(2)k + \log(\mathtt{r}^{\circ}) + \log(n)/2 + 3\log(m) + \log(1-1/\sqrt{2})\right)p^*\right\}.$$

Set

$$\boldsymbol{T}(n,m) \overset{\mathrm{def}}{=} \log(\mathtt{r}^{\circ}) + \log(n)/2 + 3\log(m) + \log(1-1/\sqrt{2},$$
$$t \overset{\mathrm{def}}{=} \sqrt{\mathtt{x} + 1 + \log(2) + p^*\left(\log(2) + \boldsymbol{T}(n,m)\right)},$$

then we infer with Lemma A.2

$$\begin{aligned}
&\mathbb{P}\left(\sup_{\boldsymbol{\upsilon}\in\varUpsilon_{\circ}(\sqrt{n}\mathtt{r}^{\circ})} |\mathbb{E}_{\epsilon}\mathcal{L}(\boldsymbol{\upsilon},\boldsymbol{\upsilon}^*) - \mathbb{E}\mathcal{L}(\boldsymbol{\upsilon},\boldsymbol{\upsilon}^*)| \ge \mathtt{C}t\right)\\
&\quad\le \mathbb{P}\left(n \sup_{\boldsymbol{\upsilon}\in\varUpsilon_{\circ}(\sqrt{n}\mathtt{r}^{\circ})} |\boldsymbol{\zeta}_{\mathbf{X}}(\boldsymbol{\upsilon}) - \boldsymbol{\zeta}_{\mathbf{X}}(\boldsymbol{\upsilon}^*)| \ge \mathtt{C}m\log(m)t\right)\\
&\quad\le \sum_{k=1}^{\infty}\exp\left\{p^*\left[(\log(2)k + \boldsymbol{T}(n,m)) - 2^{k-1}\left(\log(2) + \boldsymbol{T}(n,m)\right)\right]\right.\\
&\qquad\left.-2^{k-1}(\mathtt{x} + 1 + \log(2))\right\}\\
&\quad\le \mathrm{e}^{-\mathtt{x}}.
\end{aligned}$$

□

We have as in the proof of Lemma A.2 of [1]

$$-\mathbb{E}\mathcal{L}(\boldsymbol{\upsilon}^*,\boldsymbol{\upsilon}^*_m) = \mathbb{E}\mathcal{L}(\boldsymbol{\upsilon}^*_m,\boldsymbol{\upsilon}^*) \ge \mathbb{E}\mathcal{L}(\varPi_{p^*}\boldsymbol{\upsilon}^*,\boldsymbol{\upsilon}^*) \ge -\mathtt{r}^{*2}. \qquad \text{(A.36)}$$

Combining this lemma and Equation (A.36) with Lemma A.7 and Lemma 5.2 we find for $\|\mathcal{D}_m(\boldsymbol{\upsilon}-\boldsymbol{\upsilon}^*_m)\|^2 = \mathtt{r}^2 \ge 2\mathtt{r}^{*2}$that with probability greater than $1-2\mathrm{e}^{-\mathtt{x}}$

$$-\mathbb{E}_{\epsilon}\mathcal{L}(\boldsymbol{\upsilon},\boldsymbol{\upsilon}^*_m) \ge \mathtt{b}\mathtt{r}^2/2 - \sqrt{\mathtt{x} + \mathtt{C}p^*[\log(p^*) + \log(n)]} - \mathtt{r}^{*2}.$$

Consequently we get for $\mathtt{r}$ that additionally satisfies

$$\mathtt{r}^2 \ge \sqrt{\mathtt{x} + \mathtt{C}p^*[\log(p^*) + \log(n)]}/\mathtt{b} \vee 2\mathtt{r}^{*2},$$

that

$$-\mathbb{E}_{\epsilon}\mathcal{L}(\boldsymbol{\upsilon},\boldsymbol{\upsilon}^*_m) \ge \mathtt{b}\mathtt{r}^2/4 \overset{\mathrm{def}}{=} \mathtt{b}_{bias}\mathtt{r}^2.$$

Finally observe that by definition $\mathcal{L}(\boldsymbol{\upsilon},\boldsymbol{\upsilon}^*_m) = \mathcal{L}_m(\boldsymbol{\upsilon},\boldsymbol{\upsilon}^*_m)$.

A.9. Proof of Lemma 3.1

Proof. Note that with the definitions and with some $\boldsymbol{\upsilon} \in \Upsilon_{m,0}(\mathtt{r})$, $\boldsymbol{\gamma}_0 \in \mathbb{R}^{p^*}$ with $\|\boldsymbol{\gamma}_0\| = 1$

$$
\begin{aligned}
&\|\mathcal{D}_m^{-1}\nabla(\mathbb{E}-\mathbb{E}_\varepsilon)[\mathcal{L}_m(\boldsymbol{\upsilon}_m^*)-\mathcal{L}_m(\boldsymbol{\upsilon})]\| \\
&\quad\le \sup_{\boldsymbol{\upsilon}\in\Upsilon_{m,0}(\mathtt{r})} \|\mathcal{D}_m^{-1}(\mathbb{E}-\mathbb{E}_\varepsilon)\left[\nabla^2\mathcal{L}_m(\boldsymbol{\upsilon})\right]\mathcal{D}_m^{-1}\|\mathtt{r} \\
&\quad\le \frac{1}{\sqrt{n}c_{\mathcal{D}}}\|(\mathbb{E}-\mathbb{E}_\varepsilon)\left[\mathcal{D}_m^{-1}\nabla^2\mathcal{L}_m(\boldsymbol{\upsilon}_m^*)\right]\|\mathtt{r} \\
&\qquad + \sup_{\boldsymbol{\upsilon}\in\Upsilon_{m,0}(\mathtt{r})} \left\|(\mathbb{E}-\mathbb{E}_\varepsilon)\left[\mathcal{D}_m^{-1}\left(\left[\nabla^2\mathcal{L}_m(\boldsymbol{\upsilon})\right]-\left[\nabla^2\mathcal{L}_m(\boldsymbol{\upsilon}_m^*)\right]\right)\mathcal{D}_m^{-1}\right]\right\|\mathtt{r}.
\end{aligned}
$$

For the first term we obtain with Lemma A.31 and with some constant $\mathtt{C} > 0$

$$
\mathbb{P}\left(\frac{1}{\sqrt{n}c_{\mathcal{D}}}\|(\mathbb{E}-\mathbb{E}_\varepsilon)\left[\mathcal{D}^{-1}\nabla^2\mathcal{L}_m(\boldsymbol{\upsilon}_m^*)\right]\|\mathtt{r} \ge \mathtt{C}\sqrt{\log(p^*)+\mathtt{x}}\mathtt{r}/\sqrt{n}\right) \le \mathrm{e}^{-\mathtt{x}}.
$$

For the second term we can use similar arguments to those of Lemma 5.2 to find with some constant $\mathtt{C} > 0$ that

$$
\begin{aligned}
\mathbb{P}\Bigg(&\sup_{\boldsymbol{\upsilon}\in\Upsilon_{m,0}(\mathtt{r})}\left\|(\mathbb{E}-\mathbb{E}_\varepsilon)\left[\mathcal{D}_m^{-1}\left(\left[\nabla^2\mathcal{L}_m(\boldsymbol{\upsilon})\right]-\left[\nabla^2\mathcal{L}_m(\boldsymbol{\upsilon}_m^*)\right]\right)\mathcal{D}_m^{-1}\right]\right\| \\
&\ge \mathtt{C}\sqrt{\mathtt{x}+p^*\log(p^*)}/\sqrt{n}\Bigg) \le \mathrm{e}^{-\mathtt{x}}.
\end{aligned}
$$

Adding $\log(2)$ to $\mathtt{x}$ in the above bounds we get the claim after increasing the constants appropriately. □

A.10. Condition (*bias*$''$) is satisfied

Lemma A.26. *Under the conditions of Proposition 2.4 condition* (***bias***$''$) *is satisfied.*

Proof. It suffices to show that

$$
\operatorname{Cov}(\nabla_{\boldsymbol{\theta}}\left(\ell_i(\boldsymbol{\upsilon}_m^*)-\ell_i(\boldsymbol{\upsilon}^*)\right)) \to 0, \quad \operatorname{Cov}(\nabla_{(\eta_1,\ldots,\eta_m)}\left(\ell_i(\boldsymbol{\upsilon}_m^*)-\ell_i(\boldsymbol{\upsilon}^*)\right)) \to 0.
$$

We calculate

$$
\begin{aligned}
&\|\operatorname{Cov}(\nabla_{\boldsymbol{\theta}}\left(\ell_i(\boldsymbol{\upsilon}_m^*)-\ell_i(\boldsymbol{\upsilon}^*)\right))\| \\
&\quad\le \mathbb{E}\|\left(\boldsymbol{f}'_{\boldsymbol{\eta}_m^*}(\mathbf{X}_i^\top\boldsymbol{\theta}_m^*)-\boldsymbol{f}'_{\boldsymbol{\eta}^*}(\mathbf{X}_i^\top\boldsymbol{\theta}^*)\right)\nabla\Phi(\boldsymbol{\theta})^\top\mathbf{X}_i\|^2 \\
&\quad\le s_{\mathbf{X}}^2\mathbb{E}\|\boldsymbol{f}'_{\boldsymbol{\eta}_m^*}(\mathbf{X}_i^\top\boldsymbol{\theta}_m^*)-\boldsymbol{f}'_{\boldsymbol{\eta}^*}(\mathbf{X}_i^\top\boldsymbol{\theta}^*)\|^2 \\
&\quad\le 4s_{\mathbf{X}}^2\left(\mathbb{E}\|\boldsymbol{f}'_{\boldsymbol{\eta}_m^*-\boldsymbol{\eta}^*}(\mathbf{X}_i^\top\boldsymbol{\theta}^*)\|^2+\mathbb{E}\|\boldsymbol{f}'_{\boldsymbol{\eta}_m^*}(\mathbf{X}_i^\top\boldsymbol{\theta}_m^*)-\boldsymbol{f}'_{\boldsymbol{\eta}_m^*}(\mathbf{X}_i^\top\boldsymbol{\theta}^*)\|^2\right)
\end{aligned}
$$

$$\le 4 s_{\mathbf{X}}^2 \left(\sum_{k=0}^{\infty} \|\boldsymbol{e}_k'\|_\infty (\eta_{mk}^* - \eta_k^*)\right)^2 + 4 s_{\mathbf{X}}^4 \left(\sum_{k=0}^{m-1} \|\boldsymbol{e}_k''\|_\infty \eta_{mk}^*\right)^2 \|\boldsymbol{\theta}_m^* - \boldsymbol{\theta}^*\|^2.$$

We estimate separately

$$\begin{aligned}
\sum_{k=0}^{\infty} \|\boldsymbol{e}_k'\|_\infty (\eta_{mk}^* - \eta_k^*) &\le \mathtt{C}\|\psi'\|_\infty \left(\sum_{k=0}^{m-1} k^{3/2}(\eta_{mk}^* - \eta_k^*) + \sum_{k=m}^{\infty} k^{3/2}\eta_k^*\right) \\
&\le \mathtt{C}\|\psi'\|_\infty \left(m^2 \|\boldsymbol{\eta}_m^* - \boldsymbol{\eta}^*\| + \left(\sum_{k=m}^{\infty} k^{-2\alpha-3}\right)^{1/2} \left(\sum_{k=m}^{\infty} 2^\alpha \eta_k^{*2}\right)^{1/2} \right) \\
&\le \mathtt{C}\|\psi'\|_\infty \Bigg(m^2 \frac{1}{\sqrt{n} c_{\mathcal{D}}} \|\mathcal{D}_m(\varPi_{p^*}\boldsymbol{\upsilon}^* - \boldsymbol{\upsilon}_m^*)\| \\
&\qquad + \sqrt{(2\alpha-3)/(2\alpha-4)} \left(\sum_{k=m}^{\infty} k^{2\alpha}\eta_k^{*2}\right)^{1/2} \Bigg)
\end{aligned}$$

The last term tends to 0 because of Lemma 5.1, because $m^2 \mathtt{r}^*/\sqrt{n} \to 0$ and because $\sum_k 2^\alpha \eta_k^{*2} < \infty$. Furthermore we get with similar steps

$$\begin{aligned}
\left(\sum_{k=0}^{m-1} \|\boldsymbol{e}_k''\|_\infty \eta_{mk}^*\right) \|\boldsymbol{\theta}_m^* - \boldsymbol{\theta}^*\| &\le \|\psi''\|_\infty \|\boldsymbol{\theta}_m^* - \boldsymbol{\theta}^*\| \left(\sum_{k=0}^{m-1} k^{5/2}\eta_{mk}^*\right) \\
&\le \|\psi''\|_\infty \|\boldsymbol{\theta}_m^* - \boldsymbol{\theta}^*\| \Bigg\{ \left(\sum_{k=0}^{m-1} k^{2\alpha-5}\right)^{1/2} \left(\sum_{k=0}^{m-1} k^{2\alpha}\eta_k^*\right)^{1/2} \\
&\qquad + \frac{1}{\sqrt{n} c_{\mathcal{D}}} \left(\sum_{k=0}^{m-1} k^5\right)^2 \|\mathcal{D}(\boldsymbol{\upsilon}^* - \boldsymbol{\upsilon}_m^*)\| \Bigg\} \\
&\le \|\psi''\|_\infty \|\boldsymbol{\theta}_m^* - \boldsymbol{\theta}^*\| \left\{ m \mathtt{C}_{\|\boldsymbol{\eta}^*\|} + \frac{1}{\sqrt{n} c_{\mathcal{D}}} m^3 \|\mathcal{D}_m(\varPi_{p^*}\boldsymbol{\upsilon}^* - \boldsymbol{\upsilon}_m^*)\| \right\} \\
&\le \|\mathcal{D}_m(\varPi_{p^*}\boldsymbol{\upsilon}^* - \boldsymbol{\upsilon}_m^*)\| \frac{1}{\sqrt{n} c_{\mathcal{D}}} m \mathtt{C}_{\|\boldsymbol{\eta}^*\|} \|\psi''\|_\infty \\
&\qquad + \frac{1}{n c_{\mathcal{D}}^2} m^3 \|\mathcal{D}_m(\varPi_{p^*}\boldsymbol{\upsilon}^* - \boldsymbol{\upsilon}_m^*)\|^2 \|\psi''\|_\infty .
\end{aligned}$$

Again the last term tends to 0. Similarly we calculate

$$\begin{aligned}
\mathrm{Cov}(\nabla_{(\eta_1,\ldots,\eta_m)} (\ell_i(\boldsymbol{\upsilon}_m^*) - \ell_i(\boldsymbol{\upsilon}^*))) &\le \mathbb{E}\|\boldsymbol{e}(\mathbf{X}_i^\top \boldsymbol{\theta}_m^*) - \boldsymbol{e}(\mathbf{X}_i^\top \boldsymbol{\theta}^*)\|^2 \\
&\le s_{\mathbf{X}}^2 \|\psi'\|_\infty^2 \|\boldsymbol{\theta}_m^* - \boldsymbol{\theta}^*\|^2 \left(\sum_{k=0}^{m-1} k^{3/2}\right)^2 \\
&\le s_{\mathbf{X}}^2 \|\psi'\|_\infty \frac{1}{n c_{\mathcal{D}}} m^3 \|\mathcal{D}_m(\varPi_{p^*}\boldsymbol{\upsilon}^* - \boldsymbol{\upsilon}_m^*)\|^2,
\end{aligned}$$

which again is a zero sequence. This gives the claim. □

A.11. Proof of Lemma 5.6

Proof. Define

$$\boldsymbol{\theta}_{l^*} \overset{\text{def}}{=} \operatorname*{argmin}_{\boldsymbol{\theta}\in G_N} \|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|.$$

Then by definition

$$\begin{aligned}\max_{\boldsymbol{\eta}} \mathcal{L}_m(\widetilde{\boldsymbol{\upsilon}}^{(0)},\boldsymbol{\upsilon}_m^*) &\geq \mathcal{L}_m((\boldsymbol{\theta}_{l^*},\widetilde{\boldsymbol{\eta}}_{l^*}^{(0)}),\boldsymbol{\upsilon}_m^*) \geq \mathcal{L}_m((\boldsymbol{\theta}_{l^*},\boldsymbol{\eta}_m^*),\boldsymbol{\upsilon}_m^*)\\ &= -\sum_{i=1}^n (\boldsymbol{f}_{\boldsymbol{\eta}_m^*}(\mathbf{X}_i^\top\boldsymbol{\theta}_m^*)-\boldsymbol{f}_{\boldsymbol{\eta}_m^*}(\mathbf{X}_i^\top\boldsymbol{\theta}_{l^*}))^2\\ &\quad +(g(\mathbf{X}_i)-\boldsymbol{f}_{\boldsymbol{\eta}_m^*}(\mathbf{X}_i^\top\boldsymbol{\theta}_m^*))(\boldsymbol{f}_{\boldsymbol{\eta}_m^*}(\mathbf{X}_i^\top\boldsymbol{\theta}_m^*)-\boldsymbol{f}_{\boldsymbol{\eta}_m^*}(\mathbf{X}_i^\top\boldsymbol{\theta}_{l^*}))\\ &\quad -(\boldsymbol{f}_{\boldsymbol{\eta}_m^*}(\mathbf{X}_i^\top\boldsymbol{\theta}_m^*)-\boldsymbol{f}_{\boldsymbol{\eta}_m^*}(\mathbf{X}_i^\top\boldsymbol{\theta}_{l^*}))\varepsilon_i.\end{aligned}$$

We estimate using the smoothness of $\boldsymbol{f}_{\boldsymbol{\eta}^*}$

$$|\boldsymbol{f}_{\boldsymbol{\eta}_m^*}(\mathbf{X}_i^\top\boldsymbol{\theta}_m^*)-\boldsymbol{f}_{\boldsymbol{\eta}^*}(\mathbf{X}_i^\top\boldsymbol{\theta}_{l^*})| \leq \mathtt{C}s_{\mathbf{X}}\|\boldsymbol{\theta}_{l^*}-\boldsymbol{\theta}_m^*\| \leq \mathtt{C}s_{\mathbf{X}}\tau.$$

Furthermore the first order criteria of maximality give for some $\boldsymbol{\theta}^\circ \in {\boldsymbol{\theta}_m^*}^\perp$

$$\mathbb{E}\left[(g(\mathbf{X}_i)-\boldsymbol{f}_{\boldsymbol{\eta}_m^*}(\mathbf{X}_i^\top\boldsymbol{\theta}_m^*))\boldsymbol{f}'_{\boldsymbol{\eta}_m^*}(\mathbf{X}_i^\top\boldsymbol{\theta}_m^*)\mathbf{X}^\top\boldsymbol{\theta}^\circ\right]=0,$$

We estimate with Taylor expansion

$$\begin{aligned}&\left\|(\boldsymbol{f}_{\boldsymbol{\eta}_m^*}(\mathbf{X}_i^\top\boldsymbol{\theta}_m^*)-\boldsymbol{f}_{\boldsymbol{\eta}_m^*}(\mathbf{X}_i^\top\boldsymbol{\theta}_{l^*}))-\boldsymbol{f}'_{\boldsymbol{\eta}_m^*}(\mathbf{X}_i^\top\boldsymbol{\theta}_m^*)\mathbf{X}^\top\nabla\varPhi_{\boldsymbol{\theta}_m^*}(\varphi_{\boldsymbol{\theta}_{l^*}}-\varphi_{\boldsymbol{\theta}_m^*})\right\|\\ &\quad\leq \mathtt{C}\sqrt{m}\|\boldsymbol{\theta}_{l^*}-\boldsymbol{\theta}_m^*\|.\end{aligned}$$

Furthermore with the bounded differences inequality

$$\begin{aligned}\mathbb{P}\Big(n\Big|(P_n-\mathbb{P})(g(\mathbf{X}_i)-\boldsymbol{f}_{\boldsymbol{\eta}_m^*}(\mathbf{X}_i^\top\boldsymbol{\theta}_m^*))(\boldsymbol{f}_{\boldsymbol{\eta}_m^*}(\mathbf{X}_i^\top\boldsymbol{\theta}_m^*)-\boldsymbol{f}_{\boldsymbol{\eta}_m^*}(\mathbf{X}_i^\top\boldsymbol{\theta}_{l^*}))\Big|&\\ \geq \sqrt{\mathtt{x}}\mathtt{C}_{bias}\mathtt{C}s_{\mathbf{X}}\tau\Big) \leq \mathrm{e}^{-\mathtt{x}}.&\end{aligned}$$

Consequently with probability greater than $1-\mathrm{e}^{-\mathtt{x}}$

$$\begin{aligned}\mathcal{L}(\widetilde{\boldsymbol{\upsilon}}^{(0)},\boldsymbol{\upsilon}^*) &\geq -n\mathtt{C}^2 s_{\mathbf{X}}^2\tau^2-\mathtt{C}_{bias}\mathtt{C}\left(s_{\mathbf{X}}\tau\sqrt{\mathtt{x}}+n\sqrt{m}\tau^2\right)\\ &\quad+\sum_{i=1}^n(\boldsymbol{f}_{\boldsymbol{\eta}_m^*}(\mathbf{X}_i^\top\boldsymbol{\theta}_m^*)-\boldsymbol{f}_{\boldsymbol{\eta}_m^*}(\mathbf{X}_i^\top\boldsymbol{\theta}_{l^*}))\varepsilon_i.\end{aligned}$$

Clearly we have due to $(\mathbf{Cond}_{\varepsilon})$ for $\lambda \leq \sqrt{n}\widetilde{\mathtt{g}}/(\mathtt{C}s_{\mathbf{X}}\tau)$

$$\mathbb{P}^{\varepsilon}\left(\sum_{i=1}^n(\boldsymbol{f}_{\boldsymbol{\eta}_m^*}(\mathbf{X}_i^\top\boldsymbol{\theta}_m^*)-\boldsymbol{f}_{\boldsymbol{\eta}_m^*}(\mathbf{X}_i^\top\boldsymbol{\theta}_{l^*}))\varepsilon_i \geq \sqrt{n}t\right)$$

$$\begin{aligned}
&\leq \exp\{-\lambda t\}\mathbb{E}^{\varepsilon}\left[\exp\{\lambda \sum_{i=1}^{n}(\boldsymbol{f}_{\boldsymbol{\eta}_m^*}(\mathbf{X}_i^\top \boldsymbol{\theta}_m^*) - \boldsymbol{f}_{\boldsymbol{\eta}_m^*}(\mathbf{X}_i^\top \boldsymbol{\theta}_{l^*}))\varepsilon_i/\sqrt{n}\}\right] \\
&\leq \exp\{-\lambda t\}\prod_{i=1}^{n}\mathbb{E}^{\varepsilon}\left[\exp\{\lambda(\boldsymbol{f}_{\boldsymbol{\eta}_m^*}(\mathbf{X}_i^\top \boldsymbol{\theta}_m^*) - \boldsymbol{f}_{\boldsymbol{\eta}_m^*}(\mathbf{X}_i^\top \boldsymbol{\theta}_{l^*}))\varepsilon_i/\sqrt{n}\}\right] \\
&\leq \exp\{-\lambda t + \widetilde{\nu}^2 \mathtt{C}^2 s_{\mathbf{X}}^2 \tau^2 \lambda^2/2\}.
\end{aligned}$$

Setting $\lambda = \frac{t}{\widetilde{\nu}^2 \mathtt{C}^2 s_{\mathbf{X}}^2 \tau^2}$ we get

$$\mathbb{P}^{\varepsilon}\left(\sum_{i=1}^{n}(\boldsymbol{f}_{\boldsymbol{\eta}_m^*}(\mathbf{X}_i^\top \boldsymbol{\theta}_m^*) - \boldsymbol{f}_{\boldsymbol{\eta}_m^*}(\mathbf{X}_i^\top \boldsymbol{\theta}_{l^*}))\varepsilon_i \geq \sqrt{n}t\right) \leq \exp\left\{-\frac{t^2}{2\widetilde{\nu}^2 \mathtt{C}^2 s_{\mathbf{X}}^2 \tau^2}\right\}.$$

With $t = \widetilde{\nu}\mathtt{C}s_{\mathbf{X}}\tau\sqrt{2\mathtt{x}}$ and $\mathtt{x} \leq 2\widetilde{\nu}^2\widetilde{\mathtt{g}}^2 n/(\mathtt{C}^2 s_{\mathbf{X}}^2\tau^2)$ this gives

$$\mathbb{P}^{\varepsilon}\left(\sum_{i=1}^{n}(\boldsymbol{f}_{\boldsymbol{\eta}_m^*}(\mathbf{X}_i^\top \boldsymbol{\theta}_m^*) - \boldsymbol{f}_{\boldsymbol{\eta}_m^*}(\mathbf{X}_i^\top \boldsymbol{\theta}_{l^*}))\varepsilon_i \geq \widetilde{\nu}\mathtt{C}s_{\mathbf{X}}\tau\sqrt{2n\mathtt{x}}\right) \leq \mathrm{e}^{-\mathtt{x}}.$$

Consequently

$$\mathbb{P}\left(\mathcal{L}_m(\widetilde{\boldsymbol{\upsilon}}^{(0)}, \boldsymbol{\upsilon}_m^*) \geq -\mathtt{C}\left\{(1+\mathtt{C}_{bias}\sqrt{m})n\tau^2 + (1+\mathtt{C}_{bias})\sqrt{\mathtt{x}}\tau\sqrt{n}\right\}\right) \leq 2\mathrm{e}^{-\mathtt{x}}.$$

For the second claim note that by Lemma 5.3 the conditions $(\mathcal{ED}_1)$ and $(\mathcal{L}_0)$ from Section 4.1 hold for all $\mathtt{r} \leq \sqrt{n}\mathtt{r}^{\circ}$. We define

$$\mathrm{K}_0(\mathtt{x}) \stackrel{\text{def}}{=} \mathtt{C}\left\{(1+\mathtt{C}_{bias}\sqrt{m})n\tau^2 + (1+\mathtt{C}_{bias})\sqrt{\mathtt{x}}\tau\sqrt{n}\right\}.$$

This implies with Lemma 5.3 and Theorem 4.2 that

$$\begin{aligned}
\mathrm{R}_0(\mathtt{x}) &\leq \mathtt{C}m^{3/2}\sqrt{p^*(1+\mathtt{C}_{bias}\log(n)) + \mathtt{x} + (1+\mathtt{C}_{bias}\sqrt{m})n\tau^2 + \sqrt{n}\tau\sqrt{\mathtt{x}}} \\
&\leq \mathtt{C}m^{3/2}\sqrt{p^*(1+\mathtt{C}_{bias}\log(n)) + \mathtt{x}} \\
&\quad + \mathtt{C}m^{3/2}\sqrt{(1+\mathtt{C}_{bias}\sqrt{m})n\tau^2 + \sqrt{n}\tau\sqrt{\mathtt{x}}}.
\end{aligned}$$

We use that $\tau = o(m^{-3/2})$ if $\mathtt{C}_{bias} = 0$ and $\tau = o(m^{-11/4})$ if $\mathtt{C}_{bias} > 0$ to find

$$\mathrm{R}_0(\mathtt{x}) \leq \mathtt{C}m^{3/2}\sqrt{p^*(1+\mathtt{C}_{bias}\log(n)) + \mathtt{x}} + \mathtt{C}(\sqrt{n} + m^{1/2}n^{1/4}).$$

Repeating the same arguments as in Section 5.2 we can infer that with probability greater than $1 - 2\mathrm{e}^{-\mathtt{x}}$ the sequence satisfies $(\boldsymbol{\upsilon}_{k,k(+1)}) \subset \varUpsilon_{\circ}(\mathrm{R}_0)$ where

$$\mathrm{R}_0(\mathtt{x}) \leq \mathtt{C}\sqrt{p^*(1+\mathtt{C}_{bias}\log(n)) + \mathtt{x} + (1+\mathtt{C}_{bias}\sqrt{m})n\tau^2 + \sqrt{n}\tau\sqrt{\mathtt{x}}}.$$

Furthermore with Lemma 5.3

$$\epsilon \stackrel{\text{def}}{=} \delta(\mathtt{r})/\mathtt{r} + \omega = \mathtt{C}\frac{m^{3/2} + \mathtt{C}_{bias}m^{5/2}}{\sqrt{n}}.$$

Consequently for moderate $\mathtt{x}$ we find if $\mathtt{C}_{bias} = 0$ that

$$\epsilon \mathrm{R}_0(\mathtt{x}) = O\left(m^{3/2}/\sqrt{n}\right) O\left(\tau\sqrt{n} + \sqrt{\tau} n^{1/4}\right) + O(m^2/\sqrt{n}),$$

such that $\epsilon \mathrm{R}_0(\mathtt{x}) \to 0$ if $\tau = o(m^{-3/2})$. While $\epsilon\sqrt{\mathfrak{z}(\mathtt{x})} = O(m^2/\sqrt{n}) \to 0$. If $\mathtt{C}_{bias} > 0$ we find

$$\epsilon \mathrm{R}_0(\mathtt{x}) = O\left(m^{5/2}/\sqrt{n}\right) O\left(\tau m^{1/4}\sqrt{n} + \sqrt{\tau} n^{1/4}\right) + O(m^3 \log(n)/\sqrt{n}),$$

such that it suffices to ensure that $\tau = o(m^{-11/4})$ since then $m^{5/2}\sqrt{\tau} n^{-1/4} = o(m^{-3/8}) \to 0$, due to $n \ge O(m^6 \log(n)^2)$. In this case $\epsilon\sqrt{\mathfrak{z}(\mathtt{x})} = O(m^6/\sqrt{n}) \to 0$. This gives (A_3) and completes the proof. □

A.12. Proof of Lemma 5.7

A.12.1. Auxiliary results

First we need the following uniform bounds:

Lemma A.27. *There is a generic constant* $\mathtt{C} > 0$ *such that for any pair* $\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^\circ \in \Upsilon_\circ(R_0)$ *with* $\varsigma_{i,m}$ *from* (A.1)

$$\|\nabla \varsigma_{i,m}(\boldsymbol{\upsilon}^*)\| \le \mathtt{C}(\|\boldsymbol{f}'_{\boldsymbol{\eta}^*}\|_\infty + \|\boldsymbol{f}''_{\boldsymbol{\eta}^*}\|_\infty), \tag{A.37}$$

$$\begin{aligned} &\|\mathcal{D}_m^{-1/2}\nabla\varsigma_{i,m}(\boldsymbol{\upsilon}) - \mathcal{D}_m^{-1/2}\nabla\varsigma_{i,m}(\boldsymbol{\upsilon}^\circ)\| \\ &\quad \le \frac{\mathtt{C}}{c_{\mathcal{D}}\sqrt{n}} m \left(m^{3/2} + \left(\mathtt{C} + \frac{m^2(R_0 + \mathtt{r}^*)}{n c_{\mathcal{D}}^2}\right) m^{1/2}\right) \|\mathcal{D}_m(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^\circ)\|. \end{aligned} \tag{A.38}$$

Proof. Since $\nabla^2_{\boldsymbol{\eta}}\zeta(\boldsymbol{\upsilon}) = 0$ we can estimate with help of Lemma A.8

$$\|\nabla\varsigma_{i,m}(\boldsymbol{\upsilon}^*)\| \le \|\nabla_{\boldsymbol{\theta}}\varsigma_{i,m}(\boldsymbol{\upsilon}^*)\| + \|\nabla_{\boldsymbol{\eta}}\varsigma_{i,m}(\boldsymbol{\upsilon}^*)\|.$$

We estimate separately

$$\begin{aligned} \|\nabla_{\boldsymbol{\theta}}\varsigma_{i,m}(\boldsymbol{\upsilon}^*)\| \le\ & \|\boldsymbol{f}''_{\boldsymbol{\eta}^*}(\mathbf{X}_i^\top\boldsymbol{\theta}^*)\nabla\Phi(\boldsymbol{\theta}^*)^\top\mathbf{X}_i\mathbf{X}_i^\top\nabla\Phi(\boldsymbol{\theta}^*)\| \\ & +\|\boldsymbol{f}'_{\boldsymbol{\eta}^*}(\mathbf{X}_i^\top\boldsymbol{\theta}^*)\mathbf{X}_i\nabla^2\Phi(\boldsymbol{\theta}^\top\mathbf{X}_i)[\mathbf{X}_i,\cdot,\cdot]\| \\ \le\ & \mathtt{C}_0 s_{\mathbf{X}}^2\left(|\boldsymbol{f}'_{\boldsymbol{\eta}^*}(\mathbf{X}_i^\top\boldsymbol{\theta})| + |\boldsymbol{f}''_{\boldsymbol{\eta}^*}(\mathbf{X}_i^\top\boldsymbol{\theta})|\right) \le \mathtt{C}(\|\boldsymbol{f}'_{\boldsymbol{\eta}^*}\|_\infty + \|\boldsymbol{f}''_{\boldsymbol{\eta}^*}\|_\infty), \end{aligned}$$

This gives (A.37). For the proof of (A.38) we again use $\nabla^2_{\boldsymbol{\eta}}\zeta(\boldsymbol{\upsilon}) = 0$ and estimate with help of Lemma A.8

$$\begin{aligned} &\|\mathcal{D}_m^{-1/2}\nabla\varsigma_{i,m}(\boldsymbol{\upsilon}) - \mathcal{D}_m^{-1/2}\nabla\varsigma_{i,m}(\boldsymbol{\upsilon}^\circ)\| \le \frac{1}{c_{\mathcal{D}}\sqrt{n}}\|\nabla\varsigma_{i,m}(\boldsymbol{\upsilon}) - \nabla\varsigma_{i,m}(\boldsymbol{\upsilon}^\circ)\| \\ &\quad \le \frac{1}{c_{\mathcal{D}}\sqrt{n}}\left(\|\nabla_{\boldsymbol{\theta}}\varsigma_{i,m}(\boldsymbol{\upsilon}) - \nabla_{\boldsymbol{\theta}}\varsigma_{i,m}(\boldsymbol{\upsilon}^\circ)\| + 2\|\nabla_{\boldsymbol{\eta}}\varsigma_{i,m}(\boldsymbol{\upsilon}) - \nabla_{\boldsymbol{\eta}}\varsigma_{i,m}(\boldsymbol{\upsilon}^\circ)\|\right). \end{aligned}$$

We calculate separately

$$
\begin{aligned}
&\|\nabla_{\boldsymbol{\theta}}\varsigma_{i,m}(\boldsymbol{\upsilon}) - \nabla_{\boldsymbol{\theta}}\varsigma_{i,m}(\boldsymbol{\upsilon}^{\circ})\| \\
&\quad \leq s_{\mathbf{X}}^2 \|\boldsymbol{f}_{\boldsymbol{\eta}}''(\mathbf{X}_i^\top\boldsymbol{\theta})\nabla\varPhi(\boldsymbol{\theta})\nabla\varPhi(\boldsymbol{\theta})^\top - \boldsymbol{f}_{\boldsymbol{\eta}^{\circ}}''(\mathbf{X}_i^\top\boldsymbol{\theta}^{\circ})\nabla\varPhi(\boldsymbol{\theta}^{\circ\top}\mathbf{X}_i)\nabla\varPhi(\boldsymbol{\theta}^{\circ\top}\mathbf{X}_i)^\top\| \\
&\qquad + s_{\mathbf{X}}^2\|\boldsymbol{f}_{\boldsymbol{\eta}}'(\mathbf{X}_i^\top\boldsymbol{\theta})\mathbf{X}_i^\top\nabla^2\varPhi(\boldsymbol{\theta}^\top\mathbf{X}_i) - \boldsymbol{f}_{\boldsymbol{\eta}^{\circ}}'(\mathbf{X}_i^\top\boldsymbol{\theta}^{\circ})\mathbf{X}_i^\top\nabla^2\varPhi(\boldsymbol{\theta}^{\circ\top}\mathbf{X}_i)\|.
\end{aligned}
$$

We again separately estimate

$$
\begin{aligned}
&\|\boldsymbol{f}_{\boldsymbol{\eta}}''(\mathbf{X}_i^\top\boldsymbol{\theta})\nabla\varPhi(\boldsymbol{\theta})\nabla\varPhi(\boldsymbol{\theta})^\top - \boldsymbol{f}_{\boldsymbol{\eta}^{\circ}}''(\mathbf{X}_i^\top\boldsymbol{\theta}^{\circ})\nabla\varPhi(\boldsymbol{\theta}^{\circ\top}\mathbf{X}_i)\nabla\varPhi(\boldsymbol{\theta}^{\circ\top}\mathbf{X}_i)^\top\| \\
&\quad \leq \|[\boldsymbol{f}_{\boldsymbol{\eta}}''(\mathbf{X}_i^\top\boldsymbol{\theta}) - \boldsymbol{f}_{\boldsymbol{\eta}^{\circ}}''(\mathbf{X}_i^\top\boldsymbol{\theta}^{\circ})]\nabla\varPhi(\boldsymbol{\theta})\nabla\varPhi(\boldsymbol{\theta})^\top\| \\
&\qquad + \|\boldsymbol{f}_{\boldsymbol{\eta}^{\circ}}''(\mathbf{X}_i^\top\boldsymbol{\theta}^{\circ})[\nabla\varPhi(\boldsymbol{\theta}) - \nabla\varPhi(\boldsymbol{\theta}^{\circ\top}\mathbf{X}_i)]\nabla\varPhi(\boldsymbol{\theta})^\top\| \\
&\qquad + \|\boldsymbol{f}_{\boldsymbol{\eta}^{\circ}}''(\mathbf{X}_i^\top\boldsymbol{\theta}^{\circ})\nabla\varPhi(\boldsymbol{\theta}^{\circ\top}\mathbf{X}_i)[\nabla\varPhi(\boldsymbol{\theta})^\top - \nabla\varPhi(\boldsymbol{\theta}^{\circ\top}\mathbf{X}_i)^\top]\|.
\end{aligned}
$$

We estimate using that $\|\nabla\varPhi(\boldsymbol{\theta})\nabla\varPhi(\boldsymbol{\theta})^\top\| \leq 1$

$$
\begin{aligned}
&\|[\boldsymbol{f}_{\boldsymbol{\eta}}''(\mathbf{X}_i^\top\boldsymbol{\theta}) - \boldsymbol{f}_{\boldsymbol{\eta}^{\circ}}''(\mathbf{X}_i^\top\boldsymbol{\theta}^{\circ})]\nabla\varPhi(\boldsymbol{\theta})\nabla\varPhi(\boldsymbol{\theta})^\top\| \\
&\quad \leq \|\boldsymbol{f}_{\boldsymbol{\eta}}''(\mathbf{X}_i^\top\boldsymbol{\theta}) - \boldsymbol{f}_{\boldsymbol{\eta}^{\circ}}''(\mathbf{X}_i^\top\boldsymbol{\theta}^{\circ})\| \\
&\quad \leq \|\boldsymbol{f}_{\boldsymbol{\eta}}''(\mathbf{X}_i^\top\boldsymbol{\theta}) - \boldsymbol{f}_{\boldsymbol{\eta}^{\circ}}''(\mathbf{X}_i^\top\boldsymbol{\theta})\| + \|\boldsymbol{f}_{\boldsymbol{\eta}^{\circ}}''(\mathbf{X}_i^\top\boldsymbol{\theta}) - \boldsymbol{f}_{\boldsymbol{\eta}^{\circ}}''(\mathbf{X}_i^\top\boldsymbol{\theta}^{\circ})\|.
\end{aligned}
$$

Remember that due to the structure of the basis

$$
\begin{aligned}
|N(j)| &\overset{\mathrm{def}}{=} \Big|\Big\{k \in \{(2^j-1)17, \ldots, 2^{j+1} + (j+1)17 - 1 - 1\} : \\
&\quad |\boldsymbol{e}_k'(\mathbf{X}_i^\top\boldsymbol{\theta}') - \boldsymbol{e}_k'(\mathbf{X}_i^\top\boldsymbol{\theta})| \vee |\boldsymbol{e}_k''(\mathbf{X}_i^\top\boldsymbol{\theta}') - \boldsymbol{e}_k''(\mathbf{X}_i^\top\boldsymbol{\theta})| \vee |\boldsymbol{e}_k'(\mathbf{X}_i^\top\boldsymbol{\theta})| > 0\Big\}\Big| \\
&\quad \leq 34.
\end{aligned}
$$

We get with the same arguments as in the proof of Lemma A.11

$$
\begin{aligned}
&\|[\boldsymbol{f}_{\boldsymbol{\eta}}''(\mathbf{X}_i^\top\boldsymbol{\theta}) - \boldsymbol{f}_{\boldsymbol{\eta}^{\circ}}''(\mathbf{X}_i^\top\boldsymbol{\theta}^{\circ})]\nabla\varPhi(\boldsymbol{\theta})\nabla\varPhi(\boldsymbol{\theta})^\top\| \\
&\quad \leq \frac{\sqrt{34}}{\sqrt{n}c_{\mathcal{D}}}\left(\|\psi''\|m^{5/2} + \|\psi'''\|\left(\mathtt{C} + \frac{m^2(\mathrm{R}_0 + \mathtt{r}^*)}{\sqrt{n}c_{\mathcal{D}}}\right)m^{3/2}\right) \\
&\qquad \|\mathcal{D}_m(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^{\circ})\|.
\end{aligned}
$$

For the other two summands we estimate

$$
\begin{aligned}
&\|\boldsymbol{f}_{\boldsymbol{\eta}^{\circ}}''(\mathbf{X}_i^\top\boldsymbol{\theta}^{\circ})[\nabla\varPhi(\boldsymbol{\theta}) - \nabla\varPhi(\boldsymbol{\theta}^{\circ\top}\mathbf{X}_i)]\nabla\varPhi(\boldsymbol{\theta})^\top\| \\
&\quad \leq \|\boldsymbol{f}_{\boldsymbol{\eta}^{\circ}}''(\mathbf{X}_i^\top\boldsymbol{\theta}^{\circ})\|\left\|\left\{\nabla\varPhi(\boldsymbol{\theta}) - \nabla\varPhi(\boldsymbol{\theta}^{\circ\top}\mathbf{X}_i)\right\}\nabla\varPhi(\boldsymbol{\theta}^{\circ\top}\mathbf{X}_i)\right\|.
\end{aligned}
$$

We can use the smoothness of $\phi : \mathbb{R}^{p-1} \to S_1 \subset \mathbb{R}^p$ to find a constant $\mathtt{C}_1$ such that

$$\begin{aligned}
&\|\boldsymbol{f}''_{\boldsymbol{\eta}^\circ}(\mathbf{X}_i^\top \boldsymbol{\theta}^\circ)[\nabla \varPhi(\boldsymbol{\theta}) - \nabla \varPhi({\boldsymbol{\theta}^\circ}^\top \mathbf{X}_i)]\nabla \varPhi(\boldsymbol{\theta})^\top\| \\
&\quad \le \|\boldsymbol{f}''_{\boldsymbol{\eta}^\circ}(\mathbf{X}_i^\top \boldsymbol{\theta}^\circ)\|\mathtt{C}_2\|\boldsymbol{\theta} - \boldsymbol{\theta}^\circ\| \\
&\quad \le \mathtt{C}_1\|\boldsymbol{\theta} - \boldsymbol{\theta}^\circ\|\|\psi''\| \sum_{j=0}^{j_m-1} \sum_{k\in N(j)} \eta_k^\circ 2^{5j/2} \\
&\quad \le 17\mathtt{C}_1\|\boldsymbol{\theta} - \boldsymbol{\theta}^\circ\|\|\psi''\| \left(\mathtt{C} + \frac{m^2(\mathrm{R}_0 + \mathtt{r}^*)}{\sqrt{n}c_{\mathcal{D}}}\right) m^{1/2}.
\end{aligned}$$

We continue with

$$\begin{aligned}
&\|\boldsymbol{f}'_{\boldsymbol{\eta}}(\mathbf{X}_i^\top \boldsymbol{\theta})\mathbf{X}_i^\top \nabla^2 \varPhi(\boldsymbol{\theta}^\top \mathbf{X}_i) - \boldsymbol{f}'_{\boldsymbol{\eta}^\circ}(\mathbf{X}_i^\top \boldsymbol{\theta}^\circ)\mathbf{X}_i^\top \nabla^2 \varPhi({\boldsymbol{\theta}^\circ}^\top \mathbf{X}_i)\| \\
&\quad \le \|\boldsymbol{f}'_{\boldsymbol{\eta}}(\mathbf{X}_i^\top \boldsymbol{\theta}) - \boldsymbol{f}'_{\boldsymbol{\eta}^\circ}(\mathbf{X}_i^\top \boldsymbol{\theta}^\circ)\|\|\mathbf{X}_i^\top \nabla^2 \varPhi({\boldsymbol{\theta}^\circ}^\top \mathbf{X}_i)\| \\
&\qquad + \|\boldsymbol{f}'_{\boldsymbol{\eta}}(\mathbf{X}_i^\top \boldsymbol{\theta})\|\mathbf{X}_i^\top \nabla^2 \varPhi(\boldsymbol{\theta}^\top \mathbf{X}_i) - \mathbf{X}_i^\top \nabla^2 \varPhi({\boldsymbol{\theta}^\circ}^\top \mathbf{X}_i)\|.
\end{aligned}$$

Using the smoothness of $\phi : \mathbb{R}^{p-1} \to S_1 \subset \mathbb{R}^p$ we find constants $\mathtt{C}_2, \mathtt{C}_3$ such that with the same argument as in the proof of Lemma A.11

$$\begin{aligned}
&\|\boldsymbol{f}'_{\boldsymbol{\eta}}(\mathbf{X}_i^\top \boldsymbol{\theta})\mathbf{X}_i^\top \nabla^2 \varPhi(\boldsymbol{\theta}^\top \mathbf{X}_i) - \boldsymbol{f}'_{\boldsymbol{\eta}^\circ}(\mathbf{X}_i^\top \boldsymbol{\theta}^\circ)\mathbf{X}_i^\top \nabla^2 \varPhi({\boldsymbol{\theta}^\circ}^\top \mathbf{X}_i)\| \\
&\quad \le \frac{\sqrt{34}}{\sqrt{n}c_{\mathcal{D}}} m^{1/2} \left(\mathtt{C}_2 s_{\mathbf{X}}\|\psi''\| + \|\psi'\| + s_{\mathbf{X}}^2 \mathtt{C}_3\right) \left(\mathtt{C} + \frac{m^2(\mathrm{R}_0 + \mathtt{r}^*)}{\sqrt{n}c_{\mathcal{D}}}\right) \\
&\qquad \|\mathcal{D}_m(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^\circ)\|.
\end{aligned}$$

Finally

$$\begin{aligned}
&\|\nabla_{\boldsymbol{\eta}} \varsigma_{i,m}(\boldsymbol{\upsilon}) - \nabla_{\boldsymbol{\eta}} \varsigma_{i,m}(\boldsymbol{\upsilon}^\circ)\| \\
&\quad \le \|\left(\nabla \varPhi(\boldsymbol{\theta})^\top - \nabla \varPhi({\boldsymbol{\theta}^\circ}^\top \mathbf{X}_i)^\top\right)\mathbf{X}_i\|\|\boldsymbol{e}'(\boldsymbol{\theta}^\top \mathbf{X}_i)^\top\| \\
&\qquad + \|\nabla \varPhi({\boldsymbol{\theta}^\circ}^\top \mathbf{X}_i)^\top \mathbf{X}_i\|\|\boldsymbol{e}'(\boldsymbol{\theta}^\top \mathbf{X}_i) - \boldsymbol{e}'({\boldsymbol{\theta}^\circ}^\top \mathbf{X}_i)\|.
\end{aligned}$$

We estimate separately

$$\|\left(\nabla \varPhi(\boldsymbol{\theta})^\top - \nabla \varPhi(\boldsymbol{\theta}^\circ)^\top\right)\mathbf{X}_i\| \le \mathtt{C}_4 s_{\mathbf{X}}^2 \frac{1}{\sqrt{n}c_{\mathcal{D}}}\|\mathcal{D}_m(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^\circ)\|,$$

$$\|\boldsymbol{e}'(\boldsymbol{\theta}^\top \mathbf{X}_i)^\top\| \le \|\psi'\|_\infty \left(\sum_{j=0}^{j_m} 2^{3j}|N(j)|\right)^{1/2} \le \|\psi'\|_\infty \sqrt{34} m^{3/2}.$$

Furthermore

$$\|\nabla\Phi(\boldsymbol{\theta}^{\circ\top}\mathbf{X}_i)^\top\mathbf{X}_i\| \le \mathtt{C}_5 s_{\mathbf{X}},$$
$$\|\boldsymbol{e}'(\boldsymbol{\theta}^\top\mathbf{X}_i) - \boldsymbol{e}'(\boldsymbol{\theta}^{\circ\top}\mathbf{X}_i)\| \le \|\psi''\|_\infty\sqrt{34}m^{5/2}\frac{1}{\sqrt{n}c_{\mathcal{D}}}\|\mathcal{D}_m(\boldsymbol{\upsilon}-\boldsymbol{\upsilon}^\circ)\|.$$

Putting all estimates together gives (A.38). □

A.12.2. Condition $(\mathcal{E}\mathcal{D}_2)$

Just as for the conditions $(\mathcal{E}\mathcal{D}_1)$ and $(\mathcal{E}\mathcal{D}_0)$ we can show:

Lemma A.28. *We have* $(\mathcal{E}\mathcal{D}_2)$ *with*

$$\omega_2 = \frac{1}{\sqrt{n}c_{\mathcal{D}}},\ \ \mathtt{g}_2 = \sqrt{n}\widetilde{\mathtt{g}}c_{\mathcal{D}}m^{-1}\mathtt{C}(R_0,p^*)^{-1},\ \ \nu_2^2 = \frac{\widetilde{\nu}^2 m^2\mathtt{C}(R_0,p^*)^2}{2c_{\mathcal{D}}},$$

where with $\mathtt{C}>0$ *in* (A.38)

$$\mathtt{C}(R_0,m) \overset{\text{def}}{=} \mathtt{C}\left(m^{3/2} + \left(\mathtt{C} + \frac{m^2(R_0+\mathtt{r}^*)}{\sqrt{n}c_{\mathcal{D}}}\right)m^{1/2}\right).$$

Proof. Lemma A.27 gives for any $\boldsymbol{\upsilon},\boldsymbol{\upsilon}^\circ \in \Upsilon(\mathtt{r})$ with $\varsigma_{i,m}$ from (A.1)

$$\begin{aligned}
&\|\mathcal{D}_m^{-1}\nabla\varsigma_{i,m}(\boldsymbol{\upsilon}) - \mathcal{D}^{-1}\nabla\varsigma_{i,m}(\boldsymbol{\upsilon}^\circ)\| \\
&\quad\le \frac{\mathtt{C}}{c_{\mathcal{D}}\sqrt{n}}m\left(m^{3/2} + \left(\mathtt{C} + \frac{m^2(\mathrm{R}_0+\mathtt{r}^*)}{\sqrt{n}c_{\mathcal{D}}}\right)m^{1/2}\right)\|\mathcal{D}_m(\boldsymbol{\upsilon}-\boldsymbol{\upsilon}^\circ)\| \\
&\quad\overset{\text{def}}{=} \frac{1}{\sqrt{n}c_{\mathcal{D}}}m\mathtt{C}(\mathrm{R}_0,p^*)\|\mathcal{D}_m(\boldsymbol{\upsilon}-\boldsymbol{\upsilon}^\circ)\|.
\end{aligned} \tag{A.39}$$

We get with $\mu \le \mathtt{g}_2$ and assumption $(\mathbf{Cond}_\varepsilon)$ for any pair $\boldsymbol{\gamma}_1,\boldsymbol{\gamma}_2 \in \{\|\boldsymbol{\gamma}\|=1\}$

$$\begin{aligned}
&\mathbb{E}_\varepsilon\exp\left\{\frac{\mu}{\omega_2\|\mathcal{D}_m(\boldsymbol{\upsilon}-\boldsymbol{\upsilon}^\circ)\|}\boldsymbol{\gamma}_1^\top\left(\mathcal{D}_m^{-1}\nabla^2\left\{\zeta(\boldsymbol{\upsilon})-\zeta(\boldsymbol{\upsilon}^\circ)\right\}\right)\boldsymbol{\gamma}_2\right\} \\
&= \mathbb{E}_\varepsilon\exp\left\{\frac{\mu}{\omega_2\|\mathcal{D}_m(\boldsymbol{\upsilon}-\boldsymbol{\upsilon}^\circ)\|}\sum_{i=1}^n\varepsilon_i\boldsymbol{\gamma}_1^\top\left(\mathcal{D}_m^{-1}\nabla\left\{\varsigma_{i,m}(\boldsymbol{\upsilon})-\varsigma_{i,m}(\boldsymbol{\upsilon}^\circ)\right\}\right)\boldsymbol{\gamma}_2\right\} \\
&= \prod_{i=1}^n\mathbb{E}_\varepsilon\exp\left\{\frac{\mu}{\omega_2\|\mathcal{D}_m(\boldsymbol{\upsilon}-\boldsymbol{\upsilon}^\circ)\|}\varepsilon_i\boldsymbol{\gamma}_1^\top\left(\mathcal{D}_m^{-1}\nabla\left\{\varsigma_{i,m}(\boldsymbol{\upsilon})-\varsigma_{i,m}(\boldsymbol{\upsilon}^\circ)\right\}\right)\boldsymbol{\gamma}_2\right\} \\
&\le \prod_{i=1}^n\exp\left\{\frac{\widetilde{\nu}^2\mu^2}{2\omega_2^2\|\mathcal{D}_m(\boldsymbol{\upsilon}-\boldsymbol{\upsilon}^\circ)\|^2}\left(\boldsymbol{\gamma}_1^\top\left(\mathcal{D}_m^{-1}\nabla\left\{\varsigma_{i,m}(\boldsymbol{\upsilon})-\varsigma_{i,m}(\boldsymbol{\upsilon}^\circ)\right\}\right)\boldsymbol{\gamma}_2\right)^2\right\}.
\end{aligned}$$

With (A.39) this implies

$$\sup_{\substack{\boldsymbol{\gamma}_1,\boldsymbol{\gamma}_2\in\mathbb{R}^{p^*}\\ \|\boldsymbol{\gamma}_i\|=1}} \log \mathbb{E}_{\varepsilon} \exp\left\{\frac{\mu}{\omega_2\|\mathcal{D}_m(\boldsymbol{\upsilon}-\boldsymbol{\upsilon}^{\circ})\|}\boldsymbol{\gamma}_1^{\top}\left(\mathcal{D}_m^{-1}\nabla^2\zeta(\boldsymbol{\upsilon})-\mathcal{D}^{-1}\nabla^2\zeta(\boldsymbol{\upsilon}^{\circ})\right)\boldsymbol{\gamma}_2\right\}$$

$$\leq \frac{\widetilde{\nu}^2\mu^2}{2c_{\mathcal{D}}} m^2 \mathtt{C}(\mathrm{R}_0,p^*)^2. \qquad \square$$

A.12.3. Bound for Hessian

To control the deviation of $\mathcal{D}^{-1}\nabla\zeta(\boldsymbol{\upsilon}^*)$ we apply the following Theorem of [22]:

Theorem A.29 (Corollary 3.7 of [22])**.** *Consider a finite sequence* $(\boldsymbol{M}_i)_{i=1}^n \subset \mathbb{R}^{p^*\times p^*}$ *of independent, selfadjoint, random matrices. Assume that there is a function* $g:(0,\infty)\to\mathbb{R}_+$ *and a sequence of matrices* $(\boldsymbol{A}_i)\subset\mathbb{R}^{p^*\times p^*}$ *that satisfy for all* $\mu>0$

$$\mathbb{E}\mathrm{e}^{\mu\boldsymbol{M}_i} \preceq \mathrm{e}^{g(\mu)\boldsymbol{A}_i}, \quad \textit{where} \quad \boldsymbol{M}\preceq\boldsymbol{M'} \Leftrightarrow \boldsymbol{\gamma}^{\top}\boldsymbol{M}\boldsymbol{\gamma}\leq\boldsymbol{\gamma}^{\top}\boldsymbol{M}\boldsymbol{\gamma},\,\forall\boldsymbol{\gamma}\in\mathbb{R}^{p^*}.$$

Then for all $t\in\mathbb{R}$

$$\mathbb{P}\left(\left\|\sum_{i=1}^n \boldsymbol{M}_i\right\|\geq t\right)\leq p^*\inf_{\mu}\exp\left\{-t\mu+g(\mu)\tau\right\}, \quad \textit{where} \quad \tau \overset{\mathrm{def}}{=} \left\|\sum_{i=1}^n \boldsymbol{A}_i\right\|.$$

Lemma A.30. *We have for* $\mu\leq\widetilde{\mathtt{g}}$

$$\mathbb{E}\exp\left\{\mu\mathcal{D}^{-1}\nabla^2\zeta(\boldsymbol{\upsilon}^*)\right\}\preceq\exp\left\{g(\mu)\,\mathrm{diag}(1,\ldots,1)\right\},$$

where

$$g(\mu)=\begin{cases}\frac{\widetilde{\nu}^2\mathtt{C}^2(\|\boldsymbol{f}'_{\boldsymbol{\eta}^*}\|_\infty+\|\boldsymbol{f}''_{\boldsymbol{\eta}^*}\|_\infty)^2\mu^2}{2}, & \textit{if } \mu\leq\sqrt{n}\widetilde{\mathtt{g}}\mathtt{C}^{-1}(\|\boldsymbol{f}'_{\boldsymbol{\eta}^*}\|_\infty+\|\boldsymbol{f}''_{\boldsymbol{\eta}^*}\|_\infty)^{-1}\\ \infty, & \textit{otherwise.}\end{cases}$$

Proof. Due to Lemma A.27

$$\begin{aligned}&\mathcal{D}^{-1}\nabla\varsigma_{i,m}(\boldsymbol{\upsilon}^*)\\ &\quad\preceq \mathrm{diag}\left(\frac{1}{\sqrt{n}}\mathtt{C}(\|\boldsymbol{f}'_{\boldsymbol{\eta}^*}\|_\infty+\|\boldsymbol{f}''_{\boldsymbol{\eta}^*}\|_\infty),\ldots,\frac{1}{\sqrt{n}}\mathtt{C}(\|\boldsymbol{f}'_{\boldsymbol{\eta}^*}\|_\infty+\|\boldsymbol{f}''_{\boldsymbol{\eta}^*}\|_\infty)\right).\end{aligned}$$

Thus denoting $\mathtt{C}_1\overset{\mathrm{def}}{=}\mathtt{C}(\|\boldsymbol{f}'_{\boldsymbol{\eta}^*}\|_\infty+\|\boldsymbol{f}''_{\boldsymbol{\eta}^*}\|_\infty)$

$$\begin{aligned}\exp\left\{\mu\mathcal{D}^{-1}\nabla^2\zeta(\boldsymbol{\upsilon}^*)\right\} &= \exp\left\{\mu\sum_{i=1}^n\mathcal{D}^{-1}\nabla\varsigma_{i,m}(\boldsymbol{\upsilon}^*)\varepsilon_i\right\}\\ &\preceq \exp\left\{\mu\sum_{i=1}^n\varepsilon_i\,\mathrm{diag}\left(\frac{1}{\sqrt{n}}\mathtt{C}_1,\ldots,\frac{1}{\sqrt{n}}\mathtt{C}_1\right)\right\}.\end{aligned}$$

Consequently we obtain due to the independence of the $\varsigma_{i,m}(\boldsymbol{\upsilon}^*)$ and assumption $(\mathbf{Cond}_\varepsilon)$ for $\mu \le \sqrt{n}\widetilde{\mathtt{g}}\mathtt{C}_1^{-1}$

$$\begin{aligned}
&\mathbb{E}\exp\left\{\mu\mathcal{D}^{-1}\nabla^2\zeta(\boldsymbol{\upsilon}^*)\right\}\\
&\quad\le \prod_{i=1}^n \operatorname{diag}\left(\mathbb{E}\exp\left\{\frac{\mu}{\sqrt{n}}\varepsilon_i\mathtt{C}_1\right\},\dots,\mathbb{E}\exp\left\{\frac{\mu}{\sqrt{n}}\varepsilon_i\mathtt{C}_1\right\}\right)\\
&\quad\le \operatorname{diag}\left(\exp\left\{\frac{\widetilde{\nu}^2\mu^2}{2}\mathtt{C}_1^2\right\},\dots,\exp\left\{\frac{\widetilde{\nu}^2\mu^2}{2}\mathtt{C}_1^2\right\}\right)\\
&\quad= \exp\left\{\frac{\widetilde{\nu}^2\mathtt{C}_1^2\mu^2}{2}\operatorname{diag}(1,\dots,1)\right\}. \qquad \square
\end{aligned}$$

Lemma A.31. *We have with* $\mathtt{C}(\|\boldsymbol{f}'_{\boldsymbol{\eta}^*}\|_\infty + \|\boldsymbol{f}''_{\boldsymbol{\eta}^*}\|_\infty)$ *and if* $\mathtt{x} \le \frac{1}{2}(\widetilde{\nu}^2 n\widetilde{\mathtt{g}}^2 - \log(p^*))$

$$\mathbb{P}\left(\left\|\mathcal{D}^{-1}\nabla^2\zeta(\boldsymbol{\upsilon}^*)\right\| \ge \widetilde{\nu}\mathtt{C}(\|\boldsymbol{f}'_{\boldsymbol{\eta}^*}\|_\infty + \|\boldsymbol{f}''_{\boldsymbol{\eta}^*}\|_\infty)\sqrt{2\mathtt{x}+\log(p^*)}\right) \le \mathrm{e}^{-2\mathtt{x}}.$$

Proof. With Lemma A.30 and Theorem A.29 we obtain for

$$t \le \sqrt{n}\widetilde{\mathtt{g}}\mathtt{C}^{-1}(\|\boldsymbol{f}'_{\boldsymbol{\eta}^*}\|_\infty + \|\boldsymbol{f}''_{\boldsymbol{\eta}^*}\|_\infty)^{-1},$$

that

$$\begin{aligned}
\mathbb{P}\left(\left\|\mathcal{D}^{-1}\nabla^2\zeta(\boldsymbol{\upsilon}^*)\right\| \ge t\right) &\le p^*\inf_\mu \exp\left\{-t\mu + \frac{\widetilde{\nu}^2\mathtt{C}^2(\|\boldsymbol{f}'_{\boldsymbol{\eta}^*}\|_\infty + \|\boldsymbol{f}''_{\boldsymbol{\eta}^*}\|_\infty)^2\mu^2}{2}\right\}\\
&= \inf_\mu \exp\left\{-t\mu + \widetilde{\nu}^2\mathtt{C}(\|\boldsymbol{f}'_{\boldsymbol{\eta}^*}\|_\infty + \|\boldsymbol{f}''_{\boldsymbol{\eta}^*}\|_\infty)^2\frac{\mu^2}{2}\right\}\\
&= \exp\left\{-\frac{t^2}{2\widetilde{\nu}^2\mathtt{C}(\|\boldsymbol{f}'_{\boldsymbol{\eta}^*}\|_\infty + \|\boldsymbol{f}''_{\boldsymbol{\eta}^*}\|_\infty)^2}\right\}.
\end{aligned}$$

Defining $t(\mathtt{x})$ via

$$\mathbb{P}\left(\left\|\mathcal{D}^{-1}\nabla^2\zeta(\boldsymbol{\upsilon}^*)\right\| \ge t(\mathtt{x})\right) = \mathrm{e}^{-\mathtt{x}},$$

we find

$$t(\mathtt{x}) \le \widetilde{\nu}\mathtt{C}(\|\boldsymbol{f}'_{\boldsymbol{\eta}^*}\|_\infty + \|\boldsymbol{f}''_{\boldsymbol{\eta}^*}\|_\infty)\sqrt{2\mathtt{x}+\log(p^*)}, \quad \text{if } \mathtt{x} \le \frac{1}{2}\left(\widetilde{\nu}^2 n\widetilde{\mathtt{g}}^2 - \log(p^*)\right). \ \square$$

A.12.4. Proof of Lemma

Lemma A.31 together with Lemma A.28 gives that in this setting

$$\begin{aligned}\frac{\sqrt{1-\rho}}{2\sqrt{2}(1+\sqrt{\rho})}\kappa(\mathtt{x},\mathrm{R}_0) \;\le\; & \mathtt{C}(\|\boldsymbol{f}'_{\boldsymbol{\eta}^*}\|_\infty+\|\boldsymbol{f}''_{\boldsymbol{\eta}^*}\|_\infty)^2\sqrt{2\mathtt{x}+\log(p^*)}/\sqrt{n}\\ & +\frac{\mathtt{C}\mathfrak{z}_1(\mathtt{x},3p^*)}{n}\left(m^{5/2}+\frac{m^{7/2}(\mathrm{R}_0+\mathtt{r}^*)}{\sqrt{n}c_{\mathcal{D}}}\right)\mathrm{R}_0\\ & +\delta(\mathrm{R}_0+\mathtt{r}_0),\end{aligned}$$

if $\mathtt{x}$ is chosen moderately. As above

$$\begin{aligned}&\mathfrak{z}_1(\mathtt{x},3p^*)=O(\sqrt{\mathtt{x}+p^*})=O(\mathtt{r}_0), \qquad \|\mathcal{D}^{-1}\|\le 1/(\sqrt{n}c_{\mathcal{D}})\\ &\delta(\mathtt{r})/\mathtt{r}=O(p^{*3/2}+\mathtt{C}_{bias}m^{5/2})/\sqrt{n}.\end{aligned}$$

In both cases $\mathtt{C}_{bias}=0$ and $\mathtt{C}_{bias}>0$ the dominating term is the third summand $\delta(\mathrm{R}_0+\mathtt{r}_0)$.

Lemma 5.6 tells us that

$$\mathrm{R}_0=O\left(\sqrt{p^*(1+\mathtt{C}_{bias}\log(n))+n\tau^2+\sqrt{\mathtt{x}n}\tau}\right).$$

In case $\mathtt{C}_{bias}=0$ this means that for moderate $\mathtt{x}$

$$\kappa(\mathtt{x},\mathrm{R}_0)\le\mathtt{C}\left(\frac{p^{*2}}{\sqrt{n}}+p^{*3/2}\tau+\frac{\sqrt{\tau}p^{*3/2}}{n^{1/4}}\right)(1+o(1)),$$

which tends to zero if $p^{*4}/n\to 0$ and $\tau=o(p^{*-3/2})$.

In case $\mathtt{C}_{bias}>0$ we have

$$\mathtt{r}_0=\mathtt{C}\sqrt{p^*\log(n)}, \qquad \mathrm{R}_0=\mathtt{C}\sqrt{p^*\log(n)+\sqrt{p^*}n\tau^2+\sqrt{\mathtt{x}n}\tau}.$$

Consequently

$$\begin{aligned}\kappa(\mathtt{x},\mathrm{R}_0)\;\le\;&\mathtt{C}\Big(p^{*3}\log(n)/\sqrt{n}+p^{*11/4}\tau\\ &+n^{-1/4}m^{5/2}\sqrt{\tau})\Big)(1+o(1)),\end{aligned}$$

which tends to 0 if $m^3\log(n)/n\to 0$ and $\tau=o(p^{*-11/4})$ since then

$$n^{-1/4}m^{5/2}\sqrt{\tau}=o(m^{-3/8}).$$

Aknowledgements

I am grateful to the referees and editors of Electronic Journal of Statistics for very helpful remarks and comments.

References

[1] A. Andresen. A note on the bias of sieve profile estimation. *arXiv:1406.4045*, 2014.

[2] A. Andresen and V. Spokoiny. Critical dimension in profile semiparametric estimation. *Electron. J. Statist.*, 8(2):3077–3125, 2014. MR3301302

[3] A. Andresen and V. Spokoiny. Two convergence results for an alternation maximization procedure. *arXiv:1501.01525v1*, 2014.

[4] A. Cohen, I. Daubechies, and P. Vial. Wavelets on the interval and fast wavelet transforms. *Applied and computational harmonic analysis*, 1:54–81, 1993. MR1256527

[5] M. Delecroix, W. Haerdle, and M. Hristache. Efficient estimation in single-index regression. Technical report, SFB 373, Humboldt Univ. Berlin, 1997.

[6] R. M. Dudley. The sizes of compact subsets of hilbert space and continuity of gaussian processes. *Journal of Functional Analysis*, 1:290–330, 1967. MR0220340

[7] Jerome H. Friedman and Werner Stuetzle. Projection pursuit regression. *Journal of the American Statistical Association*, 76(376):817–823, 1981. MR0650892

[8] W. Haerdle, P. Hall, and H. Ichimura. Optimal smoothing in single-index models. *Ann. Statist.*, 21:157–178, 1993. MR1212171

[9] Peter Hall. Estimating the direction in which a data set is most interesting. *Probability Theory and Related Fields*, 80:51–77, 1988. MR0970471

[10] M. Hristache, A. Juditski, J. Polzehl, and V. Spokoiny. Structure adaptive approach for dimension reduction. *Annals of Statistics*, 29:595–623, 2001. MR1865333

[11] Peter J. Huber. Projection pursuit. *The Annals of Statistics*, 13(2):435–475, 1985. MR0790553

[12] H. Ichimura. Semiparametric least squares (sls) and weighted sls estimation of single-index models. *J Econometrics*, 58:71–120, 1993. MR1230981

[13] Lee K. Jones. On a conjecture of huber concerning the convergence of projection pursuit regression. *Ann. Statist*, 15(2):880–882, 1987. MR0888447

[14] M.R. Kosorok. *Introduction to Empirical Processes and Semiparametric Inference*. Springer in Statistics, 2005. MR2724368

[15] S. Mendelson. Learning without concentration. *arXiv:1401.0304*, 2014. MR3367000

[16] Whitney K Newey. Convergence rates and asymptotic normality for series estimators. *Journal of Econometrics*, 79(1):147–168, 1997. MR1457700

[17] Jammes L. Powell, James H. Stock, and Thomas M. Stoker. Semi-

parametric estimation of index coefficients. *Econometrica*, 57(6):1403–1430, 1989. MR1035117

[18] Xiaotong Shen. On methods of sieves and penalization. *Ann. Statist.*, 25(6):2555–2591, 1997. MR1604416

[19] Vladimir Spokoiny. Parametric estimation. Finite sample theory. *Ann. Statist.*, 40(6):2877–2909, 2012. MR3097963

[20] C. J. Stone. Optimal rates of convergence for nonparametric estimators. *Ann. Statist.*, 8(6):1348–1360, 1980. MR0594650

[21] M. Talagrand. Majorizing measures: the generic chaining. *Ann. Statist.*, 24(3):1049–1103, 1996. MR1411488

[22] J. A. Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12:389–434, 2012. MR2946459

[23] Yingcun Xia. Asymptotic distributions for two estimators of the single-index model. *Econometric Theory*, 22:1112–1137, 2006. MR2328530

[24] Yingcun Xia, H. Tong, W.K. Li, and L. Zhu. An adaptive estimation of dimension reduction space. *Journal of the Royal Statistical Society*, pages 363–410, 2002. MR1924297