Finite sample analysis of profile M-estimators

Dissertation

Zur Erlangung des akademischen Grades Dr. rer. nat.

Im Fach Mathematik

eingereicht an der

Mathematisch-Naturwissenschaftlichen Fakultät
der Humboldt-Universität zu Berlin

von
Diplom-Mathematiker Andreas Andresen

Präsident der Humboldt-Universität zu Berlin

Prof. Dr. Christian Waldhoff

Dekan der Mathematisch-Naturwissenschaftlichen Fakultät

Prof. Dr. Elmar Kulke

Gutachter

1. Prof. Dr. Vladimir Spokoiny

2. Prof. Dr. Gilles Blanchard

3. Dr. Richard Nickl

Tag der Verteidigung: 19.08.2015

Erklärung

Ich erkläre, dass ich die Dissertation selbständig und nur unter Verwendung der von mir gemäß § 7 Abs. 3 der Promotionsordnung der Mathematisch-Naturwissenschaftlichen Fakultät, veröffentlicht im Amtlichen Mitteilungsblatt der Humboldt-Universität zu Berlin Nr. 126/2014 am 18.11.2014 angegebenen Hilfsmittel angefertigt habe. Ich versichere, dass alle Stellen der Arbeit, die wörtlich oder sinngemäß aus anderen Quellen übernommen wurden, als solche kenntlich gemacht sind und dass die Arbeit in gleicher oder ähnlicher Form noch keiner Prüfungsbehörde vorgelegt wurde. Ich habe mich nicht anderwärts um einem Doktorgrad im Promotionsfach Mathematik beworben und besitze keinen Doktorgrad im Promotionsfach Mathematik. Die Promotionsordnung der Mathematisch-Naturwissenschaftlichen Fakultät, veröffentlicht im Amtlichen Mitteilungsblatt der Humboldt-Universität zu Berlin Nr. 126/2014 am 18.11.2014 habe ich zur Kenntnis genommen.

Danksagungen

Ich bedanke mich bei meinem Betreuer Prof. Vladimir Spokoiny für seine Unterstützung, seine sehr hilfreichen Anregungen und für seine Geduld. Weiter möchte ich mich bei meinen Kollegen in der Forschungsgruppe 6 am Weierstrass Institut für angewandte Analysis und Stochastik (WIAS) - und damit auch beim WIAS selbst - für die Unterstützung und das tolle Arbeitsumfeld bedanken. Besonders möchte ich hierbei Maya Zhilova, Niklas Willrich und Sebastian Holtz (HU Berlin) für ihre offenen Ohren und sehr hilfreiche Diskussionen hervorheben. Last but not least bin ich der Forschergruppe FOR1735 und damit der *Deutschen Forschungsgemeinschaft (DFG)* für die Unterstützung und Finanzierung meiner Arbeit sehr dankbar.

Abstract

This thesis presents a new approach to analyze profile M-Estimators for finite samples. The results are inspired by the ideas of [52]. The results of [52] are refined and adapted to the estimation of components of a finite dimensional parameter using the maximization of a criterion functional. A finite sample versions of the Wilks phenomenon and Fisher expansion are obtained and the critical ratio of parameter dimension $p^* \in \mathbb{N}$ to sample size $n \in \mathbb{N}$ of $p^*/\sqrt{n} \ll 1$ is derived in the setting of i.i.d. samples and a smooth criterion functional. The results are extended to parameters in infinite dimensional Hilbert spaces using the sieve approach of [22]. The sieve bias is controlled via common regularity assumptions on the parameter and functional. But our results do not rely on an orthogonal basis in the inner product induced by the model. Furthermore the thesis presents two convergence results for the alternating maximization procedure. All results are exemplified in an application to the Projection Pursuit Procedure of [20]. Under a set of natural and common assumptions all theoretical results can be applied using Daubechies wavelets.

Zusammenfassung

In dieser Arbeit wird ein neuer Ansatz für die Analyse von Profile Maximierungsschätzern präsentiert. Die Resultate sind von den Ideen aus [52] inspiriert. Es werden die Ergebnisse von [52] verfeinert und ange- passt für die Schätzung von Komponenten von endlich dimensionalen Parametern mittels der Maximierung eines Kriteriumfunktionals. Dabei werden Versionen des Wilks Phänomens und der Fisher-Erweiterung für endliche Stichproben hergeleitet und die dafür kritische Relation der Parameterdimension $p^* \in \mathbb{N}$ zum Stichprobenumfang $n \in \mathbb{N}$ von $p^*/\sqrt{n} \ll 1$ gekennzeichnet für den Fall von identisch unabhängig verteilten Beobachtungen und eines hinreichend glatten Funktionals. Die Ergebnisse werden ausgeweitet für die Behandlung von Parametern in unendlich dimensionalen Hilberträumen. Dabei wir die Sieve-Methode von [22] verwendet. Der Sieve-Bias wird durch übliche Regularitätsannahmen an den Parameter und das Funktional kontrolliert. Es wird jedoch keine Basis benötigt, die orthogonal in dem vom Model induzierten Skalarprodukt ist. Weitere Hauptresultate sind zwei Konvergenzaussagen für die alternierende Maximisierungsprozedur zur approximation des Profile-Schätzers. Alle Resultate werden anhand der Analyse der Projection Pursuit Prozedur von [20] veranschaulicht. Die Verwendung von Daubechies-Wavelets erlaubt es unter natür- lichen und üblichen Annahmen alle theoretischen Resultate der Arbeit anzuwenden.

Notation

Before we begin we list some important notations used in this work.

If not specified otherwise we use the following convention for norms

$$
\|\boldsymbol{u}\| \overset{\text{def}}{=} \|\boldsymbol{u}\|_2 \overset{\text{def}}{=} \sqrt{\sum_{k=1}^{p} u_k^2}, \qquad \text{if } \boldsymbol{u} \in \mathbb{R}^p,
$$

$$
\|A\| \overset{\text{def}}{=} \sup_{\boldsymbol{u}\in\mathbb{R}^p, \boldsymbol{v}\in\mathbb{R}^m} \frac{\boldsymbol{u}^\top A \boldsymbol{v}}{\|\boldsymbol{u}\|\|\boldsymbol{v}\|} \qquad \text{if } A \in \mathbb{R}^{p\times m}.
$$

$\varUpsilon \subset \mathbb{R}^p \times \mathcal{X}_{\boldsymbol{\eta}}$ denotes the parameter set, where $(\mathcal{X}_{\boldsymbol{\eta}}\|\cdot\|)$ is a separable Hilbert space, with norm $\|\cdot\|$ induced by the inner product $\langle\cdot,\cdot\rangle$. The elements of this set are denoted by $\boldsymbol{\upsilon}$, which can be decomposed into $\boldsymbol{\upsilon} = (\boldsymbol{\theta}, \boldsymbol{\eta}) \in \mathbb{R}^p \times \mathcal{X}_{\boldsymbol{\eta}}$.

B^c denotes the complement of a set $B \subseteq \varUpsilon$.

$\mathbf{Conv}\{B\}$ denotes the convex hull of a set $B \subseteq \varUpsilon$

$\mathcal{X}^*$ denotes the dual Hilbert space of $\mathcal{X}$. Using Riesz representation $\boldsymbol{u}^* \cdot \boldsymbol{v} = \langle \boldsymbol{u}, \boldsymbol{v}\rangle$, for $\boldsymbol{u}^* \in \mathcal{X}^*$ and $\boldsymbol{v} \in \mathcal{X}$. We ease notation and write $\boldsymbol{u}^\top \boldsymbol{v}$ instead of $\boldsymbol{u}^* \cdot \boldsymbol{v}$.

$(\boldsymbol{e}_k)_{k\in\mathbb{N}}$ denotes a countable basis of $(\mathcal{X}, \|\cdot\|)$. Sometimes we will abuse notation and denote the vector $(\boldsymbol{e}_k)_{k=1}^m \in \mathcal{X}^m$ by $\boldsymbol{e}$, if the context allows this.

$\varPi_{\boldsymbol{\theta}}, \varPi_{\boldsymbol{\eta}}$ denote the projections onto $\mathbb{R}^p$ or $\mathcal{X}$ respectively.

$\varPi_m : \mathcal{X} \to \operatorname{span}\{\boldsymbol{e}_1, \ldots, \boldsymbol{e}_m\}$ denotes the orthogonal projection onto the span of $\{\boldsymbol{e}_1, \ldots, \boldsymbol{e}_m\} \subset \mathcal{X}$ for $m \in \mathbb{N}$. In case $\mathcal{X}_l^2$ it denotes the projection onto the span of the first m canonical basis elements.

$B_{\mathtt{r}}(\boldsymbol{u}) \subset \mathcal{X}$ denotes the ball of radius $\mathtt{r} > 0$ around $\boldsymbol{u} \in \mathcal{X}$.

$\overline{A}$ denotes the closure of a set $A \subseteq \mathcal{X}$.

$int(A)$ denotes the interior of a set $A \subseteq \mathcal{X}$.

$Im(O) \subseteq \mathcal{Z}$ denotes the image of the operator $O : \mathcal{X} \to \mathcal{Z}$, where $\mathcal{Z}$ is some vector space.

$\operatorname{supp} f \subseteq \mathcal{X}$ denotes the set on which the function $f : \mathcal{X} \to \mathcal{Z}$ does not take the value $0 \in \mathcal{Z}$.

$L(\mathcal{X}, \mathcal{Y})$ denotes the space of linear maps from $\mathcal{X}$ to $\mathcal{Y}$.

l^2 denotes the set of square sum-able sequences $\{(u_k)_{k\in\mathbb{N}} : \sum_{k=1}^{\infty} u_k^2 < \infty\}$.

$L^2(\varOmega, \boldsymbol{\nu})$ denotes the set of Lebesgue functions $h : \varOmega \to \mathbb{R}$ with $\int_{\varOmega} h^2 d\boldsymbol{\nu} < \infty$.

$1_A : \varOmega \to \mathbb{R}$ denotes the indicator function of a set $A \subset \varOmega$.

$\boldsymbol{\mathcal{Y}}$ denotes the space of the random observations $\mathbb{Y} \in \boldsymbol{\mathcal{Y}}$. Further $\mathcal{M}(\boldsymbol{\mathcal{Y}}, \mathcal{F})$ denotes the class of probability distributions on the space $\boldsymbol{\mathcal{Y}}$ with sigma algebra $\mathcal{F}$.

$\mathbb{P}^*$ denotes the true underlying probability distribution of the observations $\mathbb{Y}$. When the context allows we drop the ”*” and simply write $\mathbb{P}$.

$\mathcal{L} : \varUpsilon \times \boldsymbol{\mathcal{Y}} \to \mathbb{R}$ denotes the criterion functional. In the case of maximum likelihood estimation for $n \in \mathbb{N}$ i.i.d. observations and $\mathbb{P}_{\boldsymbol{\upsilon}} \ll \boldsymbol{\nu}$ it equals

$$\mathcal{L}(\boldsymbol{\upsilon}, \mathbb{Y}) = \sum_{i=1}^{n} \log\left(\frac{d\mathbb{P}_{\boldsymbol{\upsilon}}}{d\boldsymbol{\nu}}(\boldsymbol{Y}_i)\right).$$

$\mathcal{N}(\boldsymbol{u}, \mathcal{V})$ denotes the Gaussian distribution with mean $\boldsymbol{u} \in \mathbb{R}^p$ and covariance matrix $\mathcal{V} \in \mathbb{R}^{p\times p}$.

$\varPhi : \mathbb{R} \to [0, 1]$ we denote the cumulative distribution function of $\mathcal{N}(0, 1)$.

χ_p^2 denotes the Chi-square distribution with $p \in \mathbb{N}$ degrees of freedom, i.e. the law of $\|\boldsymbol{\xi}\|^2 \geq 0$ for $\boldsymbol{\xi} \sim \mathcal{N}(0, I_p)$, where $I_p \in \mathbb{R}^{p\times p}$ denotes the identity operator.

$\chi_p^2(\mathbb{B})$ denotes the generalized Chi-square distribution with $p \in \mathbb{N}$ degrees of freedom, i.e. the law of $\|\boldsymbol{\xi}\|^2 \geq 0$ for $\boldsymbol{\xi} \sim \mathcal{N}(0, \mathbb{B})$ with some positive semi definite symmetric matrix $\mathbb{B} \in \mathbb{R}^{p\times p}$.

$\xrightarrow{w}$ denotes convergence in distribution and $\xrightarrow{\mathbb{P}}$ convergence in probability.

P_n denotes the empirical process of the sample $\{\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n\} \subset \boldsymbol{\mathcal{Y}}$ for some $n \in \mathbb{N}$, i.e. with some space $\mathcal{Z}$ the empirical process is defined as

$$P_n : \{f : \boldsymbol{\mathcal{Y}} \to \mathcal{Z}\} \to \mathcal{Z}, \quad f \mapsto \frac{1}{n}\sum_{i=1}^{n} f(\boldsymbol{Y}_i).$$

$\mathfrak{L}(\mathbf{X})$ denotes the law of the random variable $\mathbf{X} \in \mathcal{X}$.

$\mathfrak{L}(\mathbf{X}) * \mathfrak{L}(\boldsymbol{Y})$ denotes the convolution of the two laws $\mathfrak{L}(\mathbf{X}), \mathfrak{L}(\boldsymbol{Y})$ i.e. the law of $\mathbf{X} + \boldsymbol{Y}$.

$I_p \in \mathbb{R}^{p\times p}$ denotes the identity matrix.

For two matrices $A, B \in \mathbb{R}^{p\times p}$ we denote $A \geq B$ if $A - B \in \mathbb{R}^{p\times p}$ is positive definite.

$\boldsymbol{\theta}^* \in \mathbb{R}^p$ denotes the target of estimation.

$\widetilde{\boldsymbol{\theta}} \in \mathbb{R}^p$ denotes the profile M-estimator.

$\boldsymbol{\theta}^\perp \subset \mathbb{R}^p$ denotes for some $\boldsymbol{\theta} \in \mathbb{R}^p$ the subspace $\{\boldsymbol{\theta}^\circ;\, \boldsymbol{\theta}^\top \boldsymbol{\theta}^\circ = 0\}$.

$S_1^{p,+} \stackrel{\text{def}}{=} \{\boldsymbol{\theta} \in \mathbb{R}^p :\, \|\boldsymbol{\theta}\| = 1,\, \theta_1 > 0\} \subset \mathbb{R}^p$ denotes the upper half sphere.

$\llcorner x \lrcorner \in \mathbb{Z}$ denotes the largest integer smaller than or equal to $x \in \mathbb{R}$.

$\emptyset$ denotes the empty set.

Contents

Chapter 1

Introduction

Consider observations $\mathbb{Y} \in \boldsymbol{\mathcal{Y}}$ with *true distribution*

$$\mathbb{Y} \sim I\!P^*,$$

where $I\!P^* \in \mathcal{M}(\boldsymbol{\mathcal{Y}}, \boldsymbol{\mathcal{F}})$ and $\mathcal{M}(\boldsymbol{\mathcal{Y}}, \boldsymbol{\mathcal{F}})$ denotes the class of probability distributions on the space $\boldsymbol{\mathcal{Y}}$ with sigma-algebra $\mathcal{F}$. As an example one might think of an i.i.d. sample, i.e. with some law $I\!P_{\boldsymbol{Y}}$ on $(\mathcal{Y}, \mathcal{F})$

$$\mathbb{Y} = (\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n) \in \boldsymbol{\mathcal{Y}} = \bigotimes_{i=1}^{n} \mathcal{Y}, \quad I\!P^* = I\!P_{\boldsymbol{Y}}^{\otimes n}, \quad \boldsymbol{\mathcal{F}} \stackrel{\text{def}}{=} \mathcal{F}^{\otimes n}.$$

Assume that the statistical task is to infer some "parameter" $\boldsymbol{\theta}^* = \psi(I\!P^*)$ with

$$\psi : \mathcal{P} \subseteq \mathcal{M}(\boldsymbol{\mathcal{Y}}, \mathcal{F}) \to \varTheta, \;\; I\!P \mapsto \psi(I\!P) \stackrel{\text{def}}{=} \boldsymbol{\theta},$$

where $\varTheta$ is some set and $\mathcal{P} \subseteq \mathcal{M}(\boldsymbol{\mathcal{Y}}, \mathcal{F})$ is a set of measures on which the above map is defined. These types of statistical tasks can be divided into three classes. There are parametric models, where the image set $\varTheta \subset \mathbb{R}^p$ with $p \in \mathbb{N}$ a fixed dimension and where

$$\mathcal{P} = \{I\!P_{\boldsymbol{\theta}}, \, \boldsymbol{\theta} \in \varTheta\},$$

that is, the pre-image of ψ is a parametric family with $\psi(I\!P_{\boldsymbol{\theta}}) = \boldsymbol{\theta}$, see [57] for an asymptotic treatment of these models. The second class of models are fully nonparametric. In this case, the image $\varTheta$ is infinite dimensional. A prominent example would be density estimation, i.e. ψ maps an absolutely continuous probability distribution - with respect to some dominating measure $\boldsymbol{\nu}$ - to its density, see [48]. The third class lies in between these two and is called *semiparametric* estimation tasks. Normally the image set satisfies $\varTheta \subseteq \mathbb{R}^p$ for some finite $p \in \mathbb{N}$ while the set $\mathcal{P} \subseteq \mathcal{M}$ does not need to be a parametric family but usually is still parametrized, i.e.

$$\mathcal{P} = \{I\!\!P_{\boldsymbol{\upsilon}},\, \boldsymbol{\upsilon} \in \varUpsilon\},$$

where $\varUpsilon$ is some infinite dimensional set. If possible the parametrization is chosen such that $\psi(I\!\!P_{\boldsymbol{\upsilon}}) = \varPi_{\theta}\boldsymbol{\upsilon} = \boldsymbol{\theta}$, where $\boldsymbol{\upsilon} = (\boldsymbol{\theta}, \boldsymbol{\eta}) \in \varUpsilon \subseteq \mathbb{R}^p \times \mathcal{X}$ for some space $\mathcal{X}$ and $\varPi_{\boldsymbol{\theta}} : \mathbb{R}^p \times \mathcal{X} \to \mathbb{R}^p$ denoting the projection onto the $\boldsymbol{\theta}$-component. One example that we will address in some detail in this work is the so called single-index model (see [30]). In this model the observations are $\mathbb{Y} = (Y_i, \mathbf{X}_i)_{i=1}^n \subset \mathbb{R} \times \mathbb{R}^p$ with

$$Y_i = f(\mathbf{X}_i^\top \boldsymbol{\theta}^*) + \varepsilon_i \in \mathbb{R}, \qquad i = 1, ..., n, \tag{1.0.1}$$

for some non-constant $f : \mathbb{R} \to \mathbb{R}$ and $\boldsymbol{\theta}^* \in S_1^{p,+} \subset \mathbb{R}^p$ and with real valued i.i.d errors $\varepsilon_i \sim I\!\!P_\varepsilon$, $I\!\!E\varepsilon_i = 0$, $\mathrm{Var}(\varepsilon_i) = \sigma^2$ and i.i.d random variables $\mathbf{X}_i \in \mathbb{R}^p$ with distribution denoted by $I\!\!P_{\mathbf{X}}$. To ensure identifiability of $\boldsymbol{\theta}^* \in \mathbb{R}^p$ it is assumed that it lies in the half sphere $S_1^{p,+} \stackrel{\text{def}}{=} \{\boldsymbol{\theta} \in \mathbb{R}^p : \|\boldsymbol{\theta}\| = 1, \theta_1 > 0\} \subset \mathbb{R}^p$. This means that with some function space $\mathcal{X}$

$$\mathcal{P} = \left\{ \left(I\!\!P_{f(\mathbf{X}^\top \boldsymbol{\theta})} * I\!\!P_\varepsilon \right)^{\otimes n},\, \boldsymbol{\theta} \in S_1^{p,+},\, f \in \mathcal{X} \right\}.$$

We will discuss this example extensively in Chapter 6. In Section 2.1 we will briefly summarize some of the most fundamental general results about the class of *semiparametric* problems (see [34] for a rather recent monograph). This thesis deals with the analysis of a special type of such tasks, namely the case that the "target" $\boldsymbol{\theta}^* = \psi(I\!\!P)$ can be expressed as

$$\boldsymbol{\theta}^* = \varPi_{\boldsymbol{\theta}} \operatorname*{argmax}_{\boldsymbol{\upsilon} \in \varUpsilon} I\!\!E_{I\!\!P^*} \mathcal{L}(\mathbb{Y}, \boldsymbol{\upsilon}),$$

where $\mathcal{L} : \boldsymbol{\mathcal{Y}} \times \varUpsilon \to \mathbb{R}$ is some functional and $I\!\!E_{I\!\!P}$ denotes the expectation operator under the measure $I\!\!P \in \mathcal{M}$. This means that

$$\psi : \mathcal{P} \subseteq \mathcal{M}(\boldsymbol{\mathcal{Y}}, \mathcal{F}) \to \varTheta, \;\; I\!\!P \mapsto \varPi_{\boldsymbol{\theta}} \operatorname*{argmax}_{\boldsymbol{\upsilon} \in \varUpsilon} I\!\!E_{I\!\!P} \mathcal{L}(\boldsymbol{\upsilon}). \tag{1.0.2}$$

A natural way to solve this problem is to simply use the data and define as estimator

$$\widetilde{\boldsymbol{\theta}} \stackrel{\text{def}}{=} \varPi_{\boldsymbol{\theta}} \operatorname*{argmax}_{\boldsymbol{\upsilon} \in \varUpsilon} \mathcal{L}(\mathbb{Y}, \boldsymbol{\upsilon}) = \operatorname*{argmax}_{\boldsymbol{\theta} \in \varUpsilon_{\boldsymbol{\theta}}} \max_{\boldsymbol{\eta} \in \varUpsilon_{\boldsymbol{\eta}}} \mathcal{L}(\mathbb{Y}, \boldsymbol{\upsilon}), \tag{1.0.3}$$

where $\varUpsilon_{\boldsymbol{\theta}} \stackrel{\text{def}}{=} \{\varPi_{\boldsymbol{\theta}} \boldsymbol{\upsilon} : \boldsymbol{\upsilon} \in \varUpsilon\} \subset \mathbb{R}^p$ and $\varUpsilon_{\boldsymbol{\eta}} \stackrel{\text{def}}{=} \{\varPi_{\boldsymbol{\eta}} \boldsymbol{\upsilon} : \boldsymbol{\upsilon} \in \varUpsilon\} \subset \mathcal{X}$ with $\varPi_{\boldsymbol{\eta}}$ denoting the projection onto the $\boldsymbol{\eta}$-component. These estimators are called *profile Maximization Estimators* (profile ME) since $\widetilde{\boldsymbol{\theta}} \in \mathbb{R}^p$ maximizes the functional $\mathcal{L}$ after the nuisance component $\boldsymbol{\eta}$ has been "profiled out". In case of i.i.d observations a natural example for $\mathcal{L}$ would be

$$\mathcal{L}(\mathbb{Y},\boldsymbol{\upsilon}) = \mathcal{L}_n(\mathbb{Y},\boldsymbol{\upsilon}) = \sum_{i=1}^{n} \ell(\boldsymbol{Y}_i,\boldsymbol{\upsilon}), \quad \mathbb{E}_{\mathbb{P}^*}\mathcal{L}_n(\mathbb{Y},\boldsymbol{\upsilon}) = n\mathbb{E}_{\mathbb{P}_{\boldsymbol{Y}}}\ell(\boldsymbol{Y}_1,\boldsymbol{\upsilon}).$$

where $\ell : \mathcal{Y}\times\varUpsilon \to \mathbb{R}$ is a suitable functional. In case of the model in (1.0.1) assume that $\mathcal{X} \subseteq L^2(\mathbb{R})$. With some suitable function basis $(\boldsymbol{e}_k) \subset \mathcal{X}$ and parameters $\boldsymbol{\upsilon} = (\boldsymbol{\theta},\boldsymbol{\eta}) \in \mathbb{R}^p \times l^2$ one could use

$$\mathcal{L}_n(\boldsymbol{\theta},\boldsymbol{\eta}) = -\frac{1}{2}\sum_{i=1}^{n}\Big|Y_i - \sum_{k=0}^{\infty}\boldsymbol{\eta}_k\boldsymbol{e}_k(\mathbf{X}_i^\top\boldsymbol{\theta})\Big|^2, \tag{1.0.4}$$

since indeed $\boldsymbol{\theta}^* = \varPi_{\boldsymbol{\theta}}\operatorname{argmax}_{\boldsymbol{\upsilon}\in\varUpsilon} -\mathbb{E}_{\mathbb{P}^{\mathbf{X}}}\|\sum_{k=0}^{\infty}\boldsymbol{\eta}_k\boldsymbol{e}_k(\mathbf{X}^\top\boldsymbol{\theta}) - f(\mathbf{X}^\top\boldsymbol{\theta}^*)\|^2$. Of course the estimator resulting in (1.0.4) would perform arbitrarily bad because its variance is unbounded. Below we will circumvent this using the *sieve approach.*

If the functional $\mathcal{L}$ is the loglikelihood of the observations $\mathbb{Y}$ the estimator (1.0.3) becomes the so called *profile Maximum Likelihood Estimator* (pMLE). In Section 2.2 we will present in more detail some of the known results about this class of estimators, most prominently those of [40]. Here we briefly mention that even though the full model is nonparametric the estimation of $\boldsymbol{\theta} \in \mathbb{R}^p$ can in many cases be achieved with $\sqrt{n}$-rate. Given a sample $(\boldsymbol{Y}_1,\ldots,\boldsymbol{Y}_n)$ the usual approach in the analysis of these estimators consists in finding conditions on the functional $\mathcal{L}$, the true distribution $\mathbb{P}^*$, and $\varUpsilon$ that allow to derive statements of the kind

$$\sqrt{n}\breve{d}(\widetilde{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*) \xrightarrow{w} \mathcal{N}(0,\breve{d}^{-1}\breve{v}^2\breve{d}^{-1}), \tag{1.0.5}$$

$$\max_{\boldsymbol{\eta}\in\varUpsilon_{\boldsymbol{\eta}}}\mathcal{L}(\widetilde{\boldsymbol{\theta}}_n,\boldsymbol{\eta}) - \max_{\boldsymbol{\eta}\in\varUpsilon_{\boldsymbol{\eta}}}\mathcal{L}(\boldsymbol{\theta}^*,\boldsymbol{\eta}) \xrightarrow{w} \chi_p^2(\breve{d}^{-1}\breve{v}^2\breve{d}^{-1}), \tag{1.0.6}$$

where $\breve{v}^2,\breve{d}^2 \in \mathbb{R}^{p\times p}$ are some symmetric positive definite matrices. In the context of maximum likelihood estimation the matrices $\breve{v}^2 = \breve{d}^2 \in \mathbb{R}^{p\times p}$ equal the covariance matrix of the *efficient influence function*, see Section 2.2.1. (1.0.5) states the asymptotic normality of the profile ME and is based on the local linearity

$$\sqrt{n}\breve{d}(\widetilde{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*) - \breve{\boldsymbol{\xi}}_n \xrightarrow{\mathbb{P}} 0,$$

which we refer to as the "Fisher expansion", where with some sequence of random variables $\breve{\boldsymbol{\xi}}_n \xrightarrow{w} \mathcal{N}(0,\breve{d}^{-1}\breve{v}^2\breve{d}^{-1})$. It is important to note that in the right-hand side of (1.0.6) the degrees of freedom are determined by the dimension of the target $p \in \mathbb{N}$ and that it is unaffected by the full complexity of the set $\varUpsilon$ as long as it is not growing with $n \in \mathbb{N}$. The convergence (1.0.6) was first observed in [58] which is why we call it "Wilks phenomenon". Various extensions of this result can be found e.g. in [19, 18, 10].

Usually - in the i.i.d. setting - (1.0.5) and (1.0.6) are derived in three steps. First it is shown that with growing sample size $n \in \mathbb{N}$ the M-estimator $\widetilde{\boldsymbol{\upsilon}}_n$ for the full parameter $\boldsymbol{\upsilon}^*$, i.e.

$$\widetilde{\boldsymbol{\upsilon}}_n = \operatorname*{argmax}_{\boldsymbol{\upsilon} \in \varUpsilon} \mathcal{L}_n(\boldsymbol{\upsilon}), \qquad \boldsymbol{\upsilon}^* = \operatorname*{argmax}_{\boldsymbol{\upsilon} \in \varUpsilon} I\!E_{I\!P^*} \mathcal{L}(\boldsymbol{\upsilon}), \tag{1.0.7}$$

is consistent with the right rate $\mathtt{r}_n \to 0$, i.e. $I\!P^*(\widetilde{\boldsymbol{\upsilon}}_n \in B_{\mathtt{r}_n}(\boldsymbol{\upsilon}^*)) \to 1$, for some euclidean ball around $\boldsymbol{\upsilon}^*$. The second step is to use empirical process techniques to establish a uniform quadratic approximation of the kind

$$\Big| \max_{\boldsymbol{\eta}} \mathcal{L}_n(\boldsymbol{\theta}, \boldsymbol{\eta}) - \max_{\boldsymbol{\eta}} \mathcal{L}_n(\boldsymbol{\theta}^*, \boldsymbol{\eta}) - \breve{\nabla} \mathcal{L}_n(\boldsymbol{\upsilon}^*)(\boldsymbol{\theta} - \boldsymbol{\theta}^*)$$
$$-n \|\breve{d}(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\|^2 \Big| = o_{I\!P}(1), \tag{1.0.8}$$

on the set $\{\boldsymbol{\theta};\, (\boldsymbol{\theta}, \operatorname{argmax}_{\boldsymbol{\eta}} \mathcal{L}_n(\boldsymbol{\theta}, \boldsymbol{\eta})) \in B_{\mathtt{r}_n}(\boldsymbol{\upsilon}^*)\}$ with some "projected gradient" $\breve{\nabla} = \nabla_{\boldsymbol{\theta}} - \varPi \nabla_{\boldsymbol{\eta}}$, with some linear map $\varPi : \mathcal{X}_{\boldsymbol{\eta}} \to \mathbb{R}^p$ and with matrix $\breve{d} \in \mathbb{R}^{p \times p}$. The last step consists in showing that

$$n^{-1/2} \breve{d}^{-1} \breve{\nabla} \mathcal{L}_n(\boldsymbol{\upsilon}^*) \xrightarrow{w} \mathcal{N}(0, \breve{d}^{-1} \breve{v}^2 \breve{d}^{-1}).$$

The results (1.0.5) and (1.0.6) can be used for the construction of asymptotic confidence sets that yield statistical tests. The construction works as follows. Let $q_\alpha^2 > 0$ be an $\alpha-$ level quantile of a $\chi_p^2(\breve{d}^{-1} \breve{v}^2 \breve{d}^{-1})$-distribution. Set

$$\mathcal{E}(q_\alpha) = \big\{\boldsymbol{\theta} : \sqrt{n} \|\breve{d}(\widetilde{\boldsymbol{\theta}}_n - \boldsymbol{\theta})\| \le q_\alpha \big\}; \tag{1.0.9}$$

then one can use (1.0.5) to show

$$I\!P\left\{\boldsymbol{\theta}^* \notin \mathcal{E}\left(q_\alpha\right)\right\} = I\!P\left\{\sqrt{n} \|\breve{d}(\widetilde{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*)\| \ge q_\alpha\right\} \to 1 - \alpha.$$

The last step uses Slutsky's Lemma and relies on two things. First the weak convergence of $\boldsymbol{\xi}_n = n^{-1/2} \breve{d}^{-1} \breve{\nabla} \mathcal{L}_n(\boldsymbol{\upsilon}^*)$ to a $\mathcal{N}(0, \breve{d}^{-1} \breve{v}^2 \breve{d}^{-1})$-distributed random variable and secondly on the disappearance of the error term in (1.0.8). Although these results appear to be accurate in many practical finite sample situations, it is unsatisfactory from a theoretical point of view that the construction of confidence sets for the actual finite sample data set at hand remains out of reach. Relying on the asymptotic results implies ignoring the $o_{I\!P}(1)$ terms and the distance between the finite sample distribution of $\|n^{-1/2} \breve{d}^{-1} \breve{\nabla} \mathcal{L}_n(\boldsymbol{\upsilon}^*)\|^2$ and the chi square distribution with p degrees of freedom. The later can be accounted for using the Berry Esseen theorem (Berry [8]) or Edgeworth expansions (Hall [25]) but - to the authors knowledge - there is no general theory that serves a finite sample bound for the $o_{I\!P}(1)$ term in (1.0.8). As we show in Remark 4.2.19 this term can have a tremendous effect on the confidence sets. Bounding this term is rather involved because - among other reasons - it also depends on the consistency of $\widetilde{\boldsymbol{\upsilon}}$ i.e. on the rate $\mathtt{r}_n$. To get finite sample bounds one needs - besides stronger conditions on the smoothness and moments of the functional $\mathcal{L}$ - finite sample a priori bounds for the deviation of $\widetilde{\boldsymbol{\upsilon}}$.

In this thesis we present a new non-asymptotic approach based on ideas of [52] (see Chapter 3), that allows to quantify probabilistic upper bounds for the term in (1.0.8) for finite sample size. The underlying tools rely on assuming a finite full dimension $p^* \in \mathbb{N}$, i.e. $\varUpsilon \subseteq \mathbb{R}^{p^*}$. To account for infinite dimensional parameter spaces this makes using the *sieve approach* (see below) necessary. The finite sample approach yields results of the following kind: With probability greater than $1 - 2\mathrm{e}^{-\mathtt{x}}$

$$\left\| \breve{D}(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) - \breve{\boldsymbol{\xi}} \right\| \le \breve{\diamondsuit}(\mathtt{x}), \tag{1.0.10}$$

$$\left| \max_{\boldsymbol{\eta} \in \varUpsilon_{\boldsymbol{\eta}}} \mathcal{L}(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\eta}) - \max_{\boldsymbol{\eta} \in \varUpsilon_{\boldsymbol{\eta}}} \mathcal{L}(\boldsymbol{\theta}^*, \boldsymbol{\eta}) - \|\breve{\boldsymbol{\xi}}\|^2 \right| \le \sqrt{p}\breve{\diamondsuit}(\mathtt{x}). \tag{1.0.11}$$

The symbol $\breve{\diamondsuit}(\mathtt{x})$ denotes a bound for the accuracy of the above approximations. It is a central object of this work and will be discussed in detail in Chapter 4. $\breve{D} \in \mathbb{R}^{p\times p}$ is a matrix related to $\sqrt{n}\breve{d} \in \mathbb{R}^{p\times p}$ from above. The random variable $\breve{\boldsymbol{\xi}} \in \mathbb{R}^p$ possesses desirable properties, such as good tail bounds of the kind $I\!P(\|\breve{\boldsymbol{\xi}}\| \ge \mathfrak{z}(\mathtt{x})) \le 2\mathrm{e}^{-\mathtt{x}}$, with some deviation bound $\mathfrak{z}(\mathtt{x}) \le \mathtt{C}\sqrt{p + \mathtt{x}}$. These results are presented in Chapter 4. Using the scheme in (1.0.9) the bounds (1.0.10) and (1.0.11) allow the construction of (conservative) "confidence sets":

$$\mathcal{A}(\mathfrak{z}(\mathtt{x}) + \breve{\diamondsuit}(\mathtt{x})) \stackrel{\mathrm{def}}{=} \left\{ \boldsymbol{\theta} : \|\breve{D}(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta})\| \le \mathfrak{z}(\mathtt{x}) + \breve{\diamondsuit}(\mathtt{x}) \right\}, \tag{1.0.12}$$

$$I\!P(\boldsymbol{\theta}^* \in \mathcal{A}(\mathfrak{z}(\mathtt{x}) + \breve{\diamondsuit}(\mathtt{x}))) \ge 1 - 4\mathrm{e}^{-\mathtt{x}}.$$

If (approximate) quantiles q_α for $\|\breve{\boldsymbol{\xi}}\|$ are available, the construction could be refined. Assume for instance that with some small $\epsilon > 0$ and any $\alpha \in [0, 1]$

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

then (see Remark 4.2.12)

$$\alpha + \epsilon + 2\mathrm{e}^{-\mathtt{x}} \le I\!P\left\{\boldsymbol{\theta}^* \in \mathcal{A}(q_\alpha + \diamondsuit(\mathtt{x})\right\},$$

$$I\!P\left\{\boldsymbol{\theta}^* \in \mathcal{A}(q_\alpha - \diamondsuit(\mathtt{x}))\right\} \le \alpha - \epsilon - 2\mathrm{e}^{-\mathtt{x}}.$$

The important achievement of (1.0.10) and (1.0.11) is that these bounds allow to make approximate confidence statements even in the finite sample case, without ignoring "hopefully small enough" terms. As mentioned such terms appear in this or a similar form also in the asymptotic approaches (for example [40]) but they are shown to be a zero sequence in the sample size $n \in \mathbb{N}$ under certain complexity and smoothness assumptions on the set of scores $\{\breve{\nabla}\mathcal{L}(\boldsymbol{\upsilon}),\, \boldsymbol{\upsilon} \in \varUpsilon\}$. The obtained "confidence sets" (1.0.12) are more conservative, i.e. larger than the asymptotic ones, but guarantee that the claimed coverage probability is attained. Note however that on this level the contribution is rather theoretical: as in case of the asymptotic results in [40], crucial objects such as the matrix $\breve{D}$ are unknown and would have to be estimated as well. An honest real data application of these results, where all model specific constants are unknown, is not possible yet and would be well beyond the scope of this work.

In the derivation of (1.0.10) and (1.0.11) we do not simply assume that the profile ME is consistent but give conditions that ensure the right concentration behavior. This particularly allows to address the crucial question of the largest dimension of the nuisance parameter for which the Wilks and Fisher expansions still hold. As we point out in Section 4.2.5 in the smooth i.i.d case with a fixed dimension of the target parameter, both Fisher and Wilks results apply up to an error of order $p^*/n^{1/2}$. This is an improvement with respect to a naive application of the results of [52] from Chapter 3, which would lead to an error of order $p^{*3/2}/n^{1/2}$. In particular, we obtain that the error term in the Fisher expansion can be smaller than the similar error term in the Wilks Theorem, namely by a factor of the order $\sqrt{p^*}$. This ratio $p^*/n^{1/2}$ is the critical bound for the quality of the Fisher and Wilks expansions under the imposed conditions which is confirmed by a specific counter-example in Section 4.2.5. It is of interest to compare our statements with the existing literature on the growing parameter asymptotics. We particularly mention [35, 36, 37] and a series of papers by S. Portnoy, see e.g. [43, 44, 45]. The typical dimensional asymptotic appearing in those works indeed is $p^* = o(n^{1/2})$, which corresponds to our results.

Once the maximal allowed growth rate of p^* as a function of sample size is determined, the results (1.0.10) and (1.0.11) can be applied to the setting where the nuisance $\boldsymbol{\eta}$ lies in an infinite dimensional separable Hilbert space $\mathcal{X}$ via the *sieve approach*; see [22], Chapter 8. For this, let $(\boldsymbol{e}_k)$ be a suitable basis of $\mathcal{X}$ and define for some $m \in \mathbb{N}$ the sieve profile ME via

$$\widetilde{\boldsymbol{\theta}}_m \overset{\text{def}}{=} \varPi_{\boldsymbol{\theta}} \operatorname*{argmax}_{\substack{\boldsymbol{\theta}\in\mathbb{R}^p \\ \boldsymbol{\eta}\in\mathbb{R}^m}} \mathcal{L}\left(\boldsymbol{\theta}, \sum_{k=1}^{m} \eta_k \boldsymbol{e}_k\right). \tag{1.0.13}$$

By abuse of notation we denote this estimator by $\boldsymbol{\theta}_m$, where in asymptotic settings $m \in \mathbb{N}$ depends on the sample size $n \in \mathbb{N}$, such that in that context we suppress the sub index $\cdot_n$ to ease notation. This type of estimators are studied in [12] as well with a lot of examples and special cases. In case of the model in (1.0.1) and assuming that $\mathcal{X} \subseteq L^2(\mathbb{R})$ this means that we use the functional

$$\mathcal{L}_m(\boldsymbol{\theta}, \boldsymbol{\eta}) = -\frac{1}{2} \sum_{i=1}^{n} \Big| Y_i - \sum_{k=0}^{m} \boldsymbol{\eta}_k \boldsymbol{e}_k(\mathbf{X}_i^\top \boldsymbol{\theta}) \Big|^2, \tag{1.0.14}$$

instead of that in (1.0.4). The crucial part in this context is to incorporate and bound the bias " $\boldsymbol{\upsilon}^* - \boldsymbol{\upsilon}_m^*$ " where

$$\boldsymbol{\upsilon}_m^* \stackrel{\text{def}}{=} \operatorname*{argmax}_{\substack{\boldsymbol{\theta} \in \mathbb{R}^p \\ \boldsymbol{\eta} \in \mathbb{R}^m}} I\!\!E \mathcal{L}\left(\boldsymbol{\theta}, \sum_{k=1}^{m} \eta_k \boldsymbol{e}_k \right).$$

In Section 4.3 we explain in detail how this can be done. To convey the idea define $\boldsymbol{\eta}^* = \Pi_{\boldsymbol{\eta}} \boldsymbol{\upsilon}^*$. The approach is based on the decay behavior of $\langle \boldsymbol{\eta}^*, \boldsymbol{e}_k \rangle$ as $k \to \infty$ and based on the properties of the operator

$$A_{\boldsymbol{\varkappa}\boldsymbol{\upsilon}} \stackrel{\text{def}}{=} \nabla_{(I-\Pi_m)\boldsymbol{\eta}} \nabla_{(\boldsymbol{\theta}, \Pi_m \boldsymbol{\eta})} I\!\!E \mathcal{L}(\boldsymbol{\upsilon}^*) : \mathbb{R}^{p^*} \to (I - \Pi_m)\mathcal{X},$$

where $\Pi_m : \mathcal{X} \to \mathbb{R}^m$ denotes the projection onto the span of the first m basis elements $(\boldsymbol{e}_k)_{k=1}^m$. Once the bias is controlled this allows to apply the finite dimensional results for each $m \in \mathbb{N}$ to obtain

$$\Big\| \breve{D}(\widetilde{\boldsymbol{\theta}}_m - \boldsymbol{\theta}^*) - \breve{\boldsymbol{\xi}}_m \Big\| \le \breve{\diamondsuit}(\mathtt{x}) + \alpha(m),$$

$$\Big| \max_{\boldsymbol{\eta} \in \Pi_m \Upsilon_{\boldsymbol{\eta}}} \mathcal{L}(\widetilde{\boldsymbol{\theta}}_m, \boldsymbol{\eta}) - \max_{\boldsymbol{\eta} \in \Pi_m \Upsilon_{\boldsymbol{\eta}}} \mathcal{L}(\boldsymbol{\theta}^*, \boldsymbol{\eta}) - \|\breve{\boldsymbol{\xi}}_m\|^2/2 \Big| \le \sqrt{p}\breve{\diamondsuit}(\mathtt{x}) + \alpha(m),$$

where $\alpha(m) \ge 0$ quantifies the impact of the bias " $\boldsymbol{\upsilon}^* - \boldsymbol{\upsilon}_m^*$ ". The choice of $m \in \mathbb{N}$ then has to balance the two terms $\breve{\diamondsuit}(\mathtt{x})$ and $\alpha(m)$. For statistical inference the term $\alpha(m)$ would have to be added to $\breve{\diamondsuit}(\mathtt{x})$ in the bounds in (1.0.12). In Section 2.2.3 we present a representative asymptotic approach to this type of estimators from [12] and in Section 4.3.3 we will explain how the related results can be derived in our framework. As it turns out, the careful analysis of $A_{\boldsymbol{\varkappa}\boldsymbol{\upsilon}}$ allows to address bias effects that occur when the used basis $(\boldsymbol{e}_k)_{k=1}^\infty$ is not orthogonal in the inner product induced by the covariance structure of the model as commonly assumed (cf. [50] and [12]). The example of Chapter 6 shows that this assumption can be misleading in interesting cases.

Another important question is how to actually calculate $\widetilde{\boldsymbol{\theta}}$ in (1.0.3). In situations where $\mathcal{L} : \Upsilon_{\boldsymbol{\theta}} \times \Upsilon_{\boldsymbol{\eta}} \to \mathbb{R}$ is not convex, the maximization task might become computationally very hard. In case of the single-index model with $\mathcal{L}$ in (1.0.14) the maximization problem is high dimensional and non-convex. But for fixed $\boldsymbol{\theta} \in S_1 \subset \mathbb{R}^p$ maximization with respect to $\boldsymbol{\eta} \in \mathbb{R}^m$

is rather simple while for fixed $\boldsymbol{\eta} \in \mathbb{R}^m$ the maximization with respect to $\boldsymbol{\theta} \in \mathbb{R}^p$ can be feasible for low $p \in \mathbb{N}$. A widely used workaround in such a setting is to start with some initial guess $\widetilde{\boldsymbol{\eta}}^{(0)}$ and to alternate for $k \in \mathbb{N}$

$$\widetilde{\boldsymbol{\eta}}^{(k+1)} \stackrel{\text{def}}{=} \operatorname*{argmax}_{\boldsymbol{\eta} \in \Upsilon_{\boldsymbol{\eta}}} \mathcal{L}(\widetilde{\boldsymbol{\theta}}^{(k)}, \boldsymbol{\eta}), \;\; \widetilde{\boldsymbol{\theta}}^{(k)} \stackrel{\text{def}}{=} \operatorname*{argmax}_{\boldsymbol{\theta} \in \Upsilon_{\boldsymbol{\theta}}} \mathcal{L}(\boldsymbol{\theta}, \widetilde{\boldsymbol{\eta}}^{(k)}).$$

This method is called "alternation maximization (minimization) procedure". Although it is employed in many parameter estimation tasks no satisfactory and general "convergence" result is available except for the treatment of specific models (see for example [31], [41], [33] or [62]). A convergence result would be satisfactory if it stated that the elements of the limit set of the procedure posses the same statistical properties as the full maximizer $\widetilde{\boldsymbol{\upsilon}}$, or even that the limit set equals $\{\widetilde{\boldsymbol{\upsilon}}\}$. The alternation maximization procedure can be understood as a special case of the Expectation Maximization algorithm (EM algorithm) as we will illustrate in Chapter 5. There are convergence result for the EM algorithm - one of the first and most popular by [59] - but these results normally only imply that the limit point is a fixed point of the procedure. Generally it is not ensured the sequence of estimator converges to the global maximizer. For instance [59] ensures that with some $\mathcal{L}^* \le \max \mathcal{L}(\boldsymbol{\upsilon})$

$$(\widetilde{\boldsymbol{\theta}}^{(k)}, \widetilde{\boldsymbol{\eta}}^{(k)}) \to \{\boldsymbol{\upsilon} \in \Upsilon, \mathcal{L}(\boldsymbol{\upsilon}) = \mathcal{L}^*\},$$

but he can not ensure that on the set $\{\boldsymbol{\upsilon} \in \Upsilon, \mathcal{L}(\boldsymbol{\upsilon}) = \mathcal{L}^*\}$ a finite sample Wilks or Fisher expansion as in (1.0.10) or (1.0.11) applies. Similarly in a more recent work [6] derive conditions that ensure that

$$\|\widetilde{\boldsymbol{\theta}}^{(k)} - \boldsymbol{\theta}^*\| \le \nu^k \|\widetilde{\boldsymbol{\theta}}^{(0)} - \boldsymbol{\theta}^*\| + \mathtt{C}\epsilon_n,$$

with some ϵ_n that depends on the sample size and on the complexity of the parameter set. Again neither convergence to the actual profile estimator nor desired statistical properties can be guaranteed.

The second part of this thesis deals with the analysis of this procedure. In Chapter 5 we present conditions under which the sequence $(\widetilde{\boldsymbol{\theta}}^{(k)})$ converges to a limit that satisfies the same statistical properties as the profile estimator in (1.0.3) and we specify how many iterations are necessary to obtain accurate results. Furthermore we refine those conditions to obtain a guarantee that the sequence actually converges to the global maximizer $\widetilde{\boldsymbol{\upsilon}}$. More precisely with similar tools as those underlying (1.0.10) and (1.0.11) we manage to show that if the initial guess is good enough and with probability greater than $1 - 8\mathrm{e}^{-\mathtt{x}}$

$$\left\|\breve{D}(\widetilde{\boldsymbol{\theta}}^{(k)} - \boldsymbol{\theta}^*) - \breve{\boldsymbol{\xi}}\right\| \le \breve{\diamondsuit}(\mathtt{x}) + \delta(k), \tag{1.0.15}$$

$$\left|\max_{\boldsymbol{\eta}\in\varUpsilon_{\boldsymbol{\eta}}} \mathcal{L}(\widetilde{\boldsymbol{\theta}}^{(k)}, \boldsymbol{\eta}) - \max_{\boldsymbol{\eta}\in\varUpsilon_{\boldsymbol{\eta}}} \mathcal{L}(\boldsymbol{\theta}^*, \boldsymbol{\eta}) - \|\breve{\boldsymbol{\xi}}\|^2\right| \le p\breve{\diamondsuit}(\mathtt{x}) + \delta(k), \tag{1.0.16}$$

where $\delta(k) \approx \nu^k$ with $\nu < 1$; see Chaper 5. This means that the construction (1.0.12) applies to the estimator $\widetilde{\boldsymbol{\theta}}^{(k)}$ as well if $\mathcal{A}(\mathfrak{z}(\mathtt{x}) + \breve{\diamondsuit}(\mathtt{x}) + \delta(k))$ is used. In other words the sequence $(\widetilde{\boldsymbol{\theta}}^{(k)})$ attains the same statistical properties as $\widetilde{\boldsymbol{\theta}}$. Note that for statistical inference this is all that is needed, as an actual convergence to the profile ME $\widetilde{\boldsymbol{\theta}}$ is not necessary as long as (1.0.15) and (1.0.16) are met with small error $\diamondsuit(\mathtt{x})$.

We also manage to show that $(\widetilde{\boldsymbol{\theta}}^{(k)}, \widetilde{\boldsymbol{\eta}}^{(k)}) \to \widetilde{\boldsymbol{\upsilon}}$, i.e. we find conditions that ensure that with probability greater than $1 - 3\mathrm{e}^{-\mathtt{x}}$, with $\mathcal{D}^2 \stackrel{\mathrm{def}}{=} \nabla^2 I\!\!E\mathcal{L}(\boldsymbol{\upsilon}^*)$ and some $\tau(\mathtt{x}) < 1$

$$\|\mathcal{D}((\widetilde{\boldsymbol{\theta}}^{(k)}, \widetilde{\boldsymbol{\eta}}^{(k)}) - \widetilde{\boldsymbol{\upsilon}})\| \le \tau(\mathtt{x})^{k/\log(k)},$$

if the initial guess is good enough. So we obtain nearly linear convergence of $(\widetilde{\boldsymbol{\theta}}^{(k)}, \widetilde{\boldsymbol{\eta}}^{(k)})$ to $\widetilde{\boldsymbol{\upsilon}}$.

Finally we present an application of the new results to the single-index model and the Projection Pursuit Procedure of [20]. Assume observations $(Y_i, \mathbf{X}_i) \in \mathbb{R} \times \mathbb{R}^p$

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

where $g : \mathbb{R}^p \to \mathbb{R}$ is some continuous function, $(\varepsilon_i)_{i=1,\ldots,n} \subset \mathbb{R}$ are additive centered errors independent of random regressors $(\mathbf{X}_i)$. Consider the task of estimating the conditional expectation

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

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 $n \in \mathbb{N}$. For instance [54] shows that the rate is bounded from below by $n^{-\alpha/(2\alpha+p)}$, where $\alpha > 1/2$ quantifies the smoothness of $g : \mathbb{R}^p \to \mathbb{R}$. [20] propose to use a projection pursuit approach to circumvent this problem in situations where

$$g(\mathbf{X}) \approx \sum_{l=1}^{M} f_l(\mathbf{X}^\top \boldsymbol{\theta}_l^*),$$

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 $M \in \mathbb{N}$. A special case would be $M = 1$, i.e.

observations $\mathbb{Y} = (\mathbf{X}_i, y_i)_{i=1}^n$ from the model (1.0.1). Under a set of natural conditions on the smoothness of the true function g, on the distribution $I\!P_{\mathbf{X}}$ of $\mathbf{X} \in \mathbb{R}^p$ and tail assumptions on the additive i.i.d noise $\varepsilon \in \mathbb{R}$ we manage to show in Chapter 6 that the results from above apply for the sieve M-estimator $\widetilde{\boldsymbol{\theta}} \in \mathbb{R}^p$ derived via the functional in (1.0.14). That is we manage to show that if $m^7/n \to 0$ ($m^5/n \to 0$ if $M = 1$)

$$\diamondsuit(\mathtt{x}) + \alpha(m) \to 0, \qquad n \to \infty,$$

and that there is a feasible initial guess for which the alternating procedure converges in statistical and absolute sense. This also allows us to derive a rather crude assessment of the performance of the Projection Pursuit Procedure of [20]. Unfortunately the results on the critical ratio of dimension to sample size are rather restrictive and the derivations very technical and tedious such that Chapter 6 is more a proof of concept and an illustration of the theory than a presentation of results that are of scientific interest by themselves.

The Thesis is organized as follows: In Chapter 2 we present some important known results on semiparametric models, such as lower bounds for regular estimators, and on M-estimators. Chapter 3 contains a brief synopsis of the ideas and results of [52] and a collection of tools from that paper, which we will use throughout this work. It is followed by the new results for profile M-estimators in a finite dimensional setting and on sieve profile estimators in Chapter 4. Chapter 5 contains the results on the statistical properties and on the convergence of the alternating procedure. Finally in Chapter 6 we apply - for the purpose of illustration - the results to the model (1.0.1).

Chapter 2

Semiparametric models and profile M-estimators

In this chapter we will present some of the fundamental results on semiparametric models and *profile Maximization Estimators* (profile ME). Everything in this chapter - except the section on sieve M-estimators and the treatment of the single-index model - is taken from the books [34] and [57] and from the paper [40].

2.1 Results on semiparametric estimation

In this section we want to briefly summarize the results on efficiency of regular estimators in regular semiparametric models. For simplicity consider the following estimation task: Given i.i.d. observations $\mathbb{Y} = (\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n) \subset \mathcal{Y}$ with $\boldsymbol{Y}_i \sim I\!\!P$ we search for $\psi(I\!\!P) = \boldsymbol{\theta}^* \in \mathbb{R}^p$ with

$$\psi : \mathcal{P} \overset{\text{def}}{=} \{I\!\!P_{\boldsymbol{\upsilon}},\, \boldsymbol{\upsilon} \in \varUpsilon\} \to \mathbb{R}^p, \;\; I\!\!P_{\boldsymbol{\upsilon}} \mapsto \varPi_{\boldsymbol{\theta}}\boldsymbol{\upsilon} = \varPi_{\boldsymbol{\theta}}(\boldsymbol{\theta}, \boldsymbol{\eta}) = \boldsymbol{\theta},$$

where we assume that $\mathcal{P}$ possesses a dominating measure $\boldsymbol{\nu}$. We do not claim that all semiparametric estimation problems can be formulated in this way. But this setting simplifies a lot of the terms in this section and covers all examples of this thesis. Nonetheless the presentation still involves some of concepts that will not be used again in this work. We still present them in full detail to make this section self contained.

In the following we will present some definitions and results from the book [34] for general semiparametric models. For ease of notation we write $I\!\!P^* \overset{\text{def}}{=} I\!\!P_{\boldsymbol{\upsilon}^*}$.

Definition 2.1.1. *A set* $\{I\!\!P_t,\, t \in [0, \epsilon)\}$ *with* $\epsilon > 0$, $I\!\!P_0 = I\!\!P^*$ *and* $I\!\!P_t \in \mathcal{P}$ *for all* $t \in [0, \epsilon)$ *is called* one-dimensional parametric submodel *of* $\mathcal{P}$ *at* $I\!\!P^*$.

Definition 2.1.2. *A one-dimensional parametric submodel* $\{\mathbb{P}_t,\, t \in [0,\epsilon)\}$ *is called* differentiable in quadratic mean *at* $t = 0$ *with* score function $g : \mathcal{Y} \to \mathbb{R}$ *if the densities* $(p_t)_t$ *of* $(\mathbb{P}_t)_t$ *with respect to* $\boldsymbol{\nu}$ *satisfy*

$$\lim_{t\to 0} \int \left(\frac{\sqrt{p_t(y)} - \sqrt{p_0(y)}}{t} - \frac{1}{2} g(y) \sqrt{p_0(y)} \right)^2 \boldsymbol{\nu}(dy) = 0 \qquad (2.1.1)$$

Remark 2.1.1. One can show that $\mathbb{E}_{\mathbb{P}^*} g = 0$ and $\mathbb{E}_{\mathbb{P}^*} g^2 < \infty$ such that $g \in L_0^2(\mathcal{Y}, \mathbb{P}^*) \overset{\text{def}}{=} \{h \in L^2(\mathcal{Y}, \mathbb{P}^*),\, \mathbb{E}_{\mathbb{P}^*}[h] = 0\}$. Note that (2.1.1) is related to the Hellinger distance between measures (see [57], Chapter 14.5).

Definition 2.1.3. *If there exists an open neighborhood* $U(\boldsymbol{\upsilon}^*) \subset \varUpsilon$ *of* $\boldsymbol{\upsilon}^* \in \varUpsilon$ *such that for all* $\boldsymbol{\upsilon} \in U(\boldsymbol{\upsilon}^*)$ *there exists a smooth one-dimensional parametric submodel* $\{\mathbb{P}_t,\, t \in [0,\epsilon)\}$ *with* $\mathbb{P}_{t_0} = \mathbb{P}_{\boldsymbol{\upsilon}}$ *for some* $t \in [0,\epsilon)$ *then* $\mathcal{P}$ is smooth at $\mathbb{P}^* = \mathbb{P}_{\boldsymbol{\upsilon}^*} \in \mathcal{P}$.

Definition 2.1.4. *If* $\mathcal{P}$ *is smooth at* $\mathbb{P}^*$ *the collection of score functions* g *of all one-dimensional parametric submodels in* $\mathbb{P}^*$ *is called* tangent set of the model $\mathcal{P}$ at $\mathbb{P}^*$ *and is denoted by* $\dot{\mathcal{P}}_{\mathbb{P}^*} \subseteq L_0^2(\mathcal{Y})$.

Definition 2.1.5. *If* $\mathcal{P}$ *is smooth at* $\mathbb{P}^*$ *and if there exists a bounded linear operator* $\dot{\psi}_{\mathbb{P}^*} : L_0^2 \to \mathbb{R}^p$ *such that for any one-dimensional parametric submodel* $\{\mathbb{P}_t,\, t \in [0,\epsilon)\}$ *with score function* $g \in \dot{\mathcal{P}}_{\mathbb{P}^*}$

$$\frac{1}{t}\left(\psi(\mathbb{P}_t) - \psi(\mathbb{P}_0)\right) = \dot{\psi}_{\mathbb{P}^*}(g),$$

the map $\psi : \mathcal{P} \to \mathbb{R}^p$ *is called* differentiable at $\mathbb{P}$ relative to $\dot{\mathcal{P}}_{\mathbb{P}^*}$.

Definition 2.1.6. *A sequence of estimators* T_n *for* $\psi(\mathbb{P}^*)$ *is called* asymptotically linear *with* influence function $\breve{\psi}_{\mathbb{P}^*} : \mathcal{Y} \to \mathbb{R}^p$ *if*

$$\sqrt{n}(T_n - \psi(\mathbb{P}^*)) - \sqrt{n} P_n \breve{\psi}_{\mathbb{P}^*} = o_{\mathbb{P}^*}(1).$$

Definition 2.1.7. *An estimator sequence* T_n *for* $\psi(\mathbb{P}^*)$ *is called* regular at $\mathbb{P}^*$ *if for any one-dimensional submodel* $\{\mathbb{P}_t,\, t \in [0,\epsilon)\}$ *and any sequence* $t_n = O(n^{-1/2})$

$$\sqrt{n}(T_n - \psi(\mathbb{P}_{t_n})) \xrightarrow{\mathbb{P}_{t_n}} \boldsymbol{Z},$$

for some tight Borel random variable $\boldsymbol{Z}$ *that does not depend on the submodel or sequence* (t_n).

Assume for now that $\dot{\mathcal{P}}_{\mathbb{P}} \subseteq L_0^2(\mathbb{P})$ is a linear space. Then one can show with the Riesz representation theorem that there exists a function

$$\widetilde{\psi}_{\mathbb{P}^*} \in \overline{\dot{\mathcal{P}}_{\mathbb{P}^*}} \subseteq L_0^2(\mathbb{P}^*),$$

such that

$$\dot{\psi}_{I\!P}(g) = I\!E_{I\!P^*}[\widetilde{\psi}_{I\!P} g] \in \mathbb{R}^p. \tag{2.1.2}$$

The function $\widetilde{\psi}_{I\!P^*}$ is called efficient influence function. Theorem 18.3 of [34] reads:

Theorem 2.1.8. *(Convolution theorem) Assume that $\mathcal{P}$ is smooth at $I\!P^*$ and that $\psi : \mathcal{P} \to \mathbb{R}^k$ is differentiable at $I\!P^*$ relative to $\dot{\mathcal{P}}_{I\!P^*}$ with efficient influence function $\widetilde{\psi}_{I\!P^*}$. Let T_n be a regular estimator sequence for $\psi(I\!P^*)$ with $\boldsymbol{Z}$ being the weak tight limit of $\sqrt{n}(T_n - \psi(I\!P^*))$ under $I\!P^*$. Then the law of $\boldsymbol{Z}$ satisfies $\mathfrak{L}(\boldsymbol{Z}) = \mathfrak{L}(\boldsymbol{Z}_0) * \mathfrak{L}(M)$ where $M \in \mathbb{R}^p$ is some tight Borel random variable and where*

$$\boldsymbol{Z}_0 \sim \mathcal{N}\left(0, I\!E_{I\!P^*}[\widetilde{\psi}_{I\!P^*}\widetilde{\psi}_{I\!P^*}^\top]\right).$$

In other words if the model and the estimator are regular the lower bounds of parametric estimation problems - in particular those derived from the *local asymptotic normality* (LAN) of regular parametric models, see [57], Chapters 7 and 8- carry over to the semiparametric setting. In Section 4.3.2 we will analyse a particular estimator in the model $\mathbb{Y} = (\boldsymbol{Y}_i) \in \mathcal{Y}^{\otimes n}$ and $I\!P^* = I\!P_{\boldsymbol{\upsilon}^*}^{\otimes n}$

$$I\!P^* \in \mathcal{P} = \{I\!P_{\boldsymbol{\upsilon}}, \boldsymbol{\upsilon} = (\boldsymbol{\theta}, \boldsymbol{\eta}) \in \varUpsilon \subseteq \mathbb{R}^p \times \mathcal{X}\},$$

where $\mathcal{X}$ is assumed to be a separable Hilbert space. The target of estimation is $\boldsymbol{\theta}^* \in \mathbb{R}^p$, i.e. the parameter function $\psi(\cdot)$ becomes

$$\psi(I\!P_{\boldsymbol{\theta},\boldsymbol{\eta}}) = \boldsymbol{\theta} \in \mathbb{R}^p.$$

In reference to Definition 2.1.3 it suffices to consider the finite dimensional submodels of the form

$$\{I\!P_t,\, t \in [0,\epsilon)\} = \{I\!P_{\boldsymbol{\upsilon}^* + t\boldsymbol{\upsilon}},\, t \in [0,\epsilon)\}, \quad \boldsymbol{\upsilon} = (\boldsymbol{\theta}, \boldsymbol{\eta}) \in \mathbb{R}^p \times \mathcal{X}. \tag{2.1.3}$$

Define the Fréchet-derivative of $f : \varUpsilon \to \mathbb{R}$ in $\boldsymbol{\upsilon}^* \in \varUpsilon$ as a linear operator $\nabla f(\boldsymbol{\upsilon}^*) : \overline{\operatorname{span}}(\varUpsilon) \to \mathbb{R}$ such that for every $\boldsymbol{\upsilon} \in \varUpsilon$

$$\lim_{t \to 0}\left|\frac{f(\boldsymbol{\upsilon}^*) - f(\boldsymbol{\upsilon}^* + t\boldsymbol{\upsilon}) - t\nabla f(\boldsymbol{\upsilon}^*)\boldsymbol{\upsilon}}{t}\right| \to 0.$$

We call a function Fréchet-differentiable if its Fréchet-derivative exists. Assume that there is a dominating measure $\boldsymbol{\nu}$ such that $p_t = dI\!P_t/d\nu$ are well defined and $\sqrt{p_t}$ is almost everywhere Fréchet-differentiable. Also assume that $\dot{p}_t/p_t \in L^2(I\!P_t)$ with covariance that is continuous in t for all $t \in [0,\epsilon)$

and all submodels in (2.1.3). Then one can show that $\mathcal{P}$ is smooth at $I\!\!P^*$ with

$$\dot{\mathcal{P}}_{I\!\!P^*} = \left\{ \frac{1}{p_0} \nabla \frac{dI\!\!P^*}{d\nu} \boldsymbol{\upsilon}, \boldsymbol{\upsilon} \in \mathbb{R}^p \times \mathcal{X} \right\},$$

where by abuse of notation we denote for $\boldsymbol{\upsilon} \in \mathbb{R}^p \times \mathcal{X}$ with $\boldsymbol{\upsilon}^\top$ its dual element - in the sense of Riesz' Representation Theorem - and where ∇ denotes the Fréchet-gradient. Note that $\dot{\mathcal{P}}_{I\!\!P^*}$ is a linear space. Define the operator $\mathbb{F}^2_{\boldsymbol{\upsilon}^*} : \mathbb{R}^p \times \mathcal{X} \to Im(\mathbb{F}^2_{\boldsymbol{\upsilon}^*})$ as the operator that satisfies for any pair $\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^\circ \in \mathbb{R}^p \times \mathcal{X}$

$$\boldsymbol{\upsilon}^\top \mathbb{F}^2_{\boldsymbol{\upsilon}^*} \boldsymbol{\upsilon}^\circ \stackrel{\text{def}}{=} I\!\!E \left[\frac{1}{p_0^2} \left(\nabla \frac{dI\!\!P^*}{d\nu} \boldsymbol{\upsilon} \right) \left(\nabla \frac{dI\!\!P^*}{d\nu} \boldsymbol{\upsilon}^\circ \right) \right]. \tag{2.1.4}$$

and assume that it is invertible on its image $Im(\mathbb{F}^2_{\boldsymbol{\upsilon}^*})$ with inverse $\mathbb{F}^{-2}_{\boldsymbol{\upsilon}^*} : Im(\mathbb{F}^2_{\boldsymbol{\upsilon}^*}) \to \mathbb{R}^p \times \mathcal{X}$. One can represent

$$\frac{1}{t} \left[\psi(I\!\!P_t) - \psi(I\!\!P_0) \right] \Big|_{t=0} = \boldsymbol{\theta}^* = I\!\!E \left[\frac{1}{p_0^2} \left(\nabla \frac{dI\!\!P^*}{d\nu} \mathbb{F}^{-2}_{\boldsymbol{\upsilon}^*} \mathit{\Pi}_{\boldsymbol{\theta}}^\top \right) \left(\nabla \frac{dI\!\!P^*}{d\nu} \boldsymbol{\upsilon} \right) \right].$$

Consequently if $\frac{1}{p_0} \nabla \frac{dI\!\!P^*}{d\nu} \in Im(\mathbb{F}^2_{\boldsymbol{\upsilon}^*})$ almost surely this gives

$$\widetilde{\psi}_{I\!\!P^*} = \mathit{\Pi}_{\boldsymbol{\theta}} \mathbb{F}^{-2}_{\boldsymbol{\upsilon}^*} \frac{1}{p_0} \nabla \frac{dI\!\!P^*}{d\nu}, \quad I\!\!E_{I\!\!P}[\widetilde{\psi}_{I\!\!P} \widetilde{\psi}_{I\!\!P}^\top] = \mathit{\Pi}_{\boldsymbol{\theta}} \mathbb{F}^{-2}_{\boldsymbol{\upsilon}^*} \mathit{\Pi}_{\boldsymbol{\theta}}^\top,$$

where $\mathit{\Pi}_{\boldsymbol{\theta}} : \mathbb{R}^p \times \mathcal{X} \to \mathbb{R}^p$ is the orthogonal projection onto the $\boldsymbol{\theta}$-components, and $\mathit{\Pi}_{\boldsymbol{\theta}}^\top$ its adjoint operator. Note that with $\ell(\boldsymbol{\upsilon}) \stackrel{\text{def}}{=} \log(dI\!\!P_{\boldsymbol{\upsilon}}/d\nu)$ we have in case $\sqrt{p_t}$ is differentiable thanks to the chain rule

$$\nabla \ell(\boldsymbol{\upsilon}^*) = \frac{1}{p_0} \nabla \frac{dI\!\!P^*}{d\nu}, \quad \mathbb{F}^2_{\boldsymbol{\upsilon}^*} = I\!\!E \left[\nabla \ell(\boldsymbol{\upsilon}^*) \nabla \ell(\boldsymbol{\upsilon}^*)^\top \right].$$

With Theorem 2.1.8 this gives

Corollary 2.1.9. *Assume that there is a dominating measure* $\boldsymbol{\nu}$ *such that* $p_t(\boldsymbol{y}) = dI\!\!P_t/d\nu$ *are well defined and* $\sqrt{p_t}$ *differentiable and that* $\dot{p}_t/p_t \in L^2(I\!\!P_t)$ *with covariance that is continuous in* t *for all* $t \in [0, \epsilon)$ *and all submodels from 2.1.3. Furthermore assume that* $\mathbb{F}^2_{\boldsymbol{\upsilon}^*} : \mathbb{R}^p \times \mathcal{X} \to Im(\mathbb{F}^2_{\boldsymbol{\upsilon}^*})$ *in* (2.1.4) *is invertible and* $\frac{1}{p_0} \nabla \frac{dI\!\!P^*}{d\nu} \in Im(\mathbb{F}^2_{\boldsymbol{\upsilon}^*})$ *almost surely. Then all regular estimators* T_n *of* $\boldsymbol{\theta}^*$ *obey*

$$\lim_{n \to \infty} \sqrt{n}(T_n - \boldsymbol{\theta}^*) \sim \mathcal{N}\left(0, \mathit{\Pi}_{\boldsymbol{\theta}}^\top \mathbb{F}^{-2}_{\boldsymbol{\upsilon}^*} \mathit{\Pi}_{\boldsymbol{\theta}}\right) * \mathfrak{L}(M),$$

with $\mathfrak{L}(M)$ *denoting the law of some independent random variable* $M \in \mathbb{R}^p$.

Remark 2.1.2. Clearly the assumptions that $\mathbb{F}^2_{\boldsymbol{\upsilon}^*} : \mathbb{R}^p \times \mathcal{X} \to Im(\mathbb{F}^2_{\boldsymbol{\upsilon}^*})$ in (2.1.4) is invertible and that almost surely $\frac{1}{p_0}\nabla \frac{d I\!\!P^*}{d\nu} \in Im(\mathbb{F}^2_{\boldsymbol{\upsilon}^*})$ are not necessary. One could generalize the result using different concepts of inverting $\mathbb{F}^2_{\boldsymbol{\upsilon}^*}$ and via projecting $\frac{1}{p_0}\nabla \frac{d I\!\!P^*}{d\nu}$ onto a subspace on which $\mathbb{F}^2_{\boldsymbol{\upsilon}^*}$ is "invertible". But to make this excursion as focused as possible we restrict ourselves to the simplest formulation.

2.1.1 Application to single-index model

We want to apply the above to the special case of the single-index model (1.0.1). Denote $f_{\boldsymbol{\eta}}(\boldsymbol{\theta}^\top \mathbf{X}) = \sum_{k=1}^\infty \eta_k \boldsymbol{e}_k(\mathbf{X}^\top \boldsymbol{\theta})$, i.e. the $\boldsymbol{\eta}$-component $\boldsymbol{f_\eta} \in L^2(\mathbb{R})$ is identified with its Fourier coefficients $\boldsymbol{\eta} \in l^2$. The family of measures becomes

$$\mathcal{P} = \left\{ \left(I\!\!P_{\sum_{k=1}^\infty \eta_k \boldsymbol{e}_k(\mathbf{X}^\top \boldsymbol{\theta})} * I\!\!P_\varepsilon \right)^{\otimes n}, \, \boldsymbol{\theta} \in S_1^{p,+}, \, \boldsymbol{\eta} \in l^2 \right\},$$

and the parameter function $\psi(\cdot)$ remains

$$\psi(I\!\!P_{\boldsymbol{\theta},\boldsymbol{\eta}}) = \boldsymbol{\theta} \in \mathbb{R}^p.$$

Let the path $\boldsymbol{\theta}(t) \in S_1^+$ for $[0,\epsilon)$ be the geodesic satisfying $\lim_{t\to 0} \frac{1}{t}(\boldsymbol{\theta}(t) - \boldsymbol{\theta}^*) = h_{\boldsymbol{\theta}} \in \boldsymbol{\theta}^{*\perp}$ and let $h_{\boldsymbol{\eta}} \in l^2$. In reference to Definition 2.1.3 it suffices to consider the finite dimensional submodels of the form

$$\{I\!\!P_t \in \mathcal{P}, \, t \in [0,\epsilon)\} = \{I\!\!P_{\boldsymbol{\theta}(t), \boldsymbol{\eta}^* + t h_{\boldsymbol{\eta}}} \in \mathcal{P}, \, t \in [0,\epsilon)\}. \tag{2.1.5}$$

Lemma 2.1.10. *Assume that both the error distribution and the distribution of* $\mathbf{X}$ *possess a density with respect to the Lebesgue measure denoted by* p_ϵ *and* $p_{\mathbf{X}}$. *Furthermore assume that* p_ϵ *is continuously differentiable with*

$$\frac{\dot{p_\epsilon}(\varepsilon)}{p_\epsilon(\varepsilon)} \in L^2(I\!\!P_\varepsilon), \tag{2.1.6}$$

Assume that for any $\boldsymbol{\theta} \in \mathbb{R}^p$ *and* $h_{\boldsymbol{\eta}} \in l^2$

$$\int \left(\dot{f}_{\boldsymbol{\eta}^*}(\boldsymbol{x}^\top \boldsymbol{\theta}^*) \boldsymbol{\theta}^\top \boldsymbol{x} + f_{h_{\boldsymbol{\eta}}}(\boldsymbol{x}^\top \boldsymbol{\theta}^*) \right)^2 p_{\mathbf{X}}(\boldsymbol{x}) d\boldsymbol{x} < \infty \tag{2.1.7}$$

then the submodel (2.1.5) *is smooth in* $I\!\!P_{\boldsymbol{\theta}^*,\boldsymbol{\eta}^*}$ *with influence function*

$$g(y,\boldsymbol{x}) = \frac{\dot{p_\epsilon}(y - f_{\boldsymbol{\eta}^*}(\boldsymbol{x}^\top \boldsymbol{\theta}^*)) \left(\dot{f}_{\boldsymbol{\eta}^*}(\boldsymbol{x}^\top \boldsymbol{\theta}^*) h_{\boldsymbol{\theta}}^\top \boldsymbol{x} + f_{h_{\boldsymbol{\eta}}}(\boldsymbol{x}^\top \boldsymbol{\theta}^*) \right)}{p_\epsilon(y - f_{\boldsymbol{\eta}^*}(\boldsymbol{x}^\top \boldsymbol{\theta}^*))} 1_{p_{\mathbf{X}}>0}.$$

Remark 2.1.3. One way to ensure (2.1.7) is to impose that the support of $\mathbf{X}$ is bounded and that $\boldsymbol{\eta}^* \in l^2$ decays in a way that ensures that $\int \dot{f}_{\boldsymbol{\eta}^*}(t)^2 dt < \infty$.

Again $\dot{\mathcal{P}}_{I\!P^*} \subseteq L_0^2(I\!P_{\boldsymbol{\upsilon}})$ is a linear space. Take the submodel in (2.1.5), then

$$\lim_{t\to 0}\frac{1}{t}\left(\psi(I\!P_t)-\psi(I\!P_0)\right) = \lim_{t\to 0}\frac{1}{t}\left(\boldsymbol{\theta}(t)-\boldsymbol{\theta}^*\right) = h_{\boldsymbol{\theta}}^{\top}.$$

Note that

$$\begin{aligned}\Big(\dot{f}_{\boldsymbol{\eta}^*}(\boldsymbol{x}^\top\boldsymbol{\theta}^*)\boldsymbol{x}^\top h_{\boldsymbol{\theta}} + f_{h_{\boldsymbol{\eta}}}(\boldsymbol{x}^\top\boldsymbol{\theta}^*)\Big) &= \Big(\dot{f}_{\boldsymbol{\eta}^*}(\boldsymbol{x}^\top\boldsymbol{\theta}^*)\boldsymbol{x}, \boldsymbol{e}_1(\boldsymbol{x}^\top\boldsymbol{\theta}^*),\ldots\Big)^\top (h_{\boldsymbol{\theta}}, h_{\boldsymbol{\eta}})\\ &\overset{\text{def}}{=} \nabla\ell(\boldsymbol{\upsilon}^*)^\top(h_{\boldsymbol{\theta}}, h_{\boldsymbol{\eta}}),\end{aligned}$$

where for Lebesgue almost every $\boldsymbol{x}\in\mathbb{R}^p$ one has $\nabla\ell(\boldsymbol{\upsilon}^*)^\top \in (\mathbb{R}^p\times l^2)^*$. The - abuse of - notation $\nabla\ell(\boldsymbol{\upsilon}^*)$ is motivated by the fact that it is strongly related to the gradient of the functional

$$\ell(\boldsymbol{\theta},\boldsymbol{\eta}) \overset{\text{def}}{=} \Big|Y_i - \sum_{k=0}^{\infty}\boldsymbol{\eta}_k\boldsymbol{e}_k(\mathbf{X}_i^\top\boldsymbol{\theta})\Big|^2.$$

Set with $\epsilon = y - f_{\boldsymbol{\eta}^*}(\boldsymbol{x}^\top\boldsymbol{\theta}^*)$

$$w_{\varepsilon,\mathbf{X}}(\boldsymbol{x},y) \overset{\text{def}}{=} \frac{\dot{p}_\epsilon(\varepsilon)}{p_\epsilon(\varepsilon)},$$

and define for any pair $\boldsymbol{\upsilon},\boldsymbol{\upsilon}^\circ\in\mathbb{R}^p\times l^2$ the operator

$$\boldsymbol{\upsilon}^\top\widehat{\mathcal{V}}^2\boldsymbol{\upsilon}^\circ \overset{\text{def}}{=} I\!E_{I\!P^*}\Big[w_{\varepsilon,\mathbf{X}}(\boldsymbol{x},y)^2\Big(\nabla\ell(\boldsymbol{\upsilon}^*)\boldsymbol{\upsilon}^\circ\Big)\Big(\nabla\ell(\boldsymbol{\upsilon}^*)\boldsymbol{\upsilon}\Big)\Big].$$

Note that

$$g(y,\boldsymbol{x}) = w_{\varepsilon,\mathbf{X}}(\boldsymbol{x},y)\nabla\ell(\boldsymbol{\upsilon}^*)^\top(h_{\boldsymbol{\theta}},h_{\boldsymbol{\eta}})\in\mathbb{R}.$$

Setting

$$\widetilde{\psi}_{I\!P} = \Pi_{\boldsymbol{\theta}}^\top\widehat{\mathcal{V}}^{-2}w_{\varepsilon,\mathbf{X}}(\boldsymbol{x},y)\nabla\ell(\boldsymbol{\upsilon}^*)^\top\in\mathbb{R}^p,$$

we find

$$\begin{aligned}&I\!E_{I\!P^*}\widetilde{\psi}_{I\!P}g(y,\boldsymbol{x})\\ &\quad= I\!E_{I\!P^*}\Big[w_{\varepsilon,\mathbf{X}}(\boldsymbol{x},y)^2\Big(\nabla\ell(\boldsymbol{\upsilon}^*)\widehat{\mathcal{V}}^{-2}\Pi_{\boldsymbol{\theta}}^\top\Big)\Big(\nabla\ell(\boldsymbol{\upsilon}^*)(h_{\boldsymbol{\theta}},h_{\boldsymbol{\eta}})\Big)\Big] = h_{\boldsymbol{\theta}}.\end{aligned}$$

Consequently we infer with Theorem 2.1.8 that the lower bound for the covariance of regular estimators is given by

$$I\!E_{I\!P}[\widetilde{\psi}_{I\!P}\widetilde{\psi}_{I\!P}^\top] = \Pi_{\boldsymbol{\theta}}\widehat{\mathcal{V}}^{-2}\Pi_{\boldsymbol{\theta}}^\top.$$

In the special case of a Gaussian error distribution with covariance σ^2 the operator $n\widehat{\mathcal{V}}^2$ becomes equal to the operator in (6.2.4).

2.2 M-estimators in semiparametric models

In this Section introduce some important asymptotic results on M-estimation. This will allow us to relate our results of Chapter 4 to the existing theory. Consider $\mathbb{Y} \in \boldsymbol{\mathcal{Y}}$ and some criterion functional $\mathcal{L} : \boldsymbol{\mathcal{Y}} \times \varUpsilon \to \mathbb{R}$ for some set $\varUpsilon$. Then the associated M-estimator and its target are defined as

$$\widetilde{\boldsymbol{\upsilon}} \overset{\text{def}}{=} \operatorname*{argmax}_{\boldsymbol{\upsilon} \in \varUpsilon} \mathcal{L}(\mathbb{Y}, \boldsymbol{\upsilon}), \quad \boldsymbol{\upsilon}^* \overset{\text{def}}{=} \operatorname*{argmax}_{\boldsymbol{\upsilon} \in \varUpsilon} I\!\!E \mathcal{L}(\mathbb{Y}, \boldsymbol{\upsilon}). \tag{2.2.1}$$

A prominent special case is Maximum Likelihood Estimation (MLE) when

$$\mathcal{L}(\mathcal{Y}, \boldsymbol{\upsilon}) = \log \frac{dI\!\!P_{\boldsymbol{\upsilon}}}{d\boldsymbol{\nu}}(\mathcal{Y}, \boldsymbol{\upsilon}), \quad I\!\!P \in \{I\!\!P_{\boldsymbol{\upsilon}}, \, \boldsymbol{\upsilon} \in \varUpsilon\},$$

for some dominating measure $\boldsymbol{\nu}$. As noted above we are interested in profile M-estimators as defined in (1.0.3). The approach we will present in Chapter 4 is derived for the finite dimensional setting, i.e. $\varUpsilon \subseteq \mathbb{R}^{p^*}$ for some $p^* \in \mathbb{N}$. To compare our results we cite the following Theorem from [57] for i.i.d samples $(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n) = \mathbb{Y} \in \mathcal{Y}^{\otimes n}$ and $\mathcal{L}(\mathbb{Y}, \boldsymbol{\upsilon}) = \sum_{i=1}^n \ell(\boldsymbol{y}_i, \boldsymbol{\upsilon})$.

Theorem 2.2.1 ([57], Theorem 5.23). *Let $\ell(\cdot, \boldsymbol{\upsilon}) : \mathcal{Y} \to \mathbb{R}$ be measurable in an open vicinity of $\boldsymbol{\upsilon}^* \in \varUpsilon$ and let $\ell(\boldsymbol{y}, \cdot) : \varUpsilon \to \mathbb{R}$ be differentiable at $\boldsymbol{\upsilon}^* \in int(\varUpsilon)$ for almost every $\boldsymbol{y} \in \mathcal{Y}$. Assume that a measurable function $\dot{\ell} : \boldsymbol{\mathcal{Y}} \to \mathbb{R}$ exists such that for every pair $\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^\circ \in \varUpsilon$ in a neighborhood of $\boldsymbol{\upsilon}^* \in \varUpsilon$*

$$\|\ell(\boldsymbol{y}, \boldsymbol{\upsilon}) - \ell(\boldsymbol{y}, \boldsymbol{\upsilon}^\circ)\| \leq \dot{\ell}(\boldsymbol{y}) \|\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^\circ\|.$$

Furthermore assume that the map $I\!\!E \ell(\boldsymbol{y}, \cdot) : \varUpsilon \to \mathbb{R}$ admits a second order Taylor expansion at a point of maximum $\boldsymbol{\upsilon}^* \in \varUpsilon$ with nonsingular symmetric second derivative matrix $-\mathcal{D}^2 \in \mathbb{R}^{p^* \times p^*}$. If the M-estimator is consistent $\widetilde{\boldsymbol{\upsilon}}_n \xrightarrow{I\!\!P} \boldsymbol{\upsilon}^*$, then

$$\begin{aligned} \sqrt{n}(\widetilde{\boldsymbol{\upsilon}}_n - \boldsymbol{\upsilon}^*) &= \mathcal{D}^{-2} \frac{1}{\sqrt{n}} \sum_{i=1}^n \nabla \ell(\boldsymbol{y}_i, \boldsymbol{\upsilon}^*) + o_{I\!\!P}(1), \\ \sqrt{n}\mathcal{D}(\widetilde{\boldsymbol{\upsilon}}_n - \boldsymbol{\upsilon}^*) &\xrightarrow{w} \mathcal{N}(0, \mathcal{D}^{-1}\mathcal{V}^2\mathcal{D}^{-1}), \end{aligned}$$

where

$$\mathcal{V}^2 = \operatorname{Cov}(\dot{\ell}(\boldsymbol{\upsilon}^*)) \in \mathbb{R}^{p^* \times p^*}.$$

Let $I\!\!P_{\boldsymbol{\upsilon}} = (I\!\!P_{\boldsymbol{\upsilon}}^\circ)^{\otimes n}$ be the probability distribution the i.i.d. sample $(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n) = \mathbb{Y} \in \mathcal{Y}^{\otimes n}$. Assume that

$$\mathcal{L}(\mathbb{Y}, \boldsymbol{\upsilon}) = \sum_{i=1}^n \ell(\boldsymbol{y}_i, \boldsymbol{\upsilon}), \quad \ell(\boldsymbol{y}, \boldsymbol{\upsilon}) = \log p(\boldsymbol{y}, \boldsymbol{\upsilon}), \quad p(\boldsymbol{y}, \boldsymbol{\upsilon}) \overset{\text{def}}{=} \frac{dI\!\!P_{\boldsymbol{\upsilon}}^\circ}{d\boldsymbol{\nu}}(\boldsymbol{y}, \boldsymbol{\upsilon}).$$

Then the estimator defined in (2.2.1) becomes the Maximum Likelihood Estimator (MLE) for i.i.d. samples. It turns out that in this special case the existence of a second order Taylor expansion of $I\!\!E\ell$ and that $\ell(\boldsymbol{y},\cdot):\varUpsilon\to\mathbb{R}$ is differentiable at $\boldsymbol{\upsilon}^*\in int(\varUpsilon)$ is ensured by a natural condition, namely differentiability in quadratic mean.

Theorem 2.2.2 ([57], Theorem 5.39)**.** *Let $\ell(\cdot,\boldsymbol{\upsilon}):\mathcal{Y}\to\mathbb{R}$ be measurable in an open vicinity of $\boldsymbol{\upsilon}^*\in\varUpsilon$ and let $\ell(\boldsymbol{y},\cdot):\varUpsilon\to\mathbb{R}$ be differentiable in quadratic mean at $\boldsymbol{\upsilon}^*\in int(\varUpsilon)$, i.e. there exists a function $\nabla\ell(\cdot,\boldsymbol{\upsilon}^*):\mathcal{Y}\to\mathbb{R}^{p^*}$ such that as $\boldsymbol{\upsilon}\to\boldsymbol{\upsilon}^*$*

$$\int\left(\sqrt{p(\boldsymbol{y},\boldsymbol{\upsilon})}-\sqrt{p(\boldsymbol{y},\boldsymbol{\upsilon}^*)}-\frac{1}{2}\nabla\ell(\boldsymbol{y},\boldsymbol{\upsilon}^*)(\boldsymbol{\upsilon}-\boldsymbol{\upsilon}^*)\sqrt{p(\boldsymbol{y},\boldsymbol{\upsilon}^*)}\right)^2\boldsymbol{\nu}(dy)$$
$$=o(\|\boldsymbol{\upsilon}-\boldsymbol{\upsilon}^*\|^2).$$

Assume that a measurable function $\dot{\ell}:\boldsymbol{\mathcal{Y}}\to\mathbb{R}$ exists such that for every pair $\boldsymbol{\upsilon},\boldsymbol{\upsilon}^\circ\in\varUpsilon$ in a neighborhood of $\boldsymbol{\upsilon}^*\in\varUpsilon$

$$\|\ell(\boldsymbol{y},\boldsymbol{\upsilon})-\ell(\boldsymbol{y},\boldsymbol{\upsilon}^\circ)\|\le\dot{\ell}(\boldsymbol{y})\|\boldsymbol{\upsilon}-\boldsymbol{\upsilon}^\circ\|.$$

If the matrix $\mathcal{V}^2\overset{\text{def}}{=}I\!\!E[\nabla\ell(\boldsymbol{y},\boldsymbol{\upsilon}^*)\nabla\ell(\boldsymbol{y},\boldsymbol{\upsilon}^*)^\top]\in\mathbb{R}^{p^*\times p^*}$ is nonsingular and if the MLE $\widetilde{\upsilon}_n\xrightarrow{I\!\!P}\boldsymbol{\upsilon}^*$ then

$$\begin{aligned}\sqrt{n}(\widetilde{\boldsymbol{\upsilon}}_n-\boldsymbol{\upsilon}^*)&=\mathcal{V}^{-2}\frac{1}{\sqrt{n}}\sum_{i=1}^n\nabla\ell(\boldsymbol{y}_i,\boldsymbol{\upsilon}^*)+o_{I\!\!P}(1),\\ \sqrt{n}\mathcal{V}(\widetilde{\boldsymbol{\upsilon}}_n-\boldsymbol{\upsilon}^*)&\xrightarrow{w}\mathcal{N}(0,I_{p^*}).\end{aligned}$$

Remark 2.2.1. We will see in Chapter 4 that the (strong) conditions in Section 4.2.1 are rather similar to those of Theorem 2.2.1 (see discussion in Section 4.2.1). But the conditions of our approach are not sensitive to the peculiarities of maximum likelihood estimation: it is treated as an usual M-estimator.

Remark 2.2.2. The above Theorems impose that consistency of the estimator $\widetilde{\boldsymbol{\theta}}$ is already established. [57] explains how to attain it using the argmax theorem (see Section 2.2.2). But the technique presented there only gives consistency in probability and not a result of the kind

$$I\!\!P(d(\widetilde{\boldsymbol{\upsilon}}_n,\boldsymbol{\upsilon}^*)\ge\mathtt{r}_n(\mathtt{x}))\le\mathrm{e}^{-\mathtt{x}},$$

for some function $\mathtt{r}_n(\cdot)$ and metric $d(\cdot,\cdot)$ on $\varUpsilon$ as is needed for our finite sample approach in Chapter 4.

For the i.i.d model the definition (1.0.3) becomes

$$\widetilde{\boldsymbol{\theta}}_n \overset{\text{def}}{=} \Pi_{\boldsymbol{\theta}} \operatorname*{argmax}_{\boldsymbol{\upsilon}\in\varUpsilon} I\!P_n \ell(\boldsymbol{y},\boldsymbol{\upsilon}) = \operatorname*{argmax}_{\boldsymbol{\theta}\in\varUpsilon_{\boldsymbol{\theta}}} \max_{\boldsymbol{\eta}\in\varUpsilon_{\boldsymbol{\eta}}} I\!P_n \ell(\boldsymbol{y},\boldsymbol{\upsilon}).$$

For a finite dimensional full model $\varUpsilon \subset \mathbb{R}^{p^*}$ the two theorems above could be modified to yield similar results for $\widetilde{\boldsymbol{\theta}} \overset{\text{def}}{=} \Pi_{\boldsymbol{\theta}}\widetilde{\boldsymbol{\upsilon}}$:

$$\sqrt{n}(\widetilde{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) = \sqrt{n}(\Pi_{\boldsymbol{\theta}}\widetilde{\boldsymbol{\upsilon}}_n - \Pi_{\boldsymbol{\theta}}\boldsymbol{\upsilon}^*) \xrightarrow{w} \mathcal{N}(0, \Pi_{\boldsymbol{\theta}}\mathcal{D}^{-2}\mathcal{V}^2\mathcal{D}^{-2}\Pi_{\boldsymbol{\theta}}^\top).$$

For infinite dimensional parameters things become a bit more involved (see below). Besides asymptotic normality we also want to address the behavior of the (quasi) log-likelihood ratio statistic

$$\begin{aligned}\sup_{\boldsymbol{\theta}\in\varUpsilon_{\boldsymbol{\theta}}} \breve{\mathcal{L}}(\boldsymbol{\theta},\boldsymbol{\theta}^*) &\overset{\text{def}}{=} \sup_{\boldsymbol{\theta}\in\varUpsilon_{\boldsymbol{\theta}}} \breve{\mathcal{L}}(\boldsymbol{\theta}) - \breve{\mathcal{L}}(\boldsymbol{\theta}^*)\\ &\overset{\text{def}}{=} \sup_{\boldsymbol{\theta}\in\varUpsilon_{\boldsymbol{\theta}}} \sup_{\substack{\boldsymbol{\eta}\in\varUpsilon_{\boldsymbol{\eta}}\\(\boldsymbol{\theta},\boldsymbol{\eta})\in\varUpsilon}} \mathcal{L}(\boldsymbol{\theta},\boldsymbol{\eta}) - \sup_{\substack{\boldsymbol{\eta}\in\varUpsilon_{\boldsymbol{\eta}}\\(\boldsymbol{\theta}^*,\boldsymbol{\eta})\in\varUpsilon}} \mathcal{L}(\boldsymbol{\theta}^*,\boldsymbol{\eta}).\end{aligned}$$

Define

$$\widetilde{\boldsymbol{\upsilon}}_{\boldsymbol{\theta}^*} \overset{\text{def}}{=} \operatorname*{argmax}_{(\boldsymbol{\theta}^*,\boldsymbol{\eta})\in\varUpsilon} \mathcal{L}(\boldsymbol{\upsilon}).$$

Under i.i.d. conditions and assuming that $\mathcal{L}(\boldsymbol{\upsilon}) = \sum_{i=1}^n \log \frac{dI\!P_{\boldsymbol{\upsilon}}}{d\boldsymbol{\nu}}(\boldsymbol{Y}_i)$ for a parametric family $\{I\!P_{\boldsymbol{\upsilon}},\, \boldsymbol{\upsilon}\in\varUpsilon\}$ with $\varUpsilon \subseteq \mathbb{R}^{p^*}$ Theorem 16.7 of [57] reads:

Theorem 2.2.3. *Under the same assumption as in in Theorem 2.2.2. If the estimator* $\widetilde{\upsilon}_{\boldsymbol{\theta}^*} \xrightarrow{I\!P} \boldsymbol{\upsilon}^*$ *then*

$$2\sup_{\boldsymbol{\theta}\in\varUpsilon_{\boldsymbol{\theta}}} \breve{\mathcal{L}}(\boldsymbol{\theta},\boldsymbol{\theta}^*) \xrightarrow{w} \chi^2_p. \tag{2.2.2}$$

Remark 2.2.3. As noted in the Introduction the result of Theorem 2.2.3 is often referred to as the *Wilks phenomenon* as it was originally observed by Wilks ([58]) but derived in a somewhat informal fashion which is why we present the result from [57]. As pointed out the degrees of freedom are determined by the dimension $p\in\mathbb{N}$ of the target and it is unaffected by the size of the full dimension $p^*\in\mathbb{N}$. As we will see later in Chapter 4 the result becomes sensitive to the full dimension once $p^*(n)\to\infty$: (2.2.2) holds if $p{p^*}^2/n \to 0$.

Remark 2.2.4. [51] showed that in the setting of M-estimation for i.i.d. samples

$$2\sup_{\boldsymbol{\theta}\in\varUpsilon_{\boldsymbol{\theta}}} \breve{\mathcal{L}}(\boldsymbol{\theta},\boldsymbol{\theta}^*) \xrightarrow{w} \|\breve{D}\Pi_{\boldsymbol{\theta}}\mathcal{D}^{-2}\mathcal{V}\boldsymbol{Z}\|^2,$$

where $\boldsymbol{Z} \sim \mathcal{N}(0, I_{p^*})$ and where $\breve{D}^2 = (\mathit{\Pi}_{\boldsymbol{\theta}} \mathcal{D}^{-2} \mathit{\Pi}_{\boldsymbol{\theta}}^\top)^{-1} \in \mathbb{R}^p$ and $\mathcal{D}^2, \mathcal{V}^2 \in \mathbb{R}$ from Theorem 2.2.1. The assumptions in [51] are a quite technical so we simply remark that the conditions of Theorem 2.2.1 combined with $\widetilde{\upsilon}_{\boldsymbol{\theta}^*} \xrightarrow{I\!\!P} \boldsymbol{\upsilon}^*$ yield the same result. This can be proved using the same arguments as in the proof of Theorem 16.7 of [57] and some matrix algebra.

Let us now turn to profile M-estimators for i.i.d variables for infinite dimensional parameters. We follow closely Chapter 21 of [34]. Denote by

$$\dot{\mathcal{P}}^{(\boldsymbol{\eta})}_{I\!\!P^*} \subset L^2_0(I\!\!P^*),$$

the tangent set at the point $I\!\!P^* = I\!\!P_{\boldsymbol{\upsilon}^*}$ with respect to the one-dimensional submodels of the form

$$\{I\!\!P_{\boldsymbol{\theta}^*,\boldsymbol{\eta}(t)},\, t \in [0,\epsilon)\}, \quad (\boldsymbol{\theta}^*, \boldsymbol{\eta}(\cdot)) : [0,\epsilon) \to \mathit{\Upsilon}, \quad \boldsymbol{\eta}(0) = \boldsymbol{\eta}^*.$$

The results presented below do not rely on the specific structure of the tangent space. Alternatively one could assume - as we do in Chapter 4 - that $\frac{d}{dt}\boldsymbol{\eta}(t)|_{t=0} \stackrel{\text{def}}{=} \dot{\boldsymbol{\eta}}(0) \in \mathcal{X}_{\boldsymbol{\eta}}$ for some Hilbert space $\mathcal{X}_{\boldsymbol{\eta}}$ and that

$$\frac{d}{dt}\ell(\boldsymbol{\theta}^*, \boldsymbol{\eta}(t))|_{t=0} \stackrel{\text{def}}{=} \nabla_{\boldsymbol{\eta}}\ell(\boldsymbol{\theta}^*, \boldsymbol{\eta}^*)[\dot{\boldsymbol{\eta}}(0)],$$

i.e. $\nabla_{\boldsymbol{\eta}}\ell(\boldsymbol{\theta}^*, \boldsymbol{\eta}^*)$ exists and lies in the dual of the space $\mathcal{X}_{\boldsymbol{\eta}}$. Then we set $\dot{\mathcal{P}}^{(\boldsymbol{\eta})}_{I\!\!P_{\boldsymbol{\upsilon}}} = \mathcal{X}_{\boldsymbol{\eta}}$. This leads to a very similar set of assumptions as those we use in Chapter 4, which is why we adapt in the following the results of Chapter 21 of [34] to that setting.

Assume that the functional $\mathcal{L} : \mathit{\Upsilon} \to \mathbb{R}$ admits two Fréchet-derivatives. Using the Riesz representation we obtain

$$\nabla\mathcal{L}(\boldsymbol{\upsilon}) = (\nabla_{\boldsymbol{\theta}}\mathcal{L}, \nabla_{\boldsymbol{\eta}}\mathcal{L}) \in \mathbb{R}^p \times \mathcal{X}_{\boldsymbol{\eta}}, \quad \nabla^2\mathcal{L} : \mathbb{R}^p \times \mathcal{X}_{\boldsymbol{\eta}} \to \mathbb{R}^p \times \mathcal{X}_{\boldsymbol{\eta}}.$$

Define

$$\breve{\nabla}_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\upsilon}) = \nabla_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\upsilon}) - A\mathcal{H}^{-2}\nabla_{\boldsymbol{\eta}}\mathcal{L}(\boldsymbol{\upsilon}), \quad \breve{D}^2(\boldsymbol{\upsilon}^*) = D^2 - A\mathcal{H}^{-2}A^\top,$$

$$\breve{V}^2 \stackrel{\text{def}}{=} \operatorname{Cov}(\breve{\nabla}\mathcal{L}(\boldsymbol{\upsilon}^*)),$$

where

$$\begin{pmatrix} D^2 & A \\ A^\top & \mathcal{H}^2 \end{pmatrix} \stackrel{\text{def}}{=} -\nabla^2 I\!\!E\mathcal{L}(\boldsymbol{\upsilon}^*),$$

which coincides with the definitions in Chapter 4.

Remark 2.2.5. The element $\breve{\nabla}\mathcal{L}(\boldsymbol{\upsilon})$ is related to the efficient influence function in (2.1.2). If the model is correctly specified and $\mathcal{L} = \sum_{i=1}^n \ell_i$ is the log-likelihood then $\breve{\nabla}\ell(\boldsymbol{\upsilon}) = \widetilde{\psi}_{I\!\!P} \in \dot{\mathcal{P}}_{I\!\!P}$.

Consider the following list of conditions from [34] Chapter 21, which we adapted to our setting:

(A.1) (Consistency and rate of convergence) Assume that for some $c_1 > 0$

$$\|\widetilde{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*\| = o_{\mathbb{P}}(1), \;\; \|\widetilde{\boldsymbol{\eta}}_n - \boldsymbol{\eta}^*\| = O_{\mathbb{P}}(n^{-c_1}).$$

(A.2) (Finite variance) $0 < \det(\breve{D}(\boldsymbol{\upsilon}^*)^{-2}\breve{V}^2(\boldsymbol{\upsilon}^*)\breve{D}(\boldsymbol{\upsilon}^*)^{-2}) < \infty$.

(B.3) (Stochastic equicontinuity) For any $\delta_n \to 0$ amd $C > 0$

$$\begin{aligned}
&\sup_{\substack{\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|\le\delta_n,\\ \|\boldsymbol{\eta}-\boldsymbol{\eta}^*\|\le Cn^{-c_1}}} \left\| \frac{1}{\sqrt{n}}(1-\mathbb{E})(\nabla_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta},\boldsymbol{\eta}) - \nabla\mathcal{L}(\boldsymbol{\theta}^*,\boldsymbol{\eta}^*)) \right\| \\
&\quad = o_{\mathbb{P}}(1), \\
&\sup_{\substack{\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|\le\delta_n,\\ \|\boldsymbol{\eta}-\boldsymbol{\eta}^*\|\le Cn^{-c_1}}} \left\| \frac{1}{\sqrt{n}}(1-\mathbb{E})A\mathcal{H}^{-2}(\nabla_{\boldsymbol{\eta}}\mathcal{L}(\boldsymbol{\theta},\boldsymbol{\eta}) - \nabla_{\boldsymbol{\eta}}\mathcal{L}(\boldsymbol{\theta}^*,\boldsymbol{\eta}^*)) \right\| \\
&\quad = o_{\mathbb{P}}(1),
\end{aligned}$$

where we use the shorthand notation $(1-\mathbb{E})\mathbf{X} \stackrel{\text{def}}{=} \mathbf{X} - \mathbb{E}[\mathbf{X}]$.

(B.4) (Smoothness of the model) For some $c_2 > 1$ satisfying $c_1 c_2 > 1/2$, where $c_1 > 0$ is from condition A.1 and for all $(\boldsymbol{\theta},\boldsymbol{\eta}) \in \{\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\| \le \delta_n,\, \|\boldsymbol{\eta}-\boldsymbol{\eta}^*\| \le Cn^{-c_1}\})$

$$\begin{aligned}
&\frac{1}{n}\left| \mathbb{E}\left\{ \breve{\nabla}_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta},\boldsymbol{\eta}) - \breve{\nabla}_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta}^*,\boldsymbol{\eta}^*) \right\} + \breve{D}^2(\boldsymbol{\theta}-\boldsymbol{\theta}^*) \right| \\
&\quad = o(\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|) + O\left(\|\boldsymbol{\eta}-\boldsymbol{\eta}^*\|^{c_2}\right).
\end{aligned}$$

Then Corollary 21.1 of [34] reads

Theorem 2.2.4. *Suppose that the conditions (A.1), (A.2), (B.3) and (B.4) are met. Then*

$$\sqrt{n}(\widetilde{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*) = \breve{D}(\boldsymbol{\upsilon}^*)^{-1}\breve{\nabla}\mathcal{L}(\mathbf{Y}_i,\boldsymbol{\upsilon}) + o_{\mathbb{P}}(1),$$

in particular $\widetilde{\boldsymbol{\theta}}_n$ *is asymptotically normal with covariance*

$$\breve{D}(\boldsymbol{\upsilon}^*)^{-2}\breve{V}^2(\boldsymbol{\upsilon}^*)\breve{D}(\boldsymbol{\upsilon}^*)^{-2} \in \mathbb{R}^{p\times p}.$$

Remark 2.2.6. Here we only cite the weaker result, i.e. the one that needs stronger conditions. See Chapter 21 of [34] how the assumptions B.3 and B.4 can be relaxed.

Remark 2.2.7. The "Wilks"-type result

$$2 \sup_{\boldsymbol{\theta} \in \varUpsilon_{\boldsymbol{\theta}}} \breve{\mathcal{L}}(\boldsymbol{\theta}, \boldsymbol{\theta}^*) \xrightarrow{w} \|\breve{D} \varPi_{\boldsymbol{\theta}} \mathcal{D}^{-2} \mathcal{V} \boldsymbol{Z}\|^2,$$

again can be attained under similar conditions translating the arguments of the proof of Theorem 2.2.7 to the M-estimation setting.

Remark 2.2.8. In our setting of Chapter 4 condition (B.3) and (B.4) are substituted by conditions $(\breve{\mathcal{L}}_0)$ and $(\breve{E}\mathcal{D}_1)$.

2.2.1 Profile Maximum Likelihood Estimation

Again a prominent special case is the profile Maximum Likelihood Estimator. Using the true structure of the underlying family and the assumption that the observation actually are distributed according to an element of that family leads to slightly weaker conditions on the smoothness of the functional $\mathcal{L}(\boldsymbol{y}, \cdot) : \varUpsilon \to \mathbb{R}$ which in this case is the log-likelihood corresponding to the family $\{I\!\!P_{\boldsymbol{\upsilon}}, \boldsymbol{\upsilon} \in \varUpsilon\}$.

But before we list the conditions we need to introduce two additional concepts that play a central role in empirical process theory. They are related to the law of large numbers and the central limit theorem. For this consider a sample $(\boldsymbol{Y}_i)_{i=1}^n \subset \mathcal{Y}$ and underlying measure $I\!\!P$ with associated empirical process P_n indexed by a class $\mathcal{F}$ of functions from $\mathcal{Y}$ to $\mathbb{R}^p$.

Definition 2.2.5. *A function class $\mathcal{F}$ is called $I\!\!P$-*Glivenko-Cantelli *if*

$$\sup_{\boldsymbol{f} \in \mathcal{F}} \|P_n(\boldsymbol{f}) - I\!\!E \boldsymbol{f}(\boldsymbol{Y})\| \xrightarrow{I\!\!P} 0.$$

Definition 2.2.6. *A function class $\mathcal{F}$ is called $I\!\!P$-*Donsker *if the process*

$$\boldsymbol{G}_n \overset{\text{def}}{=} \left\{\sqrt{n}\left(P_n(\boldsymbol{f}) - I\!\!E \boldsymbol{f}(\boldsymbol{Y})\right), \, \boldsymbol{f} \in \mathcal{F}\right\},$$

satisfies for any continuous and compactly supported map $h \in C_c(l^\infty(\mathcal{F}))$ that

$$I\!\!E[h(\boldsymbol{G}_n)] \to I\!\!E[h(\mathbb{G})],$$

where $\mathbb{G}$ is a centered Gaussian process indexed by $\mathcal{F}$ with covariance structure

$$I\!\!E(\mathbb{G}(\boldsymbol{f}^\circ)\mathbb{G}(\boldsymbol{f})^\top) = I\!\!E[\boldsymbol{f}^\circ \boldsymbol{f}^\top] - I\!\!E[\boldsymbol{f}^\circ] I\!\!E[\boldsymbol{f}]^\top.$$

The conditions in [40] read:

(B.4)' (Least favorable smooth submodel) For each $\boldsymbol{\upsilon} = (\boldsymbol{\theta}^{\circ}, \boldsymbol{\eta}) \in \varUpsilon$ there exists a map

$$\boldsymbol{\theta} \to \boldsymbol{\eta}_{\boldsymbol{\theta}}(\boldsymbol{\theta}^{\circ}, \boldsymbol{\eta}),\ (\boldsymbol{\theta}, \boldsymbol{\eta}_{\boldsymbol{\theta}}(\boldsymbol{\theta}^{\circ}, \boldsymbol{\eta})) \in \varUpsilon,\ \boldsymbol{\eta}_{\boldsymbol{\theta}}(\boldsymbol{\theta}, \boldsymbol{\eta}) = \boldsymbol{\eta},$$

such that

$$\boldsymbol{\theta} \to l(\boldsymbol{\theta}, \boldsymbol{\theta}^{\circ}, \boldsymbol{\eta}) = \mathcal{L}\Big(\boldsymbol{\theta}, \boldsymbol{\eta}_{\boldsymbol{\theta}}(\boldsymbol{\theta}^{\circ}, \boldsymbol{\eta})\Big),$$

is twice continuously differentiable, where we denote the derivatives by $\nabla_{\boldsymbol{\theta}} l(\boldsymbol{\theta}, \boldsymbol{\theta}^{\circ}, \boldsymbol{\eta})$ and $\nabla^2_{\boldsymbol{\theta}} l(\boldsymbol{\theta}, \boldsymbol{\theta}^{\circ}, \boldsymbol{\eta})$. Furthermore the map $\boldsymbol{\theta} \to \boldsymbol{\eta}_{\boldsymbol{\theta}}(\boldsymbol{\theta}^{\circ}, \boldsymbol{\eta})$ should be such that

$$\nabla_{\boldsymbol{\theta}} l(\boldsymbol{\theta}^*, \boldsymbol{\theta}^*, \boldsymbol{\eta}^*) = \widetilde{\psi}_{I\!\!P} \in \dot{\mathcal{P}}_{I\!\!P},$$

where $\widetilde{\psi}_{I\!\!P}$ is the efficient influence function in (2.1.2) and such that

$$(\boldsymbol{\theta}, \boldsymbol{\theta}^{\circ}, \boldsymbol{\eta}) \to \Big(l(\boldsymbol{\theta}, \boldsymbol{\theta}^{\circ}, \boldsymbol{\eta}), \nabla_{\boldsymbol{\theta}} l(\boldsymbol{\theta}, \boldsymbol{\theta}^{\circ}, \boldsymbol{\eta}), \nabla^2_{\boldsymbol{\theta}} l(\boldsymbol{\theta}, \boldsymbol{\theta}^{\circ}, \boldsymbol{\eta})\Big),$$

is continuous in $(\boldsymbol{\theta}^*, \boldsymbol{\theta}^*, \boldsymbol{\eta}^*)$.

(A.1)' (Nobias) Assume that the pMLE satisfies $\widetilde{\boldsymbol{\theta}} \overset{I\!\!P}{\to} \boldsymbol{\theta}^*$ and that $\widetilde{\boldsymbol{\eta}} \overset{I\!\!P}{\to} \boldsymbol{\eta}^*$ and

$$I\!\!E \nabla_{\boldsymbol{\theta}_1} l(\boldsymbol{\theta}^*, \widetilde{\boldsymbol{\theta}}, \widetilde{\boldsymbol{\eta}}) = o(\|\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\| + n^{-1/2}).$$

(A.2)' (Regularity) Assume that $\breve{V}^2(\boldsymbol{\upsilon}^*) \overset{\text{def}}{=} I\!\!E_{I\!\!P}[\widetilde{\psi}_{I\!\!P} \widetilde{\psi}_{I\!\!P}^{\top}] \in \mathbb{R}^{p \times p}$ is invertible.

(B.3)' (Complexity) Assume that there exists a neighborhood $U \subseteq \varPi_{\boldsymbol{\theta}} \varUpsilon \times \varUpsilon$ of $(\boldsymbol{\theta}^*, \boldsymbol{\theta}^*, \boldsymbol{\eta}^*)$ such that the class $\{\nabla_{\boldsymbol{\theta}} l(\boldsymbol{\theta}, \boldsymbol{\theta}^{\circ}, \boldsymbol{\eta}),\ (\boldsymbol{\theta}, \boldsymbol{\theta}^{\circ}, \boldsymbol{\eta}) \in U\}$ is $I\!\!P$-Donsker with square-integrable envelope and the class

$$\{\nabla^2_{\boldsymbol{\theta}} l(\boldsymbol{\theta}, \boldsymbol{\theta}^{\circ}, \boldsymbol{\eta}),\ (\boldsymbol{\theta}, \boldsymbol{\theta}^{\circ}, \boldsymbol{\eta}) \in U\},$$

is $I\!\!P$-Glivenko-Cantelli and bounded in $L_1(I\!\!P)$.

Remark 2.2.9. The condition (B.3)' together with (B.4) replaces condition (B.3) from the previous section. As indicated by their labels condition (A.1)' replaces condition (A.1), (A.2)' replaces (A.2) and condition (B.4)' corresponds to (B.4). Consequently in our setting of Chapter 4 condition (B.3)' and (B.4)' are substituted by conditions $(\breve{\mathcal{L}}_0)$ and $(\breve{E}\mathcal{D}_1)$.

Theorem 2.2.7 ([40], Corollaries 4 and 5)**.** *Suppose that the conditions (A.1)', (A.2)', (B.3)' and (B.4)' are met. Then*

$$\sqrt{n}(\widetilde{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*) = -\frac{1}{\sqrt{n}}\sum_{i=1}^{n} \breve{V}^{-2}(\boldsymbol{\upsilon}^*)\breve{\nabla}\ell(\boldsymbol{Y}_i, \boldsymbol{\upsilon}) + o_{\mathbb{P}}(1),$$

in particular $\widetilde{\boldsymbol{\theta}}_n$ *is asymptotically normal with optimal covariance* $\breve{V}^{-2}(\boldsymbol{\upsilon}^*)$. *Furthermore*

$$2 \sup_{\boldsymbol{\theta}\in\varUpsilon_{\boldsymbol{\theta}}} \breve{\mathcal{L}}(\boldsymbol{\theta}, \boldsymbol{\theta}^*) \xrightarrow{w} \chi^2_p.$$

As our approach uses some of the ideas underlying this result we briefly explain the main step of the proof. The arguments in [40] can be sketched as follows: Let $\boldsymbol{\theta} \overset{\mathbb{P}}{\to} \boldsymbol{\theta}^*$ such that $(\boldsymbol{\theta}, \boldsymbol{\theta}, \boldsymbol{\eta}^*), (\boldsymbol{\theta}, \boldsymbol{\theta}^*, \boldsymbol{\eta}^*) \in V$. Define

$$\widetilde{\boldsymbol{\eta}}_{\boldsymbol{\theta}} \overset{\text{def}}{=} \operatorname*{argmax}_{\boldsymbol{\eta}} \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\eta}).$$

Then

$$\begin{aligned} \max_{\boldsymbol{\eta}} \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\eta}) &= l(\boldsymbol{\theta}, \boldsymbol{\theta}, \widetilde{\boldsymbol{\eta}}_{\boldsymbol{\theta}}) \geq l(\boldsymbol{\theta}, \boldsymbol{\theta}^*, \widetilde{\boldsymbol{\eta}}_{\boldsymbol{\theta}^*}), \\ \max_{\boldsymbol{\eta}} \mathcal{L}(\boldsymbol{\theta}^*, \boldsymbol{\eta}) &= l(\boldsymbol{\theta}^*, \boldsymbol{\theta}^*, \widetilde{\boldsymbol{\eta}}_{\boldsymbol{\theta}^*}) \geq l(\boldsymbol{\theta}^*, \boldsymbol{\theta}, \widetilde{\boldsymbol{\eta}}_{\boldsymbol{\theta}}), \end{aligned}$$

such that

$$l(\boldsymbol{\theta}, \boldsymbol{\theta}^*, \widetilde{\boldsymbol{\eta}}_{\boldsymbol{\theta}^*}) - l(\boldsymbol{\theta}^*, \boldsymbol{\theta}^*, \widetilde{\boldsymbol{\eta}}_{\boldsymbol{\theta}^*}) \leq \breve{\mathcal{L}}(\boldsymbol{\theta}) - \breve{\mathcal{L}}(\boldsymbol{\theta}^*) \leq l(\boldsymbol{\theta}, \boldsymbol{\theta}, \widetilde{\boldsymbol{\eta}}_{\boldsymbol{\theta}}) - l(\boldsymbol{\theta}^*, \boldsymbol{\theta}, \widetilde{\boldsymbol{\eta}}_{\boldsymbol{\theta}}).$$

Using the imposed conditions a second order Taylor expansion gives that for any $\boldsymbol{\theta} \overset{\mathbb{P}}{\to} \boldsymbol{\theta}^*$

$$\begin{aligned} \breve{\mathcal{L}}(\boldsymbol{\theta}) - \breve{\mathcal{L}}(\boldsymbol{\theta}^*) = &\sum_{i=1}^{n} \widetilde{\psi}_{\mathbb{P}}(\boldsymbol{y}_i)(\boldsymbol{\theta} - \boldsymbol{\theta}^*) - \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}^*)^\top \breve{D}^2(\boldsymbol{\theta} - \boldsymbol{\theta}^*) \\ &+ o_{\mathbb{P}}(\sqrt{n}\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\| + 1)^2. \end{aligned}$$

Consistency of $\widetilde{\boldsymbol{\theta}}$ allows to derive the claims of Theorem 2.2.7 after some calculations (see Corollary 1 of [40]).

Remark 2.2.10. The term $o_{\mathbb{P}}(\sqrt{n}\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\| + 1)^2$ is a bound for

$$\begin{aligned} &\sup_{(\boldsymbol{\theta}^*, \boldsymbol{\upsilon}^\circ)\in V} \left\| nP_n \left(\nabla_{\boldsymbol{\theta}} l(\boldsymbol{\theta}^*, \boldsymbol{\upsilon}^\circ) - \widetilde{\psi}_{\mathbb{P}}\right)\right\| \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\| \\ &+ \sup_{(\boldsymbol{\theta}^*, \boldsymbol{\upsilon}^\circ)\in V} \|\breve{V}^2 - nP_n \nabla^2_{\boldsymbol{\theta}} l(\boldsymbol{\theta}^*, \boldsymbol{\upsilon}^\circ)\| \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|^2, \end{aligned}$$

which corresponds to the bound in (1.0.8) and is derived using the assumptions in (B.3)'. This is where our derivations in Chapter 4 will deviate from the arguments of [40], as we do not strive for zero sequence in probability but actual finite sample bounds for the above term that reveal the impact of the full dimension (or the complexity). This is why we do not simply assume that the crucial terms have desirable properties - in the sense of (B.3)' - but impose more specific smoothness and moment conditions that allow to derive precise statements about the deviation behavior of the term in (1.0.8).

2.2.2 Consistency of the ME

In this section we want to explain how the consistency results $I\!P(d(\tilde{\boldsymbol{\upsilon}}, \boldsymbol{\upsilon}^*) \geq \epsilon) \to 0$ can be established. These arguments are usually based on the Argmax Theorem:

Theorem 2.2.8 (Theorem 5.7 of [57]). *Let $\mathcal{L}_n : \boldsymbol{\mathcal{Y}} \times \varUpsilon \to \mathbb{R}$ be a random functional such that for every $\epsilon > 0$*

$$\sup_{\boldsymbol{\upsilon} \in \varUpsilon} |\mathcal{L}_n(\boldsymbol{\upsilon}) - \mathcal{L}^*(\boldsymbol{\upsilon})| \xrightarrow{I\!P} 0, \qquad \sup_{\upsilon:\, d(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^*) \geq \epsilon} \mathcal{L}^*(\boldsymbol{\upsilon}) < \mathcal{L}^*(\boldsymbol{\upsilon}^*).$$

Then any sequence of estimators $(\boldsymbol{\upsilon}_n)$ with $\mathcal{L}_n(\boldsymbol{\upsilon}_n) \geq \mathcal{L}_n(\boldsymbol{\upsilon}^*) - o_{I\!P}(1)$ converges in probability to $\boldsymbol{\upsilon}^*$.

In the context of i.i.d M-estimation $\mathcal{L}_n(\boldsymbol{\upsilon}) = \frac{1}{n}\sum_{i=1}^n \ell(\boldsymbol{Y}_i, \boldsymbol{\upsilon})$ and $\mathcal{L}^* = I\!E\ell$. The convergence $\sup_{\boldsymbol{\upsilon}\in\varUpsilon} |\mathcal{L}_n(\boldsymbol{\upsilon}) - \mathcal{L}^*(\boldsymbol{\upsilon})| \xrightarrow{I\!P} 0$ is usually established using empirical process theory, i.e. showing that $\{\mathcal{L}(\cdot, \boldsymbol{\upsilon}),\, \boldsymbol{\upsilon} \in \varUpsilon\}$ is $I\!P$-Glivenko-Cantelli. Then the above result applies as $\mathcal{L}_n(\tilde{\boldsymbol{\upsilon}}) \geq \mathcal{L}_n(\boldsymbol{\upsilon}^*)$ by definition. In Chapter 4 we need some consistency result of the kind

$$I\!P(d(\tilde{\boldsymbol{\upsilon}}, \boldsymbol{\upsilon}^*) \geq \mathtt{r}(\mathtt{x})) \leq \mathrm{e}^{-\mathtt{x}},$$

for a function $\mathtt{r}(\mathtt{x})$. For this purpose Theorem 2.2.8 would not suffice as it only gives convergence in probability. An alternative that would yield such a bound is the following result which only applies to the correctly specified i.i.d. maximum likelihood estimator. Let $\{p(\cdot, \boldsymbol{\upsilon}),\, \boldsymbol{\upsilon} \in \varUpsilon\}$ be the family of densities $dI\!P_{\boldsymbol{\upsilon}}/d\boldsymbol{\nu}$ of the parametric family $\{I\!P_{\boldsymbol{\upsilon}},\, \boldsymbol{\upsilon} \in \varUpsilon\}$ with respect to some dominating measure $\boldsymbol{\nu}$.

Theorem 2.2.9 (Theorem 5.8 of [29]). *Let $\varUpsilon \subset \mathbb{R}^{p^*}$ be a bounded and convex set and let the functions $p(\boldsymbol{Y}, \boldsymbol{\upsilon})$ be continuous on the closure of $\varUpsilon$ for $\boldsymbol{\nu}$-almost every $\boldsymbol{Y}$ and let the following conditions be satisfied*

1. *There exists a number $\alpha > 0$ such that*

$$\int_{\mathcal{Y}} \sup_{\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^\circ \in \varUpsilon} \frac{\sqrt{p}(\boldsymbol{Y}, \boldsymbol{\upsilon}) - \sqrt{p}(\boldsymbol{Y}, \boldsymbol{\upsilon}^\circ)}{\|\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^\circ\|^\alpha} d\boldsymbol{\nu} = A \leq \infty.$$

2. There exists a number $\beta > 0$ and a function $a(\boldsymbol{\upsilon}) > 0$ such that for all $\boldsymbol{\upsilon} \in \varUpsilon$ and all $h \in \mathbb{R}^{p^*}$ with sufficiently small norm

$$\|\sqrt{p}(\boldsymbol{\upsilon}) - \sqrt{p}(\boldsymbol{\upsilon} + h)\|^2_{L^2(\boldsymbol{\nu})} \geq a(\boldsymbol{\upsilon}) \frac{\|h\|^\beta}{1 + \|h\|^\beta}.$$

Then the MLE $\boldsymbol{\upsilon}$ satisfies for any $\lambda < \beta^{-1}$ and with constants $\mathtt{C}, \mathtt{C}_1$ that only depend on α, β, A and $\operatorname{diam}(\varUpsilon)$

$$I\!\!P\left(n^{-\lambda}\|\widetilde{\boldsymbol{\upsilon}} - \boldsymbol{\upsilon}^*\| \geq \mathfrak{z}(\mathtt{x}, \lambda)\right) \leq \mathtt{C}\mathrm{e}^{-\mathtt{x}}, \tag{2.2.3}$$

$$\mathfrak{z}(\mathtt{x}, \lambda) = \mathtt{C}_1 n^{\lambda - 1/\beta}\left(\mathtt{x} + \log(n)\left(\beta^{-1} - (2\alpha)^{-1}\right)\right)^{1/\beta} a(\boldsymbol{\upsilon})^{-1/\beta}.$$

Equation (2.2.3) looks a lot like what we desire. But there are still some problems which make us use Theorem 3.3.2 instead. The first one is that the constants $\mathtt{C}, \mathtt{C}_1$ need a finite diameter $\operatorname{diam}(\varUpsilon)$ to be finite themselves. In general this is not needed to apply Theorem 3.3.2, although in chapter 6 we will need a finite diameter in order to satisfy the conditions of Theorem 3.3.2. Another issue is that the resulting bound $\mathfrak{z}(\mathtt{x}, \lambda)$ is of a rather complicated form and would not easily allow to extract for instance the effect of the full dimension $p^* \in \mathbb{N}$ on the a priori accuracy. Finally the proof of Theorem 2.2.9 relies on the correct specification and the i.i.d. structure. Our approach in Chapter 4 is designed for general M-estimation tasks such that Theorem 2.2.9 would not be general enough, despite its appeal due to its weak conditions.

2.2.3 Sieve profile M-estimators

Obviously in many models the profile M estimator $\widetilde{\boldsymbol{\theta}} \in \mathbb{R}^p$ from Equation (1.0.3) cannot be calculated in practice if the full model is infinite dimensional. There are various ways to circumvent this problem. Next to non parametric estimation and plugin of the nuisance $\boldsymbol{\eta} \in \mathcal{X}$ a prominent approach is the so called sieve technique that we want to use in this work.

The sieve approach was introduced systematically in Chapter 8 of [22] and consists in choosing a suitable sequence of subsets $(\varUpsilon_m)_{m=1}^\infty \subset \varUpsilon$ such that for each $\boldsymbol{\upsilon} \in \varUpsilon$ there exists a sequence $\varPi_m(\boldsymbol{\upsilon}) \subset \varUpsilon_m$ with $\|\boldsymbol{\upsilon} - \varPi_m(\boldsymbol{\upsilon})\| \to 0$ as $m \to \infty$. Furthermore the sets $\varUpsilon_m \subset \varUpsilon$ have to be such that $\sup_{\boldsymbol{\upsilon} \in \varUpsilon_m} \mathcal{L}(\boldsymbol{\upsilon})$ or $\operatorname{argmax}_{\boldsymbol{\upsilon} \in \varUpsilon_m} \mathcal{L}(\boldsymbol{\upsilon})$ can be calculated in practice. In Section 4.3 we will analyze the case where $\varUpsilon = \varUpsilon_{\boldsymbol{\theta}} \times \varUpsilon_{\boldsymbol{\eta}} \subseteq \mathbb{R}^p \times \mathcal{X}$ with some infinite dimensional separable Hilbert space $\mathcal{X}$ and countable basis $\{\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots\} \subset \mathcal{X}$. In that case we set $\varUpsilon_m = \varUpsilon_{\boldsymbol{\theta}} \times \varPi_m \varUpsilon_{\boldsymbol{\eta}}$, where $\varPi_m : \mathcal{X} \to \mathcal{X}_m$ denotes the orthogonal projection onto $\mathcal{X}_m \overset{\text{def}}{=} \operatorname{span}(\boldsymbol{e}_1, \ldots, \boldsymbol{e}_m)$. This means that for each $m \in \mathbb{N}$ the sieve profile M-estimator is defined as

$$\widetilde{\boldsymbol{\theta}}_m \overset{\text{def}}{=} \Pi_{\boldsymbol{\theta}} \widetilde{\boldsymbol{\upsilon}}_m \overset{\text{def}}{=} \Pi_{\boldsymbol{\theta}} \operatorname*{argmax}_{\substack{\boldsymbol{\theta} \in \mathbb{R}^p \\ \boldsymbol{\eta} \in \mathbb{R}^m}} \mathcal{L}\left(\boldsymbol{\theta}, \sum_{k=1}^m \eta_k \boldsymbol{e}_k\right). \tag{2.2.4}$$

Clearly the size of $m \in \mathbb{N}$ has to balance the variance of the estimator which usually increases with m and the bias of the estimator which generally decreases with growing m. In asymptotic settings m consequently will depend on the sample size $n \in \mathbb{N}$, which we suppress in the notation.

As mentioned in the introduction this type of estimators is studied in [12], where it is referred to as *finite dimensional linear series estimation* in Section 2.2.3 of that work. [12] contains also results on the asymptotic properties of such estimators that we want to briefly present in the following. The first results concern the consistency of $\widetilde{\boldsymbol{\theta}}_m$ in (2.2.4). Let $\|\cdot\|_{\mathcal{Y}}$ be some norm on Υ, that is restricted by (2.2.5).

Remark 2.2.11. A natural candidate for the norm $\|\cdot\|_{\mathcal{Y}}$ is to use $\|\mathcal{D}\cdot\|$, where $\mathcal{D}^2 \overset{\text{def}}{=} -\nabla^2 I\!\!E \mathcal{L}(\boldsymbol{\upsilon}^*)$.

As in one of the first treatments of this type of estimators by [21] - dealing with sieve maximum likelihood estimators - consistency is generally rooted in three ingredients

(Identification) The functional $\mathcal{L}(\mathbb{Y}, \cdot)$ satisfies for every $m \in \mathbb{N}$

$$\inf\left\{\mathtt{r} > 0, \sup_{\boldsymbol{\upsilon} \in \Upsilon_m :\, \|\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^*\|_{\mathcal{Y}} \geq \mathtt{r}} I\!\!E \mathcal{L}(\mathbb{Y}, \boldsymbol{\upsilon}) < I\!\!E \mathcal{L}(\mathbb{Y}, \boldsymbol{\upsilon}^*)\right\} = 0.$$

(Continuity) The functional $\mathcal{L}(\mathbb{Y}, \cdot)$ is upper semi-continuous for all Υ_m and $m \in \mathbb{N}$.

(Compactness) The sets Υ_m are compact for all $m \in \mathbb{N}$.

In [12] conditions of this type lead to $\widetilde{\boldsymbol{\upsilon}}_m \xrightarrow{I\!\!P} \boldsymbol{\upsilon}^*$.

Remark 2.2.12. The continuity and compactness assumption yield that $\widetilde{\boldsymbol{\theta}}_m$ in (2.2.4) is well defined and measurable as is pointed out in Remark 2.1 of [12]. In Chapter 4 we do not need compactness of the sets Υ_m as we can ensure that the set $\widetilde{\boldsymbol{\theta}}$ is not empty and contained in a compact ball $\Upsilon_\circ \subset \mathbb{R}^{p^*}$ via condition $(\mathcal{L}\mathtt{r})$ and $(\mathcal{E}\mathtt{r})$ with Theorem 3.3.2.

[12] also gives results on the rate of convergence of the full sieve estimator $\widetilde{\boldsymbol{\upsilon}}_m$ (cf. Section 3.3). For this they assume

(Smoothness of expectation) $t \mapsto I\!\!E\mathcal{L}(\boldsymbol{\upsilon}^* + t\boldsymbol{\upsilon})$ is twice continuously differentiable with

$$\frac{d^2}{dt^2} I\!\!E\mathcal{L}(\boldsymbol{\upsilon}^* + t(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^*))|_{t=0} \asymp -\|\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^*\|_{\mathcal{Y}}^2.$$

Together with the assumptions to ensure that the sieve estimator is consistent this already allows to infer that

$$\|\boldsymbol{\upsilon}^* - \boldsymbol{\upsilon}_m^*\|_{\mathcal{Y}} \le \mathtt{C}\|\boldsymbol{\upsilon}^* - \mathit{\Pi}_m^{\mathcal{Y}}\boldsymbol{\upsilon}^*\|_{\mathcal{Y}},$$

where $\mathit{\Pi}_m^{\mathcal{Y}}\boldsymbol{\upsilon}^* \in \varUpsilon_m$ denotes the closest element of $\varUpsilon_m$ to $\boldsymbol{\upsilon}^* \in \mathbb{R}^p \times \mathcal{X}$ in the metric induced by $\|\cdot\|_{\mathcal{Y}}$ and where

$$\boldsymbol{\upsilon}_m^* \stackrel{\text{def}}{=} \operatorname*{argmax}_{\boldsymbol{\upsilon}\in\varUpsilon_m} I\!\!E\mathcal{L}(\boldsymbol{\upsilon}).$$

To get rates for the estimator $\widetilde{\boldsymbol{\upsilon}}_m$ it remains to control the random component. For this [12] assume

(Derivatives) The Fréchet-derivative $\nabla\mathcal{L}(\boldsymbol{\upsilon}^*) \in \mathcal{X}^*$ of $\mathcal{L} : \varUpsilon \to \mathbb{R}$ in $\boldsymbol{\upsilon}^* \in \varUpsilon$ exists.

Furthermore they impose smoothness and deviation constraints (Condition 3.13 of [12]) on the gradient $\nabla_m\mathcal{L}(\boldsymbol{\upsilon}_m^*)$ defined as

$$\nabla_m\mathcal{L}(\boldsymbol{\upsilon}_m^*)\boldsymbol{\gamma} \stackrel{\text{def}}{=} \frac{d}{dt}\mathcal{L}(\boldsymbol{\upsilon}_m^* + t\boldsymbol{\gamma})|_{t=0},\ \boldsymbol{\gamma} \in \varUpsilon_m.$$

Theorem 3.5 of [12] states that under such conditions

$$\|\widetilde{\boldsymbol{\upsilon}}_m - \boldsymbol{\upsilon}^*\|_{\mathcal{Y}} = O_{I\!\!P}\left(\sqrt{m} + \|\mathit{\Pi}_m^{\mathcal{Y}}\boldsymbol{\upsilon}^* - \boldsymbol{\upsilon}^*\|_{\mathcal{Y}}\right).$$

Concerning the asymptotic distribution of $\sqrt{n}(\widetilde{\boldsymbol{\theta}}_m - \boldsymbol{\theta}^*)$ [12] give a result that is based on [50]. It involves a list of conditions, which adapted to fit to our setting reads as follows:

(Rate) The full sieve estimator $\widetilde{\boldsymbol{\upsilon}}_m$ in (2.2.4) satisfies with some $\mathtt{r}_m > 0$

$$I\!\!P\left(\widetilde{\boldsymbol{\upsilon}}_m \in \varUpsilon_{\circ,m}(\mathtt{r}_m)\right) \to 1,$$

where

$$\varUpsilon_{\circ,m}(\mathtt{r}_m) \stackrel{\text{def}}{=} \{\boldsymbol{\upsilon} \in \varUpsilon_m,\ \|\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^*\|_{\mathcal{Y}} < \mathtt{r}_m\}.$$

(Stochastic equicontinuity) With $\boldsymbol{\upsilon}^{\circ}(\boldsymbol{\upsilon}) = \boldsymbol{\upsilon} + \varepsilon_n(\boldsymbol{\theta}, 0)$ for any $\boldsymbol{\theta} \in \mathbb{R}^p$ and $\varepsilon_n = o(n^{-1/2})$

$$\sup_{\boldsymbol{\upsilon} \in \varUpsilon_{\circ,m}(\mathtt{r}_m)} (1 - \mathbb{E}) \left(\mathcal{L}(\boldsymbol{\upsilon}) - \mathcal{L}(\boldsymbol{\upsilon}^{\circ}) - \nabla\mathcal{L}(\boldsymbol{\upsilon}^*)(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^{\circ})\right) = o_{\mathbb{P}}(1).$$

(Expectation of criterion) It holds that

$$\sup_{\boldsymbol{\upsilon} \in \varUpsilon_{\circ,m}(\mathtt{r}_m)} \left\{ \mathbb{E}\mathcal{L}(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^*) + \frac{1}{2}\|\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^*\|^2_{\mathcal{Y}} \right\} = o(\mathtt{r}_m^2). \qquad (2.2.5)$$

(Approximation accuracy) With some positive $h_m \to 0$

$$|h_m|\|\boldsymbol{\upsilon}^* - \varPi_m^{\mathcal{Y}}\boldsymbol{\upsilon}^*\|^2_{\mathcal{Y}} = O(\sqrt{n}).$$

(Gradient) The linear operator $\nabla\mathcal{L}(\boldsymbol{\upsilon}^*)$ satisfies

$$\sup_{\boldsymbol{\upsilon} \in \varUpsilon_{\circ,m}(\mathtt{r}_m)} (1 - \mathbb{E})\, \nabla\mathcal{L}(\boldsymbol{\upsilon}^*)(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^*) = o(\sqrt{n}).$$

A very important condition - which we cite separately for that reason - is that the norm $\|\cdot\|_{\mathcal{Y}}$ induces an inner product $\langle\cdot,\cdot\rangle_{\mathcal{Y}}$ on $\varUpsilon - \boldsymbol{\upsilon}^*$ and that the sieve basis $(\boldsymbol{e}_k)_{k\in\mathbb{N}}$ satisfies

$$\langle \boldsymbol{e}_k, \boldsymbol{e}_l \rangle_{\mathcal{Y}} = \delta_{l,k}. \qquad (2.2.6)$$

Remark 2.2.13. We are not completely precise here as $(\boldsymbol{e}_k)_{k\in\mathbb{N}}$ is a basis for $\mathcal{X}$ but not $\mathbb{R}^p \times \mathcal{X}$. A complete sieve basis is $\{\boldsymbol{b}_1 \times 0, \ldots, \boldsymbol{b}_p \times 0, 0 \times \boldsymbol{e}_1, 0 \times \boldsymbol{e}_2, \ldots\}$, where $\boldsymbol{b}_1, \ldots, \boldsymbol{b}_p$ is a basis for $\mathbb{R}^p$. The above condition means that such a pair of bases $(\boldsymbol{b}_k)_{k\le p}, (\boldsymbol{e}_k)_{k\in\mathbb{N}}$ has to be chosen such that the resulting complete sieve basis is orthogonal in the inner product $\langle\cdot,\cdot\rangle_{\mathcal{Y}}$.

[50] present the following result:

Theorem 2.2.10 (Corollary 2, [50])**.** *If the above list of conditions holds, if* $\operatorname{Var}(\nabla_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\upsilon}^*)) < \infty$ *and if the basis satisfies* (2.2.6)*, then*

$$\sqrt{n}(\widetilde{\boldsymbol{\theta}}_{m(n)} - \boldsymbol{\theta}^*) \to \mathcal{N}(0, \operatorname{Var}(\nabla_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\upsilon}^*))).$$

Remark 2.2.14. In [49] a Wilks type result for sieve M-Estimation is derived in a quite general setting under similar conditions.

Unfortunately condition (2.2.6) is not easily satisfied in practice as the inner product $\langle\cdot,\cdot\rangle_{\mathcal{Y}}$ induced by condition (Expectation of criterion) may depend on the unknown true parameter. We will encounter such an example in Chapter 6 namely the model from Equation (1.0.1). If a general basis is used without prior knowledge of the norm from Equation (1.0.1), one needs more conditions as we want to show with the following example.

Let with i.i.d $\varepsilon_i \in \mathbb{R}$

$$Y_i = \boldsymbol{e}_i^\top \boldsymbol{\upsilon}^* + \varepsilon_i, \quad \boldsymbol{\upsilon}^* \in l^2, \quad (\boldsymbol{e}_i)_{i=1}^n \in l^2, \text{ linearly independent.}$$

Let

$$\frac{1}{n}\mathcal{D}^2 \overset{\text{def}}{=} \frac{1}{n}\sum_{i=1}^n \boldsymbol{e}_i \boldsymbol{e}_i^\top \sim d^2 = \begin{pmatrix} d_m^2 & a_{\boldsymbol{\varkappa}\boldsymbol{\upsilon}_m}^\top \\ a_{\boldsymbol{\varkappa}\boldsymbol{\upsilon}_m} & h_{\boldsymbol{\varkappa}\boldsymbol{\varkappa}}^2 \end{pmatrix},$$

where $d_m^2 \in \mathbb{R}^{m\times m}$, $a_{\boldsymbol{\varkappa}\boldsymbol{\upsilon}_m} : \mathbb{R}^m \to l^2$. Define the sieve ME

$$\begin{aligned} \widetilde{\boldsymbol{\upsilon}}_m &\overset{\text{def}}{=} \operatorname*{argmax}_{\boldsymbol{\upsilon}\in\mathbb{R}^m} \mathcal{L}_m(\boldsymbol{\upsilon}) = \operatorname*{argmax}_{\boldsymbol{\upsilon}\in\mathbb{R}^m} \sum_{i=1}^n (Y_i - \boldsymbol{e}_i^\top \boldsymbol{\upsilon})^2 \\ &= \left(\frac{1}{n}\sum_{i=1}^n \mathit{\Pi}_m \boldsymbol{e}_i \boldsymbol{e}_i^\top \mathit{\Pi}_m^\top\right)^{-1} \left(\frac{1}{n}\sum_{i=1}^n \boldsymbol{e}_i \boldsymbol{e}_i^\top\right) \boldsymbol{\upsilon}^* \\ &\quad + \left(\frac{1}{n}\sum_{i=1}^n \mathit{\Pi}_m \boldsymbol{e}_i \boldsymbol{e}_i^\top \mathit{\Pi}_m^\top\right)^{-1} \frac{1}{n}\sum_{i=1}^n \varepsilon_i \\ &\approx \mathit{\Pi}_m \boldsymbol{\upsilon}^* + \left(\frac{1}{n}\sum_{i=1}^n \mathit{\Pi}_m \boldsymbol{e}_i \boldsymbol{e}_i^\top \mathit{\Pi}_m^\top\right)^{-1} \frac{1}{n}\sum_{i=1}^n \varepsilon_i + d_m^{-2} a_{\boldsymbol{\varkappa}\boldsymbol{\upsilon}_m} \boldsymbol{\varkappa}^*. \end{aligned}$$

If $\boldsymbol{\theta} = \mathit{\Pi}_{\operatorname{span}\{\boldsymbol{e}_1\}}\boldsymbol{\upsilon}$ - where $(\boldsymbol{e}_k)_{k\in\mathbb{N}} \subset l^2$ is the canonical basis - this gives

$$\sqrt{n}(\widetilde{\boldsymbol{\theta}}_m - \boldsymbol{\theta}^*) = \sqrt{n}\mathit{\Pi}_1 d_m^{-2} a_{\boldsymbol{\varkappa}\boldsymbol{\upsilon}_m} \boldsymbol{\varkappa}^* + \mathit{\Pi}_1 \left(\frac{1}{n}\sum_{i=1}^n \mathit{\Pi}_m \boldsymbol{e}_i \boldsymbol{e}_i^\top \mathit{\Pi}_m^\top\right)^{-1} \frac{1}{n}\sum_{i=1}^n \varepsilon_i.$$

Thus to get asymptotic normality one needs that $\sqrt{n}\mathit{\Pi}_1 d_m^{-2} a_{\boldsymbol{\varkappa}\boldsymbol{\upsilon}_m} \boldsymbol{\varkappa}^* \to 0$, which is not implied by any of the conditions of [50] except (2.2.6). We address such effects in Section 4.3.3. In Chapter 6 we present a basis that - under mild regularity conditions - ensures $\sqrt{n}\mathit{\Pi}_1 d_m^{-2} a_{\boldsymbol{\varkappa}\boldsymbol{\upsilon}_m} \boldsymbol{\varkappa}^* \to 0$ in the context of the model (1.0.1), without any prior knowledge about the true data distribution.

2.A Proof of differentiability in quadratic mean for the single-index model

Proof. We prove this claim using the arguments of the proof of Lemma 7.6 in [57]. By definition $p_t(y, \boldsymbol{x}) = p_\epsilon(y - f_{\boldsymbol{\eta}+th_{\boldsymbol{\eta}}}(\boldsymbol{x}^\top \boldsymbol{\theta}(t)))p_{\mathbf{X}}(\boldsymbol{x})$. We can bound

$$\int \left(\frac{\sqrt{p_t(y,\boldsymbol{x})} - \sqrt{p_0(y,\boldsymbol{x})}}{t} - \frac{1}{2} g(y,\boldsymbol{x}) \sqrt{p_0(y,\boldsymbol{x})} \right)^2 d(y,\boldsymbol{x})$$

$$\leq 2 \int \left(\frac{\sqrt{p_t(y,\boldsymbol{x})} - \sqrt{p_0(y,\boldsymbol{x})}}{t} \right)^2 d(y,\boldsymbol{x})$$

$$+ \int g(y,\boldsymbol{x})^2 p_0(y) d(y,\boldsymbol{x}).$$

Denote

$$\widehat{g}(\boldsymbol{x},s) \overset{\text{def}}{=} \left(\dot{f}_{\boldsymbol{\eta}^*+sh_{\boldsymbol{\eta}}}(\boldsymbol{x}^\top \boldsymbol{\theta}(s)) \dot{\boldsymbol{\theta}}(s)^\top \boldsymbol{x} + f_{h_{\boldsymbol{\eta}}}(\boldsymbol{x}^\top \boldsymbol{\theta}^*) \right) \sqrt{p_{\mathbf{X}}(\boldsymbol{x})}.$$

We have by pointwise differentiation

$$\left. \frac{d\sqrt{p_t(y,\boldsymbol{x})}}{dt} \right|_{t=s} = \frac{\dot{p_\epsilon}(y - f_{\boldsymbol{\eta}^*+sh_{\boldsymbol{\eta}}}(\boldsymbol{x}^\top \boldsymbol{\theta}(s)))}{\sqrt{p_\epsilon(y - f_{\boldsymbol{\eta}^*+sh_{\boldsymbol{\eta}}}(\boldsymbol{x}^\top \boldsymbol{\theta}(s)))}} \widehat{g}(\boldsymbol{x})$$

and by the continuity of the derivative

$$\sqrt{p_t(y,\boldsymbol{x})} - \sqrt{p_0(y,\boldsymbol{x})} = \int_0^t \frac{\dot{p_\epsilon}(y - f_{\boldsymbol{\eta}^*+sh_{\boldsymbol{\eta}}}(\boldsymbol{x}^\top \boldsymbol{\theta}(s)))}{\sqrt{p_\epsilon(y - f_{\boldsymbol{\eta}^*+sh_{\boldsymbol{\eta}}}(\boldsymbol{x}^\top \boldsymbol{\theta}(s)))}} \widehat{g}(\boldsymbol{x},s) ds$$

$$= t \int_0^1 \frac{\dot{p_\epsilon}(y - f_{\boldsymbol{\eta}^*+rh_{\boldsymbol{\eta}}}(\boldsymbol{x}^\top \boldsymbol{\theta}(r)))}{\sqrt{p_\epsilon(y - f_{\boldsymbol{\eta}^*+rh_{\boldsymbol{\eta}}}(\boldsymbol{x}^\top \boldsymbol{\theta}(r)))}} \widehat{g}(\boldsymbol{x},s) dr.$$

This gives with Jensen's inequality, Fubini's theorem and assumptions (2.1.6) and (2.1.7)

$$\int \left(\frac{\sqrt{p_t(y,\boldsymbol{x})} - \sqrt{p_0(y,\boldsymbol{x})}}{t} \right)^2 d(y,\boldsymbol{x})$$

$$\leq \int \int_0^1 \left(\frac{\dot{p_\epsilon}(y - f_{\boldsymbol{\eta}^*+t_0h_{\boldsymbol{\eta}}}(\boldsymbol{x}^\top \boldsymbol{\theta}(s)))}{\sqrt{p_\epsilon(y - f_{\boldsymbol{\eta}^*+sh_{\boldsymbol{\eta}}}(\boldsymbol{x}^\top \boldsymbol{\theta}(s)))}} \widehat{g}(\boldsymbol{x}) \right)^2 ds d(y,\boldsymbol{x})$$

$$= \int_0^1 \int \left(\frac{\dot{p_\epsilon}(y - f_{\boldsymbol{\eta}^*+t_0h_{\boldsymbol{\eta}}}(\boldsymbol{x}^\top \boldsymbol{\theta}(s)))}{\sqrt{p_\epsilon(y - f_{\boldsymbol{\eta}^*+sh_{\boldsymbol{\eta}}}(\boldsymbol{x}^\top \boldsymbol{\theta}(s)))}} \widehat{g}(\boldsymbol{x},s) \right)^2 d(y,\boldsymbol{x}) ds$$

$$= \int_0^1 \int_{\mathbb{R}^p} \int_{\mathbb{R}} \left(\frac{\dot{p_\epsilon}(y - f_{\boldsymbol{\eta}^*+t_0h_{\boldsymbol{\eta}}}(\boldsymbol{x}^\top \boldsymbol{\theta}(s)))}{\sqrt{p_\epsilon(y - f_{\boldsymbol{\eta}^*+sh_{\boldsymbol{\eta}}}(\boldsymbol{x}^\top \boldsymbol{\theta}(s)))}} \right)^2 dy \, \widehat{g}(\boldsymbol{x},s)^2 d\boldsymbol{x} \, ds < \infty$$

Furthermore again using assumptions (2.1.6) and (2.1.7)

$$\int g(y,\boldsymbol{x})^2 p_0(y) d(y,\boldsymbol{x})$$

$$= \int_0^1 \int_{\mathbb{R}^p} \int_{\mathbb{R}} \left(\frac{\dot{p_\epsilon}(y - f_{\boldsymbol{\eta}^*}(\boldsymbol{x}^\top \boldsymbol{\theta}^*))}{\sqrt{p_\epsilon(y - f_{\boldsymbol{\eta}^*}(\boldsymbol{x}^\top \boldsymbol{\theta}^*))}} \right)^2 dy\, \widehat{g}(\boldsymbol{x},0)^2 d\boldsymbol{x}\, ds < \infty.$$

Clearly we have almost everywhere pointwise convergence

$$\frac{\sqrt{p_t(y,\boldsymbol{x})} - \sqrt{p_0(y,\boldsymbol{x})}}{t} \to \frac{1}{2} g(y,\boldsymbol{x}) \sqrt{p_0(y,\boldsymbol{x})}.$$

This gives the claim with the dominated convergence theorem. □

Chapter 3

Parametric estimation, finite sample theory

In this chapter we want to give a synopsis of the results of [52]. Many ideas and tools of this paper are used for our approach presented in Chapter 4, so we present the results in some detail. Also in the Sections 3.5.2 and 3.5.3 we present two new Theorems that are derived with the empirical process techniques of [52] and that are central for the subsequent chapters.

3.1 Basic idea

[52] deals with M-estimators in parametric models, i.e. the arguments of the functional $\mathcal{L}$ from equation (1.0.2) are finite dimensional objects $\boldsymbol{\upsilon} \in \varUpsilon \subset \mathbb{R}^{p^*\times p^*}$. Remember the definitions of the full target parameter and the M-Estimator (ME)

$$\boldsymbol{\upsilon}^* \stackrel{\text{def}}{=} \operatorname*{argmax}_{\boldsymbol{\upsilon}\in\varUpsilon} I\!\!E_{I\!\!P}\mathcal{L}(\boldsymbol{\upsilon}), \quad \widetilde{\boldsymbol{\upsilon}} \stackrel{\text{def}}{=} \operatorname*{argmax}_{\boldsymbol{\upsilon}\in\varUpsilon} \mathcal{L}(\boldsymbol{\upsilon}). \tag{3.1.1}$$

Introduce the functional gap

$$\mathcal{L}(\boldsymbol{\upsilon},\boldsymbol{\upsilon}^*) = \mathcal{L}(\boldsymbol{\upsilon}) - \mathcal{L}(\boldsymbol{\upsilon}^*), \tag{3.1.2}$$

and define the total information matrix $\mathcal{D}_0^2 = -\nabla^2 I\!\!E\mathcal{L}(\boldsymbol{\upsilon}^*) \in \mathbb{R}^{p^*}$. The results of [52] can be summarized as

$$\left|\mathcal{L}(\widetilde{\boldsymbol{\upsilon}},\boldsymbol{\upsilon}^*) - \|\boldsymbol{\xi}_{\boldsymbol{\epsilon}}\|^2/2\right| \le \varDelta_{\boldsymbol{\epsilon}}(\mathtt{x}), \tag{3.1.3}$$

$$\left\|\mathcal{D}_{\boldsymbol{\epsilon}}(\widetilde{\boldsymbol{\upsilon}} - \boldsymbol{\upsilon}^*) - \boldsymbol{\xi}_{\boldsymbol{\epsilon}}\right\|^2 \le 2\varDelta_{\boldsymbol{\epsilon}}(\mathtt{x}), \tag{3.1.4}$$

where $\mathcal{D}_{\boldsymbol{\epsilon}}^2 \approx \mathcal{D}_0^2$, $\boldsymbol{\xi}_{\boldsymbol{\epsilon}} \stackrel{\text{def}}{=} \mathcal{D}_{\boldsymbol{\epsilon}}^{-1}\nabla\mathcal{L}(\boldsymbol{\upsilon}^*)$, and $\varDelta_{\boldsymbol{\epsilon}}(\mathtt{x})$ is a random term called the *spread* which is small with probability greater $1-4\mathrm{e}^{-\mathtt{x}}$. In the smooth i.i.d. case $\varDelta_{\boldsymbol{\epsilon}}(\mathtt{x})$ is of order $p^{*3/2}/n^{1/2}$, where p^{*} is the total parameter dimension. If the model is correctly specified, which means that $\mathcal{L}(\mathbb{Y},\boldsymbol{\upsilon}) = \log\prod_{i=1}^{n} p(\boldsymbol{Y}_i,\boldsymbol{\upsilon})$, where $p(y_i,\boldsymbol{\upsilon})$ is the the density of $I\!\!P_{\boldsymbol{\upsilon}}$, $\boldsymbol{\xi}_{\boldsymbol{\epsilon}}$ is nearly standard normal such that $2\mathcal{L}(\widetilde{\boldsymbol{\upsilon}},\boldsymbol{\upsilon}^{*})$ is nearly $\chi^2_{p^*}$ if $p^{*3/2}/n^{1/2}$ is small. The results also allow to infer that the MLE $\widetilde{\boldsymbol{\upsilon}}$ is asymptotically normal and efficient.

This result is derived in two steps. First it is proven that the ME lies with a high probability in a neighborhood of the target $\boldsymbol{\upsilon}^{*} \in \varUpsilon$. In the second step the functional gap of equation (3.1.2) is sandwiched by two quadratic processes motivated by a second order Taylor expansion. For some radius $\mathtt{r} > 0$ the local neighborhood $\varUpsilon_{\circ}(\mathtt{r})$ is defined as

$$\varUpsilon_{\circ}(\mathtt{r}) \overset{\mathrm{def}}{=} \Bigl\{\boldsymbol{\upsilon} \in \mathbb{R}^{p^*} : \|\mathcal{D}_0(\boldsymbol{\upsilon}-\boldsymbol{\upsilon}^{*})\| \le \mathtt{r}\Bigr\}, \tag{3.1.5}$$

which is a ball in the intrinsic norm $\|\mathcal{D}_0(\cdot)\|$. [52] derives a deviation bound of the form

$$I\!\!P(\widetilde{\boldsymbol{\upsilon}} \in \varUpsilon_{\circ}(\mathtt{r}_0(\mathtt{x}))) \ge 1 - \mathrm{e}^{-\mathtt{x}}, \tag{3.1.6}$$

where $\mathtt{r}_0(\mathtt{x})$ grows almost linearly with $\mathtt{x}$.

On this local neighborhood one could approximate

$$\mathcal{L}(\boldsymbol{\upsilon},\boldsymbol{\upsilon}^{*}) = \nabla\mathcal{L}(\boldsymbol{\upsilon}^{*})(\boldsymbol{\upsilon}-\boldsymbol{\upsilon}^{*}) - \|\mathcal{D}_0(\boldsymbol{\upsilon}-\boldsymbol{\upsilon}^{*})\|^2/2 + \alpha_0(\boldsymbol{\upsilon},\boldsymbol{\upsilon}^{*}),$$

where

$$\alpha_0(\boldsymbol{\upsilon},\boldsymbol{\upsilon}^{*}) = \mathcal{L}(\boldsymbol{\upsilon},\boldsymbol{\upsilon}^{*}) - \nabla\mathcal{L}(\boldsymbol{\upsilon}^{*})(\boldsymbol{\upsilon}-\boldsymbol{\upsilon}^{*}) + \|\mathcal{D}_0(\boldsymbol{\upsilon}-\boldsymbol{\upsilon}^{*})\|^2/2. \tag{3.1.7}$$

The remaining task would be to derive deviation bounds for $\sup_{\boldsymbol{\upsilon}\in\varUpsilon_{\circ}(\mathtt{r})}$ $|\alpha_0(\boldsymbol{\upsilon},\boldsymbol{\upsilon}^{*})|$. This is possible using conditions $(\mathcal{L}_0)$ and $(\mathcal{E}\mathcal{D}_1)$ of Section 4.2.1 and would lead to a bound of the kind (see proof of Theorem 4.2.2 for the case that $\boldsymbol{\theta} = \boldsymbol{\upsilon} \in \mathbb{R}^{p^*}$)

$$I\!\!P\left\{\sup_{\boldsymbol{\upsilon}\in\varUpsilon(\mathtt{r})} |\alpha_0(\boldsymbol{\upsilon},\boldsymbol{\upsilon}^{*})| \ge \mathtt{r}^2\mathtt{C}\left(\delta(\mathtt{r}) + \omega\sqrt{p^{*}+\mathtt{x}}\right)\right\} \le \mathrm{e}^{-\mathtt{x}},$$

where δ,ω are small terms from those conditions. It is important to note that the error term is of order $\mathtt{r}^2\delta(\mathtt{r}) + \mathtt{r}^2\sqrt{p^*}\omega$, which could be too large for big values of $\mathtt{r}$. It turns out that a small trick enhances these bounds substantially in situations where $\mathtt{r} > 0$ is large in comparison to $p^{*} \in \mathbb{N}$. Instead of approximating the functional by a quadratic term and accounting for the uniform error $\sup_{\boldsymbol{\upsilon}\in\varUpsilon_{\circ}(\mathtt{r})}|\alpha_0(\boldsymbol{\upsilon},\boldsymbol{\upsilon}^{*})|$, [52] sandwiches the process $\mathcal{L}(\boldsymbol{\upsilon},\boldsymbol{\upsilon}^{*})$ between two different quadratic processes $\mathbb{L}_{\underline{\boldsymbol{\epsilon}}}, \mathbb{L}_{\boldsymbol{\epsilon}}$ up to uniform errors that are substantially smaller than $\sup_{\boldsymbol{\upsilon}\in\varUpsilon(\mathtt{r})}|\alpha_0(\boldsymbol{\upsilon},\boldsymbol{\upsilon}^{*})|$. To gain insight into the properties of the maximizer and maximum of $\mathcal{L}(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^*)$ with respect to $\boldsymbol{\upsilon} \in \varUpsilon$, [52] analyzes and compares the maximizers and maxima of $\mathbb{L}_{\underline{\boldsymbol{\epsilon}}}, \mathbb{L}_{\boldsymbol{\epsilon}}$, which again differ by less than $\sup_{\boldsymbol{\upsilon} \in \varUpsilon(\mathtt{r})} |\alpha_0(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^*)|$. When accounting for all approximation errors this leads to sharper results in terms of how sensitive the error terms are to the size of $\mathtt{r}$. [52] uses the following altered approximation

$$\begin{aligned} &\mathcal{L}(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^*) \qquad\qquad (3.1.8)\\ &\quad \leq \nabla\mathcal{L}(\boldsymbol{\upsilon}^*)(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^*) - \frac{1}{2}(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^*)^\top \{(1-\delta)\mathcal{D}_0^2 - \omega\mathcal{V}_0^2\}(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^*)\\ &\qquad +\alpha_\epsilon(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^*), \end{aligned}$$

where

$$\alpha_\epsilon(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^*) = (\mathcal{L} - I\!E\mathcal{L})(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^*) - \nabla\mathcal{L}(\boldsymbol{\upsilon}^*)(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^*) - \omega\|\mathcal{V}_0(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^*)\|^2/2,$$

and $\mathcal{V}_0 \overset{\text{def}}{=} I\!E[\nabla\mathcal{L}(\boldsymbol{\upsilon})\nabla\mathcal{L}(\boldsymbol{\upsilon})^\top]$. The inequality (3.1.8) is valid as [52] assumes that

$$|I\!E\mathcal{L}(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^*) - \|\mathcal{D}_0(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^*)\|^2/2| \leq \delta\|\mathcal{D}_0(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^*)\|^2/2.$$

One can bound due to $I\!E\nabla\mathcal{L}(\boldsymbol{\upsilon}^*) = 0$

$$\begin{aligned} \alpha_\epsilon(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^*) \leq \sup_{\substack{\boldsymbol{\gamma} \in \mathbb{R}^{p^*} \\ \|\boldsymbol{\gamma}\| = \mathtt{r}}} \sup_{\boldsymbol{\upsilon}^\circ \in \varUpsilon_\circ(\mathtt{r})} &\mathcal{D}_0^{-1}\left\{\nabla(\mathcal{L} - I\!E\mathcal{L})(\boldsymbol{\upsilon}^\circ) - \nabla(\mathcal{L} - I\!E\mathcal{L})(\boldsymbol{\upsilon}^*)\right\}\boldsymbol{\gamma}\\ &- \omega\|\mathcal{V}_0\mathcal{D}_0^{-1}\boldsymbol{\gamma}\|^2/2. \end{aligned}$$

The important difference to bounding $\sup_{\boldsymbol{\upsilon} \in \varUpsilon_\circ(\mathtt{r})} |\alpha_0(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^*)|$ is that the additional quadratic drift component $-\omega\|\mathcal{V}_0\mathcal{D}_0^{-1}\boldsymbol{\gamma}\|^2/2$ allows to derive - utilizing Theorem 3.5.6 - a bound for α_ϵ which is of order $\omega(\mathtt{r})p^*$. The dependence on the radius now is only through $\omega(\mathtt{r})$ which in many settings is linear in $\mathtt{r}$. This can make a tremendous difference if $\mathtt{r}_0 > 0$ from equation (3.1.6) is large in comparison to $p^* \in \mathbb{N}$. The same is done for an upper bound using $(1+\delta)\mathcal{D}_0^2 + \omega\mathcal{V}_0^2$ in (3.1.8) instead. This leads to the key result of [52], that the functional gap $\mathcal{L}(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^*)$ can be sandwiched on $\varUpsilon_\circ(\mathtt{r})$ by two processes $\mathbb{L}_{\boldsymbol{\epsilon}}(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^*)$ and $\mathbb{L}_{\underline{\boldsymbol{\epsilon}}}(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^*)$ that are quadratic in $\boldsymbol{\upsilon}$, that is for $\boldsymbol{\upsilon} \in \varUpsilon_\circ(\mathtt{r})$

$$\mathbb{L}_{\underline{\boldsymbol{\epsilon}}}(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^*) - \diamondsuit_{\underline{\boldsymbol{\epsilon}}}(\mathtt{r}) \leq \mathcal{L}(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^*) \leq \mathbb{L}_{\boldsymbol{\epsilon}}(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^*) + \diamondsuit_{\boldsymbol{\epsilon}}(\mathtt{r}), \qquad (3.1.9)$$

where $\diamondsuit_{\boldsymbol{\epsilon}}(\mathtt{r}) > 0$ and $\diamondsuit_{\underline{\boldsymbol{\epsilon}}}(\mathtt{r}) > 0$ are with high probability of order $\omega(\mathtt{r})p^*$.

Furthermore it turns out that

$$
\begin{aligned}
|\max_{\Upsilon_\circ(\mathtt{r})} \mathbb{L}_{\underline{\boldsymbol{\epsilon}}}(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^*) - \max_{\Upsilon_\circ(\mathtt{r})} \mathbb{L}_{\boldsymbol{\epsilon}}(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^*)| &\le \mathtt{C}(\delta(\mathtt{r}) + \omega(\mathtt{r}))p^*, \\
|\operatorname*{argmax}_{\Upsilon_\circ(\mathtt{r})} \mathbb{L}_{\underline{\boldsymbol{\epsilon}}}(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^*) - \operatorname*{argmax}_{\Upsilon_\circ(\mathtt{r})} \mathbb{L}_{\boldsymbol{\epsilon}}(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^*)| &\le \mathtt{C}(\delta(\mathtt{r}) + \omega(\mathtt{r}))p^*.
\end{aligned}
$$

The bracketing result (3.1.9) and the last two equations combined with the local concentration of the M-estimator (3.1.6) give (3.1.3) and (3.1.4), which again yield a number of important and informative corollaries.

Remark 3.1.1. We will also exploit this improvement in comparison to the bounds for $\sup_{\boldsymbol{\upsilon}\in\Upsilon(\mathtt{r})} |\alpha_0(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^*)|$ from (3.1.7) in Remark 4.2.15.

3.2 Wilks and Fisher via local quadratic bracketing

In the following we will briefly present the arguments employed in [52] that lead to the results (3.1.3) and (3.1.4). We do this in some detail to highlight what is new and different in our approach in the subsequent Chapters.

3.2.1 Conditions

Below we cite the conditions that are used in [52]. We will not explain them here in detail as we will restate and discuss a slightly altered list of conditions in Chapter 4 that is more relevant to our work.

Local conditions

Local conditions describe the properties of $\mathcal{L}(\boldsymbol{\upsilon})$ in a vicinity of the central point $\boldsymbol{\upsilon}^*$ from (3.1.1).

Define the stochastic component $\zeta(\boldsymbol{\upsilon})$:

$$
\zeta(\boldsymbol{\upsilon}) \overset{\text{def}}{=} \mathcal{L}(\boldsymbol{\upsilon}) - I\!\!E\mathcal{L}(\boldsymbol{\upsilon}).
$$

Below we suppose that the random function $\zeta(\boldsymbol{\upsilon})$ is differentiable in $\boldsymbol{\upsilon}$ and its gradient $\nabla\zeta(\boldsymbol{\upsilon}) = \partial\zeta(\boldsymbol{\upsilon})/\partial\boldsymbol{\upsilon} \in \mathbb{R}^{p^*}$ has some exponential moments. Our first condition describes the property of the gradient $\nabla\zeta(\boldsymbol{\upsilon}^*)$ at the central point $\boldsymbol{\upsilon}^*$.

$(\mathcal{E}\mathfrak{D}_0)$ *There exist a positive symmetric matrix* $\mathcal{V}_0^2$, *and constants* $\mathtt{g} > 0$, $\nu_0 \ge 1$ *such that* $\operatorname{Var}\bigl\{\nabla\zeta(\boldsymbol{\upsilon}^*)\bigr\} \le \mathcal{V}_0^2$ *and for all* $|\lambda| \le \mathtt{g}$

$$
\sup_{\boldsymbol{\gamma}\in\mathbb{R}^p} \log I\!\!E \exp\biggl\{\lambda \frac{\boldsymbol{\gamma}^\top \nabla\zeta(\boldsymbol{\upsilon}^*)}{\|\mathcal{V}_0\boldsymbol{\gamma}\|}\biggr\} \le \nu_0^2\lambda^2/2.
$$

The following two conditions are restricted to neighborhoods $\varUpsilon_{\circ}(\mathtt{r}) \subset \varUpsilon$ from (3.1.5).

($\mathcal{E}\mathcal{D}_1'$) *For some* $\mathtt{r}^* > 0$ *and each* $\mathtt{r} \le \mathtt{r}^*$, *there exists a constant* $\omega(\mathtt{r}) \le 1/2$ *such that it holds for all* $\boldsymbol{\upsilon} \in \varUpsilon_{\circ}(\mathtt{r})$

$$\sup_{\gamma \in \mathbb{R}^p} \log I\!\!E \exp\biggl\{ \lambda \frac{\gamma^\top \{\nabla\zeta(\boldsymbol{\upsilon}) - \nabla\zeta(\boldsymbol{\upsilon}^*)\}}{\omega(\mathtt{r}) \|\mathcal{D}_0 \gamma\|} \biggr\} \le \nu_0^2 \lambda^2/2, \qquad |\lambda| \le \mathtt{g}.$$

Here the constant $\mathtt{g}$ is the same as in (ED_0).

($\mathcal{L}_0'$) *There are a symmetric strictly positive-definite matrix* $\mathcal{D}_0^2$ *and forsome* $\mathtt{r}^* > 0$ *and each* $\mathtt{r} \le \mathtt{r}^*$ *a constant* $\delta(\mathtt{r}) \le 1/2$, *such that it holds on the set* $\varUpsilon_{\circ}(\mathtt{r})$

$$\left| 1 + \frac{2 I\!\!E \mathcal{L}(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^*)}{\|\mathcal{D}_0(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^*)\|^2} \right| \le \delta(\mathtt{r}).$$

Remark 3.2.1. We denote these conditions as $(\mathcal{L}_0')$ and $(\mathcal{E}\mathcal{D}_1')$ because we will introduce variants of these in Chapter 4. The new versions $(\mathcal{L}_0)$ and $(\mathcal{E}\mathcal{D}_1)$ will be slightly stronger but allow the mentioned improvement from $p^{*3/2}/n^{1/2}$ to $p^*/n^{1/2}$ of the bounds for terms related to the approximation error (1.0.8).

The *identifiability condition* relates the matrices $\mathcal{D}_0^2$ and $\mathcal{V}_0^2$.

($\mathcal{I}$) There is a constant $\mathfrak{a} > 0$ such that $\mathfrak{a}^2 \mathcal{D}_0^2 \ge \mathcal{V}_0^2$, i.e. such that $\mathfrak{a}^2 \mathcal{D}_0^2 - \mathcal{V}_0^2$ is positive definite.

Global conditions

The global conditions are needed to control the large deviations of the ME. They are chosen to ensure that the event $\{\|\mathcal{D}_0(\widetilde{\boldsymbol{\upsilon}} - \boldsymbol{\upsilon}^*)\| \le \mathtt{r}\}$ is of high probability for not to large $\mathtt{r} > 0$.

($\mathcal{E}\mathtt{r}$) *For any* $\mathtt{r}$, *there exist a constant* $\nu_{\mathtt{r}} > 0$ *and a value* $\mathtt{g}(\mathtt{r}) > 0$ *such that for all* $\lambda \le \mathtt{g}(\mathtt{r})$

$$\sup_{\boldsymbol{\upsilon} \in \varUpsilon_{\circ}(\mathtt{r})} \sup_{\gamma \in \mathbb{R}^p} \log I\!\!E \exp\biggl\{ \lambda \frac{\gamma^\top \nabla\zeta(\boldsymbol{\upsilon})}{\|\mathcal{V}_0 \gamma\|} \biggr\} \le \nu_{\mathtt{r}}^2 \lambda^2/2.$$

Also remember the radius $\mathtt{r}_0(\mathtt{x})$ from (3.1.6).

($\mathcal{L}\mathtt{r}$) *There is a function* $\mathtt{b}(\mathtt{r})$ *such that* $\mathtt{r}\mathtt{b}(\mathtt{r})$ *monotonously increases in* $\mathtt{r}$ *and for each* $\mathtt{r} \ge \mathtt{r}_0$

$$\inf_{\boldsymbol{\upsilon}:\, \|\mathcal{V}_0(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^*)\| = \mathtt{r}} \bigl| I\!\!E \mathcal{L}(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^*) \bigr| \ge \mathtt{b}(\mathtt{r}) \mathtt{r}^2.$$

3.2.2 Local quadratic bracketing

In this section we present the key results of [52], i.e. the local quadratic approximation of $\mathcal{L}(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^*)$ given by Theorem 3.2.1 below and its implications. We follow closely the original text of that paper.

Consider $\delta(\mathtt{r}), \omega(\mathtt{r})$ from $(\mathcal{E}\mathcal{D}_1')$ and $(\mathcal{L}_0')$ in Section 3.2.1. Introduce a vector $\boldsymbol{\epsilon}(\mathtt{r}) = (\delta(\mathtt{r}), \omega(\mathtt{r})) \in \mathbb{R}^2$ to define the quadratic process:

$$\begin{aligned}\mathbb{L}_{\boldsymbol{\epsilon}}(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^*) &\overset{\text{def}}{=} (\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^*)^\top \nabla \mathcal{L}(\boldsymbol{\upsilon}^*) - \|\mathcal{D}_{\boldsymbol{\epsilon}}(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^*)\|^2/2 \\ &= \boldsymbol{\xi}_{\boldsymbol{\epsilon}}^\top \mathcal{D}_{\boldsymbol{\epsilon}}(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^*) - \|\mathcal{D}_{\boldsymbol{\epsilon}}(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^*)\|^2/2, \qquad (3.2.1)\end{aligned}$$

where $\zeta(\boldsymbol{\upsilon}) = \mathcal{L}(\boldsymbol{\upsilon}) - I\!\!E\mathcal{L}(\boldsymbol{\upsilon})$ and $\nabla \mathcal{L}(\boldsymbol{\upsilon}^*) = \nabla \zeta(\boldsymbol{\upsilon}^*)$ by $\nabla I\!\!E \mathcal{L}(\boldsymbol{\upsilon}^*) = 0$. Further define with $\mathcal{V}_0^2 \geq \operatorname{Cov}(\nabla \zeta(\boldsymbol{\upsilon}^*))$

$$\mathcal{D}_{\boldsymbol{\epsilon}}^2 = \mathcal{D}_0^2(1 - \delta) - \omega \mathcal{V}_0^2, \qquad \boldsymbol{\xi}_{\boldsymbol{\epsilon}} \overset{\text{def}}{=} \mathcal{D}_{\boldsymbol{\epsilon}}^{-1} \nabla \mathcal{L}(\boldsymbol{\upsilon}^*).$$

$\mathbb{L}_{\underline{\boldsymbol{\epsilon}}}(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^*)$ is defined analogously via replacing $\boldsymbol{\epsilon} = (\delta, \omega)$ with $\underline{\boldsymbol{\epsilon}} = (-\delta, -\omega)$. [52] presents the following central sandwiching result:

Theorem 3.2.1 ([52], Theorem 3.1)**.** *Assume* $(\mathcal{E}\mathcal{D}_1')$ *and* $(\mathcal{L}_0')$. *Let for some* $\mathtt{r}$, *the values* $\omega \geq 3\nu_0\, \omega(\mathtt{r})$ *and* $\delta \geq \delta(\mathtt{r})$ *be such that* $\mathcal{D}_0^2(1 - \delta) - \omega V_0^2 \geq 0$. *Then for any* $\boldsymbol{\upsilon} \in \Upsilon_\circ(\mathtt{r})$ *with* $\mathtt{r} \leq \mathtt{r}*$

$$\mathbb{L}_{\underline{\boldsymbol{\epsilon}}}(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^*) - \diamondsuit_{\underline{\boldsymbol{\epsilon}}}(\mathtt{r}) \leq \mathcal{L}(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^*) \leq \mathbb{L}_{\boldsymbol{\epsilon}}(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^*) + \diamondsuit_{\boldsymbol{\epsilon}}(\mathtt{r}), \quad \boldsymbol{\upsilon} \in \Upsilon_\circ(\mathtt{r}), \quad (3.2.2)$$

with $\mathbb{L}_{\boldsymbol{\epsilon}}(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^*)$, $\mathbb{L}_{\underline{\boldsymbol{\epsilon}}}(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^*)$ *defined by* (3.2.1). *The error terms* $\diamondsuit_{\boldsymbol{\epsilon}}(\mathtt{r})$ *and* $\diamondsuit_{\underline{\boldsymbol{\epsilon}}}(\mathtt{r})$ *satisfy*

$$I\!\!P\bigl\{\omega^{-1}\diamondsuit_{\boldsymbol{\epsilon}}(\mathtt{r}) \geq \mathfrak{z}_Q(\mathtt{x}, \mathbb{Q})\bigr\} \leq \exp\bigl(-\mathtt{x}\bigr)$$

with $\mathfrak{z}_Q(\mathtt{x}, \mathbb{Q})$ *given for* $\mathtt{g}_0 = \mathtt{g}\nu_0 \geq 3$ *by*

$$\mathfrak{z}_Q(\mathtt{x}, \mathbb{Q}) \overset{\text{def}}{=} \begin{cases} \bigl(1 + \sqrt{\mathtt{x} + \mathbb{Q}}\bigr)^2 & \textit{if } 1 + \sqrt{\mathtt{x} + \mathbb{Q}} \leq \mathtt{g}_0, \\ 1 + \bigl\{2\mathtt{g}_0^{-1}(\mathtt{x} + \mathbb{Q}) + \mathtt{g}_0\bigr\}^2 & \textit{otherwise,} \end{cases}$$

where $\mathbb{Q} \leq \mathfrak{c}_1 p$ *with* $\mathfrak{c}_1 = 2$ *for* $p \geq 2$ *and* $\mathfrak{c}_1 = 2.7$ *for* $p = 1$. *Similarly for* $\diamondsuit_{\underline{\boldsymbol{\epsilon}}}(\mathtt{r})$.

Remark 3.2.2. The proof of this theorem is based on an uniform exponential deviation bound for $\mathcal{U}(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^*) - \|\mathcal{V}_0(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^*)\|^2/2$ with

$$\mathcal{U}(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^*) = \frac{1}{\omega(\mathtt{r})}\bigl\{\zeta(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^*) - (\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^*)^\top \nabla \zeta(\boldsymbol{\upsilon}^*)\bigr\}, \qquad \boldsymbol{\upsilon} \in \Upsilon_\circ(\mathtt{r}),$$

served by Theorem 3.5.6.

Proposition 3.4.1 allows to ensure that under $(\mathcal{E}\mathcal{D}_0)$, the norm $\|\boldsymbol{\xi}_0\|$ posses essentially the same deviation behavior as the norm of a Gaussian vector with the same covariance matrix. It gives for some nonrandom real value $\mathfrak{z}(\mathtt{x}, I\!\!B) > 0$ defined in Section 3.4 that

$$I\!\!P\bigl(\|\boldsymbol{\xi}\|^2 \geq \mathfrak{z}(\mathtt{x}, I\!\!B)\bigr) \leq 2\mathrm{e}^{-\mathtt{x}}. \tag{3.2.3}$$

Together with the *bracketing result* (3.2.2), the *geometric structure* of the processes $\mathbb{L}_{\boldsymbol{\epsilon}}$ and $\mathbb{L}_{\underline{\boldsymbol{\epsilon}}}$ and (3.2.3) this allow to derive a non-asymptotic versions of Fisher's and Wilks' theorems in Corollary 3.2.2. Define $\tau_{\boldsymbol{\epsilon}}(\mathtt{r}) \stackrel{\operatorname{def}}{=} \delta(\mathtt{r}) + \omega(\mathtt{r})\mathfrak{a}^2 < 1$, the value $\alpha_{\boldsymbol{\epsilon}}(\mathtt{r}) \stackrel{\operatorname{def}}{=} \frac{2\tau_{\boldsymbol{\epsilon}}}{1-\tau_{\boldsymbol{\epsilon}}^2}$ and the *spread* $\varDelta_{\boldsymbol{\epsilon}}(\mathtt{r})$ by

$$\varDelta_{\boldsymbol{\epsilon}}(\mathtt{r}) \stackrel{\operatorname{def}}{=} 2\omega(\mathtt{r})\mathfrak{z}_Q(\mathtt{x}, \mathbb{Q}) + \alpha_{\boldsymbol{\epsilon}}(\mathtt{r})\mathfrak{z}(\mathtt{x}, I\!\!B), \tag{3.2.4}$$

and note that it only depends on the radius $\mathtt{r} > 0$ through $\omega(\mathtt{r})$ and $\alpha_{\boldsymbol{\epsilon}}(\mathtt{r})$.

Further define the set $C_{\boldsymbol{\epsilon}}(\mathtt{r}) \subset \varOmega$ which is contained in the sigma algebra of the underlying probability space

$$\begin{aligned} C_{\boldsymbol{\epsilon}}(\mathtt{r}) \stackrel{\operatorname{def}}{=} \bigl\{&\tilde{\boldsymbol{\upsilon}} \in \varUpsilon_{\circ}(\mathtt{r}),\ \|V_0\mathcal{D}_{\underline{\boldsymbol{\epsilon}}}^{-1}\boldsymbol{\xi}_{\underline{\boldsymbol{\epsilon}}}\| \leq \mathtt{r}, \\ &\diamondsuit_{\boldsymbol{\epsilon}}(\mathtt{r}) \leq \omega\mathfrak{z}_Q(\mathtt{x}, \mathbb{Q}),\ \|\boldsymbol{\xi}_0\|^2 \leq \mathfrak{z}(\mathtt{x}, I\!\!B)\bigr\}. \end{aligned} \tag{3.2.5}$$

The implication of the results of [52] can be summarized as follows:

Corollary 3.2.2. *On the random set* $C_{\boldsymbol{\epsilon}}(\mathtt{r})$ *from* (3.2.5), *it holds*

$$\begin{aligned} \bigl|\mathcal{L}(\tilde{\boldsymbol{\upsilon}}, \boldsymbol{\upsilon}^*) - \|\boldsymbol{\xi}\|^2/2\bigr| &\leq \diamondsuit_{\boldsymbol{\epsilon}}(\mathtt{r}) \vee \diamondsuit_{\underline{\boldsymbol{\epsilon}}}(\mathtt{r}) + \alpha_{\boldsymbol{\epsilon}}(\mathtt{r})\|\boldsymbol{\xi}_{\boldsymbol{\epsilon}}\|, \\ \bigl\|\mathcal{D}_{\boldsymbol{\epsilon}}\bigl(\tilde{\boldsymbol{\upsilon}} - \boldsymbol{\upsilon}^*\bigr) - \boldsymbol{\xi}_{\boldsymbol{\epsilon}}\bigr\|^2 &\leq 2\varDelta_{\boldsymbol{\epsilon}}(\mathtt{r}). \end{aligned}$$

Remark 3.2.3. Define

$$\mathtt{r}_0(\mathtt{x}) \stackrel{\operatorname{def}}{=} \min\{\mathtt{r} > 0, I\!\!P(\tilde{\boldsymbol{\upsilon}} \in \varUpsilon_{\circ}(\mathtt{r})) \geq 1 - \mathrm{e}^{-\mathtt{x}}\}.$$

This gives that the set $C_{\boldsymbol{\epsilon}}(\mathtt{r}_0) \subset \varOmega$ is of probability greater than $1 - 4\mathrm{e}^{-\mathtt{x}}$ due to Theorem 3.2.1 combined with Equation (3.2.3). The bounds for the large deviation of the MLE derived in Section 3.3 give that $\mathtt{r}_0 \leq \mathtt{C}\sqrt{p^* + \mathtt{x}}$, which implies for $\delta(\mathtt{r}) + \omega(\mathtt{r}) \leq \mathtt{C}\mathtt{r}/\sqrt{n}$ that

$$\varDelta_{\boldsymbol{\epsilon}}(\mathtt{r}_0) \leq \mathtt{C}\frac{p^{*3/2}}{n^{1/2}}.$$

This is exactly the claimed sufficient ratio of dimension to sample size for the Wilks and Fisher expansions for the smooth i.i.d. setting if the approach [52] is used. As mentioned we will improve this sufficient ratio to $p^*/\sqrt{n}$ in Chapter 4.

Proof. From Corollary 3.3 of [52] we obtain

$$\|\boldsymbol{\xi}_{\underline{\boldsymbol{\epsilon}}}\|^2/2 - \diamondsuit_{\underline{\boldsymbol{\epsilon}}}(\mathtt{r}) \le \mathcal{L}(\widetilde{\boldsymbol{\upsilon}}, \boldsymbol{\upsilon}^*) \le \|\boldsymbol{\xi}_{\boldsymbol{\epsilon}}\|^2/2 + \diamondsuit_{\boldsymbol{\epsilon}}(\mathtt{r}).$$

We get

$$\begin{aligned}|\mathcal{L}(\widetilde{\boldsymbol{\upsilon}}, \boldsymbol{\upsilon}^*) - \|\boldsymbol{\xi}\|^2/2| \le{}& \diamondsuit_{\boldsymbol{\epsilon}}(\mathtt{r}) \vee \diamondsuit_{\underline{\boldsymbol{\epsilon}}}(\mathtt{r}) \\ &+ \left(\|\boldsymbol{\xi}_{\boldsymbol{\epsilon}}\|^2/2 - \|\boldsymbol{\xi}\|^2/2\right) \vee \left(\|\boldsymbol{\xi}_{\underline{\boldsymbol{\epsilon}}}\|^2/2 - \|\boldsymbol{\xi}\|^2/2\right).\end{aligned}$$

Now Lemma 3.2.3 gives the first claim. The second claim is Corollary 3.4 of [52]. □

Lemma 3.2.3 ([52], Lemma 3.9)**.** *Suppose* $(\mathcal{I})$ *and let* $\tau_{\boldsymbol{\epsilon}} \stackrel{\mathrm{def}}{=} \delta + \omega\mathfrak{a}^2 < 1$. *Then*

$$\mathcal{D}_{\boldsymbol{\epsilon}}^2 \ge (1-\tau_{\boldsymbol{\epsilon}})\mathcal{D}_0^2, \qquad \mathcal{D}_{\underline{\boldsymbol{\epsilon}}}^2 \le (1+\tau_{\boldsymbol{\epsilon}})\mathcal{D}_0^2,$$

$$\|I_p - \mathcal{D}_{\boldsymbol{\epsilon}}\mathcal{D}_{\underline{\boldsymbol{\epsilon}}}^{-2}\mathcal{D}_{\boldsymbol{\epsilon}}\|_\infty \le \alpha_{\boldsymbol{\epsilon}} \stackrel{\mathrm{def}}{=} \frac{2\tau_{\boldsymbol{\epsilon}}}{1-\tau_{\boldsymbol{\epsilon}}^2}.$$

Moreover,

$$\|\boldsymbol{\xi}_{\boldsymbol{\epsilon}}\|^2 - \|\boldsymbol{\xi}\|^2 \le \frac{\tau_{\boldsymbol{\epsilon}}}{1-\tau_{\boldsymbol{\epsilon}}}\|\boldsymbol{\xi}\|^2, \quad \|\boldsymbol{\xi}\|^2 - \|\boldsymbol{\xi}_{\underline{\boldsymbol{\epsilon}}}\|^2 \le \frac{\tau_{\boldsymbol{\epsilon}}}{1+\tau_{\boldsymbol{\epsilon}}}\|\boldsymbol{\xi}\|^2,$$

$$\|\boldsymbol{\xi}_{\boldsymbol{\epsilon}}\|^2 - \|\boldsymbol{\xi}_{\underline{\boldsymbol{\epsilon}}}\|^2 \le \alpha_{\boldsymbol{\epsilon}}\|\boldsymbol{\xi}\|^2.$$

3.3 Concentration of the qMLE

The result of Corollary 3.2.2 only holds true on the set $C(\mathtt{x}) \subset \Omega$. In Remark 3.2.3 we noted that this set is of very high probability. This particularly concerns the large deviation behavior of the estimator $\widetilde{\boldsymbol{\upsilon}} \in \mathbb{R}^{p^*}$. [52] presents one possible way of determining a radius $\mathtt{r}_0(\mathtt{x}) > 0$ such that

$$I\!P(\widetilde{\boldsymbol{\upsilon}} \in \varUpsilon_\circ(\mathtt{r}_0)) \ge \mathrm{e}^{-\mathtt{x}}. \tag{3.3.1}$$

In this section we present this approach. Before we explain it in more detail let us try to understand the idea. It involves the conditions $(\mathcal{E}\mathtt{r})$ and $(\mathcal{L}\mathtt{r})$ from Section 3.2.1. Let $\varUpsilon(\mathtt{r})$ be a family of nested sets, that shrink with decreasing radius $\mathtt{r}$, than due to the definition of $\widetilde{\boldsymbol{\upsilon}}$

$$I\!P\left(\widetilde{\boldsymbol{\upsilon}} \notin \varUpsilon(\mathtt{r})\right) \le I\!P\left(\max_{\boldsymbol{\upsilon}\in\varUpsilon(\mathtt{r})^c} \mathcal{L}(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^*) \ge 0\right),$$

because $\mathcal{L}(\widetilde{\boldsymbol{\upsilon}}) \ge \mathcal{L}(\boldsymbol{\upsilon}^*)$. Now one can decompose

$$\mathcal{L}(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^*) = \mathcal{L}(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^*) - I\!E[\mathcal{L}(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^*)] + I\!E[\mathcal{L}(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^*)],$$

such that with condition $(\mathcal{L}\mathtt{r})$ and Taylor expansion

$$\mathcal{L}(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^*) \le \nabla\boldsymbol{\zeta}(\widehat{\boldsymbol{\upsilon}})(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^*) - \mathtt{b}(\mathtt{r})\|\mathcal{D}(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^*)\|^2, \tag{3.3.2}$$

for some $\widehat{\boldsymbol{\upsilon}}(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^*)$ in the convex hull of $\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^* \in \varUpsilon$. Assume that the supremum of the norm of $\mathcal{D}^{-1}\nabla\boldsymbol{\zeta}(\boldsymbol{\upsilon})$ on $\varUpsilon(\mathtt{r})$ grows slower with the radius $\mathtt{r}$ than the quadratic term does. Then the term on the right hand side in (3.3.2) should be below zero with a high probability once the radius exceeds a certain bound $\mathtt{r}_0$. So the task is to find some $\mathtt{r}_0 > 0$ such that under condition $(\mathcal{E}\mathtt{r})$ for any $\mathtt{r} \ge \mathtt{r}_0$

$$I\!\!P\left(\max_{\boldsymbol{\upsilon}\in\varUpsilon(\mathtt{r})^c}\bigl\{\nabla\boldsymbol{\zeta}(\widehat{\boldsymbol{\upsilon}})(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^*) - \mathtt{b}(\mathtt{r})\|\mathcal{D}(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^*)\|^2\bigr\} \ge 0\right) \le \mathrm{e}^{-\mathtt{x}}.$$

This is done in Theorem 3.3.1, which again utilizes Theorem 3.5.6.

3.3.1 Upper function approach

Everything in this Section is taken from [52] except the remarks and slight variations in the formulation of the Theorems. The idea of the upper function device is to find a deterministic *upper function* $\mathfrak{u} : \varUpsilon \to \mathbb{R}$ such that

$$I\!\!P\left(\sup_{\boldsymbol{\upsilon}\in\varUpsilon}\bigl\{\mathcal{L}(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^*) + \mathfrak{u}(\boldsymbol{\upsilon}, \mathtt{x})\bigr\} < 0\right) \ge 1 - \mathrm{e}^{-\mathtt{x}}.$$

If this function $\mathfrak{u} : \varUpsilon \times \mathbb{R} \to \mathbb{R}$ satisfies $\mathfrak{u}(\cdot, \mathtt{x}) \ge 0$ on $\varUpsilon_\circ(\mathtt{r}_0(\mathtt{x}))^c \subset \varUpsilon$ for some $\mathtt{r}_0(\mathtt{x}) > 0$ then we can easily infer

$$I\!\!P(\widetilde{\boldsymbol{\upsilon}} \notin \varUpsilon_\circ(\mathtt{r}_0(\mathtt{x}))) \le I\!\!P(\mathcal{L}(\widetilde{\boldsymbol{\upsilon}}, \boldsymbol{\upsilon}^*) + \mathfrak{u}(\widetilde{\boldsymbol{\upsilon}}, \mathtt{x}) \ge 0) \le \mathrm{e}^{-\mathtt{x}},$$

because $\mathcal{L}(\widetilde{\boldsymbol{\upsilon}}, \boldsymbol{\upsilon}^*) \ge 0$ by definition.

Take a geometric sequence $\mu_k = \mu_0 2^{-k}$ with any fixed μ_0 and define $\mathfrak{t}(\mu_k) = k$ for $k \ge 0$. Define also for each $\mathtt{r} > 0$

$$\mathbb{M}(\mathtt{r}) \stackrel{\text{def}}{=} \bigl\{\mu_k : 1 + \sqrt{\mathtt{x} + \mathbb{Q} + \mathfrak{t}(\mu)} \le \nu_0 \mathtt{g}(\mathtt{r})/\mu_k\bigr\}, \tag{3.3.3}$$

with $\mathbb{Q} = \mathfrak{c}_1 p$. Theorem 2.8 of the supplement of [52] reads:

Theorem 3.3.1. *Suppose* $(\mathcal{E}\mathtt{r})$. *Let a function* $\mathfrak{u}(\boldsymbol{\upsilon})$ *be fixed. Define for any* $\boldsymbol{\upsilon}$

$$\begin{aligned}&\mathfrak{M}^*(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^*)\\ &\quad\stackrel{\text{def}}{=} \max_{\mu\in\mathbb{M}(\mathtt{r})}\Bigl\{-\frac{\mu}{3\nu_{\mathtt{r}}}\bigl[I\!\!E\mathcal{L}(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^*) + \mathfrak{u}(\boldsymbol{\upsilon})\bigr] - \frac{1}{2}\mu^2\mathtt{r}^2 - 2\mathfrak{t}(\mu)\Bigr\},\end{aligned} \tag{3.3.4}$$

where $\mathtt{r} = \|\mathcal{D}(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^*)\|$. *Take some* $\mathtt{x}$ *with* $\mathtt{x} + \mathbb{Q} \geq 2.5$. *If it holds*

$$\mathfrak{M}^*(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^*) \geq 2(\mathtt{x} + \mathbb{Q}), \qquad \boldsymbol{\upsilon} \in \varUpsilon_\circ(\mathtt{r})^c.$$

then

$$I\!\!P\left(\sup_{\boldsymbol{\upsilon} \in \varUpsilon_\circ(\mathtt{r})^c} \bigl\{\mathcal{L}(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^*) + \mathfrak{u}(\boldsymbol{\upsilon})\bigr\} \geq 0\right) \leq \mathrm{e}^{-\mathtt{x}}.$$

Remark 3.3.1. This result is proved using the following equations. The first one is this simple inequality

$$\begin{aligned} &\sup_{\boldsymbol{\upsilon} \in \varUpsilon_\circ(\mathtt{r})} \bigl\{\mathcal{L}(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^*) + \mathfrak{u}(\boldsymbol{\upsilon})\bigr\} \\ &\qquad \leq \sup_{\boldsymbol{\upsilon} \in \varUpsilon_\circ(\mathtt{r})} \{I\!\!E\mathcal{L}(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^*) + \mathfrak{u}(\boldsymbol{\upsilon})\} + \sup_{\boldsymbol{\upsilon} \in \varUpsilon_\circ(\mathtt{r})} \boldsymbol{\zeta}(\boldsymbol{\upsilon}) - \boldsymbol{\zeta}(\boldsymbol{\upsilon}^*). \end{aligned}$$

The second one is an application of Theorem 3.5.6 in Section 3.5, i.e. that with $(\mathcal{E}\mathtt{r})$ for $\mu \in \mathbb{M}(\mathtt{r})$

$$I\!\!P\left(\frac{\mu}{3\nu_0} \sup_{\boldsymbol{\upsilon} \in \varUpsilon_\circ(\mathtt{r})} \boldsymbol{\zeta}(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^*) - \frac{1}{2}\mu^2\mathtt{r}^2 \geq \{1 - \sqrt{\mathtt{x} + \mathfrak{t}(\mu) + \mathbb{Q}}\}^2\right) \leq \mathrm{e}^{-\mathtt{x}-\mathfrak{t}(\mu)}.$$

It remains to find a way to ensure that condition (3.3.4) is satisfied. This is done via a lower quadratic bound for the negative expectation $-I\!\!E\mathcal{L}(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^*) \geq \mathtt{b}(\mathtt{r})\|\mathcal{D}_0(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^*)\|^2/2$ given in condition $(\mathcal{L}\mathtt{r})$ from Section 3.2.1. Further the bounds of the exponential moments from condition $(\mathcal{L}\mathtt{r})$ have to be qualified to ensure that the set $\mathbb{M}(\mathtt{r})$ from (3.3.3) contains $\mathtt{b}/3\nu_0$. [52] presents two different results. The first one adresses the case that $\mathtt{b}(\mathtt{r}) \equiv \mathtt{b} > 0$ for all $\mathtt{r} \geq \mathtt{r}_0$. We present Theorem 4.2 in [52], in the following modified version, which is proved in the same way:

Theorem 3.3.2. *Suppose* $(\mathcal{E}\mathtt{r})$ *and* $(\mathcal{L}\mathtt{r})$ *with* $\mathtt{b}(\mathtt{r}) \equiv \mathtt{b}$. *Let, for* $\mathtt{r} \geq \mathtt{r}_0$,

$$1 + \sqrt{\mathtt{x} + \mathbb{Q}} \leq 3\nu_{\mathtt{r}}^2\mathtt{g}(\mathtt{r})/\mathtt{b}, \tag{3.3.5}$$

$$6\nu_{\mathtt{r}}\sqrt{\mathtt{x} + \mathbb{Q} + \frac{\mathtt{b}}{9\nu_{\mathtt{r}}}K} \leq \mathtt{r}\mathtt{b}, \tag{3.3.6}$$

with $\mathtt{x} + \mathbb{Q} \geq 2.5$ *and* $\mathbb{Q} = 2p^*$. *Then*

$$I\!\!P\bigl(\varUpsilon_{\mathcal{L}}(K) \not\subseteq \varUpsilon_\circ(\mathtt{r}_0)\bigr) \leq 2\mathrm{e}^{-\mathtt{x}},$$

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

Proof. The result follows from Theorem 3.3.1 with $\mathfrak{u}(\boldsymbol{\upsilon}) = K$, $\mathbb{M}(\mathtt{r}) = \{\frac{\mathtt{b}}{3\nu_{\mathtt{r}}}\}$, $\mathfrak{t}(\mu) \equiv 0$. □

Remark 3.3.2. Note that this Theorem also ensures that the maximum of $\mathcal{L}:\mathbb{R}^{p^*}\to\mathbb{R}$ is actually attained. Clearly $\boldsymbol{\upsilon}^*\in\varUpsilon(0)$ such that it is nonempty. Further

$$I\!\!P\left(\varUpsilon(0)\subseteq\varUpsilon_\circ(\mathtt{r}_0)\right)\geq 1-\mathrm{e}^{-\mathtt{x}},$$

such that $\varUpsilon(0)\subseteq\varUpsilon_\circ(\mathtt{r}_0)\subset\mathbb{R}^{p^*}$ is compact and thus $\mathcal{L}$ attains its maximum on $\varUpsilon(0)$, which will be the global maximum $\widetilde{\boldsymbol{\upsilon}}$. The same holds for $\widetilde{\boldsymbol{\upsilon}}_{\boldsymbol{\theta}^*}\in\mathbb{R}^{p^*}$, which is defined as

$$\widetilde{\boldsymbol{\upsilon}}_{\boldsymbol{\theta}^*}\stackrel{\mathrm{def}}{=}\operatorname*{argmax}_{\substack{\boldsymbol{\upsilon}\in\varUpsilon\\ \varPi_{\boldsymbol{\theta}}\boldsymbol{\upsilon}=\boldsymbol{\theta}^*}}\mathcal{L}(\boldsymbol{\upsilon}).$$

Remark 3.3.3. The condition (3.3.6) helps to understand which $\mathtt{r}_0>0$ ensures the prescribed concentration properties of $\widetilde{\boldsymbol{\upsilon}}\in\mathbb{R}^{p^*}$ and $\widetilde{\boldsymbol{\upsilon}}_{\boldsymbol{\theta}^*}\in\mathbb{R}^{p^*}$ because by definition both are in the set $\varUpsilon(0)$. Consequently, if $\mathtt{g}(\mathtt{r})>0$ is large enough, (3.3.6) follows from the bound

$$\mathtt{r}_0\geq 6\mathtt{b}^{-1}\nu_{\mathtt{r}}\sqrt{\mathtt{x}+p^*}. \tag{3.3.7}$$

Remark 3.3.4. The condition (3.3.5) qualifies the lowest admitted decay of $\mathtt{g}(\mathtt{r})$ from condition $(\mathcal{E}\mathtt{r})$, i.e. that $\mathtt{g}^2(\mathtt{r})\geq C(\mathtt{x}+p)$. This is similar to requiring finite polynomial moments for the score function. Condition (3.3.6) is derived from condition $(\mathcal{L}\mathtt{r})$. It tells us the necessary size of $\mathtt{r}_0(\mathtt{x})$ to ensure (3.3.1), namely $\mathtt{r}_0^2(\mathtt{x})\geq C(\mathtt{x}+p^*)$.

If $\mathtt{b}(\mathtt{r})$ decreases with $\mathtt{r}$, it has to be ensured that $\mathtt{b}(\mathtt{r})$ does not decrease too fast. More precisely we need that the product $\mathtt{r}\mathtt{b}(\mathtt{r})$ grows to infinity with $\mathtt{r}$. The result is given in Theorem 4.3 in [52]:

Theorem 3.3.3. *Suppose* $(\mathcal{E}\mathtt{r})$ *and* $(\mathcal{L}\mathtt{r})$. *Let* $\mathtt{r}_k$ *be such that* $\mathtt{b}(\mathtt{r}_k)\geq\mathtt{b}(\mathtt{r}_0)2^{-k}$ *for* $k\geq 1$. *If the conditions*

$$1+\sqrt{\mathtt{x}+\mathbb{Q}+ck}\leq 3\nu_{\mathtt{r}}^2\mathtt{g}(\mathtt{r}_k)/\mathtt{b}(\mathtt{r}_k),$$

$$6\nu_{\mathtt{r}}\sqrt{\mathtt{x}+\mathbb{Q}+ck+\frac{\mathtt{b}(\mathtt{r}_k)}{9\nu_{\mathtt{r}}}K}\leq\mathtt{r}_k\mathtt{b}(\mathtt{r}_k),$$

are fulfilled for $c=\log(2)$, *then it holds*

$$I\!\!P\bigl(\varUpsilon_{\mathcal{L}}(K)\not\subseteq\varUpsilon_\circ(\mathtt{r}_0)\bigr)\leq 2\mathrm{e}^{-\mathtt{x}},$$

where $\varUpsilon_{\mathcal{L}}(K)\stackrel{\mathrm{def}}{=}\{\boldsymbol{\upsilon}\in\varUpsilon:\,\mathcal{L}(\boldsymbol{\upsilon},\boldsymbol{\upsilon}^*)\geq -K\}$.

3.4 Deviation bounds for quadratic forms

In this section we present the important result by [52] on the deviation behavior of quadratic forms. It is rather technical and involves a list of constants. In the subsequent chapters we will use the corresponding bounds frequently. Everything in this Section is based on [52].

3.4.1 The idea behind the result

First let us try to get some intuition. The aim is to control the deviation for quadratic forms of type $\|M\boldsymbol{\xi}\|^2$ for a given symmetric matrix M and a random vector $\boldsymbol{\xi}$.

Remark 3.4.1. In this Chapter

$$\boldsymbol{\xi} = \mathcal{V}_0^{-1}\nabla\boldsymbol{\zeta}(\boldsymbol{\upsilon}^*), \quad M = \mathcal{D}_0^{-1}\mathcal{V}_0,$$

while in Chapters 4, 5 and 6 we could also have

$$\boldsymbol{\xi} = \breve{V}^{-1}\breve{\nabla}\boldsymbol{\zeta}(\boldsymbol{\upsilon}^*), \quad M = \breve{D}^{-1}\breve{V},$$

where the vector $\breve{\nabla}\boldsymbol{\zeta}(\boldsymbol{\upsilon}^*) \in \mathbb{R}^p$ and matrices $\breve{D}, \breve{V} \in \mathbb{R}^{p\times p}$ are introduced in Chapter 4.

The proof of the result of this section is based as usually in this chapter on the exponential Markov inequality

$$\mathbb{P}\left(\|M\boldsymbol{\xi}\|^2 \geq C(\mathtt{x})\right) \leq \mathrm{e}^{-\lambda C(\mathtt{x})/2}\mathbb{E}\left[\exp\{\lambda\|M\boldsymbol{\xi}\|^2/2\}\right].$$

It would be to restrictive to assume bounded exponential moments for the squared norm $\|M\boldsymbol{\xi}\|^2$. Using a small trick this is indeed not necessary:

$$\begin{aligned}
&\mathbb{E}\left[\exp\{\lambda\|M\boldsymbol{\xi}\|^2/2\}\right] \\
&\quad = \mathbb{E}\left[\exp\{\lambda\|M\boldsymbol{\xi}\|^2/2\}\left(\frac{\lambda}{\pi}\right)^{p^*/2}\int_{\mathbb{R}^{p^*}}\exp\{-\lambda\|M\boldsymbol{\xi}-\boldsymbol{\gamma}\|^2/2\}d\boldsymbol{\gamma}\right] \\
&\quad = \left(\frac{\lambda}{\pi}\right)^{p^*/2}\mathbb{E}\left[\int_{\mathbb{R}^{p^*}}\exp\{\lambda\boldsymbol{\gamma}^\top M\boldsymbol{\xi}-\lambda\|\boldsymbol{\gamma}\|^2/2\}d\boldsymbol{\gamma}\right]. \qquad (3.4.1)
\end{aligned}$$

To bound the right hand side of (3.4.1) one can use assumption (3.4.2) below to find

$$\begin{aligned}
\mathbb{E}\left[\exp\{\lambda\|M\boldsymbol{\xi}\|^2/2\}\right] &\leq \left(\frac{\lambda}{\pi}\right)^{p^*/2}\left[\int_{\mathbb{R}^{p^*}}\exp\{\lambda^2\|M\boldsymbol{\gamma}\|^2/2-\lambda\|\boldsymbol{\gamma}\|^2/2\}d\boldsymbol{\gamma}\right] \\
&= \left(\frac{\lambda}{\pi}\right)^{p^*/2}\left[\int_{\mathbb{R}^{p^*}}\exp\{-\lambda\boldsymbol{\gamma}^\top(I_{p^*}-\lambda M^2)\boldsymbol{\gamma}/2\}d\boldsymbol{\gamma}\right] \\
&= \det(I_{p^*}-\lambda M^2)^{-1/2},
\end{aligned}$$

which leads to the bounds for $\mathtt{C}(\mathtt{x})$ as in Proposition 3.4.1. Again we did not address the difficulty arising from the fact that in assumption (3.4.2) the moment bounds only exist up to a certain $\mathtt{g} > 0$. The actual result follows after a series of tedious calculations involving slicing arguments and the right truncation inside of the above integral.

3.4.2 Formulation of the result

Suppose that

$$\log \mathbb{E} \exp\bigl(\boldsymbol{\gamma}^{\top} \boldsymbol{\xi}\bigr) \le \|\boldsymbol{\gamma}\|^2/2, \qquad \boldsymbol{\gamma} \in \mathbb{R}^p,\, \|\boldsymbol{\gamma}\| \le \mathtt{g}. \tag{3.4.2}$$

For the symmetric matrix $\mathbb{B}^2 = MM^{\top}$ define

$$\mathtt{p}_{\mathbb{B}} = \operatorname{tr}(\mathbb{B}^2), \qquad \mathtt{v}_{\mathbb{B}}^2 = 2\operatorname{tr}(\mathbb{B}^4), \qquad \lambda_{\mathbb{B}}^* \overset{\text{def}}{=} \|\mathbb{B}^2\|_{\infty} \overset{\text{def}}{=} \lambda_{\max}(\mathbb{B}^2).$$

To ease notation suppose that $\mathtt{g}^2 \ge 2\mathtt{p}$. The other case only changes the constants in the definition of $\mathfrak{z}(\mathtt{x}, \mathbb{B}) > 0$ below. Define $\mu_c = 2/3$ and

$$\begin{aligned} \mathtt{g}_c &\overset{\text{def}}{=} \sqrt{\mathtt{g}^2 - \mu_c \mathtt{p}_{\mathbb{B}}}, \\ 2(\mathtt{x}_c + 2) &\overset{\text{def}}{=} (\mathtt{g}^2/\mu_c - \mathtt{p}_{\mathbb{B}})/\lambda_{\mathbb{B}}^* + \log\det\bigl(I_p - \mu_c \mathbb{B}/\lambda^*\bigr). \end{aligned}$$

The following proposition is a variant of Corollary 1.7 of the supplement of [52]:

Proposition 3.4.1. *Let* (3.4.2) *and* $\mathtt{g}^2 \ge 2\mathtt{p}_{\mathbb{B}}$. *Then for each* $\mathtt{x} > 0$

$$\mathbb{P}\bigl(\|M\boldsymbol{\xi}\| \ge \mathfrak{z}(\mathtt{x}, \mathbb{B})\bigr) \le 2\mathrm{e}^{-\mathtt{x}},$$

where $\mathfrak{z}(\mathtt{x}, \mathbb{B})$ *is defined by*

$$\mathfrak{z}^2(\mathtt{x} - 1, \mathbb{B}) \overset{\text{def}}{=} \begin{cases} \mathtt{p}_{\mathbb{B}} + 2\mathtt{v}_{\mathbb{B}}(\mathtt{x})^{1/2}, & \mathtt{x} \le \frac{\mathtt{v}_{\mathbb{B}}}{18\lambda_{\mathbb{B}}^*}, \\ \mathtt{p}_{\mathbb{B}} + 6\lambda^*(\mathtt{x}), & \frac{\mathtt{v}_{\mathbb{B}}}{18\lambda_{\mathbb{B}}^*} < \mathtt{x} \le \mathtt{x}_c, \\ \bigl|\mathtt{y}_c + 2\frac{\lambda_{\mathbb{B}}^*}{\mathtt{g}_c}(\mathtt{x} - \mathtt{x}_c)\bigr|^2, & \mathtt{x} > \mathtt{x}_c, \end{cases} \tag{3.4.3}$$

with $\mathtt{y}_c^2 \le \mathtt{p}_{\mathbb{B}} + 6\lambda_{\mathbb{B}}^*(\mathtt{x}_c + 2)$.

Remark 3.4.2. The definitions above are presented for the sake of completeness. They arise from the proof of the proposition. One important thing to note is that $\mathtt{x}_c \cong \mathtt{g}_n$, where in many cases $\mathtt{g}_n \to \infty$ as $n \to \infty$. This means that $\mathtt{x}_c \to \infty$ such that in most cases one can bound $\mathfrak{z}^2(\mathbb{B}, \mathtt{x}) \le \mathtt{p}_{\mathbb{B}} + 6\lambda^*(\mathtt{x} + 1)$. In not to degenerate cases one can expect that $\operatorname{tr}(MM^{\top}) \le \mathtt{C}p^*$ for some constant $\mathtt{C} > 0$, such that we obtain

$$\mathbb{P}\left(\|M\boldsymbol{\xi}\|^2 \ge \mathtt{C}(p^* + \mathtt{x})\right) \le 2\mathrm{e}^{-\mathtt{x}}.$$

3.5 Some results for empirical processes

In this section we summarize the results of the supplement of [52] concerning empirical processes. We will concentrate rather on the explanation of the ideas and correct citation of the necessary results than on the technical details. This is why we present a simple proof of the central Theorem 3.5.5 below. Also we only address the finite dimensional case. In Sections 3.5.2 and 3.5.3 we also present two new results for bounds of suprema of norms of random vector- or matrix-valued processes from [4] and [5].

We are interested in bounds for

$$\sup_{\boldsymbol{\upsilon}\in\varUpsilon_\circ(\mathtt{r})}\zeta(\boldsymbol{\upsilon},\boldsymbol{\upsilon}^*),\quad \sup_{\boldsymbol{\upsilon}\in\varUpsilon_\circ(\mathtt{r})}\omega(\mathtt{r})^{-1}\left\{\zeta(\boldsymbol{\upsilon},\boldsymbol{\upsilon}^*)-\nabla\zeta(\boldsymbol{\upsilon}^*)(\boldsymbol{\upsilon}-\boldsymbol{\upsilon}^*)\right\}. \tag{3.5.1}$$

The term $\mathcal{U}(\boldsymbol{\upsilon})$ can stand for $\mathcal{U}(\boldsymbol{\upsilon})=\zeta(\boldsymbol{\upsilon})$ or $\mathcal{U}(\boldsymbol{\upsilon})=\omega^{-1}(\zeta(\boldsymbol{\upsilon})-\nabla\zeta(\boldsymbol{\upsilon}^*)\boldsymbol{\upsilon})$ depending on context. The approach we will present here consists of two steps. First one derives a bound for the exponential moments of $\sup_{\boldsymbol{\upsilon}\in\varUpsilon_\circ(\mathtt{r})}|\mathcal{U}(\boldsymbol{\upsilon})-\mathcal{U}(\boldsymbol{\upsilon}^*)|$ using chaining and the conditions $(\mathcal{E}\mathtt{r})$ or $(\mathcal{E}\mathcal{D}_1)$:

$$\log I\!\!E\exp\Big\{\frac{\lambda}{2\nu_0\mathtt{r}}\sup_{\boldsymbol{\upsilon}\in\varUpsilon_\circ(\mathtt{r})}\big|\mathcal{U}(\boldsymbol{\upsilon})-\mathcal{U}(\boldsymbol{\upsilon}^*)\big|\Big\}\le\lambda^2/2+\mathbb{Q}(\varUpsilon_\circ(\mathtt{r})),$$

where $\mathbb{Q}(\varUpsilon_\circ(\mathtt{r}))$ is the entropy of the set $\varUpsilon_\circ(\mathtt{r})$ from (3.1.5), which is a measure of the complexity and is related to the Dudley integral (see [17]). The second step is the exponential Markov inequality

$$\begin{aligned} I\!\!P&\left(\sup_{\boldsymbol{\upsilon}\in\varUpsilon_\circ(\mathtt{r})}\big|\mathcal{U}(\boldsymbol{\upsilon})-\mathcal{U}(\boldsymbol{\upsilon}^*)\big|\ge 2\nu_0\mathtt{r}\mathfrak{z}(\mathtt{x})\right) \\ &\le\exp\left\{\lambda^2/2+\mathbb{Q}(\varUpsilon_\circ(\mathtt{r}))-\lambda\mathfrak{z}(\mathtt{x})\right\}.\end{aligned}\tag{3.5.2}$$

It remains to minimize the exponent on the right hand side with respect to $\lambda>0$ and to find a constant $\mathfrak{z}(\mathtt{x})$ such that

$$\min_{\lambda}\{\lambda^2/2+\mathbb{Q}(\varUpsilon_\circ(\mathtt{r}))-\lambda\mathfrak{z}(\mathtt{x})\}=\mathtt{x},$$

which gives $\mathfrak{z}(\mathtt{x})=\sqrt{2(\mathbb{Q}(\varUpsilon_\circ(\mathtt{r}))+\mathtt{x})}$ if there is no constraint on the size of $|\lambda|$. In some cases the dependence on $\mathtt{r}$ as in (3.5.2) is not desirable as $\mathtt{r}$ might be too big. Also in Chapter 5 we are interested in a bound for

$$\sup_{\mathtt{r}\le\mathtt{r}^*}\sup_{\boldsymbol{\upsilon}\in\varUpsilon_\circ(\mathtt{r})}\big|\mathcal{U}(\boldsymbol{\upsilon})-\mathcal{U}(\boldsymbol{\upsilon}^*)\big|.$$

To address this problem the idea is to subtract a quadratic drift term that dominates the "linear in $\mathtt{r}$" deviations of $\mathcal{U}(\boldsymbol{\upsilon})-\mathcal{U}(\boldsymbol{\upsilon}^*)$ for large $\mathtt{r}>0$, which means that one bounds

$$\sup_{\mathtt{r}\leq\mathtt{r}*}\sup_{\boldsymbol{\upsilon}\in\varUpsilon_{\circ}(\mathtt{r})}\left\{\left|\mathcal{U}(\boldsymbol{\upsilon})-\mathcal{U}(\boldsymbol{\upsilon}^{*})\right|-d(\boldsymbol{\upsilon},\boldsymbol{\upsilon}^{*})^{2}\right\},$$

for some adequate distance $d(\cdot,\cdot)$. This is done in Theorem 3.5.6 using similar arguments, which yields a bound of the kind

$$I\!\!P\left(\sup_{\mathtt{r}\leq\mathtt{r}*}\sup_{\boldsymbol{\upsilon}\in\varUpsilon_{\circ}(\mathtt{r})}\left\{\frac{1}{3\nu_0}\left|\mathcal{U}(\boldsymbol{\upsilon})-\mathcal{U}(\boldsymbol{\upsilon}^{*})\right|-d(\boldsymbol{\upsilon},\boldsymbol{\upsilon}^{*})^{2}\right\}\geq\mathfrak{z}(\mathtt{x})^{2}\right)\leq\mathrm{e}^{-\mathtt{x}},$$

where $\mathfrak{z}(\mathtt{x})=1+\sqrt{2(\mathbb{Q}(\varUpsilon_{\circ}(\mathtt{r}))+\mathtt{x})}$. The analysis becomes more involved once the exponential moments only exist for $|\lambda|\leq\mathtt{g}_0$ for some $\mathtt{g}_0>0$. This is why in the following the term $\mathfrak{z}(\mathtt{x})$ becomes more complicated.

3.5.1 A bound for local fluctuations

Everything in this Section except the proof of Lemma 3.5.4 is taken from [52]. We infer from $(\mathcal{E}\mathtt{r})$ and $(\mathcal{E}\mathcal{D}_1)$ from section 3.2.1 or 4.2.1.

$(\mathcal{E}D)$ *There exist* $\mathtt{g}>0$, $\nu_0\geq 1$ *for each* $\boldsymbol{\upsilon}\in\varUpsilon_{\circ}(\mathtt{r})$ *such that for any* $\lambda\leq\mathtt{g}(\mathtt{r})$ *and any unit vector* $\boldsymbol{\gamma}\in\mathbb{R}^{p^*}$, *it holds*

$$\log I\!\!E\exp\left\{\lambda\frac{\gamma^{\top}\nabla\mathcal{U}(\boldsymbol{\upsilon})}{\|\mathcal{D}_0\gamma\|}\right\}\leq\nu_0^2\lambda^2/2.$$

The following lemma turns out to be very useful:

Lemma 3.5.1 ([52], Lemma 2.9). *Assume that* $(\mathcal{E}D)$ *holds with some* $\mathtt{g}$ *for each* $\boldsymbol{\upsilon}\in\varUpsilon_{\circ}(\mathtt{r})$. *Consider any* $\boldsymbol{\upsilon},\boldsymbol{\upsilon}^{\circ}\in\varUpsilon_{\circ}(\mathtt{r}_0)$. *Then it holds for* $|\lambda|\leq\mathtt{g}$

$$\log I\!\!E\exp\left\{\lambda\frac{\mathcal{U}(\boldsymbol{\upsilon},\boldsymbol{\upsilon}^{\circ})}{\|\mathcal{D}_0(\boldsymbol{\upsilon}-\boldsymbol{\upsilon}^{\circ})\|}\right\}\leq\frac{\nu_0^2\lambda^2}{2}. \tag{3.5.3}$$

Thus Lemma 3.5.1 shows that condition $(\mathcal{E}D)$ implies $(\mathcal{E}d)$ with $d(\boldsymbol{\upsilon},\boldsymbol{\upsilon}^{\circ})=\|\mathcal{D}_0(\boldsymbol{\upsilon}-\boldsymbol{\upsilon}^{\circ})\|$:

$(\mathcal{E}d)$ *There exist* $\mathtt{g}>0$, $\mathtt{r}_0>0$, $\nu_0\geq 1$, *such that for any* $\lambda\leq\mathtt{g}$ *and* $\boldsymbol{\upsilon},\boldsymbol{\upsilon}^{\circ}\in\varUpsilon$ *with* $d(\boldsymbol{\upsilon},\boldsymbol{\upsilon}^{\circ})\leq\mathtt{r}_0$

$$\log I\!\!E\exp\left\{\lambda\frac{\mathcal{U}(\boldsymbol{\upsilon})-\mathcal{U}(\boldsymbol{\upsilon}^{\circ})}{d(\boldsymbol{\upsilon},\boldsymbol{\upsilon}^{\circ})}\right\}\leq\nu_0^2\lambda^2/2. \tag{3.5.4}$$

Remark 3.5.1. In the setting of Theorems 4.2.2 and 5.2.1 we have

$$\mathcal{U}(\boldsymbol{\upsilon})=\mathcal{D}^{-1}\Big(\nabla\boldsymbol{\zeta}(\boldsymbol{\upsilon})-\nabla\boldsymbol{\zeta}(\boldsymbol{\upsilon}^{*})\Big),$$

and condition $(\mathcal{E}d)$ becomes $(\mathcal{E}\mathcal{D}_1)$ from 4.2.1.

To derive bounds for the terms in (3.5.1) we only have to apply Theorem 3.5.5 or 3.5.6 from below. For this we use the basic chaining device as it was introduced by [17]. Let $(\varUpsilon_k)$ be a sequence of subsets $\varUpsilon_k \subset \varUpsilon_\circ(\mathtt{r})$ with minimal cardinality while satisfying $\sup_{\boldsymbol{\upsilon}\in\varUpsilon_\circ(\mathtt{r})}\inf_{\boldsymbol{\upsilon}'\in\varUpsilon_k} d(\boldsymbol{\upsilon}',\boldsymbol{\upsilon}) \le \mathtt{r}2^{-k}$ and $\varUpsilon_0 = \{\boldsymbol{\upsilon}^*\}$. This allows to define the entropy of $\varUpsilon_\circ(\mathtt{r})$

$$\mathbb{Q}(\varUpsilon_\circ(\mathtt{r})) := \sum_{k=1}^{\infty} 2^{-k}\log(2|\varUpsilon_k|), \tag{3.5.5}$$

and we remark, that for $\varUpsilon_\circ(\mathtt{r}) \subset \mathbb{R}^{p^*}$ we have $\mathbb{Q}(\varUpsilon_\circ(\mathtt{r})) = 2p^*$ due to the following Lemma:

Lemma 3.5.2 ([52], Lemma 2.10)**.** *Let $\varUpsilon^\circ = \{\boldsymbol{\upsilon} \in \varUpsilon : d(\boldsymbol{\upsilon},\boldsymbol{\upsilon}^*) \le \mathtt{r}\}$ for some $\boldsymbol{\upsilon} \in \mathbb{R}^{p^*}$. Under the conditions of Lemma 3.5.1, it holds $\mathbb{Q}(\varUpsilon^\circ) \le \mathfrak{c}_1 p^*$, where $\mathfrak{c}_1 = 2$ for $p^* \ge 2$, and $\mathfrak{c}_1 = 2.7$ for $p^* = 1$.*

For the derivation of Theorem 3.5.5 and Theorem 3.5.6 we need a series of Lemmas. First we need the Hölder inequality.

Lemma 3.5.3. *For any r.v.'s ξ_k and $\lambda_k \ge 0$ such that $\varLambda = \sum_k \lambda_k \le 1$*

$$\log I\!\!E \exp\Bigl(\sum_k \lambda_k \xi_k\Bigr) \le \sum_k \lambda_k \log I\!\!E \mathrm{e}^{\xi_k}.$$

Now we state the central lemma of this section where we use the sequence of sets $(\varUpsilon_k)$ to apply the chaining method.

Lemma 3.5.4 ([52], Theorem 2.1)**.** *Suppose $(\mathcal{E}d)$. If $\varUpsilon_\circ(\mathtt{r})$ is a set with finite entropy and center $\boldsymbol{\upsilon}^*$ and the radius $\mathtt{r}$, i.e. $d(\boldsymbol{\upsilon},\boldsymbol{\upsilon}^*) \le \mathtt{r}$ for all $\boldsymbol{\upsilon} \in \varUpsilon_\circ(\mathtt{r})$, then for $\lambda \le \mathtt{g}_0 \overset{\mathrm{def}}{=} \nu_0\mathtt{g}$*

$$\log I\!\!E \exp\Bigl\{\frac{\lambda}{2\nu_0\mathtt{r}} \sup_{\boldsymbol{\upsilon}\in\varUpsilon_\circ(\mathtt{r})} \bigl|\mathcal{U}(\boldsymbol{\upsilon}) - \mathcal{U}(\boldsymbol{\upsilon}^*)\bigr|\Bigr\} \le \lambda^2/2 + \mathbb{Q}(\varUpsilon_\circ(\mathtt{r})).$$

Remark 3.5.2. This Lemma is Theorem 2.1 from [52]. We give a simple and short proof because this result is fundamental for this work. The original proof is based on generic chaining and is slightly more complicated. Because we will only use results from this section for finite dimensional sets $\varUpsilon \subset \mathbb{R}^{p^*}$ usual chaining is sufficient. See Section 2 of [55] for a concise description of chaining and generalizations of the idea.

Proof. A simple change $\mathcal{U}(\cdot)$ with $\nu_0^{-1}\mathcal{U}(\cdot)$ and $\mathtt{g}$ with $\mathtt{g}_0 = \nu_0\mathtt{g}$ allows to reduce the result to the case with $\nu_0 = 1$ which we assume below. We have with our sequence $(\varUpsilon_k)$

$$\sup_{\boldsymbol{\upsilon}\in\varUpsilon_\circ(\mathtt{r})} \bigl|\mathcal{U}(\boldsymbol{\upsilon}) - \mathcal{U}(\boldsymbol{\upsilon}^\circ)\bigr| \le \sum_{k=1}^{\infty} \sup_{\boldsymbol{\upsilon}\in\varUpsilon_k} \inf_{\boldsymbol{\upsilon}'\in\varUpsilon_{k-1}} |\mathcal{U}(\boldsymbol{\upsilon}) - \mathcal{U}(\boldsymbol{\upsilon}')|.$$

We denote $\sup_{\boldsymbol{v}\in\Upsilon_k}\inf_{\boldsymbol{v}'\in\Upsilon_{k-1}}|\mathcal{U}(\boldsymbol{v})-\mathcal{U}(\boldsymbol{v}')| =: \xi_k^*$. Denote $c_k = 2^{-k}$ for $k \ge 1$. Then $\sum_{k=1}^{\infty} c_k = 1$. It holds by the Hölder inequality; see Lemma 3.5.3:

$$\log \mathbb{E}\exp\biggl(\frac{\lambda}{2\mathtt{r}}\sum_{k=1}^{\infty}\xi_k^*\biggr) \le \sum_{k=1}^{\infty} c_k \log \mathbb{E}\exp\biggl(\frac{\lambda}{2^{-k+1}\mathtt{r}}\xi_k^*\biggr)$$

We have with $\boldsymbol{v}_{k-1}'(\boldsymbol{v}) = \operatorname{argmin}_{\boldsymbol{v}'\in\Upsilon_{k-1}} d(\boldsymbol{v},\boldsymbol{v}')$ and with $(\mathcal{E}d)$

$$\begin{aligned}
&\log \mathbb{E}\exp\biggl(\frac{\lambda}{2^{-k+1}\mathtt{r}_0}\xi_k^*\biggr)\\
&\quad\le \log\sum_{\boldsymbol{v}\in\Upsilon_k}\mathbb{E}\exp\biggl(\lambda\frac{|\mathcal{U}(\boldsymbol{v})-\mathcal{U}(\boldsymbol{v}_{k-1}'(\boldsymbol{v}))|}{d(\boldsymbol{v}_{k-1}'(\boldsymbol{v}),\boldsymbol{v})}\biggr)\\
&\quad\le \log\biggl\{\sum_{\boldsymbol{v}\in\Upsilon_k}\mathbb{E}\exp\biggl(\lambda\frac{\mathcal{U}(\boldsymbol{v})-\mathcal{U}(\boldsymbol{v}_{k-1}'(\boldsymbol{v}))}{d(\boldsymbol{v}_{k-1}'(\boldsymbol{v}),\boldsymbol{v})}\biggr)\\
&\qquad\qquad+\mathbb{E}\exp\biggl(-\lambda\frac{\mathcal{U}(\boldsymbol{v})-\mathcal{U}(\boldsymbol{v}_{k-1}'(\boldsymbol{v}))}{d(\boldsymbol{v}_{k-1}'(\boldsymbol{v}),\boldsymbol{v})}\biggr)\biggr\}\\
&\quad\le \log(2|\Upsilon_k|)+\lambda^2/2.
\end{aligned}$$

which gives the claim. □

The exponential bound of Lemma 3.5.3 can be used for obtaining a probability bound on the maximum of the increments $\mathcal{U}(\boldsymbol{v})-\mathcal{U}(\boldsymbol{v}^\circ)$ over $\Upsilon_\circ(\mathtt{r})$. We restate Corollary 2.2 of the supplement of [52]:

Theorem 3.5.5. *Suppose* $(\mathcal{E}d)$. *If* $\Upsilon_\circ(\mathtt{r})$ *is a central set with center* $\boldsymbol{v}^*$ *and radius* $\mathtt{r} > 0$, *then it holds for any* $\mathtt{x} > 0$

$$\mathbb{P}\biggl(\sup_{\boldsymbol{v}\in\Upsilon_\circ(\mathtt{r})}\mathcal{U}(\boldsymbol{v})-\mathcal{U}(\boldsymbol{v}^*) > 3\nu_0\mathtt{r}\mathfrak{z}_1(\mathtt{x},\mathbb{Q})\biggr) \le \exp\bigl(-\mathtt{x}\bigr),$$

where with $\mathtt{g}_0 = \nu_0\mathtt{g}$ *and* $\mathbb{Q} = \mathbb{Q}(\Upsilon_\circ(\mathtt{r}))$

$$\mathfrak{z}_1(\mathtt{x},\mathbb{Q}) \overset{\mathrm{def}}{=} \begin{cases}\sqrt{2(\mathtt{x}+\mathbb{Q})}, & \textit{if } \sqrt{2(\mathtt{x}+\mathbb{Q})}\le\mathtt{g}_0,\\ \mathtt{g}_0^{-1}(\mathtt{x}+\mathbb{Q})+\mathtt{g}_0/2, & \textit{otherwise.}\end{cases} \tag{3.5.7}$$

Remark 3.5.3. The proof is a simple application of the exponential Markov inequality.

The previous Lemma yields a bound that depends linearly on the radius $\mathtt{r} > 0$ of the local set over which the supremum is taken. Subtracting a quadratic process as done in the proof of Theorem 3.2.1 allows to obtain a bound that is independent of $\mathtt{r} > 0$ and is preferable in situations where the radius $\mathtt{r}_0 > 0$ - needed to ensure $I\!\!P(\widetilde{\boldsymbol{\upsilon}} \in \varUpsilon_\circ(\mathtt{r})) \geq 1 - \mathrm{e}^{-\mathtt{x}}$ - is too large in comparison to $\sqrt{\mathfrak{z}_Q(\mathtt{x}, \mathbb{Q})} > 0$ from (3.5.8). For this purpose we restate Corollary 2.5 of the supplement of [52] as a Theorem:

Theorem 3.5.6. *Let* $(\varUpsilon_\circ(\mathtt{r}))_{0 \leq \mathtt{r} \leq \mathtt{r}^*} \subset \mathbb{R}^{p^*}$ *be a sequence of balls around* $\boldsymbol{\upsilon}^*$ *induces by the metric* $d(\cdot,\cdot)$. *Let a random* p *-vector process* $\mathcal{U}(\mathtt{r}, \boldsymbol{\upsilon})$ *fulfill* $\mathcal{U}(\mathtt{r}, \boldsymbol{\upsilon}^*) = 0$ *and* $(\mathcal{E}d)$ *for each* $0 \leq \mathtt{r} \leq \mathtt{r}^*$. *Finally assume that* $\sup_{\boldsymbol{\upsilon} \in \varUpsilon_\circ(\mathtt{r})} \mathcal{U}(\mathtt{r}, \boldsymbol{\upsilon})$ *increases in* $\mathtt{r}$. *Then for each* $0 \leq \mathtt{r} \leq \mathtt{r}^*$, *on a set of probability greater than* $1 - \mathrm{e}^{-\mathtt{x}}$

$$\sup_{\boldsymbol{\upsilon} \in \varUpsilon_\circ(\mathtt{r})} \left\{ \frac{1}{3\nu_1} \mathcal{U}(\mathtt{r}, \boldsymbol{\upsilon}) - d(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^*)^2 \right\} \leq \mathfrak{z}_Q(\mathtt{x}, 2p^*),$$

where with $\mathtt{g}_0 = \nu_0 \mathtt{g}$

$$\mathfrak{z}_Q(\mathtt{x}, \mathbb{Q}) \overset{\mathrm{def}}{=} \begin{cases} (1 + \sqrt{\mathtt{x} + \mathbb{Q}})^2, & \textit{if } 1 + \sqrt{\mathtt{x} + \mathbb{Q}} \leq \mathtt{g}_0, \\ 1 + \{2\mathtt{g}_0^{-1}(\mathtt{x} + \mathbb{Q}) + \mathtt{g}_0\}^2, & \textit{otherwise.} \end{cases} \tag{3.5.8}$$

Remark 3.5.4. The proof of this result is based on Lemma 3.5.4 and uses a peeling argument to derive an even better bound for the exponential moments of $\frac{1}{3\nu_0}\mathcal{U}(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^*) - \frac{\varrho}{2} d^2(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^*)$ that allows to get rid of the dependence on $\mathtt{r} > 0$. Finally Lemma 3.5.2 allows to replace $\mathbb{Q}$ by $2p^*$.

Remark 3.5.5. The generalization that the process $\mathcal{U}(\mathtt{r}, \boldsymbol{\upsilon})$ is allowed to depend on the radius $\mathtt{r}$ is possible because we impose that $\sup_{\boldsymbol{\upsilon} \in \varUpsilon_\circ(\mathtt{r})} \mathcal{U}(\mathtt{r}, \boldsymbol{\upsilon})$ increases in $\mathtt{r}$, such that the assertion remains valid.

3.5.2 A bound for the norm of a random process

This and the following section are based on [4] and [5]. In Chapter 5 we need for a random process $\mathcal{Y}(\boldsymbol{\upsilon}) \in \mathbb{R}^{p^*}$ a bound of the kind

$$I\!\!P\left(\sup_{\mathtt{r} \leq \mathtt{r}^*} \sup_{\boldsymbol{\upsilon} \in \varUpsilon_\circ(\mathtt{r})} \left\{ \frac{1}{\omega} \|\mathcal{Y}(\boldsymbol{\upsilon})\| - 2\mathtt{r}^2 \right\} \geq \mathtt{C}\mathfrak{z}_Q(\mathtt{x}, p^*) \right) \leq \mathrm{e}^{-\mathtt{x}}.$$

We want to derive it using the results of the previous section.

Let $\mathcal{Y}(\boldsymbol{\upsilon})$ be a smooth centered random vector process with values in $\mathbb{R}^{p^*}$ and let $\mathcal{D} : \mathbb{R}^{p^*} \to \mathbb{R}^{p^*}$ be some linear operator. We aim at bounding the maximum of the norm $\|\mathcal{Y}(\boldsymbol{\upsilon})\|$ over a vicinity $\varUpsilon_\circ(\mathtt{r}) \overset{\mathrm{def}}{=} \{\|\mathcal{D}(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^*)\| \leq \mathtt{r}\}$ of $\boldsymbol{\upsilon}^*$. Suppose that $\mathcal{Y}(\boldsymbol{\upsilon})$ satisfies $(\mathcal{E}d)$ with norm $d(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^\circ) = \omega \|\mathcal{D}_0(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^\circ)\|$.

Theorem 3.5.7. *Let a random* p^* *-vector process* $\mathcal{Y}(\boldsymbol{\upsilon})$ *fulfill* $\mathcal{Y}(\boldsymbol{\upsilon}^*) = 0$, $I\!\!E\mathcal{Y}(\boldsymbol{\upsilon}) \equiv 0$, *and suppose that* $\mathcal{Y}(\boldsymbol{\upsilon})$ *satisfies* $(\mathcal{E}d)$ *from* (3.5.4) *with norm*

$d(\boldsymbol{v},\boldsymbol{v}^{\circ}) = \omega\|\mathcal{D}_0(\boldsymbol{v}-\boldsymbol{v}^{\circ})\|$. *Then for each* $0 \le \mathtt{r} \le \mathtt{r}^*$, *on a set of probability greater* $1-\mathrm{e}^{-\mathtt{x}}$

$$\sup_{\mathtt{r}\le\mathtt{r}^*}\sup_{\boldsymbol{v}\in\varUpsilon_{\circ}(\mathtt{r})}\left\{\frac{1}{6\omega\nu_1}\|\mathcal{Y}(\boldsymbol{v})\| - 2\mathtt{r}^2\right\} \le \mathfrak{z}_Q(\mathtt{x},4p^*),$$

Remark 3.5.6. Note that only twice the entropy of the original set $\varUpsilon_{\circ}(\mathtt{r}) \subset \mathbb{R}^{p^*}$ enters the bound. Thus in order to control the norm $\|\mathcal{Y}(\boldsymbol{v})\|$ one only pays with this factor.

Proof. In what follows, we use the representation

$$\|\mathcal{Y}(\boldsymbol{v})\| = \omega \sup_{\|\boldsymbol{u}\|\le\|\mathcal{D}(\boldsymbol{v}-\boldsymbol{v}^*)\|}\frac{1}{\omega\|\mathcal{D}(\boldsymbol{v}-\boldsymbol{v}^*)\|}\boldsymbol{u}^{\top}\mathcal{Y}(\boldsymbol{v}).$$

Due to Lemma 3.5.8 the process $\mathcal{U}(\mathtt{r},\boldsymbol{v},\boldsymbol{u}) \overset{\mathrm{def}}{=} \frac{1}{2\omega\mathtt{r}}\boldsymbol{u}^{\top}\mathcal{Y}(\boldsymbol{v})$ satisfies for every $\mathtt{r}$ condition $(\mathcal{E}d)$ (see (3.5.4)) as process on

$$U(\mathtt{r}^*) \overset{\mathrm{def}}{=} \varUpsilon_{\circ}(\mathtt{r}^*) \times B_{\mathtt{r}^*}(0). \tag{3.5.9}$$

Further $\sup_{(\boldsymbol{v},\boldsymbol{u})\in U(\mathtt{r}))}\mathcal{U}(\mathtt{r},\boldsymbol{v},\boldsymbol{u})$ increases in $\mathtt{r}$. This allows to apply Theorem 3.5.6 to obtain the desired result. Set

$$d((\boldsymbol{v},\boldsymbol{u}),(\boldsymbol{v}^{\circ},\boldsymbol{u}^{\circ}))^2 = \|\mathcal{D}(\boldsymbol{v}-\boldsymbol{v}^{\circ})\|^2 + \|\boldsymbol{u}-\boldsymbol{u}^{\circ}\|^2.$$

We get on a set of probability greater than $1-\mathrm{e}^{-\mathtt{x}}$

$$\begin{aligned}&\sup_{(\boldsymbol{v},\boldsymbol{u})\in U(\mathtt{r}^*)}\left\{\frac{1}{6\omega\nu_1\mathtt{r}}\boldsymbol{u}^{\top}\mathcal{Y}(\boldsymbol{v}) - \|\mathcal{D}(\boldsymbol{v}-\boldsymbol{v}^*)\|^2 - \|\boldsymbol{u}\|^2\right\}\\ &\quad\le \mathfrak{z}_Q\Big(\mathtt{x},\mathbb{Q}\big(U(\mathtt{r}^*)\big)\Big)^2.\end{aligned}$$

The constant $\mathbb{Q}\big(U(\mathtt{r}^*)\big) > 0$ quantifies the complexity of the set $U(\mathtt{r}^*) \subset \mathbb{R}^{p^*}\times\mathbb{R}^{p^*}$. We point out that due to Lemma 3.5.2 we have $\mathbb{Q}(M) \le 2p^*$ for compact $M \subset \mathbb{R}^{p^*}$. This gives $\mathbb{Q}\big(U\big) = 4p^*$. Finally observe that

$$\begin{aligned}&\sup_{\mathtt{r}\le\mathtt{r}^*}\sup_{\boldsymbol{v}\in\varUpsilon_{\circ}(\mathtt{r})}\left\{\frac{1}{6\omega\nu_1}\|\mathcal{Y}(\boldsymbol{v})\| - 2\mathtt{r}^2\right\}\\ &\quad\le \sup_{\mathtt{r}\le\mathtt{r}^*}\sup_{(\boldsymbol{v},\boldsymbol{u})\in U(\mathtt{r})}\left\{\frac{1}{6\omega\nu_1\mathtt{r}}\boldsymbol{u}^{\top}\mathcal{Y}(\boldsymbol{v}) - \|\mathcal{D}(\boldsymbol{v}-\boldsymbol{v}^*)\|^2 - \|\boldsymbol{u}\|^2\right\}\\ &\quad\le \sup_{(\boldsymbol{v},\boldsymbol{u})\in U(\mathtt{r}^*)}\left\{\frac{1}{6\omega\nu_1\mathtt{r}}\boldsymbol{u}^{\top}\mathcal{Y}(\boldsymbol{v}) - \|\mathcal{D}(\boldsymbol{v}-\boldsymbol{v}^*)\|^2 - \|\boldsymbol{u}\|^2\right\}.\end{aligned}$$

□

Lemma 3.5.8. *Suppose that* $\mathcal{Y}(\boldsymbol{v})$ *satisfies* $(\mathcal{E}d)$ *from* (3.5.4) *with norm*

$$d(\boldsymbol{v},\boldsymbol{v}^{\circ}) = \omega\|\mathcal{D}_0(\boldsymbol{v}-\boldsymbol{v}^{\circ})\|,$$

for any $\boldsymbol{u}\in\mathbb{R}^{p^*}$ *with* $\|\boldsymbol{u}\|=1$. *Then the process* $\mathcal{U}(\mathtt{r},\boldsymbol{v},\boldsymbol{u}) = \frac{1}{2\omega\mathtt{r}}\mathcal{Y}(\boldsymbol{v})^{\top}\boldsymbol{u}$ *satisfies* $(\mathcal{E}d)$ *with* $|\lambda|\le\mathtt{g}/2$, $\nu^2 = 4\nu_0^2$, *with norm*

$$d((\boldsymbol{v},\boldsymbol{u}),(\boldsymbol{v}^{\circ},\boldsymbol{u}^{\circ}))^2 = \|\mathcal{D}(\boldsymbol{v}-\boldsymbol{v}^{\circ})\|^2 + \|\boldsymbol{u}-\boldsymbol{u}^{\circ}\|^2,$$

and $U\subset\mathbb{R}^{2p^*}$ *defined in* (3.5.9), *i.e. for any* $(\boldsymbol{v},\boldsymbol{u}_1),(\boldsymbol{v}^{\circ},\boldsymbol{u}_2)\in U$

$$\log\mathbb{E}\exp\biggl\{\lambda\frac{\mathcal{U}(\boldsymbol{v},\boldsymbol{u}_1)-\mathcal{U}(\boldsymbol{v}^{\circ},\boldsymbol{u}_2)}{d((\boldsymbol{v},\boldsymbol{u}_1),(\boldsymbol{v}^{\circ},\boldsymbol{u}_2))}\biggr\} \le \frac{\nu_0^2\lambda^2}{2}, \qquad |\lambda|\le\mathtt{g}/2.$$

Proof. Let $(\boldsymbol{v},\boldsymbol{u}_1),(\boldsymbol{v}^{\circ},\boldsymbol{u}_2)\in U$ and w.l.o.g. $\boldsymbol{u}_1\le\|\mathcal{D}(\boldsymbol{v}-\boldsymbol{v}^*)\|\le\|\mathcal{D}(\boldsymbol{v}^{\circ}-\boldsymbol{v}^*)\|$. By the Hölder inequality and (3.5.3), we find

$$\begin{aligned}
&\log\mathbb{E}\exp\biggl\{\lambda\frac{\mathcal{U}(\boldsymbol{v},\boldsymbol{u}_1)-\mathcal{U}(\boldsymbol{v},\boldsymbol{u}_2)}{d((\boldsymbol{v},\boldsymbol{u}_1),(\boldsymbol{v}^{\circ},\boldsymbol{u}_2))}\biggr\}\\
&\quad=\log\mathbb{E}\exp\biggl\{\lambda\frac{\mathcal{U}(\boldsymbol{v},\boldsymbol{u}_1)-\mathcal{U}(\boldsymbol{v}^{\circ},\boldsymbol{u}_1)+\mathcal{U}(\boldsymbol{v}^{\circ},\boldsymbol{u}_1)-\mathcal{U}(\boldsymbol{v}^{\circ},\boldsymbol{u}_2)}{d((\boldsymbol{v},\boldsymbol{u}_1),(\boldsymbol{v}^{\circ},\boldsymbol{u}_2))}\biggr\}\\
&\quad\le\frac{1}{2}\log\mathbb{E}\exp\biggl\{2\lambda\frac{\boldsymbol{u}_1^{\top}\bigl(\mathcal{Y}(\boldsymbol{v})-\mathcal{Y}(\boldsymbol{v}^{\circ})\bigr)}{\omega\|\mathcal{D}(\boldsymbol{v}-\boldsymbol{v}^{\circ})\|\mathtt{r}}\biggr\}\\
&\qquad+\frac{1}{2}\log\mathbb{E}\exp\biggl\{2\lambda\frac{(\boldsymbol{u}_1^{\top}-\boldsymbol{u}_2^{\top})\mathcal{Y}(\boldsymbol{v}^{\circ})}{\omega\|\boldsymbol{u}_1-\boldsymbol{u}_2\|\mathtt{r}}\biggr\}\\
&\quad\le\sup_{\|\boldsymbol{u}\|\le1}\frac{1}{2}\log\mathbb{E}\exp\biggl\{2\lambda\frac{\boldsymbol{u}^{\top}\bigl(\mathcal{Y}(\boldsymbol{v})-\mathcal{Y}(\boldsymbol{v}^{\circ})\bigr)}{\omega\|\mathcal{D}(\boldsymbol{v}-\boldsymbol{v}^{\circ})\|}\biggr\}\\
&\qquad+\sup_{\|\boldsymbol{u}\|\le1}\frac{1}{2}\log\mathbb{E}\exp\biggl\{2\lambda\frac{\boldsymbol{u}^{\top}\bigl(\mathcal{Y}(\boldsymbol{v}^{\circ})-\mathcal{Y}(\boldsymbol{v}^{*})\bigr)}{\omega\|\mathcal{D}(\boldsymbol{v}-\boldsymbol{v}^{*})\|}\biggr\}\\
&\quad\le\frac{4\nu_0^2\lambda^2}{2},\qquad\lambda\le\mathtt{g}/2.
\end{aligned}$$

□

With the same arguments one can prove the following slightly different version of the previous theorem for the case that for a random process $\breve{\mathcal{Y}}(\boldsymbol{v})\in\mathbb{R}^{p}$ we need a bound of the kind

$$\mathbb{P}\Biggl(\sup_{\boldsymbol{v}\in\varUpsilon_{\circ}(\mathtt{r}^*)}\|\breve{\mathcal{Y}}(\boldsymbol{v})\|\ge\omega\mathfrak{z}_1(\mathtt{x},2p^*+2p)\mathtt{r}\Biggr)\le\mathrm{e}^{-\mathtt{x}}.$$

Theorem 3.5.9. *Let a random* p*-vector process* $\breve{\mathcal{Y}}(\boldsymbol{v})$ *fulfill* $\breve{\mathcal{Y}}(\boldsymbol{v}^*) = 0$*,* $\mathbb{E}\breve{\mathcal{Y}}(\boldsymbol{v}) \equiv 0$*. Furthermore assume that* $\breve{\mathcal{U}}(\boldsymbol{v}) \stackrel{\text{def}}{=} \breve{\mathcal{Y}}(\boldsymbol{v})$ *satisfies* $(\mathcal{E}d)$ *from* (3.5.4) *with norm*

$$d(\boldsymbol{v}, \boldsymbol{v}^\circ) = \omega \|\mathcal{D}_0(\boldsymbol{v} - \boldsymbol{v}^\circ)\|.$$

Then for each $\mathtt{r} > 0$*, on a set of probability greater* $1 - \mathrm{e}^{-\mathtt{x}}$

$$\sup_{\boldsymbol{v} \in \Upsilon_\circ(\mathtt{r})} \|\breve{\mathcal{Y}}(\boldsymbol{v})\| \le 6\omega\nu_1\mathfrak{z}_1(\mathtt{x}, 2p^* + 2p)\mathtt{r}.$$

Remark 3.5.7. In cases when $\mathfrak{z}_1(\mathtt{x}, 2p^* + 2p) \ll \mathtt{r}$ this version can be substantially sharper than Theorem 3.5.7.

Proof. The proof uses the representation

$$\|\breve{\mathcal{Y}}(\boldsymbol{v})\| = \sup_{\|\boldsymbol{u}\| \le \mathtt{r}} \frac{1}{\mathtt{r}} \boldsymbol{u}^\top \breve{\mathcal{Y}}(\boldsymbol{v}).$$

This implies

$$\sup_{\boldsymbol{v} \in \Upsilon_\circ(\mathtt{r})} \|\breve{\mathcal{Y}}(\boldsymbol{v})\| = 2 \sup_{\boldsymbol{v} \in \Upsilon_\circ(\mathtt{r})} \sup_{\|\boldsymbol{u}\| \le \mathtt{r}} \frac{1}{2\mathtt{r}} \boldsymbol{u}^\top \breve{\mathcal{Y}}(\boldsymbol{v}).$$

Just as shown in Lemma 3.5.8 the process $\breve{\mathcal{U}}(\boldsymbol{v}, \boldsymbol{u}) \stackrel{\text{def}}{=} \frac{1}{2\mathtt{r}} \boldsymbol{u}^\top \breve{\mathcal{Y}}(\boldsymbol{v})$ satisfies condition $(\mathcal{E}d)$ as process on $\mathbb{R}^{p^*} \times \mathbb{R}^p$. This allows to apply Theorem 3.5.5 to obtain the desired result. We get on a set of probability greater $1 - \mathrm{e}^{-\mathtt{x}}$

$$\begin{aligned} \sup_{\boldsymbol{v} \in \Upsilon_\circ(\mathtt{r})} \|\breve{\mathcal{Y}}(\boldsymbol{v})\| &= 2 \sup_{\boldsymbol{v} \in \Upsilon_\circ(\mathtt{r})} \sup_{\|\boldsymbol{u}\| \le \mathtt{r}} \left\{ \frac{1}{2\mathtt{r}} \boldsymbol{u}^\top \breve{\mathcal{Y}}(\boldsymbol{v}) \right\} \\ &\le 3\nu_1\mathfrak{z}_1\Big(\mathtt{x}, \mathbb{Q}\big(\Upsilon_\circ(\mathtt{r}) \times \mathcal{B}_{\mathtt{r}}(0)\big)\Big) 2\mathtt{r}. \end{aligned}$$

It remains to note that due to Lemma 3.5.2 $\mathbb{Q}\big(\Upsilon_\circ(\mathtt{r}) \times \mathcal{B}_{\mathtt{r}}(0)\big) \le 2(p^* + p)$. □

3.5.3 A bound for the spectral norm of a random matrix process

We want to derive for a random process $\breve{\mathcal{Y}}(\boldsymbol{v}) \in \mathbb{R}^{p^* \times p^*}$ a bound of the kind

$$\mathbb{P}\left(\sup_{\boldsymbol{v} \in \Upsilon_\circ(\mathtt{r})} \left\{ \|\breve{\mathcal{Y}}(\boldsymbol{v})\| \right\} \ge \mathtt{C}\omega_2\mathfrak{z}_1(\mathtt{x}, p^*)\mathtt{r} \right) \le \mathrm{e}^{-\mathtt{x}}.$$

We derive such a bound in a very similar manner to Theorem 3.5.9. For this let $\mathcal{Y}(\boldsymbol{v})$ be a smooth centered random process with values in $\mathbb{R}^{p^* \times p^*}$ and let

$\mathcal{D} : \mathbb{R}^{p^*} \to \mathbb{R}^{p^*}$ be some linear operator. We aim at bounding the maximum of the spectral norm $\|\mathcal{Y}(\boldsymbol{v})\|$ over a vicinity $\varUpsilon_\circ(\mathtt{r}) \overset{\mathrm{def}}{=} \{\|\mathcal{D}(\boldsymbol{v} - \boldsymbol{v}^*)\| \le \mathtt{r}\}$ of $\boldsymbol{v}^*$. Suppose that $\mathcal{Y}(\boldsymbol{v})$ satisfies for each $0 < \mathtt{r} < \mathtt{r}^*$ and for all pairs $\boldsymbol{v}, \boldsymbol{v}^\circ \in \varUpsilon_\circ(\mathtt{r}) = \bigl\{\boldsymbol{v} \in \varUpsilon \colon \|\boldsymbol{v} - \boldsymbol{v}^*\| \le \mathtt{r}\bigr\} \subset \mathbb{R}^{p^*}$

$$\sup_{\|\boldsymbol{u}_1\| \le 1} \sup_{\|\boldsymbol{u}_2\| \le 1} \log \mathbb{E} \exp\biggl\{ \lambda \frac{\boldsymbol{u}_1^\top \bigl(\mathcal{Y}(\boldsymbol{v}) - \mathcal{Y}(\boldsymbol{v}^\circ)\bigr) \boldsymbol{u}_2}{\omega_2 \|\mathcal{D}(\boldsymbol{v} - \boldsymbol{v}^\circ)\|} \biggr\} \le \frac{\nu_2^2 \lambda^2}{2}. \quad (3.5.10)$$

Remark 3.5.8. In the setting of Theorem 5.2.3 we have

$$\mathcal{Y}(\boldsymbol{v}) = \mathcal{D}^{-1} \nabla^2 \boldsymbol{\zeta}(\boldsymbol{v}) - \mathcal{D}^{-1} \nabla^2 \boldsymbol{\zeta}(\boldsymbol{v}^*),$$

and condition (3.5.10) becomes $(\mathcal{E}\mathcal{D}_2)$ from 4.2.1.

Theorem 3.5.10. *Let a random process* $\mathcal{Y}(\boldsymbol{v}) \in \mathbb{R}^{p^* \times p^*}$ *fulfill* $\mathcal{Y}(\boldsymbol{v}^*) = 0$ *and let condition* (3.5.10) *be satisfied. Then for each* $0 \le \mathtt{r} \le \mathtt{r}^*$, *on a set of probability greater than* $1 - \mathrm{e}^{-\mathtt{x}}$

$$\sup_{\boldsymbol{v} \in \varUpsilon_\circ(\mathtt{r})} \|\mathcal{Y}(\boldsymbol{v})\| \le 9 \omega_2 \nu_2 \mathfrak{z}_1(\mathtt{x}, 6p^*) \mathtt{r},$$

with $\mathtt{g}_0 = \nu_2 \mathtt{g}$.

Remark 3.5.9. Note that the entropy of the original set $\varUpsilon_\circ(\mathtt{r}) \subset \mathbb{R}^{p^*}$ is multiplied by 3. Thus in order to control the spectral norm $\|\mathcal{Y}(\boldsymbol{v})\|$ one only pays with this factor.

Proof. The proof is nearly identical to that of Theorem 3.5.9 once one uses Lemma 3.5.11 and the representation

$$\|\mathcal{Y}(\boldsymbol{v})\| = \omega_2 \sup_{\|\boldsymbol{u}_2\| \le \mathtt{r}} \sup_{\|\boldsymbol{u}_2\| \le \mathtt{r}} \frac{1}{\omega_2 \mathtt{r}^2} \boldsymbol{u}_1^\top \breve{\mathcal{Y}}(\boldsymbol{v}) \boldsymbol{u}_2.$$

We omit the details. □

Lemma 3.5.11. *Suppose that* $\mathcal{Y}(\boldsymbol{v}) \in \mathbb{R}^{p^* \times p^*}$ *satisfies for each* $\|\boldsymbol{u}_1\| \le 1$, $\|\boldsymbol{u}_2\| \le 1$ *and* $|\lambda| \le \mathtt{g}$ *the inequality* (3.5.10). *Then the process* $\mathcal{U}(\boldsymbol{v}, \boldsymbol{u}_1, \boldsymbol{u}_2) = \frac{1}{3\omega_2 \mathtt{r}^2} \boldsymbol{u}_1^\top \mathcal{Y}(\boldsymbol{v})^\top \boldsymbol{u}_2$ *satisfies* $(\mathcal{E}d)$ *from* (3.5.4) *with* $|\lambda| \le \mathtt{g}/3$, $\nu_0 = \nu_2/3$ *and* $d((\boldsymbol{v}, \boldsymbol{u}_1, \boldsymbol{u}_2), (\boldsymbol{v}^\circ, \boldsymbol{u}_1^\circ, \boldsymbol{u}_2^\circ))^2 = \|\mathcal{D}(\boldsymbol{v} - \boldsymbol{v}^\circ)\|^2 + \|\boldsymbol{u}_1 - \boldsymbol{u}_1^\circ\|^2 + \|\boldsymbol{u}_2 - \boldsymbol{u}_2^\circ\|^2$ *as a process on*

$$U(\mathtt{r}) \overset{\mathrm{def}}{=} \varUpsilon_\circ(\mathtt{r}) \times B_1(0) \times B_1(0) \subset \mathbb{R}^{3p^*}.$$

This means that for any $(\boldsymbol{v}, \boldsymbol{u}_1, \boldsymbol{u}_2), (\boldsymbol{v}^\circ, \boldsymbol{u}_1^\circ, \boldsymbol{u}_2^\circ) \in U$

$$\log \mathbb{E} \exp\biggl\{ \frac{\lambda}{3} \frac{\mathcal{U}(\boldsymbol{v}, \boldsymbol{u}_1, \boldsymbol{u}_2) - \mathcal{U}(\boldsymbol{v}^\circ, \boldsymbol{u}_1^\circ, \boldsymbol{u}_2^\circ)}{d((\boldsymbol{v}, \boldsymbol{u}_1, \boldsymbol{v}_2), (\boldsymbol{v}^\circ, \boldsymbol{u}_1^\circ, \boldsymbol{u}_2^\circ))} \biggr\} \le \frac{\nu_2^2 \lambda^2}{2}, \qquad |\lambda| \le \mathtt{g}.$$

Remark 3.5.10. The proof is nearly identical to that of Lemma 3.5.8, which is why we omit it.

Chapter 4

A new approach to analyze profile M-estimators

4.1 Introduction

In this chapter we present an alternative finite sample approach to analyze the properties of the estimator defined in (1.0.3). It is largely based on [4], [3] and [2]. As in the Chapter 3 some parts of this chapter are rather technical so we fist want to convey some intuition about the central steps. Similarly to [52] the approach consists of two parts. In the first one we again control the large deviations of

$$\widetilde{\boldsymbol{\upsilon}} \stackrel{\mathrm{def}}{=} \operatorname*{argmax}_{\boldsymbol{\upsilon} \in \varUpsilon} \mathcal{L}(\boldsymbol{\upsilon}),$$

i.e. we seek for a radius $\mathtt{r}_0(\mathtt{x}) > 0$ such that

$$I\!\!P(\widetilde{\boldsymbol{\upsilon}} \in \varUpsilon_{\circ}(\mathtt{r}_0)) \geq 1 - \mathrm{e}^{-\mathtt{x}}, \tag{4.1.1}$$

where $\varUpsilon_{\circ}(\mathtt{r})$ is a ball of radius $\mathtt{r} > 0$ in the intrinsic semi-metric corresponding to the process $\mathcal{L}(\boldsymbol{\upsilon})$. For this we employ the technique presented in Section 3.3 which is why we include the conditions $(\mathcal{L}\mathtt{r})$ and $(\mathcal{E}\mathtt{r})$ into the list in Section 4.2.1. See Section 4.2.3 for a precise formulation.

The second part consists in the careful analysis of the properties of $\widetilde{\boldsymbol{\upsilon}}$ and $\widetilde{\boldsymbol{\theta}}$ in the local elliptic set $\varUpsilon_{\circ}(\mathtt{r}_0)$ around $\boldsymbol{\upsilon}^*$ in (1.0.7). This step is similar to the ideas behind Theorem 2.2.7. Simplified it works as follows. On the local neighborhood $\varUpsilon_{\circ}(\mathtt{r})$ we approximate

$$\begin{aligned} \nabla\mathcal{L}(\boldsymbol{\upsilon}) &= \nabla\mathcal{L}(\boldsymbol{\upsilon}^*) - \mathcal{D}^2(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^*) + \tau(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^*), \\ \tau(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^*) &= \nabla\mathcal{L}(\boldsymbol{\upsilon}) - \nabla\mathcal{L}(\boldsymbol{\upsilon}^*) + \mathcal{D}^2(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^*). \end{aligned}$$

Using that $\nabla \mathcal{L}(\widetilde{\boldsymbol{\upsilon}}) = 0$ this means that

$$\varPi_{\boldsymbol{\theta}} \mathcal{D}^{-2} \tau(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^*) = \boldsymbol{\theta} - \boldsymbol{\theta}^* - \varPi_{\boldsymbol{\theta}} \mathcal{D}^{-2} \nabla \mathcal{L}(\boldsymbol{\upsilon}^*).$$

Section 4.A.2 provides the following bound on a set of probability of at least $1 - \mathrm{e}^{-\mathtt{x}}$:

$$\sup_{\boldsymbol{\upsilon} \in \varUpsilon_\circ(\mathtt{r})} \| \breve{D}^{-1} \varPi_{\boldsymbol{\theta}} \mathcal{D}^{-2} \tau(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^*) \| \leq \breve{\Diamond}(\mathtt{r}, \mathtt{x}),$$

where $\breve{\Diamond}(\mathtt{r}, \mathtt{x})$ is a small error and where $\breve{D}^{-2} = \varPi_{\boldsymbol{\theta}} \mathcal{D}^{-2} \varPi_{\boldsymbol{\theta}}^\top \in \mathbb{R}^{p \times p}$ with the full information matrix $\mathcal{D}^2 = \nabla^2 I\!\!E \mathcal{L}(\boldsymbol{\upsilon}^*)$. In combination with the deviation bound (4.1.1) this leads to the following Fisher and Wilks type expansions: With probability greater than $1 - 2\mathrm{e}^{-\mathtt{x}}$

$$\begin{aligned} \big\| \breve{D}(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) - \breve{\boldsymbol{\xi}} \big\| &\leq \breve{\Diamond}(\mathtt{r}_0, \mathtt{x}), \\ \big| \max_{\boldsymbol{\eta}} \mathcal{L}(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\eta}) - \max_{\boldsymbol{\eta}} \mathcal{L}(\boldsymbol{\theta}^*, \boldsymbol{\eta}) - \|\breve{\boldsymbol{\xi}}\|^2/2 \big| &\leq \mathtt{C}\sqrt{p + \mathtt{x}}\, \breve{\Diamond}(\mathtt{r}_0, \mathtt{x}). \end{aligned}$$

In case of correctly specified i.i.d models $\breve{D}^2$ is the covariance matrix of the efficient influence function; see Section 2.1. The random vector

$$\breve{\boldsymbol{\xi}} = \breve{D}^{-1} \varPi_{\boldsymbol{\theta}} \mathcal{D}^{-2} \nabla \mathcal{L}(\boldsymbol{\upsilon}^*),$$

satisfies $I\!\!E \breve{\boldsymbol{\xi}} = 0$ and $I\!\!E \|\breve{\boldsymbol{\xi}}\|^2 \asymp p$. Moreover the general deviation bounds for the deviation of quadratic forms from Section 3.4 apply to $\|\breve{\boldsymbol{\xi}}\|^2$ (see Remark 4.2.11 for details). In the important i.i.d. case, the error term $\breve{\Diamond}(\mathtt{r}_0, \mathtt{x})$ can be bounded by $\mathtt{C}(p^* + \mathtt{x})/\sqrt{n}$ and $\breve{\boldsymbol{\xi}}$ is asymptotically normal.

We begin with developing the results for the case that the full parameter space $\varUpsilon$ is a subset of the Euclidean space of dimension p^*. In Section 4.3 we will exemplify how to extend our approach to the case when $\boldsymbol{\upsilon}$ is a functional parameter using the so called sieve approach; see e.g. [49].

The chapter is organized as follows. First we present the conditions employed for our results in Section 4.2.1. Section 4.2.2 introduces the objects and tools of the analysis and presents the main result. In Section 4.2.3 we explain how to obtain the radius $\mathtt{r}_0$ in (4.1.1) and how to improve the main result under slightly stronger conditions. Section 4.2.4 explains how the results translate to the case of i.i.d. samples. Section 4.2.5 addresses the question of critical dimensions and contains an example that shows that the ratio ${p^*}^2/n \to 0$ is critical to obtain the Wilks phenomenon and the Fisher expansion on the class of models that satisfy the conditions of Section 4.2.1 with $\delta/(\mathtt{r}) = \omega = 1/\sqrt{n}$. Section 4.3 discusses how the results can be extended to the case with the infinite full dimension via the sieve approach. We present further conditions on the correlation structure of the full gradient $\nabla \mathcal{L}(\boldsymbol{\upsilon}^*) \in \mathcal{X}$ which will allow to controll the bias induced by the sieve approach in (1.0.13).

4.2 Finite dimensional full parameter space

This section presents our main results on the semiparametric profile estimator which include the Wilks expansion of the profile maximum likelihood $\breve{L}(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}^*) \in \mathbb{R}$ and the Fisher expansion of the profile ME $\widetilde{\boldsymbol{\theta}} \in \mathbb{R}^p$.

4.2.1 Conditions

This section collects the conditions imposed on the model. Let the full dimension of the problem be finite, i.e. $p^* < \infty$. Our conditions involve the symmetric positive definite information matrix $\mathcal{D}_0^2 \in \mathbb{R}^{p^* \times p^*}$ and a central point $\boldsymbol{\upsilon}^\circ \in \mathbb{R}^{p^*}$. In typical situations for $p^* < \infty$, one can set $\boldsymbol{\upsilon}^\circ = \boldsymbol{\upsilon}^*$ where $\boldsymbol{\upsilon}^*$ is the "true point" from (1.0.7). The matrix $\mathcal{D}_0^2 \in \mathbb{R}^{p^* \times p^*}$ can be defined as follows:

$$\mathcal{D}_0^2 = -\nabla^2 I\!\!E \mathcal{L}(\boldsymbol{\upsilon}^\circ).$$

It is worth mentioning that $-\nabla^2 I\!\!E \mathcal{L}(\boldsymbol{\upsilon}^\circ) = \operatorname{Cov}(\nabla \mathcal{L}(\boldsymbol{\upsilon}^*))$ if the model $\boldsymbol{Y} \sim I\!\!P_{\boldsymbol{\upsilon}^*} \in (I\!\!P_{\boldsymbol{\upsilon}})$ is correctly specified and sufficiently regular; see e.g. [29].

Remark 4.2.1. This is not the only possible choice for $\mathcal{D}_0^2$ and $\boldsymbol{\upsilon}^\circ$. In general there is no restriction for the choice of $\mathcal{D}_0^2$, as long as the following list of conditions can be satisfied. The same holds for the matrix $\mathcal{V}_0^2 \in \mathbb{R}^{p^* \times p^*}$ that we introduce below.

In the context of semiparametric estimation, it is convenient to represent the information matrix in block form:

$$\mathcal{D}_0^2 = \begin{pmatrix} D_0^2 & A_0 \\ A_0^\top & \mathrm{H}_0^2 \end{pmatrix}. \tag{4.2.1}$$

Using the matrix $\mathcal{D}_0^2$ and the central point $\boldsymbol{\upsilon}^\circ \in \mathbb{R}^{p^*}$, we define the local set $\varUpsilon_\circ(\mathtt{r}) \subset \varUpsilon \subseteq \mathbb{R}^{p^*}$ with some $\mathtt{r} \geq 0$:

$$\varUpsilon_\circ(\mathtt{r}) \stackrel{\text{def}}{=} \bigl\{\boldsymbol{\upsilon} = (\boldsymbol{\theta}, \boldsymbol{\eta}) \in \varUpsilon \colon \|\mathcal{D}_0(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^\circ)\| \leq \mathtt{r}\bigr\}. \tag{4.2.2}$$

Remark 4.2.2. For readers familiar with [52] we remark that the use of $\mathcal{D}_0$ instead of $\mathcal{V}_0$ in the above definition has no deeper reason but is a choice of convenience.

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

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

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

$$\mathtt{r}_0(\mathtt{x}) \overset{\text{def}}{=} \inf_{\mathtt{r}>0} \left\{ I\!\!P(\widetilde{\boldsymbol{\upsilon}}, \widetilde{\boldsymbol{\upsilon}}_{\boldsymbol{\theta}^*} \in \varUpsilon_\circ(\mathtt{r})) \geq 1 - \mathrm{e}^{-\mathtt{x}} \right\}, \tag{4.2.3}$$

which we set to infinity if $\widetilde{\boldsymbol{\upsilon}} = \emptyset$ or $\widetilde{\boldsymbol{\upsilon}}_{\boldsymbol{\theta}^*} = \emptyset$. Under the conditions $(\mathcal{L}\mathtt{r})$ and $(\mathcal{E}\mathtt{r})$ Theorem 3.3.2 in Section 3.3 states that $\mathtt{r}_0 = \mathtt{r}_0(\mathtt{x}) \approx \mathtt{C}\sqrt{\mathtt{x} + p^*} > 0$.

We assume that the functional $\mathcal{L}(\boldsymbol{\upsilon})\colon \mathbb{R}^{p^*} \to \mathbb{R}$ is sufficiently smooth in $\boldsymbol{\upsilon} \in \mathbb{R}^{p^*}$, $\nabla\mathcal{L}(\boldsymbol{\upsilon}) \in \mathbb{R}^{p^*}$ stands for the gradient and $\nabla^2 I\!\!E\mathcal{L}(\boldsymbol{\upsilon}) \in \mathbb{R}^{p^* \times p^*}$ for the Hessian of the expectation $I\!\!E\mathcal{L} : \mathbb{R}^{p^*} \to \mathbb{R}$ at $\boldsymbol{\upsilon} \in \mathbb{R}^{p^*}$. By smooth enough we mean that all appearing derivatives exist and that we can interchange $\nabla I\!\!E\mathcal{L}(\boldsymbol{\upsilon}) = I\!\!E\nabla\mathcal{L}(\boldsymbol{\upsilon})$ on $\varUpsilon_\circ(\mathtt{r}_0)$, where $\mathtt{r}_0 > 0$ is defined in Equation (4.2.3) and $\varUpsilon_\circ(\mathtt{r})$ in equation (4.2.2).

A sufficient list of conditions

The following three conditions ensure that $\mathcal{D}_0^2$ is not degenerated and further they quantify the smoothness properties on $\varUpsilon_\circ(\mathtt{r})$ of the expected log-likelihood value $I\!\!E\mathcal{L}(\boldsymbol{\upsilon})$ and of the stochastic component $\breve{\nabla}_{\boldsymbol{\theta}}\zeta(\boldsymbol{\upsilon}) \in \mathbb{R}^p$ where

$$\zeta(\boldsymbol{\upsilon}) \overset{\text{def}}{=} \mathcal{L}(\boldsymbol{\upsilon}) - I\!\!E\mathcal{L}(\boldsymbol{\upsilon}), \tag{4.2.4}$$

$$\breve{\nabla}_{\boldsymbol{\theta}} \overset{\text{def}}{=} \nabla_{\boldsymbol{\theta}} - A_0 \mathrm{H}_0^{-2} \nabla_{\boldsymbol{\eta}}. \tag{4.2.5}$$

First we state an *identifiability condition.*

$(\boldsymbol{\mathcal{I}})$ The block matrices in (4.2.1) satisfy for some $\nu < 1$

$$\|\mathrm{H}_0^{-1} A_0^\top D_0^{-1}\| \leq \nu.$$

Remark 4.2.3. The condition $(\mathcal{I})$ allows to define the important $p \times p$ efficient information matrix $\breve{D}_0^2$ which is defined as the inverse of the $\boldsymbol{\theta}$-block of the inverse of the full dimensional matrix $\mathcal{D}_0^2$. The exact formula is given by

$$\breve{D}_0^2 \overset{\text{def}}{=} \left(\varPi_{\boldsymbol{\theta}} \mathcal{D}^{-2} \varPi_{\boldsymbol{\theta}}^\top \right)^{-1} = D_0^2 - A_0 \mathrm{H}_0^{-2} A_0^\top,$$

and $(\mathcal{I})$ ensures that the matrix $\breve{D}_0^2$ is well posed, see for instance [9], Chapter 2.4. In fact $(\mathcal{I})$ is equivalent to the conditions (A.2) and (A.2)' from Section 2.2 as can be seen with Lemma 4.A.6.

$(\breve{\mathcal{L}}_{\mathbf{0}})$ For some $\mathtt{r}^* > 0$ and each $\mathtt{r} \le \mathtt{r}^*$, there is a constant $\breve{\delta}(\mathtt{r})$ such that it holds on the set $\Upsilon_\circ(\mathtt{r})$:

$$\begin{aligned} \|D_0^{-1} D^2(\boldsymbol{\upsilon}) D_0^{-1} - I_p\| &\le \breve{\delta}(\mathtt{r}), \\ \|D_0^{-1} (\mathrm{A}(\boldsymbol{\upsilon}) - A_0) \mathrm{H}_0^{-1}\| &\le \breve{\delta}(\mathtt{r}) \\ \left\|D_0^{-1} A_0 \mathrm{H}_0^{-1} \left(I_m - \mathrm{H}_0^{-1} \mathrm{H}^2(\boldsymbol{\upsilon}) \mathrm{H}_0^{-1}\right)\right\| &\le \breve{\delta}(\mathtt{r}), \end{aligned}$$

where

$$\mathcal{D}(\boldsymbol{\upsilon})^2 \stackrel{\text{def}}{=} -\nabla^2 I\!\!E \mathcal{L}(\boldsymbol{\upsilon}), \quad \mathcal{D}(\boldsymbol{\upsilon}) = \begin{pmatrix} D^2(\boldsymbol{\upsilon}) & \mathrm{A}(\boldsymbol{\upsilon}) \\ \mathrm{A}^\top(\boldsymbol{\upsilon}) & \mathrm{H}^2(\boldsymbol{\upsilon}) \end{pmatrix}.$$

$(\breve{\mathcal{E}}\mathcal{D}_{\mathbf{1}})$ The projected gradient $\breve{\nabla}_{\boldsymbol{\theta}} \boldsymbol{\zeta} : \Upsilon \to \mathbb{R}^p$ of $\boldsymbol{\zeta} : \Upsilon \to \mathbb{R}$ from (4.2.4) almost surely satisfies $\breve{\nabla}_{\boldsymbol{\theta}} \zeta(\boldsymbol{\upsilon}) \to \breve{\nabla}_{\boldsymbol{\theta}} \zeta(\boldsymbol{\upsilon}')$ as $\boldsymbol{\upsilon} \to \boldsymbol{\upsilon}'$ with $\breve{\nabla}_{\boldsymbol{\theta}}$ defined in (4.2.5). Furthermore for all $0 < \mathtt{r} < 4\mathtt{r}_0$, there exist constants $\omega \le 1/2$ and $\breve{\mathtt{g}} > 0$ such that for all $|\mu| \le \breve{\mathtt{g}}$ and $\boldsymbol{\upsilon}, \boldsymbol{\upsilon}' \in \Upsilon_\circ(\mathtt{r})$ for each $\mathtt{r} \le \mathtt{r}^*$ and some $\mathtt{r}^* > 0$

$$\sup_{\boldsymbol{\upsilon}, \boldsymbol{\upsilon}' \in \Upsilon_\circ(\mathtt{r})} \sup_{\substack{\boldsymbol{\gamma} \in \mathbb{R}^p \\ \|\boldsymbol{\gamma}\| \le 1}} \log I\!\!E \exp\left\{ \frac{\mu}{\breve{\omega}} \frac{\boldsymbol{\gamma}^\top \breve{D}^{-1} \{\breve{\nabla}_{\boldsymbol{\theta}} \boldsymbol{\zeta}(\boldsymbol{\upsilon}) - \breve{\nabla}_{\boldsymbol{\theta}} \boldsymbol{\zeta}(\boldsymbol{\upsilon}')\}}{\|\mathcal{D}(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}')\|} \right\} \le \frac{\breve{\nu}_1^2 \mu^2}{2}.$$

Remark 4.2.4. $(\breve{\mathcal{L}}_0)$ describes the local smoothness properties of the function $I\!\!E\mathcal{L}(\boldsymbol{\upsilon})$. In particular, it allows to bound the error of local linear approximation of the gradient $\breve{\nabla}_{\boldsymbol{\theta}} I\!\!E \mathcal{L}(\boldsymbol{\upsilon})$ where the projected gradient $\breve{\nabla}_{\boldsymbol{\theta}}$ is defined in (4.2.5). Under condition $(\breve{\mathcal{L}}_0)$ it follows from the second order Taylor expansion for any $\boldsymbol{\upsilon}, \boldsymbol{\upsilon}' \in \Upsilon_\circ(\mathtt{r})$ (see Lemma 4.A.1)

$$\left\| \breve{D}^{-1} \left(\breve{\nabla} I\!\!E \mathcal{L}(\boldsymbol{\upsilon}) - \breve{\nabla} I\!\!E \mathcal{L}(\boldsymbol{\upsilon}^*) \right) + \breve{D}(\boldsymbol{\theta} - \boldsymbol{\theta}^*) \right\| \le \mathtt{C} \breve{\delta}(\mathtt{r}) \mathtt{r}. \tag{4.2.6}$$

In the proofs we actually only need the inequality (4.2.6) which in some cases can be weaker than $(\breve{\mathcal{L}}_0)$. This reveals that condition $(\breve{\mathcal{L}}_0)$ is strongly related to conditions (B.4) and (B.4)' of Section 2.2. The term $\breve{\delta}(\mathtt{r})$ quantifies how smooth the second derivative is. We impose such a qualified smoothness in order to give finite sample deviation bounds as a function of the radius of the local set $\Upsilon_\circ(\mathtt{r})$.

Remark 4.2.5. Condition $(\breve{\mathcal{E}}\mathcal{D}_1)$ takes the place of (B.3) or (B.3)' of Section 2.2. We show in the proof of Theorem 4.2.2 that it implies (B.3) or at least something very similar, namely that with high probability the term

$$\sup_{\boldsymbol{\upsilon} \in \Upsilon_\circ(\mathtt{r})} \left\| \frac{1}{\sqrt{n}} (1 - I\!\!E) \left([\nabla_{\boldsymbol{\theta}} - A_0 \mathrm{H}_0^{-2} \nabla_{\boldsymbol{\eta}}] \mathcal{L}(\boldsymbol{\upsilon}) - [\nabla_{\boldsymbol{\theta}} - A_0 \mathrm{H}_0^{-2} \nabla_{\boldsymbol{\eta}}] \mathcal{L}(\boldsymbol{\upsilon}^*) \right) \right\|,$$

is small. It is also strongly related to the assumption of Donsker- and and Glivenko-Cantelli properties, i.e. (B.3)' of Section 2.2. In fact one can use $(\breve{\mathcal{E}}\mathcal{D}_1)$ to show - using similar arguments to those in Section 3.5- that the Dudley integral based on covering numbers is finite. This allows to infer that the class

$$\left\{\breve{\nabla}_{\boldsymbol{\theta}}\boldsymbol{\zeta}(\boldsymbol{\upsilon}),\, \boldsymbol{\upsilon} \in \Upsilon_{\circ}(\mathtt{r})\right\},$$

is $I\!\!P$-Donsker (for example Theorem 2.5 of [34]). Note that in linear models or regressions with bounded regressors this condition is automatically satisfied. In the single-index example this condition becomes a condition on the smoothness of the employed basis functions $\boldsymbol{e}_k : \mathbb{R} \to \mathbb{R}$ and a subexponential moment bound on the additive noise $\varepsilon \in \mathbb{R}$, see condition $(\mathbf{Cond}_{\varepsilon})$ in Chapter 6.

The above conditions are sufficient to prove our main results. But we include another condition that allows to control the deviations of $\|\breve{D}^{-1}\breve{\nabla}\boldsymbol{\zeta}(\boldsymbol{\upsilon}^*)\|$.

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

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

Remark 4.2.6. One possible and natural choice for the matrices $\breve{V}^2 \in \mathbb{R}^{p\times p}$ and $\mathcal{V}_0^2 \in \mathbb{R}^{p^*\times p^*}$ (see $(\mathcal{E}\mathcal{D}_0)$ below) is

$$\mathcal{V}_0^2 \overset{\mathrm{def}}{=} \operatorname{Var}\bigl\{\nabla\mathcal{L}(\boldsymbol{\upsilon}^{\circ})\bigr\}, \quad \breve{V}^2 = \operatorname{Cov}(\breve{\nabla}_{\boldsymbol{\theta}}\zeta(\boldsymbol{\upsilon}^{\circ})).$$

But also other matrices could be used as long as $(\breve{\mathcal{E}}\mathcal{D}_0)$ or $(\mathcal{E}\mathcal{D}_0)$ can be satisfied.

Stronger conditions for the full model

In many situations the following, stronger conditions, are easier to verify and allow a further improvement of the results of Theorem 4.2.2 with the help of Proposition 4.2.3:

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

$$\bigl\|\mathcal{D}_0^{-1}\bigl\{\nabla^2 I\!\!E\mathcal{L}(\boldsymbol{\upsilon})\bigr\}\mathcal{D}_0^{-1} - I_{p^*}\bigr\| \le \delta(\mathtt{r}),$$

where $I_p \in \mathbb{R}^{p\times p}$ denotes the identity matrix.

$(\mathcal{E}\mathcal{D}_1)$ There exists a constant $\omega \le 1/2$, such that for all $|\mu| \le \mathtt{g}$ and some $\mathtt{r}^* > 0$ and each $\mathtt{r} \le \mathtt{r}^*$

$$\sup_{\boldsymbol{\upsilon},\boldsymbol{\upsilon}' \in \varUpsilon_\circ(\mathtt{r})} \sup_{\|\gamma\|=1} \log I\!\!E \exp\left\{ \frac{\mu\, \gamma^\top \mathcal{D}_0^{-1}\{\nabla\boldsymbol{\zeta}(\boldsymbol{\upsilon}) - \nabla\boldsymbol{\zeta}(\boldsymbol{\upsilon}')\}}{\omega\, \|\mathcal{D}_0(\boldsymbol{\upsilon}-\boldsymbol{\upsilon}')\|} \right\} \le \frac{\nu_1^2\mu^2}{2}.$$

Remark 4.2.7. Observe the difference to $(\mathcal{L}_0')$ and $(\mathcal{E}\mathcal{D}_1')$ from Chapter 3. The new versions $(\mathcal{L}_0)$ and $(\mathcal{E}\mathcal{D}_1)$ allow the mentioned improvement from $p^{*3/2}/n^{1/2}$ to $p^*/n^{1/2}$ of the bounds for terms related to the approximation error (1.0.8).

$(\mathcal{E}\mathcal{D}_0)$ 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 $|\mu| \le \mathtt{g}$

$$\sup_{\gamma \in \mathbb{R}^{p^*}} \log I\!\!E \exp\left\{ \mu \frac{\langle \nabla\zeta(\boldsymbol{\upsilon}^\circ), \gamma\rangle}{\|\mathcal{V}_0\gamma\|} \right\} \le \frac{\nu_0^2\mu^2}{2}.$$

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

Lemma 4.2.1. *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*

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

Remark 4.2.8. Note that with $(\breve{\mathcal{L}}_0)$, $(\breve{\mathcal{E}}\mathcal{D}_0)$ and $(\breve{\mathcal{E}}\mathcal{D}_1)$ the smoothness and moment conditions do not have to be satisfied for the full gradient $\nabla\mathcal{L}(\cdot)$ but only for the projected one $\left(\nabla_{\boldsymbol{\theta}} + \mathrm{A}\mathrm{H}^{-1}\nabla_{\boldsymbol{\eta}}\right)\mathcal{L}(\cdot)$. This can make a tremendous difference to $(\mathcal{L}_0)$, $(\mathcal{E}\mathcal{D}_0)$ and $(\mathcal{E}\mathcal{D}_1)$ if $\mathrm{A}(\cdot) \in \mathbb{R}^{p\times m}$ is small while $\nabla_{\boldsymbol{\eta}}\mathcal{L}(\cdot)$ is rather rough or possesses bad moment properties. In that case $(\mathcal{E}\mathcal{D}_0)$ and $(\mathcal{E}\mathcal{D}_1)$ might not be satisfied or $\breve{\delta}(\mathtt{r})$, $\breve{\omega}$ and $\breve{\nu}_1$ would be considerably smaller than their counterparts $\delta(\mathtt{r})$, ω and ν_1. This is particularly obvious if $\mathrm{A}(\cdot) \equiv 0$.

Conditions to ensure concentration of the ME

Finally we present two conditions that allow a specific approach to determine a radius $\mathtt{r}_0(\mathtt{x}) > 0$ such that $I\!\!P(\widetilde{\boldsymbol{\upsilon}} \in \varUpsilon(\mathtt{r}_0)) \ge 1 - \mathrm{e}^{\mathtt{x}}$ (see Section 4.2.3). These conditions have to be satisfied on the whole set $\varUpsilon \subseteq \mathbb{R}^{p^*}$. Note, however, that the conditions $(\mathcal{L}\mathtt{r})$ and $(\mathcal{E}\mathtt{r})$ can be substituted with any other set of conditions that allow to determine a value $\mathtt{r}_0$ ensuring $I\!\!P(\widetilde{\boldsymbol{\upsilon}} \in \varUpsilon(\mathtt{r}_0)) \ge 1 - \mathrm{e}^{\mathtt{x}}$.

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

$$\frac{-I\!\!E\mathcal{L}(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^\circ)}{\|\mathcal{D}_0(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^\circ)\|^2} \geq \mathtt{b}(\mathtt{r}), \qquad \boldsymbol{\upsilon} \in \varUpsilon_\circ(\mathtt{r}).$$

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

$$\sup_{\boldsymbol{\upsilon} \in \varUpsilon_\circ(\mathtt{r})} \sup_{\mu \leq \mathtt{g}(\mathtt{r})} \sup_{\gamma \in \mathbb{R}^{p^*}} \log I\!\!E \exp\left\{\mu \frac{\langle \nabla\zeta(\boldsymbol{\upsilon}), \gamma\rangle}{\|\mathcal{D}_0 \gamma\|}\right\} \leq \frac{\nu_{\mathtt{r}}^2 \mu^2}{2}.$$

Remark 4.2.9. These two conditions serve a qualified a priori concentration result for the full estimator $\widetilde{\boldsymbol{\upsilon}}$, of the type $I\!\!P\{\widetilde{\boldsymbol{\upsilon}} \in \varUpsilon_\circ(\mathtt{r}_0(\mathtt{x}))\} \geq 1 - \mathrm{e}^{-\mathtt{x}}$. Condition $(\mathcal{L}\mathtt{r})$ is satisfied for many estimators that employ some least square functional as we do for the single-index model in Chapter 6. In a more general setting it could be combined with yet another even rougher a priori consistency result $I\!\!P(\widetilde{\boldsymbol{\upsilon}} \in U(\boldsymbol{\upsilon}^*))$ for some open neighborhood $U(\boldsymbol{\upsilon}^*) \subset \varUpsilon$. Then $(\mathcal{L}\mathtt{r})$ is automatically satisfied as smooth functions are quadratic around their maximum, in this case $I\!\!E\mathcal{L}$ around $\boldsymbol{\upsilon}^*$. Furthermore the condition can be relaxed to $-I\!\!E\mathcal{L}(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^\circ)$ growing with super linear speed in the distance $\|\mathcal{D}(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^*)\|$, see Theorem 2.1 in [53]. In this case the calculations become technically more involved which is why we focus on $(\mathcal{L}\mathtt{r})$ for the sake of readability. $(\mathcal{E}\mathtt{r})$ is a global exponential moment condition and ensures that the norm of the stochastic component $\nabla\zeta(\boldsymbol{\upsilon}) \in \mathbb{R}^{p^*}$ is bounded with high probability. For example in the least square setting with additive noise this is satisfied with $\mathtt{g}(\mathtt{r}) = \infty$ if the additive noise is subgaussian.

Discussion of the Conditions

We want to discuss how restrictive these conditions are and relate them to the assumptions (A.1),(A.2), etc. presented Section 2.2.

Condition $(\mathcal{I})$ actually is equivalent to (A.2) and (A.2)'. Consider the smoothness criteria $(\mathcal{E}\mathcal{D}_1)$ and $(\mathcal{L}_0)$. These become necessary for our approach if the target is the full parameter $\boldsymbol{\upsilon}^* \in \mathbb{R}^{p^*}$, if the accuracy of results needs to be increased (see 4.2.3) or for the convergence of an alternation maximization procedure (see Chapter 5). Our conditions compare well with those of Theorem 2.2.1. One difference is that we specify in $(\mathcal{L}_0)$ how accurate a second order Taylor expansion of $I\!\!E\mathcal{L}$ is, which we quantify with the term $\delta(\mathtt{r})$. Furthermore instead of mere differentiability of $\mathcal{L} - I\!\!E\mathcal{L}$ we need to impose something like Lipschitz continuity of the gradient in $(\mathcal{E}\mathcal{D}_1)$. Similarly the conditions $(\breve{\mathcal{E}}\mathcal{D}_1)$ and $(\breve{\mathcal{L}}_0)$ compare with (B.3) and (B.4) of Theorem 2.2.4 and (B.3)' and (B.4)' of Theorem 2.2.7. Very similar to (B.4) we qualify the smoothness of $\breve{\nabla}\mathcal{L}$ via $\breve{\omega}$ and $\breve{\delta}(\mathtt{r})$. But instead of (B.3) or (B.3)' we assume the exponential bound in $(\breve{\mathcal{E}}\mathcal{D}_1)$ and can exploit the finite dimensional parameter set to obtain the desired uniform bounds in Theorem 3.5.7. We aim not only for vanishing approximation error terms but for expressions that reveal the interplay of full dimension, smoothness of the functional $\mathcal{L}$ and moments of the score. In both settings the quantification of the smoothness enables us to specify the impact of a large or even growing full parameter dimension p^* in Theorem 4.2.2. As we show in Section 4.2.5 a relaxation of these conditions leads to stronger conditions on the ratio of p^* to n. So in terms of smoothness our conditions do not differ substantially from the established theory, and where they differ they do not seem to be stronger than necessary.

A mayor and obvious difference is that we do not only impose smoothness conditions on $\mathcal{L}$ but also rather strong exponential moment conditions in $(\breve{\mathcal{E}}\mathcal{D}_1)$ or $(\mathcal{E}\mathcal{D}_1)$. Usually one only assumes some finite moments of the errors; cf. [29], Chapter 2. We impose more moments for rather pragmatic reasons. The first and most obvious one is that we derive finite sample results and one needs qualified moment bounds to do this in a general setting as the one we work in. Our condition is a bit more restrictive but it allows to obtain finite sample bounds of the kind that with some small $\epsilon > 0$

$$I\!P\left\{\left\|\breve{D}\big(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\big) - \breve{\boldsymbol{\xi}}\right\| \geq \epsilon(p^* + \mathrm{x})\right\} \geq \mathrm{e}^{-\mathtt{x}},$$

i.e. the bounds depend linearly on the exponent $\mathtt{x}$. Without comparable moment bounds these results do not seem to be attainable in such a general setting. Consider for instance the simple model

$$y = \sqrt{\upsilon^*} + \varepsilon \in \mathbb{R}, \quad \widetilde{\upsilon} = \operatorname*{argmax}_{\boldsymbol{\upsilon} \in \mathbb{R}} (y - \sqrt{\upsilon})^2/2,$$

with $\upsilon^* \neq 0$, $\sqrt{x} \stackrel{\text{def}}{=} \operatorname{sign}(x)\sqrt{|x|}$, $I\!E\varepsilon = 0$ and $\operatorname{Cov}(\varepsilon) = 1$. Then up to the exponential moments all conditions from above are met with $\breve{D}^2 = \mathcal{D}^2 = \frac{1}{4\boldsymbol{\upsilon}^*}$ and $\breve{\boldsymbol{\xi}} = \varepsilon$. We find

$$|\breve{D}(\widetilde{\upsilon} - \upsilon^*) - \breve{\boldsymbol{\xi}}| = \left|\frac{1}{2\sqrt{\upsilon^*}}\left(y^2 - \upsilon^*\right) - \varepsilon\right| = \left|\left(\frac{2\sqrt{\upsilon^*} + \varepsilon}{2\sqrt{\upsilon^*}} - 1\right)\varepsilon\right| = \frac{\varepsilon^2}{2\sqrt{\upsilon^*}}.$$

If $\log I\!E[\exp(\lambda\varepsilon)] < \lambda^2/2$ we can derive

$$I\!P(|\breve{D}(\widetilde{\upsilon} - \upsilon^*) - \breve{\boldsymbol{\xi}}| \geq 8\sqrt{\upsilon^*}\mathrm{x}) \leq \mathrm{e}^{-\mathtt{x}},$$

while obviously without comparable moment criteria such a result - a linear relation between the exponent on the right-hand side and the bound on the left-hand side - could not be attained.

Secondly, exponential bounds simplify the proofs in Section 3.5 considerably. If the exponential were replaced by polynomial moments the calculations would be by far more tedious and the substitute of Lemma 3.5.4

would involve a more complicated term than the entropy $\mathbb{Q}(\varUpsilon)$ in (3.5.5). Finally exponential bounds allow to state results that hold with a probability greater than $1-\mathtt{C}e^{-\mathtt{x}}$ instead of $1-\mathtt{C}p^{-1}(\mathtt{x})$ for some moment function $p:\mathbb{R}\to\mathbb{R}$, which is a question of taste. Without comparable moment and smoothness bounds our results do not seem to be attainable in such a general setting.

A final difference lies in the consistency assumptions. Instead of (A.1) or (A.1)' we use conditions $(\mathcal{E}\mathtt{r})$ and $(\mathcal{L}\mathtt{r})$. In a way these are the strongest conditions in our list, as they are formulated to hold on the full set $\varUpsilon$. Obviously they represent only one among many options on how to restrict the model to ensure the desired type of a priori consistency of M-estimators. This is why we present the results in a way such that the particular way of obtaining a concentration behavior based on $(\mathcal{E}\mathtt{r})$ and $(\mathcal{L}\mathtt{r})$ and Theorem 3.3.2 can be replaced by any other available technique. An advantage of the approach we follow here is that the obtained bounds are of the same type as those we present for quadratic forms or those we derive in Theorem 4.2.2. This means that if the conditions are all satisfied, all deviation bounds are of similar order.

Section 4.2.4 explains how to satisfy the above conditions in case of i.i.d. observations and a smooth criterion functional $\mathcal{L}$ and hopefully serves some intuition. In Chapter 6 we present a rather sophisticated model for which all conditions can be satisfied under very natural and common assumptions on the model.

4.2.2 Wilks and Fisher expansions

This section states the main results in a finite dimensional framework. First we introduce the main elements of the approach. Let the *information matrix* $\mathcal{D}_0^2\in\mathbb{R}^{p^*\times p^*}$ be from the condition in Section 4.2.1. For the semiparametric $(\boldsymbol{\theta},\boldsymbol{\eta})$-setup, we consider the block representation of the vector $\nabla\stackrel{\text{def}}{=}\nabla\mathcal{L}(\boldsymbol{\upsilon}^*)$ and of the matrix $\mathcal{D}_0^2$

$$\nabla=\begin{pmatrix}\nabla_{\boldsymbol{\theta}}\\ \nabla_{\boldsymbol{\eta}}\end{pmatrix},\quad \mathcal{D}_0^2=\begin{pmatrix}D_0^2 & A_0\\ A_0^\top & \mathrm{H}_0^2\end{pmatrix}.$$

We repeat also the definition of the $p\times p$ matrix $\breve{D}^2$

$$\breve{D}^2=D^2-A_0\mathrm{H}_0^{-2}A_0^\top,$$

and p-vectors $\breve{\nabla}_{\boldsymbol{\theta}}$ and $\breve{\boldsymbol{\xi}}\in\mathbb{R}^p$

$$\breve{\nabla}_{\boldsymbol{\theta}}\stackrel{\text{def}}{=}\nabla_{\boldsymbol{\theta}}\boldsymbol{\zeta}(\boldsymbol{\upsilon}^*)-A_0\mathrm{H}_0^{-2}\nabla_{\boldsymbol{\eta}}\boldsymbol{\zeta}(\boldsymbol{\upsilon}^*),\quad \breve{\boldsymbol{\xi}}\stackrel{\text{def}}{=}\breve{D}^{-1}\breve{\nabla}_{\boldsymbol{\theta}}.$$

The random variable $\breve{\nabla}_{\boldsymbol{\theta}}\in\mathbb{R}^p$ is related to the efficient influence function in semiparametric estimation and the matrix $\breve{D}^2\in\mathbb{R}^{p\times p}$ equals its covariance in case of correct specification.

Remark 4.2.10. It seems worthy to point our that $\breve{D}^{-2}\breve{\nabla}_{\boldsymbol{\theta}} = \varPi_{\boldsymbol{\theta}}\mathcal{D}^{-2}\nabla$, see again [9], Chapter 2.4.

Define the *semiparametric spread* $\breve{\diamondsuit}(\mathtt{r},\mathtt{x}) > 0$ as

$$\breve{\diamondsuit}(\mathtt{r},\mathtt{x}) \stackrel{\mathrm{def}}{=} 4\left(\frac{4}{(1-\nu^2)^2}\breve{\delta}(4\mathtt{r}) + 6\breve{\nu}_1\breve{\omega}\mathfrak{z}_1(\mathtt{x},2p^* + 2p)\right)\mathtt{r}, \qquad (4.2.8)$$

where $\breve{\delta}(\mathtt{r})$ is shown in the condition $(\breve{\mathcal{L}}_0)$ and the constants $\breve{\omega}$, $\breve{\nu}_1$ are from condition $(\breve{\mathcal{E}}\mathcal{D}_1)$ in Section 4.2.1. The value $\mathfrak{z}_1(\mathtt{x},2p^*+2p)$ is related to the entropy of the unit ball in a $\mathbb{R}^{p^*+p}$-dimensional Euclidean space. It is defined in (3.5.7) and one can apply $\mathfrak{z}_1(\mathtt{x},p^*) \cong \sqrt{\mathtt{x}+p^*}$ as long as $\mathtt{x} > 0$ is not too large; see Section 3.5.2. The *semiparametric spread* $\breve{\diamondsuit}(\mathtt{r},\mathtt{x})$ measures the quality of a linear approximation to $\breve{\nabla}\mathcal{L}(\boldsymbol{\upsilon}) - \breve{\nabla}\mathcal{L}(\boldsymbol{\upsilon}^*)$ in the local vicinity the local vicinity $\varUpsilon_\circ(\mathtt{r}) = \bigl\{\boldsymbol{\upsilon} \in \varUpsilon\colon \|\mathcal{D}_0(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^\circ)\| \le \mathtt{r}\bigr\}$, i.e. it provides a bound for the term in (1.0.8). Our results become accurate if $\diamondsuit(\mathtt{r}_0,\mathtt{x})$ is small. The spread will be evaluated in the i.i.d. case in Section 4.2.4 below.

Theorem 4.2.2. *Assume* $(\breve{\mathcal{E}}\mathcal{D}_1)$, $(\breve{\mathcal{L}}_0)$, *and* $(\mathcal{I})$ *with a central point* $\boldsymbol{\upsilon}^\circ = \boldsymbol{\upsilon}^*$ *and some matrix* $\mathcal{D}_0^2$ *and* $4\mathtt{r}_0 \le \mathtt{r}^*$. *Assume further that the sets of maximizers* $\widetilde{\boldsymbol{\upsilon}}$, $\widetilde{\boldsymbol{\upsilon}}_{\boldsymbol{\theta}^*}$ *are not empty. Then it holds on a set* $\varOmega(\mathtt{x}) \subseteq \varOmega$ *of probability greater than* $1 - 2\mathrm{e}^{-\mathtt{x}}$ *for the profile ME* $\widetilde{\boldsymbol{\theta}}$ *in* (1.0.3)

$$\bigl\|\breve{D}\bigl(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\bigr) - \breve{\boldsymbol{\xi}}\bigr\| \le \breve{\diamondsuit}(\mathtt{r}_0,\mathtt{x}), \qquad (4.2.9)$$

$$\bigl|2\breve{L}(\widetilde{\boldsymbol{\theta}},\boldsymbol{\theta}^*) - \|\breve{\boldsymbol{\xi}}\|^2\bigr| \le 9\Bigl(\|\breve{\boldsymbol{\xi}}\| + \breve{\diamondsuit}(\mathtt{r}_0,\mathtt{x})\Bigr)\breve{\diamondsuit}(\mathtt{r}_0,\mathtt{x}), \qquad (4.2.10)$$

where the spread $\breve{\diamondsuit}(\mathtt{r}_0,\mathtt{x})$ *is defined in* (4.2.7) *and where* $\mathtt{r}_0 > 0$ *is defined in* (4.2.3).

Remark 4.2.11. The Wilks expansion claims that the profile maximum likelihood $\breve{L}(\widetilde{\boldsymbol{\theta}},\boldsymbol{\theta}^*) \stackrel{\mathrm{def}}{=} \breve{L}(\widetilde{\boldsymbol{\theta}}) - \breve{L}(\boldsymbol{\theta}^*)$ can be approximated by a quadratic form $\|\breve{\boldsymbol{\xi}}\|^2/2$ with $\breve{\boldsymbol{\xi}} = \breve{D}^{-1}\breve{\nabla}_{\boldsymbol{\theta}}$. In the correctly specified i.i.d setting the vector $\breve{\boldsymbol{\xi}}$ is asymptotically standard normal and the quadratic form $\|\breve{\boldsymbol{\xi}}\|^2 = \|\breve{D}^{-1}\breve{\nabla}_{\boldsymbol{\theta}}\|^2$ converges weakly to a chi-square random variable with $p \in \mathbb{N}$ degrees of freedom, which follows from the central limit theorem and the fact that then $\mathrm{Cov}(\breve{\boldsymbol{\xi}}) = I_p$. In the general case, the behavior of the quadratic form $\|\breve{\boldsymbol{\xi}}\|^2$ depends on the characteristics of the matrix $\breve{I\!\!B} \stackrel{\mathrm{def}}{=} \breve{D}^{-1}\breve{V}^2\breve{D}^{-1}$ where $\breve{V}^2 \in \mathbb{R}^{p\times p}$ is from $(\breve{\mathcal{E}}\mathcal{D}_0)$ and in many cases equals $\breve{V}^2 = \mathrm{Cov}(\breve{\nabla}_{\boldsymbol{\theta}})$. More precisely, one can find an upper quantile function $\mathfrak{z}(\mathtt{x},\breve{I\!\!B})$ of this quadratic form ensuring

$$I\!\!P\bigl(\|\breve{\boldsymbol{\xi}}\| > \mathfrak{z}(\mathtt{x},\breve{I\!\!B})\bigr) \le 2\mathrm{e}^{-\mathtt{x}};$$

see Proposition 3.4.1. One can use the bound $\mathfrak{z}^2(\mathtt{x}, \breve{\mathbb{B}}) \le \mathtt{C}(p + \mathtt{x})$ in most situations. We call $\breve{\mathbb{B}} \in \mathbb{R}^{p \times p}$ *semiparametric misspecification matrix* as it is related to the misspecification matrix introduced in [27]. $\breve{\mathbb{B}}$ is equal to the identity matrix if a correctly specified log likelihood is used.

Remark 4.2.12. One can use the expansion (4.2.9) for the construction of elliptic confidence sets

$$\mathcal{A}(\mathfrak{z}) = \bigl\{\boldsymbol{\theta} : \|\breve{D}(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta})\| \le \mathfrak{z}\bigr\};$$

for some $\mathfrak{z}(\mathtt{x}) > 0$. For this assume that the quantiles of $\|\breve{\boldsymbol{\xi}}\|$ are available or that they can be given up to a small error based on the Berry Esseen theorem (Berry [8]) or Edgeworth expansions (Hall [25]). Let $q_\alpha > 0$ be the $\alpha-$ level quantile of $\|\breve{\boldsymbol{\xi}}\|$. Then we find with the triangular inequality and (4.2.9)

$$\begin{aligned} \mathbb{P}\Bigl\{\boldsymbol{\theta}^* \notin \mathcal{E}\Bigl(q_\alpha + \breve{\diamondsuit}(\mathtt{r}_0, \mathtt{x})\Bigr)\Bigr\} &= \mathbb{P}\Bigl\{\|\breve{D}(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\| \ge q_\alpha + \breve{\diamondsuit}(\mathtt{r}_0, \mathtt{x})\Bigr\} \\ &\le \mathbb{P}\Bigl\{\|\breve{\boldsymbol{\xi}}\| \ge q_\alpha\Bigr\} + 2\mathrm{e}^{-\mathtt{x}} = 1 - \bigl(\alpha - 2\mathrm{e}^{-\mathtt{x}}\bigr), \end{aligned}$$

and

$$\begin{aligned} \mathbb{P}\Bigl\{\boldsymbol{\theta}^* \in \mathcal{E}\Bigl(q_\alpha - \breve{\diamondsuit}(\mathtt{r}_0, \mathtt{x})\Bigr)\Bigr\} &= \mathbb{P}\Bigl\{\|\breve{D}(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\| \le q_\alpha - \breve{\diamondsuit}(\mathtt{r}_0, \mathtt{x})\Bigr\} \\ &\le \mathbb{P}\Bigl\{\|\breve{\boldsymbol{\xi}}\| \le q_\alpha\Bigr\} + 2\mathrm{e}^{-\mathtt{x}} = \alpha + 2\mathrm{e}^{-\mathtt{x}}. \end{aligned}$$

Consequently up to $\breve{\diamondsuit}(\mathtt{r}_0, \mathtt{x})$ and $2\mathrm{e}^{-\mathtt{x}}$ the set $\mathcal{E}\left(q_\alpha\right)$ serves as a confidence set. The choice of $\mathtt{x}$ determines the trade off between the closeness of $q_\alpha \pm \breve{\diamondsuit}(\mathtt{r}_0, \mathtt{x})$ to q_α and the probability level $\alpha + 2\mathrm{e}^{-\mathtt{x}}$ to α.

Remark 4.2.13. The profile maximum likelihood process $\breve{L}(\boldsymbol{\theta})$ can be used for defining the likelihood-based confidence sets of the form

$$\mathcal{E}(\mathfrak{z}) = \{\boldsymbol{\theta} : \breve{L}(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}) \le \mathfrak{z}\}$$

The bound (4.2.10) helps to evaluate the coverage probability $\mathbb{P}\bigl(\boldsymbol{\theta}^* \notin \mathcal{E}(\mathfrak{z})\bigr)$ in terms of deviation properties of the quadratic form $\|\breve{\boldsymbol{\xi}}\|^2$; cf. Corollary 3.2 in [52].

Remark 4.2.14. In the classical finite dimensional case a usual choice for the central point $\boldsymbol{\upsilon}^\circ$ is $\boldsymbol{\upsilon}^\circ = \boldsymbol{\upsilon}^* = \operatorname{argmax}_{\boldsymbol{\upsilon} \in \varUpsilon} \mathbb{E}\mathcal{L}(\boldsymbol{\upsilon})$ and one can define the matrix $\mathcal{D}_0^2$ as $\mathcal{D}_0^2 = -\nabla^2 \mathbb{E}\mathcal{L}(\boldsymbol{\upsilon}^*)$. However, for the sieve semiparametric problem in Section 4.3, we use another definition related to the infinite dimensional model.

4.2.3 Large deviation bounds

In this section we want to present a way to determine a value $\mathtt{r}_0 > 0$ such that the full ME $\widetilde{\boldsymbol{\upsilon}} \in \mathbb{R}^{p^*}$ belongs to the local vicinity $\Upsilon_\circ(\mathtt{r}_0) \subset \mathbb{R}^{p^*}$ with high probability. As a first step we apply Theorem 3.3.2. It is important to note that Theorem 3.3.2 is one particular approach which could be replaced by any other appropriate technique. For instance, in the model with i.i.d. observations, Theorem 5.3 of [29] might serve as a tool. The required conditions can be substantially weakened to upper and lower bounds on the Hellinger distance between models for distinct parameters. We use Theorem 3.3.2 because it applies to M-estimators and finite samples.

But the upper function approach in Theorem 3.3.2 of showing the consistency for an M-estimator can be rather rough and the bound (3.3.7) could lead to quite large values of $\mathtt{r}_0 > 0$. As the obtained value $\mathtt{r}_0 > 0$ enters into the error term $\check{\diamondsuit}(\mathtt{r}_0, \mathtt{x}) > 0$ of Theorem 4.2.2 it is desirable to obtain a general refined bound for $\mathtt{r}_1 \le \mathtt{r}_0$ that still ensures that $I\!\!P(\widetilde{\boldsymbol{\upsilon}} \in \Upsilon_\circ(\mathtt{r}_1)) \ge 1 - \mathtt{C}\mathrm{e}^{-\mathtt{x}}$ with a small constant $\mathtt{C} > 0$. Such an improvement is possible as the following proposition shows. Define the *parametric uniform spread*:

$$\diamondsuit_Q(\mathtt{r}, \mathtt{x}) \stackrel{\mathrm{def}}{=} \delta(\mathtt{r})\mathtt{r} + 6\omega\nu_1 \left(2\mathtt{r}^2 + \mathfrak{z}_Q(\mathtt{x}, 4p^*)^2\right), \tag{4.2.11}$$

with $\mathfrak{z}_Q(\mathtt{x}, 4p^*)$ in (3.5.8). Furthermore with $\mathcal{V}^2 \in \mathbb{R}^{p^* \times p^*}$ from condition $(\mathcal{E}\mathcal{D}_0)$ introduce the *misspecification matrix* $I\!\!B \in \mathbb{R}^{p^* \times p^*}$ given by the famous sandwich formula; see [27]:

$$I\!\!B = \mathcal{D}_0^{-1}\mathcal{V}_0^2\mathcal{D}_0^{-1}.$$

In case of correct model specification with $\mathcal{D}_0^2 = \mathcal{V}_0^2$, the *misspecification matrix* $I\!\!B$ becomes the identity: $I\!\!B = I_{p^*}$. Theorem 3.4.1 tells us that

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

where $\mathfrak{z}(\mathtt{x}, I\!\!B) \le \mathtt{C}\sqrt{\operatorname{tr}(I\!\!B^2) + \mathtt{x}}$ for moderate choice of $\mathtt{x} > 0$, see (3.4.3).

Proposition 4.2.3. *Assume the conditions of Theorem 4.2.2 and additionally assume* $(\mathcal{L}_0)$, $(\mathcal{E}\mathcal{D}_1)$ *and* $(\mathcal{E}\mathcal{D}_0)$ *with* $\mathcal{V}^2 \in \mathbb{R}^{p^* \times p^*}$ *and* $4\mathtt{r}_0 \le \mathtt{r}^*$. *Let* $\mathtt{r}_0(\mathtt{x}) > 0$ *be such that* (4.2.3) *holds. Then*

$$I\!\!P\left\{\widetilde{\boldsymbol{\upsilon}}, \widetilde{\boldsymbol{\upsilon}}_{\boldsymbol{\theta}^*} \in \Upsilon_\circ(\mathtt{r}_1)\right\} \ge 1 - 4\mathrm{e}^{-\mathtt{x}}$$

where

$$\mathtt{r}_1 \le \mathfrak{z}(\mathtt{x}, I\!\!B) + \diamondsuit_Q(R_0, \mathtt{x}) \wedge \mathtt{r}_0(\mathtt{x}).$$

Furthermore if there is some $\epsilon > 0$ *such that* $\delta(\mathtt{r})/\mathtt{r} \vee 6\nu_1\omega \le \epsilon$ *for all* $\mathtt{r} \le \mathtt{r}_0$ *and with* $4\epsilon\left(\mathfrak{z}(\mathtt{x},\mathbb{B}) + \epsilon\mathfrak{z}_Q(\mathtt{x},4p^*)^2\right)^2 \le c < 1$ *and* $4\epsilon\mathtt{r}_0(\mathtt{x}) < 1$ *then*

$$\mathbb{P}\{\widetilde{\boldsymbol{\upsilon}}, \widetilde{\boldsymbol{\upsilon}}_{\boldsymbol{\theta}^*} \in \varUpsilon_\circ(\mathtt{r}_0^*)\} \ge 1 - 4\mathrm{e}^{-\mathtt{x}}$$

where

$$\mathtt{r}_0^* \le \mathfrak{z}(\mathtt{x},\mathbb{B}) + \epsilon\mathfrak{z}_Q(\mathtt{x},4p^*)^2 + \epsilon\frac{2c}{1-c}. \tag{4.2.12}$$

Remark 4.2.15. In cases where $\mathtt{r}_0 \gg p^*$ it might happen that $\diamondsuit(\mathtt{r}_0,\mathtt{x}) = \delta(\mathtt{r}_0)\mathtt{r}_0 + 6\nu_1\omega\mathfrak{z}_1(\mathtt{x},2p^*+2p)\mathtt{r}_0 \gg \mathfrak{z}(\mathtt{x},\mathbb{B})$. This again may be caused by the fact that a multiple of $\delta(\mathtt{r}_0)\mathtt{r}_0$ determines the size of $\diamondsuit(\mathtt{r}_0,\mathtt{x})$. Thanks to the bracketing device the spread $\varDelta_{\boldsymbol{\epsilon}}(\mathtt{r}_0)$ from Equation (3.2.4) only depends linearly and through $\delta(\mathtt{r}_0)$ on $\mathtt{r}_0 > 0$ and could in some settings be significantly smaller than $\diamondsuit(\mathtt{r}_0,\mathtt{x})$ (see Section 3.1). We can exploit this in the following way. Define with the *score covariance matrix* $\mathcal{V}^2 = \mathbb{E}[\nabla\boldsymbol{\zeta}(\boldsymbol{\upsilon}^*)\nabla\boldsymbol{\zeta}(\boldsymbol{\upsilon}^*)^\top] \in \mathbb{R}^{p^*\times p^*}$

$$\mathfrak{a} \stackrel{\mathrm{def}}{=} \inf\{c \in \mathbb{R} : c^2\mathcal{D}^2 \ge \mathcal{V}^2\}.$$

If the initial radius $\mathtt{r}_0 > 0$ from Proposition 4.2.3 additionally satisfies $\mathtt{r}_0(1-\epsilon(\mathtt{r}_0)) \ge \mathfrak{z}(\mathtt{x},\mathbb{B})$ where $\epsilon(\mathtt{r}) = \delta(\mathtt{r}) + 3\nu_1\mathfrak{a}^2\omega\mathtt{r}$ then we can set in Proposition 4.2.3

$$\mathtt{r}_1 \stackrel{\mathrm{def}}{=} \left(\frac{\mathfrak{z}(\mathtt{x},\mathbb{B})}{(1-\epsilon(\mathtt{r}_0))^2(1-\nu)} + \diamondsuit(\mathtt{r}_0,\mathtt{x}) \wedge \sqrt{\varDelta_{\boldsymbol{\epsilon}}(\mathtt{r}_0,\mathtt{x})}\right) \wedge \mathtt{r}_0. \tag{4.2.13}$$

In situations where $\mathtt{r}_0 \gg p^*$ we can expect that $\diamondsuit(\mathtt{r}_0,\mathtt{x}) \gg \sqrt{\varDelta_{\boldsymbol{\epsilon}}(\mathtt{r}_0)}$ such that (4.2.13) can significantly decrease the size of $\mathtt{r}_1 > 0$. The proof of this claim is presented along with the proof of Proposition 4.2.3.

4.2.4 The i.i.d. case

In this section we want to illustrate the results for the case of a smooth i.i.d. model. As explained in the introduction in Chapter 1 this means $\mathbb{Y} = (\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n) \in \bigotimes_{i=1}^n \mathcal{Y}$ and

$$\mathcal{L}(\mathbb{Y},\boldsymbol{\upsilon}) = \frac{1}{n}\sum_{i=1}^n \ell(\boldsymbol{Y}_i,\boldsymbol{\upsilon}), \quad \mathbb{E}_{\mathbb{P}}\mathcal{L}(\boldsymbol{\upsilon}) = \mathbb{E}_{\mathbb{P}_{\boldsymbol{Y}}}\ell(\boldsymbol{Y}_1,\boldsymbol{\upsilon}),$$

where $\ell : \mathcal{Y} \times \varUpsilon \to \mathbb{R}$ is a suitable functional. As above we omit the data in the following and write $\ell_i(\boldsymbol{\upsilon}) \stackrel{\text{def}}{=} \ell(\boldsymbol{Y}_i, \boldsymbol{\upsilon})$. Note that

$$\begin{aligned}
\boldsymbol{\upsilon}^* &\stackrel{\text{def}}{=} \operatorname*{argmax}_{\boldsymbol{\upsilon}\in\varUpsilon} I\!\!E\mathcal{L}(\boldsymbol{\upsilon}) = \operatorname*{argmax}_{\boldsymbol{\upsilon}\in\varUpsilon} I\!\!E\ell(\boldsymbol{\upsilon}),\\
\mathcal{D}^2 &\stackrel{\text{def}}{=} \nabla^2 I\!\!E\mathcal{L}(\boldsymbol{\upsilon}^*) = nd^2 \stackrel{\text{def}}{=} n\nabla^2 I\!\!E\ell(\boldsymbol{\upsilon}^*),\\
\mathcal{V}^2 &\stackrel{\text{def}}{=} \operatorname{Cov}\Bigl(\nabla\boldsymbol{\zeta}(\boldsymbol{\upsilon}^*)\Bigr) = nv^2 \stackrel{\text{def}}{=} n\operatorname{Cov}\Bigl(\nabla(\ell - I\!\!E\ell)(\boldsymbol{\upsilon}^*)\Bigr).
\end{aligned}$$

One way to check the conditions of Section 4.2.1 is to assume that they are met with $\mathcal{L}, \mathcal{D}$ replaced by ℓ, d with some ν_0^*, ω_1^*, $\delta(\mathtt{r}) = \delta^*\mathtt{r}$, $\mathtt{b}(\mathtt{r}) = \mathtt{b}^*$ and $\mathtt{g} = \mathtt{g}_1$. In that case one can easily check the conditions in Section 4.2.1 for the full functional $\mathcal{L}(\boldsymbol{\upsilon}) = \sum_{i=1}^n \ell(y_i, \boldsymbol{\upsilon})$ with $\omega = \omega_1 n^{-1/2}$, $\delta(\mathtt{r}) = \delta^*\mathtt{r}n^{-1/2}$, $\mathtt{b}(\mathtt{r}) = \mathtt{b}^*$, and $\mathtt{g} = \mathtt{g}_1 n^{1/2}$; cf. Lemma 5.1 in [52]. To gain a bit more intuition let us consider the following stronger sufficient list of conditions. Abbreviate

$$\zeta_\ell(\cdot, \boldsymbol{Y}) \stackrel{\text{def}}{=} (\ell - I\!\!E\ell)(\cdot, \boldsymbol{Y}) : \varUpsilon \to \mathbb{R}.$$

(ℓ_0) The matrix valued function $\nabla^2 I\!\!E[\ell(\cdot)] : \varUpsilon \to \mathbb{R}^{p^*\times p^*}$ is locally Lipschitz continuous with Lipschitz constant δ^* in an open neighborhood $U \ni \boldsymbol{\upsilon}^*$.

(ed_1) There are constants $\nu_0^*, \mathtt{g}^* > 0$ such that for all $\boldsymbol{\upsilon} \in U$ the random matrix valued function $\nabla^2\zeta_\ell(\cdot, \boldsymbol{Y}) : \varUpsilon \to \mathbb{R}^{p^*\times p^*}$ satisfies for all $|\lambda| \le \mathtt{g}^*$

$$\begin{aligned}
&\sup_{\substack{\boldsymbol{\gamma}_1,\boldsymbol{\gamma}_2\in\mathbb{R}^{p^*}\\ \|\boldsymbol{\gamma}_1\|=\|\boldsymbol{\gamma}_2\|=1}} \log I\!\!E \sup_{\boldsymbol{\upsilon}^\circ\in\operatorname{conv}(\boldsymbol{\upsilon},\boldsymbol{\upsilon}^*)} \exp\Bigl\{\lambda\boldsymbol{\gamma}_1^\top d^{-1}\nabla^2\zeta_\ell(\boldsymbol{\upsilon}^\circ)d^{-1}\boldsymbol{\gamma}_2\Bigr\}\\
&\le \nu_0^*\lambda^2/2.
\end{aligned}$$

(ed_0) The random vector valued function $\nabla\zeta_\ell(\cdot, \boldsymbol{Y}) : \varUpsilon \to \mathbb{R}^{p^*\times p^*}$ satisfies for all $|\lambda| \le \mathtt{g}^*$ and all $\boldsymbol{\upsilon} \in \varUpsilon$

$$\sup_{\substack{\boldsymbol{\gamma}\in\mathbb{R}^{p^*}\\ \|\boldsymbol{\gamma}\|=1}} \log I\!\!E \exp\Bigl\{\lambda\boldsymbol{\gamma}^\top d^{-1}\nabla\zeta_\ell(\boldsymbol{\upsilon})\Bigr\} \le \nu_0^*\lambda^2/2.$$

$(\ell_{\mathtt{r}})$ There is a constant $\mathtt{b}^* > 0$ such that

$$I\!\!E\left[\ell(\boldsymbol{\upsilon}) - \ell(\boldsymbol{\upsilon}^*)\right] \ge \mathtt{b}^*\|d(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^*)\|^2.$$

$(\boldsymbol{\iota})$ There is a constant $c_d > 0$ such that the matrix $d^2 \stackrel{\text{def}}{=} \nabla^2 I\!\!E\ell(\boldsymbol{\upsilon}^*)$ satisfies $\boldsymbol{\gamma}^\top d^2\boldsymbol{\gamma} \ge c_d\|\boldsymbol{\gamma}\|^2$ for all $\boldsymbol{\gamma} \in \mathbb{R}^{p^*}$.

Lemma 4.2.4. *Assume that* $n \in \mathbb{N}$ *is large enough to ensure that the local neighborhood* $U \subset \varUpsilon$ *of* $\boldsymbol{\upsilon}^*$ *from the conditions* (ℓ_0) *and* (ed_1) *satisfies*

$$\begin{aligned}\varUpsilon_\circ(\mathtt{r}^*) &\stackrel{\mathrm{def}}{=} \{\boldsymbol{\upsilon} \in \varUpsilon : \|\mathcal{D}(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^*)\| \le \mathtt{r}^*\} \\ &= \frac{1}{\sqrt{n}}\{\boldsymbol{\upsilon} \in \varUpsilon : \|d(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^*)\| \le \mathtt{r}^*\} \subseteq U.\end{aligned}$$

Then the conditions (ℓ_0), (ed_1), (ed_0), $(\ell_{\mathtt{r}})$ *and* (ι) *imply* $(\mathcal{L}_0)$, $(\mathcal{E}\mathcal{D}_1)$, $(\mathcal{E}\mathcal{D}_0)$, $(\mathcal{E}\mathcal{D}_{\mathtt{r}})$, $(\mathcal{L}_0)$ *and* $(\mathcal{I})$ *with* $\delta(\mathtt{r}) = \frac{\delta^*}{\sqrt{n}c_d^3}\mathtt{r}$, $\omega = \frac{1}{\sqrt{n}}$, $\mathtt{g} = \sqrt{n}\mathtt{g}^*$, $\nu_1 = \nu_0 = \nu_0^*$, $\mathtt{g}(\mathtt{r}) = \sqrt{n}\mathtt{g}^*$, $\mathtt{b} = \mathtt{b}^*$ *for all* $\mathtt{r} \le \mathtt{r}^*$. *Furthermore* $\nu^2 \ge 1 - \frac{c_d}{\|d_{\boldsymbol{\theta}}^2\| \vee \|h^2\|}$ *where* $d_{\boldsymbol{\theta}}^2 = \varPi_{\boldsymbol{\theta}}^\top d \varPi_{\boldsymbol{\theta}} \in \mathbb{R}^{p \times p}$ *and* $h^2 = \varPi_{\boldsymbol{\eta}}^\top d \varPi_{\boldsymbol{\eta}} \in \mathbb{R}^{m \times m}$.

Remark 4.2.16. To keep things simple we do not elaborate on how to check $(\breve{\mathcal{L}}_0)$, $(\breve{\mathcal{E}}\mathcal{D}_1)$, $(\breve{\mathcal{E}}\mathcal{D}_0)$ but refer to Lemma 4.2.1.

Noting that $\mathcal{L}(\widetilde{\boldsymbol{\upsilon}}, \boldsymbol{\upsilon}^*) \ge 0$ and $\mathcal{L}(\widetilde{\boldsymbol{\upsilon}}_{\boldsymbol{\theta}^*}, \boldsymbol{\upsilon}^*) \ge 0$ Theorem 3.3.2 yields that

$$I\!\!P\left(\widetilde{\boldsymbol{\upsilon}}, \widetilde{\boldsymbol{\upsilon}}_{\boldsymbol{\theta}^*} \in \varUpsilon_\circ(\mathtt{r}_0)\right) \ge 1 - \mathrm{e}^{-\mathtt{x}}, \quad \text{with} \quad \mathtt{r}_0(\mathtt{x}) = 6\frac{\nu_0^*}{\mathtt{b}^*}\sqrt{2p^* + \mathtt{x}}.$$

Theorem 4.2.2 applies with $\mathcal{D}^2 = n\nabla^2 I\!\!E\ell(\boldsymbol{\upsilon}^*)$ and $\boldsymbol{\upsilon}^\circ = \boldsymbol{\upsilon}^*$. We immediately obtain the following result.

Corollary 4.2.5. *Let* $Y_1, \ldots, Y_n$ *be i.i.d. and let the conditions* (ℓ_0), (ed_1), (ed_0), $(\ell_{\mathtt{r}})$ *and* (ι) *be met. Assume that* $\mathtt{r}_0(\mathtt{x}) = 6\frac{\nu_0^*}{\mathtt{b}^*}\sqrt{2p^* + \mathtt{x}} \le \mathtt{r}^*$. *Then we get the Fisher and Wilks results of Theorem 4.2.2 for* $\mathtt{x} \ll \sqrt{n}\mathtt{g}^*$ *with*

$$\breve{\diamondsuit}(\mathtt{r}_0, \mathtt{x}) \le \frac{36\nu_0^*}{\sqrt{n}\mathtt{b}^*}\left(\frac{4}{(1-\nu^2)^2}\frac{\delta^*}{c_d^3}\frac{\nu_0}{\mathtt{b}^*}(\mathtt{x} + 2p^*) + \nu_0\mathfrak{z}_1(\mathtt{x}, 2p^* + 2p)\sqrt{\mathtt{x} + 2p^*}\right).$$

Remark 4.2.17. The definition of $\mathfrak{z}_1(\mathtt{x}, 2p^* + 2p)$ in (3.5.6) implies for moderate values of $\mathtt{x} > 0$ that

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

with some fixed constant $\mathtt{C}_{\diamondsuit}$. The Fisher result (4.2.9) is meaningful if $\breve{\diamondsuit}(\mathtt{r}_0, \mathtt{x})$ is small yielding the constraint $p^* \ll n^{1/2}$. If the target dimension p is fixed, the same condition is sufficient for the Wilks expansion in (4.2.10). However, if the target dimension p is of order p^*, the constraint for the Wilks theorem becomes $p^* = o(n^{1/3})$. See Section 4.2.5 for an example that shows, that this difference actually occurs in certain examples.

4.2.5 Impact of the full dimension

This section discusses the effects of a full dimension p^* that grows with the sample size n. The results of Theorem 4.2.2 refined by Proposition 4.2.3 are accurate if the *parametric uniform spread* $\diamondsuit(\mathtt{r},\mathtt{x})$ in (4.2.11) fulfills $\diamondsuit(\mathtt{r}_0,\mathtt{x}) \le \mathfrak{z}(\mathtt{x}, I\!\!B)$ and $\breve{\diamondsuit}(\mathtt{r}_1,\mathtt{x})$ is small, with $\mathtt{r}_1 = 2\mathfrak{z}(\mathtt{x}, I\!\!B)$. Usually $\mathfrak{z}(\mathtt{x}, I\!\!B) \le \mathtt{C}\sqrt{\mathtt{x}+p^*}$ which means that

$$\breve{\diamondsuit}(\mathtt{r}_1,\mathtt{x}) \asymp \breve{\delta}(\mathtt{r}_1)\mathtt{r}_1 + \breve{\omega}\mathtt{r}_1^2 \quad \text{which is small for } \mathtt{r}_1^2 \asymp p^*. \tag{4.2.14}$$

The critical size of p^* then depends on the exact bounds for $\breve{\delta}(\cdot), \breve{\omega}$. If $\breve{\delta}(\mathtt{r})/\mathtt{r} \asymp \breve{\omega} \asymp 1/\sqrt{n}$ (as in Corrolary 4.2.5) the condition (4.2.14) reads " $\breve{\diamondsuit}(\mathtt{r}_1,\mathtt{x}) \asymp p^*/\sqrt{n}$ is small". This means that one needs that "${p^*}^2/n$ is small" to obtain an accurate non-asymptotic version of the Wilks phenomenon and the Fisher Theorem. Similar conclusions were obtained by Portnoy in several papers on growing dimension in generalized linear models and for natural exponential families, see e.g. [43, 44, 45] as well as by Mammen in [35, 36, 37]. Improvements of the critical relation such as $p^* = o(n^{3/2})$ in [44] are rooted heavily in the structure of the particular model. For instance [44] is limited to linear or generalized linear regression with independent observations, which is exploited extensively in the derivation.

Thus the typical sufficient dimensional asymptotic is $p^* = o(n^{1/2})$. In [46] these are derived for natural exponential families with correct specification. In this setting [46] uses the linearity in the stochastic component of the loglikelihood and the correct specification to obtain both Fisher and Wilks phenomenon when ${p^*}^2/n \to 0$. Our results apply in a rather general situation and deliver some useful information even in the case when the model is misspecified and when the stochastic component of $\nabla\mathcal{L}(\cdot)$ is nonlinear.

Remark 4.2.18. Note that in Corollary 3.2.2 if $\breve{\delta}(\mathtt{r})/\mathtt{r} \asymp \breve{\omega} \asymp 1/\sqrt{n}$ the condition on the ratio of dimension to sample size is that ${p^*}^3/n$ is small. One reason for the improvement in this chapter - the results obviously apply for the case $\widetilde{\boldsymbol{\theta}} = \widetilde{\boldsymbol{\upsilon}}$, i.e. the full ME - is that we do not first carry out a local quadratic approximation of $\mathcal{L}(\boldsymbol{\upsilon},\boldsymbol{\upsilon}^*)$ and then bound the displacement of $\widetilde{\boldsymbol{\upsilon}}$. To the contrary we take the first derivative and then carry out a local linear approximation of $\breve{\nabla}\mathcal{L}(\boldsymbol{\upsilon}) - \breve{\nabla}\mathcal{L}(\boldsymbol{\upsilon}) \in \mathbb{R}^p$. This means that we do not pay in the accuracy with uniform bounds for the term in (1.0.8), but only with uniform deviation bounds for its derivative, i.e. for

$$\left\| \breve{D}^{-1}\left(\breve{\nabla}\mathcal{L}(\widetilde{\boldsymbol{\upsilon}}) - \breve{\nabla}\mathcal{L}(\boldsymbol{\upsilon}^*)\right) + \breve{D}(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) \right\|.$$

This is motivated by the observation that the accuracy of the Fisher expansion depends on the displacement of the ME through the second order terms in a Taylor expansion of the gradient $\nabla\mathcal{L}(\boldsymbol{\upsilon})$ and not on the third order terms in a Taylor expansion of $\mathcal{L}(\boldsymbol{\upsilon})$. For the Wilks expansion (4.2.10) this is different as here indeed the Taylor expansion of $\mathcal{L}(\boldsymbol{\upsilon})$ is the key. But a trick from [40] allows us to only pay with an additional factor proportional to the square root of the target dimension $p \in \mathbb{N}$ and not to the full dimension $p^* \in \mathbb{N}$. The key observation is that one can carry out a kind of Taylor expansion for the profile functional $\max_{\boldsymbol{\eta}} \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\eta})$ in a neighborhood of $\boldsymbol{\theta}^* \in \mathbb{R}^p$. For details see Lemma 4.A.2.

Critical dimension

To see that under the conditions of Section 4.2.1 with $\delta(\mathtt{r})/\mathtt{r} \asymp \omega \asymp 1/\sqrt{n}$ we can not do better than $p^* \ll n^{1/2}$ we present the following example. We write $p^* = p_n$. Consider the single observation model

$$\begin{aligned}
\boldsymbol{Y} &= f(\boldsymbol{\upsilon}) + \boldsymbol{\varepsilon},\\
f(\boldsymbol{\upsilon}) &= f(\theta, \boldsymbol{\eta}) \overset{\text{def}}{=} \begin{pmatrix} \theta \\ \boldsymbol{\eta}_1 \\ \vdots \\ \boldsymbol{\eta}_{p_n-1} \end{pmatrix} + \begin{pmatrix} \|\boldsymbol{\eta}\|^2 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \in \mathbb{R}^{p_n},
\end{aligned}$$

with $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \frac{1}{n} I_{p_n})$ and $\boldsymbol{\upsilon} = (\theta, \boldsymbol{\eta}) \in \mathbb{R} \times \mathbb{R}^{p_n - 1}$. This model is equivalent to the i.i.d. observations in the same model with the errors $\boldsymbol{\varepsilon}_i \sim \mathcal{N}(0, I_{p_n})$. Assume that the parameter of interest is $\theta \in \mathbb{R}$ and that the true point satisfies $\boldsymbol{\upsilon}^* = 0 \in \mathbb{R}^{p^*}$.

Proposition 4.2.6. *Under $p_n/\sqrt{n} \to 0$, the Fisher expansion is accurate and the profile MLE asymptotically standard normal. If $p_n/\sqrt{n} \not\to 0$ the profile MLE in the above model is not root-n consistent. For $\sqrt{n} = o(p_n)$ the root-n bias tends to infinity almost surely. Finally, the Wilks phenomenon occurs if and only if $p_n = o(\sqrt{n})$.*

Remark 4.2.19. The above example can also be used to illustrate the difference between a finite sample approach and using asymptotic normality for the construction of confidence sets. For fixed dimension the profile MLE is asymptotically standard normal, i.e. with $q_\alpha > 0$ denoting the α-level quantile of a chi-square distribution with one degree of freedom

$$I\!\!P\left(\theta^* \in \left\{|\widetilde{\theta} - \theta|^2 \le q_\alpha/n\right\}\right) \to \alpha. \tag{4.2.15}$$

But the proof of Proposition 4.2.6 gives

$$|\widetilde{\theta} - \theta^*| = \left|\varepsilon_\theta - \|\boldsymbol{\varepsilon}_{\boldsymbol{\eta}}\|^2\right|,$$

where $n\|\boldsymbol{\varepsilon}_{\boldsymbol{\eta}}\|^2 \sim \chi^2_{p_n-1}$ and $\varepsilon_\theta \sim \mathcal{N}(0,1/n)$. It is known that the ratio of the median of a chi-square distribution and its degrees of freedom converges to 1 if the degrees of freedom tend to infinity. This means that for any $0<\epsilon<1$ the set

$$C \overset{\text{def}}{=} \left\{n\|\boldsymbol{\varepsilon}_{\boldsymbol{\eta}}\|^2 \geq (1-\epsilon)p_n\right\},$$

is of probability greater than $1/2$ for p_n large enough. Let $f_{\chi^2_{p_n-1}} : [0,\infty)\to\mathbb{R}$ denote the Lebesgue density of a $\chi^2_{p_n-1}$ random variable. We can use the independence of $\|\boldsymbol{\varepsilon}_{\boldsymbol{\eta}}\|$ and ε_θ and Fubini's Theorem to estimate

$$\begin{aligned}
I\!\!P\left(\theta^* \in \left\{|\widetilde{\theta}-\theta|^2 \leq q_\alpha/n\right\}\right) &= \int_0^\infty I\!\!P\left(|\varepsilon_\theta - z/n|^2 \leq q_\alpha/n\right) f_{\chi^2_m}(z)dz \\
&= \int_0^\infty \left[\varPhi\left(\frac{z}{\sqrt{n}}+\sqrt{q_\alpha}\right)-\varPhi\left(\frac{z}{\sqrt{n}}-\sqrt{q_\alpha}\right)\right] f_{\chi^2_m}(z)dz \\
&< \frac{1}{2}\left[\alpha+\varPhi\left((1-\epsilon)\frac{p_n}{\sqrt{n}}+\sqrt{q_\alpha}\right)-\varPhi\left((1-\epsilon)\frac{p_n}{\sqrt{n}}-\sqrt{q_\alpha}\right)\right],
\end{aligned}$$

where $\varPhi:\mathbb{R}\to[0,1]$ denotes the distribution function of a standard normal random variable. If $p_n/\sqrt{n}$ is significantly larger than 0, the value

$$\varPhi\left(\frac{(1-\epsilon)p_n}{\sqrt{n}}+\sqrt{q_\alpha}\right)-\varPhi\left(\frac{(1-\epsilon)p_n}{\sqrt{n}}-\sqrt{q_\alpha}\right),$$

is distinctively smaller α. For example for $\alpha=0.95$ and $(1-\epsilon)p_n/\sqrt{n} = 11/12$ we get

$$I\!\!P\left(\theta^* \in \left\{|\widetilde{\theta}-\theta|^2 \leq q_{0.95}/n\right\}\right) < 0.9.$$

In other words the asymptotic confidence statement in (4.2.15) is very inaccurate in the finite sample case because the error term in the local linear approximation is not addressed. In this way the full dimension has an impact on the behavior of the estimator. Our results in Theorem 4.2.2 quantify the size of these terms for a large set of models and give a guideline for how to correct confidence sets to avoid this effect. The price are more conservative sets, but their coverage property is ensured.

Remark 4.2.20. There is an interesting connection between the condition $p^*/\sqrt{n}\to 0$ and the general theory on semiparametric M-estimators. In the common asymptotic approach to semiparametric M-estimators one assumes a priori consistency of the estimator $\widetilde{\boldsymbol{\upsilon}} = (\widetilde{\boldsymbol{\theta}},\widetilde{\boldsymbol{\eta}})$. More precisely, in case the functional $\breve{\nabla}\mathcal{L}(\cdot)$ is smooth enough, one assumes that $\|\widetilde{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\| = o_{I\!\!P}(1)$ and $\|\widetilde{\boldsymbol{\eta}}-\boldsymbol{\eta}^*\| = O_{I\!\!P}(n^{-1/4})$, see Section 2.1. On the other hand the results of Theorem 4.2.2 are accurate if $\breve{\diamondsuit}(\mathtt{r}_0,\mathtt{x})$ is small. As explained above this means in the i.i.d setting that $\check{\Diamond}(\mathtt{r}_0,\mathtt{x}) = o(1)$. Neglecting the contribution of $\|\widetilde{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\|$ to $\mathtt{r}_0 = O(\sqrt{n}\|\widetilde{\boldsymbol{\upsilon}}-\boldsymbol{\upsilon}^*\|)$ this can be ensured if

$$\check{\Diamond}(\mathtt{r}_0,\mathtt{x}) \leq \mathtt{C}(p^*+\mathtt{r}_0^2)/\sqrt{n} \leq o(1) + \mathtt{C}\sqrt{n}\|\widetilde{\boldsymbol{\eta}}-\boldsymbol{\eta}^*\|^2 \to 0,$$

i.e. if $\|\widetilde{\boldsymbol{\eta}}-\boldsymbol{\eta}^*\| = o(n^{-1/4})$. But consider the radius $\mathtt{r}_1 > 0$ from Proposition 4.2.3. It is of order $\sqrt{p^*+m}$ if $\Diamond(\mathtt{r}_0) = O(\sqrt{p^*+\mathtt{x}})$. In that case in the i.i.d. setting the constraint on the a priori deviation bound becomes $\Diamond(\mathtt{r}_0,\mathtt{x}) = O(\sqrt{p^*+\mathtt{x}})$. This can be ensured if

$$\Diamond(\mathtt{r}_0,\mathtt{x}) \leq \mathtt{C}(p^*+\mathtt{r}_0^2)/\sqrt{n} \leq o(1) + \mathtt{C}\sqrt{n}\|\widetilde{\boldsymbol{\eta}}-\boldsymbol{\eta}^*\|^2 = O(\sqrt{p^*+\mathtt{x}}).$$

Consequently, if $p^*+\mathtt{x} = o(\sqrt{n})$ we only need $\|\widetilde{\boldsymbol{\eta}}-\boldsymbol{\eta}^*\| = o(n^{-1/8})$, which is a considerably weaker constraint. Using the second part of Proposition 4.2.3, the constraint on $\|\widetilde{\boldsymbol{\eta}}-\boldsymbol{\eta}^*\|$ becomes

$$\mathtt{C}\mathtt{r}_0/\sqrt{n} \leq \mathtt{C}(\|\widetilde{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\| + \|\widetilde{\boldsymbol{\eta}}-\boldsymbol{\eta}^*\|) \to 0,$$

i.e. we only need consistency of $\widetilde{\boldsymbol{\upsilon}}$. But note that these bounds only concern the finite dimensional case. In the infinite dimensional setting treated in Section 4.3.2 we have to impose conditions that ensure that the bias induced by the sieve approach is small enough. Section 4.3.3 serves such conditions for the Hilbert space setting. One of these conditions reads that $\|\mathcal{H}(\boldsymbol{\eta}^*-\varPi_m\boldsymbol{\eta}^*)\|^2 \leq \mathtt{C}m$, i.e. the true nuisance component $\boldsymbol{\eta}^* \in \mathcal{X}$ is well approximated by its projection onto the span of the first $m \in \mathbb{N}$ basis elements $(\boldsymbol{e}_k) \subset \mathcal{X}$. If we represent with some $\alpha > 0$

$$\boldsymbol{\eta}^* = \sum_{k=1}^{\infty}\eta_k^*\boldsymbol{e}_k, \qquad \sum_{k=1}^{\infty}\eta_k^{*2}k^{2\alpha} < \infty,$$

and if $\mathcal{H} = I_{\mathcal{X}}$ we obtain from $\|\mathcal{H}(\boldsymbol{\eta}^*-\varPi_m\boldsymbol{\eta}^*)\|^2 \leq \mathtt{C}m$ the constraint $n \leq m^{2\alpha+1}$, which means that we need $\alpha > 1/2$ if $m = o(n^{1/2})$. On the other hand in the setting of one dimensional nonparametric regression $\alpha > 1/2$ means that $\boldsymbol{\eta}^* \in \mathcal{X}$ is nonparametrically estimable with rate $o(n^{-1/4})$.

Critical smoothness

Here we address the necessary smoothness to ensure that the condition that ${p^*}^2/n \ll 1$ suffices to ensure that the Fisher expansion is accurate. We show that the slightly weaker version $(\mathcal{L}_0')$ from 3.2.1 of $(\mathcal{L}_0)$ already allows to find examples that satisfy all conditions of Section 4.2.1 but for which the critical ratio is ${p^*}^3/n \to 0$. Namely, we present an example in which the behavior of the profile ME $\widetilde{\boldsymbol{\theta}}$ heavily depends on the value $\beta_n = \sqrt{p_n^3/n} \geq \beta > 0$. If $\beta_n \to 0$, then we can prove asymptotic efficiency of $\widetilde{\boldsymbol{\theta}}$. On the other hand if $\beta_n \geq \beta > 0$ we can show that the ME $\widetilde{\boldsymbol{\theta}}$ is not anymore root-n consistent.

Assume that $p_n/\sqrt{n} \to 0$. Let a random vector $\mathbf{X} \in \mathbb{R}^{p_n}$ follow $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\upsilon}^*, n^{-1} I_{p_n})$. Take for simplicity $\boldsymbol{\upsilon}^* = 0$ and let $I\!\!P = I\!\!P_0$ denote the distribution of $\mathbf{X}$. Introduce a special set $\mathcal{S} \subset \mathbb{R}^{p_n}$ with

$$\begin{aligned} \mathcal{S} &\stackrel{\text{def}}{=} \Bigl\{ \boldsymbol{\upsilon} = (\upsilon_1, \ldots, \upsilon_{p_n}) : \upsilon_1 = \frac{z}{2}\sqrt{\beta_n/n},\, z \in \mathbb{Z} \Bigr\} \\ &\quad \cap \Upsilon_\circ \left(\sqrt{2p_n/n} + \frac{1}{2}\sqrt{\beta_n/n} \right). \end{aligned} \tag{4.2.16}$$

We denote by $\mathcal{S}_\delta$ its δ-vicinity:

$$\mathcal{S}_\delta \stackrel{\text{def}}{=} \{\boldsymbol{\upsilon} : d(\boldsymbol{\upsilon}, \mathcal{S}) < \delta\},$$

where $d(\boldsymbol{\upsilon}, \mathcal{S})$ is the Euclidean distance from the point $\boldsymbol{\upsilon}$ to the set $\mathcal{S}$. Also $\mathcal{S}_\delta^c$ stands for the complement of $\mathcal{S}_\delta$. Below we fix $\delta = 1/n$. Consider a special parametric quasi log-likelihood ratio $\mathcal{L}(\boldsymbol{\upsilon}, 0)$ defined as

$$\mathcal{L}(\boldsymbol{\upsilon}, 0) = n\mathbf{X}^\top \boldsymbol{\upsilon} - n\|\boldsymbol{\upsilon}\|^2/2 + nf(\boldsymbol{\upsilon})\|\boldsymbol{\upsilon}\|^3.$$

Here $f : \mathbb{R} \mapsto \mathbb{R}$ is a smooth function with

$$f(\boldsymbol{\upsilon}) = \begin{cases} 1 & \boldsymbol{\upsilon} \in \mathcal{S}, \\ 0 & \boldsymbol{\upsilon} \in \mathcal{S}_\delta^c. \end{cases}$$

Below we consider the problem of estimating the first component $\theta \stackrel{\text{def}}{=} \upsilon_1 \in \mathbb{R}$. Since by assumption $p_n/\sqrt{n} \to 0$ it holds for n large enough and for any $\boldsymbol{\upsilon}$ with $\|\boldsymbol{\upsilon}\|^2 \leq 4p_n/n + \beta_n/n$ that $n\|\boldsymbol{\upsilon}\|^2/2 \geq nf(\boldsymbol{\upsilon})\|\boldsymbol{\upsilon}\|^3$ and thus

$$\operatorname*{argmax}_{\boldsymbol{\upsilon}} I\!\!E\mathcal{L}(\boldsymbol{\upsilon}) = \operatorname*{argmin}_{\boldsymbol{\upsilon}} \bigl\{ n\|\boldsymbol{\upsilon}\|^2/2 - nf(\boldsymbol{\upsilon})\|\boldsymbol{\upsilon}\|^3 \bigr\} = 0.$$

It is easy to see that all conditions from Section 4.2.1 except $(\boldsymbol{\mathcal{L}_0})$ are satisfied with $\omega \cong 1/\sqrt{n}$ and

$$\mathcal{D}^2 = \mathcal{V}^2 = nI_{p_n}, \quad \Upsilon_\circ(\mathtt{r}) = \{\|\boldsymbol{\upsilon}\| \leq \mathtt{r}/\sqrt{n}\}.$$

But clearly $(\mathcal{L}_0)'$ is met with $\delta(\mathtt{r}) = \mathtt{r}/\sqrt{n}$. It is straightforward to see that

$$\breve{D}_0 = \sqrt{n}, \qquad \breve{\nabla}(\mathcal{L} - I\!\!E\mathcal{L}) = \nabla_{\boldsymbol{\theta}}(\mathcal{L} - I\!\!E\mathcal{L}) = nX_1, \text{ and } \breve{\boldsymbol{\xi}} = \sqrt{n}X_1.$$

The next result shows that in this example the critical ratio reads $\beta_n = \sqrt{p_n^3/n}$, i.e. iff it is not small, the profile ME $\widetilde{\theta}$ is not root-n consistent.

Proposition 4.2.7. *If* $\beta_n^2 = p_n^3/n \to 0$ *then*

$$\|\breve{D}_0(\widetilde{\theta} - \theta^*) - \breve{\boldsymbol{\xi}}\| = \sqrt{n}|\widetilde{v}_1 - X_1| \to 0.$$

Suppose that $\beta_n \to (6c)^2$ *for some* $c > 0$. *Let also* n *be large enough to ensure*

$$\frac{2^{1/3} - 1}{2^{1/6}} \sqrt{p_n/n} \geq \frac{1}{2} (p_n/n)^{3/4}.$$

There exists a positive $\alpha > 0$ *such that it holds with a probability exceeding* α

$$\|\breve{D}_0(\widetilde{\theta} - \theta^*) - \breve{\boldsymbol{\xi}}\| \geq \frac{1}{6}\beta_n^{1/2} - \frac{1}{\sqrt{n}} \geq c - o_n(1).$$

If $\beta_n \to \infty$, *then*

$$\|\breve{D}_0(\widetilde{\theta} - \theta^*) - \breve{\boldsymbol{\xi}}\| \xrightarrow{I\!P} +\infty,$$

where $\xrightarrow{I\!P}$ *means convergence in probability.*

In short: we have shown that - everything else left unchanged - a smoothness condition of the kind of $(\boldsymbol{\mathcal{L}_0})$, i.e. qualified smoothness of second derivatives, is necessary to ensure that "$p^*/\sqrt{n}$ is small" suffices to get accurate results in Theorem 4.2.2 for $\delta(\mathtt{r})/\mathtt{r} \approx \omega \approx 1/\sqrt{n}$.

Difference between Wilks and Fisher

This section discusses the issue of *critical dimensions* if the target dimension $p = cp^*$ for some $c > 0$. We again write $p^* = p_n$. In this case Theorem 4.2.2 - assuming that $\delta(\mathtt{r})/\mathtt{r} \cong \omega \cong 1/\sqrt{n}$ - requires that $p_n = o(n^{1/3})$ or $p_n = o(n^{1/2})$ to obtain nonasymptotic versions of the Wilks phenomenon and the Fisher Theorem respectively. Here we show that this difference actually occurs on the class of models satisfying the conditions of Section 4.2.1. We present an example that shows critical behavior in the following sense. When $p_n^3/n \nrightarrow 0$ we find for each $n \in \mathbb{N}$ large enough a set $\mathcal{A} \subset \Omega$ of positive probability on which the profile log likelihood ratio does not converge to a chi-square random variable. In accordance with the results of Theorem 4.2.2 the estimator is efficient if $p_n^2/n \to 0$ and the Wilks phenomenon occurs if $p_n^3/n \to 0$.

Assume $p_n = 2m$ and take as target $\boldsymbol{\theta} := \Pi_1 \boldsymbol{\upsilon} \in \mathbb{R}^m$, where $\Pi_1 : \mathbb{R}^{p_n} \to \mathbb{R}^m$ denotes the orthogonal projection onto the first $m \in \mathbb{N}$ components. Assume further that $p_n^2/n \to 0$. We use a missspecified model, i.e. we take standard normal observations on $\mathbb{R}^{p_n}$ but assume that the ME is derived from the correct loglikelihood function altered by an additional term. Consider

$$\mathcal{L}(\boldsymbol{v}) = n\boldsymbol{Y}^\top \boldsymbol{v} - n\|\boldsymbol{v}\|^2/2 + f(\boldsymbol{v})n\|\boldsymbol{v}\|^3/3,$$

where

$$\boldsymbol{Y} \sim N\left(0, \frac{1}{n} I_{p_n}\right),$$

and where $f : \mathbb{R}^{p_n} \to \mathbb{R}$ is some smooth function with - for some $L > 0$ -

$$f(\boldsymbol{v}) = \begin{cases} 1, & \mathcal{S} := \left\{\|\mathit{\Pi}_1 \boldsymbol{v}\| \geq \frac{2}{L}\sqrt{\frac{p_n}{n}}\right\} \cap B_{2\sqrt{\frac{p_n}{n}}}(0), \\ 0, & \mathit{otherwise.} \end{cases}$$

More precisely we set for any $\boldsymbol{v}^\circ \in \mathbb{R}^{p_n}$

$$f(\boldsymbol{v}^\circ) = \varphi_{\{\|\mathit{\Pi}_1 \boldsymbol{v}\| \geq \frac{2}{L}\sqrt{\frac{p_n}{n}}\}}(\boldsymbol{v}^\circ) 1_{B_{2\sqrt{\frac{p_n}{n}}}(0)}(\boldsymbol{v}^\circ). \tag{4.2.17}$$

The "smooth" factor is defined as

$$\varphi_{\{\|\mathit{\Pi}_1 \boldsymbol{v}\| \geq \frac{2}{L}\sqrt{\frac{p_n}{n}}\}}(\boldsymbol{v}^\circ) = \int_{\mathbb{R}} 1_{\{\|\mathit{\Pi}_1 \boldsymbol{v}\| \leq \frac{1}{L}\sqrt{\frac{p_n}{n}}\}}(\boldsymbol{v}_1) K_{\frac{1}{L}\sqrt{\frac{p_n}{n}}}(\boldsymbol{v}_1^\circ - \boldsymbol{v}_1) d\boldsymbol{v}_1,$$

where K is a smooth kernel with support on $[-1, 1]$ and

$$K_h(x) := \frac{1}{h} K\left(\frac{x}{h}\right).$$

Proposition 4.2.8. *In the above model the conditions of Section 4.2.1 are satisfied yielding* $\diamondsuit(\mathtt{r}_0, \mathtt{x}) = o(p_n/\sqrt{n})$*. The Fisher theorem holds true if* $p_n^2/n \to 0$*. Furthermore the Wilks phenomenon occurs iff* $p_n^3/n \to 0$*.*

4.3 Infinite dimensional nuisance parameter

This section discusses how the approach can be extended to the infinite dimensional case. First the basic idea of projecting the infinite dimensional problem down to a finite dimensional one is explained using a suggestion by [22] namely the *sieve* profile ME. The particular type of sieve we are using was also studied in [50] and we try to relate our results to that paper (see Section 2.2.3). We prove under bias constraints that the projected *sieve* estimator is nearly normal and efficient. To avoid further technical distractions (or obstacles) we present the case of a separable Hilbert space.

4.3.1 Sieve approach

To make this chapter self contained we repeat how to construct the sieve estimator and how to use the Hilbert space in order to reduce the problem to parameters $\boldsymbol{\upsilon} \in l^2 \overset{\text{def}}{=} \{(x_j)_{j\in\mathbb{N}} \subset \mathbb{R}, \sum_{k=j}^{\infty} x_j^2 < \infty\}$. Consider the $(\boldsymbol{\theta}, \boldsymbol{\eta})$-setup with $\boldsymbol{\theta} \in \varTheta \subseteq \mathbb{R}^p$ and $\boldsymbol{\eta} \in \mathcal{X}$, where $\mathcal{X}$ is an infinite dimensional separable Hilbert space. As always the target parameter $\boldsymbol{\theta}^*$ is defined as

$$\boldsymbol{\theta}^* = \varPi_{\boldsymbol{\theta}} \operatorname*{argmax}_{\boldsymbol{\upsilon}\in\varUpsilon} I\!\!E\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\eta}). \tag{4.3.1}$$

The Hilbert space $\mathcal{X}$ is assumed to be separable such that it possesses a countable orthonormal basis $\{\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots\} \subset \mathcal{X}$. Any vector $\boldsymbol{\eta} \in \mathcal{X}$ admits a unique decomposition of the form

$$\boldsymbol{\eta} = \sum_{k=1}^{\infty} \eta_k \boldsymbol{e}_k,$$

where $\eta_j = \langle \boldsymbol{\eta}, \boldsymbol{e}_j \rangle$ is the usual Fourier coefficient. In the *sieve* approach one assumes that for any $m \in \mathbb{N}$ a finite set $\boldsymbol{e}_1, \ldots, \boldsymbol{e}_m$ of elements in $\mathcal{X}$ is fixed and the vector $\boldsymbol{\eta}$ can be approximated by a finite linear combination $\boldsymbol{\eta}_m$ of the basis functions $(\boldsymbol{e}_k)_{k\in\mathbb{N}}$:

$$\left\| \boldsymbol{\eta}^* - \sum_{k=1}^{m} \eta_k^* \boldsymbol{e}_k \right\| = \tau(m) \to 0, \text{ as } m \to \infty.$$

By abuse of notation we denote $\boldsymbol{\upsilon} = (\boldsymbol{\theta}, \boldsymbol{\eta}) \in \mathbb{R}^p \times l^2$ and modify the parameter set such that $\varUpsilon \subseteq \mathbb{R}^p \times l^2$ via identifying each $\boldsymbol{\eta} \in \mathcal{X}$ with its Fourier coefficients. Redefine $\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\eta})$ such that it is a function of the Fourier coefficients of the nuisance component.

$$\begin{aligned} \mathcal{L}(\boldsymbol{\upsilon}) &\overset{\text{def}}{=} \mathcal{L}\left(\boldsymbol{\theta}, \sum_{j=1}^{\infty} \eta_j \boldsymbol{e}_j\right) \\ \boldsymbol{\upsilon}^* &\overset{\text{def}}{=} \operatorname*{argmax}_{(\boldsymbol{\theta},\boldsymbol{\eta})\in l^2} I\!\!E\left[\mathcal{L}\left(\boldsymbol{\theta}, \sum_{k=1}^{\infty} \eta_j \boldsymbol{e}_j\right)\right], \end{aligned}$$

and define the m-dimensional sieve approximation $\mathcal{L}_m(\boldsymbol{\upsilon})$ of $\mathcal{L}(\boldsymbol{\upsilon})$ by

$$\begin{aligned} \mathcal{L}_m(\boldsymbol{\theta}, \boldsymbol{\eta}) &\overset{\text{def}}{=} \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\eta}_m(\boldsymbol{\eta})), \\ (\boldsymbol{\theta}, \boldsymbol{\eta}) \in \varUpsilon_m &\overset{\text{def}}{=} \{\boldsymbol{\upsilon} = (\boldsymbol{\theta}, \boldsymbol{\eta}) \in \mathbb{R}^{p^*} : (\boldsymbol{\theta}, \boldsymbol{\eta}_m) \in \varUpsilon\}. \end{aligned}$$

The corresponding sieve profile estimator $\widetilde{\boldsymbol{\theta}}_m$ and its target $\boldsymbol{\theta}^*_m$ for this parametric m-submodel are defined in the usual way:

$$\widetilde{\boldsymbol{\theta}}_m \overset{\text{def}}{=} \Pi_{\boldsymbol{\theta}} \widetilde{\boldsymbol{\upsilon}}_m \overset{\text{def}}{=} \Pi_0 \operatorname*{argmax}_{\boldsymbol{\upsilon} \in \Upsilon_m} \mathcal{L}_m(\boldsymbol{\theta}, \boldsymbol{\eta}), \tag{4.3.2}$$

$$\boldsymbol{\theta}^*_m \overset{\text{def}}{=} \Pi_{\boldsymbol{\theta}} \boldsymbol{\upsilon}^*_m \overset{\text{def}}{=} \Pi_{\boldsymbol{\theta}} \operatorname*{argmax}_{\boldsymbol{\upsilon} \in \Upsilon_m} I\!E \mathcal{L}_m(\boldsymbol{\theta}, \boldsymbol{\eta}).$$

The question we are interested in can be formulated as follows: is $\widetilde{\boldsymbol{\theta}}_m$ a good (efficient) estimator of $\boldsymbol{\theta}^*$ from (4.3.1) under a proper choice of m?

4.3.2 Bias constraints and efficiency

The parametric results obtained in Section 4.2 claim that $\widetilde{\boldsymbol{\theta}}_m \in \mathbb{R}^p$ is a good estimator for $\boldsymbol{\theta}^*_m \in \mathbb{R}^p$ if the spread $\breve{\diamondsuit}(\mathtt{r}_0, \mathtt{x}) > 0$ is small. More precisely, we have the following: define for fixed $\mathtt{x} > 0$ the value $\mathtt{r}_0 > 0$ by

$$\mathtt{r}_0(\mathtt{x}) \overset{\text{def}}{=} \inf_{\mathtt{r} \ge 0} \left\{ I\!P\left\{ \widetilde{\boldsymbol{\upsilon}}_m, \widetilde{\boldsymbol{\upsilon}}_{\boldsymbol{\theta}^*_m, m} \in \Upsilon_{0,m}(\mathtt{r}) \right\} \ge 1 - \mathrm{e}^{-\mathtt{x}} \right\},$$

$$\text{and set } \Upsilon_{0,m}(\mathtt{r}) \overset{\text{def}}{=} \left\{ \boldsymbol{\upsilon} \in \Upsilon_m,\ \| \mathcal{D}_m(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^*_m) \| \le \mathtt{r} \right\},$$

where $\boldsymbol{\upsilon}^*_m = (\boldsymbol{\theta}^*_m, \boldsymbol{\eta}^*_m) = \operatorname{argmax}_{\boldsymbol{\upsilon}} I\!E \mathcal{L}_m(\boldsymbol{\upsilon})$ and

$$\widetilde{\boldsymbol{\upsilon}}_{\boldsymbol{\theta}, m} \overset{\text{def}}{=} \operatorname*{argmax}_{\substack{\boldsymbol{\upsilon} \in \Upsilon_m \\ \Pi_0 \boldsymbol{\upsilon} = \boldsymbol{\theta}}} \mathcal{L}_m(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^*).$$

Furthermore define the matrix $\breve{D}^2_m$ as

$$\breve{D}^2_m(\boldsymbol{\upsilon}^*_m) \overset{\text{def}}{=} \left(\Pi_{\boldsymbol{\theta}} \mathcal{D}_m^{-2} \Pi_{\boldsymbol{\theta}}^\top \right)^{-1} \in \mathbb{R}^{p \times p}, \ \ \mathcal{D}^2_m \overset{\text{def}}{=} \nabla^2_{p+m} I\!E[\mathcal{L}(\boldsymbol{\upsilon}^*_m)] \in \mathbb{R}^{p^* \times p^*},$$

i.e. the derivatives of $I\!E[\mathcal{L}]$ are only taken with respect to the first $p+m \in \mathbb{N}$ coordinates of $\boldsymbol{\upsilon} \in l^2$ and the Hessian is evaluated in $\boldsymbol{\upsilon}^*_m \in \mathbb{R}^{p^*}$. Applying Theorem 4.2.2 to $\widetilde{\boldsymbol{\theta}}_m$ in (4.3.2) we find that with probability greater than $1 - 2\mathrm{e}^{-\mathtt{x}}$

$$\| \breve{D}_m \big(\widetilde{\boldsymbol{\theta}}_m - \boldsymbol{\theta}^*_m \big) - \breve{\boldsymbol{\xi}}_m(\boldsymbol{\upsilon}^*_m) \| \le \breve{\diamondsuit}(\mathtt{r}_0, \mathtt{x}). \tag{4.3.3}$$

The result (4.3.3) involves two kinds of bias. The first concerns the difference $\boldsymbol{\theta}^*_m - \boldsymbol{\theta}^* \in \mathbb{R}^p$ and the second arises with the difference between $\breve{D}_m \in \mathbb{R}^{p \times p}$ and $\breve{D} \in \mathbb{R}^{p \times p}$ where

$$\breve{D}^2 \overset{\text{def}}{=} \left(\Pi_{\boldsymbol{\theta}} \nabla^2 I\!E[\mathcal{L}(\boldsymbol{\upsilon}^*)]^{-1} \Pi_{\boldsymbol{\theta}}^\top \right)^{-1} \in \mathbb{R}^{p \times p}.$$

This means that in the case of $\breve{D}^2$ the derivatives of $I\!E[\mathcal{L}]$ are taken with respect to all coordinates of $\boldsymbol{\upsilon} \in l^2$ and the Hessian is calculated in the "true point" $\boldsymbol{\upsilon}^* \in l^2$.

Remark 4.3.1. To be more precise we assume that $I\!E\mathcal{L} : \varUpsilon \to \mathbb{R}$ is Fréchet differentiable and that each element of the gradient $\langle \nabla I\!E\mathcal{L}, \boldsymbol{e}_k \rangle$ again is Fréchet differentiable as well. We denote the resulting operator by $\mathcal{D}^2 = \nabla^2 I\!E[\mathcal{L}(\boldsymbol{\upsilon}^*)] : \overline{\operatorname{span}\varUpsilon} \to \overline{\operatorname{span}\varUpsilon}$.

The second bias - i.e. bounds for $\|I - \breve{D}_m^{-1}(\boldsymbol{\upsilon}_m^*)\breve{D}^2(\boldsymbol{\upsilon}^*)\breve{D}_m^{-1}(\boldsymbol{\upsilon}_m^*)\|$ - will be neglected for now, as only the operator $\breve{D}_m^2(\boldsymbol{\upsilon}_m^*) \in \mathbb{R}^{p\times p}$ is available in practice. We will come back to it, when we derive efficiency for the sieve profile estimator $\widetilde{\boldsymbol{\theta}}_m \in \mathbb{R}^p$.

For the first type of bias we impose the following condition:

(*bias*) There exists a function $\alpha : \mathbb{N} \to \mathbb{R}_+$ such that

$$\|\breve{D}_m(\boldsymbol{\upsilon}_m^*)(\boldsymbol{\theta}_m^* - \boldsymbol{\theta}^*)\| \le \alpha(m), \quad \alpha(m) \to 0, \text{ as } m \to \infty.$$

Remark 4.3.2. Section 4.3.3 presents conditions on the structure of $\mathcal{D} : l^2 \to l^2$ and on the sequence $\boldsymbol{\eta}^* \in l^2$ that yield **(*bias*)**.

We represent

$$\mathcal{D}_m^2(\boldsymbol{\upsilon}_m^*) = \begin{pmatrix} D^2(\boldsymbol{\upsilon}_m^*) & \mathrm{A}_m^\top(\boldsymbol{\upsilon}_m^*) \\ \mathrm{A}_m(\boldsymbol{\upsilon}_m^*) & \mathrm{H}_m^2(\boldsymbol{\upsilon}_m^*) \end{pmatrix} \in \mathbb{R}^{(p+m)\times(p+m)}.$$

With Theorem 4.2.2 and **(*bias*)** we directly get the following corollary:

Corollary 4.3.1. *Assume (bias) and that the conditions* $(\breve{\mathcal{E}}\mathcal{D}_0)$, $(\breve{\mathcal{E}}\mathcal{D}_1)$ *and* $(\breve{\mathcal{L}}_0)$ *from Section 4.2.1 are satisfied for all* $m \ge m_0$ *for some* $m_0 \in \mathbb{N}$ *and with* $\mathcal{D}_0^2 = \nabla_{p+m}^2 I\!E\mathcal{L}_m(\boldsymbol{\upsilon}_m^*) \in \mathbb{R}^{p^*\times p^*}$, $\mathcal{V}_0^2 = \operatorname{Cov}[\nabla_{p+m}\mathcal{L}_m(\boldsymbol{\upsilon}_m^*)] \in \mathbb{R}^{p^*\times p^*}$ *and* $\boldsymbol{\upsilon}^\circ = \boldsymbol{\upsilon}_m^* \in \mathbb{R}^{p^*}$. *Assume that* $\widetilde{\boldsymbol{\upsilon}}_m \ne \emptyset$ *and* $\widetilde{\boldsymbol{\upsilon}}_{\boldsymbol{\theta}_m^*} \ne \emptyset$. *Choose* $\mathtt{r}_0(\mathtt{x}) > 0$ *such that* $I\!P(\widetilde{\boldsymbol{\upsilon}}_m, \widetilde{\boldsymbol{\upsilon}}_{\boldsymbol{\theta}_m^*,m} \in \varUpsilon_{0,m}(\mathtt{r}_0(\mathtt{x}))) \ge 1 - \mathrm{e}^{-\mathtt{x}}$. *Then it holds for any* $m \ge m_0$ *with probability greater than* $1 - 2\mathrm{e}^{-\mathtt{x}}$

$$\big\|\breve{D}_m(\boldsymbol{\upsilon}_m^*)\big(\widetilde{\boldsymbol{\theta}}_m - \boldsymbol{\theta}^*\big) - \breve{\boldsymbol{\xi}}_m(\boldsymbol{\upsilon}_m^*)\big\| \le \breve{\diamondsuit}(\mathtt{r}_0, \mathtt{x}) + \alpha(m),$$

where

$$\breve{\boldsymbol{\xi}}_m(\boldsymbol{\upsilon}_m^*) \stackrel{\text{def}}{=} \breve{D}_m^{-1}(\nabla_{\boldsymbol{\theta}} - A_m H_m^{-1}\nabla_{\boldsymbol{\eta}})\mathcal{L}_m(\boldsymbol{\upsilon}_m^*).$$

Define

$$\breve{L}(\boldsymbol{\theta}) \stackrel{\text{def}}{=} \max_{\boldsymbol{\eta}\in\mathbb{R}^m} \mathcal{L}_m(\boldsymbol{\theta}, \boldsymbol{\eta}),$$

where it is important to note that the maximization is restricted to the finite dimensional space $\mathbb{R}^m$. As above abbreviate $\breve{L}(\boldsymbol{\theta},\boldsymbol{\theta}^*) \stackrel{\text{def}}{=} \breve{L}(\boldsymbol{\theta}) - \breve{L}(\boldsymbol{\theta}^*)$. For the bias in the Wilks result a bit more work is needed. We can show the following:

Theorem 4.3.2. *Assume the same as in Corollary 4.3.1. Pick a radius* $0 < \mathtt{r}_0^\circ$ *such that*

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

Then we get with probability greater than $1 - 2\mathrm{e}^{-\mathtt{x}}$

$$\begin{aligned}&\left|\max_{\boldsymbol{\eta}\in\Pi_m\Upsilon_{\boldsymbol{\eta}}} \mathcal{L}(\widetilde{\boldsymbol{\theta}}_m,\boldsymbol{\eta}) - \max_{\boldsymbol{\eta}\in\Pi_m\Upsilon_{\boldsymbol{\eta}}} \mathcal{L}(\boldsymbol{\theta}^*,\boldsymbol{\eta}) - \|\breve{\boldsymbol{\xi}}_m\|^2/2\right| \\ &\qquad \le 9\left(\|\breve{\boldsymbol{\xi}}_m(\boldsymbol{\upsilon}_m^*)\| + \breve{\diamondsuit}(\mathtt{r}_0^\circ,\mathtt{x})\right)\left(\breve{\diamondsuit}(\mathtt{r}_0^\circ,\mathtt{x}) + \alpha(m)\right) + 2\alpha(m).\end{aligned}$$

Remark 4.3.3. With condition $(\breve{\mathcal{E}}\mathcal{D}_0)$ we can use Theorem 3.4.1 to obtain

$$\mathbb{P}\left(\|\breve{\boldsymbol{\xi}}_m(\boldsymbol{\upsilon}_m^*)\| \ge \mathfrak{z}(\mathtt{x},\breve{\mathbb{B}})\right) \le \mathrm{e}^{-\mathtt{x}}.$$

Remark 4.3.4. The radius $\mathtt{r}_0^\circ \in \mathbb{R}$ can be determined again using the tools of Section 4.2.3. Clearly Theorem 3.3.2 can be applied to find some $\mathtt{r}_0 \le \mathtt{r}_0^\circ$ such that

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

Furthermore note that by the mean value theorem

$$\begin{aligned}\mathcal{L}_m(\boldsymbol{\theta}^*,\widetilde{\boldsymbol{\eta}}_{\boldsymbol{\theta}^*,m}) - \mathcal{L}_m(\boldsymbol{\upsilon}_m^*) &\ge \mathcal{L}_m(\boldsymbol{\theta}^*,\boldsymbol{\eta}_m^*) - \mathcal{L}_m(\boldsymbol{\upsilon}_m^*) \\ &\ge -(1+\nu)\alpha(m) \sup_{\boldsymbol{\upsilon}\in\Upsilon_\circ((1+\nu)\alpha(m))} \|D^{-1}\nabla_{\boldsymbol{\theta}}\mathcal{L}_m(\boldsymbol{\upsilon})\|.\end{aligned}$$

With condition $(\breve{\mathcal{E}}\mathcal{D}_0)$ and $(\breve{\mathcal{E}}\mathcal{D}_1)$ the right-hand side can be bounded by some constant $-\alpha(m)\mathtt{C}(p^*+\mathtt{x}) \in \mathbb{R}$ with probability greater than $1-2\mathrm{e}^{-\mathtt{x}}$ using the tools of Section 4.2.3. Combining this with Theorem 3.3.2 gives that with

$$\mathtt{r}_0^\circ = 6\mathtt{b}^{-1}\nu_{\mathtt{r}}\sqrt{\mathtt{x} + \log(4) + p^* + \frac{\mathtt{b}}{9\nu_{\mathtt{r}}^2}\alpha(m)\mathtt{C}(p^*+\mathtt{x})},$$

it holds that

$$\begin{aligned}\mathbb{P}\left(\left\{\widetilde{\boldsymbol{\upsilon}}_m, \widetilde{\boldsymbol{\upsilon}}_{\boldsymbol{\theta}^*_m,m} \in \Upsilon_{0,m}(\mathtt{r}_0)\right\} \cap \left\{\widetilde{\boldsymbol{\upsilon}}_{\boldsymbol{\theta}^*,m} \in \Upsilon_{0,m}(\mathtt{r}_0^\circ)\right\}\right) &> 1 - 4\mathrm{e}^{-\mathtt{x}-\log(4)} \\ &= 1 - \mathrm{e}^{-\mathtt{x}}.\end{aligned}$$

This means that $\mathtt{r}_0^\circ \approx \mathtt{r}_0$ as long as $\alpha(m) \to 0$.

Now we want to show how this approach allows to prove the classical weak convergence statements for the sieve profile ME and efficiency of the sieve profile MLE $\widetilde{\boldsymbol{\theta}}_m \in \mathbb{R}^p$. From this point on we focus on the i.i.d. model in which n denotes the sample size and the functional is of the form $\mathcal{L} = \sum_{i=1}^n \ell(\boldsymbol{\theta}, \boldsymbol{\eta}, \boldsymbol{Y}_i)$. As in Section 4.2.4 this gives that $\mathcal{D}_m^2 = n d_m$, $\breve{D}_m^2 = n\breve{d}_m$ and $\breve{D}^2 = n\breve{d}$. As the efficient covariance is derived for the score evaluated at the true full target $\boldsymbol{\upsilon}^* \in l^2$ we need further assumptions on the bias:

$(\boldsymbol{bias'})$ With $\|\cdot\|$ denoting the spectral norm and with some function $\beta(m) \to 0$ as $m \to \infty$ it holds that

$$\|I - \breve{D}_m(\boldsymbol{\upsilon}^*)^{-1}\breve{D}(\boldsymbol{\upsilon}^*)^2\breve{D}_m(\boldsymbol{\upsilon}^*)^{-1}\| \leq \beta(m),$$
$$\|I - \breve{D}_m(\boldsymbol{\upsilon}_m^*)^{-1}\breve{D}_m(\boldsymbol{\upsilon}^*)^2\breve{D}_m(\boldsymbol{\upsilon}_m^*)^{-1}\| \leq \beta(m).$$

Remark 4.3.5. Again we postpone the question how to satisfy the above condition to Section 4.3.3 which presents conditions on the structure of $\mathcal{D} : l^2 \to l^2$ and on the sequence $\boldsymbol{\eta}^* \in l^2$ that yield $(\boldsymbol{bias'})$.

Furthermore we need convergence of the covariance of the weighted score. For this define

$$\breve{v}_{m,\mathcal{D}}^2(\boldsymbol{\upsilon}_m^*) \stackrel{\text{def}}{=} \operatorname{Cov}\left(\nabla_{\boldsymbol{\theta}}\ell_1(\boldsymbol{\upsilon}_m^*) - \mathrm{A}_m\mathrm{H}_m^{-2}\nabla_{\boldsymbol{\eta}}\ell_1(\boldsymbol{\upsilon}_m^*)\right),$$
$$\breve{v}^2(\boldsymbol{\upsilon}^*) \stackrel{\text{def}}{=} \operatorname{Cov}\left(\nabla_{\boldsymbol{\theta}}\ell_1(\boldsymbol{\upsilon}^*) - \mathrm{A}\mathrm{H}^{-2}\nabla_{\boldsymbol{\eta}}\ell_1(\boldsymbol{\upsilon}^*)\right).$$

$(\boldsymbol{bias''})$ As $m \to \infty$ with $\|\cdot\|$ denoting the spectral norm

$$\|\breve{D}_m^{-1}(\boldsymbol{\upsilon}_m^*)\breve{V}_{m,\mathcal{D}}^2(\boldsymbol{\upsilon}_m^*)\breve{D}_m^{-1}(\boldsymbol{\upsilon}_m^*) - \breve{d}^{-1}\breve{v}^2\breve{d}^{-1}\| \to 0.$$

Remark 4.3.6. This is a condition on how the covariance operator of $\nabla_{p+m}\mathcal{L}(\boldsymbol{\upsilon}) \in \mathbb{R}^{p+m}$ is affected when it is evaluated in $\boldsymbol{\upsilon}_m^* \in \mathbb{R}^{p+m}$ instead of $\boldsymbol{\upsilon}^* \in l^2$. In the single-index example we get $(bias'')$ due to the smoothness of the functional.

Corollary 4.3.1 and Theorem 4.3.2 allow to derive the following corollary which yields the asymptotic efficiency of $\widetilde{\boldsymbol{\theta}}_m$ and the classical Wilks phenomenon.

Corollary 4.3.3. *Assume that we are given iid observations from* $I\!\!P = I\!\!P_{\boldsymbol{\theta}^*,\boldsymbol{\eta}^*}$. *Assume that for some* $m_0 \in \mathbb{N}$ *any* $m \geq m_0$ *the conditions of Theorem 4.3.1 and the condition* $(\breve{\mathcal{E}}\mathcal{D}_0)$ *are satisfied with* $\boldsymbol{\upsilon}^\circ = \boldsymbol{\upsilon}_m^*$. *Furthermore let the conditions* $(bias')$ *and* $(bias'')$ *be satisfied. Assume that for any* $\mathtt{r} > 0$ *that* $\breve{\delta}_n(\mathtt{r}) \to 0$ *as* $n \in \mathbb{N}$ *tends to infinity and that* $\breve{\omega}_n \to 0$. *Finally assume that* $\mathtt{r}_0(\mathtt{x}) < \infty$ *for any* $\mathtt{x} > 0$, $m, n \in \mathbb{N}$, *where* $\mathtt{r}_0(\mathtt{x})$ *is* *chosen such that* $I\!P(\widetilde{\boldsymbol{\upsilon}}_m, \widetilde{\boldsymbol{\upsilon}}_{\boldsymbol{\theta}^*_m,m}, \widetilde{\boldsymbol{\upsilon}}_{\boldsymbol{\theta}^*,m} \in \varUpsilon_\circ(\mathtt{r}_0)) \ge 1 - \mathrm{e}^{-\mathtt{x}}$ *. Then there is a sequence* $m_n \to \infty$ *such that - with convergence as* $n \to \infty$ *-*

$$\begin{aligned}
\sqrt{n}\breve{d}\big(\widetilde{\boldsymbol{\theta}}_{m_n} - \boldsymbol{\theta}^*\big) - \breve{\boldsymbol{\xi}} \qquad\qquad &\xrightarrow{I\!P} 0,\\
\sqrt{n}\breve{d}\big(\widetilde{\boldsymbol{\theta}}_{m_n} - \boldsymbol{\theta}^*\big) \qquad\qquad &\xrightarrow{w} \mathcal{N}(0, \breve{d}^{-1}\breve{v}^2\breve{d}^{-1}),\\
\max_{\boldsymbol{\eta}\in \varPi_m\varUpsilon_{\boldsymbol{\eta}}} \mathcal{L}(\widetilde{\boldsymbol{\theta}}_m, \boldsymbol{\eta}) - \max_{\boldsymbol{\eta}\in \varPi_m\varUpsilon_{\boldsymbol{\eta}}} \mathcal{L}(\boldsymbol{\theta}^*, \boldsymbol{\eta}) &\xrightarrow{w} \mathfrak{L}(\|\breve{\boldsymbol{\xi}}_\infty\|^2/2),\\
&\quad \breve{\boldsymbol{\xi}}_\infty \sim \mathcal{N}(0, \breve{d}^{-1}\breve{v}^2\breve{d}^{-1}).
\end{aligned}$$

Remark 4.3.7. On this level of generality we can not specify the right choice of $m_n \in \mathbb{N}$ that ensures the convergence. But in Chapter 6 we manage to show that it equals the optimal choice for a series estimator of the nuisance component $\boldsymbol{\eta}^* \in l^2$ for known $\boldsymbol{\theta}^*$. As is pointed out in [42], the best choice is $m = n^{1/(2\alpha+1)}$, with $\alpha > 1/2$ quantifying the "smoothness" of $\boldsymbol{\eta}^*$ - is admissible.

Remark 4.3.8. For the case of the profile MLE $\ell(\boldsymbol{\theta}, \boldsymbol{\eta}, \boldsymbol{Y}_i)$ is the log-likelihood for a single observation. In that case assume that the linear operator $\mathbb{F}^2_{\boldsymbol{\upsilon}^*} \stackrel{\mathrm{def}}{=} \operatorname{Cov}\big\{\nabla\ell(\boldsymbol{\upsilon}^*)\big\} : l^2 \to Im(\mathbb{F}^2_{\boldsymbol{\upsilon}^*})$ is invertible and that $\nabla\ell(\boldsymbol{\upsilon}^*) \in Im(\mathbb{F}^2_{\boldsymbol{\upsilon}^*})$. With Corollary 2.1.9 we infer that the asymptotically optimal variance for regular estimators is given by the inverse of the partial information matrix

$$\breve{\mathbb{F}}_{\boldsymbol{\upsilon}^*} = \Big(\varPi_{\boldsymbol{\theta}} \operatorname{Cov}\big\{\nabla\ell(\boldsymbol{\upsilon}^*)\big\}^{-1}\varPi_{\boldsymbol{\theta}}^\top\Big)^{-1},$$

where as above $\varPi_{\boldsymbol{\theta}}$ is the orthogonal projection onto the $\boldsymbol{\theta}$-components, and $\varPi_{\boldsymbol{\theta}}^\top$ its adjoint operator. In case of correct specification we have that $\breve{v}^2 = \breve{d}^{-1} = \breve{\mathbb{F}}_{\boldsymbol{\theta},\boldsymbol{\eta}}$, such that

$$\breve{d}^{-1}\breve{v}^2\breve{d}^{-1} = I_p.$$

In that case Corollary 4.3.3 yields the efficiency of the sieve profile MLE and we recover the Wilks phenomenon for that estimator.

4.3.3 One way to control the sieve bias

In this section we present a particular way to derive the conditions $(bias)$ and $(bias')$. For this we assume that $\mathcal{X} = l^2 \stackrel{\mathrm{def}}{=} \{(x_k)_{k=1}^\infty \subset \mathbb{R}, \sum_{k=1}^\infty x_k^2 < \infty\}$. Denote by $\varPi_{p^*} : l^2 \to \mathbb{R}^{p^*}$ the projection to the first $p^* \in \mathbb{N}$ coordinates of an element of l^2. By abuse of notation we denote by $(Id_{l^2} - \varPi_{p^*})$ the orthogonal projection onto $\{\boldsymbol{x} \in l^2 : x_k = 0, k = 1, \ldots, m\}$. Furthermore denote by $\nabla_{\boldsymbol{\upsilon}_m}$ the differentiation with respect to the $m \in \mathbb{N}$ first coordinates. We represent

$$\mathcal{D}(\boldsymbol{\upsilon}) \overset{\text{def}}{=} -\nabla^2 I\!\!E\mathcal{L}(\boldsymbol{\upsilon}) = \begin{pmatrix} \mathcal{D}_m^2(\boldsymbol{\upsilon}) & A_{\boldsymbol{\varkappa}\boldsymbol{\upsilon}_m}^\top(\boldsymbol{\upsilon}) \\ A_{\boldsymbol{\varkappa}\boldsymbol{\upsilon}_m}(\boldsymbol{\upsilon}) & \mathcal{H}_{\boldsymbol{\varkappa}\boldsymbol{\varkappa}}^2(\boldsymbol{\upsilon}) \end{pmatrix}.$$

To bound the bias $\|\breve{D}_m(\boldsymbol{\upsilon}_m^*)(\boldsymbol{\theta}_m^* - \boldsymbol{\theta}^*)\| > 0$ we present the following condition:

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

$$\|\mathcal{D}_m^{-1} A_{\boldsymbol{\varkappa}\boldsymbol{\upsilon}_m}^\top \boldsymbol{\varkappa}^*\| \le \widehat{\alpha}(m). \tag{4.3.4}$$

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

$$\|\mathcal{D}_m^{-1}\left(\nabla_{\boldsymbol{\upsilon}_m\boldsymbol{\varkappa}} I\!\!E\mathcal{L}(\mathit{\Pi}_{p^*}\boldsymbol{\upsilon}^*, \lambda\boldsymbol{\varkappa}^*) - A_{\boldsymbol{\varkappa}\boldsymbol{\upsilon}_m}^\top\right)\boldsymbol{\varkappa}^*\| \le \tau(m),$$

$$\left|\boldsymbol{\varkappa}^{*\top}(\mathcal{H}_{\boldsymbol{\varkappa}\boldsymbol{\varkappa}} - \nabla_{\boldsymbol{\varkappa}\boldsymbol{\varkappa}} I\!\!E\mathcal{L}\left((\mathit{\Pi}_{p^*}\boldsymbol{\upsilon}^*, \lambda\boldsymbol{\varkappa}^*)\right)\boldsymbol{\varkappa}^*\right| \le \mathtt{C}_{\boldsymbol{\varkappa}^*} m. \tag{4.3.5}$$

Remark 4.3.9. This condition corresponds to condition (approximation accuracy) and condition (2.2.6) of Section 2.2.3, which are taken from [50]. The later means $A_{\boldsymbol{\varkappa}\boldsymbol{\upsilon}_m} \equiv 0$ such that the most interesting results of this subsection can not be related to that paper. As was pointed out in Section 2.2.3 conditions of the type (4.3.4) are crucial in settings where $\nabla^2 I\!\!E\mathcal{L}(\boldsymbol{\upsilon}^*)$ is unknown. In those cases a basis that is orthogonal with respect to the inner product $\langle \mathcal{D}^2\cdot, \cdot\rangle$ cannot be constructed.

To ensure that $\breve{D}_m(\boldsymbol{\upsilon}_m^*)$ is close to $\breve{D}(\boldsymbol{\upsilon}^*)$ we impose the following second condition.

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

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

Furthermore we introduce the following infinite dimensional version of $(\mathcal{L}\mathtt{r}_\infty)$ from Section 4.2.1:

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

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

Remark 4.3.10. Conditions $(\boldsymbol{\mathcal{L}}\mathtt{r}_\infty)$, the smoothness of $\mathcal{L} : \varUpsilon \to \mathbb{R}$ and the assumption that

$$\|\mathcal{H}_{\boldsymbol{\varkappa}\boldsymbol{\varkappa}}\boldsymbol{\varkappa}^*\|^2 \le \mathtt{C}_{\boldsymbol{\varkappa}^*} m,$$

are strongly related to conditions (identification), (approximation accuracy), and (compactness) from Section 2.2.3. Both sets of conditions allow to bound

$$\|\mathrm{H}_m(\boldsymbol{\eta}^*_m - \mathit{\Pi}_m\boldsymbol{\eta}^*)\| \le \mathtt{C}\sqrt{m}.$$

Theorem 4.3.4. *Let the condition* $(\mathcal{L}\mathtt{r}_\infty)$ *with* $\mathtt{b}(\mathtt{r}) \equiv \mathtt{b} > 0$, $(\boldsymbol{\varkappa})$ *and condition* $(\mathcal{I})$ *from Section 4.2.1 be satisfied for both* $\mathcal{D}_m(\boldsymbol{\upsilon}^*)$ *and* $\mathcal{D}_m(\boldsymbol{\upsilon}^*_m)$ *and for* $I\!E\mathcal{L} : l^2 \to \mathbb{R}$. *Set* $\mathtt{r}^{*2} = 4\mathtt{C}^2_{\boldsymbol{\varkappa}^*} m/\mathtt{b}$ *and let for some* $m_0 \in \mathbb{N}$ *and all* $m \ge m_0$ *the condition* $(\breve{\mathcal{L}}_0)$ *be fulfilled for* $\mathcal{D}_0 = \mathcal{D}_m(\boldsymbol{\upsilon}^*_m)$, $\boldsymbol{\upsilon}^\circ = \boldsymbol{\upsilon}^*_m$ *and for any* $\mathtt{r} \le \mathtt{r}^*$. *If further* $(\boldsymbol{\varkappa})$ *is fulfilled then* $\boldsymbol{(bias)}$ *is satisfied with*

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

If further the condition $(\boldsymbol{\upsilon\varkappa})$ *is fulfilled then* $\boldsymbol{(bias')}$ *is satisfied with a constant* $\mathtt{C}(\nu, \breve{\delta}(\mathtt{r}^*)) > 0$ *and*

$$\|I - \breve{D}_m(\boldsymbol{\upsilon}^*)^{-1}\breve{D}(\boldsymbol{\upsilon}^*)^2\breve{D}_m(\boldsymbol{\upsilon}^*)^{-1}\| \le \frac{1+\nu^2+\beta^2(m)}{1-\nu^2}\frac{\beta^2(m)}{1-\beta^2(m)},$$

and

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

4.A Proofs

This section collects the proofs in chronological order.

4.A.1 Proof of Lemma 4.2.1

Proof. Take any $\boldsymbol{\gamma} \in \mathbb{R}^p$ with $\|\boldsymbol{\gamma}\| = 1$ then

$$\begin{aligned}\boldsymbol{\gamma}^\top\breve{D}^{-1}\breve{\nabla}_{\boldsymbol{\theta}}\zeta(\boldsymbol{\upsilon}) &= \boldsymbol{\gamma}^\top\begin{pmatrix}\breve{D}^{-1}D & \breve{D}^{-1}A\mathrm{H}^{-1}\end{pmatrix}\begin{pmatrix}D^{-1} & 0\\ 0 & \mathrm{H}^{-1}\end{pmatrix}\nabla\zeta(\boldsymbol{\upsilon})\\ &\stackrel{\mathrm{def}}{=} \widehat{\boldsymbol{\gamma}}^\top\mathcal{D}^{-1}\nabla\zeta(\boldsymbol{\upsilon}),\end{aligned}$$

where

$$\|\widehat{\gamma}\| \le \left\|\begin{pmatrix}\breve{D}^{-1}D & \breve{D}^{-1}A\mathrm{H}^{-1}\end{pmatrix}\right\|\left\|\begin{pmatrix}D^{-1} & 0\\ 0 & \mathrm{H}^{-1}\end{pmatrix}\mathcal{D}\right\| \le \frac{(1+\nu)\sqrt{1+\nu^2}}{\sqrt{1-\nu^2}}.$$

This gives that $\boldsymbol{(\mathcal{E}\mathfrak{D}_1)}$ implies $\boldsymbol{(\breve{\mathcal{E}}\mathfrak{D}_1)}$ and $\boldsymbol{(\mathcal{E}\mathfrak{D}_0)}$ implies $\boldsymbol{(\breve{\mathcal{E}}\mathfrak{D}_0)}$ with

$$\breve{\mathrm{g}} = \frac{\sqrt{1-\nu^2}}{(1+\nu)\sqrt{1+\nu^2}}\mathrm{g}, \quad \breve{\nu} = \frac{(1+\nu)\sqrt{1+\nu^2}}{\sqrt{1-\nu^2}}\nu.$$

Furthermore for any $\boldsymbol{\upsilon} \in \varUpsilon_\circ(\mathtt{r})$

$$\begin{aligned} \|I_p - D^{-1}D^2(\boldsymbol{\upsilon})D^{-1}\| &= \|D^{-1}(D^2 - D^2(\boldsymbol{\upsilon}))D^{-1}\| \\ &= \|D^{-1}\varPi_{\boldsymbol{\theta}}(\mathcal{D}^2 - \mathcal{D}^2(\boldsymbol{\upsilon}))\varPi_{\boldsymbol{\theta}}^\top D^{-1}\| \\ &= \|D^{-1}\varPi_{\boldsymbol{\theta}}\mathcal{D}(I_{p^*} - \mathcal{D}^{-1}\mathcal{D}^2(\boldsymbol{\upsilon})\mathcal{D}^{-1})\mathcal{D}\varPi_{\boldsymbol{\theta}}^\top D^{-1}\| \\ &\le \|D^{-1}\varPi_{\boldsymbol{\theta}}\mathcal{D}\|^2\|I_{p^*} - \mathcal{D}^{-1}\mathcal{D}^2(\boldsymbol{\upsilon})\mathcal{D}^{-1}\| = \delta(\mathtt{r}). \end{aligned}$$

Also

$$\begin{aligned} \|D^{-1}(\mathrm{A}(\boldsymbol{\upsilon}) - \mathrm{A})\mathrm{H}^{-1}\| &= \|D^{-1}\varPi_{\boldsymbol{\theta}}(\mathcal{D}^2(\boldsymbol{\upsilon}) - \mathcal{D}^2)\varPi_{\boldsymbol{\eta}}^\top \mathrm{H}^{-1}\| \\ &= \|D^{-1}\varPi_{\boldsymbol{\theta}}\mathcal{D}(\mathcal{D}^{-1}\mathcal{D}^2(\boldsymbol{\upsilon})\mathcal{D}^{-1} -)\mathcal{D}\varPi_{\boldsymbol{\eta}}^\top \mathrm{H}^{-1}\| \\ &\le \|D^{-1}\varPi_{\boldsymbol{\theta}}\mathcal{D}\|\|\mathrm{H}^{-1}\varPi_{\boldsymbol{\eta}}\mathcal{D}\|\|I_{p^*} - \mathcal{D}^{-1}\mathcal{D}^2(\boldsymbol{\upsilon})\mathcal{D}^{-1}\| \\ &\le \delta(\mathtt{r}). \end{aligned}$$

With the same arguments

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

□

4.A.2 Proof of Theorem 4.2.2

For $\zeta(\boldsymbol{\upsilon}) = \mathcal{L}(\boldsymbol{\upsilon}) - \mathbb{E}\mathcal{L}(\boldsymbol{\upsilon})$ remember the semiparametric normalized stochastic gradient gap

$$\breve{\mathcal{Y}}(\boldsymbol{\upsilon}) = \breve{D}^{-1}\Big(\breve{\nabla}_{\boldsymbol{\theta}}\zeta(\boldsymbol{\upsilon}) - \breve{\nabla}_{\boldsymbol{\theta}}\zeta(\boldsymbol{\upsilon}^*)\Big). \tag{4.A.1}$$

Fix the radius $\mathtt{r}_0(\mathtt{x}) > 0$ that ensures $\mathbb{P}\{\widetilde{\boldsymbol{\upsilon}}, \widetilde{\boldsymbol{\upsilon}}_{\boldsymbol{\theta}^*} \in \varUpsilon_\circ(\mathtt{r}_0)\} \ge 1 - \mathrm{e}^{-\mathtt{x}}$. Define $C(\mathtt{r}_0, \mathtt{x}) \subseteq \varOmega$ as

$$C(\mathtt{r}_0, \mathtt{x}) \overset{\mathrm{def}}{=} \big\{\widetilde{\boldsymbol{\upsilon}}, \widetilde{\boldsymbol{\upsilon}}_{\boldsymbol{\theta}^*} \in \varUpsilon_\circ(\mathtt{r}_0)\big\} \cap \left\{\sup_{\boldsymbol{\upsilon}\in\varUpsilon_\circ(4\mathtt{r}_0)} \|\breve{\mathcal{Y}}(\boldsymbol{\upsilon})\| \le 6\nu_1\breve{\omega}\mathfrak{z}(\mathtt{x}, \mathbb{Q})4\mathtt{r}_0\right\}.$$

In the following we will derive statements that hold true on this set $C(\mathtt{r}_0, \mathtt{x}) \subseteq \varOmega$ which is of probability greater than $1 - 2\mathrm{e}^{-\mathtt{x}}$. Indeed it follows right away from the definition of $\mathtt{r}_0 > 0$ that

$$\mathbb{P}\big\{\widetilde{\boldsymbol{\upsilon}}, \widetilde{\boldsymbol{\upsilon}}_{\boldsymbol{\theta}^*} \notin \varUpsilon_\circ(\mathtt{r}_0)\big\} \le \mathrm{e}^{-\mathtt{x}}.$$

By Theorem 3.5.9 - which is applicable because $(\breve{\mathcal{E}}\mathcal{D}_1)$ implies (3.5.4) with $\|\cdot\|_{\mathcal{Y}} = \|\mathcal{D}(\cdot)\|$ - we infer

$$\mathbb{P}\left(\sup_{\boldsymbol{\upsilon} \in \Upsilon_\circ(\mathtt{r}_0)} \|\breve{\mathcal{Y}}(\boldsymbol{\upsilon})\| \le 6\nu_1 \breve{\omega}\mathfrak{z}_1(\mathtt{x}, 2p^* + 2p)\mathtt{r}_0 \right) \ge 1 - \mathrm{e}^{-\mathtt{x}}.$$

Proof of claim on $C(\mathtt{r}_0, \mathtt{x}) \subseteq \Omega$

Before we prove the claim we prove the following useful lemma:

Lemma 4.A.1. *Assume that* $(\breve{\mathcal{L}}_0)$ *is fulfilled. Then*

$$\sup_{\boldsymbol{\upsilon} \in \Upsilon_\circ(\mathtt{r})} \left\| \breve{D}^{-1}\left(\breve{\nabla}\mathbb{E}\mathcal{L}(\boldsymbol{\upsilon}) - \breve{\nabla}\mathbb{E}\mathcal{L}(\boldsymbol{\upsilon}^*) \right) + \breve{D}(\boldsymbol{\theta} - \boldsymbol{\theta}^*) \right\| \le \frac{4}{(1-\nu^2)^2}\mathtt{r}\breve{\delta}(\mathtt{r}).$$

Proof. We have with Taylor expansion and some $\widehat{\boldsymbol{\upsilon}} \in \Upsilon_\circ(\mathtt{r})$

$$\begin{aligned} \nabla\mathbb{E}\mathcal{L}(\boldsymbol{\upsilon}) - \nabla\mathbb{E}\mathcal{L}(\boldsymbol{\upsilon}^*) &= \nabla^2\mathbb{E}\mathcal{L}(\widehat{\boldsymbol{\upsilon}})(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^*) \\ &\overset{\text{def}}{=} -\mathcal{D}^2(\widehat{\boldsymbol{\upsilon}})(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^*) \\ &= -\begin{pmatrix} D^2(\widehat{\boldsymbol{\upsilon}}) & \mathrm{A}(\widehat{\boldsymbol{\upsilon}}) \\ \mathrm{A}^\top(\widehat{\boldsymbol{\upsilon}}) & \mathrm{H}^2(\widehat{\boldsymbol{\upsilon}}) \end{pmatrix}(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^*). \end{aligned}$$

This gives

$$\begin{aligned} &-\breve{D}^{-1}\left(\breve{\nabla}\mathbb{E}\mathcal{L}(\boldsymbol{\upsilon}) - \breve{\nabla}\mathbb{E}\mathcal{L}(\boldsymbol{\upsilon}^*) \right) \\ &\quad = \breve{D}^{-1}\left(\; D^2(\widehat{\boldsymbol{\upsilon}}) - A\mathrm{H}^{-2}\mathrm{A}^\top(\widehat{\boldsymbol{\upsilon}}) \quad \mathrm{A}(\widehat{\boldsymbol{\upsilon}}) - \mathrm{A}\mathrm{H}^{-2}\mathrm{H}^2(\widehat{\boldsymbol{\upsilon}}) \; \right)(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^*) \\ &\quad = \breve{D}^{-1}\left(D^2(\widehat{\boldsymbol{\upsilon}}) - A\mathrm{H}^{-2}\mathrm{A}^\top(\widehat{\boldsymbol{\upsilon}}) \right)\breve{D}^{-1}\breve{D}(\boldsymbol{\theta} - \boldsymbol{\theta}^*) \\ &\qquad + \left(\breve{D}^{-1}\mathrm{A}(\widehat{\boldsymbol{\upsilon}}) - \breve{D}^{-1}\mathrm{A}\mathrm{H}^{-2}\mathrm{H}^2(\widehat{\boldsymbol{\upsilon}}) \right)(\boldsymbol{\eta} - \boldsymbol{\eta}^*). \end{aligned}$$

We estimate separately using $(\breve{\mathcal{L}}_0)$ and $(\mathcal{I})$

$$\begin{aligned} &\|\breve{D}^{-1}\left(D^2(\widehat{\boldsymbol{\upsilon}}) - A\mathrm{H}^{-2}\mathrm{A}^\top(\widehat{\boldsymbol{\upsilon}}) \right)\breve{D}^{-1} - I_p\| \\ &\quad = \left\| \breve{D}^{-1}\left(D^2(\widehat{\boldsymbol{\upsilon}}) - D^2 - \left\{ A\mathrm{H}^{-2}(\mathrm{A}^\top(\widehat{\boldsymbol{\upsilon}}) - \mathrm{A}^\top) \right\} \right)\breve{D}^{-1} \right\| \\ &\quad \le \|\breve{D}^{-1}D\|^2\left(\|D^{-1}D^2(\widehat{\boldsymbol{\upsilon}})D^{-1} - I_p\| \right. \\ &\qquad \left. + \|D^{-1}A\mathrm{H}^{-1}\|\|D^{-1}(\mathrm{A}(\widehat{\boldsymbol{\upsilon}}) - \mathrm{A})\mathrm{H}^{-1}\| \right) \\ &\quad \le \frac{1+\nu}{1-\nu^2}\breve{\delta}_0(\mathtt{r}), \end{aligned}$$

and

$$\begin{aligned}
&\left\| \left(\breve{D}^{-1}\mathrm{A}(\widehat{\boldsymbol{\upsilon}}) - \breve{D}^{-1}\mathrm{A}\mathrm{H}^{-2}\mathrm{H}^{2}(\widehat{\boldsymbol{\upsilon}}) \right) (\boldsymbol{\eta} - \boldsymbol{\eta}^{*}) \right\| \\
&\quad \le \left\| \breve{D}^{-1}\mathrm{A}(\widehat{\boldsymbol{\upsilon}})\mathrm{H}^{-1} - \breve{D}^{-1}\mathrm{A}\mathrm{H}^{-2}\mathrm{H}^{2}(\widehat{\boldsymbol{\upsilon}})\mathrm{H}^{-1} \right\| \|\mathrm{H}(\boldsymbol{\eta} - \boldsymbol{\eta}^{*})\| \\
&\quad \le \|\breve{D}^{-1}D\| \left\{ \|D^{-1}(\mathrm{A}(\widehat{\boldsymbol{\upsilon}}) - \mathrm{A})\mathrm{H}^{-1}\| \right. \\
&\qquad \left. + \|D^{-1}\mathrm{A}\mathrm{H}^{-1}\left(I_m - \mathrm{H}^{-1}\mathrm{H}^{2}(\widehat{\boldsymbol{\upsilon}})\mathrm{H}^{-1} \right)\| \right\} \|\mathrm{H}(\boldsymbol{\eta} - \boldsymbol{\eta}^{*})\| \\
&\quad \le \frac{2}{\sqrt{1-\nu^{2}}} \breve{\delta}_{0}(\mathtt{r}) \|\mathrm{H}(\boldsymbol{\eta} - \boldsymbol{\eta}^{*})\|.
\end{aligned}$$

Furthermore

$$\|\breve{D}(\boldsymbol{\theta} - \boldsymbol{\theta}^{*})\| \vee \|\mathrm{H}(\boldsymbol{\eta} - \boldsymbol{\eta}^{*})\| \le \frac{1}{\sqrt{1-\nu^{2}}} \frac{1}{\sqrt{1-\nu^{2}}} \|\mathcal{D}(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^{*})\| \le \frac{1}{1-\nu^{2}} \mathtt{r}.$$

Together this gives that

$$\begin{aligned}
&\sup_{\boldsymbol{\upsilon} \in \Upsilon_{\circ}(\mathtt{r})} \left\| \breve{D}^{-1}\left(\breve{\nabla}\mathbb{E}\mathcal{L}(\boldsymbol{\upsilon}) - \breve{\nabla}\mathbb{E}\mathcal{L}(\boldsymbol{\upsilon}^{*}) \right) + \breve{D}(\boldsymbol{\theta} - \boldsymbol{\theta}^{*}) \right\| \\
&\quad \le \left(\frac{1+\nu}{1-\nu^{2}} + \frac{2}{\sqrt{1-\nu^{2}}} \right) \frac{1}{1-\nu^{2}} \mathtt{r}\breve{\delta}(\mathtt{r}) \\
&\quad \le \frac{4}{(1-\nu^{2})^{2}} \mathtt{r}\breve{\delta}(\mathtt{r}).
\end{aligned}$$

□

The next lemma already completes the proof of (4.2.9) and (4.2.10) on $C(\mathtt{r}_0, \mathtt{x}) \subset \Omega$:

Lemma 4.A.2. *Assume that the condition* $(\breve{\mathcal{L}}_0)$ *is fulfilled. Then on the set* $C(\mathtt{r}_0, \mathtt{x}) \subset \Omega$ *the approximations* (4.2.9) *and* (4.2.10) *are valid.*

Proof. Using $\breve{\nabla}_{\boldsymbol{\theta}}\mathcal{L}(\widetilde{\boldsymbol{\upsilon}}) = 0$, that by assumption $\breve{\nabla}\mathbb{E}\mathcal{L} = \mathbb{E}\breve{\nabla}\mathcal{L}$ and the triangular inequality we find

$$\begin{aligned}
\|\breve{D}(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^{*}) - \breve{\boldsymbol{\xi}}\| &= \left\| \breve{D}^{-1}\left\{ \breve{\nabla}\mathcal{L}(\widetilde{\boldsymbol{\upsilon}}) - \breve{\nabla}\mathcal{L}(\boldsymbol{\upsilon}^{*}) \right\} + \breve{D}(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^{*}) \right\| \\
&\le \left\| \breve{D}^{-1}\left(\breve{\nabla}\mathbb{E}\mathcal{L}(\widetilde{\boldsymbol{\upsilon}}) - \breve{\nabla}\mathbb{E}\mathcal{L}(\boldsymbol{\upsilon}^{*}) \right) + \breve{D}(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^{*}) \right\| \\
&\quad + \left\| \breve{D}^{-1}\{ \breve{\nabla}_{\boldsymbol{\theta}}\zeta(\widetilde{\boldsymbol{\upsilon}}) - \breve{\nabla}_{\boldsymbol{\theta}}\zeta(\boldsymbol{\upsilon}^{*}) \} \right\|.
\end{aligned}$$

As we assume that $\widetilde{\boldsymbol{\upsilon}} \in \Upsilon_{\circ}(\mathtt{r}_0)$ we get with $(\breve{\mathcal{L}}_0)$ by Lemma 4.A.1

$$\left\| \breve{D}^{-1}\left(\breve{\nabla}\mathbb{E}\mathcal{L}(\widetilde{\boldsymbol{\upsilon}}) - \breve{\nabla}\mathbb{E}\mathcal{L}(\boldsymbol{\upsilon}^{*}) \right) + \breve{D}(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^{*}) \right\| \le \frac{4}{(1-\nu^{2})^{2}} \mathtt{r}_0\breve{\delta}(\mathtt{r}_0).$$

For the remainder we use that on $C(\mathtt{r}_0, \mathtt{x}) \subset \Omega$

$$\left\| \breve{D}^{-1}\bigl\{\breve{\nabla}_{\boldsymbol{\theta}}\zeta(\widetilde{\boldsymbol{\upsilon}}) - \breve{\nabla}_{\boldsymbol{\theta}}\zeta(\boldsymbol{\upsilon}^*)\bigr\}\right\| \le \sup_{\boldsymbol{\upsilon} \in \Upsilon_\circ(4\mathtt{r}_0)} \|\breve{\mathcal{Y}}(\boldsymbol{\upsilon})\| \le 6\breve{\nu}_1\breve{\omega}\mathfrak{z}_1(\mathtt{x}, \mathbb{Q})4\mathtt{r}_0.$$

This gives (4.2.9) on $C(\mathtt{r}_0, \mathtt{x}) \subset \Omega$. For (4.2.10) we will first show that on $C(\mathtt{r}_0, \mathtt{x}) \subset \Omega$

$$\begin{aligned} &\Bigl|\breve{L}(\widetilde{\boldsymbol{\theta}}) - \breve{L}(\boldsymbol{\theta}^*) - \Bigl(\breve{\nabla}\zeta(\boldsymbol{\upsilon}^*)(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) - \|\breve{D}(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\|^2/2\Bigr)\Bigr| \\ &\quad \le \Bigl(\|\breve{D}^{-1}\breve{\nabla}\| + \breve{\diamondsuit}(\mathtt{r}_0, \mathtt{x})\Bigr)\breve{\diamondsuit}(\mathtt{r}_0, \mathtt{x}), \end{aligned} \tag{4.A.2}$$

where $\breve{L}(\boldsymbol{\theta}) \stackrel{\text{def}}{=} \max_{\boldsymbol{\eta} \in \Upsilon_{\boldsymbol{\eta}}} \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\eta})$. To show this we use some ideas of the proof of Theorem 1 of [40], that is we define

$$l : \mathbb{R}^p \times \Upsilon \to \mathbb{R}, \ \ (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\eta}) \mapsto \mathcal{L}(\boldsymbol{\theta}_1, \boldsymbol{\eta} + \mathrm{H}^{-2}A^\top(\boldsymbol{\theta}_2 - \boldsymbol{\theta}_1)). \tag{4.A.3}$$

Note that

$$\begin{aligned} \nabla_{\boldsymbol{\theta}_1} l(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\eta}) &= \breve{\nabla}_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta}_1, \boldsymbol{\eta} + \mathrm{H}^{-2}A^\top(\boldsymbol{\theta}_2 - \boldsymbol{\theta}_1)), \\ \text{i.e. } \nabla_{\boldsymbol{\theta}_1} l(\boldsymbol{\theta}^*, \boldsymbol{\theta}^*, \boldsymbol{\eta}^*) &= \breve{\nabla}\zeta(\boldsymbol{\upsilon}^*). \end{aligned}$$

Remark 4.A.1. If the model was correctly specified and $\mathcal{L}$ the true log likelihood $\nabla_{\boldsymbol{\theta}_1} l(\boldsymbol{\theta}^*, \boldsymbol{\theta}^*, \boldsymbol{\eta}^*)$ would be equal to $\sum_{i=1}^n \widetilde{\psi}_{I\!\!P}(\boldsymbol{Y}_i)$, with $\widetilde{\psi}_{I\!\!P}$ the efficient influence function from (2.1.2).

We can represent:

$$\breve{L}(\widetilde{\boldsymbol{\theta}}) - \breve{L}(\boldsymbol{\theta}^*) = l(\widetilde{\boldsymbol{\theta}}, \widetilde{\boldsymbol{\theta}}, \widetilde{\boldsymbol{\eta}}) - l(\boldsymbol{\theta}^*, \boldsymbol{\theta}^*, \widetilde{\boldsymbol{\eta}}_{\boldsymbol{\theta}^*}), \quad \widetilde{\boldsymbol{\eta}}_{\boldsymbol{\theta}^*} \stackrel{\text{def}}{=} \Pi_{\boldsymbol{\eta}} \operatorname*{argmax}_{\substack{\boldsymbol{\upsilon} \in \Upsilon, \\ \Pi_0\boldsymbol{\upsilon} = \boldsymbol{\theta}^*}} \mathcal{L}(\boldsymbol{\upsilon}).$$

This allows to bound from above

$$\begin{aligned} \breve{L}(\widetilde{\boldsymbol{\theta}}) - \breve{L}(\boldsymbol{\theta}^*) &\le l(\widetilde{\boldsymbol{\theta}}, \widetilde{\boldsymbol{\theta}}, \widetilde{\boldsymbol{\eta}}) - l(\boldsymbol{\theta}^*, \widetilde{\boldsymbol{\theta}}, \widetilde{\boldsymbol{\eta}}) \\ &= \nabla_{\boldsymbol{\theta}_1} l(\boldsymbol{\theta}^*, \boldsymbol{\theta}^*, \boldsymbol{\eta}^*)(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) - \|\breve{D}(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\|^2/2 + \breve{\alpha}(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}^*), \end{aligned}$$

where

$$\begin{aligned} \breve{\alpha}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) \stackrel{\text{def}}{=} \ & l(\boldsymbol{\theta}_1, \widetilde{\boldsymbol{\theta}}, \widetilde{\boldsymbol{\eta}}) - l(\boldsymbol{\theta}_2, \widetilde{\boldsymbol{\theta}}, \widetilde{\boldsymbol{\eta}}) - \nabla_{\boldsymbol{\theta}_1} l(\boldsymbol{\theta}^*, \boldsymbol{\theta}^*, \boldsymbol{\eta}^*)(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2) \\ & + \|\breve{D}(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2)\|^2/2. \end{aligned}$$

We will show

$$\breve{\alpha}(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}^*) \le \Bigl(\|\breve{D}^{-1}\breve{\nabla}\| + \breve{\diamondsuit}(\mathtt{r}_0, \mathtt{x})\Bigr)\breve{\diamondsuit}(\mathtt{r}_0, \mathtt{x}), \tag{4.A.4}$$

which gives the upper bound of (4.A.2). Note that $\breve{\alpha}(\boldsymbol{\theta}^*, \boldsymbol{\theta}^*) = 0$ such that we get with Taylor expansion

$$\breve{\alpha}(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}^*) \leq \|\breve{D}(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\| \sup_{\boldsymbol{\theta} \in \varPi_{\boldsymbol{\theta}} \varUpsilon_{\circ}(\mathtt{r}_0)} |\breve{D}^{-1} \nabla_{\boldsymbol{\theta}_1} \breve{\alpha}(\boldsymbol{\theta}, \boldsymbol{\theta}^*)|.$$

We find

$$\begin{aligned} \nabla_{\boldsymbol{\theta}_1} \breve{\alpha}(\boldsymbol{\theta}, \boldsymbol{\theta}^*) &= \nabla_{\boldsymbol{\theta}_1} l(\boldsymbol{\theta}, \widetilde{\boldsymbol{\theta}}, \widetilde{\boldsymbol{\eta}}) - \nabla_{\boldsymbol{\theta}_1} l(\boldsymbol{\theta}^*, \boldsymbol{\theta}^*, \boldsymbol{\eta}^*) + \breve{D}(\boldsymbol{\theta} - \boldsymbol{\theta}^*) \\ &= \breve{\nabla}\zeta(\boldsymbol{\upsilon}^{\circ}) - \breve{\nabla}\zeta(\boldsymbol{\upsilon}^*) + I\!\!E\left[\breve{\nabla}\mathcal{L}(\boldsymbol{\upsilon}^{\circ}) - \breve{\nabla}\mathcal{L}(\boldsymbol{\upsilon}^*)\right] + \breve{D}(\boldsymbol{\theta} - \boldsymbol{\theta}^*), \end{aligned}$$

where

$$\begin{aligned} \boldsymbol{\upsilon}^{\circ} &\stackrel{\text{def}}{=} (\boldsymbol{\theta}, \widetilde{\boldsymbol{\eta}} + \mathrm{H}^{-2} A^{\top}(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta})), \\ \|\mathcal{D}(\boldsymbol{\upsilon}^{\circ} - \boldsymbol{\upsilon}^*)\| &\leq \|D(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\| + \|\mathrm{H}(\widetilde{\boldsymbol{\eta}} - \boldsymbol{\eta}^*)\| + \nu\|D(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta})\| \\ &\leq 2(1+\nu)\mathtt{r}_0 < 4\mathtt{r}_0. \end{aligned}$$

Using Lemma 4.A.1 and the definition of $C(\mathtt{r}_0, \mathtt{x})$ we can bound

$$\sup_{\boldsymbol{\theta} \in \varPi_{\boldsymbol{\theta}} \varUpsilon_{\circ}(\mathtt{r}_0)} |\breve{D}^{-1} \nabla_{\boldsymbol{\theta}_1} \breve{\alpha}(\boldsymbol{\theta}, \boldsymbol{\theta}^*)| \leq \breve{\diamondsuit}(\mathtt{r}_0, \mathtt{x}).$$

Using (4.2.9) we find on $C(\mathtt{r}_0, \mathtt{x})$

$$\|\breve{D}(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\| \leq \|\breve{D}^{-1}\breve{\nabla}\| + \breve{\diamondsuit}(\mathtt{r}_0, \mathtt{x}).$$

This gives (4.A.4). Similarly we can bound from below:

$$\breve{L}(\widetilde{\boldsymbol{\theta}}) - \breve{L}(\boldsymbol{\theta}^*) \geq l(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}^*, \widetilde{\boldsymbol{\eta}}_{\boldsymbol{\theta}^*}) - l(\boldsymbol{\theta}^*, \boldsymbol{\theta}^*, \widetilde{\boldsymbol{\eta}}_{\boldsymbol{\theta}^*}),$$

and repeat the same arguments using that $\widetilde{\boldsymbol{\upsilon}}_{\boldsymbol{\theta}^*} \in \varUpsilon_{\circ}(\mathtt{r}_0)$ on $C(\mathtt{r}_0, \mathtt{x}) \subset \varOmega$ to obtain the lower bound of (4.A.2). Plugging (4.2.9) into (4.A.2) this gives

$$\begin{aligned} \left|2\breve{L}(\widetilde{\boldsymbol{\theta}}) - 2\breve{L}(\boldsymbol{\theta}^*) - \|\breve{D}^{-1}\breve{\nabla}\zeta(\boldsymbol{\upsilon}^*)\|^2\right| &\leq 4\left(\|\breve{D}^{-1}\breve{\nabla}\| + \breve{\diamondsuit}(\mathtt{r}_0, \mathtt{x})\right)\breve{\diamondsuit}(\mathtt{r}_0, \mathtt{x}) \\ &\quad + \breve{\diamondsuit}(\mathtt{r}_0, \mathtt{x})^2. \end{aligned}$$

□

4.A.3 Proof of Proposition 4.2.3

We start with an auxiliary result. Define the *parametric gradient gap*

$$\mathcal{Y}(\boldsymbol{\upsilon}) = \mathcal{D}^{-1}\Big(\nabla\zeta(\boldsymbol{\upsilon}) - \nabla\zeta(\boldsymbol{\upsilon}^*)\Big),$$

and the set

$$C(\nabla) = \bigcap_{\mathtt{r}\leq \mathrm{R}_0(\mathtt{x})} \left\{ \sup_{\boldsymbol{\upsilon}\in\varUpsilon_\circ(\mathtt{r})} \left\{ \frac{1}{6\omega\nu_1}\|\mathcal{Y}(\boldsymbol{\upsilon})\| - 2\mathtt{r}^2 \right\} \leq \mathfrak{z}_Q(\mathtt{x},4p^*)^2 \right\}$$
$$\cap\left\{ \max\{\|\mathcal{D}^{-1}\nabla\mathcal{L}\|, \|D^{-1}\nabla_{\boldsymbol{\theta}}\mathcal{L}\|, \|\mathrm{H}^{-1}\nabla_{\boldsymbol{\eta}}\mathcal{L}\|\} \leq \mathfrak{z}(\mathtt{x}) \right\}.$$

Lemma 4.A.3. *We have that for any* $\mathtt{r}_0^{(l)} \leq \mathtt{r}_0$

$$C(\nabla) \cap \{\widetilde{\boldsymbol{\upsilon}}, \widetilde{\boldsymbol{\upsilon}}_{\boldsymbol{\theta}^*} \in \varUpsilon_\circ(\mathtt{r}_0^{(l)})\} \subseteq C(\nabla) \cap \{\widetilde{\boldsymbol{\upsilon}}, \widetilde{\boldsymbol{\upsilon}}_{\boldsymbol{\theta}^*} \in \varUpsilon_\circ(\mathtt{r}_0^{(l+1)})\},$$

where

$$\mathtt{r}_0^{(l+1)} = \mathfrak{z}(\mathtt{x}, I\!\!B) + \diamondsuit_Q(\mathtt{r}^{(l)}, \mathtt{x}) \wedge \mathtt{r}_0.$$

Proof. Since $\nabla\mathcal{L}(\widetilde{\boldsymbol{\upsilon}}) = 0$ we find with the triangular inequality

$$\|\mathcal{D}(\widetilde{\boldsymbol{\upsilon}} - \boldsymbol{\upsilon}^*) - \mathcal{D}^{-1}\nabla\mathcal{L}(\boldsymbol{\upsilon}^*)\| \leq \|\mathcal{D}^{-1}\Big(\nabla\zeta(\widetilde{\boldsymbol{\upsilon}}) - \nabla\zeta(\boldsymbol{\upsilon}^*)\Big)\|$$
$$+\|\mathcal{D}^{-1}I\!\!E\nabla\mathcal{L}(\boldsymbol{\upsilon}) - \mathcal{D}^{-1}I\!\!E\nabla\mathcal{L}(\boldsymbol{\upsilon}^*) + \mathcal{D}\left(\widetilde{\boldsymbol{\upsilon}} - \boldsymbol{\upsilon}^*\right)\|.$$

In Section 4.2.1 we assume that $\mathcal{L} : \mathbb{R}^{p^*} \to \mathbb{R}$ is smooth enough such that we can interchange $\nabla I\!\!E\mathcal{L}(\boldsymbol{\upsilon}) = I\!\!E\nabla\mathcal{L}(\boldsymbol{\upsilon})$ on $\varUpsilon_\circ(\mathtt{r}_0)$. This gives by condition $(\mathcal{L}_0)$ and Taylor expansion

$$\sup_{\boldsymbol{\upsilon}\in\varUpsilon_\circ(\mathtt{r})} \|\mathcal{D}^{-1}I\!\!E\nabla\mathcal{L}(\boldsymbol{\upsilon}) - \mathcal{D}^{-1}I\!\!E\nabla\mathcal{L}(\boldsymbol{\upsilon}^*) + \mathcal{D}\left(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^*\right)\|$$
$$\leq \sup_{\boldsymbol{\upsilon}\in\varUpsilon_\circ(\mathtt{r})} \|\mathcal{D}^{-1}\nabla^2 I\!\!E\mathcal{L}(\boldsymbol{\upsilon})\mathcal{D}^{-1} + I_{p^*}\|\mathtt{r} \leq \delta(\mathtt{r})\mathtt{r}.$$

For the remainder we use the definition of $C(\nabla)$. This gives

$$\|\mathcal{D}(\widetilde{\boldsymbol{\upsilon}} - \boldsymbol{\upsilon}^*) - \mathcal{D}^{-1}\nabla\mathcal{L}(\boldsymbol{\upsilon}^*)\| \leq \delta(\mathtt{r})\mathtt{r} + 6\omega\nu_1\left(2\mathtt{r}^2 + \mathfrak{z}_Q(\mathtt{x}, 4p^*)^2\right).$$

By the triangular inequality this implies

$$\widetilde{\boldsymbol{\upsilon}} \in \varUpsilon_\circ\left(\mathfrak{z}(\mathtt{x}, I\!\!B) + \diamondsuit_Q(\mathtt{r}^{(l)}, \mathtt{x}) \wedge \mathtt{r}_0\right).$$

For $\widetilde{\boldsymbol{\upsilon}}_{\boldsymbol{\theta}^*}$ we repeat the same arguments with the restriction to the set

$$\varUpsilon_{\circ,\boldsymbol{\theta}^*}(\mathtt{r}) \overset{\text{def}}{=} \{(\boldsymbol{\theta}, \boldsymbol{\eta}) \in \varUpsilon_\circ(\mathtt{r}) : \boldsymbol{\theta} = \boldsymbol{\theta}^*\}.$$

We bound on $\varUpsilon_{\circ,\boldsymbol{\theta}^*}(\mathtt{r})$

$$\begin{aligned}&\|\mathrm{H}^{-1}\big\{\nabla_{\boldsymbol{\eta}}\mathcal{L}(\boldsymbol{\upsilon})-\nabla_{\boldsymbol{\eta}}\mathcal{L}(\boldsymbol{\upsilon}^*)+\mathrm{H}^2\left(\boldsymbol{\eta}-\boldsymbol{\eta}^*\right)\big\}\|\\&\quad\le\|\mathrm{H}^{-1}\big\{\nabla_{\boldsymbol{\eta}}\mathcal{L}(\boldsymbol{\upsilon})-\nabla_{\boldsymbol{\eta}}\mathcal{L}(\boldsymbol{\upsilon}^*)\big\}\|\\&\qquad+\|\mathrm{H}^{-1}\big\{\nabla_{\boldsymbol{\eta}}I\!\!E\mathcal{L}(\boldsymbol{\upsilon})-\nabla_{\boldsymbol{\eta}}I\!\!E\mathcal{L}(\boldsymbol{\upsilon}^*)+\mathrm{H}^2\left(\boldsymbol{\eta}-\boldsymbol{\eta}^*\right)\big\}\|.\end{aligned}$$

Take any $\boldsymbol{\gamma}\in\mathbb{R}^m$ with $\|\boldsymbol{\gamma}\|=1$ then

$$\boldsymbol{\gamma}^\top\mathrm{H}^{-1}\nabla_{\boldsymbol{\eta}}\mathcal{L}(\boldsymbol{\upsilon})=(0,\mathrm{H}^{-1}\boldsymbol{\gamma})^\top\nabla_{\boldsymbol{\upsilon}}\mathcal{L}(\boldsymbol{\upsilon})=(0,\mathrm{H}^{-1}\boldsymbol{\gamma})^\top\mathcal{D}\mathcal{D}^{-1}\nabla_{\boldsymbol{\upsilon}}\mathcal{L}(\boldsymbol{\upsilon}).$$

Note that $\|\mathcal{D}(0,\mathrm{H}^{-1}\boldsymbol{\gamma})\|^2=\|\boldsymbol{\gamma}\|^2=1$ such that

$$\begin{aligned}\|\mathrm{H}^{-1}\big\{\nabla_{\boldsymbol{\eta}}\zeta(\boldsymbol{\upsilon})-\nabla_{\boldsymbol{\eta}}\zeta(\boldsymbol{\upsilon}^*)\big\}\|&=\sup_{\substack{\boldsymbol{\gamma}\in\mathbb{R}^m\\\|\boldsymbol{\gamma}\|=1}}\boldsymbol{\gamma}^\top\mathrm{H}^{-1}\left\{\nabla_{\boldsymbol{\eta}}\zeta(\boldsymbol{\upsilon})-\nabla_{\boldsymbol{\eta}}\zeta(\boldsymbol{\upsilon}^*)\right\}\\&\le\sup_{\substack{\boldsymbol{\gamma}\in\mathbb{R}^{p^*}\\\|\boldsymbol{\gamma}\|=1}}\boldsymbol{\gamma}^\top\mathcal{D}^{-1}\left\{\nabla_{\boldsymbol{\upsilon}}\zeta(\boldsymbol{\upsilon})-\nabla_{\boldsymbol{\upsilon}}\zeta(\boldsymbol{\upsilon}^*)\right\}\\&=\|\mathcal{D}^{-1}\left\{\nabla_{\boldsymbol{\upsilon}}\zeta(\boldsymbol{\upsilon})-\nabla_{\boldsymbol{\upsilon}}\zeta(\boldsymbol{\upsilon}^*)\right\}\|\\&\le 6\nu_1\omega_{\mathfrak{z}1}(\mathtt{x},4p^*)\mathtt{r}.\end{aligned}$$

As above we find with Taylor expansion

$$\begin{aligned}&\sup_{\boldsymbol{\upsilon}\in\varUpsilon_{\circ,\boldsymbol{\theta}^*}(\mathtt{r})}\|\mathrm{H}^{-1}\big\{\nabla_{\boldsymbol{\eta}}I\!\!E\mathcal{L}(\boldsymbol{\upsilon})-\nabla_{\boldsymbol{\eta}}I\!\!E\mathcal{L}(\boldsymbol{\upsilon}^*)+\mathrm{H}^2\left(\boldsymbol{\eta}-\boldsymbol{\eta}^*\right)\big\}\|\\&\quad\le\sup_{\boldsymbol{\upsilon}\in\varUpsilon_{\circ,\boldsymbol{\theta}^*}(\mathtt{r})}\|\mathrm{H}^{-1}\mathrm{H}^2(\boldsymbol{\upsilon})\mathrm{H}^{-1}-\boldsymbol{I}_m\|\mathtt{r}.\end{aligned}$$

We can bound using $\|\mathcal{D}(0,\mathrm{H}^{-1}\boldsymbol{\gamma})\|^2=\|\boldsymbol{\gamma}\|^2$ and $(\mathcal{L}_0)$

$$\begin{aligned}\|\mathrm{H}^{-1}\mathrm{H}^2(\boldsymbol{\upsilon})\mathrm{H}^{-1}-\boldsymbol{I}_m\|&=\sup_{\substack{\boldsymbol{\gamma}\in\mathbb{R}^m\\\|\boldsymbol{\gamma}\|=1}}\left(\mathrm{H}^{-1}\boldsymbol{\gamma}\right)^\top\left\{\mathrm{H}^2(\boldsymbol{\upsilon})-\mathrm{H}^2\right\}\mathrm{H}^{-1}\boldsymbol{\gamma}\\&=\sup_{\substack{\boldsymbol{\gamma}\in\mathbb{R}^m\\\|\boldsymbol{\gamma}\|=1}}\left(0,\mathrm{H}^{-1}\boldsymbol{\gamma}\right)^\top\left\{\mathcal{D}^2(\boldsymbol{\upsilon})-\mathcal{D}^2\right\}\left(0,\mathrm{H}^{-1}\boldsymbol{\gamma}\right)\\&\le\sup_{\substack{\boldsymbol{\gamma}\in\mathbb{R}^{p^*}\\\|\boldsymbol{\gamma}\|=1}}\boldsymbol{\gamma}\left\{\mathcal{D}^{-1}\mathcal{D}^2(\boldsymbol{\upsilon})\mathcal{D}^{-1}-\boldsymbol{I}_{p^*}\right\}\boldsymbol{\gamma}\le\delta(\mathtt{r}).\end{aligned}$$

This gives the claim. □

Lemma 4.A.4. *We have*

$$I\!\!P\left(C(\nabla)\cap\{\widetilde{\boldsymbol{\upsilon}},\widetilde{\boldsymbol{\upsilon}}_{\boldsymbol{\theta}^*}\in\varUpsilon_\circ(\mathtt{r}_0(\mathtt{x}))\}\right)\ge 1-4\mathrm{e}^{-\mathtt{x}}.$$

Proof. By definition of $\mathtt{r}_0(\mathtt{x})$

$$I\!\!P\left\{\widetilde{\boldsymbol{\upsilon}}, \widetilde{\boldsymbol{\upsilon}}_{\boldsymbol{\theta}^*} \in \varUpsilon_\circ(\mathtt{r}_0(\mathtt{x}))\right\} \geq 1 - \mathrm{e}^{-\mathtt{x}}.$$

Lemma 4.A.5 yields

$$\|\mathrm{H}^{-1}\nabla_{\boldsymbol{\eta}}\|^2 \leq \|\mathcal{D}^{-1}\nabla\|^2,$$

which implies that

$$\{\|\mathcal{D}^{-1}\nabla\| \leq \mathfrak{z}(\mathtt{x}, I\!\!B)\} \subseteq \{\|\mathrm{H}^{-1}\nabla_{\boldsymbol{\eta}}\| \leq \mathfrak{z}(\mathtt{x}, I\!\!B)\}.$$

To control the probability $I\!\!P\big(\|\mathcal{D}^{-1}\nabla\| > \mathfrak{z}(\mathtt{x}, I\!\!B)\big)$ we apply Proposition 3.4.1 with

$$I\!\!B = \mathcal{D}^{-1}\mathcal{V}_0^2\mathcal{D}^{-1}.$$

We obtain

$$I\!\!P\left(\|\mathcal{D}^{-1}\nabla\| > \mathfrak{z}(\mathtt{x}, I\!\!B)\right) \leq 2\mathrm{e}^{-\mathtt{x}}.$$

By Theorem 3.5.7 with $p = p^*$ we have

$$I\!\!P\left(\bigcap_{\mathtt{r}\leq \mathrm{R}_0(\mathtt{x})}\left\{\sup_{\boldsymbol{\upsilon}\in\varUpsilon_\circ(\mathtt{r})}\left\{\frac{1}{6\omega\nu_1}\|\mathcal{Y}(\boldsymbol{\upsilon})\| - 2\mathtt{r}^2\right\} \leq \mathfrak{z}_Q(\mathtt{x}, 4p^*)^2\right\}\right) \geq 1 - \mathrm{e}^{-\mathtt{x}}.$$

This gives that $I\!\!P(C'(\mathtt{r}_0, \mathtt{x})) \geq 1 - 4\mathrm{e}^{-\mathtt{x}}$. □

Proof. Now we can proof the claim of proposition 4.2.3. We proof this claim via induction. On

$$\varOmega(\mathtt{x}) \overset{\mathrm{def}}{=} C(\nabla) \cap \{\widetilde{\boldsymbol{\upsilon}}, \widetilde{\boldsymbol{\upsilon}}_{\boldsymbol{\theta}^*} \in \varUpsilon_\circ(\mathtt{r}_0(\mathtt{x}))\},$$

we have

$$\widetilde{\boldsymbol{\upsilon}}, \widetilde{\boldsymbol{\upsilon}}_{\boldsymbol{\theta}^*} \in \varUpsilon_\circ(\mathtt{r}_0), \quad \text{set } \mathtt{r}^{(0)} \overset{\mathrm{def}}{=} \mathtt{r}_0.$$

With Lemma 4.A.3 we find that

$$\varOmega(\mathtt{x}) \subseteq \left\{\boldsymbol{\upsilon}, \widetilde{\boldsymbol{\upsilon}}_{\boldsymbol{\theta}^*} \in \varUpsilon_\circ(\mathtt{r}^{(l)})\right\} \text{ implies } \varOmega(\mathtt{x}) \subseteq \left\{\widetilde{\boldsymbol{\upsilon}}, \widetilde{\boldsymbol{\upsilon}}_{\boldsymbol{\theta}^*} \in \varUpsilon_\circ(\mathtt{r}^{(l+1)})\right\},$$

where

$$\mathtt{r}^{(l)} \leq \mathfrak{z}(\mathtt{x}, I\!\!B) + \diamondsuit_Q\left(\mathtt{r}^{(l-1)}, \mathtt{x}\right).$$

Setting $l = 1$ this gives the first claim. For the second claim we show that

$$\varOmega(\mathtt{x}) \subseteq \left\{ \widetilde{\boldsymbol{\upsilon}}, \widetilde{\boldsymbol{\upsilon}}_{\boldsymbol{\theta}^*} \in \varUpsilon_\circ \left(\limsup_{l\to\infty} \mathtt{r}^{(l)} \right) \right\} \subseteq \{ \widetilde{\boldsymbol{\upsilon}}, \widetilde{\boldsymbol{\upsilon}}_{\boldsymbol{\theta}^*} \in \varUpsilon_\circ(\mathtt{r}_0^*) \} .$$

Consequently we have to show that $\limsup_{l\to\infty} \mathtt{r}^{(l)} \le \mathtt{r}_0^*$ with $\mathtt{r}_0^*$ defined in (4.2.12). For this we use $\delta(\mathtt{r})/\mathtt{r} \vee 4\nu_1\omega \le \epsilon$ to estimate further

$$\begin{aligned} \mathtt{r}^{(l)} &\le \mathfrak{z}(\mathtt{x}, \mathbb{B}) + \diamondsuit_Q \left(\mathtt{r}^{(l-1)}, \mathtt{x} \right) \\ &\le \mathfrak{z}(\mathtt{x}, \mathbb{B}) + \epsilon \left[\mathtt{r}^{(l-1)^2} + \mathfrak{z}_Q(\mathtt{x}, 4p^*)^2 \right] \\ &\le \mathfrak{z}(\mathtt{x}, \mathbb{B}) + \epsilon \mathfrak{z}_Q(\mathtt{x}, 4p^*)^2 + \epsilon \mathtt{r}^{(l-1)^2}, \end{aligned}$$

such that

$$\mathtt{r}^{(l)^2} \le 2 \left(\mathfrak{z}(\mathtt{x}, \mathbb{B}) + \epsilon \mathfrak{z}_Q(\mathtt{x}, 4p^*)^2 \right)^2 + 2\epsilon^2 \mathtt{r}^{(l-1)^4}. \tag{4.A.5}$$

Abbreviate $\mathfrak{z}_\epsilon(\mathtt{x}) \overset{\text{def}}{=} \left(\mathfrak{z}(\mathtt{x}, \mathbb{B}) + \epsilon \mathfrak{z}_Q(\mathtt{x}, 4p^*)^2 \right)^2$. We will show via induction that

$$\begin{aligned} \mathtt{r}^{(l)^2} \le{}& 2 \sum_{s=0}^{r} 4^{\sum_{k=0}^{s} 2^k} \epsilon^{-\sum_{k=0}^{s-1} 2^{k+1}} \mathfrak{z}_\epsilon(\mathtt{x})^{2^{s+1}} \\ &+ 2(4)^{\sum_{k=0}^{r} 2^k} \epsilon^{\sum_{k=0}^{r} 2^{k+1}} \mathtt{r}^{(l-r)^{2^{r+1}}}. \end{aligned} \tag{4.A.6}$$

Equation (4.A.5) already serves the claim for $r = 1$. Assume that the claim is shown for $r \in \mathbb{N}$ then we plug (4.A.5) into (4.A.6) to find

$$\begin{aligned} \mathtt{r}^{(l)^2} \le{}& 2 \sum_{s=0}^{r} 4^{\sum_{k=0}^{s} 2^k} \epsilon^{\sum_{k=0}^{s-1} 2^{k+1}} \mathfrak{z}_\epsilon(\mathtt{x})^{2^{s+1}} \\ &+ 2(4)^{\sum_{k=0}^{r} 2^k} \epsilon^{\sum_{k=0}^{r} 2^{k+1}} \left(2\mathfrak{z}_\epsilon(\mathtt{x})^2 + 2\epsilon^2 \mathtt{r}^{(l-r-1)^4} \right)^{2^{r+1}} \\ \le{}& 2 \sum_{s=0}^{r} 4^{\sum_{k=0}^{r} 2^k} \epsilon^{\sum_{k=0}^{s-1} 2^{k+1}} \mathfrak{z}_\epsilon(\mathtt{x})^{2^{s+1}} \\ &+ 2(4)^{\sum_{k=0}^{r} 2^k} \epsilon^{\sum_{k=0}^{r} 2^{k+1}} \left(4^{2^{r+1}} \mathfrak{z}_\epsilon(\mathtt{x})^{2^{r+2}} + 4^{2^{r+1}} \epsilon^{2^{r+2}} \mathtt{r}^{(l-r-1)^{2^{r+2}}} \right) \\ \le{}& 2 \sum_{s=0}^{r+1} 4^{\sum_{k=0}^{s} 2^k} \epsilon^{\sum_{k=0}^{s-1} 2^{k+1}} \mathfrak{z}_\epsilon(\mathtt{x})^{2^{s+1}} + 2(4)^{2^{r+1}} \epsilon^{2^{r+1}} \mathtt{r}^{(l-r-1)^{2^{r+2}}}, \end{aligned}$$

which gives (4.A.6) for all $r \le l$. We can bound

$$\begin{aligned} \sum_{s=0}^{r+1} 4^{\sum_{k=0}^{s} 2^k} \epsilon^{\sum_{k=0}^{s-1} 2^{k+1}} \mathfrak{z}_\epsilon(\mathtt{x})^{2^{s+1}} &\le 4 \sum_{s=0}^{r+1} (4\epsilon)^{\sum_{k=1}^{s} 2^k} \mathfrak{z}_\epsilon(\mathtt{x})^{2^{s+1}} \\ &\le 4\mathfrak{z}_\epsilon(\mathtt{x}) \sum_{s=0}^{r+1} \{4\epsilon\mathfrak{z}_\epsilon(\mathtt{x})\}^{2^{s+1}-1}. \end{aligned}$$

Using that $4\epsilon\mathfrak{z}_\epsilon(\mathtt{x}) < c$ we find

$$2\sum_{s=0}^{l} 4^{\sum_{k=0}^{s} 2^k} \epsilon^{\sum_{k=0}^{s-1} 2^{k+1}} \mathfrak{z}_\epsilon(\mathtt{x})^{2^{s+1}} \le \frac{8}{1-c}\mathfrak{z}_\epsilon(\mathtt{x}).$$

Clearly because $4\epsilon\mathtt{r}_0(\mathtt{x}) < 1$

$$(4)^{2^{l-1}} \epsilon^{2^{l-1}} {\mathtt{r}_0}^{2^l} = \mathtt{r}_0 (4\epsilon\mathtt{r}_0)^{2^{l-1}} \to 0.$$

Consequently

$$\begin{aligned} \limsup_{l\to\infty} \mathtt{r}^{(l)} &\le \mathfrak{z}(\mathtt{x}, \mathbb{B}) + \epsilon\mathfrak{z}_Q(\mathtt{x}, 4p^*)^2 + \limsup_{l\to\infty} \epsilon {\mathtt{r}^{(l-1)}}^2 \\ &\le \mathfrak{z}(\mathtt{x}, \mathbb{B}) + \epsilon\mathfrak{z}_Q(\mathtt{x}, 4p^*)^2 + \epsilon^2 \frac{8}{1-c}\mathfrak{z}_\epsilon(\mathtt{x}). \end{aligned}$$

Again using $4\epsilon\mathfrak{z}_\epsilon(\mathtt{x}) < c$ gives the claim. □

Lemma 4.A.5. *Let $\mathcal{D} \in \mathbb{R}^{(p+p)\times(p+p)}$ be invertible and*

$$\mathcal{D}^2 = \begin{pmatrix} D^2 & A \\ A^\top & H^2 \end{pmatrix} \in \mathbb{R}^{(p+p)\times(p+p)}, \qquad D \in \mathbb{R}^{p\times p},\ H \in \mathbb{R}^{m\times m} \text{ invertible,.}$$

Then for any $\boldsymbol{\upsilon} = (\boldsymbol{\theta}, \boldsymbol{\eta}) \in \mathbb{R}^{p+m}$ we have $\|H^{-1}\boldsymbol{\eta}\| \vee \|D^{-1}\boldsymbol{\theta}\| \le \|\mathcal{D}^{-1}\boldsymbol{\upsilon}\|$.

Proof. With $\boldsymbol{\upsilon} = (\boldsymbol{\theta}, \boldsymbol{\eta}) \in \mathbb{R}^{p+m}$

$$\|D^{-1}\boldsymbol{\theta}\| = \|D^{-1}\Pi_{\boldsymbol{\theta}}\mathcal{D}\mathcal{D}^{-1}\boldsymbol{\upsilon}\| \le \|D^{-1}\Pi_{\boldsymbol{\theta}}\mathcal{D}\|\|\mathcal{D}^{-1}\boldsymbol{\upsilon}\| \le \|\mathcal{D}^{-1}\boldsymbol{\upsilon}\|,$$

because

$$\|D^{-1}\Pi_{\boldsymbol{\theta}}\mathcal{D}\|^2 = \sup_{\|\boldsymbol{\gamma}\|=1} \boldsymbol{\gamma}^\top D^{-1}\Pi_{\boldsymbol{\theta}}\mathcal{D}^2\Pi_{\boldsymbol{\theta}}^\top D^{-1}\boldsymbol{\gamma} = \|\boldsymbol{\gamma}\| = 1.$$

The same argument works for $\|\mathrm{H}^{-1}\boldsymbol{\eta}\|$. □

Remark 4.A.2. To address the claim of remark 4.2.15 note that we have $C'(\mathtt{r}_0, \mathtt{x}) \subseteq \{\|(1-\epsilon(\mathtt{r}_0))^{-1}\mathcal{D}_0^{-1}\nabla\| \le (1-\epsilon(\mathtt{r}_0))^{-1}\mathfrak{z}(\mathtt{x}, I\!B) \le \mathtt{r}_0\}$ by the choice of $\mathtt{r}_0 > 0$ such that by Corollary 3.2.2

$$\|(1-\epsilon(\mathtt{r}_0))\mathcal{D}_0(\widetilde{\boldsymbol{\upsilon}} - \boldsymbol{\upsilon}^*) - (1-\epsilon(\mathtt{r}_0))^{-1}\mathcal{D}_0^{-1}\nabla\| \le \sqrt{2\varDelta_{\boldsymbol{\epsilon}}(\mathtt{r}_0)}.$$

Consequently

$$\|\mathcal{D}_0(\widetilde{\boldsymbol{\upsilon}} - \boldsymbol{\upsilon}^*)\| \le (1-\epsilon(\mathtt{r}_0))^{-2}\|\mathcal{D}_0^{-1}\nabla\| + \sqrt{2\varDelta_{\boldsymbol{\epsilon}}(\mathtt{r}_0)}.$$

The same can be done for $\widetilde{\boldsymbol{\eta}}_{\boldsymbol{\theta}^*}$ which gives the claim.

4.A.4 Proof of Lemma 4.2.4

Proof. First note that due to (ι) we have

$$\|\mathcal{D}^{-1}\| = \frac{1}{\sqrt{n}}\|d^{-1}\| \le \frac{1}{\sqrt{n}c_d}. \tag{4.A.7}$$

Now we prove the implications.

$(\mathcal{L}_0)$ As by assumption $\varUpsilon_\circ(\mathbf{r}^*) \subset U$ we simply estimate using (4.A.7) and (ℓ_0) for any $\boldsymbol{\upsilon} \in \varUpsilon_\circ(\mathbf{r}^*)$

$$\begin{aligned}\|I - \mathcal{D}^{-1}\nabla^2 I\!E\mathcal{L}(\boldsymbol{\upsilon})\mathcal{D}^{-1}\| &\le \frac{1}{nc_d^2}\|\mathcal{D}^2 - \nabla^2 I\!E\mathcal{L}(\boldsymbol{\upsilon})\| \\ &= \frac{1}{c_d^2}\|\nabla^2 I\!E\ell(\boldsymbol{\upsilon}^*) - \nabla^2 I\!E\ell(\boldsymbol{\upsilon})\| \le \frac{\delta^*}{\sqrt{n}c_d^3}\mathtt{r}.\end{aligned}$$

$(\mathcal{ED}_1)$ Abbreviate $\zeta_i = (\ell_i - I\!E\ell_i)$ and $\zeta = (\mathcal{L} - I\!E\mathcal{L})$. Take any $\boldsymbol{\gamma} \in \mathbb{R}^{p^*}$ and $\boldsymbol{\upsilon}, \boldsymbol{\upsilon}' \in \varUpsilon_\circ(\mathbf{r}^*) \subset U$ and use the mean value theorem to find some $\widehat{\boldsymbol{\upsilon}} \in \operatorname{conv}(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}') \subset U$

$$\begin{aligned}&\log I\!E\exp\left\{\frac{\mu\boldsymbol{\gamma}^\top\mathcal{D}_0^{-1}\{\nabla\zeta(\boldsymbol{\upsilon}) - \nabla\zeta(\boldsymbol{\upsilon}')\}}{\omega\,\|\mathcal{D}_0(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}')\|}\right\} \\ &\quad= \log I\!E\exp\left\{\frac{\mu}{\omega n}\boldsymbol{\gamma}^\top d^{-1}\left\{\sum_{i=1}^n \nabla^2\zeta_i(\widehat{\boldsymbol{\upsilon}})\right\}d^{-1}\frac{d(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}')}{\|d(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}')\|}\right\}.\end{aligned}$$

Using independence and (ed_1) this gives with $\omega = \frac{1}{\sqrt{n}}$ and $|\mu| \le$

$\sqrt{n}\mathrm{g}_0^*$

$$
\begin{aligned}
&\log I\!\!E \exp\left\{\frac{\mu\, \gamma^\top \mathcal{D}_0^{-1}\{\nabla\zeta(\boldsymbol{\upsilon}) - \nabla\zeta(\boldsymbol{\upsilon}')\}}{\omega\, \|\mathcal{D}_0(\boldsymbol{\upsilon}-\boldsymbol{\upsilon}')\|}\right\} \\
&\quad \le \sum_{i=1}^{n} \sup_{\substack{\gamma\in\mathbb{R}^{p^*} \\ \|\gamma\|=1}} \log I\!\!E \exp\left\{\frac{\mu}{\sqrt{n}}\gamma_1^\top d^{-1}\nabla^2\zeta_i(\widehat{\boldsymbol{\upsilon}})d^{-1}\gamma_2\right\} \le \nu_0^*\mu^2/2.
\end{aligned}
$$

Furthermore $(\mathcal{I})$ is a consequence of Lemma 4.A.6 and (ι). The other claims can be shown with the same argument or follow trivially from the setting. □

Lemma 4.A.6. *For a positive definite symmetric matrix*

$$
\mathcal{D}^2 = \begin{pmatrix} D^2 & A \\ A^\top & H^2 \end{pmatrix},
$$

with $c_{\mathcal{D}}\|\boldsymbol{\upsilon}\|^2 \le \boldsymbol{\upsilon}^\top\mathcal{D}\boldsymbol{\upsilon}$ *for some* $c_{\mathcal{D}} > 0$ *we have that*

$$
\|D^{-1}AH^{-2}A^\top D^{-1}\| =: \nu^2 \le 1 - \frac{c_{\mathcal{D}}}{\|D\|^2 \wedge \|H\|^2}.
$$

Proof. For any $\boldsymbol{\upsilon} = (\boldsymbol{\theta},\boldsymbol{\eta}) \in \mathbb{R}^{p+m}$ we have

$$
\begin{aligned}
\boldsymbol{\upsilon}^\top\mathcal{D}^2\boldsymbol{\upsilon} &= (\boldsymbol{\theta}^\top,\boldsymbol{\eta}^\top)\begin{pmatrix} D^2 & A \\ A^\top & H^2 \end{pmatrix}\begin{pmatrix} \boldsymbol{\theta} \\ \boldsymbol{\eta} \end{pmatrix} \\
&= (\boldsymbol{\theta}^\top D^\top,\boldsymbol{\eta}^\top H^\top)\begin{pmatrix} I_p & D^{-1}AH^{-1} \\ H^{-1}A^\top D^{-1} & I_m \end{pmatrix}\begin{pmatrix} D\boldsymbol{\theta} \\ H\boldsymbol{\eta} \end{pmatrix} \\
&= \|D\boldsymbol{\theta}\|^2 + \|H\boldsymbol{\eta}\|^2 + 2\langle H\boldsymbol{\eta}, H^{-1}A^\top D^{-1}\boldsymbol{\theta}\rangle.
\end{aligned}
$$

Minimized with respect to $\boldsymbol{\eta}$, i.e. with $H\boldsymbol{\eta} = -H^{-1}A^\top D^{-1}D\boldsymbol{\theta}$ we find

$$
\begin{aligned}
\boldsymbol{\upsilon}^\top\mathcal{D}^2\boldsymbol{\upsilon} &= \|D\boldsymbol{\theta}\|^2 - \|H^{-1}A^\top D^{-1}D\boldsymbol{\theta}\|^2 \\
&= (D\boldsymbol{\theta})^\top(I_p - D^{-1}AH^{-2}A^\top D^{-1})D\boldsymbol{\theta},
\end{aligned}
$$

which gets minimal - i.e. equal to $(1-\nu^2)\|D\boldsymbol{\theta}\|$ - if

$$
D^{-1}AH^{-2}A^\top D^{-1}D\boldsymbol{\theta} = \|D^{-1}AH^{-2}A^\top D^{-1}\|D\boldsymbol{\theta} = \nu^2 D\boldsymbol{\theta},
$$

i.e. if $D\boldsymbol{\theta} \in \mathbb{R}^p$ is a maximal eigenvalue of $D^{-1}AH^{-2}A^\top D^{-1} \in \mathbb{R}^{p\times p}$. With the assumption $c_{\mathcal{D}}\|\boldsymbol{\upsilon}\|^2 \le \boldsymbol{\upsilon}^\top\mathcal{D}\boldsymbol{\upsilon}$ this gives

$$
c_{\mathcal{D}}\|\boldsymbol{\upsilon}\|^2 \le \boldsymbol{\upsilon}^\top\mathcal{D}^2\boldsymbol{\upsilon} = (1-\nu^2)\|D\boldsymbol{\theta}\|^2, \quad \|\boldsymbol{\upsilon}\|^2 = \|\boldsymbol{\theta}\|^2 + \|H^{-2}A^\top\boldsymbol{\theta}\|^2,
$$

such that

$$\nu^2 \le 1 - c_{\mathfrak{D}} \frac{\|\boldsymbol{\theta}\|^2}{\|D\boldsymbol{\theta}\|^2} \le 1 - \frac{c_{\mathfrak{D}}}{\|D\|^2}.$$

With analogous arguments we can obtain

$$\nu^2 \le 1 - c_{\mathfrak{D}} \frac{\|\boldsymbol{\eta}\|^2}{\|H\boldsymbol{\eta}\|^2} \le 1 - \frac{c_{\mathfrak{D}}}{\|H\|^2}.$$

This completes the proof. □

4.A.5 Proof of Proposition 4.2.6

The profile MLE can be calculated easily

$$\widetilde{\boldsymbol{\theta}} = \Pi_{\boldsymbol{\theta}} f^{-1}(\boldsymbol{Y}) = \Pi_{\boldsymbol{\theta}} f^{-1}(f(\boldsymbol{\upsilon}^*) + \boldsymbol{\varepsilon}_i) = \boldsymbol{\theta}^* + \varepsilon_{\boldsymbol{\theta}} - \|\boldsymbol{\varepsilon}_{\boldsymbol{\eta}}\|^2,$$

where $\varepsilon = (\varepsilon_{\boldsymbol{\theta}}, \boldsymbol{\varepsilon}_{\boldsymbol{\eta}}) \in \mathbb{R} \times \mathbb{R}^{p_n-1}$. It is straight forward to show, that the conditions of Section 4.2.1 are satisfied with $\mathcal{D}_0^2 = n\mathbb{E}[\nabla f \nabla f^\top(\boldsymbol{\upsilon}^*)] = Id_{p^*}$, $\breve{D}^2 = n$ and $\breve{\boldsymbol{\xi}} = \sqrt{n}\varepsilon_{\boldsymbol{\theta}}$. But we immediately see that

$$\sqrt{n}(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) - \sqrt{n}\varepsilon_{\boldsymbol{\theta}} = -\sqrt{n}\|\boldsymbol{\varepsilon}_{\boldsymbol{\eta}}\|^2 \sim \frac{-\chi^2_{p_n-1}}{\sqrt{n}}.$$

This means that if $p_n = O(n^{1/2})$ the estimator is not root-n consistent. For $\sqrt{n} = o(p_n)$ the root-n bias goes to infinity almost surely. Clearly if $p_n = o(n^{1/2})$ the Fisher expansion is accurate.

Concerning the Wilks phenomenon note that $\mathcal{L}(\widetilde{\boldsymbol{\upsilon}}) = 0$. On the other hand, with probability tending to one as $n \to \infty$

$$\begin{aligned} -\max_{\boldsymbol{\eta}\in\mathbb{R}^{p_n-1}} \mathcal{L}(\boldsymbol{\theta}^*, \boldsymbol{\eta}) &= n \min_{\lambda\in\mathbb{R}} \left\{ \left(y_{\boldsymbol{\theta}} - \lambda^2 \|\boldsymbol{Y}_{\boldsymbol{\eta}}\|^2\right)^2 + (1-\lambda)^2 \|\boldsymbol{Y}_{\boldsymbol{\eta}}\|^2 \right\} \\ &= n \min_{\lambda\in\mathbb{R}} \left\{ \left(\varepsilon_{\boldsymbol{\theta}} - \lambda^2 \|\boldsymbol{\varepsilon}_{\boldsymbol{\eta}}\|^2\right)^2 + (1-\lambda)^2 \|\boldsymbol{\varepsilon}_{\boldsymbol{\eta}}\|^2 \right\} \\ &= n \min_{\lambda\in\mathbb{R}} \left\{ \varepsilon_{\boldsymbol{\theta}}^2 + \|\boldsymbol{\varepsilon}_{\boldsymbol{\eta}}\|^2 \left(\lambda^4 \|\boldsymbol{\varepsilon}_{\boldsymbol{\eta}}\|^2 - \lambda^2 \varepsilon_{\boldsymbol{\theta}} + (1-\lambda)^2\right) \right\}, \end{aligned}$$

where $\boldsymbol{Y} = (y_{\boldsymbol{\theta}}, \boldsymbol{Y}_{\boldsymbol{\eta}}) \in \mathbb{R} \times \mathbb{R}^{p_n-1}$ and $\boldsymbol{\varepsilon} = (\varepsilon_{\boldsymbol{\theta}}, \boldsymbol{\varepsilon}_{\boldsymbol{\eta}}) \in \mathbb{R} \times \mathbb{R}^{p_n-1}$. Clearly $\|\boldsymbol{\varepsilon}_{\boldsymbol{\eta}}\|^2 = O(p_n/n) \to 0$ a.s. and $\varepsilon_{\boldsymbol{\theta}} \to 0$ a.s. such that the sequence of minimizers satisfies $\lambda_n \to 1$ a.s.. This gives for any $\tau > 0$ and $n \ge n_\tau \in \mathbb{N}$ large enough

$$-\max_{\boldsymbol{\eta}} \mathcal{L}(\boldsymbol{\theta}^*, \boldsymbol{\eta}) \ge n\varepsilon_{\boldsymbol{\theta}}^2 + (1-\tau)n \|\boldsymbol{\varepsilon}_{\boldsymbol{\eta}}\|^4 - (1+\tau)n\varepsilon_{\boldsymbol{\theta}} \|\boldsymbol{\varepsilon}_{\boldsymbol{\eta}}\|^2. \quad (4.A.8)$$

Furthermore we get setting $\lambda = 1$

$$-\max_{\boldsymbol{\eta}} \mathcal{L}(\boldsymbol{\theta}^*, \boldsymbol{\eta}) \le n\varepsilon_{\boldsymbol{\theta}}^2 + n \left\| \boldsymbol{\varepsilon}_{\boldsymbol{\eta}} \right\|^4 - n\varepsilon_{\boldsymbol{\theta}} \left\| \boldsymbol{\varepsilon}_{\boldsymbol{\eta}} \right\|^2 . \tag{4.A.9}$$

As $\breve{L}(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}^*) = -\max_{\boldsymbol{\eta}} \mathcal{L}(\boldsymbol{\theta}^*, \boldsymbol{\eta})$ the inequalities (4.A.8) and (4.A.9) combine to

$$\begin{aligned} n\varepsilon_{\boldsymbol{\theta}}^2 + n(1-\tau) \left\| \boldsymbol{\varepsilon}_{\boldsymbol{\eta}} \right\|^4 - (1+\tau) n\varepsilon_{\boldsymbol{\theta}} \left\| \boldsymbol{\varepsilon}_{\boldsymbol{\eta}} \right\|^2 &\le \breve{L}(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}^*) \\ &\le n\varepsilon_{\boldsymbol{\theta}}^2 + n \left\| \boldsymbol{\varepsilon}_{\boldsymbol{\eta}} \right\|^4 - n\varepsilon_{\boldsymbol{\theta}} \left\| \boldsymbol{\varepsilon}_{\boldsymbol{\eta}} \right\|^2 . \end{aligned}$$

This gives the Wilks phenomenon if $p_n^2/n \to 0$. If $p_n^2/n \to \infty$ the right-hand side in (4.A.8) diverges since with $\tau = 1/2$

$$\begin{aligned} & n\varepsilon_{\boldsymbol{\theta}}^2 + \frac{n}{2} \left\| \boldsymbol{\varepsilon}_{\boldsymbol{\eta}} \right\|^4 - 2n\varepsilon_{\boldsymbol{\theta}} \left\| \boldsymbol{\varepsilon}_{\boldsymbol{\eta}} \right\|^2 \\ & \quad \sim \chi_1^2 * (\chi_{p_n-1}^4/2n) * \left\{ -\mathcal{N}(0,1)(2\chi_{p_n-1}^2/\sqrt{n}) \right\} \xrightarrow{w} \delta_\infty . \end{aligned}$$

If $p_n^2/n \to \mathtt{C}$ then $\breve{L}(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}^*)$ can not converge to a χ^2-distribution with one degree of freedom as one can let $\tau > 0$ tend 0. This completes the proof.

4.A.6 Proof of Proposition 4.2.7

We only sketch the proof of the first claim as it is rather uninteresting. Note that

$$I\!\!P(n\|\boldsymbol{Y}\|^2 \ge 4p_n) \to 0,$$

which implies that $\widetilde{\boldsymbol{\upsilon}} \in \varUpsilon_\circ(2\sqrt{p_n})$. On $\varUpsilon_\circ(2\sqrt{p_n})$

$$\begin{aligned} n\boldsymbol{\upsilon}^\top \boldsymbol{Y} - n\|\boldsymbol{\upsilon}\|^2/2 - \sqrt{\beta_n} &\le \mathcal{L}(\boldsymbol{\upsilon}) \\ &\le n\boldsymbol{\upsilon}^\top \boldsymbol{Y} - n\|\boldsymbol{\upsilon}\|^2/2 + \sqrt{\beta_n}. \end{aligned} \tag{4.A.10}$$

Maximizing on the left-hand side of (4.A.10) and plugging in $\widetilde{\boldsymbol{\upsilon}}$ on the right-hand side we get

$$\|\mathcal{D}(\widetilde{\boldsymbol{\upsilon}} - \boldsymbol{Y})\|^2/2 = n\|\boldsymbol{Y}\|^2/2 - n\widetilde{\boldsymbol{\upsilon}}^\top \boldsymbol{Y} + n\|\widetilde{\boldsymbol{\upsilon}}\|^2/2 \le 2\sqrt{\beta_n}.$$

This gives the claim:

$$\|\breve{D}_0(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) - \breve{\boldsymbol{\xi}}\|^2 \le \|\mathcal{D}(\widetilde{\boldsymbol{\upsilon}} - \boldsymbol{Y})\|^2 \le 2\sqrt{\beta_n} \to 0.$$

For the other claims we first show that for n large enough, the MLE $\widetilde{\boldsymbol{\upsilon}} \in \mathbb{R}^{p_n}$ belongs with probability close to one to the $\delta = 1/n$ vicinity $\mathcal{S}_\delta$ of the set $\mathcal{S}$ in (4.2.16). The second step is to show that with a probability exceeding a fixed constant $\alpha > 0$, the profile MLE $\widetilde{\theta}$ differs significantly from y_1 which is the profile MLE in the linear Gaussian model. The third step focuses on the case $\beta_n \to \infty$.

1. First we show that for n large enough, the MLE $\widetilde{\boldsymbol{\upsilon}} \in \mathbb{R}^{p_n}$ lies in $\mathcal{S}_\delta$ with probability close to one. For this we check that the maximum of $\mathcal{L}(\boldsymbol{\upsilon})$ on $\mathcal{S}_\delta^c$ is smaller than a similar maximum on $\mathcal{S}$ for "typical" values of $\boldsymbol{Y}$ and n large enough. Indeed, for any point $\boldsymbol{\upsilon} \in \mathcal{S}_\delta^c$

$$\begin{aligned}\mathcal{L}(\boldsymbol{\upsilon}, 0) &\le \max_{\boldsymbol{\upsilon} \in \mathcal{S}_\delta^c} \mathcal{L}(\boldsymbol{\upsilon}, 0) = \max_{\boldsymbol{\upsilon} \in \mathcal{S}_\delta^c} \bigl\{ n\boldsymbol{Y}^\top \boldsymbol{\upsilon} - n\|\boldsymbol{\upsilon}\|^2/2 \bigr\} \\ &\le \max_{\boldsymbol{\upsilon} \in \mathbb{R}^{p_n}} \bigl\{ n\boldsymbol{Y}^\top \boldsymbol{\upsilon} - n\|\boldsymbol{\upsilon}\|^2/2 \bigr\} = \frac{n}{2}\|\boldsymbol{Y}\|^2.\end{aligned}$$

Furthermore, introduce a set of "typical" values $\boldsymbol{Y}$:

$$C_1 \stackrel{\mathrm{def}}{=} \left\{ \boldsymbol{Y} : \frac{1}{2}\left(\frac{p_n}{n}\right)^{3/2} < \|\boldsymbol{Y}\|^3 < \left(\frac{2p_n}{n}\right)^{3/2}, \text{ and } |y_1| \le 1 \right\}.$$

It is straightforward to see that $I\!P\bigl(\boldsymbol{Y} \in C_1\bigr)$ is exponentially close to one for n large. Below we assume that $\boldsymbol{Y} \in C_1$ and study the value $\mathcal{L}(\boldsymbol{\upsilon}, 0)$ for $\boldsymbol{\upsilon} \in \mathcal{S}$. By assumption n is large enough to ensure that

$$\frac{2^{1/3}-1}{2^{1/6}}\left(\frac{p_n}{n}\right)^{1/2} \ge \frac{1}{2}\left(\frac{p_n}{n}\right)^{3/4} = \frac{1}{2}\sqrt{\beta_n/n}. \tag{4.A.11}$$

Introduce $\boldsymbol{Y}_{\mathcal{S}}$ as the closest point in $\mathcal{S}$ to $\boldsymbol{Y}$ with $|\upsilon_1| \ge |y_1|$. This point always exists by the definition of $\mathcal{S}$. Denote

$$\delta(\boldsymbol{Y}) = \|\boldsymbol{Y} - \boldsymbol{Y}_{\mathcal{S}}\| = |y_1 - \upsilon_1|.$$

By construction of $\mathcal{S}$, it holds $\delta(\boldsymbol{Y}) \le 0.5\sqrt{\beta_n/n}$ for $\boldsymbol{Y} \in C_1$. For n satisfying (4.A.11) this also yields $\bigl[\|\boldsymbol{Y}\| - \delta(\boldsymbol{Y})\bigr]^3 \ge 1/2\|\boldsymbol{Y}\|^3$. We have for $\boldsymbol{Y} \in C_1$

$$\begin{aligned}\max_{\boldsymbol{\upsilon} \in \mathcal{S}} \mathcal{L}(\boldsymbol{\upsilon}, 0) &\ge \mathcal{L}(\boldsymbol{Y}_{\mathcal{S}}, 0) \\ &\ge n\|\boldsymbol{Y}\|^2 - n|y_1|\delta(\boldsymbol{Y}) - \frac{n}{2}\bigl\{\|\boldsymbol{Y}\|^2 - 2|y_1|\delta(\boldsymbol{Y}) + \delta^2(\boldsymbol{Y})\bigr\} \\ &\qquad + n\bigl\{\|\boldsymbol{Y}\|^2 - 2|y_1|\delta(\boldsymbol{Y}) + \delta^2(\boldsymbol{Y})\bigr\}^{3/2} \\ &\ge \frac{n}{2}\|\boldsymbol{Y}\|^2 - n\delta^2(\boldsymbol{Y}) + n\bigl\{\|\boldsymbol{Y}\| - \delta(\boldsymbol{Y})\bigr\}^3 \\ &> \frac{n}{2}\|\boldsymbol{Y}\|^2 - \frac{\beta_n}{4} + \frac{n}{2}\|\boldsymbol{Y}\|^3 > \frac{n}{2}\|\boldsymbol{Y}\|^2 \ge \max_{\boldsymbol{\upsilon} \in \mathcal{S}_\delta^c} \mathcal{L}(\boldsymbol{\upsilon}, 0).\end{aligned}$$

This implies $\widetilde{\boldsymbol{\upsilon}} \in \mathcal{S}_\delta$.

2. Now we discuss the case when $\beta_n^2 = p_n^3/n \to (6c)^4$ for some $c \geq 0$ and show that the profile MLE $\widetilde{\boldsymbol{\theta}}$ deviates significantly from y_1 on a set of positive probability. Define for each $n \in \mathbb{N}$

$$C_n \stackrel{\text{def}}{=} C_1 \cap \left\{ \|\boldsymbol{Y} - \boldsymbol{Y}_{\mathbb{S}}\| \geq \frac{1}{6}\sqrt{\beta_n/n} \right\} = C_1 \cap \left\{ |y_1 - y_{\mathbb{S},1}| \geq \frac{1}{6}\sqrt{\beta_n/n} \right\}.$$

It is easy to see that $I\!\!P(C_n) \geq \alpha$ for some fixed $\alpha > 0$ and all n. It remains to note that on the set C_n it holds under (4.A.11)

$$\begin{aligned} \|\breve{D}_0(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) - \breve{\boldsymbol{\xi}}\| &= \sqrt{n}|\widetilde{v}_1 - y_1| \\ &\geq \sqrt{n}|y_1 - y_{\mathbb{S},1}| - \sqrt{n}/n \\ &\geq \frac{1}{6}\beta_n^{1/2} - \frac{1}{\sqrt{n}} \to \begin{cases} \infty & p_n^3/n \to \infty, \\ c & p_n^3/n \to (6c)^4, \end{cases} \end{aligned}$$

which yields the claim.

3. Finally consider the case when $\beta_n \to \infty$. Take the sequence $c_n = \beta_n^{-1/4} \to 0$ .Consider the set

$$C_n \stackrel{\text{def}}{=} C_1 \cap \left\{ \|\boldsymbol{Y} - \boldsymbol{Y}_{\mathbb{S}}\| \geq \frac{c_n}{6}\sqrt{\beta_n/n} \right\} = C_1 \cap \left\{ |y_1 - y_{\mathbb{S},1}| \geq \frac{1}{6}\sqrt{\beta_n^{1/2}/n} \right\}.$$

Then $I\!\!P(C_n) \to 1$ and on C_n

$$\|\breve{D}_0(\widetilde{\theta} - \theta^*) - \breve{\boldsymbol{\xi}}\| \geq \frac{c_n}{6}\beta_n^{1/2} - \frac{1}{\sqrt{n}} = \frac{1}{6}\beta_n^{1/4} - \frac{1}{\sqrt{n}} \to \infty,$$

as required.

4.A.7 Proof of Theorem 4.2.8

Since by assumption $p_n^2/n \to 0$ and the support of $f(\boldsymbol{v})$ is contained in $B_{2\sqrt{p_n}/\sqrt{n}}(0)$ by the choice of $\widehat{K}$ it holds for n large enough and for any $\boldsymbol{v}$ with $\|\boldsymbol{v}\|^2 \leq 4p_n/n$ that $n\|\boldsymbol{v}\|^2/2 \geq nf(\boldsymbol{v})\|\boldsymbol{v}\|^3$ and thus

$$\operatorname*{argmax}_{\boldsymbol{v}} I\!\!E\mathcal{L}(\boldsymbol{v}) = \operatorname*{argmin}_{\boldsymbol{v}}\left\{n\|\boldsymbol{v}\|^2/2 - nf(\boldsymbol{v})\|\boldsymbol{v}\|^3/3\right\} = 0 \in \mathbb{R}^{p_n}.$$

Apart from $(\mathcal{L}_0)$ it is easy to see that all conditions are satisfied with $\mathtt{b} = 1$ and $\delta(\mathtt{r})/\mathtt{r} \cong \omega \cong 1/\sqrt{n}$ if we set

$$\mathcal{D}^2 = \mathcal{V}^2 = nI_{p_n}.$$

It is straightforward to see that

$$\breve{D}_0 = \sqrt{n}, \qquad \breve{\nabla}(\mathcal{L} - I\!\!E\mathcal{L}) = \nabla_{\boldsymbol{\theta}}(\mathcal{L} - I\!\!E\mathcal{L}) = n\boldsymbol{Y}_{\boldsymbol{\theta}}, \text{ and } \breve{\boldsymbol{\xi}} = \sqrt{n}\boldsymbol{Y}_{\boldsymbol{\theta}},$$

where $\boldsymbol{Y} = (\boldsymbol{Y}_{\boldsymbol{\theta}}, \boldsymbol{Y}_{\boldsymbol{\eta}}) \in \mathbb{R}^m \times \mathbb{R}^m$. Consequently Theorem 4.2.2 gives efficiency of the profile if $p_n^2/n \to 0$ and the Wilks phenomenon if $p_n^3 n \to 0$. In the following we will first show that condition $(\mathcal{L}_0)$ is satisfied, then that the Fisher theorem holds if $p_n^2/n \to 0$ and finally that $p_n^3/n \to 0$ is indeed necessary to obtain the Wilks phenomenon.

Condition $(\mathcal{L}_0)$

We will show that $\nabla^2 I\!\!E\mathcal{L}$ is Lipschitz continuous on $\varUpsilon_\circ(\mathtt{r}_0) = B_{2\sqrt{p_n/n}}(0)$ with Lipschitz constant $n\widetilde{L} > 0$ where $\widetilde{L}$ is independent of n, p_n. This gives $(\mathcal{L}_0)$ with $\delta(\mathtt{r}) = \widetilde{L}\mathtt{r}/\sqrt{n}$. For this purpose it suffices to consider the Lipschitz continuity of $\nabla^2 g(\boldsymbol{v}) := \nabla^2(f(\boldsymbol{v})n\|\boldsymbol{v}\|^3)$. We neglect the indicator $1_{B_{2\sqrt{p_n/n}}(0)}(\cdot)$ as we only have to consider smoothness on $\varUpsilon_\circ(\mathtt{r}_0)$. We have for two points $\boldsymbol{v}, \boldsymbol{v}^\circ \in \varUpsilon_\circ$

$$\begin{aligned}\frac{1}{n}\|\nabla^2 g(\boldsymbol{v}) - \nabla^2 g(\boldsymbol{v}^\circ)\| \le \|&\nabla^2 f(\boldsymbol{v})\|\boldsymbol{v}\|^3 - \nabla^2 f(\boldsymbol{v}^\circ)\|\boldsymbol{v}^\circ\|^3\| \\ &+\|\nabla f(\boldsymbol{v})\|\boldsymbol{v}\|\boldsymbol{v}^\top - \nabla f(\boldsymbol{v}^\circ)\|\boldsymbol{v}^\circ\|\boldsymbol{v}^{\circ\top}\| \\ &+\|\frac{f(\boldsymbol{v})}{\|\boldsymbol{v}\|}\boldsymbol{v}\boldsymbol{v}^\top - \frac{f(\boldsymbol{v}^\circ)}{\|\boldsymbol{v}^\circ\|}\boldsymbol{v}^\circ\boldsymbol{v}^{\circ\top}\|.\end{aligned}$$

Denote by $L_{\|\cdot\|^3|\varUpsilon_\circ}$ the Lipschitz constant of $\|\cdot\|^3$ restricted to $\varUpsilon_\circ(\mathtt{r}_0)$, which is independent of $n, p_n \in \mathbb{N}$ because the set $\varUpsilon_\circ(\mathtt{r}_0) \subset B_1(0)$ for $n \in \mathbb{N}$ large enough. We estimate

$$\begin{aligned}&\|\nabla^2 f(\boldsymbol{v})\|\boldsymbol{v}\|^3 - \nabla^2 f(\boldsymbol{v}^\circ)\|\boldsymbol{v}^\circ\|^3\| \\ &\quad\le \|\nabla^2 f(\boldsymbol{v}) - \nabla^2 f(\boldsymbol{v}^\circ)\|\|\boldsymbol{v}\|^3 + \|\nabla^2 f(\boldsymbol{v}^\circ)\|\|\|\boldsymbol{v}\|^3 - \|\boldsymbol{v}^\circ\|^3\| \\ &\quad\le 8\left(\frac{p_n}{n}\right)^{3/2}\|\nabla^3 f\|_\infty\|\boldsymbol{v} - \boldsymbol{v}^\circ\| + \|\nabla^2 f\|_\infty L_{\|\cdot\|^3|\varUpsilon_\circ}\|\boldsymbol{v} - \boldsymbol{v}^\circ\|.\end{aligned}$$

By the definition (4.2.17) we find that

$$\|\nabla^3 f\|_\infty \le L^3\left(\frac{n}{p_n}\right)^{3/2}\left\|\int_{\mathbb{R}} K^{(3)}(\boldsymbol{v})d\boldsymbol{v}\right\| =: \mathtt{C}\left(\frac{n}{p_n}\right)^{3/2},$$

with a constant $\mathtt{C} \in \mathbb{R}$ that does not depend on $n, p_n \in \mathbb{N}$. With similar arguments for the other terms we find

$$\|\nabla^2 I\!\!E\mathcal{L}(\boldsymbol{v}) - \nabla^2 I\!\!E\mathcal{L}(\boldsymbol{v}^\circ)\| \le n\widetilde{L}\|\boldsymbol{v} - \boldsymbol{v}^\circ\|.$$

Fisher theorem

We control the deviations of the maximizer of $\mathcal{L}$. The gradient reads

$$\nabla\mathcal{L}(\boldsymbol{v}) = n\boldsymbol{Y} - n\boldsymbol{v} + nf(\boldsymbol{v})\frac{1}{2}\|\boldsymbol{v}\|\boldsymbol{v} + n\nabla f(\boldsymbol{v})\|\boldsymbol{v}\|^3/3.$$

Setting this equal to zero we find that $\widetilde{\boldsymbol{v}}$ satisfies

$$\sqrt{n}\|\boldsymbol{Y}-\widetilde{\boldsymbol{v}}\| \le \frac{\sqrt{n}}{2}\|\widetilde{\boldsymbol{v}}\|^2 + \sqrt{n}\nabla f(\widetilde{\boldsymbol{v}})\|\widetilde{\boldsymbol{v}}\|^3/3.$$

Using the fact that by Theorem 3.3.2 $I\!P(\widetilde{\boldsymbol{v}} \in B_{2\sqrt{\frac{p_n}{n}}}(0)) \ge 1-2\mathrm{e}^{-p^*}$ such that $\|\widetilde{\boldsymbol{v}}\| \cong \sqrt{p_n/n}$ and that $\|\nabla f(\widetilde{\boldsymbol{v}})\| \cong \sqrt{n/p_n}$ we obtain

$$\|\breve{D}_0(\widetilde{\boldsymbol{\theta}}-\boldsymbol{\theta}^*)-\breve{\boldsymbol{\xi}}\|^2 = n\|\widetilde{\boldsymbol{\theta}}-\boldsymbol{Y}_{\boldsymbol{\theta}}\|^2 \le n\|\widetilde{\boldsymbol{v}}-\boldsymbol{Y}\|^2 \lesssim p_n^2/n,$$

which shows that if $p_n^2/n \to 0$ we obtain the Fisher theorem.

Wilks phenomenon

Suppose for a moment that $f \equiv 1$. One can see that the unique local maximizer $\widehat{\boldsymbol{v}}$ of

$$\widehat{\mathcal{L}}(\boldsymbol{v}) = n\boldsymbol{Y}^\top\boldsymbol{v} - n\|\boldsymbol{v}\|^2/2 + n\|\boldsymbol{v}\|^3/3,$$

equals $\lambda\boldsymbol{Y}$ for some $\lambda > 0$ as only the term $n\boldsymbol{Y}^\top\boldsymbol{v}$ depends on the direction of $\boldsymbol{v}$ and is maximized on balls with finite radius on the linear space spanned by $\boldsymbol{Y}$. We will show that $\lambda = 1+\delta(\boldsymbol{Y})\|\boldsymbol{Y}\|$ where almost surely

$$\delta(\boldsymbol{Y}) \to 1.$$

To see this note that the maximization problem reduces to solving

$$\operatorname*{argmax}_{\lambda}\left\{\lambda - \lambda^2/2 + \|\boldsymbol{Y}\|\lambda^3/3\right\}.$$

The solution can easily be obtained with first and second order criteria of maximality and is given as

$$\begin{aligned}\lambda_{\max} &= \frac{1-\sqrt{1-4\|\boldsymbol{Y}\|}}{2\|\boldsymbol{Y}\|} = \frac{4\|\boldsymbol{Y}\|}{2\|\boldsymbol{Y}\|(1+\sqrt{1-4\|\boldsymbol{Y}\|})}\\ &= 1+\frac{1-\sqrt{1-4\|\boldsymbol{Y}\|}}{(1+\sqrt{1-4\|\boldsymbol{Y}\|})} = 1+\frac{4\|\boldsymbol{Y}\|}{(1+\sqrt{1-4\|\boldsymbol{Y}\|})^2} =: 1+\tau(\boldsymbol{Y})\|\boldsymbol{Y}\|.\end{aligned}$$

Consequently $\widehat{\boldsymbol{v}} = (1+\tau(\boldsymbol{Y})\|\boldsymbol{Y}\|)\boldsymbol{Y}$. If $\widehat{\boldsymbol{v}} \in \mathcal{S}$ this means that $\widetilde{\boldsymbol{v}} = \widehat{\boldsymbol{v}}$ in our model, since for any other point $\boldsymbol{v} \in \Upsilon$

$$\begin{aligned}\mathcal{L}(\boldsymbol{v}) &= n\boldsymbol{Y}^\top\boldsymbol{v} - n\|\boldsymbol{v}\|^2/2 + f(\boldsymbol{v})n\|\boldsymbol{v}\|^3\\ &\le n\boldsymbol{Y}^\top\boldsymbol{v} - n\|\boldsymbol{v}\|^2/2 + n\|\boldsymbol{v}\|^3\\ &\le \max_{\boldsymbol{v}}\left\{n\boldsymbol{Y}^\top\boldsymbol{v} - n\|\boldsymbol{v}\|^2/2 + n\|\boldsymbol{v}\|^3\right\} = \mathcal{L}(\widehat{\boldsymbol{v}}).\end{aligned}$$

The event $\{\widehat{\boldsymbol{\upsilon}} \in \mathcal{S}\}$ is of strictly positive probability that depends on the choice of $L > 0$ and grows with $n \to \infty$. Observe that if $\widehat{\boldsymbol{\upsilon}} \in \mathcal{S}$

$$\begin{aligned}
\breve{L}(\widetilde{\boldsymbol{\theta}}) &= \max_{\boldsymbol{\eta}} \mathcal{L}(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\eta}) = \mathcal{L}(\widetilde{\boldsymbol{\upsilon}}) = \mathcal{L}(\widehat{\boldsymbol{\upsilon}}) \\
&= n\left((1+\tau(\boldsymbol{Y})\|\boldsymbol{Y}\|) - (1+\tau(\boldsymbol{Y})\|\boldsymbol{Y}\|)^2/2\right)\|\boldsymbol{Y}\|^2 \\
&\quad + n(1+\tau(\boldsymbol{Y})\|\boldsymbol{Y}\|)^3/3\|\boldsymbol{Y}\|^3 \\
&= n\|\boldsymbol{Y}\|^2/2 + n\|\boldsymbol{Y}\|^3/3 + n\Bigg(\left(\frac{1}{2}+\tau(\boldsymbol{Y})\right)\|\boldsymbol{Y}\|^4 + \tau(\boldsymbol{Y})^2\|\boldsymbol{Y}\|^5 \\
&\quad + \tau(\boldsymbol{Y})^3\|\boldsymbol{Y}\|^6/3\Bigg).
\end{aligned}$$

By the definition of $\boldsymbol{Y}$ we have almost surely $\lim n\|\boldsymbol{Y}\|^2/p_n \le \mathtt{C}$, such that if $p_n^2/n \to 0$

$$n\left((\frac{1}{2}+\tau(\boldsymbol{Y}))\|\boldsymbol{Y}\|^4 + \tau(\boldsymbol{Y})^2\|\boldsymbol{Y}\|^5 + \tau(\boldsymbol{Y})^3\|\boldsymbol{Y}\|^6/3\right) = o_{I\!P}(1).$$

On the other hand we have due to $f(0,\boldsymbol{\eta}) = 0$ for all $\boldsymbol{\eta} \in \mathbb{R}^m$

$$\breve{L}(\boldsymbol{\theta}^*) = \max_{\boldsymbol{\eta}} \mathcal{L}(\boldsymbol{\theta}^*, \boldsymbol{\eta}) = \max_{\boldsymbol{\eta}} \left\{n\boldsymbol{Y}^\top(0,\boldsymbol{\eta}) - n\|\boldsymbol{\eta}\|^2/2\right\} = n\|\boldsymbol{Y}_{\boldsymbol{\eta}}\|^2/2.$$

Consequently

$$\begin{aligned}
\breve{L}(\widetilde{\boldsymbol{\theta}}) - \breve{L}(\boldsymbol{\theta}^*) &= n\|\boldsymbol{Y}\|^2/2 - n\|\boldsymbol{Y}_{\boldsymbol{\eta}}\|^2/2 + n\|\boldsymbol{Y}\|^3 + o_{I\!P}(1) \\
&= n\boldsymbol{Y}_{\boldsymbol{\theta}}^2/2 + n\|\boldsymbol{Y}\|^3 + o_{I\!P}(1).
\end{aligned}$$

It is clear that if $p_n^3/n \to 0$ also $n\|\boldsymbol{Y}\|^3 \to 0$ almost surely. Furthermore $n\boldsymbol{Y}_{\boldsymbol{\theta}}^2 \sim \chi_m^2$ for all $n \in \mathbb{N}$. But if $p_n^3/n \nrightarrow 0$ obviously $n\boldsymbol{Y}_{\boldsymbol{\theta}}^2/2 + n\|\boldsymbol{Y}\|^3 + o_{I\!P}(1)$ does not converge to a χ^2-square random variable with m degrees of freedom. In consequence the Wilks phenomenon does not occur on a set of positive probability if $p_n^3/n \nrightarrow 0$.

4.A.8 Proof of Proposition 4.3.2

Remember the definitions

$$\begin{aligned}
\widetilde{\boldsymbol{\upsilon}}_{\boldsymbol{\theta}_m^*,m} &= (\boldsymbol{\theta}_m^*, \widetilde{\boldsymbol{\eta}}_{\boldsymbol{\theta}_m^*}) \stackrel{\text{def}}{=} \operatorname*{argmax}_{\substack{\boldsymbol{\upsilon}\in\varUpsilon \\ \varPi_0\boldsymbol{\upsilon}=\boldsymbol{\theta}_m^*}} \mathcal{L}_m(\boldsymbol{\upsilon}), \\
\widetilde{\boldsymbol{\upsilon}}_{\boldsymbol{\theta}^*,m} &= (\boldsymbol{\theta}_m^*, \widetilde{\boldsymbol{\eta}}_{\boldsymbol{\theta}^*}) \stackrel{\text{def}}{=} \operatorname*{argmax}_{\substack{\boldsymbol{\upsilon}\in\varUpsilon \\ \varPi_0\boldsymbol{\upsilon}=\boldsymbol{\theta}^*}} \mathcal{L}_m(\boldsymbol{\upsilon}).
\end{aligned}$$

Define for some $0 < \mathtt{r}_0^\circ$

$$\mathcal{A}(\mathtt{x}, \mathtt{r}_0^\circ) \overset{\text{def}}{=} \{\widetilde{\boldsymbol{\upsilon}}_m, \widetilde{\boldsymbol{\upsilon}}_{\boldsymbol{\theta}_m^*, m}, \widetilde{\boldsymbol{\upsilon}}_{\boldsymbol{\theta}^*, m} \in \varUpsilon_{0,m}(\mathtt{r}_0^\circ)\} \\ \cap \left\{ \sup_{\boldsymbol{\upsilon} \in \varUpsilon_\circ(4\mathtt{r}_0^\circ)} \|\breve{\mathfrak{y}}(\boldsymbol{\upsilon})\| \le 6\nu_1 \breve{\omega}_{\mathfrak{z}1}(\mathtt{x}, 2p^* + 2p) 4\mathtt{r}_0^\circ \right\} \subset \varOmega,$$

with $\breve{\mathfrak{y}}(\boldsymbol{\upsilon}) \in \mathbb{R}^{p^*}$ in (4.A.1).

We prove this claim in a similar fashion as in Section 4.A.2. With the function $l_m : \mathbb{R}^p \times \varUpsilon \to \mathbb{R}$ defined as in (4.A.3) with $\mathcal{L}$ replaced by $\mathcal{L}_m$ we can represent:

$$\breve{L}_m(\boldsymbol{\theta}_m^*) - \breve{L}_m(\boldsymbol{\theta}^*) = l_m(\boldsymbol{\theta}_m^*, \boldsymbol{\theta}_m^*, \widetilde{\boldsymbol{\eta}}_{\boldsymbol{\theta}_m^*}) - l_m(\boldsymbol{\theta}^*, \boldsymbol{\theta}^*, \widetilde{\boldsymbol{\eta}}_{\boldsymbol{\theta}^*}),$$

where $\breve{L}_m(\boldsymbol{\theta}) \overset{\text{def}}{=} \max_{\boldsymbol{\eta} \in \varPi_m \varUpsilon_{\boldsymbol{\eta}}} \mathcal{L}_m(\boldsymbol{\theta}, \boldsymbol{\eta})$. Repeating the same arguments as in Section 4.A.2 we obtain

$$\begin{aligned} \breve{L}_m(\boldsymbol{\theta}_m^*) - \breve{L}_m(\boldsymbol{\theta}^*) &\le l_m(\boldsymbol{\theta}_m^*, \boldsymbol{\theta}_m^*, \widetilde{\boldsymbol{\eta}}_{\boldsymbol{\theta}_m^*}) - l_m(\boldsymbol{\theta}^*, \boldsymbol{\theta}_m^*, \widetilde{\boldsymbol{\eta}}_{\boldsymbol{\theta}_m^*}) \\ &= \breve{\nabla}_{\boldsymbol{\theta}} \mathcal{L}_m(\boldsymbol{\upsilon}^*)(\boldsymbol{\theta}_m^* - \boldsymbol{\theta}^*) - \|\breve{D}_m(\boldsymbol{\theta}_m^* - \boldsymbol{\theta}^*)\|^2/2 \\ &\quad + \breve{\alpha}_m^*(\boldsymbol{\theta}_m^*, \boldsymbol{\theta}^*), \end{aligned}$$

where $\breve{\alpha}_m^*(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) \in \mathbb{R}$ is defined as

$$\begin{aligned} \breve{\alpha}_m^*(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) \overset{\text{def}}{=} {} & l(\boldsymbol{\theta}_1, \boldsymbol{\theta}_m^*, \widetilde{\boldsymbol{\eta}}_{\boldsymbol{\theta}_m^*}) - l(\boldsymbol{\theta}_2, \boldsymbol{\theta}_m^*, \widetilde{\boldsymbol{\eta}}_{\boldsymbol{\theta}_m^*}) \\ & - \nabla_{\boldsymbol{\theta}_1} l(\boldsymbol{\theta}^*, \boldsymbol{\theta}^*, \boldsymbol{\eta}^*)(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2) - \|\breve{D}(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2)\|^2/2. \end{aligned}$$

and satisfies

$$\begin{aligned} \breve{\alpha}_m^*(\boldsymbol{\theta}_m^*, \boldsymbol{\theta}^*) &\le \|\breve{D}_m(\boldsymbol{\theta}_m^* - \boldsymbol{\theta}^*)\| \sup_{\boldsymbol{\theta} \in \varPi_{\boldsymbol{\theta}} \varUpsilon_\circ(4\mathtt{r}_0^\circ)} |\breve{D}_m^{-1} \nabla_{\boldsymbol{\theta}_1} \breve{\alpha}_m(\boldsymbol{\theta}, \boldsymbol{\theta}^*)| \\ &\le \alpha(m) \breve{\diamondsuit}(\mathtt{r}_0^\circ, \mathtt{x}), \end{aligned}$$

since $\mathcal{A}(\mathtt{x}, \mathtt{r}_0^\circ) \subseteq \{\widetilde{\boldsymbol{\upsilon}}_{\boldsymbol{\theta}_m^*}, \widetilde{\boldsymbol{\upsilon}}_{\boldsymbol{\theta}^*} \in \varUpsilon_\circ(\mathtt{r}_0^\circ)\}$. With similar arguments for the lower bound this gives

$$2|\breve{L}_m(\boldsymbol{\theta}_m^*) - \breve{L}_m(\boldsymbol{\theta}^*)| \le \alpha(m) \left(2\|\breve{D}^{-1} \breve{\nabla} \mathcal{L}_m(\boldsymbol{\upsilon}^*)\| + \alpha(m) + 2\breve{\diamondsuit}(\mathtt{r}_0^\circ, \mathtt{x}) \right).$$

The claim follows because the result (4.2.10) of Theorem 4.2.2 occurs on

$$\mathcal{A}(\mathtt{x}, \mathtt{r}_0^\circ) \subseteq C(\mathtt{x}, \mathtt{r}_0^\circ) \subset \varOmega.$$

It remains to note that the set $\mathcal{A}(\mathtt{x}, \mathtt{r}_0^\circ) \subset \varOmega$ is of probability greater than $1 - 2\mathrm{e}^{-\mathtt{x}}$ by the choice of $\mathtt{r}_0^\circ > 0$.

4.A.9 Proof of Corollary 4.3.3

We will only prove the asymptotic normality as the the proof the Wilks phenomenon is very similar. Define

$$\mathcal{V}_m^2(\boldsymbol{\upsilon}_m^*) = \operatorname{Cov}\bigl(\nabla_{p+m}\mathcal{L}_m(\boldsymbol{\upsilon}_m^*)\bigr),\ \mathbb{B}_m = \mathcal{D}_m^{-1}\mathcal{V}_m^2\mathcal{D}_m^{-1},$$
$$\breve{\nabla}_{\boldsymbol{\theta},m} = \nabla_{\boldsymbol{\theta}} - \mathrm{A}_m\mathrm{H}_m^{-2}\nabla_{\boldsymbol{\eta}},\ \breve{V}_m^2 = \operatorname{Cov}(\breve{\nabla}_{\boldsymbol{\theta}}\zeta(\boldsymbol{\upsilon}_m^*)),\ \breve{\mathbb{B}}_m = \breve{D}_m^{-1}\breve{V}_m^2\breve{D}_m^{-1}.$$

Remember $p^* = p+m \in \mathbb{N}$ and that the point $\boldsymbol{\upsilon}_m^* \in \mathbb{R}^p \times \mathbb{R}^m$ is defined by maximizing the expected value for the sieved functional $\mathcal{L}_m$ and the operators $\mathcal{D}_m^2 \in \mathbb{R}^{p^*\times p^*}$, $\breve{D}_m^2 \in \mathbb{R}^{p\times p}$ correspond to this point, i.e. we abbreviate $\mathcal{D}_m^2 \stackrel{\mathrm{def}}{=} \mathcal{D}_m^2(\boldsymbol{\upsilon}_m^*)$, while $\mathcal{D}^2 = \mathcal{D}^2(\boldsymbol{\upsilon}^*)$ and $\breve{D}_m^2 = \breve{D}_m^2(\boldsymbol{\upsilon}_m^*)$, $\breve{D}^2 = \breve{D}^2(\boldsymbol{\upsilon}^*)$, where $\boldsymbol{\upsilon}^* = \operatorname{argmax}_{\boldsymbol{\upsilon}\in\Upsilon}\mathbb{E}\mathcal{L}(\boldsymbol{\upsilon})$, i.e. the true full maximizer.

We get with Theorem 4.2.2 applied to $\widetilde{\boldsymbol{\theta}}_m$ in (4.3.2) that with probability greater than $1-2\mathrm{e}^{-\mathtt{x}}$

$$\|\breve{D}_m\bigl(\widetilde{\boldsymbol{\theta}}_m - \boldsymbol{\theta}_m^*\bigr) - \breve{\boldsymbol{\xi}}_m(\boldsymbol{\upsilon}_m^*)\| \le \diamondsuit(\mathtt{r}_0,\mathtt{x}). \tag{4.A.12}$$

We write

$$\begin{aligned}&\breve{D}\bigl(\widetilde{\boldsymbol{\theta}}_m - \boldsymbol{\theta}^*\bigr) - \breve{\boldsymbol{\xi}}_m(\boldsymbol{\upsilon}_m^*)\\ &\quad= \breve{D}_m\bigl(\widetilde{\boldsymbol{\theta}}_m - \boldsymbol{\theta}_m^*\bigr) - \breve{\boldsymbol{\xi}}_m(\boldsymbol{\upsilon}_m^*) + (\breve{D}_m - \breve{D})\bigl(\widetilde{\boldsymbol{\theta}}_m - \boldsymbol{\theta}_m^*\bigr) + \breve{D}_m\bigl(\boldsymbol{\theta}_m^* - \boldsymbol{\theta}^*\bigr).\end{aligned}$$

By (4.A.12) it suffices to bound $\|(\breve{D}_m - \breve{D})(\widetilde{\boldsymbol{\theta}}_m - \boldsymbol{\theta}_m^*)\|$ and $\|\breve{D}_m(\boldsymbol{\theta}_m^* - \boldsymbol{\theta}^*)\|$. With assumption (*bias*) we get

$$\|\breve{D}_m(\boldsymbol{\theta}_m^* - \boldsymbol{\theta}^*)\| \le \alpha(m).$$

Furthermore

$$\begin{aligned}&\|(\breve{D}_m - \breve{D})\bigl(\widetilde{\boldsymbol{\theta}}_m - \boldsymbol{\theta}_m^*\bigr)\|\\ &\quad\le \|(\breve{D}_m - \breve{D}_m(\boldsymbol{\upsilon}^*))\bigl(\widetilde{\boldsymbol{\theta}}_m - \boldsymbol{\theta}_m^*\bigr)\| + \|(\breve{D}_m(\boldsymbol{\upsilon}^*) - \breve{D})\bigl(\widetilde{\boldsymbol{\theta}}_m - \boldsymbol{\theta}_m^*\bigr)\|\\ &\quad\le \|\breve{D}_m\bigl(\widetilde{\boldsymbol{\theta}}_m - \boldsymbol{\theta}_m^*\bigr)\| \Bigl(\|\boldsymbol{I} - \breve{D}_m^{-1}\breve{D}_m^2(\boldsymbol{\upsilon}^*)\breve{D}_m^{-1}\|^{1/2}\\ &\qquad\quad + \|\boldsymbol{I} - \breve{D}_m(\boldsymbol{\upsilon}^*)^{-1}\breve{D}^2(\boldsymbol{\upsilon}^*)\breve{D}_m(\boldsymbol{\upsilon}^*)^{-1}\|^{1/2}\|\breve{D}_m(\boldsymbol{\upsilon}^*)\breve{D}_m^{-1}\|\Bigr).\end{aligned}$$

With (4.A.12) and the fact, that with condition $(\breve{\mathcal{E}}\mathcal{D}_0)$ it holds that (see Section 3.4)

$$\mathbb{P}(\|\breve{\boldsymbol{\xi}}_m(\boldsymbol{\upsilon}_m^*)\| \le \mathfrak{z}(\mathtt{x}_n, \breve{\mathbb{B}}_m)) \ge 1 - 2\mathrm{e}^{\mathtt{x}_n},$$

we obtain with probability greater than $1-4\mathrm{e}^{\mathtt{x}_n}$

$$\|\breve{D}_m(\widetilde{\boldsymbol{\theta}}_m - \boldsymbol{\theta}_m^*)\| \le \|\breve{\boldsymbol{\xi}}_m(\boldsymbol{\upsilon}_m^*\| + \diamondsuit(\mathtt{r}_0,\mathtt{x}) \le \mathfrak{z}(\mathtt{x}, \breve{\mathbb{B}}_m) + \diamondsuit(\mathtt{r}_0,\mathtt{x}).$$

where $\mathfrak{z}(\mathtt{x}, \breve{\mathbb{B}}_m) = O(\sqrt{p+\mathtt{x}})$. Combining these bounds gives with $(bias')$

$$\begin{aligned}\|\breve{D}\bigl(\widetilde{\boldsymbol{\theta}}_m - \boldsymbol{\theta}^*\bigr) - \breve{\boldsymbol{\xi}}_m(\boldsymbol{\upsilon}^*_m)\| \le\; & \diamondsuit(\mathtt{r}_0, \mathtt{x}) + \beta(m)\left(\mathfrak{z}(\mathtt{x}, \breve{\mathbb{B}}_m) + \diamondsuit(\mathtt{r}_0, \mathtt{x})\right)\\ & +\alpha(m),\end{aligned}$$

where $\mathtt{r}_0(\mathtt{x})$ is chosen such that $\mathbb{P}(\widetilde{\boldsymbol{\upsilon}}_n, \widetilde{\boldsymbol{\upsilon}}_{\boldsymbol{\theta}^*_m,m} \in \varUpsilon_{0,m}(\mathtt{r}_0(\mathtt{x}))) \ge 1 - \mathrm{e}^{-\mathtt{x}}$.

By assumption $\mathtt{r}_0(\mathtt{x}) < \infty$ for any $\mathtt{x} > 0$, $m, n \in \mathbb{N}$. Remember that $\breve{\diamondsuit}\bigl(\mathtt{r}_0, \mathtt{x}_n\bigr) \approx \breve{\delta}_n(\mathtt{r}_0)\mathtt{r}_0 + \breve{\omega}_n\sqrt{\mathtt{x} + p + m_n}\mathtt{r}_0$ where by assumption $\breve{\delta}_n(\mathtt{r}) \to 0$ for any $\mathtt{r} > 0$ and $\omega_n \to 0$. This implies that there exist sequences $(m_n) \subset \mathbb{N}$ with $m_n \to \infty$ and $\mathtt{x}_n \to \infty$ with

$$\diamondsuit(\mathtt{r}_0, \mathtt{x}_n) + \beta(m)\left(\mathfrak{z}(\mathtt{x}_n, \breve{\mathbb{B}}_{m_n}) + \diamondsuit(\mathtt{r}_0, \mathtt{x}_n)\right) + \alpha(m_n) \to 0 \quad (4.A.13)$$

as $n \to \infty$. Fix such sequences $m_n \to \infty$ and $\mathtt{x}_n \to \infty$. Then we have due to (4.A.13) that for any $\epsilon > 0$ there exists an $n \in \mathbb{N}$ such that

$$\mathbb{P}(\|\breve{D}(\widetilde{\boldsymbol{\theta}}_m - \boldsymbol{\theta}^*) - \breve{\boldsymbol{\xi}}_m(\boldsymbol{\upsilon}^*_m)\| \ge \epsilon) \le 4\mathrm{e}^{\mathtt{x}_n}.$$

As $\mathtt{x}_n \to \infty$ we get the claim by Slutsky's Lemma once we showed that $\breve{\boldsymbol{\xi}}_m(\boldsymbol{\upsilon}^*_m)$ is asymptotically $\mathcal{N}(0, \breve{d}^{-1}\breve{v}^2\breve{d}^{-1})$-distributed.

For this observe

$$\begin{aligned}\breve{\boldsymbol{\xi}}_m(\boldsymbol{\upsilon}^*_m) &= \breve{D}_m^{-1}(\nabla_{\boldsymbol{\theta}} - \mathrm{A}_m\mathrm{H}_m^{-2}\nabla_{\boldsymbol{\eta}})\mathcal{L}(\boldsymbol{\upsilon}^*_m)\\ &= \frac{1}{\sqrt{n}}\sum_{i=1}^n \left(\frac{1}{\sqrt{n}}\breve{D}_m\right)^{-1}\bigl(\nabla_{\boldsymbol{\theta}}\ell_i(\boldsymbol{\upsilon}^*_m) - \mathrm{A}_m\mathrm{H}_m^{-2}\nabla_{\boldsymbol{\eta}}\ell_i(\boldsymbol{\upsilon}^*_m)\bigr)\\ &\overset{\mathrm{def}}{=} \frac{1}{\sqrt{n}}\sum_{i=1}^n \mathbf{X}_i.\end{aligned}$$

Due to assumptions $\boldsymbol{(bias'')}$ we have $\operatorname{Cov}(\mathbf{X}_i) \to \breve{d}^{-1}\breve{v}^2\breve{d}^{-1} \in \mathbb{R}^{p\times p}$. Consequently

$$\breve{\boldsymbol{\xi}}_m(\boldsymbol{\upsilon}^*_m) = \frac{1}{\sqrt{n}}\sum_{i=1}^n \mathbf{X}_i,$$

where the random vectors $\mathbf{X}_i$ are i.i.d. with zero mean and covariance tending to $\breve{d}^{-1}\breve{v}^2\breve{d}^{-1}$, such that by a slightly generalized central limit theorem

$$\breve{\boldsymbol{\xi}}_m(\boldsymbol{\upsilon}^*_m) \xrightarrow{w} \mathcal{N}(0, \breve{d}^{-1}\breve{v}^2\breve{d}^{-1}).$$

4.A.10 Proof of Theorem 4.3.4

We prove this theorem in a series of lemmas.

Lemma 4.A.7. *Assume that* $(\mathcal{L}\mathtt{r}_{\infty})$ *is satisfied for any* $\mathtt{r} \geq \mathtt{r}_1$ *with* $\mathtt{b}(\mathtt{r}) \equiv \mathtt{b}$ *and that the condition* $(\boldsymbol{\varkappa})$ *is satisfied. Then we get* $\|\mathcal{D}(\boldsymbol{\upsilon}_m^* - \boldsymbol{\upsilon}^*)\| \leq \mathtt{r}^* \vee \mathtt{r}_1$ *where* $\mathtt{r}^{*2} = 4\mathtt{C}_{\boldsymbol{\varkappa}^*} m/\mathtt{b}$ *.*

Proof. Note that

$$\|\mathcal{D}(\boldsymbol{\upsilon}^* - \varPi_{p^*}\boldsymbol{\upsilon}^*)\| = \|\mathcal{H}_{\boldsymbol{\varkappa}\boldsymbol{\varkappa}}\boldsymbol{\varkappa}^*\|,$$

such that $\boldsymbol{\upsilon}^* \in \varUpsilon_\circ(\mathtt{r}^*)$. Furthermore we have $\nabla I\!\!E\mathcal{L}(\boldsymbol{\upsilon}^*) = 0$ such that by the Taylor expansion with some $\lambda \in [0,1]$

$$I\!\!E\mathcal{L}(\varPi_{p^*}\boldsymbol{\upsilon}^*, \boldsymbol{\upsilon}^*) = -\|\mathcal{H}_{\boldsymbol{\varkappa}\boldsymbol{\varkappa}}\boldsymbol{\varkappa}^*\|^2 + \boldsymbol{\varkappa}^{*\top}(\mathcal{H}_{\boldsymbol{\varkappa}\boldsymbol{\varkappa}} - \nabla_{\varkappa\varkappa} I\!\!E\mathcal{L}(\boldsymbol{\upsilon}^*, \lambda\boldsymbol{\varkappa}^*))\boldsymbol{\varkappa}^*.$$

which gives with (4.3.5) and $(\boldsymbol{\varkappa})$ on $\varUpsilon_\circ(\mathtt{r}^*)$ that

$$|I\!\!E\mathcal{L}(\varPi_{p^*}\boldsymbol{\upsilon}^*, \boldsymbol{\upsilon}^*)| \leq \|\mathcal{D}(\boldsymbol{\upsilon}^* - \varPi_{p^*}\boldsymbol{\upsilon}^*)\|^2 + \mathtt{C}_{\boldsymbol{\varkappa}^*} m \leq 2\mathtt{C}_{\boldsymbol{\varkappa}^*} m. \tag{4.A.14}$$

We show that $\boldsymbol{\upsilon}_m^*$ also belongs to $\varUpsilon_\circ(\mathtt{r}^*)$ for $\mathtt{r}^{*2} \geq 4\mathtt{C}_{\boldsymbol{\varkappa}^*} m/\mathtt{b}$. Suppose for the moment that $\big\|\mathcal{D}(\boldsymbol{\upsilon}_m^* - \boldsymbol{\upsilon}^*)\big\| > \mathtt{r}^* \vee \mathtt{r}_1$. By $(\mathcal{L}\mathtt{r}_{\infty})$, it holds

$$2\big|I\!\!E\mathcal{L}(\boldsymbol{\upsilon}_m^*, \boldsymbol{\upsilon}^*)\big| \geq \mathtt{b}\big\|\mathcal{D}(\boldsymbol{\upsilon}_m^* - \boldsymbol{\upsilon}^*)\big\|^2 > \mathtt{b}\mathtt{r}^{*2}. \tag{4.A.15}$$

This contradicts $\big|I\!\!E\mathcal{L}(\boldsymbol{\upsilon}_m^*, \boldsymbol{\upsilon}^*)\big| \leq \big|I\!\!E\mathcal{L}(\varPi_{p^*}\boldsymbol{\upsilon}^*, \boldsymbol{\upsilon}^*)\big|$ in view of

$$\mathtt{r}^{*2} \geq 4\mathtt{C}_{\boldsymbol{\varkappa}^*} m/\mathtt{b},$$

and (4.A.14), such that $\boldsymbol{\upsilon}_m^* \in \varUpsilon_\circ(\mathtt{r}^*)$. □

Lemma 4.A.8. *Assume that* $(\mathcal{L}\mathtt{r}_{\infty})$ *is satisfied with* $\mathtt{b}(\mathtt{r}) \equiv \mathtt{b}$ *. Assume further* $(\boldsymbol{\varkappa})$ *and* $(\breve{\mathcal{L}}_0)$ *with central point* $\boldsymbol{\upsilon}_m^* \in \mathbb{R}^{p^*}$ *and operator* $\mathcal{D}_m\mathbb{R}^{p^* \times p^*}$ *. Then we get with* $\mathtt{r}^{*2} = 4\mathtt{C}_{\boldsymbol{\varkappa}^*} m/\mathtt{b}$ *and some* $\mathtt{C}(\nu) > 0$

$$\|\breve{D}_m(\boldsymbol{\theta}_m^* - \boldsymbol{\theta}^*)\| \leq \mathtt{C}(\nu)\left(\alpha(m) + \tau(m) + \breve{\delta}(2\mathtt{r}^*)\mathtt{r}^*\right).$$

Proof. Using condition $(\breve{\mathcal{L}}_0)$ and Lemma 4.A.1 we have on

$$\varUpsilon_m(\mathtt{r}) = \{\|\mathcal{D}_m(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}_m^*)\| \leq \mathtt{r}\} \subset \mathbb{R}^{p+m},$$

that

$$\sup_{\boldsymbol{\upsilon} \in \varUpsilon_m(\mathtt{r})} \|\breve{D}_m^{-1}\left(\breve{\nabla}_m I\!\!E\mathcal{L}_m(\boldsymbol{\upsilon}) - \breve{\nabla}_m I\!\!E\mathcal{L}_m(\boldsymbol{\upsilon}_m^*)\right) - \breve{D}_m(\boldsymbol{\theta} - \boldsymbol{\theta}_m^*)\|$$
$$\leq \mathtt{C}\breve{\delta}(\mathtt{r})\mathtt{r}.$$

Because of Lemma 4.A.7 we know that

$$\begin{aligned}\|\mathcal{D}_m(\varPi_{p^*}\boldsymbol{\upsilon}^* - \boldsymbol{\upsilon}^*_m)\| &= \|\mathcal{D}(\varPi_{p^*}\boldsymbol{\upsilon}^* - \boldsymbol{\upsilon}^*_m)\| \\ &\le \|\mathcal{D}(\varPi_{p^*}\boldsymbol{\upsilon}^* - \boldsymbol{\upsilon}^*)\| + \|\mathcal{D}(\boldsymbol{\upsilon}^* - \boldsymbol{\upsilon}^*_m)\| \le 2\mathtt{r}^*,\end{aligned}$$

such that $\varPi_{p^*}\boldsymbol{\upsilon}^*_m \in \varUpsilon_{0,m}(2\mathtt{r}^*)$, which gives

$$\begin{aligned}&\|\breve{D}_m^{-1}\left(\breve{\nabla}_m I\!\!E\mathcal{L}_m(\boldsymbol{\upsilon}^*_m) - \breve{\nabla}_m I\!\!E\mathcal{L}_m(\varPi_{p^*}\boldsymbol{\upsilon}^*)\right) - \breve{D}_m(\boldsymbol{\theta}^* - \boldsymbol{\theta}^*_m)\| \\ &\quad\le 2\mathtt{C}\breve{\delta}(2\mathtt{r}^*)\mathtt{r}^*.\end{aligned}$$

We derive with the triangle inequality

$$\begin{aligned}\left\|\breve{D}_m(\boldsymbol{\theta}^*_m - \boldsymbol{\theta}^*)\right\| &\le 2\mathtt{C}(\nu)\breve{\delta}(2\mathtt{r}^*)\mathtt{r}^* \\ &\qquad + \left\|\breve{D}_m^{-1}\left(\breve{\nabla}_m I\!\!E\mathcal{L}_m(\boldsymbol{\upsilon}^*_m) - \breve{\nabla}_m I\!\!E\mathcal{L}_m(\varPi_{p^*}\boldsymbol{\upsilon}^*)\right)\right\|.\end{aligned}$$

It remains to bound the second term on the right-hand side. Because $\nabla_{p+m} I\!\!E\mathcal{L}(\boldsymbol{\upsilon}^*_m) = 0$ and $\nabla I\!\!E\mathcal{L}(\boldsymbol{\upsilon}^*) = 0$ we find

$$\begin{aligned}&\left\|\breve{D}_m^{-1}\left(\breve{\nabla}_m I\!\!E\mathcal{L}_m(\boldsymbol{\upsilon}^*_m) - \breve{\nabla}_m I\!\!E\mathcal{L}_m(\varPi_{p^*}\boldsymbol{\upsilon}^*)\right)\right\| \\ &\quad = \left\|\breve{D}_m^{-1}\left(\breve{\nabla}_m I\!\!E\mathcal{L}(\boldsymbol{\upsilon}^*) - \breve{\nabla}_m I\!\!E\mathcal{L}(\varPi_{p^*}\boldsymbol{\upsilon}^*)\right)\right\|.\end{aligned}$$

Using that $\|\mathcal{D}(\varPi\boldsymbol{\upsilon}^* - \boldsymbol{\upsilon}^*)\| \le \mathtt{r}^*$, Lemma 4.A.5 and condition $(\boldsymbol{\mathcal{I}})$ we may infer by Taylor expansion that with some $\lambda \in [0,1]$

$$\begin{aligned}&\left\|\breve{D}_m^{-1}\left(\breve{\nabla}_m I\!\!E\mathcal{L}(\boldsymbol{\upsilon}^*) - \breve{\nabla}_m I\!\!E\mathcal{L}(\varPi_{p^*}\boldsymbol{\upsilon}^*)\right)\right\| \\ &\quad\le \left\|\mathcal{D}_m^{-1}A^{\top}_{\boldsymbol{\varkappa}\boldsymbol{\upsilon}_m}\left(\varPi_{p^*}\boldsymbol{\upsilon}^* - \boldsymbol{\upsilon}^*\right)\right\| \\ &\qquad + \left\|\mathcal{D}_m^{-1}(\nabla_{\boldsymbol{\upsilon}\boldsymbol{\varkappa}} I\!\!E[\mathcal{L}\left((\varPi_{p^*}\boldsymbol{\upsilon}^*, \lambda\boldsymbol{\varkappa}^*)\right)] - A^{\top}_{\boldsymbol{\varkappa}\boldsymbol{\upsilon}})\boldsymbol{\varkappa}^*\right\| \\ &\quad= \left\|\mathcal{D}_m^{-1}A^{\top}_{\boldsymbol{\varkappa}\boldsymbol{\upsilon}}\boldsymbol{\varkappa}^*\right\| \\ &\qquad + \left\|\mathcal{D}_m^{-1}\left(\nabla_{\boldsymbol{\upsilon}\boldsymbol{\varkappa}} I\!\!E[\mathcal{L}\left((\varPi_{p^*}\boldsymbol{\upsilon}^*, \lambda\boldsymbol{\varkappa}^*)\right)] - A^{\top}_{\boldsymbol{\varkappa}\boldsymbol{\upsilon}}\right)\boldsymbol{\varkappa}^*\right\|.\end{aligned}$$

Due to assumption $(\boldsymbol{\varkappa})$ the last sum is bounded by $(\widehat{\alpha}(m) + \tau(m))$. Together this gives that

$$\left\|\breve{D}_m(\boldsymbol{\theta}^*_m - \boldsymbol{\theta}^*)\right\| = \mathtt{C}\left(\widehat{\alpha}(m) + \tau(m) + \breve{\delta}(2\mathtt{r}^*)\mathtt{r}^*\right).$$

□

Lemma 4.A.9. *Assume* $(\boldsymbol{\upsilon\varkappa})$ *then*

$$\|I - \breve{D}_m^{-1}\breve{D}^2\breve{D}_m^{-1}\| \le \frac{1+\nu^2+\beta^2(m)}{1-\nu^2}\frac{\beta^2(m)}{1-\beta^2(m)}.$$

Proof. Take any $\mathbf{v} \in \mathbb{R}^p$ with $\|\mathbf{v}\| \le 1$ and note that with $\boldsymbol{\upsilon} = (\boldsymbol{\theta}, \boldsymbol{\eta}, \boldsymbol{\varkappa}) \in l^2$

$$\begin{aligned}
&\breve{D}^{-2}\breve{D}_m\mathbf{v} \\
&\quad= \mathit{\Pi}_{\boldsymbol{\theta}} \operatorname*{argmax}_{\boldsymbol{\upsilon}\in l^2}\Big\{\boldsymbol{\theta}^\top\breve{D}_m\mathbf{v} - \|\mathcal{D}\boldsymbol{\upsilon}\|^2/2\Big\} \\
&\quad= \mathit{\Pi}_{\boldsymbol{\theta}} \operatorname*{argmax}_{\boldsymbol{\upsilon}\in \mathbb{R}^{p+m}}\Big\{\boldsymbol{\theta}^\top\breve{D}_m\mathbf{v} - \|\mathcal{D}_m\boldsymbol{\upsilon}\|^2/2 - \inf_{\boldsymbol{\varkappa}}(\boldsymbol{\varkappa}^\top A_{\boldsymbol{\varkappa\upsilon}}^\top\boldsymbol{\upsilon} + \|\mathcal{H}_{\boldsymbol{\varkappa\varkappa}}\boldsymbol{\varkappa}\|^2/2)\Big\} \\
&\quad= \mathit{\Pi}_{\boldsymbol{\theta}} \operatorname*{argmax}_{\boldsymbol{\upsilon}\in \mathbb{R}^{p+m}}\Big\{\boldsymbol{\theta}^\top\breve{D}_m\mathbf{v} - \|\mathcal{D}_m\boldsymbol{\upsilon}\|^2/2 - \|\mathcal{H}_{\boldsymbol{\varkappa\varkappa}}^{-1/2}A_{\boldsymbol{\varkappa\upsilon}}^\top\boldsymbol{\upsilon}\|^2/2\Big\} \\
&\quad\stackrel{\text{def}}{=} \mathit{\Pi}_{\boldsymbol{\theta}} \operatorname*{argmax}_{\boldsymbol{\upsilon}\in \mathbb{R}^{p+m}} g(\boldsymbol{\upsilon}) \stackrel{\text{def}}{=} \mathit{\Pi}_{\boldsymbol{\theta}}\boldsymbol{\upsilon}^\circ.
\end{aligned}$$

Setting the gradient of $g(\cdot)$ equal to zero gives that the maximizer $\boldsymbol{\upsilon}^\circ \in \mathbb{R}^{p+m}$ satisfies

$$\boldsymbol{\upsilon}^\circ = \mathcal{D}_m^{-1}(I_{p^*} - \mathcal{D}_m^{-1}A_{\boldsymbol{\varkappa\upsilon}}^\top\mathcal{H}_{\boldsymbol{\varkappa\varkappa}}^{-2}A_{\boldsymbol{\varkappa\upsilon}}^\top\mathcal{D}_m^{-1})^{-1}\mathcal{D}_m^{-1}\mathit{\Pi}_{\boldsymbol{\theta}}^\top\breve{D}_m\mathbf{v},$$

where $\mathit{\Pi}_{\boldsymbol{\theta}}^\top : \mathbb{R}^p \to \mathbb{R}^{p+m}$ denotes the canonical embedding of $\mathbb{R}^p$ into $\mathbb{R}^{p+m}$. By assumption we have

$$\|(I_{p^*} - \mathcal{D}_m^{-1}A_{\boldsymbol{\varkappa\upsilon}}^\top\mathcal{H}_{\boldsymbol{\varkappa\varkappa}}^{-2}A_{\boldsymbol{\varkappa\upsilon}}^\top\mathcal{D}_m^{-1})^{-1} - I_{p^*}\| \le \frac{\beta^2(m)}{1-\beta^2(m)}.$$

Note that $\breve{D}_m\mathit{\Pi}_{\boldsymbol{\theta}}\mathcal{D}_m^{-2}\mathit{\Pi}_{\boldsymbol{\theta}}^\top\breve{D}_m\mathbf{v} = \mathbf{v}$ which gives

$$\begin{aligned}
&\|(I - \breve{D}_m\breve{D}^{-2}\breve{D}_m)\mathbf{v}\| = \|\mathbf{v} - \breve{D}_m\mathit{\Pi}_{\boldsymbol{\theta}}\boldsymbol{\upsilon}^\circ\| \\
&\quad= \|\breve{D}_m\mathit{\Pi}_{\boldsymbol{\theta}}\mathcal{D}_m^{-2}\mathit{\Pi}_{\boldsymbol{\theta}}^\top\breve{D}_m\mathbf{v} - \breve{D}_m\mathit{\Pi}_{\boldsymbol{\theta}}\boldsymbol{\upsilon}^\circ\| \\
&\quad= \Bigg\|\breve{D}_m\mathit{\Pi}_{\boldsymbol{\theta}}\mathcal{D}_m^{-1} \\
&\qquad\qquad\Big((I_{p^*} - \mathcal{D}_m^{-1}A_{\boldsymbol{\varkappa\upsilon}}^\top\mathcal{H}_{\boldsymbol{\varkappa\varkappa}}^{-2}A_{\boldsymbol{\varkappa\upsilon}}^\top\mathcal{D}_m^{-1})^{-1} - I_{p^*}\Big)\mathcal{D}_m^{-2}\mathit{\Pi}_{\boldsymbol{\theta}}^\top\breve{D}_m\mathbf{v}\Bigg\| \\
&\quad\le \frac{\beta^2(m)}{1-\beta^2(m)}\|\breve{D}_m\mathit{\Pi}_{\boldsymbol{\theta}}\mathcal{D}_m^{-2}\mathit{\Pi}_{\boldsymbol{\theta}}^\top\breve{D}_m\mathbf{v}\| \\
&\quad\le \frac{\beta^2(m)}{1-\beta^2(m)}\|\mathbf{v}\| = \frac{\beta^2(m)}{1-\beta^2(m)}.
\end{aligned}$$

This implies

$$\begin{aligned}\|I - \breve{D}_m^{-1}\breve{D}^2\breve{D}_m^{-1}\| &\le \|I - \breve{D}_m\breve{D}^{-2}\breve{D}_m\|\|\breve{D}_m^{-1}\breve{D}^2\breve{D}_m^{-1}\| \\ &\le \frac{1+\nu^2+\beta^2(m)}{1-\nu^2}\frac{\beta^2(m)}{1-\beta^2(m)}.\end{aligned}$$

□

With similar arguments as in the proof of Lemma 4.A.1 we can prove the following lemma which completes the proof:

Lemma 4.A.10. *Assume* $(\boldsymbol{\upsilon\varkappa})$, $(\mathcal{I})$ *and* $(\breve{\mathcal{L}}_0)$ *then we get with* $\mathtt{r}^{*2}$ $= 4\mathtt{C}_{\boldsymbol{\varkappa}^*}m/\mathtt{b}$

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

Proof. Denote $\breve{D}_{mm} \stackrel{\text{def}}{=} \breve{D}_m(\boldsymbol{\upsilon}_m^*)$, $\mathcal{D}_{mm} \stackrel{\text{def}}{=} \mathcal{D}_m(\boldsymbol{\upsilon}_m^*)$ and $\breve{D}_m \stackrel{\text{def}}{=} \breve{D}_m(\boldsymbol{\upsilon}^*)$, $\mathcal{D}_m \stackrel{\text{def}}{=} \mathcal{D}_m(\boldsymbol{\upsilon}^*)$. We simply calculate

$$\begin{aligned}\|\breve{D}_{mm}^{-1}\breve{D}_m^2\breve{D}_{mm}^{-1} - I_p\| &= \|\breve{D}_{mm}^{-1}\left(\breve{D}_m^2 - \breve{D}_{mm}^2\right)\breve{D}_{mm}^{-1}\| \\ &\le \|\breve{D}_{mm}^{-1}\left(D(\boldsymbol{\upsilon}^*)^2 - D(\boldsymbol{\upsilon}_m^*)^2\right)\breve{D}_{mm}^{-1}\| \\ &\quad + \|\breve{D}_{mm}^{-1}\left(\mathrm{A}_m\mathrm{H}_m^{-2}\mathrm{A}_m^\top(\boldsymbol{\upsilon}^*) - \mathrm{A}_m\mathrm{H}_m^{-2}\mathrm{A}_m^\top(\boldsymbol{\upsilon}_m^*)\right)\breve{D}_{mm}^{-1}\|.\end{aligned}$$

Now with Lemma 4.A.7 the first summand can be bounded by

$$\begin{aligned}\|\breve{D}_{mm}^{-1}\left(D(\boldsymbol{\upsilon}^*)^2 - D(\boldsymbol{\upsilon}_m^*)^2\right)\breve{D}_{mm}^{-1}\| &\le \|\breve{D}_{mm}^{-1}D(\boldsymbol{\upsilon}_m^*)^1\|^2\breve{\delta}(\mathtt{r}^*) \\ &\le \frac{1}{(1-\sqrt{\nu})^2}\breve{\delta}(\mathtt{r}^*).\end{aligned}$$

We use the triangular inequality to find

$$\begin{aligned}&\|\breve{D}_{mm}^{-1}\left(\mathrm{A}_m\mathrm{H}_m^{-2}\mathrm{A}_m^\top(\boldsymbol{\upsilon}^*) - \mathrm{A}_m\mathrm{H}_m^{-2}\mathrm{A}_m^\top(\boldsymbol{\upsilon}_m^*)\right)\breve{D}_{mm}^{-1}\| \\ &\quad\le \|\breve{D}_{mm}^{-1}\left(\mathrm{A}_m(\boldsymbol{\upsilon}^*) - \mathrm{A}_m(\boldsymbol{\upsilon}_m^*)\right)\mathrm{H}_m^{-2}(\boldsymbol{\upsilon}^*)\mathrm{A}_m^\top(\boldsymbol{\upsilon}^*)\breve{D}_{mm}^{-1}\| \\ &\qquad + \|\breve{D}_{mm}^{-1}\left(\mathrm{A}_m(\boldsymbol{\upsilon}^*) - \mathrm{A}_m(\boldsymbol{\upsilon}_m^*)\right)\mathrm{H}_m^{-2}(\boldsymbol{\upsilon}_m^*)\mathrm{A}_m^\top(\boldsymbol{\upsilon}_m^*)\breve{D}_{mm}^{-1}\| \\ &\qquad + \|\breve{D}_{mm}^{-1}\mathrm{A}_m(\boldsymbol{\upsilon}_m^*)\left(\mathrm{H}_m(\boldsymbol{\upsilon}_m^*)^{-2} - \mathrm{H}_m(\boldsymbol{\upsilon}^*)^{-2}\right)\mathrm{A}_m^\top(\boldsymbol{\upsilon}^*)\breve{D}_{mm}^{-1}\|.\end{aligned}$$

With condition $(\breve{\mathcal{L}}_0)$ and $(\mathcal{I})$ we find $\|D(\boldsymbol{\upsilon}_m^*)D(\boldsymbol{\upsilon}^*)^{-1}\| \leq \sqrt{1-\breve{\delta}(\mathtt{r}^*)}$

$$\|\breve{D}_{mm}^{-1}\left(\mathrm{A}_m(\boldsymbol{\upsilon}^*) - \mathrm{A}_m(\boldsymbol{\upsilon}_m^*)\right)\mathrm{H}_m^{-2}(\boldsymbol{\upsilon}^*)\mathrm{A}_m^\top(\boldsymbol{\upsilon}^*)\breve{D}_{mm}^{-1}\| \leq \frac{\sqrt{\nu}\sqrt{1-\breve{\delta}(\mathtt{r}^*)}}{(1-\sqrt{\nu})^2}\breve{\delta}(\mathtt{r}^*),$$

$$\|\breve{D}_{mm}^{-1}\left(\mathrm{A}_m(\boldsymbol{\upsilon}^*) - \mathrm{A}_m(\boldsymbol{\upsilon}_m^*)\right)\mathrm{H}_m^{-2}(\boldsymbol{\upsilon}_m^*)\mathrm{A}_m^\top(\boldsymbol{\upsilon}_m^*)\breve{D}_{mm}^{-1}\| \leq \frac{\sqrt{\nu}}{(1-\sqrt{\nu})^2}\breve{\delta}(\mathtt{r}^*).$$

With the same argument we find

$$\begin{aligned}&\|\breve{D}_{mm}^{-1}\mathrm{A}_m(\boldsymbol{\upsilon}_m^*)\left(\mathrm{H}_m(\boldsymbol{\upsilon}_m^*)^{-2} - \mathrm{H}_m(\boldsymbol{\upsilon}^*)^{-2}\right)\mathrm{A}_m^\top(\boldsymbol{\upsilon}^*)\breve{D}_{mm}^{-1}\|\\ &\quad \leq \frac{1}{(1-\sqrt{\nu})^2}\breve{\delta}(\mathtt{r}^*)\|\mathrm{H}_m(\boldsymbol{\upsilon}_m^*)^{-1}\mathrm{A}_m^\top(\boldsymbol{\upsilon}^*)D(\boldsymbol{\upsilon}_m^*)^{-1}\| \leq \frac{\sqrt{\nu}+\breve{\delta}(\mathtt{r}^*)}{(1-\sqrt{\nu})^2}\breve{\delta}(\mathtt{r}^*).\end{aligned}$$

This gives the claim. □

Chapter 5

Convergence of an alternation maximization procedure

5.1 Introduction

This chapter presents two convergence results for an alternating maximization procedure to approximate M-estimators. It is largely based on [5]. We focus on finite dimensional parameter spaces $\varUpsilon \subseteq \mathbb{R}^{p^*}$ with $p^* = p+m \in \mathbb{N}$ being the full dimension, as infinite dimensional maximization problem are computationally anyways not feasible. As explained in Chapter 1 the alternating maximization procedure is used in situations where a direct computation of the full maximum estimator (ME) $\widetilde{\boldsymbol{\upsilon}} \in \mathbb{R}^{p^*}$ in (2.2.1) is not feasible or simply very difficult to implement. In such cases a workaround has to be found to calculate the profile in (1.0.3).

One prominent approach is - given some (data dependent) functional $\mathcal{L} : \mathbb{R}^p \times \mathbb{R}^m \to \mathbb{R}$ and an initial guess $\widetilde{\boldsymbol{\upsilon}}^{(0)} \in \mathbb{R}^{p+m}$ - to set for $k \in \mathbb{N}$

$$\begin{aligned} \widetilde{\boldsymbol{\upsilon}}^{(k,k+1)} &\overset{\text{def}}{=} (\widetilde{\boldsymbol{\theta}}^{(k)}, \widetilde{\boldsymbol{\eta}}^{(k+1)}) = \left(\widetilde{\boldsymbol{\theta}}^{(k)}, \operatorname*{argmax}_{\boldsymbol{\eta} \in \mathbb{R}^m} \mathcal{L}(\widetilde{\boldsymbol{\theta}}^{(k)}, \boldsymbol{\eta}) \right), \\ \widetilde{\boldsymbol{\upsilon}}^{(k,k)} &\overset{\text{def}}{=} (\widetilde{\boldsymbol{\theta}}^{(k)}, \widetilde{\boldsymbol{\eta}}^{(k)}) = \left(\operatorname*{argmax}_{\boldsymbol{\theta} \in \mathbb{R}^p} \mathcal{L}(\boldsymbol{\theta}, \widetilde{\boldsymbol{\eta}}^{(k)}), \widetilde{\boldsymbol{\eta}}^{(k)} \right). \end{aligned} \tag{5.1.1}$$

As we mentioned this "alternation maximization procedure" (or minimization) is a widely applied algorithm in many parameter estimation tasks (see [31], [41], [33] or [62]). Some natural questions arise: Does the sequence $(\widetilde{\boldsymbol{\theta}}^{(k)})$ converge to a limit that satisfies the same statistical properties as the profile estimator? And if the answer is yes, after how many steps does the sequence acquire these properties? Under what circumstances does the sequence actually converge to the global maximizer $\widetilde{\boldsymbol{\upsilon}}$? This problem is hard because the behavior of each step of the sequence is determined by the actual finite sample realization of the functional $\mathcal{L}(\cdot) = \mathcal{L}(\cdot, \mathbb{Y})$, where we usually suppress the data dependence to ease notation. To the authors' knowledge no general "convergence" result is available that answers the questions from above except for the treatment of specific models (see again [31], [41], [33] or [62]).

We address this difficulty via employing new finite sample techniques of Chapter 4 which allow to answer the above questions: with growing iteration number $k \in \mathbb{N}$ the estimators $\widetilde{\boldsymbol{\theta}}^{(k)}$ attain the same statistical properties as the profile M-estimator $\widetilde{\boldsymbol{\theta}}$ in (1.0.3) and Theorem 5.2.1 provides a choice of the necessary number of steps $K \in \mathbb{N}$. Under slightly stronger conditions on the structure of the model we can give a convergence result to the global maximizier that does not rely on unimodality. Further we can address the important question under which ratio of full dimension $p^* = p + m \in \mathbb{N}$ to sample size $n \in \mathbb{N}$ the sequence behaves as desired. For instance for smooth $\mathcal{L}$ our results on statistic properties of $\widetilde{\boldsymbol{\theta}}^{(k)}$ become sharp if $p^*/\sqrt{n}$ is small and convergence to the full maximizer already occurs if p^*/n is small.

We already pointed out in the introduction that the alternation maximization procedure can be understood as a special case of the Expectation Maximization algorithm (EM algorithm) as we will illustrate below. The EM algorithm itself was derived in [16] where particular versions of this approach are generalized. [16] also contains a variety of problems where an application of EM algorithm can be fruitful; for a brief history of the EM algorithm see [38] (Sect. 1.8). We briefly explain the EM algorithm. Take observations $\mathbb{X} \sim I\!P_{\boldsymbol{\theta}}$ from some parametric family $(I\!P_{\boldsymbol{\theta}}, \, \boldsymbol{\theta} \in \Theta)$. Assume that a parameter $\boldsymbol{\theta} \in \Theta$ is to be estimated as maximizer of the functional $\mathcal{L}_c(\boldsymbol{\theta}, \mathbb{X}) \in \mathbb{R}$, but that only $\mathbb{Y} \in \mathcal{Y}$ is observed, where $\mathbb{Y} = f_Y(\mathbb{X})$ is the image of the complete data set $\mathbb{X} \in \mathcal{X}$ under some map $f_Y : \mathcal{X} \to \mathcal{Y}$. Prominent examples for the map f_Y are projections onto some components of $\mathbb{X}$ if both $\mathbb{Y}$ and $\mathbb{X}$ are vector valued. The information lost under the map can be regarded as missing data or latent variables. As a direct maximization of the functional is impossible without knowledge of $\mathbb{X}$ the EM algorithm serves as a workaround. It consists of the iteration of tow steps: starting with some initial guess $\widetilde{\boldsymbol{\theta}}^{(0)}$ the k. "expectation step" derives the functional Q via

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(k)}) = I\!E_{\boldsymbol{\theta}^{(k)}}[\mathcal{L}_c(\boldsymbol{\theta}, \mathbb{X})|\mathbb{Y}],$$

which means that on the right hand side the conditional expectation is calculated under the distribution $I\!P_{\boldsymbol{\theta}^{(k)}}$. The k. "maximization step" then simply locates the maximizer $\boldsymbol{\theta}_{k+1}$ of Q .

Since the algorithm is very popular in applications a lot of research on its behavior has been done. We are only dealing with a special case of this procedure so we restrict ourselves to citing the well-known convergence result by Wu in [59]. Wu presents regularity conditions that ensure that $\mathcal{L}(\boldsymbol{\theta}^{(k+1)}|\mathbb{Y}) \geq \mathcal{L}(\boldsymbol{\theta}^{(k)}|\mathbb{Y})$ where

$$\mathcal{L}(\boldsymbol{\theta}|\mathbb{Y}) \overset{\text{def}}{=} I\!\!E\left[\mathcal{L}_c(\boldsymbol{\theta}, \mathbb{X})|\mathbb{Y} = f_Y(\mathbb{X})\right],$$

such that $\mathcal{L}(\boldsymbol{\theta}^{(k)}|\boldsymbol{Y}) \to \mathcal{L}^*(\mathbb{Y})$ for some limit value $\mathcal{L}^*(\mathbb{Y}) > 0$, that may depend on the starting point $\boldsymbol{\theta}^{(0)}$. Additionally Wu gives conditions that guarantee that the sequence $\boldsymbol{\theta}^{(k)}$ (possibly a sequence of sets) converges to $C(\mathcal{L}^*) \overset{\text{def}}{=} \{\boldsymbol{\theta}|\, \mathcal{L}(\boldsymbol{\theta}|\mathbb{Y}) = \mathcal{L}^*(\mathbb{Y})\}$. [16] show that the speed of convergence is linear in case of point valued $\boldsymbol{\theta}^{(k)}$ and of some differentiability criterion being met. A limitation of these results is that it is not clear whether $\mathcal{L}^* = \sup \mathcal{L}(\boldsymbol{\theta}|\mathbb{Y})$ and thus it is not guaranteed that $C(\mathcal{L}^*)$ is the desired MLE and not just some local maximum. Of course this problem disappears if $\mathcal{L}(\cdot|\mathbb{Y})$ is unimodal and the regularity conditions are met but this assumption may be too restrictive. See also [11] for convergence results along similar lines.

[6] is a recent work that presents a new way of addressing the properties of the EM sequence in a very general i.i.d. setting, based on concavity of $\boldsymbol{\theta} \mapsto I\!\!E_{\boldsymbol{\theta}^*}[\mathcal{L}_c(\boldsymbol{\theta}, \mathbb{X})]$. They assume that the functional $\mathcal{L}_c$ is concave and smooth enough (First order stability) and that for a sample $(\boldsymbol{Y}_i)$ with high probability an uniform bound of the kind

$$\sup_{\boldsymbol{\theta} \in B_{\mathtt{r}}(\boldsymbol{\theta}^*)} \left| \operatorname*{argmax}_{\boldsymbol{\theta}^\circ} \sum_{i=1}^n I\!\!E_{\boldsymbol{\theta}}[\mathcal{L}_c(\boldsymbol{\theta}^\circ, \mathbb{X})|\boldsymbol{Y}_i] - \operatorname*{argmax}_{\boldsymbol{\theta}^\circ} I\!\!E_{\boldsymbol{\theta}^*}[I\!\!E_{\boldsymbol{\theta}}[\mathcal{L}_c(\boldsymbol{\theta}^\circ, \mathbb{X})|\mathbb{Y}]] \right| \\ \leq \epsilon_n, \tag{5.1.2}$$

is satisfied. Under these assumptions, with high probability and some $\nu < 1$ it holds

$$\|\widetilde{\boldsymbol{\theta}}^{(k)} - \boldsymbol{\theta}^*\| \leq \nu^k \|\boldsymbol{\theta}^{(0)} - \boldsymbol{\theta}^*\| + \mathtt{C}\epsilon_n. \tag{5.1.3}$$

Unfortunately this does not answer our two questions to full satisfaction. First the bound (5.1.2) is rather high-level and has to be checked for each model, while we seek (and find) properties of the functional - such as smoothness and bounds on the moments of its gradient - that lead to comparably desirable behavior. Further with (5.1.3) it remains unclear whether for large $k \in \mathbb{N}$ the alternating sequence satisfies a Fisher expansion or whether a Wilks type phenomenon occurs. In particular it remains open which ratio of dimension to sample size ensures good performance of the procedure. Also the actual convergence of $\widetilde{\boldsymbol{\theta}}^{(k)} \to \boldsymbol{\theta}^*$ is not implied, as the right hand side in (5.1.3) is bounded from below by $\mathtt{C}\epsilon_n > 0$.

Remark 5.1.1. In the context of the alternating procedure the bound

(5.1.2) would read

$$\max_{\boldsymbol{\theta}^\circ \in B_{\mathrm{r}}(\boldsymbol{\theta}^*)} \left| \operatorname*{argmax}_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}, \widetilde{\boldsymbol{\eta}}_{\boldsymbol{\theta}^\circ}) - \operatorname*{argmax}_{\boldsymbol{\theta}} I\!E\mathcal{L}(\boldsymbol{\theta}, \widetilde{\boldsymbol{\eta}}_{\boldsymbol{\theta}^\circ}) \right| \le \epsilon_n,$$

which is still difficult to check.

To see that the procedure (5.1.1) is a special case of the EM algorithm we have to find the right triplet $(\mathbb{X}, f_Y, \mathcal{L}_c)$. For this we take $\mathbb{X} = \big(\boldsymbol{Z}, \mathbb{Y}\big)$ with $\boldsymbol{Z} \sim \operatorname{argmax}_{\boldsymbol{\eta}} \mathcal{L}\{(\boldsymbol{\theta}, \boldsymbol{\eta}), \mathbb{Y}\}$ under $I\!P_{\boldsymbol{\theta}}$. Further we set $f_Y(\mathbb{X}) = \mathbb{Y}$ and $\mathcal{L}_c(\boldsymbol{\theta}, \mathbf{X}) \overset{\text{def}}{=} \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\eta}, \mathbb{Y})$, where $\mathbf{X} = (\boldsymbol{\eta}, \mathbb{Y})$. Then we find

$$\begin{aligned} Q(\boldsymbol{\theta}, \widetilde{\boldsymbol{\theta}}^{(k-1)}) &= I\!E_{\widetilde{\boldsymbol{\theta}}^{(k-1)}}[\mathcal{L}_c(\boldsymbol{\theta}, \mathbb{X})|\mathbb{Y}] \\ &= I\!E_{\widetilde{\boldsymbol{\theta}}^{(k-1)}}\Big[\mathcal{L}_c\Big(\boldsymbol{\theta}, \operatorname*{argmax}_{\boldsymbol{\eta}} \mathcal{L}\{(\boldsymbol{\theta}^{(k-1)}, \boldsymbol{\eta}), \mathbb{Y}\}, \mathbb{Y}\Big)\Big|\mathbb{Y}\Big] \\ &= \mathcal{L}_c\Big(\boldsymbol{\theta}, \operatorname*{argmax}_{\boldsymbol{\eta}} \mathcal{L}\{(\widetilde{\boldsymbol{\theta}}^{(k-1)}, \boldsymbol{\eta}), \mathbb{Y}\}, \mathbb{Y}\Big) \\ &= \mathcal{L}(\boldsymbol{\theta}, \widetilde{\boldsymbol{\eta}}^{(k)}, \mathbb{Y}), \end{aligned}$$

and thus the resulting sequence is the same as in (5.1.1). Consequently the convergence results from above apply to our problem if the involved regularity criteria are met. But as noted these results do not tell us if the limit of the sequence $(\widetilde{\boldsymbol{\theta}}^{(k)})$ actually is the profile and the statistical properties of limit points are not clear as the error term $\mathtt{C}\epsilon_n$ in (5.1.3) determines the limit distribution and it is not obvious whether it is asymptotically $\sqrt{\chi^2}$-distributed.

The results of this chapter fill this gap for a wide range of settings. As we pointed out in the introduction we manage to establish that under mild conditions on the initial guess and the same conditions as in Chapter 4 the estimators $\widetilde{\boldsymbol{\theta}}^{(k)}$ satisfy a Fisher and Wilks expansion as shown for the profile ME in Theorem 4.2.2. Further we manage to show under slightly stronger smoothness conditions that $(\widetilde{\boldsymbol{\theta}}^{(k)}, \widetilde{\boldsymbol{\eta}}^{(k)})$ indeed approaches the ME $\widetilde{\boldsymbol{\upsilon}}$ with nearly linear convergence speed, i.e. $\|\mathcal{D}((\boldsymbol{\theta}^{(k)}, \boldsymbol{\eta}^{(k)}) - \widetilde{\boldsymbol{\upsilon}})\| \le \tau^{k/\log(k)}$ with some $0 < \tau < 1$ and $\mathcal{D}^2 = -I\!E\nabla^2\mathcal{L}(\boldsymbol{\upsilon}^*)$. The latter is not necessary for statistical inference on $\boldsymbol{\theta}^*$ but can be useful in the context of stochastic optimization, when it has to be ensured that the maximum is approached with growing number of iterations $k \in \mathbb{N}$.

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 $\widetilde{\boldsymbol{\upsilon}}^{(k,k)}$. Also we do not specify whether the elements of the resulting sequence are sets or single points. All statements made about properties of $\widetilde{\boldsymbol{\upsilon}}^{(k,k(+1))}$ are to be understood in the sense that they hold for "every point of $\widetilde{\boldsymbol{\upsilon}}^{(k,k(+1))}$ ".

It is worthy to point out two technical challenges of the analysis. First the sketched approach relies on (5.A.1). As all estimators $(\widetilde{\boldsymbol{\upsilon}}^{(k,k(+1))})$ are random this means that we need with some small $\beta > 0$

$$I\!\!P\left(\bigcap_{k\in\mathbb{N}_0}\left\{\widetilde{\boldsymbol{\upsilon}}^{(k,k)},\widetilde{\boldsymbol{\upsilon}}^{(k,k+1)}\in\{\|\mathcal{D}(\boldsymbol{\upsilon}-\boldsymbol{\upsilon}^{*})\|\le \mathrm{R}_0\}\right\}\right)\ge 1-\mathrm{e}^{-\mathtt{x}}.$$

This is not trivial but Theorem 3.3.2 serves the result thanks to the property $\mathcal{L}(\widetilde{\boldsymbol{\upsilon}}^{(k,k(+1))})\ge\mathcal{L}(\widetilde{\boldsymbol{\upsilon}}^{(0)})$. Furthermore, the main result 5.2.1 is formulated to hold for all $k\in\mathbb{N}_0$. This implies the need of a bound of the kind

$$I\!\!P\left(\bigcap_{k\in\mathbb{N}_0}\left\{\left\|\breve{D}^{-1}\{\breve{\nabla}\boldsymbol{\zeta}(\widetilde{\boldsymbol{\upsilon}}^{(k,k)})-\breve{\nabla}\boldsymbol{\zeta}(\boldsymbol{\upsilon}^{*})\}\right\|\le\epsilon(\mathtt{r}_k)\right\}\right)\ge 1-\mathrm{e}^{-\mathtt{x}},$$

with some small $\epsilon(\mathtt{r})>0$ that is decreasing if $\mathtt{r}>0$ shrinks. We manage to derive this result in the desired way using Theorem 3.5.7.

5.2 Main results

5.2.1 Introduction of important objects

In this section we introduce all objects and bounds that are relevant for Theorem 5.2.1. This section is quite technical but necessary to understand the results.

First consider the $p^{*}\times p^{*}$ matrices $\mathcal{D}^2$ and $\mathcal{V}^2$ from Section 4.2.1, which could be defined similarly to the Fisher information matrix:

$$\mathcal{D}^2\overset{\mathrm{def}}{=}-\nabla^2 I\!\!E\mathcal{L}(\boldsymbol{\upsilon}^{*}),\ \ \mathcal{V}^2\overset{\mathrm{def}}{=}\mathrm{Cov}(\nabla\mathcal{L}(\boldsymbol{\upsilon}^{*})).$$

We represent the information and covariance matrix in block form:

$$\mathcal{D}^2=\begin{pmatrix}D^2 & A\\ A^{\top} & \mathrm{H}^2\end{pmatrix},\ \ \mathcal{V}^2=\begin{pmatrix}V^2 & \mathcal{B}\\ \mathcal{B}^{\top} & Q^2\end{pmatrix}.$$

A crucial object is the constant $0\le\nu$ defined by

$$\|D^{-1}A\mathrm{H}^{-1}\|^2\overset{\mathrm{def}}{=}\nu,$$

which we assume to be smaller 1. It determines the speed of convergence of the alternating procedure (see Theorem 5.2.1). Define also the local sets

$$\Upsilon_{\circ}(\mathtt{r})\overset{\mathrm{def}}{=}\bigl\{\boldsymbol{\upsilon}:\,(\boldsymbol{\upsilon}-\boldsymbol{\upsilon}^{*})^{\top}\mathcal{D}^2(\boldsymbol{\upsilon}-\boldsymbol{\upsilon}^{*})\le\mathtt{r}^2\bigr\},$$
$$\widetilde{\Upsilon}_{\circ}(\mathtt{r})\overset{\mathrm{def}}{=}\bigl\{\boldsymbol{\upsilon}:\,(\boldsymbol{\upsilon}-\widetilde{\boldsymbol{\upsilon}})^{\top}\mathcal{D}^2(\boldsymbol{\upsilon}-\widetilde{\boldsymbol{\upsilon}})\le\mathtt{r}^2\bigr\},$$

and as in Chapter 4 define the radius $\mathtt{r}_0 > 0$ via

$$\mathtt{r}_0(\mathtt{x}) \overset{\text{def}}{=} \inf_{\mathtt{r} \ge 0} \left\{ I\!\!P \left(\operatorname*{argmax}_{\substack{\boldsymbol{\upsilon} \in \varUpsilon \\ \varPi_{\boldsymbol{\theta}} \boldsymbol{\upsilon} = \boldsymbol{\theta}^*}} \mathcal{L}(\boldsymbol{\upsilon}), \widetilde{\boldsymbol{\upsilon}} \in \varUpsilon_\circ(\mathtt{r}) \right) \ge 1 - \mathrm{e}^{-\mathtt{x}} \right\}. \tag{5.2.1}$$

Remark 5.2.1. This radius can be determined using conditions $(\mathcal{L}_{\mathtt{r}})$ and $(\mathcal{E}\mathtt{r})$ of Section 4.2.1 and Theorem 3.3.2 which would yield $\mathtt{r}_0(\mathtt{x}) = \mathtt{C}\sqrt{\mathtt{x} + p^*}$ with some constant $\mathtt{C} > 0$.

Furthermore, remember the $p \times p$ matrix $\breve{D}$ and the p-vectors $\breve{\nabla}_{\boldsymbol{\theta}}$ and $\breve{\boldsymbol{\xi}}$ as

$$\breve{D}^2 = D^2 - A\mathrm{H}^{-2}A^\top, \; \breve{\nabla}_{\boldsymbol{\theta}} = \nabla_{\boldsymbol{\theta}} - A\mathrm{H}^{-2}\nabla_{\boldsymbol{\eta}}, \; \breve{\boldsymbol{\xi}} = \breve{D}^{-1}\breve{\nabla}_{\boldsymbol{\theta}}.$$

For our estimations we need the constant

$$\mathfrak{z}(\mathtt{x}) \overset{\text{def}}{=} \mathfrak{z}(\mathtt{x}, I\!\!B) \vee \mathfrak{z}_Q(\mathtt{x}, 4p^*) \approx \sqrt{p^* + \mathtt{x}}, \quad I\!\!B^2 \overset{\text{def}}{=} \mathcal{D}^{-1}\mathcal{V}^2\mathcal{D}^{-1},$$

where $\mathfrak{z}(\mathtt{x}, \cdot)$ is explained in Section 3.4 and $\mathfrak{z}_Q(\mathtt{x}, \cdot)$ is defined in Equation (3.5.8).

Remark 5.2.2. The constant $\mathfrak{z}(\mathtt{x})$ is only introduced for ease of notation. This makes some bounds less sharp but allows to address all terms that are of order $\sqrt{p^* + \mathtt{x}}$ with one symbol. The constant $\mathfrak{z}(\mathtt{x}, I\!\!B)$ is comparable to the " $1 - \mathrm{e}^{-\mathtt{x}}$ "-quantile of the norm of $\mathcal{D}^{-1}\mathcal{V}\mathbf{X}$, where $\mathbf{X} \sim \mathcal{N}(0, I_{p^*})$, i.e. it is of order of the trace of $I\!\!B$. The constant $\mathfrak{z}_Q(\mathtt{x}, \mathbb{Q})$ arises as an exponential deviation bound for the supremum of a smooth process over a set with complexity described by $\mathbb{Q}$.

To bound the deviations of the points of the sequence $(\widetilde{\boldsymbol{\upsilon}}^{(k,k(+1))})$ we need the following radius:

$$\mathrm{R}_0(\mathtt{x}, \mathrm{K}_0) \overset{\text{def}}{=} \mathfrak{z}(\mathtt{x}) \vee \frac{6\nu_0}{\mathtt{b}(1-\nu)} \sqrt{\mathtt{x} + 2.4p^* + \frac{\mathtt{b}^2}{9\nu_0^2}\mathrm{K}_0(\mathtt{x}, \beta)}, \tag{5.2.2}$$

which will ensure $\{\widetilde{\boldsymbol{\upsilon}}^{(0)}, \widetilde{\boldsymbol{\upsilon}}^{(0,1)}, \ldots\} \subset \varUpsilon_\circ(\mathrm{R}_0)$, where $\mathrm{K}_0(\mathtt{x}) > 0$ is defined as

$$\mathrm{K}_0(\mathtt{x}, \beta) \overset{\text{def}}{=} \inf_{K > 0} \left\{ I\!\!P\left(\mathcal{L}(\widetilde{\boldsymbol{\upsilon}}^{(0)}, \boldsymbol{\upsilon}^*) \ge -K \right) \ge \beta(\mathtt{x}) \right\},$$

for some $\beta(\mathtt{x}) \to 0$ as $\mathtt{x} \to \infty$, see condition (A_1) in Section 5.2.2. Finally we (re)define the *parametric uniform spread* and the *semiparametric uniform* *spread*

$$\begin{aligned}
\diamondsuit_Q(\mathtt{r},\mathtt{x}) &\overset{\text{def}}{=} \left\{\delta(\mathtt{r})\mathtt{r} + 6\nu_1\omega(\mathfrak{z}_Q(\mathtt{x},4p^*) + 2\mathtt{r}^2)\right\},\\
\breve{\diamondsuit}_Q(\mathtt{r},\mathtt{x}) &\overset{\text{def}}{=} 6\nu_1\breve{\omega}\left(\mathfrak{z}_Q(\mathtt{x},2p^*+2p)^2 + 32\mathtt{r}^2\right) \\
&\qquad + \frac{32}{(1-\nu^2)^2}\breve{\delta}(4\mathtt{r})\mathtt{r}.
\end{aligned} \tag{5.2.3}$$

Remark 5.2.3. These objects are central to our analysis. $\breve{\diamondsuit}_Q(\mathtt{r},\mathtt{x})$ describes the accuracy of our main result of Theorem 5.2.1. It is small for not too large $\mathtt{r}$, if $\breve{\omega},\breve{\delta}$ introduced in $(\breve{\mathcal{E}}\mathcal{D}_1)$, $(\breve{\mathcal{L}}_0)$ from Section 4.2.1 are small (with Lemma 4.2.1 it suffices that ω,δ from $(\mathcal{E}\mathcal{D}_1)$, $(\mathcal{L}_0)$ are small). $\breve{\diamondsuit}_Q(\mathtt{r},\mathtt{x})$ is structurally slightly different from $\breve{\diamondsuit}(\mathtt{r},\mathtt{x})$ in (4.2.8) as it is based on Theorem 3.5.6 and allows a "uniform in k" formulation of our main result Theorem 5.2.1, but for moderate $\mathtt{x}\in\mathbb{R}_+$ they are of similar size.

5.2.2 Dependence on initial guess

Theorem 5.2.1 is only valid under the conditions from Section 4.2.1 and under some constraints on the quality of the initial guess $\tilde{\boldsymbol{\upsilon}}^{(0)}\in\mathbb{R}^{p^*}$ which we denote by (A_1), (A_2) and (A_3):

$(\mathbf{A}_1)$ With probability greater than $1-\beta_{(\mathbf{A})}(\mathtt{x})$ the initial guess satisfies

$$\mathcal{L}(\tilde{\boldsymbol{\upsilon}}^{(0)},\boldsymbol{\upsilon}^*) \geq -\mathrm{K}_0(\mathtt{x}),$$

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 4.2.1 hold for all $\mathtt{r}\leq \mathrm{R}_0(\mathtt{x},\mathrm{K}_0(\mathtt{x},\beta))$ where R_0 is defined in (5.2.2) with $\beta(\mathtt{x})=\beta_{(\mathbf{A})}(\mathtt{x})$.

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

$$c(\epsilon,\mathfrak{z}(\mathtt{x})) \overset{\text{def}}{=} \epsilon 4\mathtt{C}(\nu)\frac{1}{1-\nu}\left(\mathfrak{z}(\mathtt{x})+\epsilon\mathfrak{z}(\mathtt{x})^2\right) < 1, \tag{5.2.4}$$

$$c(\epsilon,\mathrm{R}_0) \overset{\text{def}}{=} \epsilon 4\mathtt{C}(\nu)\frac{1}{1-\nu}\mathrm{R}_0 < 1, \tag{5.2.5}$$

with

$$\mathtt{C}(\nu) \overset{\text{def}}{=} 2\sqrt{2}(1+\sqrt{\nu})(1-\sqrt{\nu})^{-1}. \tag{5.2.6}$$

Remark 5.2.4. One way of obtaining condition (A_1) is to show that $\widetilde{\boldsymbol{\upsilon}}_0 \in \varUpsilon_\circ(R_K)$ with probability greater than $1-\beta_{(\mathbf{A})}(\mathtt{x})$ for some finite $R_K(\mathtt{x}) \in \mathbb{R}$ and $0 \le \beta_{(\mathbf{A})}(\mathtt{x}) < 1$. Then (see Section 5.A.3)

$$\mathrm{K}_0(\mathtt{x}) \overset{\text{def}}{=} (1/2 + 12\nu_0\omega)R_K^2 + (\delta(R_K) + \mathfrak{z}(\mathtt{x}))R_K + 6\nu_0\omega\mathfrak{z}(\mathtt{x})^2.$$

Condition (A_1) is specified by conditions (A_2) and (A_3) and is fundamental, as it allows with dominating probability to concentrate the analysis on a local set $\varUpsilon_\circ\bigl(\mathrm{R}_0(\mathtt{x})\bigr)$ (see Theorem 3.3.2). Conditions (A_2) and (A_3) impose a bound on $\mathrm{R}_0(\mathtt{x})$ and thus on K_0 from (A_1). These conditions boil down to $\delta(\mathrm{R}_0) + \omega\mathrm{R}_0$ being significantly smaller than 1. Condition (A_3) ensures that the quality of the main results in Theorem 4.2.2, i.e. that $\breve{\diamondsuit}_Q(\mathtt{r}_k, \mathtt{x}) \approx \breve{\diamondsuit}(\mathtt{r}_0, \mathtt{x})$ under rather mild conditions on the size R_0, as we only need $\epsilon\mathrm{R}_0$ to be small. A violation of (A_2) would make it impossible to apply Theorem 3.5.6 which is the backbone of our proofs.

Remark 5.2.5. In case of iid observations with sample size $n \in \mathbb{N}$ one often has $\delta(\mathrm{R}_0) + \omega\mathrm{R}_0 \le \mathtt{C}\mathrm{R}_0(\mathtt{x})/\sqrt{n}$ which suggests at first glance that (A_2) and (A_3) are only a question of the sample size. But note that in case of iid observations the functional satisfies $n \approx -\mathcal{L}(\widetilde{\boldsymbol{\upsilon}}^{(0)}, \boldsymbol{\upsilon}^*)$ and $\mathrm{R}_0 \ge c\sqrt{-\mathcal{L}(\widetilde{\boldsymbol{\upsilon}}^{(0)}, \boldsymbol{\upsilon}^*)}$ such that the conditions (A_2) and (A_3) are not satisfied automatically with sufficiently large sample size. They are true conditions on the quality of the first guess.

5.2.3 Statistical properties of the alternating sequence

In this Section we present the first result of this Chapter, i.e. that the limit of the alternating sequence satisfies a finite sample Wilks Theorem and Fisher expansion.

To avoid distracting technicalities we impose one further merely technical condition:

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

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

Remark 5.2.6. Without this the calculation of $\mathrm{R}_0(\mathtt{x})$ in Section 5.A.3 would become technically more involved but no further insight would be gained.

Theorem 5.2.1. *Assume that the conditions* $(\mathcal{E}\mathcal{D}_0), (\mathcal{E}\mathcal{D}_1)$, $(\mathcal{L}_0)$, $(\mathcal{L}_{\mathtt{r}})$ *and* $(\mathcal{E}\mathtt{r})$ *of Section 4.2.1 are met with a constant* $\mathtt{b}(\mathtt{r}) \equiv \mathtt{b}$ *and where* $\mathcal{V}_0^2 = \operatorname{Cov}\bigl(\nabla\mathcal{L}(\boldsymbol{\upsilon}^*)\bigr)$, $\mathcal{D}_0^2 = -\nabla^2 I\!\!E\mathcal{L}(\boldsymbol{\upsilon}^*)$ *and where* $\boldsymbol{\upsilon}^\circ = \boldsymbol{\upsilon}^*$. *Assume that* $(\breve{\mathcal{E}}\mathcal{D}_1)$ *and* $(\breve{\mathcal{L}}_0)$ *are met and that* $R_0 \vee 4\mathtt{r}_0 \le \mathtt{r}^*$. *Furthermore assume* (B_1)

and that the initial guess satisfies (A_1) *and* (A_2) *of Section 5.2.2. Then it holds with probability greater than* $1 - 8\mathrm{e}^{-\mathtt{x}} - \beta_{(\mathbf{A})}$ *for all* $k \in \mathbb{N}$

$$\left\|\breve{D}\left(\widetilde{\boldsymbol{\theta}}^{(k)} - \boldsymbol{\theta}^*\right) - \breve{\boldsymbol{\xi}}\right\| \le \breve{\Diamond}_Q(\mathtt{r}_k, \mathtt{x}), \tag{5.2.7}$$

$$\left|2\breve{L}(\widetilde{\boldsymbol{\theta}}^{(k)}, \boldsymbol{\theta}^*) - \|\breve{\boldsymbol{\xi}}\|^2\right| \le 9\left(\|\breve{\boldsymbol{\xi}}\| + \breve{\Diamond}_Q(\mathtt{r}_k, \mathtt{x})\right)\breve{\Diamond}_Q(\mathtt{r}_k, \mathtt{x}), \tag{5.2.8}$$

where

$$\mathtt{r}_k \le 2\sqrt{2}(1-\sqrt{\nu})^{-1}\left\{(\mathfrak{z}(\mathtt{x}) + \Diamond_Q(R_0, \mathtt{x})) + (1+\sqrt{\nu})\nu^k R_0(\mathtt{x})\right\}.$$

If further condition (A_3) *is satisfied then* (5.2.7) *and* (5.2.8) *are met with*

$$\mathtt{r}_k \le \left(\mathtt{C}(\nu) + \frac{4\mathtt{C}(\nu)^3 c(\epsilon, \mathfrak{z}(\mathtt{x}))}{1 - c(\epsilon, \mathfrak{z}(\mathtt{x}))}\right)\left(\mathfrak{z}(\mathtt{x}) + \epsilon\mathfrak{z}(\mathtt{x})^2\right)$$
$$+\nu^k\left(\mathtt{C}(\nu) + \nu\frac{4\mathtt{C}(\nu)^3 c(\epsilon, R_0)}{1 - c(\epsilon, R_0)}\right)R_0.$$

In particular this means that if

$$k \ge \frac{2\log(\mathfrak{z}(\mathtt{x})) - \log\{2R_0(\mathtt{x}, K_0)\}}{\log(\nu)},$$

we have with $\mathfrak{z}(\mathtt{x})^2 \le \mathtt{C}_{\mathfrak{z}}(p^* + \mathtt{x})$ *and some constant* $\mathtt{C} > 0$

$$\breve{\Diamond}_Q(\mathtt{r}_k, \mathtt{x}) \approx \breve{\Diamond}_Q\left(\mathtt{C}\sqrt{p^* + \mathtt{x}}, \mathtt{x}\right).$$

Remark 5.2.7. Note that the results are very similar to those in Theorem 4.2.2 for the profile M-estimator $\widetilde{\boldsymbol{\theta}}$. This is evident after noting that (ignoring terms of the order $\epsilon\mathfrak{z}(\mathtt{x})$)

$$\mathtt{r}_k \lesssim \mathtt{C}(\nu)\left(\mathfrak{z}(\mathtt{x}) + \nu^k(\mathrm{R}_0 + \mathtt{C}\epsilon\mathrm{R}_0^2)\right),$$

which for large $k \in \mathbb{N}$ means $\mathtt{r}_k \lesssim \mathtt{C}(\nu)\mathfrak{z}(\mathtt{x}) \le \mathtt{C}'\sqrt{p^* + \mathtt{x}}$.

Remark 5.2.8. Concerning the properties of $\breve{\boldsymbol{\xi}} \in \mathbb{R}^p$ we refer to Remark 4.2.11.

Remark 5.2.9. In general an exact numerical computation of

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

is not possible. To address this define $\widehat{\theta}(\boldsymbol{\eta})$ and $\widehat{\eta}(\boldsymbol{\theta})$ as the numerical approximations to $\theta(\boldsymbol{\eta})$ and $\eta(\boldsymbol{\theta})$ and assume that

$$\|D(\widehat{\theta}(\boldsymbol{\eta}) - \theta(\boldsymbol{\eta}))\| \le \tau, \ \text{ and } \ \|\mathrm{H}(\widehat{\eta}(\boldsymbol{\theta}) - \eta(\boldsymbol{\theta}))\| \le \tau,$$

for all $\boldsymbol{\theta} \in \varUpsilon_{0,\boldsymbol{\theta}}(\mathrm{R}_0) \overset{\mathrm{def}}{=} \{\boldsymbol{\upsilon} \in \varUpsilon_{\circ}(\mathrm{R}_0),\, \varPi_{\boldsymbol{\theta}}\boldsymbol{\upsilon} = \boldsymbol{\theta}\}$ and $\boldsymbol{\eta} \in \varUpsilon_{0,\boldsymbol{\eta}}(\mathrm{R}_0) \overset{\mathrm{def}}{=} \{\boldsymbol{\upsilon} \in \varUpsilon_{\circ}(\mathrm{R}_0),\, \varPi_{\boldsymbol{\eta}}\boldsymbol{\upsilon} = \boldsymbol{\eta}\}$. Then we can easily modify the proof of Theorem 5.2.1 via adding $\mathtt{C}(\nu)\tau$ to the error terms and the radii $\mathtt{r}_k$, where $\mathtt{C}(\nu)$ is some rational function of ν.

Remark 5.2.10. Note that under condition (A_3) the size of $\mathtt{r}_k$ for $k \to \infty$ does not depend on $\mathrm{R}_0 > 0$. Thus as long as $\epsilon\mathrm{R}_0$ is small enough the quality of the initial guess no longer affects the statistical properties of the sequence $(\boldsymbol{\theta}^{(k)})$ for large $k \in \mathbb{N}$.

5.2.4 Convergence to the ME

Even though Theorem 5.2.1 tells us that the statistical properties of the alternating sequence resemble those of its target, the profile ME, it is an interesting question if the underlying approach allows to identify conditions under which the sequence actually attains the maximizer $\widetilde{\boldsymbol{\upsilon}}$. Without further assumptions Theorem 5.2.1 yields the following Corollary:

Corollary 5.2.2. *Under the assumptions of Theorem 5.2.1 it holds with probability greater than* $1 - 8\mathrm{e}^{-\mathtt{x}} - \beta_{(\mathbf{A})}$

$$\|\breve{D}(\widetilde{\boldsymbol{\theta}} - \widetilde{\boldsymbol{\theta}}^{(k)})\| \le \breve{\diamondsuit}_Q(\mathtt{r}_k, \mathtt{x}) + \breve{\diamondsuit}(\mathtt{r}_0, \mathtt{x}),$$

where $\mathtt{r}_0 > 0$ *is defined in* (5.2.1) *and*

$$\breve{\diamondsuit}(\mathtt{r}, \mathtt{x}) \overset{\mathrm{def}}{=} \frac{8}{(1-\nu^2)^2}\breve{\delta}(\mathtt{r})\mathtt{r} + 6\nu_1\breve{\omega}\mathfrak{z}_1(\mathtt{x}, 2p^* + 2p)\mathtt{r}.$$

Remark 5.2.11. The value $\mathfrak{z}_1(\mathtt{x}, \cdot)$ is defined in (3.5.6).

Corollary 5.2.2 is a first step in the direction of an actual convergence result but the gap $\breve{\diamondsuit}_Q(\mathtt{r}_k, \mathtt{x}) + \breve{\diamondsuit}(\mathtt{r}_0, \mathtt{x})$ is not a zero sequence in $k \in \mathbb{N}$. It turns out that it is possible to prove convergence to the ME at the cost of assuming more smoothness of the functional $\mathcal{L}$ and using the right bound for the maximal eigenvalue of the Hessian $\nabla^2\mathcal{L}(\boldsymbol{\upsilon}^*)$.

Consider the following condition, that basically quantifies how "well behaved" the second derivative $\nabla^2(\mathcal{L} - I\!E\mathcal{L})$ is:

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

$$\sup_{\boldsymbol{\upsilon},\boldsymbol{\upsilon}' \in \varUpsilon_{\circ}(\mathtt{r})} \sup_{\|\gamma_1\|=1} \sup_{\|\gamma_2\|=1} \log I\!E \exp\left\{\frac{\mu\gamma_1^{\top}\mathcal{D}^{-1}\{\nabla^2\boldsymbol{\zeta}(\boldsymbol{\upsilon}) - \nabla^2\boldsymbol{\zeta}(\boldsymbol{\upsilon}')\}\gamma_2}{\omega_2\|\mathcal{D}(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}')\|}\right\}$$
$$\le \frac{\nu_2^2\mu^2}{2}.$$

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

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

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

$$\varkappa(\mathtt{x}, \mathrm{R}_0) \stackrel{\text{def}}{=} \frac{2\sqrt{2}(1+\sqrt{\nu})}{\sqrt{1-\nu}}\Big[\delta(\mathrm{R}_0) + 9\omega_2\nu_2\|\mathcal{D}^{-1}\|\mathfrak{z}_1(\mathtt{x}, 6p^*)\mathrm{R}_0$$

$$+ \|\mathcal{D}^{-1}\|\mathfrak{z}\left(\mathtt{x}, \nabla^2\mathcal{L}(\boldsymbol{\upsilon}^*)\right)\Big],$$

where $\mathfrak{z}_1(\mathtt{x}, \cdot)$ is defined in (3.5.6). With these definitions we can prove the following Theorem:

Theorem 5.2.3. *Let the conditions* $(\mathcal{E}\mathcal{D}_2)$ *,* $(\mathcal{L}_0)$ *,* $(\mathcal{L}_{\mathtt{r}})$ *and* $(\mathcal{E}\mathtt{r})$ *be met with a constant* $\mathtt{b}(\mathtt{r}) \equiv \mathtt{b}$ *and where* $\mathcal{D}^2 = -\nabla^2 I\!\!E\mathcal{L}(\boldsymbol{\upsilon}^*)$ *and* $\boldsymbol{\upsilon}^* = \boldsymbol{\upsilon}^\circ$ *. Furthermore suppose* (B_1) *and that the initial guess satisfies* (A_1) *and* (A_2) *. Assume that* $\varkappa(\mathtt{x}, R_0) < (1-\nu)$ *. Then*

$$I\!\!P\left(\bigcap_{k\in\mathbb{N}}\left\{\widetilde{\boldsymbol{\upsilon}}^{(k,k(+1))} \in \widetilde{\varUpsilon_\circ}(\mathtt{r}_k^*)\right\}\right) \geq 1 - 3\mathrm{e}^{-\mathtt{x}} - \beta_{(\mathbf{A})},$$

where

$$\mathtt{r}_k^* \leq \begin{cases} \nu^k \frac{4\sqrt{2}}{1-\varkappa(\mathtt{x},R_0)k} R_0, & \varkappa(\mathtt{x}, R_0)k \leq 1, \\ \nu^{\frac{k}{\log(k)}\log\left(\frac{1-\nu}{\varkappa(\mathtt{x},R_0)}\right)} c_k R_0, & \textit{otherwise,} \end{cases} \tag{5.2.9}$$

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

Remark 5.2.12. This means that we obtain nearly linear convergence to the global maximizer $\widetilde{\boldsymbol{\upsilon}}$.

Remark 5.2.13. As in Remark 5.2.9 if no exact numerical computation of the stepwise maximizers is possible we can easily modify the proof of Theorem 5.2.3 via adding $\mathtt{C}(\nu)\tau$ to $\varkappa(\mathtt{x}, \mathrm{R}_0)$ to address that case.

Remark 5.2.14. For the case that $\mathcal{L}(\boldsymbol{\upsilon}) = \sum_{i=1}^n \ell_i(\boldsymbol{\upsilon})$ with a sum of independent marginal functionals $\ell_i : \varUpsilon \to \mathbb{R}$ we can use Corollary 3.7 of [56] to obtain

$$\mathfrak{z}\left(\mathtt{x}, \nabla^2\mathcal{L}(\boldsymbol{\upsilon}^*)\right) = \sqrt{2\tau}\nu_3\sqrt{\mathtt{x} + p^*},$$

if for some sequence of matrices $(\boldsymbol{A}_i) \subset \mathbb{R}^{p^* \times p^*}$

$$\log I\!\!E \exp \lambda \nabla^2 \ell_i(\boldsymbol{\upsilon}^*) \preceq \nu_3^2\lambda^2/2\boldsymbol{A}_i, \quad \|\sum_{i=1}^n \boldsymbol{A}_i\| \leq \tau.$$

In case of smooth i.i.d models this means that

$$\begin{aligned} \varkappa(\mathtt{x}, \mathrm{R}_0) &\le \mathtt{C}(\mathrm{R}_0 + \mathtt{x} + \log(p^*))/\sqrt{n} + \mathtt{C}\mathrm{R}_0\sqrt{\mathtt{x} + p^*}/n, \\ \text{i.e. } \varkappa(\mathtt{x}, \mathrm{R}_0) &= O((\mathtt{x} + \mathrm{R}_0 + \log(p^*))/\sqrt{n}, \text{ if } p^* + \mathtt{x} = o(n). \end{aligned}$$

Remark 5.2.15. It may happen that $\varkappa(\mathtt{x}, \mathrm{R}_0)/(1-\nu)$ is very close to 1. In that case the obtained convergence is rather slow. But a close look at the proof of Theorem 5.2.3 reveals that this can be improved using Lemma 5.A.4. For this purpose assume that $\delta(\mathtt{r})/\mathtt{r} \vee 6\nu_1\omega_2 \le \epsilon$ for some $\epsilon > 0$ and assume (A_3) from Section 5.2.2. Bound $\mathtt{r}^* \le \mathtt{C}(\mathfrak{z}(\mathtt{x}) + \nu^k \mathrm{R}_0)$ with $\mathtt{r}_k^*$ defined in (5.A.6) and with some constant $\mathtt{C} > 0$. Then the result of Theorem 5.2.3 is true with $\varkappa(\mathtt{x}, \mathtt{C}\mathfrak{z}(\mathtt{x}))$ instead of $\varkappa(\mathtt{x}, \mathrm{R}_0)$ and with probability greater $1 - 10\mathrm{e}^{-\mathtt{x}}$. See Remark 5.A.3 for more details.

5.2.5 Critical dimension

As in Section 4.2.5 we want to address the issue of *critical parameter dimensions* when the full dimension p^* grows with the sample size n. We write $p^* = p_n$. The results of Theorem 5.2.1 are accurate if the spread function $\breve{\diamondsuit}_Q(\mathtt{r}_k, \mathtt{x})$ in (5.2.3) is small. The critical size of p^* then depends on the exact bounds on $\breve{\delta}(\cdot)$ and $\breve{\omega}$. In the i.i.d setting one usually has $\breve{\delta}(\mathtt{r})/\mathtt{r} \asymp \breve{\omega} \asymp 1/\sqrt{n}$ such that $\breve{\diamondsuit}(\mathtt{r}_k, \mathtt{x}) \asymp p^*/\sqrt{n}$ for large $k \in \mathbb{N}$; see Section 4.2.4. In other words, one needs that "p^{*2}/n is small" to obtain an accurate non-asymptotic version of the Wilks phenomenon and the Fisher Theorem for the limit of the alternating sequence. This is not surprising because because good performance of the ME itself can only be guaranteed if "p^{*2}/n is small" as shown in Section 4.2.5. There are examples where the pME only satisfies a Wilks- or Fisher result if "p^{*2}/n is small", such that in any of those settings the alternating sequence started in the global maximizer does not admit an accurate Wilks- or Fisher expansion.

The constraint $\varkappa(\mathtt{x}, \mathrm{R}_0) < (1-\nu)$ of Theorem 5.2.3 for the convergence of the sequence to the global maximizer means that one needs $p^*/n \ll 1$ in the smooth i.i.d. setting if $\mathrm{R}_0 \le \mathtt{C}_{\mathrm{R}_0}\sqrt{p^* + \mathtt{x}}$. Furthermore in the smooth i.i.d. setting the speed of convergence in Theorem 5.2.3 decreases if p^*/n grows. Unfortunately we were unable to find an example that meets the conditions of Section 4.2.1 and where no convergence occurs if p^*/n tends to infinity. Whether this dimension effect on the convergence is an artifact of our proofs or indeed a property of the alternating procedure remains an open question.

5.A Proofs

5.A.1 Proof of Theorem 5.2.1

In this section we will proof Theorem 5.2.1. Before we start with the actual proof we want to explain the ideo of the proof and sketch the strategy.

5.A.2 Idea of the proof

To motivate the approach - and hopefully to ease understanding - first consider the toy model

$$\mathbb{Y} = \boldsymbol{\upsilon}^* + \boldsymbol{\varepsilon}, \text{ where } \boldsymbol{\varepsilon} \sim \mathcal{N}(0, \mathbb{F}_{\boldsymbol{\upsilon}^*}^{-2}), \quad \mathbb{F}_{\boldsymbol{\upsilon}^*}^2 =: \begin{pmatrix} \mathbb{F}_{\boldsymbol{\theta}^*}^2 & A \\ A^\top & \mathbb{F}_{\boldsymbol{\eta}^*}^2 \end{pmatrix}.$$

In this case we set $\mathcal{L}$ to be the true log likelihood of the observations

$$\mathcal{L}(\boldsymbol{\upsilon}, \mathbb{Y}) = -\|\mathbb{F}_{\boldsymbol{\upsilon}^*}(\boldsymbol{\upsilon}^* - \mathbb{Y})\|^2/2.$$

With any starting initial guess $\widetilde{\boldsymbol{\upsilon}}^{(0)} \in \mathbb{R}^{p+m}$ we obtain in (5.1.1) for $k \in \mathbb{N}$ and the usual first order criterion of maximality the following two equations

$$\begin{aligned} \mathbb{F}_{\boldsymbol{\theta}^*}(\widetilde{\boldsymbol{\theta}}^{(k)} - \boldsymbol{\theta}^*) &= \ I_{\boldsymbol{\theta}^*}\boldsymbol{\varepsilon}_{\boldsymbol{\theta}} + \mathbb{F}_{\boldsymbol{\theta}^*}^{-1} A(\widetilde{\boldsymbol{\eta}}^{(k)} - \boldsymbol{\eta}^*), \\ \mathbb{F}_{\boldsymbol{\eta}^*}(\widetilde{\boldsymbol{\eta}}^{(k+1)} - \boldsymbol{\eta}^*) &= I_{\boldsymbol{\eta}^*}\boldsymbol{\varepsilon}_{\boldsymbol{\eta}} + \mathbb{F}_{\boldsymbol{\eta}^*}^{-1} A^\top(\widetilde{\boldsymbol{\theta}}^{(k)} - \boldsymbol{\theta}^*). \end{aligned}$$

Combining these two equations we derive, assuming $\|\mathbb{F}_{\boldsymbol{\theta}^*}^{-1} A \mathbb{F}_{\boldsymbol{\eta}^*}^{-2} A^\top I_{\boldsymbol{\theta}^*}^{-1}\| =: \|\boldsymbol{M}_0\| = \nu < 1$

$$\begin{aligned} \mathbb{F}_{\boldsymbol{\theta}^*}(\widetilde{\boldsymbol{\theta}}^{(k)} - \boldsymbol{\theta}^*) &= \mathbb{F}_{\boldsymbol{\theta}^*}^{-1}(\mathbb{F}_{\boldsymbol{\theta}^*}^2 \boldsymbol{\varepsilon}_{\boldsymbol{\theta}} - A\boldsymbol{\varepsilon}_{\boldsymbol{\eta}}) + \mathbb{F}_{\boldsymbol{\theta}^*}^{-1} A \mathbb{F}_{\boldsymbol{\eta}^*}^{-1} A^\top \mathbb{F}_{\boldsymbol{\theta}^*}^{-1} \mathbb{F}_{\boldsymbol{\theta}^*}(\widetilde{\boldsymbol{\theta}}_{k-1} - \boldsymbol{\theta}^*) \\ &= \sum_{l=1}^{k} \boldsymbol{M}_0^{k-l} \mathbb{F}_{\boldsymbol{\theta}^*}^{-1}(\mathbb{F}_{\boldsymbol{\theta}^*}^2 \boldsymbol{\varepsilon}_{\boldsymbol{\theta}} - A\boldsymbol{\varepsilon}_{\boldsymbol{\eta}}) \\ &\quad + \boldsymbol{M}_0^k \mathbb{F}_{\boldsymbol{\theta}^*}(\widetilde{\boldsymbol{\theta}}^{(0)} - \boldsymbol{\theta}^*) \to \mathbb{F}_{\boldsymbol{\theta}^*}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*). \end{aligned}$$

Because the limit $\widehat{\boldsymbol{\theta}}$ is independent of the initial point $\widetilde{\boldsymbol{\upsilon}}^{(0)}$ and because the profile $\widetilde{\boldsymbol{\theta}}$ is a fix point of the procedure the unique limit satisfies $\widehat{\boldsymbol{\theta}} = \widetilde{\boldsymbol{\theta}}$. The argument is based on the fact that in this setting the functional is quadratic such that the gradient satisfies

$$\nabla\mathcal{L}(\boldsymbol{\upsilon}) = \mathbb{F}_{\boldsymbol{\upsilon}^*}^2(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^*) + \mathbb{F}_{\boldsymbol{\upsilon}^*}^2 \boldsymbol{\varepsilon}.$$

Any smooth function is quadratic around its maximizer which motivates a local linear approximation of the gradient of the functional $\mathcal{L}$ to derive our results with similar arguments. This is done in the proof of Theorem 5.2.1.

First it is ensured that the whole sequence $(\widetilde{\boldsymbol{\upsilon}}^{(k,k(+1))})_{k\in\mathbb{N}_0}$ satisfies for some $\mathrm{R}_0(\mathtt{x}) > 0$ and with probability greater than $1 - \mathrm{e}^{-\mathtt{x}}$

$$\{\widetilde{\boldsymbol{\upsilon}}^{(k,k(+1))},\, k \in \mathbb{N}_0\} \subset \{\|\mathcal{D}(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^*)\| \le \mathrm{R}_0(\mathtt{x})\}, \tag{5.A.1}$$

where $\mathcal{D}^2 \overset{\mathrm{def}}{=} \nabla^2 I\!\!E\mathcal{L}(\boldsymbol{\upsilon}^*)$; here we use Theorem 3.3.2. In the second step we approximate with $\zeta = \mathcal{L} - I\!\!E\mathcal{L}$

$$\mathcal{L}(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^*) = \nabla\zeta(\boldsymbol{\upsilon}^*)(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^*) - \|\mathcal{D}(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^*)\|^2/2 + \alpha(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^*), \tag{5.A.2}$$

where $\alpha(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^*)$ is defined by (5.A.2). Similar to the toy case above this allows to use the first order criterion of maximality and (5.A.1) to obtain a bound of the kind

$$\begin{aligned}\|\mathcal{D}(\boldsymbol{\upsilon}^{(k,k)} - \boldsymbol{\upsilon}^*)\| &\le \mathtt{C}\sum_{l=0}^{k}\nu^l\Big(\|\mathcal{D}^{-1}\nabla\zeta(\boldsymbol{\upsilon}^*)\| + |\alpha(\boldsymbol{\upsilon}^{(l,l)}, \boldsymbol{\upsilon}^*)|\Big)\\ &\le \mathtt{C}_1\big(\|\mathcal{D}^{-1}\nabla\zeta(\boldsymbol{\upsilon}^*)\| + \epsilon(\mathrm{R}_0)\big) + \nu^k\mathrm{R}_0 \overset{\mathrm{def}}{=} \mathtt{r}_k.\end{aligned}$$

This is done in Lemma 5.A.3 using the results from Chapter 4 to show that $\epsilon(\mathrm{R}_0)$ is small. Having established

$$I\!\!P\left(\bigcap_{k\in\mathbb{N}}\Big\{\boldsymbol{\upsilon}^{(k,k(+1))} \subset \{\|\mathcal{D}(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^*)\| \le \mathtt{r}_k\}\Big\}\right) \ge 1 - 2\mathrm{e}^{-\mathtt{x}},$$

the same arguments as in the proof of Theorem 4.2.2 allow to obtain our first main result. For convergence to the full ME $\widetilde{\boldsymbol{\upsilon}}$ similar arguments are used. The only difference is that instead of (5.A.2) we use the approximation

$$\mathcal{L}(\boldsymbol{\upsilon}, \widetilde{\boldsymbol{\upsilon}}) = -\|\mathcal{D}(\boldsymbol{\upsilon} - \widetilde{\boldsymbol{\upsilon}})\|^2/2 + \alpha'(\boldsymbol{\upsilon}, \widetilde{\boldsymbol{\upsilon}}),$$

exploiting that $\nabla\mathcal{L}(\widetilde{\boldsymbol{\upsilon}}) \equiv 0$, which allows to obtain actual convergence to the ME.

5.A.3 A desireable set

The first step of the proof is to find a desirable set $\Omega(\mathtt{x}) \subset \Omega$ of high probability, on which a linear approximation of the gradient of the functional $\mathcal{L}(\boldsymbol{\upsilon})$ can be carried out with sufficient accuracy. Once this set is found all subsequent analysis concerns events in $\Omega(\mathtt{x}) \subset \Omega$.

For this purpose define the set

$$\begin{aligned}
\varOmega(\mathtt{x}) &= \bigcap_{k=0}^{\infty}(C^{(k,k)}\cap C^{(k,k+1)})\cap C(\nabla) \qquad (5.A.3)\\
&\quad \cap\{\mathcal{L}(\widetilde{\boldsymbol{\upsilon}}^{(0)},\boldsymbol{\upsilon}^*)\geq -\mathrm{K}_0(\mathtt{x})\},\\
C^{(k,k(+1))} &= \Big\{\|\mathcal{D}(\widetilde{\boldsymbol{\upsilon}}^{(k,k(+1))}-\boldsymbol{\upsilon}^*)\|\leq \mathrm{R}_0(\mathtt{x}),\ \|D(\widetilde{\boldsymbol{\theta}}^{(k)}-\boldsymbol{\theta}^*)\|\leq \mathrm{R}_0(\mathtt{x}),\\
&\quad \|\mathrm{H}(\widetilde{\boldsymbol{\eta}}^{(k(+1))}-\boldsymbol{\eta}^*)\|\leq \mathrm{R}_0(\mathtt{x})\Big\},\\
C(\nabla) &= \bigcap_{\mathtt{r}\leq \mathrm{R}_0(\mathtt{x})}\left\{\sup_{\boldsymbol{\upsilon}\in\varUpsilon_\circ(\mathtt{r})}\left\{\frac{1}{6\omega\nu_1}\|\mathcal{Y}(\boldsymbol{\upsilon})\|-2\mathtt{r}^2\right\}\leq \mathfrak{z}_Q(\mathtt{x},4p^*)^2\right\}\\
&\quad \bigcap_{\mathtt{r}\leq \mathrm{R}_0(\mathtt{x})}\left\{\sup_{\boldsymbol{\upsilon}\in\varUpsilon_\circ(\mathtt{r})}\left\{\frac{1}{6\breve{\omega}\breve{\nu}_1}\|\breve{\mathcal{Y}}(\boldsymbol{\upsilon})\|-2\mathtt{r}^2\right\}\leq \mathfrak{z}_Q(\mathtt{x},2p^*+2p)^2\right\}\\
&\quad \cap\Big\{\max\{\|\mathcal{D}^{-1}\nabla\mathcal{L}\|,\|D^{-1}\nabla_{\boldsymbol{\theta}}\mathcal{L}\|,\|\mathrm{H}^{-1}\nabla_{\boldsymbol{\eta}}\mathcal{L}\|\}\leq \mathfrak{z}(\mathtt{x})\Big\}\\
&\quad \cap\{\widetilde{\boldsymbol{\upsilon}},\widetilde{\boldsymbol{\upsilon}}_{\boldsymbol{\theta}^*}\in\varUpsilon_\circ(\mathtt{r}_0(\mathtt{x}))\}.
\end{aligned}$$

For $\zeta(\boldsymbol{\upsilon})=\mathcal{L}(\boldsymbol{\upsilon})-\mathbb{E}\mathcal{L}(\boldsymbol{\upsilon})$ the semiparametric normalized stochastic gradient gap is defined as

$$\breve{\mathcal{Y}}(\boldsymbol{\upsilon})=\breve{D}^{-1}\Big(\breve{\nabla}_{\boldsymbol{\theta}}\zeta(\boldsymbol{\upsilon})-\breve{\nabla}_{\boldsymbol{\theta}}\zeta(\boldsymbol{\upsilon}^*)\Big).$$

the parametric normalized stochastic gradient gap $\mathcal{Y}(\boldsymbol{\upsilon})$ is defined as

$$\mathcal{Y}(\boldsymbol{\upsilon})=\mathcal{D}_0^{-1}\Big(\nabla\zeta(\boldsymbol{\upsilon})-\nabla\zeta(\boldsymbol{\upsilon}^*)\Big),$$

and $\mathtt{r}_0(\mathtt{x})>0$ is chosen such that $\mathbb{P}(\widetilde{\boldsymbol{\upsilon}},\widetilde{\boldsymbol{\upsilon}}_{\boldsymbol{\theta}^*}\in\varUpsilon_\circ(\mathtt{r}_0))\geq 1-\mathrm{e}^{-\mathtt{x}}$, where

$$\widetilde{\boldsymbol{\upsilon}}_{\boldsymbol{\theta}^*}\overset{\mathrm{def}}{=}\operatorname*{argmax}_{\substack{\boldsymbol{\upsilon}\in\varUpsilon\\ \varPi_{\boldsymbol{\theta}}\boldsymbol{\upsilon}=\boldsymbol{\theta}^*}}\mathcal{L}(\boldsymbol{\upsilon}).$$

Remark 5.A.1. We intersect the set with the event $\{\widetilde{\boldsymbol{\upsilon}},\widetilde{\boldsymbol{\upsilon}}_{\boldsymbol{\theta}^*}\in\varUpsilon_\circ(\mathtt{r}_0)\}$ where we a priory demand $\mathtt{r}_0(\mathtt{x})>0$ to be chosen such that $\mathbb{P}(\widetilde{\boldsymbol{\upsilon}},\widetilde{\boldsymbol{\upsilon}}_{\boldsymbol{\theta}^*}\in\varUpsilon_\circ(\mathtt{r}_0))\geq 1-\mathrm{e}^{-\mathtt{x}}$. Note that condition $(\mathcal{E}\mathtt{r})$ together with $(\mathcal{L}\mathtt{r})$ allow to set $\sqrt{p^*+\mathtt{x}}\approx\mathtt{r}_0\leq\mathrm{R}_0$ (see Theorem 3.3.2).

In Section 5.A.3 we show that this set is of probability greater than $1-8\mathrm{e}^{-\mathtt{x}}-\beta_{(\boldsymbol{A})}$. We want to explain the purpose of this set along the architecture of the proof of our main theorem.

$\{\mathcal{L}(\widetilde{\boldsymbol{\upsilon}}^{(0)},\boldsymbol{\upsilon}^*)\geq -\mathbf{K}_0(\mathtt{x})\}$**:** This set ensures, that the first guess satisfies

$$\mathcal{L}(\widetilde{\boldsymbol{\upsilon}}^{(0)},\boldsymbol{\upsilon}^*)\geq -\mathrm{K}_0(\mathtt{x}),$$

which intuitively means that it is close enough to the target $\boldsymbol{\upsilon}^*\in\mathbb{R}^{p^*}$. This fact allows us to obtain an a priori bound for the deviation of the sequence $(\widetilde{\boldsymbol{\upsilon}}^{(k,k(+1))})_{k\in\mathbb{N}}\subset\varUpsilon$ from $\boldsymbol{\upsilon}^*\in\varUpsilon$ with Theorem 3.3.2.

$\{\mathcal{D}(\widetilde{\boldsymbol{\upsilon}}^{(k,k(+1))}-\boldsymbol{\upsilon}^*)\leq\mathbf{R}_0(\mathtt{x})\}$**:** As just mentioned this event is of high probability due to $\mathcal{L}(\widetilde{\boldsymbol{\upsilon}}^{(0)},\boldsymbol{\upsilon}^*)\geq -\mathrm{K}_0(\mathtt{x})$ and Theorem 3.3.2. This allows to concentrate the analysis on the set $\varUpsilon_\circ(\mathrm{R}_0)$ on which Taylor expansions of the functional $\mathcal{L}:\mathbb{R}^{p^*}\to\mathbb{R}$ become accurate.

$C(\nabla)$**:** This set ensures that on $\varOmega(\mathtt{x})\subset\varOmega$ all occurring random quadratic forms and stochastic errors are controlled by $\mathfrak{z}(\mathtt{x})\in\mathbb{R}$. Consequently we can derive in the proof of Lemma 5.A.3 an a priori bound of the form $\|\mathcal{D}(\widetilde{\boldsymbol{\upsilon}}^{(k,k(+1))}-\boldsymbol{\upsilon}^*)\|\leq\mathtt{r}_k$ for a decreasing sequence of radii $(\mathtt{r}_k)\subset\mathbb{R}_+$ satisfying $\limsup_{k\to\infty}\mathtt{r}_k=\mathtt{C}\mathfrak{z}(\mathtt{x})$. Further this set allows to obtain in Lemma 5.A.5 the bounds for all $k\in\mathbb{N}$.

On $\varOmega(\mathtt{x})\subset\varOmega$ we find for all $k\in\mathbb{N}$ that $\widetilde{\boldsymbol{\upsilon}}^{(k,k(+1))}\in\varUpsilon_\circ(\mathtt{r}_k)$ such that we can follow the arguments of Theorem 4.2.2 to obtain the desired result with accuracy measured by $\breve{\diamondsuit}_Q(\mathtt{r}_k,\mathtt{x})$.

Probability of desirable set

Here we show that the set $\varOmega(\mathtt{x})$ actually is of probability greater than $1-8\mathrm{e}^{-\mathtt{x}}-\beta_{(\mathbf{A})}$. We start with the following Lemma:

Lemma 5.A.1. *The set $C(\nabla)$ satisfies*

$$\mathbb{P}(C(\nabla))\geq 1-4\mathrm{e}^{-\mathtt{x}}.$$

Proof. Denote

$$\mathcal{A}\overset{\text{def}}{=}\bigcap_{\mathtt{r}\leq\mathrm{R}_0(\mathtt{x})}\left\{\sup_{\boldsymbol{\upsilon}\in\varUpsilon_\circ(\mathtt{r})}\left\{\frac{1}{6\omega\nu_1}\|\mathcal{Y}(\boldsymbol{\upsilon})\|-2\mathtt{r}^2\right\}\leq\mathfrak{z}_Q(\mathtt{x},4p^*)^2\right\}$$

$$\mathcal{B}\overset{\text{def}}{=}\bigcap_{\mathtt{r}\leq\mathrm{R}_0(\mathtt{x})}\left\{\sup_{\boldsymbol{\upsilon}\in\varUpsilon_\circ(\mathtt{r})}\left\{\frac{1}{6\breve{\omega}\breve{\nu}_1}\|\breve{\mathcal{Y}}(\boldsymbol{\upsilon})\|-2\mathtt{r}^2\right\}\leq\mathfrak{z}_Q(\mathtt{x},2p^*+2p)^2\right\}$$

$$\mathcal{C}\overset{\text{def}}{=}\Big\{\max\{\|\mathcal{D}^{-1}\nabla\mathcal{L}\|,\|D^{-1}\nabla_{\boldsymbol{\theta}}\mathcal{L}\|,\|\mathrm{H}^{-1}\nabla_{\boldsymbol{\eta}}\mathcal{L}\|\}\leq\mathfrak{z}(\mathtt{x})\Big\}.$$

We estimate

$$\begin{aligned}\mathbb{P}(C(\nabla))\geq{}&1-\mathbb{P}\left(\mathcal{A}^c\right)-\mathbb{P}\left(\mathcal{B}^c\right)-\mathbb{P}\left(\mathcal{C}^c\right)\\&-\mathbb{P}\left(\widetilde{\boldsymbol{\upsilon}},\widetilde{\boldsymbol{\upsilon}}_{\boldsymbol{\theta}^*}\notin\varUpsilon_\circ(\mathtt{r}_0)\right)-\mathbb{P}\left(\|\breve{D}^{-1}\breve{\nabla}_{\boldsymbol{\theta}}\|^2>\mathfrak{z}(\mathtt{x},\breve{\mathbb{B}}_{\boldsymbol{\theta}})\right),\end{aligned}$$

where $I\!\!B_{\boldsymbol{\theta}} \stackrel{\text{def}}{=} D^{-1}V^2D^{-1}$. We bound using for both terms Theorem 3.5.7 which is applicable due to $(\mathcal{E}\mathcal{D}_1)$ and $(\breve{\mathcal{E}}\mathcal{D}_1)$:

$$I\!\!P(\mathcal{A}^c) \le \mathrm{e}^{-\mathtt{x}}, \quad I\!\!P\left(\mathcal{B}^c\right) \le \mathrm{e}^{-\mathtt{x}}.$$

For the set $\mathcal{C} \subset \Omega$ observe that we can use $(\boldsymbol{\mathcal{I}})$ and Lemma 4.A.5 to find

$$\|\mathrm{H}^{-1}\nabla_{\boldsymbol{\eta}}\| \vee \|D^{-1}\nabla_{\boldsymbol{\theta}}\| \le \|\mathcal{D}^{-1}\nabla\|.$$

This implies that

$$\begin{aligned}&\{\|\mathcal{D}^{-1}\nabla\| \le \mathfrak{z}(\mathtt{x}, I\!\!B)\}\\ &\qquad \subseteq \{\|D^{-1}\nabla_{\boldsymbol{\theta}}\| \vee \|\mathrm{H}^{-1}\nabla_{\boldsymbol{\eta}}\| \le \mathfrak{z}(\mathtt{x}, I\!\!B)\}.\end{aligned}$$

Using the deviation properties of quadratic forms of Proposition 3.4.1 we find

$$I\!\!P\left(\|\mathcal{D}^{-1}\nabla\| > \mathfrak{z}(\mathtt{x}, I\!\!B)\right) \le 2\mathrm{e}^{-\mathtt{x}}, \quad I\!\!P\Big(\|\breve{D}^{-1}\breve{\nabla}\| > \mathfrak{z}(\mathtt{x}, \breve{I\!\!B})\Big) \le 2\mathrm{e}^{-\mathtt{x}}.$$

By the choice of $\mathfrak{z}(\mathtt{x}) > 0$ and $\mathtt{r}_0(\mathtt{x}) > 0$ this gives the claim. □

The next step is to show that the set $\bigcap_{k=1}^{K}(C^{(k,k)} \cap C^{(k,k+1)})$ is of high probability, which is independent of the number of necessary steps. Note that with $(\mathcal{I})$

$$\|D(\widetilde{\boldsymbol{\theta}}^{(k)} - \boldsymbol{\theta}^*)\| \vee \|\mathrm{H}(\widetilde{\boldsymbol{\eta}}^{(k(+1))} - \boldsymbol{\eta}^*)\| \le \frac{1}{1-\nu}\|\mathcal{D}(\widetilde{\boldsymbol{\upsilon}}^{(k,k(+1))} - \boldsymbol{\upsilon}^*)\|.$$

With assumption (B_1) and

$$\mathrm{R}_0(\mathtt{x}) = \frac{6\nu_0}{\mathtt{b}(1-\nu)}\sqrt{\mathtt{x} + \mathbb{Q} + \frac{\mathtt{b}}{9\nu_0^2}\mathrm{K}_0(\mathtt{x})},$$

this implies the desired result as $\mathcal{L}(\boldsymbol{\upsilon}^{(k,k(+1))}, \boldsymbol{\upsilon}^*) \ge \mathcal{L}(\widetilde{\boldsymbol{\upsilon}}^{(0)}, \boldsymbol{\upsilon}^*)$ such that with Theorem 3.3.2

$$\begin{aligned}&I\!\!P\left(\bigcap_{k=0}^{K}(C^{(k,k)} \cap C^{(k,k+1)})\right)\\ &\quad \ge I\!\!P\left(\bigcap_{k=0}^{K}(C^{(k,k)} \cap C^{(k,k+1)}) \cap \Big\{\mathcal{L}(\widetilde{\boldsymbol{\upsilon}}^{(0)}, \boldsymbol{\upsilon}^*) \ge -\mathrm{K}_0\Big\}\right)\\ &\qquad -I\!\!P(\mathcal{L}(\widetilde{\boldsymbol{\upsilon}}^{(0)}, \boldsymbol{\upsilon}^*) \le -\mathrm{K}_0)\\ &\quad \ge I\!\!P\Big\{\varUpsilon(\mathrm{K}_0(\mathtt{x})) \subset \varUpsilon_{\circ}\Big((1-\nu)\mathrm{R}_0(\mathtt{x})\Big)\Big\} - \beta_{\boldsymbol{(A)}}\\ &\quad \ge 1 - \mathrm{e}^{-\mathtt{x}} - \beta_{\boldsymbol{(A)}}.\end{aligned}$$

Remark 5.A.2. This also shows that the sets of maximizers $(\widetilde{\boldsymbol{\upsilon}}^{(k,k(+1))})$ are nonempty and well defined since the maximization always takes place on compact sets of the form $\{\boldsymbol{\theta} \in \mathbb{R}^p,\, (\boldsymbol{\theta}, \boldsymbol{\eta}) \in \varUpsilon_\circ(\mathrm{R}_0)\}$ or $\{\boldsymbol{\eta} \in \mathbb{R}^m,\, (\boldsymbol{\theta}, \boldsymbol{\eta}) \in \varUpsilon_\circ(\mathrm{R}_0)\}$.

To address the claim of remark 5.2.4 we present the following Lemma:

Lemma 5.A.2. *On the set* $C(\nabla) \cap \{\widetilde{\boldsymbol{\upsilon}}_0 \in \varUpsilon_\circ(R_K)\}$ *it holds*

$$\mathcal{L}(\boldsymbol{\upsilon}_0, \boldsymbol{\upsilon}^*) \geq -(1/2 + 12\nu_0\omega)R_K^2 - (\delta(R_K) + \mathfrak{z}(\mathtt{x}))R_K - 6\nu_0\omega\mathfrak{z}(\mathtt{x})^2.$$

Proof. With similar arguments as in the proof of Lemma 5.A.3 we have on $C(\nabla) \subset \varOmega$ that

$$\begin{aligned}
\mathcal{L}(\boldsymbol{\upsilon}_0, \boldsymbol{\upsilon}^*) &\geq \mathbb{E}[\mathcal{L}(\boldsymbol{\upsilon}_0, \boldsymbol{\upsilon}^*)] - \|\mathcal{D}^{-1}\nabla\boldsymbol{\zeta}(\boldsymbol{\upsilon}^*)\|R_K \\
&\quad -|\{\nabla\boldsymbol{\zeta}(\widehat{\boldsymbol{\upsilon}}) - \nabla\boldsymbol{\zeta}(\boldsymbol{\upsilon}^*)\}(\boldsymbol{\upsilon}_0 - \boldsymbol{\upsilon}^*)| \\
&\geq -\|\mathcal{D}(\boldsymbol{\upsilon}_0 - \boldsymbol{\upsilon}^*)\|^2/2 - \|\mathcal{D}^{-1}\nabla\boldsymbol{\zeta}(\boldsymbol{\upsilon}^*)\|R_K \\
&\quad -\|\mathcal{D}^{-1}\big\{\nabla\mathcal{L}(\widehat{\boldsymbol{\upsilon}}) - \nabla\mathcal{L}(\boldsymbol{\upsilon}^*)\big\}\|R_K - R_K\delta(R_K) \\
&\geq -(1/2 + 12\nu_0\omega)R_K^2 - (\delta(R_K) + \mathfrak{z}(\mathtt{x}))R_K - 6\nu_0\omega\mathfrak{z}(\mathtt{x})^2.
\end{aligned}$$

□

Proof of convergence

We derive the a priori bound $\widetilde{\boldsymbol{\upsilon}}^{(k,k(+1))} \in \varUpsilon_\circ(\mathtt{r}_k)$ with an adequately decreasing sequence $(\mathtt{r}_k) \subset \mathbb{R}_+$ using the argument of Section 5.A.2, where $\limsup \mathtt{r}_k \approx \mathfrak{z}(\mathtt{x})$.

Lemma 5.A.3. *Assume that*

$$\varOmega(\mathtt{x}) \subseteq \bigcap_{k\in\mathbb{N}} \left\{\boldsymbol{\upsilon}^{(k,k(+1))} \in \varUpsilon_\circ\left(\mathtt{r}_k^{(l)}\right)\right\}.$$

Then under the assumptions of Theorem 5.2.1 we get on $\varOmega(\mathtt{x})$ *for all* $k \in \mathbb{N}_0$

$$\begin{aligned}
\left\|\mathcal{D}(\widetilde{\boldsymbol{\upsilon}}^{(k,k(+1))} - \boldsymbol{\upsilon}^*)\right\| &\leq 2\sqrt{2}(1-\sqrt{\nu})^{-1}\left(\mathfrak{z}(\mathtt{x}) + (1+\sqrt{\nu})\nu^k R_0(\mathtt{x})\right) \\
&\qquad +2\sqrt{2}(1+\sqrt{\nu})\sum_{r=0}^{k-1}\nu^r\diamondsuit_Q\left(\mathtt{r}_{k-r}^{(l)}\right) \\
&=: \mathtt{r}_k^{(l+1)}.
\end{aligned}$$

Proof. 1. We first show that on $\varOmega(\mathtt{x})$

$$\begin{aligned} D(\widetilde{\boldsymbol{\theta}}^{(k)}-\boldsymbol{\theta}^{*}) &= D^{-1}\nabla_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\upsilon}^{*})-D^{-1}A(\widetilde{\boldsymbol{\eta}}^{(k)}-\boldsymbol{\eta}^{*})+\boldsymbol{\tau}\big(\mathtt{r}_k^{(l)}\big), \qquad (5.A.4)\\ \mathrm{H}(\widetilde{\boldsymbol{\eta}}^{(k)}-\boldsymbol{\eta}^{*}) &= \mathrm{H}^{-1}\nabla_{\boldsymbol{\eta}}\mathcal{L}(\boldsymbol{\upsilon}^{*})-\mathrm{H}^{-1}A^{\top}(\widetilde{\boldsymbol{\theta}}^{(k-1)}-\boldsymbol{\theta}^{*})+\boldsymbol{\tau}\big(\mathtt{r}_k^{(l)}\big), \end{aligned}$$

where

$$\|\boldsymbol{\tau}(\mathtt{r})\| \le \diamondsuit_Q(\mathtt{r},\mathtt{x})=\big\{\delta(\mathtt{r})\mathtt{r}+6\nu_1\omega(\mathfrak{z}_Q(\mathtt{x},4p^{*})+2\mathtt{r}^2)\big\}\,.$$

The proof is the same in each step for both statements such that we only prove the first one. The arguments presented here are similar to those Lemma 4.A.3. By assumption on $\varOmega(\mathtt{x})$ we have $\widetilde{\boldsymbol{\upsilon}}^{(k,k(+1))}\in\varUpsilon_{\circ}\big(\mathtt{r}_k^{(l)}\big)$. Define with $\zeta=\mathcal{L}-\mathbb{E}\mathcal{L}$

$$\alpha(\boldsymbol{\upsilon},\boldsymbol{\upsilon}^{*}):=\mathcal{L}(\boldsymbol{\upsilon},\boldsymbol{\upsilon}^{*})-\big(\nabla\zeta(\boldsymbol{\upsilon}^{*})(\boldsymbol{\upsilon}-\boldsymbol{\upsilon}^{*})-\|\mathcal{D}(\boldsymbol{\upsilon}-\boldsymbol{\upsilon}^{*})\|^2/2\big)\,.$$

Note that

$$\begin{aligned} \mathcal{L}(\boldsymbol{\upsilon},\boldsymbol{\upsilon}^{*}) &= \nabla\zeta(\boldsymbol{\upsilon}^{*})(\boldsymbol{\upsilon}-\boldsymbol{\upsilon}^{*})-\|\mathcal{D}(\boldsymbol{\upsilon}-\boldsymbol{\upsilon}^{*})\|^2/2+\alpha(\boldsymbol{\upsilon},\boldsymbol{\upsilon}^{*})\\ &= \nabla_{\boldsymbol{\theta}}\zeta(\boldsymbol{\upsilon}^{*})(\boldsymbol{\theta}-\boldsymbol{\theta}^{*})-\|D(\boldsymbol{\theta}-\boldsymbol{\theta}^{*})\|^2/2+(\boldsymbol{\theta}-\boldsymbol{\theta}^{*})^{\top}A(\boldsymbol{\eta}-\boldsymbol{\eta}^{*})\\ &\quad +\nabla_{\boldsymbol{\eta}}\zeta(\boldsymbol{\upsilon}^{*})(\boldsymbol{\eta}-\boldsymbol{\eta}^{*})-\|\mathrm{H}(\boldsymbol{\eta}-\boldsymbol{\eta}^{*})\|^2/2+\alpha(\boldsymbol{\upsilon},\boldsymbol{\upsilon}^{*}). \end{aligned}$$

Setting $\nabla_{\boldsymbol{\theta}}\mathcal{L}(\widetilde{\boldsymbol{\theta}}^{(k)},\widetilde{\boldsymbol{\eta}}^{(k)})=0$ we find

$$D(\widetilde{\boldsymbol{\theta}}^{(k)}-\boldsymbol{\theta}^{*})-\mathcal{D}^{-1}\big(\nabla_{\boldsymbol{\theta}}\zeta(\boldsymbol{\upsilon}^{*})-A(\widetilde{\boldsymbol{\eta}}^{(k)}-\boldsymbol{\eta}^{*})\big)=\mathcal{D}^{-1}\nabla_{\boldsymbol{\theta}}\alpha(\widetilde{\boldsymbol{\upsilon}}^{(k,k)},\boldsymbol{\upsilon}^{*}).$$

As we assume that $\widetilde{\boldsymbol{\upsilon}}^{(k,k)}\in\varUpsilon_{\circ}(\mathrm{R}_0)$ it suffices to show that with dominating probability

$$\sup_{(\boldsymbol{\theta},\widetilde{\boldsymbol{\eta}}^{(k)})\in\varUpsilon_{\circ}(\mathrm{R}_0)}\|\mathcal{U}_{\boldsymbol{\theta}}(\boldsymbol{\theta},\widetilde{\boldsymbol{\eta}}^{(k)})\|\le\diamondsuit(\mathtt{r}_k^{(l)}),$$

where

$$\begin{aligned} &\mathcal{U}_{\boldsymbol{\theta}}(\boldsymbol{\theta},\widetilde{\boldsymbol{\eta}}^{(k)})\\ &\quad\overset{\mathrm{def}}{=} D^{-1}\big\{\nabla_{\boldsymbol{\theta}}\mathcal{L}(\widetilde{\boldsymbol{\upsilon}}^{(k,k)})-\nabla_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\upsilon}^{*})-D^2\,(\boldsymbol{\theta}-\boldsymbol{\theta}^{*})-A(\widetilde{\boldsymbol{\eta}}^{(k)}-\boldsymbol{\eta}^{*})\big\}. \end{aligned}$$

To see this note first that with Lemma 4.A.5 $\|D^{-1}\varPi_{\boldsymbol{\theta}}\mathcal{D}\boldsymbol{\upsilon}\|\le\|\mathcal{D}^{-1}\mathcal{D}\boldsymbol{\upsilon}\|$.

This gives by condition $(\mathcal{L}_{\mathbf{0}})$, Lemma 4.A.5 and Taylor expansion

$$\begin{aligned}
&\sup_{(\boldsymbol{\theta},\widetilde{\boldsymbol{\eta}}^{(k)})\in\Upsilon_\circ(\mathtt{r})} \|\mathbb{E}\mathcal{U}(\boldsymbol{\theta},\widetilde{\boldsymbol{\eta}}^{(k)})\| \\
&\quad\le \sup_{\boldsymbol{\upsilon}\in\Upsilon_\circ(\mathtt{r})} \|D^{-1}\Pi_{\boldsymbol{\theta}}\Big(\nabla\mathbb{E}\mathcal{L}(\boldsymbol{\upsilon})-\nabla\mathbb{E}\mathcal{L}(\boldsymbol{\upsilon}^*)-\mathcal{D}\left(\boldsymbol{\upsilon}-\boldsymbol{\upsilon}^*\right)\Big)\| \\
&\quad\le \sup_{\boldsymbol{\upsilon}\in\Upsilon_\circ(\mathtt{r})} \|D^{-1}\Pi_{\boldsymbol{\theta}}\mathcal{D}\|\|\mathcal{D}^{-1}\nabla^2\mathbb{E}\mathcal{L}(\boldsymbol{\upsilon})^2\mathcal{D}^{-1}-I_{p^*}\|^{1/2}\mathtt{r} \\
&\quad\le \delta(\mathtt{r})\mathtt{r}.
\end{aligned}$$

For the remainder note that again with Lemma 4.A.5

$$\|D^{-1}\Big(\nabla_{\boldsymbol{\theta}}\zeta(\boldsymbol{\upsilon})-\nabla_{\boldsymbol{\theta}}\zeta(\boldsymbol{\upsilon}^*)\Big)\Big\| \le \|\mathcal{D}^{-1}\Big(\nabla\zeta(\boldsymbol{\upsilon})-\nabla\zeta(\boldsymbol{\upsilon}^*)\Big)\Big\|.$$

This yields that on $\Omega(\mathtt{x})$

$$\begin{aligned}
&\sup_{(\boldsymbol{\theta},\widetilde{\boldsymbol{\eta}}^{(k)})\in\Upsilon_\circ(\mathtt{r})} \Big\|\mathcal{U}_{\boldsymbol{\theta}}(\boldsymbol{\theta},\widetilde{\boldsymbol{\eta}}^{(k)})-\mathbb{E}\mathcal{U}_{\boldsymbol{\theta}}(\boldsymbol{\theta},\widetilde{\boldsymbol{\eta}}^{(k)})\Big\| \\
&\quad\le \sup_{\boldsymbol{\upsilon}\in\Upsilon_\circ(\mathtt{r})} \Big\|D^{-1}\Big(\nabla_{\boldsymbol{\theta}}\zeta(\boldsymbol{\upsilon})-\nabla_{\boldsymbol{\theta}}\zeta(\boldsymbol{\upsilon}^*)\Big)\Big\| \\
&\quad\le 6\nu_1\omega \sup_{\boldsymbol{\upsilon}\in\Upsilon_\circ(\mathtt{r})} \left\{\frac{1}{6\nu_1\omega}\|\mathcal{Y}(\boldsymbol{\upsilon})\|\right\} \le 6\nu_1\omega\big\{\mathfrak{z}_Q(\mathtt{x},4p^*)+2\mathtt{r}^2\big\}.
\end{aligned}$$

Using the same argument for $\widetilde{\boldsymbol{\eta}}^{(k)}$ gives the claim.

2. We prove the a priori bound for the distance of the kth estimator to the oracle

$$\big\|\mathcal{D}(\widetilde{\boldsymbol{\upsilon}}^{(k,k(+1))}-\boldsymbol{\upsilon}^*)\big\| \le \mathtt{r}_k^{(l+1)}.$$

To see this we first use the inequality

$$\|\mathcal{D}(\widetilde{\boldsymbol{\upsilon}}^{(k,k(+1))}-\boldsymbol{\upsilon}^*)\| \le \sqrt{2}\|D(\widetilde{\boldsymbol{\theta}}^{(k)}-\boldsymbol{\theta}^*)\|+\sqrt{2}\|\mathrm{H}(\widetilde{\boldsymbol{\eta}}^{(k(+1))}-\boldsymbol{\eta}^*)\|.$$

We find with (5.A.4)

$$\begin{aligned}
\|D(\widetilde{\boldsymbol{\theta}}^{(k)}-\boldsymbol{\theta}^*)\| &\le \|D^{-1}\nabla_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\upsilon}^*)\|+\|D^{-1}A(\widetilde{\boldsymbol{\eta}}^{(k)}-\boldsymbol{\eta}^*)\|+\|\boldsymbol{\tau}\big(\mathtt{r}_k^{(l)}\big)\| \\
&\le \|D^{-1}\nabla_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\upsilon}^*)\|+\|D^{-1}A\mathrm{H}^{-1}\|\|\mathrm{H}(\widetilde{\boldsymbol{\eta}}^{(k)}-\boldsymbol{\eta}^*)\|+\|\boldsymbol{\tau}\big(\mathtt{r}_k^{(l)}\big)\|.
\end{aligned}$$

Next we use that on $\Omega(\mathtt{x})$

$$\|D^{-1}A\mathrm{H}^{-1}\|\le\sqrt{\nu},\ \|D^{-1}\nabla_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\upsilon}^*)\|\le\mathfrak{z}(\mathtt{x}),\ \|\mathrm{H}^{-1}\nabla_{\boldsymbol{\eta}}\mathcal{L}(\boldsymbol{\upsilon}^*)\|\le\mathfrak{z}(\mathtt{x}),$$

and

$$\|\mathrm{H}(\widetilde{\boldsymbol{\eta}}^{(k)} - \boldsymbol{\eta}^*)\| \le \|\mathrm{H}^{-1}\nabla_{\boldsymbol{\eta}}\mathcal{L}(\boldsymbol{\upsilon}^*)\| + \|\mathrm{H}^{-1}A^\top(\widetilde{\boldsymbol{\theta}}^{(k-1)} - \boldsymbol{\theta}^*)\| + \|\boldsymbol{\tau}(\mathtt{r}_k^{(l)})\|,$$

to derive the recursive formula

$$\|D(\widetilde{\boldsymbol{\theta}}^{(k)} - \boldsymbol{\theta}^*)\| \le (1+\sqrt{\nu})\left(\mathfrak{z}(\mathtt{x}) + \|\boldsymbol{\tau}(\mathtt{r}_k^{(l)})\|\right) + \nu\|D(\widetilde{\boldsymbol{\theta}}^{(k-1)} - \boldsymbol{\theta}^*)\|.$$

Deriving the analogous formula for $\|\mathrm{H}(\widetilde{\boldsymbol{\eta}}^{(k)} - \boldsymbol{\eta}^*)\|$ and solving the recursions gives the claim.

□

Lemma 5.A.4. *Assume the same as in Theorem 5.2.1. Then we get*

$$\mathit{\Omega}(\mathtt{x}) \subseteq \bigcap_{k\in\mathbb{N}} \left\{\boldsymbol{\upsilon}^{(k,k(+1))} \in \varUpsilon_\circ\left(\mathtt{r}_k^{(1)}\right)\right\},$$

where

$$\mathtt{r}_k^{(1)} \le 2\sqrt{2}(1-\sqrt{\nu})^{-1}\left\{(\mathfrak{z}(\mathtt{x}) + \diamondsuit_Q(R_0,\mathtt{x})) + (1+\sqrt{\nu})\nu^k R_0(\mathtt{x})\right\} \tag{5.A.5}$$

Furthermore assume that $\delta(\mathtt{r})/\mathtt{r} \vee 12\nu_1\omega \le \epsilon$ *and that* (5.2.4) *and* (5.2.5) *are met with* $\mathtt{C}(\nu)$ *defined in* (5.2.6)*. Then*

$$\mathit{\Omega}(\mathtt{x}) \subseteq \bigcap_{k\in\mathbb{N}} \left\{\boldsymbol{\upsilon}^{(k,k(+1))} \in \varUpsilon_\circ(\mathtt{r}_k^*)\right\},$$

where

$$\begin{aligned}\mathtt{r}_k^* \le{}& \left(\mathtt{C}(\nu) + \frac{4\mathtt{C}(\nu)^3 c(\epsilon,\mathfrak{z}(\mathtt{x}))}{1-c(\epsilon,\mathfrak{z}(\mathtt{x}))}\right)\left(\mathfrak{z}(\mathtt{x}) + \epsilon\mathfrak{z}(\mathtt{x})^2\right)\\ &+\nu^k\left(\mathtt{C}(\nu) + \nu\frac{4\mathtt{C}(\nu)^3 c(\epsilon,R_0)}{1-c(\epsilon,R_0)}\right)R_0.\end{aligned} \tag{5.A.6}$$

Proof. We proof this claim via induction. On $\mathit{\Omega}(\mathtt{x})$ we have

$$\boldsymbol{\upsilon}^{(k,k(+1))} \in \varUpsilon_\circ(\mathrm{R}_0), \quad \text{set } \mathtt{r}_k^{(0)} \stackrel{\text{def}}{=} \mathrm{R}_0.$$

With Lemma 5.A.3 we find that if

$$\mathit{\Omega}(\mathtt{x}) \subseteq \bigcap_{k\in\mathbb{N}} \left\{\boldsymbol{\upsilon}^{(k,k(+1))} \in \varUpsilon_\circ(\mathtt{r}_k^{(l)})\right\},$$

that then

$$\mathit{\Omega}(\mathtt{x}) \subseteq \bigcap_{k\in\mathbb{N}} \left\{\boldsymbol{\upsilon}^{(k,k(+1))} \in \varUpsilon_\circ(\mathtt{r}_k^{(l+1)})\right\},$$

where

$$\begin{aligned}\mathtt{r}_k^{(l)} &\le 2\sqrt{2}(1-\sqrt{\nu})^{-1}\left(\mathfrak{z}(\mathtt{x}) + (1+\sqrt{\nu})\nu^k \mathrm{R}_0(\mathtt{x})\right) \\ &\quad +2\sqrt{2}(1+\sqrt{\nu})\sum_{r=0}^{k-1} \nu^r \diamondsuit_Q\left(\mathtt{r}_{k-r}^{(l-1)}, \mathtt{x}\right).\end{aligned}$$

Setting $l=1$ this gives

$$\mathtt{r}_k^{(1)} \le 2\sqrt{2}(1-\sqrt{\nu})^{-1}\left\{(\mathfrak{z}(\mathtt{x}) + \diamondsuit_Q(\mathrm{R}_0, \mathtt{x})) + (1+\sqrt{\nu})\nu^k \mathrm{R}_0(\mathtt{x})\right\},$$

which already yields (5.A.5). For the second claim we show that

$$\mathit{\Omega}(\mathtt{x}) \subseteq \bigcap_{k\in\mathbb{N}} \left\{\boldsymbol{\upsilon}^{(k,k(+1))} \in \mathit{\Upsilon}_\circ\left(\limsup_{l\to\infty} \mathtt{r}_k^{(l)}\right)\right\} \subseteq \bigcap_{k\in\mathbb{N}} \left\{\boldsymbol{\upsilon}^{(k,k(+1))} \in \mathit{\Upsilon}_\circ(\mathtt{r}_k^*)\right\}.$$

We have to show that $\limsup_{l\to\infty} \mathtt{r}_k^{(l)} \le \mathtt{r}_k^*$ in (5.A.6). For this we use $\delta(\mathtt{r})/\mathtt{r} \vee 12\nu_1\omega \le \epsilon$ to estimate further

$$\begin{aligned}\mathtt{r}_k^{(l)} &\le 2\sqrt{2}(1-\sqrt{\nu})^{-1}\left(\mathfrak{z}(\mathtt{x}) + (1+\sqrt{\nu})\nu^k \mathrm{R}_0(\mathtt{x})\right) \\ &\quad +2\sqrt{2}(1+\sqrt{\nu})\epsilon\sum_{r=0}^{k-1} \nu^r \left(\left(\mathtt{r}_{k-r}^{(l-1)}\right)^2 + \mathfrak{z}(\mathtt{x})^2\right) \\ &\le 2\sqrt{2}(1-\sqrt{\nu})^{-1}\left(\mathfrak{z}(\mathtt{x}) + \epsilon\mathfrak{z}(\mathtt{x})^2 + (1+\sqrt{\nu})\nu^k \mathrm{R}_0(\mathtt{x})\right) \\ &\quad +2\sqrt{2}(1+\sqrt{\nu})\epsilon\sum_{r=0}^{k-1} \nu^r \left(\mathtt{r}_{k-r}^{(l-1)}\right)^2 \\ &\le \mathtt{C}(\nu)\left\{\left(\mathfrak{z}(\mathtt{x}) + \epsilon\mathfrak{z}(\mathtt{x})^2\right) + \nu^k \mathrm{R}_0 + \epsilon\sum_{r=0}^{k-1} \nu^r \left(\mathtt{r}_{k-r}^{(l-1)}\right)^2\right\},\end{aligned}$$

where $\mathtt{C}(\nu) > 0$ is defined in (5.2.6). We set

$$A_{s,k}^{(l)} \overset{\mathrm{def}}{=} \sum_{r_1=0}^{k-1} \nu^{r_1}\left(\sum_{r_2=0}^{k-r_1-1} \nu^{r_2}\left(\ldots \sum_{r_s=0}^{k-r_1-\ldots-r_{s-1}-1} \nu^{r_s}\left(\mathtt{r}_{k-r_1-\ldots-r_s}^{(l-1)}\right)^2 \ldots\right)^2\right)^2.$$

We will now show that

$$
\begin{aligned}
A^{(l)}_{s,k} \le 4^{\sum_{t=0}^{s-1} 2^t} \mathtt{C}(\nu)^{2^s} \Bigg\{ & \left(\frac{1}{1-\nu}\right)^{\sum_{t=0}^{s-1} 2^t} \left(\mathfrak{z}(\mathtt{x}) + \epsilon\mathfrak{z}(\mathtt{x})^2\right)^{2^s} \\
& + \nu^k \left(\frac{1}{\nu^{-1}-1}\right)^{\sum_{t=0}^{s-1} 2^t} \mathrm{R}_0^{2^s} \Bigg\} \\
& + 4^{\sum_{t=0}^{s-1} 2^t} (\mathtt{C}(\nu)\epsilon)^{2^s} A^{(l-1)}_{s+1,k}.
\end{aligned}
\tag{5.A.7}
$$

We proof this claim via induction. Clearly

$$
\begin{aligned}
A^{(l)}_{1,k} &= \sum_{r_1=0}^{k-1} \nu^{r_1} \left(\mathtt{r}^{(l-1)}_{k-r_1}\right)^2 \le 4\mathtt{C}(\nu)^2 \sum_{r_1=0}^{k-1} \nu^{r_1} \left\{ \left(\mathfrak{z}(\mathtt{x}) + \epsilon\mathfrak{z}(\mathtt{x})^2\right)^2 + \nu^{2(k-r_1)} \mathrm{R}_0^2 \right\} \\
&\quad + 4\mathtt{C}(\nu)^2 \epsilon^2 \sum_{r_1=0}^{k-1} \nu^{r_1} \left(\sum_{r_2=0}^{k-r_1-r_2-1} \nu^{r_2} \left(\mathtt{r}^{(l-2)}_{k-r_1-r_2}\right)^2 \right)^2 \\
&\le 4\mathtt{C}(\nu)^2 \left\{ \frac{1}{1-\nu} \left(\mathfrak{z}(\mathtt{x}) + \epsilon\mathfrak{z}(\mathtt{x})^2\right)^2 + \frac{\nu^k}{\nu^{-1}-1} \mathrm{R}_0^2 \right\} \\
&\quad + 4\mathtt{C}(\nu)^2 \epsilon^2 A^{(l-1)}_{2,k}.
\end{aligned}
$$

Furthermore

$$
\begin{aligned}
A^{(l)}_{s,k} &\overset{\text{def}}{=} \sum_{r_1=0}^{k-1} \nu^{r_1} \left(\sum_{r_2=0}^{k-r_1-1} \nu^{r_2} \left(\cdots \sum_{r_s=0}^{k-r_1-\ldots-r_{s-1}-1} \nu^{r_s} \left(\mathtt{r}^{(l-1)}_{k-r_1-\ldots-r_s}\right)^2 \cdots \right)^2 \right)^2 \\
&= \sum_{r_1=0}^{k-1} \nu^{r_1} \left(A^{(l)}_{s-1,k-r_1} \right)^2 .
\end{aligned}
\tag{5.A.8}
$$

Plugging in the induction assumption (5.A.7) we get for $s \ge 2$

$$
\begin{aligned}
A^{(l)}_{s,k} \le \sum_{r_1=0}^{k-1} \nu^{r_1} \Bigg(4^{\sum_{t=0}^{s-2} 2^t} \mathtt{C}(\nu)^{2^{s-1}} \Bigg\{ & \left(\frac{1}{1-\nu}\right)^{\sum_{t=0}^{s-2} 2^t} \left(\mathfrak{z}(\mathtt{x}) + \epsilon\mathfrak{z}(\mathtt{x})^2\right)^{2^{s-1}} \\
& + \nu^k \left(\frac{1}{\nu^{-1}-1}\right)^{\sum_{t=0}^{s-2} 2^t} \mathrm{R}_0^{2^{s-1}} \Bigg\} \\
& + 4^{\sum_{t=0}^{s-2} 2^t} (\mathtt{C}(\nu)\epsilon)^{2^{s-1}} A^{(l-1)}_{s,k-r_1} \Bigg)^2 .
\end{aligned}
$$

Shifting the index this gives

$$A_{s,k}^{(l)} \le 4 \sum_{r_1=0}^{k-1} \nu^{r_1} \left(4^{\sum_{t=1}^{s-1} 2^t} \mathtt{C}(\nu)^{2^s} \left\{ \left(\frac{1}{1-\nu} \right)^{\sum_{t=1}^{s-1} 2^{t-1}} \left(\mathfrak{z}(\mathtt{x}) + \epsilon \mathfrak{z}(\mathtt{x})^2 \right)^{2^s} \right. \right.$$

$$\left. + \nu^k \left(\frac{1}{\nu^{-1}-1} \right)^{\sum_{t=1}^{s-1} 2^t} \mathrm{R}_0^{2^s} \right\}$$

$$\left. + \, 4^{\sum_{t=1}^{s-1} 2^t} (\mathtt{C}(\nu)\epsilon)^{2^s} (A_{s,k-r_1}^{(l-1)})^2 \right).$$

Direct calculation then leads to

$$A_{s,k}^{(l)} \le 4^{\sum_{t=0}^{s-1} 2^t} \mathtt{C}(\nu)^{2^s} \left\{ \left(\frac{1}{1-\nu} \right)^{\sum_{t=0}^{s-1} 2^t} \left(\mathfrak{z}(\mathtt{x}) + \epsilon \mathfrak{z}(\mathtt{x})^2 \right)^{2^s} \right.$$

$$\left. + \nu^k \left(\frac{1}{\nu^{-1}-1} \right)^{\sum_{t=0}^{s-1} 2^t} \mathrm{R}_0^{2^s} \right\}$$

$$+ 4^{\sum_{t=0}^{s-1} 2^t} (\mathtt{C}(\nu)\epsilon)^{2^s} \sum_{r_1=0}^{k-1} \nu^{r_1} (A_{s,k-r_1}^{(l-1)})^2,$$

from which we infer (5.A.7) with (5.A.8). Similarly we can prove

$$A_{s,k}^{(1)} = \left(\frac{1}{1-\nu} \right)^{2^s-1} \mathrm{R}_0^{2^s}.$$

Abbreviate

$$\lambda_s \overset{\text{def}}{=} 4^{2^s-1} \mathtt{C}(\nu)^{2^s}, \quad \beta_s \overset{\text{def}}{=} 4^{2^s-1} (\mathtt{C}(\nu)\epsilon)^{2^s},$$

$$\mathfrak{z}_s(\mathtt{x}) \overset{\text{def}}{=} \left(\frac{1}{1-\nu} \right)^{2^s-1} \left(\mathfrak{z}(\mathtt{x}) + \epsilon \mathfrak{z}(\mathtt{x})^2 \right)^{2^s}, \quad \mathrm{R}_s \overset{\text{def}}{=} \left(\frac{1}{\nu^{-1}-1} \right)^{2^s-1} \mathrm{R}_0^{2^s}.$$

Then (5.A.7) allows to bound

$$\begin{aligned} \mathtt{r}_k^{(l)} &\le \mathtt{C}(\nu) \left\{ \left(\mathfrak{z}(\mathtt{x}) + \epsilon \mathfrak{z}(\mathtt{x})^2 \right) + \nu^k \mathrm{R}_0 + \epsilon A_{1,k}^{(l)} \right\} \\ &\le \sum_{s=0}^{l-1} \lambda_s \prod_{r=0}^{s-1} \beta_r \mathfrak{z}_s(\mathtt{x}) + \nu^k \sum_{s=0}^{l-1} \lambda_s \prod_{r=0}^{s-1} \beta_r \mathrm{R}_s + \prod_{r=0}^{l-1} \beta_r \mathrm{R}_l. \end{aligned} \tag{5.A.9}$$

We estimate further

$$\begin{aligned}
\sum_{s=0}^{l-1} \lambda_s \prod_{r=0}^{s-1} \beta_r \mathfrak{z}_s(\mathtt{x}) - \mathtt{C}(\nu)\left(\mathfrak{z}(\mathtt{x}) + \epsilon\mathfrak{z}(\mathtt{x})^2\right) &= \sum_{s=1}^{l-1} \lambda_s \prod_{r=0}^{s-1} \beta_r \mathfrak{z}_s(\mathtt{x}) \\
&\le \sum_{s=1}^{l-1} 4^{2^s} \mathtt{C}(\nu)^{2^{s+1}} \epsilon^{2^s-1} \left(\frac{1}{1-\nu}\right)^{2^s-1} \left(\mathfrak{z}(\mathtt{x}) + \epsilon\mathfrak{z}(\mathtt{x})^2\right)^{2^s} \\
&= \epsilon 4^2 \mathtt{C}(\nu)^4 \left(\frac{1}{1-\nu}\right) \left(\mathfrak{z}(\mathtt{x}) + \epsilon\mathfrak{z}(\mathtt{x})^2\right)^2 \\
&\quad \sum_{s=1}^{l-1} \left(\epsilon 4\mathtt{C}(\nu)\frac{1}{1-\nu}\left(\mathfrak{z}(\mathtt{x}) + \epsilon\mathfrak{z}(\mathtt{x})^2\right)\right)^{2^s-1} .
\end{aligned}$$

Assuming (5.2.4) this gives

$$\sum_{s=0}^{l-1} \lambda_s \prod_{r=0}^{s-1} \beta_r \mathfrak{z}_s(\mathtt{x}) \le \left(\mathtt{C}(\nu) + \frac{4\mathtt{C}(\nu)^3 c(\epsilon, \mathfrak{z}(\mathtt{x}))}{1 - c(\epsilon, \mathfrak{z}(\mathtt{x}))}\right)\left(\mathfrak{z}(\mathtt{x}) + \epsilon\mathfrak{z}(\mathtt{x})^2\right).$$

With the same argument we find under (5.2.5) that

$$\nu^k \sum_{s=0}^{l-1} \lambda_s \prod_{r=0}^{s-1} \beta_r \mathrm{R}_s \le \nu^k \left(\mathtt{C}(\nu) + \nu\frac{4\mathtt{C}(\nu)^3 c(\epsilon, \mathrm{R}_0)}{1 - c(\epsilon, \mathrm{R}_0)}\right)\mathrm{R}_0.$$

Additionally (5.2.5) implies

$$\prod_{r=0}^{l-1} \beta_r \mathrm{R}_l \le \left(\epsilon 4\mathtt{C}(\nu)\frac{1}{\nu^{-1}-1}\right)^{2^{l-1}} \mathrm{R}_0^{2^l} \to 0.$$

Plugging these bounds into (5.A.9) and letting l tend to infinity gives the claim. □

Result after convergence

In the previous section we showed that

$$\begin{aligned}
\varOmega(\mathtt{x}) \subset & \bigcap_{\mathtt{r} \le \mathrm{R}_0(\mathtt{x})} \left\{ \sup_{\boldsymbol{\upsilon} \in \varUpsilon_\circ(\mathtt{r})} \left\{ \frac{1}{6\breve{\omega}\breve{\nu}_1} \|\breve{\mathfrak{Y}}(\boldsymbol{\upsilon})\| - 2\mathtt{r}^2 \right\} \le \mathfrak{z}_Q(\mathtt{x}, 2p^* + 2p)^2 \right\} \\
& \cap \bigcap_{k \in \mathbb{N}} \left\{ \boldsymbol{\upsilon}^{(k,k)} \in \varUpsilon_\circ\left(\mathtt{r}_k^{(\cdot)}\right), \, \boldsymbol{\upsilon}^{(k,k+1)} \in \varUpsilon_\circ\left(\mathtt{r}_k^{(\cdot)}\right) \right\} \\
& \cap \{\widetilde{\boldsymbol{\upsilon}}, \widetilde{\boldsymbol{\upsilon}}_{\boldsymbol{\theta}^*} \in \varUpsilon_\circ(\mathtt{r}_0)\},
\end{aligned}$$

where $\mathtt{r}_k^{(\cdot)}$ is defined in (5.A.6) or (5.A.5). The claim of Theorem 5.2.1 follows with the following lemma:

Lemma 5.A.5. *Assume* $(\breve{\mathcal{E}}\mathcal{D}_1)$, $(\breve{\mathcal{L}}_0)$, *and* $(\mathcal{I})$ *with a central point* $\boldsymbol{\upsilon}^\circ = \boldsymbol{\upsilon}^*$ *and* $\mathcal{D}_0^2 = \nabla^2 I\!E\mathcal{L}(\boldsymbol{\upsilon}^*)$. *Then it holds on* $\Omega(\mathtt{x}) \subseteq \Omega$ *that for all* $k \in \mathbb{N}$

$$\big\|\breve{D}\big(\widetilde{\boldsymbol{\theta}}_k - \boldsymbol{\theta}^*\big) - \breve{\boldsymbol{\xi}}\big\| \le \breve{\diamondsuit}_Q(\mathtt{r}_k, \mathtt{x}), \tag{5.A.10}$$

$$\big|2\breve{L}(\widetilde{\boldsymbol{\theta}}_k, \boldsymbol{\theta}^*) - \|\breve{\boldsymbol{\xi}}\|^2\big| \le 9\left(\|\breve{D}^{-1}\breve{\nabla}\| + \breve{\diamondsuit}_Q(\mathtt{r}_k, \mathtt{x})\right)\breve{\diamondsuit}_Q(\mathtt{r}_k, \mathtt{x}), \tag{5.A.11}$$

where the spread $\breve{\diamondsuit}(\mathtt{r}, \mathtt{x})$ *is defined in* (4.2.7), *where* $\breve{L}(\boldsymbol{\theta}) \stackrel{\text{def}}{=} \max_{\boldsymbol{\eta} \in \varUpsilon_{\boldsymbol{\eta}}} \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\eta})$ *and where*

$$\mathtt{r}_k \stackrel{\text{def}}{=} \mathtt{r}_k^{(\cdot)} \vee \mathtt{r}_0.$$

Proof. The proof is nearly the same as that of Lemma 4.A.2. We only sketch it and refer the reader to Lemma 4.A.2 for the skipped arguments. We define

$$l : \mathbb{R}^p \times \varUpsilon \to \mathbb{R}, \;\; (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\eta}) \mapsto \mathcal{L}(\boldsymbol{\theta}_1, \boldsymbol{\eta} + \mathrm{H}^{-2}A^\top(\boldsymbol{\theta}_2 - \boldsymbol{\theta}_1)).$$

Note that

$$\begin{aligned} \nabla_{\boldsymbol{\theta}_1} l(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\eta}) &= \breve{\nabla}_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}_1, \boldsymbol{\eta} + \mathrm{H}^{-2}A^\top(\boldsymbol{\theta}_2 - \boldsymbol{\theta}_1)), \\ \widetilde{\boldsymbol{\theta}}^{(k)} &= \operatorname*{argmax}_{\boldsymbol{\theta}} l(\boldsymbol{\theta}, \widetilde{\boldsymbol{\theta}}^{(k)}, \widetilde{\boldsymbol{\eta}}^{(k)}), \end{aligned}$$

such that $\breve{\nabla}_{\boldsymbol{\theta}}\mathcal{L}(\widetilde{\boldsymbol{\theta}}^{(k)}, \widetilde{\boldsymbol{\eta}}^{(k)}) = 0$. This gives

$$\big\|\breve{D}\big(\widetilde{\boldsymbol{\theta}}_k - \boldsymbol{\theta}^*\big) - \breve{\boldsymbol{\xi}}\big\| = \big\|\breve{D}^{-1}\breve{\nabla}\mathcal{L}(\widetilde{\boldsymbol{\theta}}^{(k)}, \widetilde{\boldsymbol{\eta}}^{(k)}) - \breve{D}^{-1}\breve{\nabla}\mathcal{L}(\boldsymbol{\upsilon}^*) + \breve{D}\big(\widetilde{\boldsymbol{\theta}}_k - \boldsymbol{\theta}^*\big)\big\|.$$

The right-hand side can be bounded just as in the proof of Lemma 4.A.2 the only difference being that $6\breve{\nu}_1\breve{\omega}\mathfrak{z}_1(\mathtt{x}, 2p^* + 2p)\mathtt{r}$ is replaced by $6\breve{\nu}_1\breve{\omega}(\mathfrak{z}_Q(\mathtt{x}, 2p^* + 2p)^2 + 2\mathtt{r}^2)$. This gives (5.A.10).

For (5.A.11) we can represent:

$$\breve{L}(\widetilde{\boldsymbol{\theta}}^{(k)}) - \breve{L}(\boldsymbol{\theta}^*) = l(\widetilde{\boldsymbol{\theta}}^{(k)}, \widetilde{\boldsymbol{\theta}}^{(k)}, \widetilde{\boldsymbol{\eta}}^{(k+1)}) - l(\boldsymbol{\theta}^*, \boldsymbol{\theta}^*, \widetilde{\boldsymbol{\eta}}_{\boldsymbol{\theta}^*}),$$

where

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

Due to the definition of $\widetilde{\boldsymbol{\theta}}^{(k)}$ and $\widetilde{\boldsymbol{\eta}}^{(k+1)}$

$$\begin{aligned} l(\widetilde{\boldsymbol{\theta}}^{(k)}, \boldsymbol{\theta}^*, \widetilde{\boldsymbol{\eta}}_{\boldsymbol{\theta}^*}) - l(\boldsymbol{\theta}^*, \boldsymbol{\theta}^*, \widetilde{\boldsymbol{\eta}}_{\boldsymbol{\theta}^*}) &\le \breve{L}(\widetilde{\boldsymbol{\theta}}^{(k)}) - \breve{L}(\boldsymbol{\theta}^*) \\ &\le l(\widetilde{\boldsymbol{\theta}}^{(k)}, \widetilde{\boldsymbol{\theta}}^{(k)}, \widetilde{\boldsymbol{\eta}}^{(k+1)}) - l(\boldsymbol{\theta}^*, \widetilde{\boldsymbol{\theta}}^{(k)}, \widetilde{\boldsymbol{\eta}}^{(k+1)}). \end{aligned}$$

Again the remaining steps are exactly the same as in the proof of Lemma 4.A.2.

□

5.A.4 Proof of Corollary 5.2.2

Proof. Note that with the argument of Section 5.A.3 $I\!\!P(\varOmega'(\mathtt{x})) \ge 1-8\mathrm{e}^{-\mathtt{x}}-\beta_{(\mathbf{A})}$ where with $\varOmega(\mathtt{x})$ in (5.A.3)

$$\varOmega'(\mathtt{x}) = \varOmega(\mathtt{x}) \cap \{\widetilde{\boldsymbol{\upsilon}} \in \varUpsilon_{\circ}(\mathtt{r}_0)\}.$$

On $\varOmega'(\mathtt{x})$ it holds due to Theorem 5.2.1 and due to Theorem 4.2.2

$$\|\breve{D}(\widetilde{\boldsymbol{\theta}}^{(k)} - \boldsymbol{\theta}^*) - \breve{\boldsymbol{\xi}}\| \le \breve{\diamondsuit}_Q(\mathtt{r}_k, \mathtt{x}), \quad \|\breve{D}(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) - \breve{\boldsymbol{\xi}}\| \le \breve{\diamondsuit}(\mathtt{r}_0, \mathtt{x}).$$

The claim follows with the triangular inequality. □

5.A.5 Proof of Theorem 5.2.3

We prove this Theorem in a similar manner to the convergence result in Lemma 5.A.3. Redefine the set $\varOmega(\mathtt{x})$

$$\begin{aligned}
\varOmega(\mathtt{x}) &\overset{\mathrm{def}}{=} \bigcap_{k=0}^{K} (C^{(k,k)} \cap C^{(k,k+1)}) \qquad\qquad (5.A.12)\\
&\quad \cap C(\nabla) \cap \{\mathcal{L}(\widetilde{\boldsymbol{\upsilon}}^{(0)}, \boldsymbol{\upsilon}^*) \ge -\mathrm{K}_0(\mathtt{x})\}, \text{ where}\\
C^{(k,k(+1))} &= \Big\{\|\mathcal{D}(\widetilde{\boldsymbol{\upsilon}}^{(k,k(+1))} - \boldsymbol{\upsilon}^*)\| \le \mathrm{R}_0(\mathtt{x}),\, \|D(\widetilde{\boldsymbol{\theta}}^{(k)} - \boldsymbol{\theta}^*)\| \le \mathrm{R}_0(\mathtt{x}),\\
&\quad \|\mathrm{H}(\widetilde{\boldsymbol{\eta}}^{(k(+1))} - \boldsymbol{\eta}^*)\| \le \mathrm{R}_0(\mathtt{x})\Big\},\\
C(\nabla) &= \left\{ \sup_{\boldsymbol{\upsilon} \in \varUpsilon_{\circ}(\mathrm{R}_0(\mathtt{x}))} \|\mathfrak{Y}(\nabla^2)(\boldsymbol{\upsilon})\| \le 9\nu_2\omega_2\mathfrak{z}_1(\mathtt{x}, 6p^*)\mathrm{R}_0(\mathtt{x}) \right\}\\
&\quad \cap \{\|\mathcal{D}^{-1}\nabla^2\boldsymbol{\zeta}(\boldsymbol{\upsilon}^*)\| \le \mathfrak{z}(\mathtt{x}, \nabla^2\boldsymbol{\zeta}(\boldsymbol{\upsilon}^*))\}.
\end{aligned}$$

where

$$\mathfrak{Y}(\nabla^2)(\boldsymbol{\upsilon}) \overset{\mathrm{def}}{=} \mathcal{D}^{-1}\left(\nabla^2\boldsymbol{\zeta}(\boldsymbol{\upsilon}) - \nabla^2\boldsymbol{\zeta}(\boldsymbol{\upsilon}^*)\right) \in \mathbb{R}^{p^{*2}}.$$

We see that on $\varOmega(\mathtt{x})$

$$\boldsymbol{\upsilon}^{(k,k(+1))} \in \widetilde{\varUpsilon}_{\circ}(\mathrm{R}_0) \overset{\mathrm{def}}{=} \{\|\mathcal{D}(\boldsymbol{\upsilon} - \widetilde{\boldsymbol{\upsilon}})\| \le \mathrm{R}_0 + \mathtt{r}_0\} \cap \varUpsilon_{\circ}(\mathrm{R}_0).$$

Lemma 5.A.6. *Under the conditions of Theorem 5.2.3*

$$I\!\!P(\varOmega(\mathtt{x})) \ge 1 - 3\mathrm{e}^{-\mathtt{x}} - \beta_{(\mathbf{A})}.$$

Proof. The proof is very similar to the one presented in Section 5.A.3, so we only give a sketch. By assumption

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

and due to $(\mathcal{E}\mathcal{D}_2)$ with Theorem 3.5.10

$$I\!P\left(\sup_{\boldsymbol{\upsilon}\in\varUpsilon_\circ(\mathrm{R}_0(\mathtt{x}))} \|\mathcal{Y}(\nabla^2)(\boldsymbol{\upsilon})\| \le 9\nu_2\omega_2\mathfrak{z}_1(\mathtt{x}, 6p^*)\mathrm{R}_0(\mathtt{x})\right) \ge 1 - \mathrm{e}^{-\mathtt{x}}.$$

□

Lemma 5.A.7. *Assume for some sequence* $(\mathtt{r}_k^{(l)})$ *that*

$$\bigcap_{k\in\mathbb{N}} \left\{\left\|\mathcal{D}(\widetilde{\boldsymbol{\upsilon}}^{(k,k(+1))} - \widetilde{\boldsymbol{\upsilon}})\right\| \le \mathtt{r}_k^{(l)}\right\} \subseteq \varOmega(\mathtt{x}).$$

Then we get on $\varOmega(\mathtt{x})$

$$\begin{aligned}\left\|\mathcal{D}(\widetilde{\boldsymbol{\upsilon}}^{(k,k(+1))} - \widetilde{\boldsymbol{\upsilon}})\right\| \;&\le\; 2\sqrt{2}(1+\sqrt{\nu})\sum_{r=0}^{k-1}\nu^r\|\boldsymbol{\tau}(\mathtt{r}_{k-r}^{(l)})\| + 2\sqrt{2}\nu^k(R_0 + \mathtt{r}_0),\\ &=:\; \mathtt{r}_k^{(l+1)}.\end{aligned} \tag{5.A.13}$$

where

$$\|\boldsymbol{\tau}(\mathtt{r})\| \le \left[\delta(R_0) + 9\nu_2\omega_2\|\mathcal{D}^{-1}\|\mathfrak{z}_1(\mathtt{x}, 6p^*)R_0 + \|\mathcal{D}^{-1}\|\mathfrak{z}(\mathtt{x}, \nabla^2\boldsymbol{\zeta}(\boldsymbol{\upsilon}^*))\right]\mathtt{r}.$$

Proof. 1. We first show that on $\varOmega(\mathtt{x})$

$$\begin{aligned} D(\widetilde{\boldsymbol{\theta}}^{(k)} - \widetilde{\boldsymbol{\theta}}) \;&=\; -D^{-1}A(\widetilde{\boldsymbol{\eta}}^{(k)} - \widetilde{\boldsymbol{\eta}}) + \boldsymbol{\tau}\big(\mathtt{r}_k^{(l)}\big),\\ \mathrm{H}(\widetilde{\boldsymbol{\eta}}^{(k)} - \boldsymbol{\eta}^*) \;&=\; -\mathrm{H}^{-1}A^\top(\widetilde{\boldsymbol{\theta}}^{(k-1)} - \widetilde{\boldsymbol{\theta}}) + \boldsymbol{\tau}\big(\mathtt{r}_k^{(l)}\big),\end{aligned}$$

The proof is very similar to that of Lemma 5.A.3. Define

$$\alpha(\boldsymbol{\upsilon}, \widetilde{\boldsymbol{\upsilon}}) := \mathcal{L}(\boldsymbol{\upsilon}, \widetilde{\boldsymbol{\upsilon}}) + \|\mathcal{D}(\boldsymbol{\upsilon} - \widetilde{\boldsymbol{\upsilon}})\|^2/2.$$

Note that

$$\begin{aligned}\mathcal{L}(\boldsymbol{\upsilon}, \widetilde{\boldsymbol{\upsilon}}) \;&=\; \nabla\mathcal{L}(\boldsymbol{\upsilon}) - \|\mathcal{D}(\boldsymbol{\upsilon} - \widetilde{\boldsymbol{\upsilon}})\|^2/2 + \alpha(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^*)\\ &=\; -\|D(\boldsymbol{\theta} - \widetilde{\boldsymbol{\theta}})\|^2/2 + (\boldsymbol{\theta} - \boldsymbol{\theta}^*)^\top A(\boldsymbol{\eta} - \widetilde{\boldsymbol{\eta}})\\ &\qquad -\|\mathrm{H}(\boldsymbol{\eta} - \widetilde{\boldsymbol{\eta}})\|^2/2 + \alpha(\boldsymbol{\upsilon}, \widetilde{\boldsymbol{\upsilon}}).\end{aligned}$$

Setting $\nabla_{\boldsymbol{\theta}}\mathcal{L}(\widetilde{\boldsymbol{\theta}}^{(k)},\widetilde{\boldsymbol{\eta}}^{(k)})=0$ we find

$$D(\widetilde{\boldsymbol{\theta}}^{(k)}-\widetilde{\boldsymbol{\theta}})=D^{-1}A(\widetilde{\boldsymbol{\eta}}^{(k)}-\widetilde{\boldsymbol{\eta}})+D^{-1}\nabla_{\boldsymbol{\theta}}\alpha(\widetilde{\boldsymbol{\upsilon}}^{(k,k)},\widetilde{\boldsymbol{\upsilon}}).$$

We want to show

$$\sup_{(\boldsymbol{\theta},\widetilde{\boldsymbol{\eta}}^{(k)})\in\widetilde{\varUpsilon}_{\circ}\left(\mathtt{r}_k^{(l)}\right)\cap\varUpsilon_{\circ}(\mathrm{R}_0)} D^{-1}\nabla_{\boldsymbol{\theta}}\alpha((\boldsymbol{\theta},\widetilde{\boldsymbol{\eta}}^{(k)}),\widetilde{\boldsymbol{\upsilon}})\le\|\boldsymbol{\tau}\big(\mathtt{r}_k^{(l)}\big)\|,$$

where

$$D^{-1}\nabla_{\boldsymbol{\theta}}\alpha(\boldsymbol{\upsilon},\widetilde{\boldsymbol{\upsilon}})\overset{\mathrm{def}}{=}D^{-1}\big\{\nabla_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\upsilon})-D^2\,(\boldsymbol{\theta}-\widetilde{\boldsymbol{\theta}})-A(\widetilde{\boldsymbol{\eta}}^{(k)}-\widetilde{\boldsymbol{\eta}})\big\}.$$

To see this note that by assumption we have

$$\varOmega(\mathtt{x})\subseteq\{\widetilde{\boldsymbol{\upsilon}}\in\varUpsilon_{\circ}(\mathtt{r}_0)\}\subseteq\{\widetilde{\boldsymbol{\upsilon}}\in\varUpsilon_{\circ}(\mathrm{R}_0)\}.$$

By condition $(\mathcal{L}_0)$, Lemma 4.A.5 and Taylor expansion we have

$$\begin{aligned}
&\sup_{(\boldsymbol{\theta},\widetilde{\boldsymbol{\eta}}^{(k)})\in\widetilde{\varUpsilon}_{\circ}(\mathtt{r}_k^{(l)})\cap\varUpsilon_{\circ}(\mathrm{R}_0)}\|\mathbb{E}\mathcal{U}(\boldsymbol{\theta},\widetilde{\boldsymbol{\eta}}^{(k)})\|\\
&\quad\le\sup_{\boldsymbol{\upsilon}\in\widetilde{\varUpsilon}_{\circ}(\mathtt{r}_k^{(l)})\cap\varUpsilon_{\circ}(\mathrm{R}_0)}\|D^{-1}\varPi_{\boldsymbol{\theta}}\Big(\nabla\mathbb{E}\mathcal{L}(\boldsymbol{\upsilon})-\nabla\mathbb{E}\mathcal{L}(\widetilde{\boldsymbol{\upsilon}})-\mathcal{D}\,(\boldsymbol{\upsilon}-\boldsymbol{\upsilon}^*)\Big)\|\\
&\quad\le\sup_{\boldsymbol{\upsilon}\in\varUpsilon_{\circ}(\mathrm{R}_0)}\|D^{-1}\varPi_{\boldsymbol{\theta}}\mathcal{D}\|\|\mathcal{D}^{-1}\nabla^2\mathbb{E}\mathcal{L}(\boldsymbol{\upsilon})\mathcal{D}^{-1}-I_{p^*}\|\mathtt{r}_k^{(l)}\\
&\quad\le\delta(\mathrm{R}_0)\mathtt{r}_k^{(l)}.
\end{aligned}$$

For the remainder note that with $\zeta=\mathcal{L}-\mathbb{E}\mathcal{L}$ on $\varOmega(\mathtt{x})$ using Lemma 4.A.5 we can bound

$$\begin{aligned}
&\sup_{(\boldsymbol{\theta},\widetilde{\boldsymbol{\eta}}^{(k)})\in\widetilde{\varUpsilon}_{\circ}(\mathtt{r}_k^{(l)})\cap\varUpsilon_{\circ}(\mathrm{R}_0)}\Big\|\mathcal{U}_{\boldsymbol{\theta}}(\boldsymbol{\theta},\widetilde{\boldsymbol{\eta}}^{(k)})-\mathbb{E}\mathcal{U}_{\boldsymbol{\theta}}(\boldsymbol{\theta},\widetilde{\boldsymbol{\eta}}^{(k)})\Big\|\\
&\quad\le\sup_{\boldsymbol{\upsilon}\in\widetilde{\varUpsilon}_{\circ}(\mathtt{r}_k^{(l)})\cap\varUpsilon_{\circ}(\mathrm{R}_0)}\Big\|D^{-1}\Big(\nabla_{\boldsymbol{\theta}}\boldsymbol{\zeta}(\boldsymbol{\upsilon})-\nabla_{\boldsymbol{\theta}}\boldsymbol{\zeta}(\widetilde{\boldsymbol{\upsilon}})\Big)\Big\|\\
&\quad\le\sup_{\boldsymbol{\upsilon}\in\varUpsilon_{\circ}(\mathrm{R}_0)}\big\|\mathcal{D}^{-1}\nabla^2\boldsymbol{\zeta}(\boldsymbol{\upsilon})\mathcal{D}^{-1}\big\|\,\mathtt{r}_k^{(l)}\\
&\quad\le\sup_{\boldsymbol{\upsilon}\in\varUpsilon_{\circ}(\mathrm{R}_0)}\bigg\{\frac{1}{9\nu_2\omega_2}\|\mathcal{D}^{-1}\left(\nabla^2\boldsymbol{\zeta}(\boldsymbol{\upsilon})-\nabla^2\boldsymbol{\zeta}(\boldsymbol{\upsilon}^*)\right)\mathcal{D}^{-1}\|\bigg\}6\nu_1\omega\mathtt{r}_k^{(l)}\\
&\qquad+\bigg\{\|\mathcal{D}^{-1}\nabla^2\boldsymbol{\zeta}(\boldsymbol{\upsilon}^*)\mathcal{D}^{-1}\|\bigg\}\mathtt{r}_k^{(l)}\\
&\quad\le\left[9\nu_2\omega_2\|\mathcal{D}^{-1}\|\mathfrak{z}_1(\mathtt{x},6p^*)\mathrm{R}_0+\|\mathcal{D}^{-1}\|\mathfrak{z}(\mathtt{x},\nabla^2\boldsymbol{\zeta}(\boldsymbol{\upsilon}^*))\right]\mathtt{r}_k^{(l)}.
\end{aligned}$$

Using the same argument for $\widetilde{\boldsymbol{\eta}}^{(k)}$ gives the claim.

The claim follows as in the proof of Lemma 5.A.3. □

Lemma 5.A.8. *Assume that* $\varkappa(\mathtt{x}, R_0) < 1 - \nu$ *where*

$$\varkappa(\mathtt{x}, R_0) \overset{\text{def}}{=} \frac{2\sqrt{2}(1+\sqrt{\nu})}{\sqrt{1-\nu}} \Bigg(\delta(R_0) + 9\omega_2\nu_2 \|\mathcal{D}^{-1}\|_{\mathfrak{z}_1}(\mathtt{x}, 6p^*) R_0 \\ + \|\mathcal{D}^{-1}\|_{\mathfrak{z}} \left(\mathtt{x}, \nabla^2 \mathcal{L}(\boldsymbol{\upsilon}^*)\right) \Bigg).$$

Then

$$\varOmega(\mathtt{x}) \subseteq \bigcap_{k\in\mathbb{N}} \left\{ \boldsymbol{\upsilon}^{(k,k(+1))} \in \widetilde{\varUpsilon_{\circ}}(\mathtt{r}_k) \right\},$$

where $(\mathtt{r}_k)_{k\in\mathbb{N}}$ *satisfy the bound* (5.2.9).

Proof. Define for all $k \in \mathbb{N}_0$ the sequence $\mathtt{r}_k^{(0)} = \mathrm{R}_0$. We estimate

$$\|\boldsymbol{\tau}(\mathtt{r}_k^{(l)})\| \le \frac{1}{\sqrt{1-\nu}} \Bigg(\delta(\mathrm{R}_0) + 6\nu_1\omega_2 \|\mathcal{D}^{-1}\|_{\mathfrak{z}_1}(\mathtt{x}, 6p^*)\mathrm{R}_0 \\ + \|\mathcal{D}^{-1}\|_{\mathfrak{z}}(\mathtt{x}, I\!\!B(\nabla^2) \Bigg) \mathtt{r}_k^{(l)},$$

such that by definition

$$2\sqrt{2}(1+\sqrt{\nu}) \sum_{r=0}^{k-1} \nu^r \|\boldsymbol{\tau}(\mathtt{r}_{k-r}^{(l)})\| \le \varkappa(\mathtt{x}, \mathrm{R}_0) \sum_{r=0}^{k-1} \nu^r \mathtt{r}_{k-r}^{(l)}.$$

Plugging in the recursive formula for $\mathtt{r}_k^{(l)}$ in (5.A.13) we find

$$\begin{aligned} \mathtt{r}_k^{(l)} &\le \varkappa(\mathtt{x}, \mathrm{R}_0) \sum_{r=0}^{k-1} \nu^r \mathtt{r}_{k-r}^{(l-1)} + 2\sqrt{2}\nu^k(\mathrm{R}_0 + \mathtt{r}_0) \\ &\le \varkappa(\mathtt{x}, \mathrm{R}_0) \sum_{r=0}^{k-1} \nu^r \left(\varkappa(\mathtt{x}, \mathrm{R}_0) \sum_{s=0}^{k-r-1} \nu^s \mathtt{r}_{k-r-s}^{(l-2)} + 2\nu^{k-r}(\mathrm{R}_0 + \mathtt{r}_0) \right) \\ &\quad + 2\sqrt{2}(\mathrm{R}_0 + \mathtt{r}_0)\nu^k \\ &\le \varkappa(\mathtt{x}, \mathrm{R}_0)^2 \sum_{r=0}^{k-1} \nu^r \sum_{s=0}^{k-r-1} \nu^s \mathtt{r}_{k-r-s}^{(l-2)} + 2\sqrt{2}\nu^k(\mathrm{R}_0 + \mathtt{r}_0)\left(\varkappa(\mathtt{x}, \mathrm{R}_0)k + 1\right). \end{aligned}$$

By induction this gives for $l \in \mathbb{N}$

$$\begin{aligned}
\mathtt{r}_k^{(l)} &\le \varkappa(\mathtt{x}, \mathrm{R}_0)^l \sum_{r_1=0}^{k-1} \nu^{r_1} \sum_{r_2=0}^{k-r_1-1} \nu^{r_2} \dots \sum_{r_l=0}^{k-\sum_{s=1}^{l-1} r_s - 1} \nu^{r_l} (\mathrm{R}_0 + \mathtt{r}_0) \\
&\quad + 2\sqrt{2}\nu^k (\mathrm{R}_0 + \mathtt{r}_0) \sum_{s=0}^{l-1} \varkappa(\mathtt{x}, \mathrm{R}_0)^s k^s \\
&\le \left(\left(\frac{\varkappa(\mathtt{x}, \mathrm{R}_0)}{1-\nu} \right)^l + 2\sqrt{2}\nu^k \sum_{s=0}^{l-1} \bigl(\varkappa(\mathtt{x}, \mathrm{R}_0) k\bigr)^s \right) (\mathrm{R}_0 + \mathtt{r}_0) \\
&\le \begin{cases} \left(\left(\frac{\varkappa(\mathtt{x},\mathrm{R}_0)}{1-\nu} \right)^l + 2\sqrt{2}\nu^k \frac{1}{1-\varkappa(\mathtt{x},\mathrm{R}_0)k} \right) (\mathrm{R}_0 + \mathtt{r}_0), & \varkappa(\mathtt{x}, \mathrm{R}_0) k \le 1, \\ \varkappa(\mathtt{x}, \mathrm{R}_0)^l \left(\left(\frac{1}{1-\nu} \right)^l + 2\sqrt{2}\nu^k \frac{k^l}{\varkappa(\mathtt{x},\mathrm{R}_0)k-1} \right) (\mathrm{R}_0 + \mathtt{r}_0), & \text{otherwise.} \end{cases}
\end{aligned}$$

By Lemma 5.A.7

$$\varOmega(\mathtt{x}) \subset \bigcap_{k \in \mathbb{N}_0} \bigcap_{l \in \mathbb{N}} \Bigl\{ \widetilde{\boldsymbol{\upsilon}}^{(k,k(+1))} \in \widetilde{\varUpsilon}_\circ\bigl(\mathtt{r}_k^{(l)}\bigr) \Bigr\}.$$

Set if $\varkappa(\mathtt{x}, \mathrm{R}_0)/(1-\nu) < 1$

$$l(k) \overset{\mathrm{def}}{=} \begin{cases} \infty, & \varkappa(\mathtt{x}, \mathrm{R}_0) k \le 1, \\ \frac{k \log(\nu) + \log(2\sqrt{2}) - \log(\varkappa(\mathtt{x},\mathrm{R}_0)k-1)}{-\log(1-\nu) - \log(k)}, & \text{otherwise.} \end{cases}$$

Then with $\mathtt{r}_k^{**} \overset{\mathrm{def}}{=} \mathtt{r}_k^{(\lfloor l(k) \rfloor)}$ we get

$$\begin{aligned}
\varOmega(\mathtt{x}) &\subset \bigcap_{k \in \mathbb{N}_0} \Bigl\{ \widetilde{\boldsymbol{\upsilon}}^{(k,k(+1))} \in \widetilde{\varUpsilon}_\circ\bigl(\mathtt{r}_k^{*}\bigr) \Bigr\}, \\
\mathtt{r}_k^{**} &\le \begin{cases} \frac{\nu^k 2\sqrt{2}}{1-\varkappa(\mathtt{x},\mathrm{R}_0)k} (\mathrm{R}_0 + \mathtt{r}_0), & \varkappa(\mathtt{x}, \mathrm{R}_0) k \le 1, \\ 2 \left(\frac{\varkappa(\mathtt{x},\mathrm{R}_0)}{1-\nu} \right)^{\frac{k}{\log(k)} L(k) - 1} (\mathrm{R}_0 + \mathtt{r}_0). & \text{otherwise,} \end{cases}
\end{aligned}$$

The sequence $L(k) > 0$ is defined as

$$L(k) \overset{\mathrm{def}}{=} \left\lfloor \frac{\log(1/\nu) - \frac{1}{k}\bigl(\log(2\sqrt{2}) - \log(\varkappa(\mathtt{x}, \mathrm{R}_0) k - 1)\bigr)}{1 + \frac{1}{\log(k)} \log(1-\nu)} \right\rfloor \in \mathbb{N},$$

where $\lfloor x \rfloor \in \mathbb{N}_0$ denotes the largest natural number smaller than $x > 0$. To ensure that $L(k) > 0$ we assume that $k \log(1/\nu) - \log(2\sqrt{2}) > k$. Further as $\varkappa(\mathtt{x}, \mathrm{R}_0) < (1-\nu)$ and $L(k)$ is only relevant once $\varkappa(\mathtt{x}, \mathrm{R}_0)k > 1$ it follows that

$$0 < 1 + \frac{1}{\log(k)} \log(1-\nu) < 1.$$

Then

$$L(k) \geq \log(1/\nu) - \frac{1}{k}\Big(\log(2\sqrt{2}) - \log(\varkappa(\mathtt{x}, \mathrm{R}_0)k - 1)\Big) > 1.$$

Consequently

$$\begin{aligned}\left(\frac{\varkappa(\mathtt{x}, \mathrm{R}_0)}{1-\nu}\right)^{\frac{k}{\log(k)}L(k)} &\leq \nu^{\frac{k}{\log(k)}\log\left(\frac{1-\nu}{\varkappa(\mathtt{x},\mathrm{R}_0)}\right)} \left(\frac{\varkappa(\mathtt{x}, \mathrm{R}_0)}{1-\nu}\right)^{-\frac{1}{\log(k)}\left(\log(2\sqrt{2})-\log(\varkappa(\mathtt{x},\mathrm{R}_0)k-1)\right)} \\ &\overset{\text{def}}{=} \nu^{\frac{k}{\log(k)}\log\left(\frac{1-\nu}{\varkappa(\mathtt{x},\mathrm{R}_0)}\right)} c_k,\end{aligned}$$

where $c_k \to \frac{\varkappa(\mathtt{x},\mathrm{R}_0)}{1-\nu}$. Finally note that $\mathrm{R}_0 + \mathtt{r}_0 \leq 2\mathrm{R}_0$ and the proof is complete. □

Remark 5.A.3. As pointed out in Remark 5.2.15 the above result can be improved. Assume that $\delta(\mathtt{r})/\mathtt{r} \vee 6\nu_1\omega_2 \leq \epsilon$ for some $\epsilon > 0$ and assume (A_3) from Section 5.2.2. Redefine $\varOmega(\mathtt{x})$ as the intersection of the two sets in (5.A.3) and (5.A.12). Then $I\!\!P(\varOmega(\mathtt{x})) \geq 1 - 10\mathrm{e}^{-\mathtt{x}}$. Also redefine

$$\begin{aligned}\varkappa(\mathtt{x}, \mathtt{r}) \overset{\text{def}}{=} \frac{2\sqrt{2}(1+\sqrt{\nu})}{\sqrt{1-\nu}}\bigg(&\epsilon\mathtt{r} + 3\epsilon\|\mathcal{D}^{-1}\|_{\mathfrak{z}1}(\mathtt{x}, 6p^*)\mathtt{r} \\ &+ \|\mathcal{D}^{-1}\|_{\mathfrak{z}}\left(\mathtt{x}, \nabla^2\mathcal{L}(\boldsymbol{\upsilon}^*)\right)\bigg).\end{aligned}$$

By the arguments of the proof of Theorem 5.2.1 we find with $\mathtt{r}_k^*$ defined in (5.A.6)

$$\bigcap_{k\in\mathbb{N}} \{\boldsymbol{\upsilon}^{(k,k(+1))} \in \varUpsilon_\circ(\mathtt{r}_k^*)\}.$$

Using this in Lemma 5.A.7 instead of $\cap_{k\in\mathbb{N}}\{\boldsymbol{\upsilon}^{(k,k(+1))} \in \varUpsilon_\circ(\mathrm{R}_0)\}$ we can bound

$$\begin{aligned}\|\boldsymbol{\tau}(\mathtt{r}_k^{(l)})\| \leq \frac{1}{\sqrt{1-\nu}}\bigg(&\delta(\mathtt{r}_k^*) + 6\nu_1\omega_2\|\mathcal{D}^{-1}\|_{\mathfrak{z}1}(\mathtt{x}, 6p^*)\mathtt{r}_k^* \\ &+\|\mathcal{D}^{-1}\|_{\mathfrak{z}}(\mathtt{x}, I\!\!B(\nabla^2)\bigg)\mathtt{r}_k^{(l)}.\end{aligned}$$

Consequently, representing $\mathtt{r}_k^* = \mathtt{C}\left(\mathfrak{z}(\mathtt{x}) + \nu^k \mathrm{R}_0\right)$ and using $\delta(\mathtt{r})/\mathtt{r} \vee 6\nu_1\omega_2 \leq \epsilon$ we find

$$\begin{aligned}
\mathtt{r}_k^{(l)} &\leq \varkappa(\mathtt{x}, \mathtt{C}\mathfrak{z}(\mathtt{x})) \sum_{r=0}^{k-1} \nu^r \mathtt{r}_{k-r}^{(l-1)} + 2\sqrt{2}\nu^k(\mathrm{R}_0 + \mathtt{r}_0) \\
&\qquad + \mathtt{C}\epsilon(1 + \|\mathcal{D}^{-1}\|\mathfrak{z}_1(\mathtt{x}, 6p^*)) \sum_{r=0}^{k-1} \nu^k \mathrm{R}_0 \mathtt{r}_{k-r}^{(l-1)} \\
&\leq \varkappa(\mathtt{x}, \mathtt{C}\mathfrak{z}(\mathtt{x})) \sum_{r=0}^{k-1} \nu^r \mathtt{r}_{k-r}^{(l-1)} + \mathtt{C}_1 \epsilon \mathrm{R}_0 k \nu^k (\mathrm{R}_0 + \mathtt{r}_0),
\end{aligned}$$

where $\mathtt{C}_1 \geq 2\sqrt{2} + \mathtt{C}(1 + \|\mathcal{D}^{-1}\|\mathfrak{z}_1(\mathtt{x}, 6p^*))$. With the same arguments as in the proof of Lemma 5.A.8 we infer

$$\begin{aligned}
&\mathtt{r}_k^{(l)}/(\mathrm{R}_0 + \mathtt{r}_0) \\
&\quad \leq \begin{cases} \left(\left(\frac{\varkappa(\mathtt{x}, \mathtt{C}\mathfrak{z}(\mathtt{x}))}{1-\nu} \right)^l + k\nu^k \frac{\mathtt{C}_1 \epsilon \mathrm{R}_0}{1 - \varkappa(\mathtt{x}, \mathtt{C}\mathfrak{z}(\mathtt{x}))k} \right), & \varkappa(\mathtt{x}, \mathtt{C}\mathfrak{z}(\mathtt{x}))k \leq 1, \\ \varkappa(\mathtt{x}, \mathtt{C}\mathfrak{z}(\mathtt{x}))^l \left(\left(\frac{1}{1-\nu} \right)^l + \nu^k \frac{k^{l+1} \mathtt{C}_1 \epsilon \mathrm{R}_0}{\varkappa(\mathtt{x}, \mathtt{C}\mathfrak{z}(\mathtt{x}))k - 1} \right), & \text{otherwise.} \end{cases}
\end{aligned}$$

Set

$$l(k) \overset{\mathrm{def}}{=} \begin{cases} \infty, & \varkappa(\mathtt{x}, \mathtt{C}\mathfrak{z}(\mathtt{x}))k \leq 1, \\ \frac{k \log(\nu) + \log(\mathtt{C}_1 \epsilon \mathrm{R}_0) - \log(\varkappa(\mathtt{x}, \mathtt{C}\mathfrak{z}(\mathtt{x}))k - 1) - \log(k)}{-\log(1-\nu) - \log(k)}, & \text{otherwise.} \end{cases}$$

Then with $\mathtt{r}_k^* \overset{\mathrm{def}}{=} \mathtt{r}_k^{(\lfloor l(k) \rfloor)}$ we get with a slight adaptation of $L(k)$

$$\mathit{\Omega}(\mathtt{x}) \subset \bigcap_{k \in \mathbb{N}_0} \left\{ \widetilde{\boldsymbol{\upsilon}}^{(k, k(+1))} \in \widetilde{\mathit{\Upsilon}}_\circ(\mathtt{r}_k^*) \right\},$$

$$\mathtt{r}_k^* \leq \begin{cases} \frac{\nu^k 2\sqrt{2}}{1 - \varkappa(\mathtt{x}\mathtt{C}\mathfrak{z}(\mathtt{x}))k} (\mathrm{R}_0 + \mathtt{r}_0), & \varkappa(\mathtt{x}, \mathtt{C}\mathfrak{z}(\mathtt{x}))k \leq 1, \\ 2 \left(\frac{\varkappa(\mathtt{x}, \mathtt{C}\mathfrak{z}(\mathtt{x}))}{1-\nu} \right)^{\frac{k}{\log(k)} L(k) - 1} (\mathrm{R}_0 + \mathtt{r}_0), & \text{otherwise.} \end{cases}$$

Chapter 6

Projection pursuit and the single index model

In this chapter we explain how to apply the results of the previous chapters to the analysis of profile M-Estimators in the single index model and how to analyse the performance of the projection pursuit procedure. It is largely based on [1].

6.1 Finding the most interesting directions of a data set

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

$$Y_i = g(\mathbf{X}_i) + \varepsilon_i, \ i = 1, ..., n, \tag{6.1.1}$$

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

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

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 [54]. [20] propose to use a projection pursuit approach to circumvent this problem in situations where

$$g(\mathbf{X}) \approx \sum_{l=1}^{M} f_{(l)}(\mathbf{X}^\top \boldsymbol{\theta}^*_{(l)}), \tag{6.1.2}$$

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 (6.1.2) 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 (6.1.2) becomes rather restrictive. On the other hand, for large $M \in \mathbb{N}$ and large $p \in \mathbb{N}$ the assumption (6.1.2) 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 $g^\circ(\boldsymbol{Z}) = I\!E[Y|\boldsymbol{Z}]$. 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 (6.1.2) 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 [28] and [32] for a more sophisticated approach of showing that smooth functions g can be well approximated as in (6.1.2). [20] suggest to estimate the pairs $(f_l, \boldsymbol{\theta}_l^*)$ iteratively. The first task is to estimate

$$\boldsymbol{\theta}^*_{(1)} \overset{\text{def}}{=} \operatorname*{argmin}_{\boldsymbol{\theta} \in S_1^{p,+}} I\!E\left[\Big(g(\mathbf{X}) - I\!E[g(\mathbf{X})|\mathbf{X}^\top \boldsymbol{\theta}]\Big)^2\right]. \tag{6.1.3}$$

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

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

In this chapter we will mainly focus on the task (6.1.3). It has been observed in [24] 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

$$g(\mathbf{X}) = f(\mathbf{X}^\top \boldsymbol{\theta}^*), \tag{6.1.4}$$

for some $f : \mathbb{R} \to \mathbb{R}$ and $\boldsymbol{\theta}^* \in S_1^{p,+} \subset \mathbb{R}^p$, the estimation problem (6.1.3) becomes the task to estimate the linear response vector in a semiparametric single-index model (see [30]). 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 $(\varepsilon_i) \in \mathbb{R}$

$$Y_i = f(h(\mathbf{X}_i)) + \varepsilon_i,\ i = 1, ..., n. \tag{6.1.5}$$

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 $h(\boldsymbol{\theta}, \boldsymbol{x}) = \boldsymbol{\theta}^\top \boldsymbol{x}$. [60] compares the asymptotic distributions of two different prominent estimation procedures for $\boldsymbol{\theta}^*$. The first is the average derivative estimation introduced by [47] and refined by [26] and is based on the fact that if (6.1.4) is correct

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

which suggests to estimate $\boldsymbol{\theta}^*$ via an estimate of $I\!E\left[f'(\boldsymbol{\theta}^* \mathbf{X})\right]$. The second one is the minimal conditional variance estimation by [61] which is inspired by [23] and aims at directly solving (6.1.3) via a local linear approximation of $I\!E[y|\mathbf{X}^\top \boldsymbol{\theta}]$. Further results are the asymptotic efficiency of a semiparametric maximum-likelihood estimator shown by [15] for particular examples and in [23] the right choice of the bandwidth for the nonparametric estimation of the link function.

In this chapter we want to use a different approach to carry out the first step (6.1.3) that allows to apply the results of the previous chapters. For this purpose denote

$$I\!E[g(\mathbf{X})|\mathbf{X}^\top \boldsymbol{\theta}^*] = f(\mathbf{X}^\top \boldsymbol{\theta}^*). \tag{6.1.6}$$

Assume that $f \in \operatorname{span}\{(\boldsymbol{e}_k)_{k\in\mathbb{N}}\}$ for a given set of basis functions $(\boldsymbol{e}_k)_{k\in\mathbb{N}} \subset \mathcal{X}$. For some $m \ge 1$ 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 $\boldsymbol{\eta} = (\eta_1, \dots, \eta_m)^\top \in \mathbb{R}^m$. Further assume that $I\!P(\mathbf{X}_i \in B_{s_{\mathbf{X}}}(0)) \approx 1$ for some $s_{\mathbf{X}} > 0$. Our aim is to analyse for $m \in \mathbb{N}$ the properties of the estimator

$$\widetilde{\boldsymbol{\theta}}_m \overset{\text{def}}{=} \varPi_{\boldsymbol{\theta}} \widetilde{\boldsymbol{\upsilon}}_m \overset{\text{def}}{=} \varPi_{\boldsymbol{\theta}} \operatorname*{argmax}_{(\boldsymbol{\theta},\boldsymbol{\eta})\in\varUpsilon_m} \mathcal{L}_m(\boldsymbol{\theta}, \boldsymbol{\eta}), \tag{6.1.7}$$

where

$$\mathcal{L}_m(\boldsymbol{\theta},\boldsymbol{\eta}) = -\sum_{\{i\colon \|\mathbf{X}_i\|\le s_{\mathbf{X}}\}} \|Y_i - \boldsymbol{f}_{\boldsymbol{\eta}}(\mathbf{X}_i^\top\boldsymbol{\theta})\|^2/2. \tag{6.1.8}$$

The set $\varUpsilon_m$ satisfies $\varUpsilon_m = S_1^{p,+} \times B_{\mathtt{r}^\circ}^m \subset \mathbb{R}^p \times \mathbb{R}^m$ where $B_{\mathtt{r}^\circ}^m \subset \mathbb{R}^m$ denotes the centered ball of radius $\mathtt{r}^\circ > 0$. Note that this is exactly the type of estimator presented in Section 4.3. In [30] a very similar estimator is analyzed based on a "leave one out" kernel estimation of $I\!\!E[Y_i|\mathbf{X}_i^\top\boldsymbol{\theta}]$ 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 6.1.1. 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 6.3.6.

Remark 6.1.2. To avoid undesirable boundary effects (see Remark 6.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 6.2.1 we assume in condition $(\mathbf{Cond}_{\mathbf{X}})$ that there is positive probability that $\mathbf{X} \in B_{s_{\mathbf{X}}+c_B}(0)\backslash B_{s_{\mathbf{X}}}(0)$ (see also Remark 6.A.5). 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.

The estimator $\widetilde{\boldsymbol{\theta}}_m$ in (6.1.7) is supposed to approach

$$\boldsymbol{\theta}^* \stackrel{\text{def}}{=} \varPi_{\boldsymbol{\theta}}(\boldsymbol{\theta}^*,\boldsymbol{\eta}^*) \stackrel{\text{def}}{=} \varPi_{\boldsymbol{\theta}} \operatorname*{argmax}_{(\boldsymbol{\theta},\boldsymbol{\eta})\in\varUpsilon} I\!\!E\mathcal{L}_\infty(\boldsymbol{\theta},\boldsymbol{\eta}), \tag{6.1.9}$$

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

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

Remark 6.1.3. To understand the motivation of this functional note that for any $\boldsymbol{\theta} \in S_1^{p,+}$ the sequence

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

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

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

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

$$\boldsymbol{f}_{\boldsymbol{\eta}^*_{\boldsymbol{\theta}}}(\mathbf{X}^\top\boldsymbol{\theta}) = I\!E[g(\mathbf{X})|\mathbf{X}^\top\boldsymbol{\theta}], \tag{6.1.10}$$

such that the target (6.1.9) indeed coincides with the most informative direction in (6.1.3).

Remark 6.1.4. Note that there is a model bias and an approximation bias of the form

$$\begin{aligned}
\text{"}\mathit{model\,bias}\text{"} &= \min_{\boldsymbol{\upsilon}\in\varUpsilon} I\!E\|g(\mathbf{X}) - \boldsymbol{f}_{\boldsymbol{\eta}}(\mathbf{X}^\top\boldsymbol{\theta})\|^2, \\
\text{"}\mathit{approximation\,bias}\text{"} &= \min_{\boldsymbol{\upsilon}\in\varUpsilon_m} I\!E\|\boldsymbol{f}_{\boldsymbol{\eta}}(\mathbf{X}^\top\boldsymbol{\theta}) - \boldsymbol{f}_{\boldsymbol{\eta}^*}(\mathbf{X}^\top\boldsymbol{\theta}^*)\|^2, \quad (6.1.11)
\end{aligned}$$

which both have to be accounted for.

As pointed out we will analyze the properties of the estimator $\widetilde{\boldsymbol{\theta}}_m$ in (6.1.7) using the results of Chapter 4 and Chapter 5. 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 we still manage to control the sieve bias in (6.1.11). Even though we assume what is necessary to apply the results of the previous chapters, the calculations needed to check the conditions from Section 4.2.1 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 6.3.1. 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 chapter. A naive approach to satisfy the important condition $(\mathcal{L}\mathtt{r})$ from Section 4.2.1 would include a bound for

$$\sup_{\boldsymbol{\upsilon}\in\varUpsilon_m} \left| I\!E[\mathcal{L}(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^*)|(\mathbf{X}_i)_{i=1,\ldots,n}] - I\!E\mathcal{L}(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^*)\right|. \tag{6.1.12}$$

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 [39]. Mendelson's crucial insight is that to obtain $I\!E[\mathcal{L}(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^*)|(\mathbf{X}_i)_{i=1,\ldots,n}] \geq \mathtt{br}^2$ one only has to ensure that

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

We follow this route in the proof of Lemma 6.3.7. But we only apply this idea in the case that $\mathtt{C}_{bias} = 0$. In the general case we derive a bound for (6.1.12) 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 get accurate results when applying Theorem 4.2.2.

Altogether this chapter is more a proof of concept than an illustration of the elegance and applicability of the theoretical results of Chapter 4 and Chapter 5. But on the other hand it has to be kept in mind, that the results attained are considerably stronger than the weak convergence results usually aimed for in this context.

6.2 Main results

6.2.1 Assumptions

To apply the technique presented in Chapter 4 and Chapter 5 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 $(\mathbf{X}_i)_{i=1,\ldots,n} \subset \mathbb{R}^p$ are i.i.d with distribution denoted by $I\!\!P^{\mathbf{X}}$ and independent of $(\varepsilon_i)_{i=1,\ldots,n} \subset \mathbb{R}$. The measure $I\!\!P^{\mathbf{X}}$ is absolutely continuous with respect to the Lebesgue measure. The Lebesgue density $p_{\mathbf{X}}$ of $I\!\!P^{\mathbf{X}}$ is Lipschitz continuous on $B_{s_{\mathbf{X}}}(0) \subset \mathbb{R}^p$ with Lipschitz constant $L_{p_{\mathbf{X}}} > 0$. Furthermore we assume that for any pair $\boldsymbol{\theta}, \boldsymbol{\theta}^\circ \in S_1^{+,p}$ with $\boldsymbol{\theta} \perp \boldsymbol{\theta}^\circ$ we have $\operatorname{Var}\left(\mathbf{X}^\top\boldsymbol{\theta} \middle| \mathbf{X}^\top\boldsymbol{\theta}^*\right) > \sigma^2_{\mathbf{X}|\perp}$ for some constant $\sigma^2_{\mathbf{X}|\perp} > 0$ that does not depend on $\mathbf{X}^\top\boldsymbol{\theta}^* \in \mathbb{R}$. Furthermore assume that for all such pairs $\left\|\frac{p_{\boldsymbol{\theta}^\circ,\boldsymbol{\theta}}}{p_{\boldsymbol{\theta}}}\right\|_\infty < \infty$ with $p_{\boldsymbol{\theta}^\circ,\boldsymbol{\theta}} : \mathbb{R}^2 \to \mathbb{R}_+$ denoting the density of $(\mathbf{X}^\top\boldsymbol{\theta}^\circ, \mathbf{X}^\top\boldsymbol{\theta}) \in \mathbb{R}^2$. Also let on $B_{s_{\mathbf{X}}+c_B}(0)$ the density satisfy $p_{\mathbf{X}} > c_{p_{\mathbf{X}}} > 0$ for some constants $c_{p_{\mathbf{X}}}, c_B > 0$.

Remark 6.2.1. $\operatorname{Var}\left(\mathbf{X}^\top\boldsymbol{\theta}^\circ \middle| \mathbf{X}^\top\boldsymbol{\theta}^*\right) = 0$ 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 $p_{\mathbf{X}} > c_{p_{\mathbf{X}}} > 0$ to ensure identifiability.

Remark 6.2.2. We assume that the support of $I\!\!P^{\mathbf{X}}$ contains 0 without loss of generality. If that was the case one could modify the sample as follows. Let $\boldsymbol{x}_0$ be an inner point of the support of $I\!\!P^{\mathbf{X}}$. Generate a new sample $(\mathbf{X}_i')_{i=1,\ldots,n} = (\mathbf{X}_i - \boldsymbol{x}_0)_{i=1,\ldots,n}$ and assume $(\mathbf{Cond_X})$ 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 (6.1.6):

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

$$f = \mathbb{E}[g(\mathbf{X})|\mathbf{X}^\top \boldsymbol{\theta}^* = \cdot] = \boldsymbol{f}_{\boldsymbol{\eta}^*} = \sum_{k=1}^\infty \eta_k^* \boldsymbol{e}_k, \tag{6.2.1}$$

where $\|\boldsymbol{f}'_{\boldsymbol{\eta}^*}\|_\infty = \mathtt{C}_{\|\boldsymbol{f}'_{\boldsymbol{\eta}^*}\|_\infty} < \infty$ and $\|\boldsymbol{f}''_{\boldsymbol{\eta}^*}\|_\infty = \mathtt{C}_{\|\boldsymbol{f}''_{\boldsymbol{\eta}^*}\|_\infty} < \infty$ and where with some $\alpha > 2$ - or $\alpha > 9/2$ if $\mathtt{C}_{bias} > 0$ see Remark 6.2.15 - and a constant $C_{\|\boldsymbol{\eta}^*\|} > 0$

$$\sum_{k=0}^\infty k^{2\alpha} \eta_k^{*2} \le C_{\|\boldsymbol{\eta}^*\|}^2 < \infty. \tag{6.2.2}$$

Remark 6.2.3. We can now specify the parameter set $\varUpsilon \subset \mathbb{R}^p \times l^2$ namely

$$\varUpsilon \stackrel{\text{def}}{=} \left\{(\boldsymbol{\theta}, \boldsymbol{\eta}) \in \mathbb{R}^p \times l^2,\, \boldsymbol{\theta} \in S_1^{p,+}\right\}.$$

Remark 6.2.4. Simply using (6.2.2) does not - easily - yield a bound for $\|\boldsymbol{f}''_{\boldsymbol{\eta}^*}\|_\infty$ since (see proof of Lemma 6.A.18)

$$|\boldsymbol{f}''_{\boldsymbol{\eta}^*}(\mathbf{X}^\top\boldsymbol{\theta})| \le \sqrt{34}\|\psi''\|_\infty \left(\sum_{k=0}^\infty k^{2\alpha}\eta_k^{*2}\right)^{1/2} \left(\sum_{j=0}^\infty 2^{5j-2\alpha}\right)^{1/2} = \infty.$$

Remark 6.2.5. In the case that the data is not from the model (6.1.4) but from the model in (6.1.1) 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

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

satisfies (6.2.1) with some $\boldsymbol{\eta}(\boldsymbol{\theta}, \boldsymbol{x})$ and $\alpha(\boldsymbol{\theta}, \boldsymbol{x}) > 9/2 + \epsilon$, where $\epsilon > 0$ is independent of $\boldsymbol{x}$. More precisely set for any $\boldsymbol{\theta} \in S_1^{p,+}$

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

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 (6.2.1) as well with some $\boldsymbol{\eta}(\boldsymbol{\theta})$ and $\alpha(\boldsymbol{\theta}) \ge \inf_{\boldsymbol{x}\in B_{s_{\mathbf{X}}}(0)\cap\boldsymbol{\theta}^\perp}\{\alpha(\boldsymbol{\theta}, \boldsymbol{x})\} > 9/2$. We proof this in Section 6.A.

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

($\mathbf{Cond}_{\mathbf{X}\boldsymbol{\theta}^*}$) On some ball $B_h(\boldsymbol{x}_0) \subseteq B_{s_{\mathbf{X}}}(0)$ with $h > 0$ it holds true that $|\boldsymbol{f}'_{\boldsymbol{\eta}^*}(\mathbf{X}^\top \boldsymbol{\theta}^*)| > c_{\boldsymbol{f}'_{\boldsymbol{\eta}^*}}$ for some $c_{\boldsymbol{f}'_{\boldsymbol{\eta}^*}} > 0$.

Remark 6.2.6. Note that 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 (6.1.3) is not defined.

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

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

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

Remark 6.2.7. Note that our assumptions in terms of moments and smoothness are quite common in this model. For instance [23] assume that the density $p_{\mathbf{X}}$ of the regressors $(\mathbf{X}_i)$ is twice continuously differentiable, that $\mathbb{E}[y|\mathbf{X}^\top\boldsymbol{\theta}^* = \cdot]$ has two bounded derivatives and that the errors (ε_i) are centered with bounded polynomial moments of arbitrary degree. In [30] even three derivatives of $\mathbb{E}[y|\mathbf{X}^\top\boldsymbol{\theta}^* = \cdot]$ are assumed.

Unfortunately these conditions do not facilitate an easy proof of our desired results in the case that $\mathtt{C}_{bias} > 0$. 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

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

for some constant $\mathtt{C}_{bias} \ge 0$. 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 $\boldsymbol{\theta} \notin B(\boldsymbol{\theta}^*)$

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

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

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

Remark 6.2.8. 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 chapter - 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 (6.1.3) is.

6.2.2 Some important objects

In this subsection we introduce some important objects that are relevant for our results.

For given $p^* = p + m$, set $\Pi_{p^*}\boldsymbol{\upsilon} = (\upsilon_1, \ldots, \upsilon_{p^*}) = (\boldsymbol{\theta}, \Pi_m\boldsymbol{\eta}) \in \mathbb{R}^{p^*}$. We represent the full parameter $\boldsymbol{\upsilon} \in \mathbb{R}^\infty$ in the form

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

where $\boldsymbol{\varkappa} = (\eta_{m+1}, \ldots)^\top$ stands for the remaining components of the expansion (6.2.1). We repeat the definitions of the sieve estimator $\widetilde{\boldsymbol{\upsilon}}_m$, its possibly biased target $\boldsymbol{\upsilon}^*_m$ and the full oracle $\boldsymbol{\upsilon}^* \in \Upsilon \subset l^2$

$$\begin{aligned}
\widetilde{\boldsymbol{\upsilon}}_m &= \operatorname*{argmax}_{\boldsymbol{\upsilon}\in\Upsilon_m} \mathcal{L}_m(\boldsymbol{\upsilon}), \\
\boldsymbol{\upsilon}^*_m &= (\boldsymbol{\theta}^*_m, \boldsymbol{\eta}^*_m) = \operatorname*{argmax}_{\boldsymbol{\upsilon}\in\Upsilon^*_m} I\!E[\mathcal{L}_m(\boldsymbol{\upsilon})], \\
\boldsymbol{\upsilon}^* &= (\Pi_{p^*}\boldsymbol{\upsilon}^*, \boldsymbol{\varkappa}^*) = \operatorname*{argmax}_{\boldsymbol{\upsilon}\in\Upsilon\subset l^2} I\!E[\mathcal{L}(\boldsymbol{\upsilon})],
\end{aligned}$$

where $\mathcal{L}(\cdot)$ is the functional in (6.1.8) for $m = \infty$. We set

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

with some $\mathtt{r}^\circ > \infty$ defined in Lemma 6.3.6.

Remark 6.2.9. 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 or 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 $\Upsilon_\circ$ is ensured via Theorem 3.3.2.

Remark 6.2.10. Note that we maximize over different sets. To control the large deviations and avoid boundary effects we have to ensure that with overwhelming probability $\widetilde{\boldsymbol{\upsilon}}_m \subset \operatorname{int}\{\Upsilon_m\} \subset \Upsilon^*_m$. We do this with Lemma 6.3.6, which tells us that we may set $\mathtt{r}^\circ \le \mathtt{C}\sqrt{m}$ 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 6.2.4 lies in $S_1^{p,+} \times B^m_{\mathtt{r}^\circ}(0)$.

We define the *information operator* $\mathcal{D}^2$ similarly to the Fisher information matrix as the Hessian operator of the expected value of the functional:

$$\mathcal{D}^2(\boldsymbol{\upsilon}) \overset{\text{def}}{=} -\nabla^2 I\!\!E\mathcal{L}(\boldsymbol{\upsilon}) = -\nabla^2 I\!\!E\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{f}).$$

Consider the following block representations of of the *information operator*:

$$\mathcal{D}^2 = n\begin{pmatrix} D^2 & A_{\boldsymbol{\theta\eta}} \\ A_{\boldsymbol{\theta\eta}}^\top & \mathcal{H}^2 \end{pmatrix} = \begin{pmatrix} \mathcal{D}_m^2 & A_{\boldsymbol{\upsilon\varkappa}} \\ A_{\boldsymbol{\upsilon\varkappa}}^\top & \mathcal{H}_{\varkappa\varkappa}^2 \end{pmatrix} = n\begin{pmatrix} D^2 & A_m & A_{\boldsymbol{\theta\varkappa}} \\ A_m^\top & \mathrm{H}_m^2 & A_{\boldsymbol{\eta\varkappa}} \\ A_{\boldsymbol{\varkappa\theta}} & A_{\boldsymbol{\varkappa\eta}} & \mathcal{H}_{\varkappa\varkappa}^2 \end{pmatrix}.$$

where $A_{\boldsymbol{\upsilon\varkappa}}$ is a - possibly unbounded - operator from l^2 to $\mathbb{R}^{p+m}$. Define $c_{\mathcal{D}} \overset{\text{def}}{=} \lambda_{\min}(\mathcal{D}_m(\boldsymbol{\upsilon}_m^*))/\sqrt{n}$, where $\lambda_{\min}(\mathcal{D}_m) \in \mathbb{R}$ denotes the smallest eigenvalue of $\mathcal{D}_m \in \mathbb{R}^{p^* \times p^*}$. In Lemma 6.A.8 we derive that $c_{\mathcal{D}} > 0$. Furthermore we introduce the influence matrix and the score

$$\breve{D}_m^{-2} = \varPi_{\boldsymbol{\theta}}\mathcal{D}_m^{-2}\varPi_{\boldsymbol{\theta}}^\top, \quad \breve{\boldsymbol{\xi}}_m = \nabla_{\boldsymbol{\theta}}\zeta(\boldsymbol{\upsilon}_m^*) - A_m\mathrm{H}_m^{-2}\nabla_{\boldsymbol{\eta}}\zeta(\boldsymbol{\upsilon}_m^*), \quad \boldsymbol{\zeta} = \mathcal{L} - I\!\!E_\varepsilon\mathcal{L},$$

where $I\!\!E_\varepsilon$ denotes the expectation operator of the law of $(\varepsilon_i)_{i=1,\ldots,n}$ given $(\mathbf{X}_i)_{i=1,\ldots,n}$.

6.2.3 Properties of the Wavelet Sieve profile M-estimator

This section presents the application of the results of Chapter 4 to the estimator $\widetilde{\boldsymbol{\theta}}_m$ in (6.1.7). 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 chapter 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. Thus whenever there appears a constant $\mathtt{C} > 0$ without further remarks it is a polynomial of $\|\psi\|_\infty, \|\psi'\|_\infty, \|\psi''\|_\infty, \mathtt{C}_{\|\boldsymbol{f}^*\|}, s_{\mathbf{X}}, \frac{1}{c_{p_{\mathbf{X}}}}$, etc. where $\psi : \mathbb{R} \to \mathbb{R}$ is introduced in Section 6.3.2.. Also - in the proofs - 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.

Define

$$\breve{\diamondsuit}(\mathtt{x}) = \mathtt{C}_{\diamond}\frac{p^{*5/2} + \mathtt{C}_{bias}p^{*7/2} + \mathtt{x}}{\sqrt{n}},$$

where $\mathtt{C}_{\diamondsuit} > 0$ is a polynomial of $\|\psi\|_\infty, \|\psi'\|_\infty, \|\psi''\|_\infty, \mathtt{C}_{\|\boldsymbol{f}^*\|}, s_{\mathbf{X}}$, etc.. We get the following result by applying Theorem 4.2.2

Proposition 6.2.1. *Assume* $(\mathcal{A})$. *If* $\mathtt{C}_{bias} = 0$ *suppose that* $m^{-(2\alpha+1)}n \to 0$ *and that* $p^{*4}/n \to 0$. *If* $\mathtt{C}_{bias} > 0$ *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* $1 - 12\mathrm{e}^{-\mathtt{x}} - \exp\left\{-m^3\mathtt{x}\right\} - \exp\left\{-nc_{(\boldsymbol{Q})}/4\right\}$

$$\left|2\max_{\boldsymbol{\eta}}\mathcal{L}(\widetilde{\boldsymbol{\theta}}_m,\boldsymbol{\eta}) - \max_{\boldsymbol{\eta}}\mathcal{L}(\boldsymbol{\theta}_m^*,\boldsymbol{\eta}) - \|\breve{\boldsymbol{\xi}}_m\|^2\right| \le \mathtt{C}\left(\sigma\sqrt{p+\mathtt{x}} + \breve{\diamondsuit}(\mathtt{x})\right)\breve{\diamondsuit}(\mathtt{x}),$$

$$\left\|\breve{D}_m(\boldsymbol{\upsilon}_m^*)\bigl(\widetilde{\boldsymbol{\theta}}_m - \boldsymbol{\theta}_m^*\bigr) - \breve{\boldsymbol{\xi}}_m\right\| \le \breve{\diamondsuit}(\mathtt{x}).$$

where $c_{(\boldsymbol{Q})}, \mathtt{C} > 0$ *.*

Remark 6.2.11. The constant $c_{(\boldsymbol{Q})} > 0$ is derived in the proof of Lemma 6.A.20 and does not depend on $\mathtt{x}$, n, p^* .

Remark 6.2.12. The necessary size of $n \in \mathbb{N}$ is determined by the speed with which $p^{*4}/n \to 0$ and $m^{-2\alpha-1}n \to 0$ or $p^{*6}\log(n)/n \to 0$ and $m^{-2(\alpha-1)}n \to 0$ respectively in the cases $\mathtt{C}_{bias} = 0$ or $\mathtt{C}_{bias} > 0$ respectively In the proof of Proposition 6.2.1 we impose conditions on $n \in \mathbb{N}$ of the kind

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

for certain constants $\mathtt{C}_1, \mathtt{C}_2 > 0$ that are polynomials of $\|\psi\|_\infty$, $\|\psi'\|_\infty$, $\|\psi''\|_\infty$, $\mathtt{C}_{\|\boldsymbol{f}^*\|}$, $L_{\nabla\varPhi}$, $s_{\mathbf{X}}$.

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 that depends on the dimension $m \in \mathbb{N}$ of the nuisance parameter $\boldsymbol{\eta} \in \mathbb{R}^m$. The next result will specify the finite sample properties of $\breve{D}\bigl(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\bigr) \in \mathbb{R}^p$ where

$$\breve{D}^{-2} = \varPi_{\boldsymbol{\theta}}\mathcal{D}^{-2}(\boldsymbol{\upsilon}^*)\varPi_{\boldsymbol{\theta}}^\top \in \mathbb{R}^{p\times p}.$$

We get the following result.

Proposition 6.2.2. *Assume* $(\mathcal{A})$ *. If* $\mathtt{C}_{bias} = 0$ *suppose that* $m^{-(2\alpha+1)}n \to 0$ *and that* $p^{*4}/n \to 0$ *. If* $\mathtt{C}_{bias} > 0$ *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* $1 - 12\mathrm{e}^{-\mathtt{x}} - \exp\left\{-m^3\mathtt{x}\right\} - \exp\left\{-nc_{(\boldsymbol{Q})}/4\right\}$

$$\left\|\breve{D}_m(\boldsymbol{\upsilon}_m^*)\bigl(\widetilde{\boldsymbol{\theta}}_m - \boldsymbol{\theta}^*\bigr) - \breve{\boldsymbol{\xi}}_m(\boldsymbol{\upsilon}_m^*)\right\| \le \breve{\diamondsuit}(\mathtt{x}) + \alpha(m),$$

and

$$\begin{aligned}&\left|\max_{\boldsymbol{\eta}\in\varPi_m\varUpsilon_{\boldsymbol{\eta}}}\mathcal{L}(\widetilde{\boldsymbol{\theta}}_m,\boldsymbol{\eta}) - \max_{\boldsymbol{\eta}\in\varPi_m\varUpsilon_{\boldsymbol{\eta}}}\mathcal{L}(\boldsymbol{\theta}^*,\boldsymbol{\eta}) - \|\breve{\boldsymbol{\xi}}_m\|^2/2\right|\\ &\quad\le \mathtt{C}\left(\sigma\sqrt{p+\mathtt{x}} + \breve{\diamondsuit}(\mathtt{x})\right)\breve{\diamondsuit}(\mathtt{x}) + \alpha(m)(\mathtt{C} + \sqrt{p+\mathtt{x}}),\end{aligned}$$

where

$$\alpha(m) \le \mathtt{C}\sqrt{n}\left(m^{-(\alpha+1/2)} + \mathtt{C}_{bias}m^{-(\alpha-1)}\right),$$

and

$$\mathtt{r}_p^* \overset{\mathrm{def}}{=} \left(2 + \frac{1+\nu}{1-\nu}\right)\breve{\diamondsuit}(\mathtt{x}) + \alpha(m).$$

Further if $\mathtt{C}_{bias} = 0$ *and* $p^{*5/2}/\sqrt{n} \to 0$ *we find as* $n \to \infty$

$$\breve{D}\big(\widetilde{\boldsymbol{\theta}}_m - \boldsymbol{\theta}^*\big) \xrightarrow{w} \mathcal{N}(0, \sigma^2 I_p),$$

$$2\max_{\boldsymbol{\eta}\in\varPi_m\varUpsilon_{\boldsymbol{\eta}}} \mathcal{L}(\widetilde{\boldsymbol{\theta}}_m, \boldsymbol{\eta}) - 2\max_{\boldsymbol{\eta}\in\varPi_m\varUpsilon_{\boldsymbol{\eta}}} \mathcal{L}(\boldsymbol{\theta}^*, \boldsymbol{\eta}) \xrightarrow{w} \chi_p^2. \tag{6.2.3}$$

Remark 6.2.13. 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 [42]. 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 6.2.14. Assume that the model (6.1.5) is correct. We will see in Section 6.A.2 that then

$$\sigma^2\mathcal{D}^2(\boldsymbol{\upsilon}^*) = \mathrm{Cov}(\nabla\mathcal{L}(\boldsymbol{\upsilon}^*)), \tag{6.2.4}$$

Such that with (6.2.3)

$$\sqrt{n}\big(\widetilde{\boldsymbol{\theta}}_m - \boldsymbol{\theta}^*\big) \xrightarrow{w} \mathcal{N}(0, \sigma^2\breve{D}^{-2}), \quad \sigma^2\breve{D}^{-2} = \varPi_{\boldsymbol{\theta}}^\top \mathrm{Cov}(\nabla\mathcal{L}(\boldsymbol{\upsilon}^*))\varPi_{\boldsymbol{\theta}}.$$

As we showed in Section 2.1.1 this 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 $\mathcal{B}_{s_{\mathbf{X}}} \subset \mathbb{R}^p$.

Remark 6.2.15. Note that we do not show any weak convergence statements for the case that $\mathtt{C}_{bias} > 0$. The approach of Section 4.3.3 is not applicable - at least not with the arguments we use for the case $\mathtt{C}_{bias} = 0$ in Lemma 6.A.6. Also note that to control the approximation bias and the size of $\breve{\diamondsuit}(\mathtt{x})$ when $\mathtt{C}_{bias} > 0$ the necessary smoothness of $I\!E[g(\mathbf{X})|\mathbf{X}^\top\boldsymbol{\theta}^* = \cdot] = \boldsymbol{f}_{\boldsymbol{\eta}^*}(\cdot) : \mathbb{R} \to \mathbb{R}$ measured in $\alpha > 0$ in (6.2.2) increases from $\alpha > 2$ to $\alpha > 9/2$ to ensure that $\alpha(m) \to 0$.

6.2.4 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. For this note that the maximization problem

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

is not convex and thus computationally involved. We propose to obtain the maximizer via the alternation maximization procedure from Chapter 5. To remind the reader this sequential algorithm works as follows: Start with some initial guess $\widetilde{\boldsymbol{\upsilon}}^{(0)} \in \varUpsilon$. Then calculate for $k \in \mathbb{N}$ iteratively

$$\begin{aligned}
\widetilde{\boldsymbol{\upsilon}}^{(k,k+1)} &\overset{\text{def}}{=} (\widetilde{\boldsymbol{\theta}}^{(k)}, \widetilde{\boldsymbol{\eta}}^{(k+1)}) = \left(\widetilde{\boldsymbol{\theta}}^{(k)}, \operatorname*{argmax}_{\boldsymbol{\eta}} \mathcal{L}_m(\widetilde{\boldsymbol{\theta}}^{(k)}, \boldsymbol{\eta}) \right), \\
\widetilde{\boldsymbol{\upsilon}}^{(k,k)} &\overset{\text{def}}{=} (\widetilde{\boldsymbol{\theta}}^{(k)}, \widetilde{\boldsymbol{\eta}}_k) = \left(\operatorname*{argmax}_{\boldsymbol{\theta}} \mathcal{L}_m(\boldsymbol{\theta}, \widetilde{\boldsymbol{\eta}}^{(k)}), \widetilde{\boldsymbol{\eta}}^{(k)} \right).
\end{aligned}$$

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 $G_N \overset{\text{def}}{=} (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_N) \subset S_1^+$ and define

$$\widetilde{\boldsymbol{\upsilon}}^{(0)} \overset{\text{def}}{=} \operatorname*{argmax}_{\substack{(\boldsymbol{\theta},\boldsymbol{\eta}) \in \varUpsilon \\ \boldsymbol{\theta} \in G_N}} \mathcal{L}_m(\boldsymbol{\upsilon}). \tag{6.2.5}$$

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

$$\begin{aligned}
\widetilde{\boldsymbol{\eta}}_l^{(0)} &\overset{\text{def}}{=} \operatorname{argmax} \mathcal{L}(\boldsymbol{\theta}_l, \boldsymbol{\eta}) \qquad (6.2.6) \\
&= \left(\frac{1}{n} \sum_{i=1}^{n} \boldsymbol{e}\boldsymbol{e}^\top (\mathbf{X}_i^\top \boldsymbol{\theta}_l) \right)^{-1} \frac{1}{n} \sum_{i=1}^{n} Y_i \boldsymbol{e}^\top (\mathbf{X}_i^\top \boldsymbol{\theta}_l) \in \mathbb{R}^m,
\end{aligned}$$

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

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

Define the fineness of the grid via $\tau \overset{\text{def}}{=} \sup_{\boldsymbol{\theta},\boldsymbol{\theta}^\circ \in G_N} \|\boldsymbol{\theta} - \boldsymbol{\theta}^\circ\|$. To asses the statistical properties of the alternating procedure we can derive the following result via an application of Theorem 5.2.1.

Proposition 6.2.3. *If* $\mathtt{C}_{bias} = 0$ *set* $\tau = o(p^{*-3/2})$ *and* $m^4 = o(n)$ *and assume that* $m^{-(2\alpha+1)} n \to 0$. *If* $\mathtt{C}_{bias} > 0$ *set* $\tau = o(m^{-11/4})$ *and*

$m^6\log(n)/n \to 0$ and assume that $m^{-2(\alpha-1)}n \to 0$. Furthermore let $\mathtt{x} \le 2\widetilde{\nu}^2\widetilde{\mathtt{g}}^2(1+\mathtt{C}_{bias})n$. With the initial guess given by Equation (6.2.5) the alternating sequence satisfies with probability greater than $1-12\mathrm{e}^{-\mathtt{x}}-\exp\left\{-m^3\mathtt{x}\right\} - \exp\left\{-nc_{(\boldsymbol{Q})}/4\right\}$

$$\left\|\breve{D}_m(\boldsymbol{\upsilon}_m^*)\big(\widetilde{\boldsymbol{\theta}}^{(k)}-\boldsymbol{\theta}^*\big)-\breve{\boldsymbol{\xi}}\right\| \le \breve{\diamondsuit}_Q(\mathtt{r}_k,\mathtt{x}), \tag{6.2.7}$$

and

$$\begin{aligned} &2\big|\max_{\boldsymbol{\eta}}\mathcal{L}(\widetilde{\boldsymbol{\theta}}^{(k)},\boldsymbol{\eta})-\max_{\boldsymbol{\eta}}\mathcal{L}(\boldsymbol{\theta}^*,\boldsymbol{\eta})-\|\breve{\boldsymbol{\xi}}\|^2/2\big| \\ &\quad\le 9\left(\|\breve{\boldsymbol{\xi}}\|+\breve{\diamondsuit}_Q(\mathtt{r}_k,\mathtt{x})\right)\breve{\diamondsuit}(\mathtt{r}_k,\mathtt{x}), \end{aligned} \tag{6.2.8}$$

where

$$\breve{\diamondsuit}_Q(\mathtt{r},\mathtt{x}) \le \mathtt{C}_{\diamondsuit}\left(\frac{\left(p^{*3/2}+\mathtt{x}+\mathtt{C}_{bias}p^{*5/2}\right)\mathtt{r}^2}{\sqrt{n}}\right),$$

and where with some constant $\mathtt{C}$

$$\begin{aligned} \mathtt{r}_k &\le (1-\sqrt{\nu})^{-1}\mathtt{C}\left(\sqrt{\mathtt{x}+p^*}+2\nu^k R_0(\mathtt{x})\right), \\ R_0(\mathtt{x}) &\le \mathtt{C}\sqrt{p^*(1+\mathtt{C}_{bias}\log(n))+\mathtt{x}+(1+\mathtt{C}_{bias}\sqrt{m})n\tau^2+\sqrt{n}\tau\sqrt{\mathtt{x}}}. \end{aligned}$$

Remark 6.2.16. The constraint $\tau = o(p^{*-3/2})$ implies that for the calculation of the initial guess the vector $\widetilde{\boldsymbol{\eta}}_{(l)}^{(0)}$ in (6.2.6) 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 (6.2.8) and (6.2.7) 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 case that $\mathtt{C}_{bias} > 0$ we need 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.

Define the local neighborhood around the ME $\widetilde{\boldsymbol{\upsilon}}$ (we suppress $\cdot_m$ here)

$$\widetilde{\varUpsilon}_{\circ}(\mathtt{r}) \stackrel{\text{def}}{=} \{\boldsymbol{\upsilon}\in\varUpsilon:\ \|\mathcal{D}(\boldsymbol{\upsilon}-\widetilde{\boldsymbol{\upsilon}})\| \le \mathtt{r}\}.$$

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 5.2.3.

Proposition 6.2.4. *Take the initial guess given by Equation* (6.2.5)*. Assume* $(\mathcal{A})$ *. If* $\mathtt{C}_{bias} = 0$ *set* $\tau = o(m^{-3/2})$ *and* $m^4 = o(n)$ *and assume that* $m^{-(2\alpha+1)}n \to 0$ *. If* $\mathtt{C}_{bias} > 0$ *set* $\tau = o(m^{-11/4})$ *and* $m^6 \log(n) = o(n)$ *and assume that* $m^{-2(\alpha-1)}n \to 0$ *. Let* $\mathtt{x} > 0$ *be chosen such that*

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

Then

$$I\!\!P\left(\bigcap_{k\in\mathbb{N}}\left\{\widetilde{\boldsymbol{\upsilon}}^{(k,k(+1))} \in \widetilde{\varUpsilon_{\circ}}(\mathtt{r}_k^*)\right\}\right) \ge 1 - 10\mathrm{e}^{-\mathtt{x}} - \exp\left\{-m^3\mathtt{x}\right\}$$
$$- \exp\left\{-nc_{(\boldsymbol{Q})}/4\right\},$$

where with $\varkappa(\mathtt{x}, R_0) = O({p^*}^2/\sqrt{n} + \mathtt{C}_{bias}{p^*}^3/\sqrt{n}) \to 0$

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

with

$$R_0 \le \mathtt{C}\sqrt{p^*(1 + \mathtt{C}_{bias}\log(n)) + \mathtt{x} + (1 + \mathtt{C}_{bias}\sqrt{m})n\tau^2 + \sqrt{n}\tau\sqrt{\mathtt{x}}}.$$

Remark 6.2.17. Note that in the case $\mathtt{C}_{bias} = 0$ the constraint on the size of the dimension $p^* \in \mathbb{N}$ for accurate results is weaker in Proposition 6.2.4 than in Proposition 6.2.3, as there are no "right-hand sides" and thus $m^4 = o(n)$ is sufficient.

6.2.5 Performance of Projection Pursuit Procedure

In this section we want to briefly assess the performance of the Projection Pursuit procedure of [20]. 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. In the previous sections we already established that for observations of the kind

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

the estimator in (6.1.7) satisfies

$$\left|I\!\!E[Y|\mathbf{X}^\top{\boldsymbol{\theta}^*}_{(1)}] - \boldsymbol{f}_{\widetilde{\boldsymbol{\eta}}_{(1)}}(\mathbf{X}^\top\widetilde{\boldsymbol{\theta}}_{(1)})\right| \tag{6.2.9}$$
$$\le \mathtt{C}\left(\mathtt{r}^* + \alpha(m) + \diamondsuit(\mathtt{x}) + \|{\mathcal{D}_{(1)}}^{-1}\nabla\mathcal{L}_{(1)}(\boldsymbol{\upsilon}^*)\|\right)/\sqrt{n},$$

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

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

where

$$\begin{aligned} g_{(l)}(\mathbf{X}_i) &\approx \sum_{s=l}^{M} \boldsymbol{f}_{\boldsymbol{\eta}^*{}_{(s)}}(\mathbf{X}_i^\top \boldsymbol{\theta}^*{}_{(s)}), \\ \tau_{i(l)} &= \sum_{s=1}^{l} \boldsymbol{f}_{\boldsymbol{\eta}^*{}_{(s)}}(\mathbf{X}_i^\top \boldsymbol{\theta}^*{}_{(s)}) - \boldsymbol{f}_{\widetilde{\boldsymbol{\eta}}_{(s)}}(\mathbf{X}_i^\top \widetilde{\boldsymbol{\theta}}_{(s)}). \end{aligned}$$

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 4.2.2 is to bound with probability $1 - \mathrm{e}^{-\mathtt{x}}$

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

and to show that $I\!P(\widetilde{\boldsymbol{\upsilon}}, \widetilde{\boldsymbol{\upsilon}}_{\boldsymbol{\theta}^*} \in \varUpsilon_\circ(\mathtt{r}_0)) \ge 1 - \mathrm{e}^{-\mathtt{x}}$. Consequently we decompose (we suppress $\cdot_m$ to ease notation)

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

and define

$$\begin{aligned} \boldsymbol{\upsilon}_{m(l)}^* &\stackrel{\text{def}}{=} \operatorname*{argmax}_{\boldsymbol{\upsilon} \in \varUpsilon_m} I\!E \mathcal{L}_{\varepsilon(l)}(\boldsymbol{\upsilon}), \\ \mathcal{D}_{m(l)}{}^2 &\stackrel{\text{def}}{=} \nabla^2 I\!E[\mathcal{L}_{\varepsilon(l)}(\boldsymbol{\upsilon}_{m(l)}^*)], \\ \boldsymbol{\zeta}_{\varepsilon(l)}(\boldsymbol{\upsilon}) &\stackrel{\text{def}}{=} \mathcal{L}_{\varepsilon(l)}(\boldsymbol{\upsilon}) - I\!E \mathcal{L}_{\varepsilon(l)}(\boldsymbol{\upsilon}). \end{aligned}$$

We assume that the condition (model bias) holds for every function $g_{(l)}$. With Remark 6.2.5, Lemma 6.3.7 and Lemma 6.A.6 this means that the conditions of Section 4.2.1 and 4.3.2 are met for $(\mathcal{L}_{\varepsilon(l)}, \varUpsilon_m, \mathcal{D}_{m(l)})$ with high probability for every $l = 1, \dots, M$. 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 following Proposition.

Proposition 6.2.5. *Assume that* $M = O(p^*)$ *and that the conditions* $(\mathcal{A})$ *hold for every* $l = 1 \dots, M$. *Assume further* $\frac{p^{*3} \log(n) M + \mathtt{x}}{\sqrt{n}} \to 0$ *and assume that* $m^{-2(\alpha-1)} n \to 0$. *With probability greater than*

$$1 - \mathrm{e}^{-\mathtt{x}} - M \left(12\mathrm{e}^{-\mathtt{x}} + \exp\left\{ -m^3 \mathtt{x} \right\} + \exp\left\{ -n c_{(\boldsymbol{Q})}/4 \right\} \right),$$

we have

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

Remark 6.2.18. Denoting the bias

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

Proposition 6.2.5 implies if $\mathtt{x} \leq p^* = o(n^{1/6})$ that

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

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

6.3 Details

In this section we lay out how to apply the results of the Chapters 4 and 5. First we will explain the implications of the regression setup with random design and explain which type of Daubechies wavelets can be used. Then we show that the conditions $(\mathcal{ED}_0)$, $(\mathcal{ED}_1)$, $(\mathcal{L}_0)$ and $(\mathcal{I})$ of Section 4.2.1 can be satisfied under the assumptions $(\mathcal{A})$. These imply - by Lemma 4.2.1 - $(\breve{\mathcal{L}}_0)$, $(\breve{\mathcal{E}}\mathcal{D}_1)$ and $(\breve{\mathcal{E}}\mathcal{D}_0)$ from Section 4.2.1, necessary for Theorem 4.2.2. Furthermore we will show that the conditions $(\mathcal{E}\mathtt{r})$ and $(\mathcal{L}\mathtt{r})$ from 4.2.1 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 in (6.2.5) and the values of $\delta(\mathtt{r}), \omega$ from $(\breve{\mathcal{L}}_0)$, $(\breve{\mathcal{E}}\mathcal{D}_1)$ allow to apply the Theorems 5.2.1 and 5.2.3.

6.3.1 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

$$\boldsymbol{\upsilon}_m^* \stackrel{\mathrm{def}}{=} \operatorname*{argmax}_{\boldsymbol{\upsilon} \in \varUpsilon_m} I\!\!E \mathcal{L}_m(\boldsymbol{\upsilon}),$$

where $I\!\!E$ 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 $I\!\!E$. But in Lemma 6.3.7 we show the conditions $(\mathcal{E}\mathcal{D}_0)$, $(\mathcal{E}\mathtt{r})$ and $(\mathcal{E}\mathcal{D}_1)$ for the random variables

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

i.e. we use only the expectation with respect to the noise (ε_i). This leads to rather weak conditions on the errors (ε_i) but the statements are in the sense that the conditions are met with high probability with respect to the distribution of the $(\mathbf{X}_i)$. 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 becomes necessary to apply the theory of the previous chapters. As becomes evident from the proof of Theorem 4.2.2, we have to bound the term

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

and add the obtained bound to $\breve{\diamondsuit}(\mathtt{r}_0, \mathtt{x})$ on the right-hand side in (4.2.9) and (4.2.10). Also the probability of the desired bound has to be subtracted from the probability of the event that (4.2.9) and (4.2.10) are valid in Theorem 4.2.2. The following lemma serves this bound.

Lemma 6.3.1. *With some constant* $\mathtt{C} > 0$

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

Remark 6.3.1. 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 $\diamondsuit(\mathtt{r})$ in the subsequent analysis. Consequently we neglect it in the following and let a constant $\mathtt{C}_{\diamondsuit} > 0$ account for its contribution in the formulation of Proposition 6.2.1.

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

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

This leads to concentration bounds for sums of i.i.d. matrices which can be handled with the results of [56]. We do this in Section 6.A.8. Again the set on which Theorem 4.2.2 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 6.2.1 the probability of the desired results can only be bounded from below 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 4.2.3.

To further address the peculiarities of the regression setting we present the following adapted versions of Theorem 4.2.2 and Proposition 4.2.3, which are proved in exactly the same way.

Theorem 6.3.2. *Assume* $(\breve{\mathcal{L}}_0)$ *and* $(\mathcal{I})$. *Also assume that on some set* $\boldsymbol{\mathcal{N}}(\mathtt{x}) \subset \Omega$ *the condition* $(\breve{\mathcal{E}}\mathcal{D}_1)$ *is met and that on* $\boldsymbol{\mathcal{N}}(\mathtt{x})$ *the sets of maximizers* $\widetilde{\boldsymbol{\upsilon}}, \widetilde{\boldsymbol{\upsilon}}_{\boldsymbol{\theta}^*}$ *are not empty. Also assume that* $4\mathtt{r}_0 \le \mathtt{r}^*$. *Assume that* $\boldsymbol{\mathcal{N}}(\mathtt{x})$ *contains with some* $\tau(\cdot) \in \mathbb{R}$ *the set*

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

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

$$\|\breve{D}(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) - \breve{\boldsymbol{\xi}}\| \le \breve{\diamondsuit}(\mathtt{r}_0, \mathtt{x}) + \tau(\mathtt{r}_0),$$

and

$$\begin{aligned} &2\big|\max_{\boldsymbol{\eta}} \mathcal{L}(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\eta}) - \max_{\boldsymbol{\eta}} \mathcal{L}(\boldsymbol{\theta}^*, \boldsymbol{\eta}) - \|\breve{\boldsymbol{\xi}}\|^2/2\big| \\ &\qquad \le 9\left(\|\breve{\boldsymbol{\xi}}\| + \breve{\diamondsuit}(\mathtt{r}_0, \mathtt{x}) + \tau(\mathtt{r}_0)\right)\left(\breve{\diamondsuit}(\mathtt{r}_0, \mathtt{x}) + \tau(\mathtt{r}_0)\right), \end{aligned}$$

where the spread $\breve{\diamondsuit}(\mathtt{r}_0, \mathtt{x})$ *is defined in* (4.2.7) *and where* $\mathtt{r}_0 > 0$ *is defined in* (4.2.3).

Proposition 6.3.3. *Assume that the conditions of Theorem 4.2.2 are satisfied and additionally that on* $\boldsymbol{\mathcal{N}}(\mathtt{x})$ *the conditions* $(\mathcal{L}_0)$ *and that* $(\mathcal{E}\mathcal{D}_1)$ *and* $(\mathcal{E}\mathcal{D}_0)$ *are met. Also assume that* $4\mathtt{r}_0 \le \mathtt{r}^*$ *. Then the results of Theorem 4.2.2 hold with* $\mathtt{r}_1 \le \mathtt{r}_0$ *instead of* $\mathtt{r}_0$ *and with probability greater than* $1 - 4\mathrm{e}^{-\mathtt{x}} - I\!\!P(\boldsymbol{\mathcal{N}}(\mathtt{x})^c)$ *where*

$$\mathtt{r}_1 \le \mathfrak{z}(\mathtt{x}, I\!\!B) + \diamondsuit_Q(R_0, \mathtt{x}) \wedge \mathtt{r}_0(\mathtt{x}).$$

Furthermore if there is some $\epsilon > 0$ *such that* $\delta(\mathtt{r})/\mathtt{r} \vee 6\nu_1\omega \le \epsilon$ *for all* $\mathtt{r} \le \mathtt{r}_0$ *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 replaced with* $\mathtt{r}_0^*$ *which is bounded by*

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

Furthermore we present an adapted version of Theorem 5.2.1.

Theorem 6.3.4. *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 4.2.1 are met with a constant* $\mathtt{b}(\mathtt{r}) \equiv \mathtt{b}$ *and where* $\mathcal{V}_0^2 = \mathrm{Cov}\left(\nabla\mathcal{L}(\boldsymbol{\upsilon}^*)\right)$, $\mathcal{D}_0^2 = -\nabla^2 I\!\!E\mathcal{L}(\boldsymbol{\upsilon}^*)$ *and* $\boldsymbol{\upsilon}^\circ = \boldsymbol{\upsilon}^*$ *. Also assume that* $R_0 \vee 4\mathtt{r}_0 \le \mathtt{r}^*$ *. Assume further that on* $\boldsymbol{\mathcal{N}}(\mathtt{x})$ *the sets* $(\widetilde{\boldsymbol{\upsilon}}^{(k,k(+1))})$ *are not empty and that it contains with some* $\tau(\cdot) \in \mathbb{R}$ *the set*

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

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

Assume further (B_1) *and that the initial guess satisfies* (A_1) *and* (A_2) *of Section 5.2.2. Then the claims* (5.2.7) *and* (5.2.8) *of Theorem 5.2.1 hold with probability greater than* $1 - 8\mathrm{e}^{-\mathtt{x}} - \beta_{(\mathbf{A})} - I\!\!P(\boldsymbol{\mathcal{N}}(\mathtt{x})^c)$ *for all* $k \in \mathbb{N}$ *. If further condition* (A_3) *with* $\delta(\mathtt{r}) + \tau(\mathtt{r}) \vee \nu_{1,m}\omega\mathtt{r} \le \epsilon\mathtt{r}$ *is satisfied then* (5.2.7) *and* (5.2.8) *are met - for some constant* $\mathtt{C} > 0$ *- with*

$$\mathtt{r}_k \le \mathtt{C}\left(\mathfrak{z}(\mathtt{x}) + \nu^k R_0\right).$$

6.3.2 Choice of basis

To control the approximation bias of the sieve estimator $\widetilde{\boldsymbol{\theta}}_m \in \mathbb{R}^p$ with the approach from Section 4.3.3 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 6.A.6 that the following terms vanish as $m \to \infty$

$$\int_{\mathbb{R}} \boldsymbol{e}_{m+k}(x)\boldsymbol{e}_{m+l}(x)p_{\mathbf{X}^\top\boldsymbol{\theta}^*}(x)dx;\ l, k \in \mathbb{N}, \tag{6.3.1}$$

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 (6.3.1) vanish for any basis of $L^2([-s_{\mathbf{X}}, s_{\mathbf{X}}])$. Of course - following [50] - we could assume that the basis is orthogonal in the inner product induced by the Hessian $\nabla^2 I\!\!E\mathcal{L}(\boldsymbol{\upsilon}^*)$. 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 (6.3.1) can be shown to diminish sufficiently fast for a Lipschitz continuous density $p_{\mathbf{X}^\top\boldsymbol{\theta}^*}$ (see Lemma 6.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 [14] and [13] such a basis $(\boldsymbol{e}_k)$ is available on $L^2([-s_{\mathbf{X}}, s_{\mathbf{X}}])$. This basis possesses all the properties needed for the proof of Lemma 6.A.6 and thus will allow us to control the approximation bias in (6.1.11).

To understand the choice of this basis $(\boldsymbol{e}_k)_{k\in\mathbb{N}}$ we first have to briefly explain how the Daubechies wavelets are derived. To ease understanding we adopt the notation of [14]. Starting with a scaling function $\phi : \mathbb{R} \to \mathbb{R}$ where $\|\phi\|_{L^2(\mathbb{R})} = 1$ one obtains a sequence of nested spaces, i.e. for $j \in \mathbb{N}$

$$V_j = \operatorname{span}\{2^{-j/2}\phi(2^{-j}\cdot -n); n \in \mathbb{Z}\} \subset L^2(\mathbb{R}),$$
$$\ldots \subset V_1 \subset V_0 \subset V_{-1} \subset \ldots \subset L^2(\mathbb{R}).$$

If the scaling function $\phi : \mathbb{R} \to \mathbb{R}$ satisfies certain properties one can show that $\overline{\bigcup_{n\in\mathbb{Z}} V_n} = L^2(\mathbb{R})$ and that $(2^{-j/2}\phi(2^{-j}\cdot -n))_{n\in\mathbb{Z}}$ is an orthonormal basis in $V_j \subset L^2(\mathbb{R})$ for every $j \in \mathbb{Z}$ (see Theorem 6.3.6 of [14]). Denote for each $j \in \mathbb{Z}$ by $W_j \subset L^2(\mathbb{R})$ the orthogonal complement of $V_{j+1} \subset L^2(\mathbb{R})$ in $V_j \subset L^2(\mathbb{R})$. This gives

$$V_j = V_{j+1} \oplus W_{j+1} = \bigoplus_{\substack{k>j,\\ k\in\mathbb{Z}}} W_k, \text{ such that } L^2(\mathbb{R}) = \bigoplus_{j\in\mathbb{Z}} W_j. \quad (6.3.2)$$

The idea of Daubechies wavelets is to find a function $\psi \in W_1$ that satisfies with $\psi_{j,n} \overset{\text{def}}{=} 2^{-j/2}\psi(2^{-j}\cdot +n)$

$$W_j = \operatorname{span}(\Psi_{j,n}; n \in \mathbb{Z}), \;\; \langle \psi_{j,n}, \psi_{j,n'} \rangle_{L^2} = \delta_{n,n'},\, n, n' \in \mathbb{Z}.$$

This is indeed possible. For this denote

$$h_n = \langle \phi, \phi(2\cdot +n) \rangle, n \in \mathbb{Z}, \text{ i.e. } \phi = \sqrt{2}\sum_{n\in\mathbb{Z}} h_n \phi(2\cdot -n),$$

and define

$$\psi = \sqrt{2}\sum_{n\in\mathbb{Z}}(-1)^{n-1}h_{-n-1}\phi(2\cdot -n).$$

Theorem 6.3.6 and Chapter 6.4 of [14] and the table 3.1 of [7] show that there exists a scaling function $\phi_9 : \mathbb{R} \to \mathbb{R}$ for which the associated family $\psi_{j,n} \overset{\text{def}}{=} 2^{-j/2}\psi(2^{-j}\cdot +n)$ satisfies

$$(\psi_{j,n})_{j,n\in\mathbb{Z}} \text{ ONB of } L^2(\mathbb{R}),\ \operatorname{supp}(\psi) \subseteq [0,17],\ \psi \in C^3(\mathbb{R}). \quad (6.3.3)$$

Thus we obtain a well-suited basis for $L^2(\mathbb{R})$ but we only need one for $L^2([-s_{\mathbf{X}}, s_{\mathbf{X}}])$. We could simply embed

$$L^2([-s_{\mathbf{X}}, s_{\mathbf{X}}]) \to L^2(\mathbb{R}),\ \ f(\cdot) \mapsto f(\cdot)1_{[-s_{\mathbf{X}}, s_{\mathbf{X}}]},$$

and use that basis but this would mean that we have to include basis functions $\psi_{j,n} \in L^2(\mathbb{R})$ for positive $j \in \mathbb{N}$ as well. We want to avoid this. We would like to do the following: First adapt the scale and support of the basis and the corresponding shift operation to the interval via redefining

$$\phi_{9,s_{\mathbf{X}}}(t) = s_{\mathbf{X}}^{-1/2}\phi_9(s_{\mathbf{X}}^{-1}t+1),\ \ \psi_{s_{\mathbf{X}}}(t) = s_{\mathbf{X}}^{-1/2}\psi(s_{\mathbf{X}}^{-1}t+1).$$

The associated wavelet basis $\psi_{j,n} \overset{\text{def}}{=} 2^{-j/2}\psi_{s_{\mathbf{X}}}(2^{-j}\cdot +ns_{\mathbf{X}})$ still satisfies all properties in (6.3.3) where the support is adapted to read $[-s_{\mathbf{X}}, 16s_{\mathbf{X}}]$. Next note that (6.3.2) and the definition of the subspaces implies

$$L^2(\mathbb{R}) = V_0 \oplus \bigoplus_{j\in\mathbb{N}} W_{-j},$$

where the definition is adapted to read $V_j = \overline{\operatorname{span}}\{2^{-j/2}\phi_{7,s_{\mathbf{X}}}(2^{-j}\cdot -ns_{\mathbf{X}}); n \in \mathbb{Z}\} \subset L^2(\mathbb{R})$. As we only have to approximate functions that are nonzero on $[-s_{\mathbf{X}}, s_{\mathbf{X}}]$ this suggest the following basis: for $k = 2^{j_k} + j_k 17 - 1 + r_k \in \mathbb{N}$ where $j_k \in \mathbb{N}_0$ and $r_k \in \{0, \dots, 2^{j_k} + 17 - 1\}$ we set

$$\boldsymbol{e}_k \overset{\text{def}}{=} \begin{cases} \phi_{9,s_{\mathbf{X}}}(t-(k-1)s_{\mathbf{X}}) & \text{if } k \leq 17, \\ \psi_{-j_k,r_k} & \text{if } k > 17. \end{cases}$$

These are all elements of a basis for $L^2(\mathbb{R})$ which have a support with nonempty intersection with $[-s_{\mathbf{X}}, s_{\mathbf{X}}]$. We end up with something that resembles a basis 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 $k = 2^{j_k} + j_k 17 - 1 + r_k \in \mathbb{N}$

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

The trouble is, that on each level $j \in \mathbb{N}$ there are 34 wavelets whose support is not contained in $[-s_{\mathbf{X}}, s_{\mathbf{X}}]$. But again there is a remedy that introduces new scaling functions $\phi_7^{left}, \phi_7^{right}$ to deal with the edges of the interval, see [13] Theorem 4.4. The technique presented in [13] allows to contruct 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 $r_k \in \{0, \ldots, 2^{j_k} + 17 - 1\}$

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

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

$$\begin{aligned} &\left| \left\{ l = 2^{j_l} - j_l 17 + r_l \middle| r_l \in \{0, \ldots, 2^{j_l} + 16\}, \, \mathrm{supp}(\boldsymbol{e}_k) \cap \mathrm{supp}(\boldsymbol{e}_l) \neq \emptyset \right\} \right| \\ &\quad \leq \ulcorner 2^{(j_l - j_k)} 17 \urcorner. \end{aligned} \tag{6.3.4}$$

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 $l \geq k$ 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 $(\boldsymbol{\varkappa})$ and $(\boldsymbol{\upsilon\varkappa})$ from Section 4.3.3 in Lemma 6.A.6.

6.3.3 Conditions satisfied

In this section we show that the conditions of section 4.2.1 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 6.3.5. *Assume $(\mathcal{A})$ then there is a constant $\mathtt{C} > 0$ that depends only on $\|p_{\mathbf{X}^\top \boldsymbol{\theta}^*}\|_\infty, \mathtt{C}_{\|\boldsymbol{f}^*\|}, s_{\mathbf{X}}, L_{p_{\mathbf{X}}}$ such that with*

$$\mathtt{r}^* = \mathtt{C}\sqrt{n} m^{-(1+2\alpha)/2} \sqrt{m}. \tag{6.3.5}$$

we get $\|\mathcal{D}_m(\boldsymbol{\upsilon}_m^* - \boldsymbol{\upsilon}^*)\| \leq \mathtt{r}^*$.

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 6.3.6. *Define*

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

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

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

Remark 6.3.2. This Lemma also ensures that the alternating sequence $(\widetilde{\boldsymbol{\theta}}_k, \widetilde{\boldsymbol{\eta}}_{k(-1)})$ introduced in Section 6.2.4 lies in $S_1^{p,+} \times B^m_{\mathtt{r}^\circ}(0)$, with

$$\mathtt{r}^\circ \leq \mathtt{C}\sqrt{p^*\log(p^*)+\mathtt{x}}. \tag{6.3.6}$$

Note that - using that by Lemma 6.A.5 we have $\mathcal{D}_m \geq c_{\mathcal{D}} > 0$ - this also means that

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

Now we show that the general conditions of section 4.2.1 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 $\mathcal{L} - I\!\!E_\varepsilon\mathcal{L}$ where $I\!\!E_\varepsilon$ denotes the expectation operator of the law of $(\varepsilon_i)_{i=1,\ldots,n}$ given $(\mathbf{X}_i)_{i=1,\ldots,n}$. 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 6.3.7. *Assume the conditions* $(\mathcal{A})$. *Then with* $\boldsymbol{\upsilon}^\circ = \boldsymbol{\upsilon}^*_m \in \mathbb{R}^{p^*}$ *and*

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

and $\mathtt{x}\leq m$ *we get the conditions of section 4.2.1 on the set*

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

with:

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

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

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

$$\begin{aligned}\mathtt{g}(\mathtt{r}) &= \sqrt{n}c_{\mathcal{D}}\widetilde{g}\mathtt{C}\left(\sqrt{m}+m^{3/2}\mathtt{r}/\sqrt{n}\right)^{-1},\\ \nu^2_{\mathtt{r},m} &= \widetilde{\nu}^2\Big(1+\mathtt{C}\left(m^{3/2}+\mathtt{r}m^2/\sqrt{n}\right)\mathtt{r}/\sqrt{n}\Big)\\ &\qquad +\mathtt{C}\left(m+m^3\mathtt{r}^2/n\right)\Big(\mathtt{x}+\log(2m)\Big)^{1/2}/\sqrt{n}\Big).\end{aligned} \tag{6.3.7}$$

$(\boldsymbol{\mathcal{E}\mathfrak{D}_1})$ *on* $\varUpsilon_\circ(\mathtt{r})$ *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*

$$\mathtt{g} \ge \frac{\sqrt{n}}{\mathtt{r}m^{3/2}C_{(\boldsymbol{\mathcal{E}\mathfrak{D}_1})}}, \qquad \omega \overset{\mathrm{def}}{=} \frac{2}{\sqrt{n}c_{\mathfrak{D}}}, \qquad \nu_{1,m}^2 = \widetilde{\nu}^2 \mathtt{C}_{(\boldsymbol{\mathcal{E}\mathfrak{D}_1})} m^2,$$

where $C_{(\boldsymbol{\mathcal{E}\mathfrak{D}_1})}$ *is some constant that only depends on* $\|\psi\|, \|\psi'\|, \|\psi''\|$, $L_{p_{\mathbf{X}}}, s_{\mathbf{X}}, c_{\mathfrak{D}}$, *etc..*

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

$$\delta(\mathtt{r}) = \frac{\mathtt{C}_{(\boldsymbol{\mathcal{L}_0})} \left\{ m^{3/2} + \mathtt{C}_{bias} m^{5/2} \right\} \mathtt{r}}{c_{\mathfrak{D}} \sqrt{n}}.$$

The constant $\mathtt{C}_{(\boldsymbol{\mathcal{L}_0})} > 0$ *is polynomial of* $\|\psi\|_\infty$ $, \|\psi'\|_\infty$, $\|\psi''\|_\infty$, $\mathtt{C}_{\|\boldsymbol{f}^*\|}$, $L_{\nabla \varPhi}$, $s_{\mathbf{X}}$, $c_{\mathfrak{D}}^{-1}$ *and* $\|p_{\mathbf{X}^\top \boldsymbol{\theta}^*}\|_\infty$ *and is independent of* $\mathtt{x}$, n, p^* .

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

$$\mathtt{r}^2 \ge (3(2 + \mathtt{C})\mathtt{r}^{*2} + \mathtt{C}_{\sum})/(c\mathtt{b}) \vee m \tag{6.3.8}$$

for certain constants $c_{(\boldsymbol{\mathcal{L}}\mathtt{r})}$, c, $\mathtt{C}$, $\mathtt{C}_{\sum} > 0$ *and with probability greater than* $1 - \exp\left\{-m^3\mathtt{x}\right\} - \exp\left\{-nc_{(\boldsymbol{Q})}/4\right\}$. *In the case that* $\mathtt{C}_{bias} \neq 0$ *we get for*

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

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

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

Remark 6.3.3. The condition $\mathtt{r}m^2/\sqrt{n} \le 1$ needed for $(\boldsymbol{\mathcal{E}\mathfrak{D}_1})$ can be relaxed to read $\mathtt{r}m^{3/2}/\sqrt{n} \le 1$ if one increases $\nu_{1,m}^2 = \widetilde{\nu}^2 \mathtt{C}_{(\boldsymbol{\mathcal{E}\mathfrak{D}_1})} m^3$. This does not change the bounds for $\diamondsuit(\mathtt{r}, \mathtt{x})$, as $\delta(\mathtt{r})$ then still is 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 (6.2.5), as we explain in Remark 6.3.6.

For the regularity condition $(\mathcal{I})$ we use the following Lemma.

Lemma 6.3.8. *Under the assumptions of the last lemma the identifiability condition* $(\mathcal{I})$ *is satisfied with*

$$\nu^2 \le 1 - \frac{c_{\mathfrak{D}}}{\mathtt{C}p}.$$

Proof. This follows from $\mathcal{D} \ge c_{\mathcal{D}} Id$ with Lemma 4.A.6 where

$$\nu^2 \le 1 - \frac{n c_{\mathcal{D}}}{\lambda_{\max} D \wedge \lambda_{\max} \mathcal{H}} \le 1 - \frac{c_{\mathcal{D}}}{\mathtt{C} p}.$$

where we used Lemma 6.A.6 to bound $\lambda_{\max} D \le \mathtt{C} p$ in the last step. □

Finally we apply the following Lemma 4.2.1 to obtain the conditions $(\breve{\mathcal{L}}_0)$, $(\breve{\mathcal{E}}\mathcal{D}_1)$ and $(\breve{\mathcal{E}}\mathcal{D}_0)$.

Remark 6.3.4. 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 $A\mathrm{H}^{-2}$ for better bounds in the proof of condition $(\breve{\mathcal{L}}_0)$. As this Chapter is very long and technical without this sophistication we postpone this improvement to future work.

6.3.4 Large deviations

Next we determine the necessary size of the radius $\mathtt{r}_0(\mathtt{x})$ defined by

$$\begin{aligned}
\mathtt{r}_0(\mathtt{x}) &\overset{\text{def}}{=} \inf\{\mathtt{r} > 0 : \ I\!P\{\widetilde{\boldsymbol{\upsilon}}_m, \widetilde{\boldsymbol{\upsilon}}_{\boldsymbol{\theta}_m^*, m} \in \varUpsilon_\circ(\mathtt{r})\} \le \mathrm{e}^{-\mathtt{x}}\},\\
\widetilde{\boldsymbol{\upsilon}}_{\boldsymbol{\theta}_m^*, m} &\overset{\text{def}}{=} \operatorname*{argmax}_{\substack{\boldsymbol{\upsilon} \in \varUpsilon_m \\ \varPi_{\boldsymbol{\theta}} \boldsymbol{\upsilon} = \boldsymbol{\theta}_m^*}} \mathcal{L}_m(\boldsymbol{\upsilon}),\\
\varUpsilon_\circ(\mathtt{r}) &\overset{\text{def}}{=} \{\boldsymbol{\upsilon} \in \mathbb{R}^{p^*} : \|\mathcal{D}_m(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}_m^*)\| \le \mathtt{r}\}.
\end{aligned}$$

We want to use Theorem 3.3.2. For this we have with Lemma 6.3.6 combined with Lemma 6.A.16 that condition $(\mathcal{E}\mathtt{r})$ is met with probability $1 - 2\mathrm{e}^{-\mathtt{x}}$ and with (setting $\mathtt{r} = \mathtt{C}\sqrt{n}\sqrt{p^* \log(p^*)}$ in (6.3.7))

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

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 $\mathtt{b} > 0$

$$I\!P\left(-I\!E_\epsilon \mathcal{L}_m(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}_m^*) \ge \mathtt{b}\mathtt{r}^2\right) \ge 1 - \mathrm{e}^{-\mathtt{x}} - \exp\left\{-m^3 \mathtt{x}\right\} - \exp\left\{-n c_{\boldsymbol{(Q)}}/4\right\}.$$

Note that the second condition (3.3.5) of Theorem 3.3.2 is satisfied in our setting for $n \in \mathbb{N}$ large enough as we assume that $p^{*5}(1 + \mathtt{C}_{bias} \log(n))/n \to 0$. Finally we only have to ensure that $\mathtt{r}_0 > 0$ is large enough to satisfy (6.3.8), then Theorem 3.3.2 yields the following corollary.

Corollary 6.3.9. *Consider the set*

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

Then it holds that

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

where

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

Repeating the same steps from above gives that on the set

$$\{(\mathcal{E}\mathtt{r}) \text{ and } (\mathcal{L}_{\mathtt{r}}) \text{ are met}\} \cap \left\{ \sup_{\boldsymbol{\theta} \in S_1^{p,+}} \left\| \widetilde{\boldsymbol{\eta}}^{(\infty)}_{m,\boldsymbol{\theta}} \right\| \le \mathtt{C}\sqrt{p^* \log(p^*) + \mathtt{x}} \right\}$$
$$\cap \left\{ \sup_{\boldsymbol{\upsilon} \in \Upsilon_m \setminus \Upsilon_\circ(\mathtt{r}_0^\circ)} \mathcal{L}(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^*_m) < 0 \right\}.$$

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

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

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

Corollary 6.3.10. *Consider the set*

$$\mathcal{B} \stackrel{\text{def}}{=} \{(\mathcal{E}\mathtt{r}) \text{ and } (\mathcal{L}_{\mathtt{r}}) \text{ are met}\} \cap \left\{ \sup_{\boldsymbol{\theta} \in S_1^{p,+}} \left\| \widetilde{\boldsymbol{\eta}}^{(\infty)}_{m,\boldsymbol{\theta}} \right\| \le \mathtt{C}\sqrt{p^* \log(p^*) + \mathtt{x}} \right\}$$
$$\cap \left\{ \sup_{\boldsymbol{\upsilon} \in \Upsilon_m \setminus \Upsilon_\circ(\mathtt{r}_0^\circ)} \mathcal{L}(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^*_m) < 0 \right\}.$$

Then it holds that

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

where

$$\mathtt{r}_0 \le \begin{cases} \mathtt{C}\sqrt{\mathtt{x} + p^*} & \text{if } \mathtt{C}_{bias} = 0, \\ \mathtt{C}\sqrt{\mathtt{x} + p^* \log(n)} & \text{if } \mathtt{C}_{bias} > 0. \end{cases} \tag{6.3.9}$$

6.3.5 Proof of finite sample Wilks and Fisher expansion

Combining Lemma 6.3.7 and Corollary 6.3.10 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:

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

where $\mathtt{C}_{\diamondsuit} > 0$ is a polynomial of $\|\psi\|_\infty, \|\psi'\|_\infty, \|\psi''\|_\infty, \mathtt{C}_{\|\boldsymbol{f}^*\|}, L_{\nabla \Phi}, s_{\mathbf{X}}$.

With these results the case $\mathtt{C}_{bias} = 0$ in Proposition 6.2.1 is merely a corollary of Theorem 6.3.2 and of Lemma 4.2.1. More precisely define the set

$$\begin{aligned}
&\boldsymbol{\mathcal{N}}(\mathtt{x})\\
&\quad \stackrel{\text{def}}{=} \left\{ \sup_{\boldsymbol{\theta} \in S_1^{p,+}} \left\| \widetilde{\boldsymbol{\eta}}^{(\infty)}_{m,\boldsymbol{\theta}} \right\| \le \mathtt{C}\sqrt{p^* \log(p^*) + \mathtt{x}} \right\}\\
&\qquad \cap \left\{ \sup_{\boldsymbol{\upsilon} \in \varUpsilon_m \setminus \varUpsilon_\circ(\mathtt{r}_0^\circ)} \mathcal{L}(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^*_m) < 0 \right\}\\
&\qquad \cap \{ \widetilde{\boldsymbol{\upsilon}}_m, \widetilde{\boldsymbol{\upsilon}}_{\boldsymbol{\theta}^*_m, m} \in \{ \|\mathcal{D}_m(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^*_m)\| \le \mathtt{r}_0 \} \}\\
&\qquad \cap \left\{ \sup_{\boldsymbol{\upsilon} \in \varUpsilon_\circ(\mathtt{r}_0)} \|\nabla(I\!\!E - I\!\!E_\varepsilon)[\mathcal{L}(\boldsymbol{\upsilon}^*_m) - \mathcal{L}(\boldsymbol{\upsilon})]\| \le \mathtt{C}(\mathtt{x} + p^*)^2 \mathtt{r}_0/\sqrt{n} \right\}\\
&\qquad \cap \{ \text{The conditions of Section 4.2.1 are met for } (\mathcal{L}, \varUpsilon_m, \mathcal{D}) \}.
\end{aligned}$$

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 Section 3.4 on the deviation behavior of quadratic forms we can bound with some constant related to the finite value $\operatorname{tr}(\breve{D}^{-1}\breve{V}^2\breve{D}^{-1})$

$$I\!\!P(\|\breve{D}^{-1}\breve{\nabla}\|) \le \mathfrak{z}(\mathtt{x}, I\!\!\breve{B})) \ge 1 - 2\mathrm{e}^{-\mathtt{x}}, \;\; \mathfrak{z}(\mathtt{x}, I\!\!\breve{B}) \le \sigma \mathtt{C}\sqrt{p^* + \mathtt{x}}.$$

Thus we get the claim with Theorem 6.3.2 via adapting the size of $\mathtt{C}_{\diamondsuit} > 0$.

For the case that $\mathtt{C}_{bias} > 0$ we want to apply Proposition 6.3.3. For this define

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

Then $\mathtt{r}_0 > 0$ in (6.3.9) satisfies by assumption

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

since $m^3 \log(n)/\sqrt{n} \to 0$. Consequently Proposition 6.3.3 applies with $\boldsymbol{\mathcal{N}}(\mathtt{x})$ from above, which yields the claim of Proposition 6.2.1 with an error term

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

where

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

6.3.6 Bounding the sieve bias

We prove this claim via showing that the conditions of Corollary 4.3.1 and Theorem 4.3.2 are met, which can be adapted to the regression set up in the same way as we did with Theorem 4.2.2 and Proposition 4.2.3. This concerns especially condition (bias) from Section 4.3. For this we use the conditions $(\mathcal{L}\mathtt{r}_\infty)$ and $(\boldsymbol{\varkappa})$ from Section 4.3.3 and then we can use Theorem 4.3.4. But exactly this is done in Lemma 6.A.6. Thus 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 $2(\|\breve{D}^{-1}\breve{\nabla}\|+\mathtt{r}_p^*(\mathtt{x}_n))\diamondsuit(\mathtt{r}_2,\mathtt{x}_n)\xrightarrow{\mathbb{P}}0$ 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 also focus on the case $\mathtt{C}_{bias}=0$ and use Corollary 4.3.3. 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 $(\boldsymbol{\upsilon\varkappa})$ - as proven in Lemma 6.A.6 - we can apply Lemma 4.A.9 to find

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

and with Lemma 4.A.10

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

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 $\operatorname{Cov}(Y_i(m))\to 0$ where

$$\begin{aligned}Y_i(m)\overset{\mathrm{def}}{=}(\frac{1}{\sqrt{n}}\breve{D}_m)^{-1}\{&\nabla_{\boldsymbol{\theta}}\left(\ell_i(\boldsymbol{\upsilon}_m^*)-\ell_i(\boldsymbol{\upsilon}^*)\right)\\ &-\mathrm{A}_m\mathrm{H}_m^{-2}\nabla_{(\eta_1,\ldots,\eta_m)}\left(\ell_i(\boldsymbol{\upsilon}_m^*)-\ell_i(\boldsymbol{\upsilon}^*)\right)\}.\end{aligned}$$

which is done with Lemma 6.A.26. This completes the proof after plugging in the bounds.

6.3.7 Proof of convergence of the alternating procedure

Here we want to explain in more detail how the Propositions 6.2.3 and 6.2.4 can be derived.

We want to use Theorem 6.3.4, i.e. the adapted version of Theorem (5.2.1). For this it remains to check the conditions $(\mathbf{A}_1)$, $(\mathbf{A}_2)$ and $(\mathbf{A}_2)$ from Section 5.2.2 for the initial guess defined in (6.2.5).

Remark 6.3.5. Condition $(\mathbf{B}_1)$ is met in our case as we pointed out in Section 6.3.4.

We can prove the following lemma:

Lemma 6.3.11. *It holds for* $\mathtt{x}\le\mathtt{C}\widetilde{\nu}^2\widetilde{\mathtt{g}}^2 n$ *that*

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

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* (5.2.2)*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 6.3.7).*

Together with Theorem 6.3.4 this implies Proposition 6.2.3 as we can bound

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

Remark 6.3.6. $\epsilon \mathrm{R}_0 \to 0$ implies $\mathrm{R}_0 m^{3/2}/\sqrt{n} \to 0$. As pointed out in Remark 6.3.3 this means that the conditions from Section 4.2.1 can be satisfied on $\Upsilon_\circ(\mathrm{R}_0)$.

For Proposition 6.2.4 we apply Theorem 5.2.3, which can be adapted to the regression setup in analogy to Theorem 6.3.4. 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

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

We derive a bound for $\mathfrak{z}(\mathtt{x}, \nabla^2\mathcal{L}(\boldsymbol{\upsilon}^*))$ in Lemma 6.A.31 which is based on Corollary 3.7 of [56], as is proposed in Remark 5.2.14. The claim of Proposition 6.2.4 is shown with the following Lemma.

Lemma 6.3.12. *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* $\varkappa(\mathtt{x}, R_0) \to 0$ *with* $n \to \infty$.

Remark 6.3.7. The bound for $\mathtt{x}$ comes from Lemma 6.A.31 but also from the definition of $\mathfrak{z}_1(\mathtt{x}, 3p^*)$ in (3.5.6) and ensures that $\mathfrak{z}_1(\mathtt{x}, 3p^*) = O(\sqrt{\mathtt{x} + p^*})$.

6.A Proofs

In the following all the technical steps necessary to prove the Lemmas of section 6.3 are presented. But first we cite an important result that will be used in our arguments, namely the bounded difference inequality:

Theorem 6.A.1 (Bounded differences inequality)**.** *Let a function* $f : \mathcal{X}^n \to \mathbb{R}$ *satisfy for any* $\mathbf{X}_1, \ldots, \mathbf{X}_n, \mathbf{X}_i' \in \mathcal{X}$

$$|f(\mathbf{X}_1, \ldots, \mathbf{X}_i, \ldots, \mathbf{X}_n) - f(\mathbf{X}_1, \ldots, \mathbf{X}_i', \ldots, \mathbf{X}_n)| \le c_i.$$

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

$$I\!\!P\left(f(\mathbf{X}) - I\!\!E f(\mathbf{X}) \ge t\right) \le \mathrm{e}^{-\frac{2t^2}{\sum_{i=1}^n c_i^2}},$$

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

Furthermore we will use the basic chaining device as it was introduced by [17] (see Section 2 of [55] for a more concise description). As we use the idea several times, we summarize the central step in the following Lemma

Lemma 6.A.2. *Let* $\{\mathcal{Y}(\boldsymbol{v})-\mathcal{Y}(\boldsymbol{v}^*),\ \boldsymbol{v}\in\Upsilon\}$ *be a family of random variables index by a set* Υ *that is contained in a normed space* $(\mathcal{X},\|\cdot\|)$. *Define* $\Upsilon_0=\{\boldsymbol{v}^*\}$ *and with some* $\mathtt{r}>0$ *the sequence* $\mathtt{r}_k=2^{-k}\mathtt{r}$ *and the sequence of sets* Υ_k *each with minimal cardinality such that*

$$\Upsilon\subset\bigcup_{\boldsymbol{v}\in\Upsilon_k}B_{\mathtt{r}_k}(\boldsymbol{v}),\qquad B_{\mathtt{r}}(\boldsymbol{v})\overset{\text{def}}{=}\{\boldsymbol{v}^\circ\in\Upsilon,\ \|\boldsymbol{v}^\circ-\boldsymbol{v}\|\le\mathtt{r}\}.$$

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

$$\begin{aligned}&I\!\!P\left(\sup_{\boldsymbol{v}\in\Upsilon}|\mathcal{Y}(\boldsymbol{v})-\mathcal{Y}(\boldsymbol{v}^*)|\ge\mathfrak{z}\right)\\&\quad\le\sum_{k=1}^{\infty}|\Upsilon_k|\sup_{\boldsymbol{v}^\circ\in\Upsilon_k}I\!\!P\left(\inf_{\boldsymbol{v}\in\Upsilon_{k-1}}|\mathcal{Y}(\boldsymbol{v})-\mathcal{Y}(\boldsymbol{v}^\circ)|\ge2^{-(k-1)/2}(1-1/\sqrt{2})\mathfrak{z}\right).\end{aligned}$$

Proof. We simply use the definition and estimate

$$\begin{aligned}&I\!\!P\left(\sup_{\boldsymbol{v}\in\Upsilon}|\mathcal{Y}(\boldsymbol{v})-\mathcal{Y}(\boldsymbol{v}^*)|\ge\mathfrak{z}\right)\\&\quad\le I\!\!P\left(\sum_{k=1}^{\infty}\sup_{\boldsymbol{v}_k\in\Upsilon_k}\inf_{\boldsymbol{v}_{k-1}\Upsilon_{k-1}}|\mathcal{Y}(\boldsymbol{v}_k)-\mathcal{Y}(\boldsymbol{v}_{k-1})|\ge\mathfrak{z}\right)\\&\quad\le\sum_{k=1}^{\infty}I\!\!P\Bigg(\sup_{\boldsymbol{v}_k\in\Upsilon_k}\inf_{\boldsymbol{v}_{k-1}\in\Upsilon_{k-1}}|\mathcal{Y}(\boldsymbol{v}_k)-\mathcal{Y}(\boldsymbol{v}_{k-1})|\\&\qquad\qquad\qquad\qquad\ge2^{-(k-1)/2}(1-1/\sqrt{2})\mathfrak{z}\Bigg)\\&\quad\le\sum_{k=1}^{\infty}|\Upsilon_k|\sup_{\boldsymbol{v}_k\in\Upsilon_k}I\!\!P\Bigg(\inf_{\boldsymbol{v}_{k-1}\in\Upsilon_{k-1}}|\mathcal{Y}(\boldsymbol{v}_k)-\mathcal{Y}(\boldsymbol{v}_{k-1})|\\&\qquad\qquad\qquad\qquad\ge2^{-(k-1)/2}2^{-(k-1)/2}(1-1/\sqrt{2})\mathfrak{z}\Bigg),\end{aligned}$$

where we used that $\sum_{k=1}^{\infty}2^{-(k-1)/2}\le1/(1-1/\sqrt{2})$. □

6.A.1 Proof of Remark 6.2.5

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

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

Note that the application of Fubini's theorem is justified since by assumtion $|\boldsymbol{f}_{\boldsymbol{\theta},\boldsymbol{x}}(t)\boldsymbol{e}_k(t)p_{\mathbf{X}|\mathbf{X}^{\top}\boldsymbol{\theta}=t}(\boldsymbol{x})| < \infty$. Furthermore with Jensen's inequality and exchanging the order integration and summation as the $\limsup$ 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 $\alpha(\boldsymbol{\theta}) \leq \alpha(\boldsymbol{\theta},\boldsymbol{x})$. □

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

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

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

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

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

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

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

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

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

$$\varsigma_{i,m}(\boldsymbol{\upsilon}) \stackrel{\text{def}}{=} \Big(\boldsymbol{f}_{\boldsymbol{\eta}}'(\mathbf{X}_i^\top\boldsymbol{\theta})\nabla\varPhi(\boldsymbol{\theta})^\top\mathbf{X}_i, \boldsymbol{e}(\mathbf{X}_i^\top\boldsymbol{\theta})\Big) \in \mathbb{R}^{p^*}. \tag{6.A.1}$$

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

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

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

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

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

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

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

$$\mathcal{D}^2(\boldsymbol{\upsilon}) = nd^2(\boldsymbol{\upsilon})+I\!\!E\left[\Big[\boldsymbol{f}_{\boldsymbol{\eta}}(\mathbf{X}^\top\boldsymbol{\theta})-g(\mathbf{X})\Big]\begin{pmatrix} v^2_{\boldsymbol{\theta}}(\boldsymbol{\upsilon}) & b^\top_\infty(\boldsymbol{\upsilon})\\ b^\top_\infty(\boldsymbol{\upsilon}) & 0\end{pmatrix}\right],$$

where with the obvious adaptations

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

Remark 6.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 $\boldsymbol{f}_{\boldsymbol{\eta}^*}(\mathbf{X}^\top\boldsymbol{\theta}^*)\overset{\text{def}}{=}I\!\!E[g(\mathbf{X})|\mathbf{X}^\top\boldsymbol{\theta}^*]$.

Furthermore we calculate - with $\varsigma_{i,m}$ from (6.A.1) -

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

where

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

6.A.3 Preliminary calculations

Lemma 6.A.3. *We have*

$$\begin{aligned} &|\mathbb{E}[\boldsymbol{e}_k \boldsymbol{e}_l(\mathbf{X}^\top \boldsymbol{\theta})]| \\ &\quad \le 17 s_{\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), \ \textit{for } l \ge k \qquad (6.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}^\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), \qquad (6.A.3) \\ &|\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 2^{3(j_l+j_k)/2 - (j_l \vee j_k)} 1_{\{I_k \cap I_l \neq \emptyset\}}(k,l), \qquad (6.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}'\|^2 2^{j_k} 2^{j_l} \|\psi'\|_\infty^2 s_{\mathbf{X}}^4 17^2 1_{\{I_k \cap I_l \neq \emptyset\}}, \qquad (6.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}'\|^2 2^{2j_k} 2^{2j_l} \|\psi''\|_\infty^2 s_{\mathbf{X}}^4 17^2 1_{\{I_k \cap I_l \neq \emptyset\}}, \qquad (6.A.6) \\ &\mathbb{E}\left[\Big(\boldsymbol{e}_l(\mathbf{X}^\top \boldsymbol{\theta}) - \boldsymbol{e}_l(\mathbf{X}^\top \boldsymbol{\theta}^*_m)\Big) \boldsymbol{e}_k(\mathbf{X}^\top \boldsymbol{\theta})\right] \\ &\quad \le \mathtt{C} \|\boldsymbol{\theta} - \boldsymbol{\theta}'\| 2^{j_l/2} 2^{(j_k \wedge j_l)/2}. \qquad (6.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 $L_{p_{\mathbf{X}^\top \boldsymbol{\theta}^*}} \le s_{\mathbf{X}}^p L_{p_{\mathbf{X}}}$. 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

$$\begin{aligned}
I\!E[\boldsymbol{e}_k\boldsymbol{e}_l(\mathbf{X}^\top\boldsymbol{\theta})] &= \int_{I_l} \boldsymbol{e}_k(x)\boldsymbol{e}_l(x)p_{\mathbf{X}^\top\boldsymbol{\theta}^*}(x)dx \\
&= \int_{I_l} \boldsymbol{e}_k(x)\boldsymbol{e}_l(x)p_{\mathbf{X}^\top\boldsymbol{\theta}^*}(x_0)dx1_{\{I_l\cap I_k\neq\emptyset\}}(k,l) \\
&\quad + \int_{I_l} \boldsymbol{e}_k(x)\boldsymbol{e}_l(x)\Big(p_{\mathbf{X}^\top\boldsymbol{\theta}^*}(x) - p_{\mathbf{X}^\top\boldsymbol{\theta}^*}(x_0)\Big)dx1_{\{I_l\cap I_k\neq\emptyset\}}(k,l),
\end{aligned}$$

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 $l = 2^{j_l} + j_l17 - 1 + r_l \in \mathbb{N}$. 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

$$\begin{aligned}
&|\int_{I_l} \boldsymbol{e}_k(x)\boldsymbol{e}_l(x)\Big(p_{\mathbf{X}^\top\boldsymbol{\theta}^*}(x) - p_{\mathbf{X}^\top\boldsymbol{\theta}^*}(x_0)\Big)dx|1_{\{I_l\cap I_k\neq\emptyset\}}(k,l) \\
&\quad \le s_{\mathbf{X}}^p L_{p_{\mathbf{X}}} 2^{-j_l-1} \int_{I_l} |\boldsymbol{e}_k(x)||\boldsymbol{e}_l(x)|dx1_{\{I_l\cap I_k\neq\emptyset\}}(k,l) \\
&\quad \le s_{\mathbf{X}}^p L_{p_{\mathbf{X}}} 2^{-j_l-1} \left(\int_{I_l} \boldsymbol{e}_l(x)^2dx \int_{I_l} \boldsymbol{e}_k(x)^2dx\right)^{1/2} 1_{\{I_l\cap I_k\neq\emptyset\}}(k,l) \\
&\quad \le s_{\mathbf{X}}^p L_{p_{\mathbf{X}}} 2^{-j_l-1} \left(\int_{I_l} \boldsymbol{e}_k(x)^2dx\right)^{1/2} 1_{\{I_l\cap I_k\neq\emptyset\}}(k,l) \\
&\quad \le 17 s_{\mathbf{X}}^{p+1} L_{p_{\mathbf{X}}} \|\psi\|_\infty 2^{-j_l-1} 2^{j_k/2 - j_l/2} 1_{\{I_l\cap I_k\neq\emptyset\}}(k,l),
\end{aligned}$$

where we used that the $(\boldsymbol{e}_k)$ form an orthonormal basis, that $\|\boldsymbol{e}_k\|_\infty \le 2^{j_k/2}\|\psi\|_\infty$ and that I_l is of length $2^{-j_l}17s_{\mathbf{X}}$. This gives (6.A.2). Using that for any $\boldsymbol{\theta} \in W_S$ it holds true that $\|\nabla\varPhi(\boldsymbol{\theta}^*)\boldsymbol{\theta}\| \le \frac{\sqrt{p+2}}{2}\pi$ we estimate similarly to before

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

The bound (6.A.4) follows with exactly the same calculations. To show (6.A.5) we calculate with $M_k \stackrel{\text{def}}{=} \{(x,y) \in \mathbb{R}^2,\, x \in I_k\} \cup \{(x,y) \in \mathbb{R}^2,\, x+y \in I_k\}$ and with $p_{\boldsymbol{\theta},(\boldsymbol{\theta}^\circ-\boldsymbol{\theta})} : \mathbb{R}^2 \to \mathbb{R}_+$ denoting the density of $(\mathbf{X}^\top\boldsymbol{\theta}, \mathbf{X}^\top(\boldsymbol{\theta}^\circ - \boldsymbol{\theta})) \in \mathbb{R}^2$

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

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

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

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 $(\mathbf{Cond_X})$

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

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

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

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

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

We have by (6.A.5)

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

As above we can bound

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

□

Lemma 6.A.4. *For any* $(\boldsymbol{\theta},\boldsymbol{\eta}) \in \mathbb{R}^{p+m}$

$$\begin{aligned}
\|\boldsymbol{e}(\boldsymbol{x})\| &\le \mathtt{C}\|\psi\|_\infty\sqrt{m}, && (6.A.8)\\
|\boldsymbol{f}_{\boldsymbol{\eta}}(\boldsymbol{x})| &\le \mathtt{C}\|\psi\|_\infty\sqrt{m}\|\boldsymbol{\eta}\|, \\
\|\boldsymbol{e}'(\boldsymbol{x})\| &\le \sqrt{17}\|\psi'\|m^{3/2}, && (6.A.9)
\end{aligned}$$

Proof. Clearly $|\boldsymbol{f}_{\boldsymbol{\eta}}(\boldsymbol{x})| \le \|\boldsymbol{\eta}\|\|\boldsymbol{e}(\mathbf{X}_i^\top\boldsymbol{\theta}^*_m)\|$. 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

$$\begin{aligned}
&|M(j)| && (6.A.10)\\
&\quad \overset{\text{def}}{=} \left|\left\{k \in \{(2^j + j17 - 1,\ldots,2^{j+1} + j17 + 15\} : |\boldsymbol{e}_k(\boldsymbol{x})| \neq 0\right\}\right| \le 17.
\end{aligned}$$

This implies

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

The proof of (6.A.9) works analogously. □

6.A.4 Lower bound for the information operator

Lemma 6.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 6.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 $\|\boldsymbol{\gamma}\| = 1$

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

Using the properties of conditional expectation we can write

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

Using assumption (model bias) we find

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

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

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

Furthermore we show in Lemma 6.A.24 that with some $\mathtt{b}^* > 0$ and $Q > 0$

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

Remark 6.A.3. A close look at the proof of Lemma 6.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

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

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

6.A.5 Regularity

Lemma 6.A.6. *Assume that the density* $p_{\mathbf{X}} : \mathbb{R}^p \to \mathbb{R}$ *is Lipshitz 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* $\lambda \in [0, 1]$

$$\begin{aligned} \|\mathcal{H}_m^{1/2} \boldsymbol{\varkappa}^*\|^2 &< \left(17 \|p_{\mathbf{X}^\top \boldsymbol{\theta}^*}\|_\infty \mathtt{C}_{\|\boldsymbol{f}^*\|} + 17^2 \sqrt{36} s_{\mathbf{X}}^{p+1} L_{p_{\mathbf{X}}} \|\psi\|_\infty \mathtt{C}_{\|\boldsymbol{f}^*\|}^2 \right) n m^{-2\alpha}, \\ \alpha(m) &\stackrel{\text{def}}{=} \|\mathcal{D}_m^{-1} A_{\boldsymbol{\upsilon}\boldsymbol{\varkappa}} \boldsymbol{\varkappa}^*\| \leq \mathtt{C}_1 \sqrt{n} \left(m^{-(\alpha+1/2)} + \mathtt{C}_{bias} m^{-(\alpha-1)} \right), \\ \tau(m) &\stackrel{\text{def}}{=} \|\mathcal{D}_m^{-1} \nabla_{\boldsymbol{\upsilon}\boldsymbol{\varkappa}} I\!E[\mathcal{L}\left((\Pi_{p^*} \boldsymbol{\upsilon}^*, \lambda \boldsymbol{\varkappa}^*) - A_{\boldsymbol{\upsilon}\boldsymbol{\varkappa}} \right)] \boldsymbol{\varkappa}^*\| \leq \mathtt{C}_1 m^{-2\alpha+1/2} \sqrt{n}, \\ 0 &= \left| \boldsymbol{\varkappa}^{*\top} (\mathcal{H}_m - \nabla_{\boldsymbol{\varkappa}\boldsymbol{\varkappa}} I\!E \mathcal{L}(\Pi_{p^*} \boldsymbol{\upsilon}^*, \lambda \boldsymbol{\varkappa}^*)) \boldsymbol{\varkappa}^* \right|, \end{aligned}$$

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

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

Furthermore we find that

$$\|D^2\| \leq n \frac{p+2}{4} \mathtt{C}_{\|f\|} \|\psi'\|_\infty^2 s_{\mathbf{X}}^2 \pi^2.$$

Proof. We have that

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

Due to Lemma 6.A.5

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

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

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

We first analyze the second summand

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

We use (6.A.2) from Lemma 6.A.3 to find

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

Note that for each $j_k = 0, \ldots, j_m$ there exists at most 17 elements $r_k(l) \in \{0, \ldots, 2^{j_k} + 16\}$ 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 $\|\boldsymbol{\eta}\| = 1$

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

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 $r_l \in \{0, \ldots, 2^{j_l} + 17 - 1\}$

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

with

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

Consequently

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

For the second summand we remind the reader that

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

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

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

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

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

The exchange of the order of summation is justified by the subsequent bounds and again the dominated convergence theorem. We again use Lemma 6.A.3 to find with (6.A.3) and with similar arguments to those from above

$$
\begin{aligned}
|\boldsymbol{\theta} a_{\boldsymbol{\theta}\boldsymbol{\varkappa}} \boldsymbol{\varkappa}^*| \leq\ & 17\frac{\sqrt{p+2}}{2}\pi \|\psi'\|_\infty s_{\mathbf{X}}^2 \|p_{\mathbf{X}}\|_\infty \\
& \sum_{k=1}^{\infty} \eta_k^* 2^{3j_k/2} \sum_{l=m+1}^{\infty} \eta_l^* 2^{-(j_l \vee j_k)/2} 1_{\{I_l \cap I_k \neq \emptyset\}}(k,l) \\
\leq\ & 17\frac{\sqrt{p+2}}{2}\pi \|\psi'\|_\infty s_{\mathbf{X}}^2 \|p_{\mathbf{X}}\|_\infty \sum_{k=1}^{\infty} \eta_k^* k^{3/2} \left(\sum_{l=m+1}^{\infty} l^{2\alpha} \eta_l^{*2} \right)^{1/2} \\
& \left(\sum_{j_l=j_m+1}^{\infty} \sum_{r_l=0}^{2^{j_l}+33} 2^{-2\alpha j_l} 2^{-(j_l \vee j_k)} 1_{\{I_l \cap I_k \neq \emptyset\}}(k,l) \right)^{1/2}.
\end{aligned}
$$

We have due to (6.3.4) that

$$
\begin{aligned}
& \sum_{j_l=j_m+1}^{\infty} \sum_{r_l=0}^{2^{j_l}+33} 2^{-2\alpha j_l} 2^{-(j_l \vee j_k)} 1_{\{I_l \cap I_k \neq \emptyset\}}(k,l) \\
&= \sum_{j_l=j_m+1}^{\infty} 2^{-2\alpha j_l} 2^{-(j_l \vee j_k)} \sum_{r_l=0}^{2^{j_l}+33} 1_{\{I_l \cap I_k \neq \emptyset\}}(k,l) \\
&= \sum_{j_l=j_m+1}^{\infty} 2^{-2\alpha j_l} 2^{-(j_l \vee j_k)} \\
&\qquad \left| \left\{ l = 2^{j_l} + j_l 17 - 1 + r_l \,\middle|\, r_l \in \{0, \ldots, 2^{j_l} + 33\},\, I_l \cap I_k \neq \emptyset \right\} \right| \\
&= \sum_{j_l=j_m+1}^{\infty} 2^{-(2\alpha+1)j_l} 2^{-(j_k - j_l)_+} \ulcorner 2^{(j_l - j_k)} 17 \urcorner \\
&\leq 2^{-(2\alpha+1)j_m} 18 \leq 17 m^{-(2\alpha+1)} 18.
\end{aligned}
$$

Which gives

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

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

Furthermore with $\boldsymbol{\theta}^\circ = \nabla\varPhi_{\boldsymbol{\theta}^*}\varphi_{\boldsymbol{\theta}} \in \boldsymbol{\theta}^{*\perp}$

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

We bound

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

We can exchange summation and expectation to find

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

We estimate

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

Such that

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

Collecting both summands

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

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

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

Remark 6.A.4. In case $\mathtt{C}_{bias} > 0$ we do not manage to get a bound for $\boldsymbol{\theta}b_{\boldsymbol{\theta}\boldsymbol{\varkappa}}\boldsymbol{\varkappa}$ for general $\boldsymbol{\varkappa} \in l^2$. 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)

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

As above we find

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

We estimate

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

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

$$
\begin{aligned}
&\sum_{l=k+1}^\infty l^{-2\alpha} 2^{-3j_l} 1_{\{I_l \cap I_k \neq \emptyset\}}(k,l) \\
&\quad \le \sum_{l=k+1}^\infty 2^{-(3+2\alpha)j_l} 1_{\{I_l \cap I_k \neq \emptyset\}}(k,l) \\
&\quad \le \sum_{j=j_{k+1}}^\infty 2^{-(3+2\alpha)j} \\
&\qquad |\{l = 2^j - 2j17 - 1, \dots, 2^{j+1} - 2(j+1)17 - 1 - 1 : \, I_l \cap I_k \neq \emptyset\}| \\
&\quad = \sum_{j=j_{k+1}}^\infty 2^{-(3+2\alpha)j} \ulcorner 2^{j-j_k} 17 \urcorner \\
&\quad \le 2^{-j_k} 18 \sum_{j=j_{k+1}}^\infty 2^{-(2+2\alpha)j} \\
&\quad = 2^{-(3+2\alpha)j_k} 18 \sum_{j=0}^\infty 2^{-(2+2\alpha)j} \le 2^{-(3+2\alpha)j_k} 36.
\end{aligned}
$$

Plugging this in we find

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

From which we obtain

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

Next we show

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

For this note that

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

We infer

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

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 (6.A.4)

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

We can estimate further

$$
\begin{aligned}
&\sum_{k,l=m+1}^{\infty} \eta_k^* \eta_l^* 2^{3(j_l+j_k)/2-(j_l\vee j_k)} 1_{\{I_k\cap I_l\neq\emptyset\}}(k,l) \\
&\quad\le \sum_{k=m+1}^{\infty} \eta_k^* 2^{3j_k/2} \sum_{l=m+1}^{\infty} \eta_l^* 2^{3j_l/2-(j_l\vee j_k)} 1_{\{I_k\cap I_l\neq\emptyset\}}(k,l) \\
&\quad\le \sum_{k=m+1}^{\infty} \eta_k^* 2^{3j_k/2} \left(\sum_{l=m+1}^{\infty} l^{2\alpha} f^{*2}_{\ l}\right)^{1/2} \\
&\qquad \left(\sum_{l=m+1}^{\infty} 2^{(3-2\alpha)2j_l-2(j_l\vee j_k)} 1_{\{I_k\cap I_l\neq\emptyset\}}(k,l)\right)^{1/2}.
\end{aligned}
$$

Observe

$$
\begin{aligned}
&\sum_{l=m+1}^{\infty} 2^{(3-2\alpha)j_l-2(j_l\vee j_k)} 1_{\{I_k\cap I_l\neq\emptyset\}}(k,l) \\
&\quad= \sum_{j=j_m+1}^{\infty} 2^{(3-2\alpha)j-2(j\vee j_k)} \\
&\qquad \left|\left\{ l = j12 + 2^j + r_l \middle| r_l \in \{0,\dots,2^j+11\},\, I_l\cap I_k \neq \emptyset \right\}\right| \\
&\quad= \sum_{j=j_m+1}^{\infty} 2^{(3-2\alpha)j-2(j\vee j_k)} \ulcorner 2^{(j-j_k)}17 \urcorner \\
&\quad\le 18 \sum_{j=j_m+1}^{\infty} 2^{(2-2\alpha)j} = 17^3 18 m^{-2\alpha+2}.
\end{aligned}
$$

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

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

Furthermore

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

Together this implies

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

Clearly

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

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

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

Furthermore we find that

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

and

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

This completes the proof. □

6.A.6 Proof or Lemma 6.3.5

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

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

We proof the claim in two lemmas. The first one concerns condition $(\mathcal{L}\mathtt{r}_\infty)$ from Section 4.3.3. It is important to note, that in the proof of Lemma 4.A.7 it is only needed for the lower bound in (4.A.15). This means that we can use the full expectation $I\!\!E$ instead of $I\!\!E_\varepsilon$:

Lemma 6.A.7. *Assume* $(\mathcal{A})$. *Then there exists a constant* $\mathtt{b}>0$ *such that*

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

Proof. As in Lemma 6.A.8 we can make the decomposition

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

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

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

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

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

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

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

Otherwise - if $\|D(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\|^2 \le \tau^2\mathtt{r}^2$ - we have as in the proof of Lemma 6.A.5

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

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

□

The claim of Lemma 6.3.5 now is a direct consequence of Lemma 4.A.7.

6.A.7 Proof of Lemma 6.3.6

Remark 6.A.5. We assume that the density of the regressors satisfies $p_{\mathbf{X}} \ge c_{p_{\mathbf{X}}} > 0$ 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

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

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 $\|\boldsymbol{e}\| \le \mathtt{C}\sqrt{m}$

$$
\begin{aligned}
\left\|\widetilde{\boldsymbol{\eta}}_{m,\boldsymbol{\theta}}^{(\infty)}\right\| &\overset{\text{def}}{=} \left\|\operatorname*{argmax}_{\boldsymbol{\eta}\in\mathbb{R}^m} \mathcal{L}_m(\boldsymbol{\theta}, \boldsymbol{\eta})\right\| \\
&= \left\|\left(\frac{1}{n}\sum_{i=1}^n \boldsymbol{e}\boldsymbol{e}^\top(\mathbf{X}_i^\top \boldsymbol{\theta})\right)^{-1} \frac{1}{n}\sum_{i=1}^n (g(\mathbf{X}_i) + \varepsilon_i)\boldsymbol{e}(\mathbf{X}_i^\top \boldsymbol{\theta})\right\| \\
&\le \left(\|g\|_\infty \mathtt{C}\sqrt{m} + \left\|\frac{1}{n}\sum_{i=1}^n \varepsilon_i \boldsymbol{e}(\mathbf{X}_i^\top \boldsymbol{\theta})\right\|\right) \\
&\qquad \left\|\left(\frac{1}{n}\sum_{i=1}^n \boldsymbol{e}\boldsymbol{e}^\top(\mathbf{X}_i^\top \boldsymbol{\theta})\right)^{-1}\right\|. \qquad (6.A.11)
\end{aligned}
$$

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

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

We want to apply Theorem 3.5.5 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 condition $(\mathcal{E}d)$ in (3.5.4) is met with $d(\boldsymbol{\eta}, \boldsymbol{\eta}^\circ) = \|\boldsymbol{\eta} - \boldsymbol{\eta}^\circ\|_{\mathbb{R}^m}$. This is indeed the case since by Lemma 6.A.4 for any pair $\boldsymbol{\eta}, \boldsymbol{\eta}^\circ \in B_1(0)$

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

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

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

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

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

This implies with Theorem 3.5.5

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

Two bound the norm of the inverse of the matrix in (6.A.11) we denote

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

Note that with Remark 6.A.5

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

while

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

We bound

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

For the first term we can use the bounded differences inequality (Theorem 6.A.1) to find

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

For the second summand we define $\boldsymbol{\zeta}_{\mathbf{X}}(\boldsymbol{\upsilon})\overset{\text{def}}{=}(P_n-I\!\!P)\boldsymbol{f}_{\boldsymbol{\eta}}(\mathbf{X}_i^\top\boldsymbol{\theta})^2$. We use the chaining method, i.e. Lemma 6.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 $\boldsymbol{\upsilon}'\in B_{\mathtt{r}_k,\mathcal{D}}(\boldsymbol{\upsilon})$

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

We estimate for an application of the bounded differences inequality

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

We have as $\|\eta\|=1$ with Lemma 6.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

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

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

$$I\!P\left(\inf_{\Upsilon_{k-1,m}} |\boldsymbol{\zeta}_{\mathbf{X}}(\boldsymbol{\upsilon}_k) - \boldsymbol{\zeta}_{\mathbf{X}}(\boldsymbol{\upsilon}_{k-1})| \ge \frac{2^{-(k-1)}(1-1/\sqrt{2})s}{\sqrt{n}}\right)$$
$$\le \mathrm{e}^{-s^2}. \tag{6.A.12}$$

Set

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

and plug it into (6.A.12), then we find with Lemma 6.A.2

$$I\!P\left(\sup_{\boldsymbol{\theta}\in S_1^p} \|\boldsymbol{M}_n(\boldsymbol{\theta}) - I\!E\left[\boldsymbol{M}_n(\boldsymbol{\theta})\right]\| \ge \mathtt{C}\sqrt{\mathtt{x}+p^*\log(p^*)}/\sqrt{n}\right)$$
$$\le I\!P\left(\sup_{\boldsymbol{\upsilon}\in\Upsilon_m} \boldsymbol{\zeta}_{\mathbf{X}}(\boldsymbol{\upsilon}) - \boldsymbol{\zeta}_{\mathbf{X}}(\boldsymbol{\upsilon}^*) \ge \mathtt{C}\sqrt{\mathtt{x}+p^*\log(p^*)}/\sqrt{n}\right)$$
$$\le \sum_{k=1}^{\infty} \exp\Bigg\{p^*[1+\log(2)k + \log(\mathtt{C}_{\boldsymbol{\zeta}} m^{3/2}) - \log(1-1/\sqrt{2})]$$
$$-2^{k-1}\Big[\mathtt{x} + \log(2) + p^*[1+\log(2)+\log(\mathtt{C}_{\boldsymbol{\zeta}} m^{3/2})$$
$$- \log(1-1/\sqrt{2})]\Big]\Bigg\} \le \mathrm{e}^{-\mathtt{x}}.$$

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

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

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

6.A.8 Proof of Lemma 6.3.7

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

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

Lemma 6.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 6.A.5 and is independent of m, n and where $\mathtt{r}^* > 0$ is defined in (6.3.5).

Remark 6.A.6. By the definition of $\mathtt{r}^* > 0$ in (6.3.5) 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 6.A.5 we can prove the following result. It is proved very similarly to Lemma 6.A.18:

Lemma 6.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 6.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 6.3.5 and Lemma 6.A.5

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

Some bounds for the score

Lemma 6.A.10. *We have*

$$\begin{aligned}
|\boldsymbol{f}'_{\boldsymbol{\eta}^*_m}(\boldsymbol{x})| &\le (C_{\|\boldsymbol{f}\|}+1)\sqrt{34}s_{\mathbf{X}}\|\psi'\|_\infty,\\
|\boldsymbol{f}_{\boldsymbol{\eta}}(\mathbf{X}^\top\boldsymbol{\theta}) - \boldsymbol{f}_{\boldsymbol{\eta}^\circ}(\mathbf{X}^\top\boldsymbol{\theta}^\circ)| &\le \mathtt{C}\frac{\|\mathcal{D}(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^\circ)\|\sqrt{m}}{\sqrt{n}}\\
&\quad + \mathtt{C}\left(\frac{\|\mathcal{D}(\boldsymbol{\upsilon}^\circ - \boldsymbol{\upsilon}^*)\|m^2}{\sqrt{n}} + 1\right). \quad (6.A.13)
\end{aligned}$$

Proof. Using assumption $(\mathbf{Cond}_{\boldsymbol{\eta}^*})$, that $|M(j)| \le 17$ (in (6.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 6.3.5 and $m \in \mathbb{N}$ large enough ($m^5/n \to 0$ and $\mathtt{r}^* \cong m$)

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

For the second claim we bound (6.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 6.A.11. *We have with* $\varsigma_{i,m}$ *from* (6.A.1)

$$\|\varsigma_{i,m}(\boldsymbol{\upsilon}_m^*)\| \le (C_{\|\boldsymbol{f}\|}+1)\sqrt{34}s_{\mathbf{X}}\|\psi'\|_\infty + \sqrt{17}\|\psi\|_\infty\sqrt{m},$$

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

$$\begin{aligned}\|\varsigma_{i,m}(\boldsymbol{\upsilon}) - \varsigma_{i,m}(\boldsymbol{\upsilon}')\| \le \sqrt{34}\Big(& s_{\mathbf{X}}\|\psi'\|_\infty m^{3/2} + 2(C_{\|\boldsymbol{f}\|}+1)\sqrt{m}\|\psi''\|_\infty s_{\mathbf{X}} \\ & +2\|\psi'\|_\infty s_{\mathbf{X}} m^{3/2} + \|\psi'\|_\infty C_{\|\boldsymbol{\eta}_m^*\|}\sqrt{2}L_{\nabla\Phi.}\Big)\frac{\|\mathcal{D}_m(\boldsymbol{\upsilon}-\boldsymbol{\upsilon}')\|}{\sqrt{n}c_{\mathcal{D}}}.\end{aligned}$$

Proof. Note

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

Such that with (6.A.8) and Lemma 6.A.10

$$\|\varsigma_{i,m}(\boldsymbol{\upsilon}_m^*)\| \le (C_{\|\boldsymbol{f}\|}+1)\sqrt{34}s_{\mathbf{X}}\|\psi'\|_\infty + \sqrt{17}\|\psi\|_\infty\sqrt{m}. \quad (6.A.14)$$

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

$$\begin{aligned}|N(j)| \overset{\text{def}}{=} \Big|\Big\{ & k \in \{2^j - 2j17 - 1, \ldots, 2^{j+1} - 2(j+1)17 - 1 - 1\}: \quad (6.A.15) \\ & |\boldsymbol{e}_k(\mathbf{X}_i^\top\boldsymbol{\theta}') - \boldsymbol{e}_k(\mathbf{X}_i^\top\boldsymbol{\theta})| \vee |\boldsymbol{e}'_k(\mathbf{X}_i^\top\boldsymbol{\theta}') - \boldsymbol{e}'_k(\mathbf{X}_i^\top\boldsymbol{\theta})| > 0\Big\}\Big| \le 34.\end{aligned}$$

Furthermore we always have that

$$|\boldsymbol{e}'_k(\mathbf{X}_i^\top\boldsymbol{\theta}') - \boldsymbol{e}'_k(\mathbf{X}_i^\top\boldsymbol{\theta})| \le 2^{j_k5/2}\|\psi''\|_\infty s_{\mathbf{X}}\|\boldsymbol{\theta}-\boldsymbol{\theta}'\|.$$

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 (6.A.15)

$$\begin{aligned}
&|\boldsymbol{f}'_{\boldsymbol{\eta}}(\boldsymbol{\theta}^\top \mathbf{X}_i) - \boldsymbol{f}'_{\boldsymbol{\eta}}(\mathbf{X}_i^\top \boldsymbol{\theta}')| \\
&\quad = |\sum_{k=1}^{m} \boldsymbol{\eta}_k (e'_k(\mathbf{X}_i^\top \boldsymbol{\theta}) - e'_k(\mathbf{X}_i^\top \boldsymbol{\theta}'))| \\
&\quad \leq \left(\sum_{j=0}^{j_m - 1} \sum_{k \in N(j)} \eta_k 2^{5j/2} \right) \|\boldsymbol{\theta} - \boldsymbol{\theta}'\| \|\psi''\|_\infty s_{\mathbf{X}} \\
&\quad \leq \left\{ \left(\sum_{j=0}^{j_m - 1} \sum_{k \in N(j)} \eta^*_{mk} 2^{5j/2} \right) + \left(\sum_{j=0}^{j_m - 1} \sum_{k \in N(j)} (\eta_k - \eta^*_{mk}) 2^{5j/2} \right) \right\} \\
&\qquad \|\boldsymbol{\theta} - \boldsymbol{\theta}'\| \|\psi''\|_\infty s_{\mathbf{X}} \\
&\quad \leq \left\{ \sqrt{34} \left(\sum_{k=0}^{m-1} \eta^{*\,2}_{mk} k^{2\alpha} \right)^{\frac{1}{2}} \left(\sum_{j=0}^{j_m - 1} 2^{(5-2\alpha)j} \right)^{\frac{1}{2}} + \frac{\mathtt{r} m^2}{n} \sqrt{m} \right\} \\
&\qquad \|\boldsymbol{\theta} - \boldsymbol{\theta}'\| \|\psi''\|_\infty s_{\mathbf{X}} \\
&\quad \leq \sqrt{34} (C_{\|\boldsymbol{f}\|} + 1) m^{3/2} \|\boldsymbol{\theta} - \boldsymbol{\theta}'\| \|\psi''\|_\infty s_{\mathbf{X}},
\end{aligned} \tag{6.A.16}$$

and with the same arguments

$$\begin{aligned}
\|\boldsymbol{e}(\mathbf{X}_i^\top \boldsymbol{\theta}) - \boldsymbol{e}(\mathbf{X}_i^\top \boldsymbol{\theta}')\| &\leq \left(\sum_{k=1}^{m} |\boldsymbol{e}_k(\mathbf{X}_i^\top \boldsymbol{\theta}) - \boldsymbol{e}_k(\mathbf{X}_i^\top \boldsymbol{\theta}')|^2 \right)^{1/2} \\
&\leq \sqrt{34} \left(\sum_{j=0}^{j_m - 1} 2^{3j} \right)^{1/2} \|\boldsymbol{\theta} - \boldsymbol{\theta}'\| \|\psi'\|_\infty s_{\mathbf{X}} \\
&\leq \sqrt{34} m^{3/2} \|\boldsymbol{\theta} - \boldsymbol{\theta}'\| \|\psi'\|_\infty s_{\mathbf{X}},
\end{aligned} \tag{6.A.17}$$

and

$$\begin{aligned}
\|f'_{\boldsymbol{\eta} - \boldsymbol{\eta}'}(\boldsymbol{\theta}^\top \mathbf{X}_i) \nabla \varphi_{\boldsymbol{\theta}}^\top \mathbf{X}_i\| &\leq s_{\mathbf{X}} \sum_{k=1}^{m} |\eta_k - \eta'_{m,k}| |\boldsymbol{e}'_k(\boldsymbol{\theta}^\top \mathbf{X}_i)| \\
&\leq \sqrt{34} \|\boldsymbol{\eta} - \boldsymbol{\eta}'\| s_{\mathbf{X}} \|\psi'\|_\infty \left(\sum_{j=0}^{j_m - 1} 2^{3j} \right)^{1/2} \\
&\leq \sqrt{34} \|\boldsymbol{\eta} - \boldsymbol{\eta}'\| s_{\mathbf{X}} \|\psi'\|_\infty m^{3/2}.
\end{aligned} \tag{6.A.18}$$

Finally similar to (6.A.14) we have with $M(j) \subset \mathbb{N}$ from (6.A.10)

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

where since $\boldsymbol{\upsilon}' \in \Upsilon_\circ(\mathtt{r})$ and $n \in \mathbb{N}$ large enough ($\mathtt{r}^2 = O(m)$ and $m^5(1 + \mathtt{C}_{bias} \log(n))/n \to 0$)

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

such that

$$\begin{aligned}
&\|\boldsymbol{f}'_{\boldsymbol{\eta}'}(\mathbf{X}_i^\top \boldsymbol{\theta}')(\nabla \varphi_{\boldsymbol{\theta}}^\top - \nabla \varPhi_{\boldsymbol{\theta}'}^\top)\mathbf{X}_i\| \qquad (6.A.19) \\
&\quad \le \|\psi'\|_\infty (C_{\|\boldsymbol{\eta}^*\|} + 1)\sqrt{34} L_{\nabla \varPhi.} \|\boldsymbol{\theta} - \boldsymbol{\theta}'\| s_{\mathbf{X}}.
\end{aligned}$$

We get combining (6.A.19) , (6.A.16), (6.A.18) and (6.A.17)

$$
\begin{aligned}
&\|\varsigma_{i,m}(\boldsymbol{\upsilon}) - \varsigma_{i,m}(\boldsymbol{\upsilon}')\| \\
&\quad = \|f'_{\boldsymbol{\eta}-\boldsymbol{\eta}'}(\boldsymbol{\theta}^\top \mathbf{X}_i)\nabla\varphi_{\boldsymbol{\theta}}^\top \mathbf{X}_i \\
&\qquad + \Big[\boldsymbol{f}'_{\boldsymbol{\eta}'}(\boldsymbol{\theta}^\top \mathbf{X}_i) - \boldsymbol{f}'_{\boldsymbol{\eta}'}(\mathbf{X}_i^\top \boldsymbol{\theta}')\Big]\nabla\varphi_{\boldsymbol{\theta}}^\top \mathbf{X}_i \\
&\qquad + \boldsymbol{f}'_{\boldsymbol{\eta}'}(\mathbf{X}_i^\top \boldsymbol{\theta}')(\nabla\varphi_{\boldsymbol{\theta}}^\top - \nabla\varPhi_{\boldsymbol{\theta}'}^\top)\mathbf{X}_i, \boldsymbol{e}(\mathbf{X}_i^\top \boldsymbol{\theta}) - \boldsymbol{e}(\mathbf{X}_i^\top \boldsymbol{\theta}'))\| \\
&\quad \le \sqrt{34}\|\boldsymbol{\eta} - \boldsymbol{\eta}'\| s_{\mathbf{X}}\|\psi'\|_\infty m^{3/2} \\
&\qquad + \sqrt{34}(C_{\|\boldsymbol{f}\|} + 1)\sqrt{m}\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|\|\psi''\|_\infty s_{\mathbf{X}} \\
&\qquad + \sqrt{34} m^{3/2}\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|\|\psi'\|_\infty s_{\mathbf{X}} \\
&\qquad + \|\psi'\|_\infty (C_{\|\boldsymbol{\eta}^*\|} + 1)\sqrt{34} L_{\nabla\varPhi.}\|\boldsymbol{\theta} - \boldsymbol{\theta}'\| \\
&\quad \le \sqrt{34}(s_{\mathbf{X}}\|\psi'\|_\infty m^{3/2} + 2(C_{\|\boldsymbol{f}\|} + 1)\sqrt{m}\|\psi''\|_\infty s_{\mathbf{X}} \\
&\qquad + \|\psi'\|_\infty s_{\mathbf{X}} m^{3/2} + \|\psi'\|_\infty C_{\|\boldsymbol{\eta}_m^*\|}\sqrt{2} L_{\nabla\varPhi.}\Big)\frac{2\|\mathcal{D}_m(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}')\|}{\sqrt{n} c_{\mathcal{D}}},
\end{aligned}
$$

where we used Lemma 6.A.8 in the last step to find that

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

□

Crude deviation bounds for sums of random matrices

The next auxiliary Lemma relies on a non-commutative Bernstein inequality; see Theorem 1.4 of [56].

Lemma 6.A.12. *Suppose that $\boldsymbol{v}_i \in \mathbb{R}^{p_1}$ are iid random vectors, where $p \in \mathbb{N}$. Define*

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

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

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

Proof. This lemma is an immediate consequence of the non-commutative Bernstein inequality (Theorem 1.4 in [56]). We only have to note that

$$\sum_{i=1}^{n} I\!\!E[\mathbf{M}_i^2] \le 2nI\!\!E[\|\boldsymbol{v}_1\|^4] = 2nB^2.$$

□

Lemma 6.A.13. *We have with* $\mathtt{x} \le 9n/2 - \log(2m)$ *that*

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

where with $\varsigma_{i,m}$ *from* (6.A.1)

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

Proof. We want to employ lemma 6.A.12. We estimate using Lemma 6.A.11

$$\|\varsigma_{i,m}(\boldsymbol{v}_m^*)\varsigma_{i,m}(\boldsymbol{v}_m^*)^\top\| \le 34\Big((C_{\|\boldsymbol{f}\|}+1)\sqrt{2}s_{\mathbf{X}}\|\psi'\|_\infty + \|\psi\|_\infty\Big)^2 m =: C_M m,$$

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

$$I\!\!E[\|\varsigma_{i,m}(\boldsymbol{v}_m^*)\|^4] \le C_M^2 m^2.$$

Plugging these bounds into Lemma 6.A.12 we get

$$I\!\!P(\|\mathbf{S}_n\| \ge n^{-1}t) \le 2m\exp\Big\{-\frac{t^2}{4nC_M^2m^2 + 2C_Mmt/3}\Big\}.$$

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

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

□

Lemma 6.A.14. *We have with* $\mathtt{x} \le 9n/2 - \log(2m)$ *that*

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

where with $\varsigma_{i,m}$ *from* (6.A.1)

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

Proof. We want to employ lemma 6.A.12. We estimate using Lemma 6.A.11 and that $\mathtt{r}^{\circ} \le \mathtt{C}\sqrt{p^* + \mathtt{x}}$

$$\begin{aligned}\|\varsigma_{i,m}(\boldsymbol{\upsilon})\varsigma_{i,m}(\boldsymbol{\upsilon})^{\top}\| &\le 3\|\varsigma_{i,m}(\boldsymbol{\upsilon}^*_m)\|^2 + 3\|\varsigma_{i,m}(\boldsymbol{\upsilon}^*_m) - \varsigma_{i,m}(\boldsymbol{\upsilon})\|^2 \\ &\le C_M m + \mathtt{C}\frac{\|\mathcal{D}_m(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^*_m)\|^2 m^3}{n} \\ &\le C_M m + \mathtt{C} m^3 \mathtt{r}^2/n.\end{aligned}$$

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

$$I\!\!E[\|\varsigma_{i,m}(\boldsymbol{\upsilon}^*_m)\|^4] \le \mathtt{C}^2(m^2 + m^6\mathtt{r}^4/n^2).$$

Plugging these bounds into Lemma 6.A.12 we get

$$\begin{aligned}&I\!\!P(\|\mathbf{S}_n\| \ge n^{-1}t) \\ &\quad \le 2m\exp\Big\{-\frac{t^2}{4n\mathtt{C}^2(m^2 + m^6\mathtt{r}^4/n^2) + 2\mathtt{C}\,(m + m^3\mathtt{r}^2/n)\,t/3}\Big\}.\end{aligned}$$

Setting $t = \sqrt{8n}\mathtt{C}\left(m + m^3\mathtt{r}^2/n\right)\Big(\mathtt{x} + \log(2m)\Big)^{1/2}/n^2$ and $\mathtt{x} \le 9n/2 - \log(2m)$ this implies

$$I\!\!P\left(\|\mathbf{S}_n\| \ge \sqrt{8}\mathtt{C}\left(m + m^3\mathtt{r}^2/n\right)\Big(\mathtt{x} + \log(2m)\Big)^{1/2}/\sqrt{n}\right) \le \mathrm{e}^{-\mathtt{x}}.$$

□

Lemma 6.A.15. *We have with*

$$t = C_M^2\|\mathcal{D}_m(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}')\|^2\sqrt{5/n}m^3\Big(\mathtt{x} + \log(2m)\Big)^{1/2},$$

and $\mathtt{x} \le 9n/2 - \log(2m)$

$$I\!\!P(\|\mathbf{S}_n\| \ge n^{-1}t) \le \mathrm{e}^{-\mathtt{x}},$$

where with $\boldsymbol{\upsilon} \in \Upsilon_{\circ}(\mathtt{r})$ *and with* $\varsigma_{i,m}$ *from* (6.A.1)

$$\begin{aligned}\mathbf{S}_n &= \frac{1}{n}\sum_{i=1}^{n}(\varsigma_{i,m}(\boldsymbol{\upsilon}') - \varsigma_{i,m}(\boldsymbol{\upsilon}))(\varsigma_{i,m}(\boldsymbol{\upsilon}') - \varsigma_{i,m}(\boldsymbol{\upsilon}))^{\top} \\ &\quad - I\!\!E(\varsigma_{i,m}(\boldsymbol{\upsilon}') - \varsigma_{i,m}(\boldsymbol{\upsilon}))(\varsigma_{i,m}(\boldsymbol{\upsilon}') - \varsigma_{i,m}(\boldsymbol{\upsilon}))^{\top} \\ C_M &= \sqrt{34}\Big(s_{\mathbf{X}}\|\psi'\|_{\infty} + 3(C_{\|\boldsymbol{f}\|} + 1)\|\psi''\|_{\infty}s_{\mathbf{X}} \\ &\quad + 3\|\psi'\|_{\infty}s_{\mathbf{X}} + \|\psi'\|_{\infty}C_{\|\boldsymbol{\eta}'\|}\sqrt{2}L_{\nabla\Phi.}\Big)\frac{2}{c_{\mathcal{D}}}.\end{aligned}$$

Proof. We estimate using Lemma 6.A.11

$$\begin{aligned}
&\|(\varsigma_{i,m}(\boldsymbol{\upsilon}') - \varsigma_{i,m}(\boldsymbol{\upsilon}))(\varsigma_{i,m}(\boldsymbol{\upsilon}') - \varsigma_{i,m}(\boldsymbol{\upsilon}))^{\top}\| \\
&\quad \le \|\varsigma_{i,m}(\boldsymbol{\upsilon}') - \varsigma_{i,m}(\boldsymbol{\upsilon})\|^2 \\
&\quad \le 34\Big(s_{\mathbf{X}}\|\psi'\|_\infty + 3(C_{\|\boldsymbol{f}\|}+1)\|\psi''\|_\infty s_{\mathbf{X}} \\
&\qquad\quad +3\|\psi'\|_\infty s_{\mathbf{X}} + \|\psi'\|_\infty C_{\|\boldsymbol{\eta}'\|}\sqrt{2}L_{\nabla\varPhi.}\Big)\frac{4\|\mathcal{D}_m(\boldsymbol{\upsilon}-\boldsymbol{\upsilon}')\|^2 m^3}{n c_{\mathcal{D}}^2} \\
&\quad =: C_M^2\frac{\|\mathcal{D}_m(\boldsymbol{\upsilon}-\boldsymbol{\upsilon}')\|^2 m^3}{n}.
\end{aligned}$$

With the same estimates we obtain

$$\mathbb{E}[\|\varsigma_{i,m}(\boldsymbol{\upsilon}') - \varsigma_{i,m}(\boldsymbol{\upsilon})\|^4] \le C_M^4 m^6\frac{\|\mathcal{D}_m(\boldsymbol{\upsilon}-\boldsymbol{\upsilon}')\|^4}{n^2}.$$

Plugging these bounds into Lemma 6.A.12 we get with $d(\boldsymbol{\upsilon},\boldsymbol{\upsilon}') \overset{\mathrm{def}}{=} \|\mathcal{D}_m(\boldsymbol{\upsilon}-\boldsymbol{\upsilon}')\|$

$$\begin{aligned}
&\mathbb{P}(\|\mathbf{S}_n\| \ge n^{-1}t) \\
&\quad \le 2m\exp\Big\{-\frac{t^2}{4d(\boldsymbol{\upsilon},\boldsymbol{\upsilon}')^4 C_M^4 n^{-1}m^6 + 2d(\boldsymbol{\upsilon},\boldsymbol{\upsilon}')^2 C_M^2 m^3 n^{-1}t/3}\Big\}.
\end{aligned}$$

Setting $t = C_M^2\|\mathcal{D}_m(\boldsymbol{\upsilon}-\boldsymbol{\upsilon}')\|^2\sqrt{8/n}m^3\Big(\mathtt{x}+\log(2m)\Big)^{1/2}$ and $\mathtt{x} \le 9n/2 - \log(2m)$ this yields

$$\mathbb{P}(\|\mathbf{S}_n\| \ge n^{-1}t) \le \mathrm{e}^{-\mathtt{x}}.$$

□

Conditions $(\mathcal{E}\mathfrak{D}_0)$, $(\mathcal{E}\mathrm{r})$ and $(\mathcal{E}\mathfrak{D}_{1,m})$

Lemma 6.A.16. *With probability greater than* $1-3\mathrm{e}^{-\mathtt{x}}$ *we have* $(\mathcal{E}\mathfrak{D}_0)$ *with*

$$\begin{aligned}
g &= \sqrt{n}\sigma^{-1}c_{\mathcal{D}}\widetilde{g}\Big((C_{\|\boldsymbol{\eta}^*\|}+1)\sqrt{34}s_{\mathbf{X}}\|\psi'\|_\infty + \sqrt{17}\|\psi\|_\infty\sqrt{m}\Big)^{-1}, \\
\nu_m^2 &= 2\widetilde{\nu}^2\sigma^2,
\end{aligned}$$

and $(\boldsymbol{\mathcal{E}\mathtt{r}})$ *with*

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

where $\mathtt{C}_{(\boldsymbol{\mathcal{E}\mathtt{r}})} > 0$ *is independent of* $n, m, \mathtt{x}, \mathtt{r}$ *.*

Proof. Lemma 6.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 6.A.11 - we get with $\mu \le \sqrt{n}\sigma^{-1} c_{\mathcal{D}} \widetilde{g}\Big((C_{\|\boldsymbol{\eta}^*\|} + 1)\sqrt{34} s_{\mathbf{X}} \|\psi'\|_\infty + \sqrt{17}\|\psi\|_\infty \sqrt{m}\Big)^{-1}$, with $\varsigma_{i,m}$ from (6.A.1) and assumption $(\mathbf{Cond}_\varepsilon)$

$$
\begin{aligned}
&\sup_{\gamma\in\mathbb{R}^{p^*}} \log \mathbb{E}_\varepsilon \exp\left\{\mu \frac{\langle \nabla\zeta(\boldsymbol{\upsilon}_m^*), \gamma\rangle}{\|\mathcal{V}_m(\boldsymbol{\upsilon}_m^*)\gamma\|}\right\} \\
&\quad\le \sum_{i=1}^n \sup_{\gamma\in\mathbb{R}^{p^*},\, \|\widetilde{\gamma}\|=1} \log \mathbb{E} \exp\left\{\mu \langle \widetilde{\gamma}, \mathcal{V}_m^{-1}(\boldsymbol{\upsilon}_m^*) \varsigma_{i,m}(\boldsymbol{\upsilon}_m^*)\rangle \varepsilon_i\right\} \\
&\quad\le \widetilde{\nu}^2 \mu^2 \widetilde{\gamma}^\top \mathcal{V}_m^{-1}(\boldsymbol{\upsilon}_m^*) \left(\sum_{i=1}^n \varsigma_{i,m}(\boldsymbol{\upsilon}_m^*) \varsigma_{i,m}(\boldsymbol{\upsilon}_m^*)^\top\right) \mathcal{V}_m^{-1}(\boldsymbol{\upsilon}^*) \widetilde{\gamma} \\
&\quad= \widetilde{\nu}^2 \mu^2 + \widetilde{\nu}^2 \mu^2 \widetilde{\gamma}^\top \mathcal{V}_m^{-1}(\boldsymbol{\upsilon}_m^*) n \mathbf{S}_n \mathcal{V}_m^{-1}(\boldsymbol{\upsilon}_m^*) \widetilde{\gamma} \\
&\quad\le \widetilde{\nu}^2 \mu^2 + \widetilde{\nu}^2 \mu^2 \varkappa_n,
\end{aligned} \tag{6.A.20}
$$

where

$$
\begin{aligned}
\varkappa_n &= \widetilde{\gamma}^\top \left(n^{-1}\mathcal{V}_m\right)^{-1/2} \mathbf{S}_n \left(n^{-1}\mathcal{V}_m\right)^{-1/2} \widetilde{\gamma}, \\
\mathbf{S}_n &= \frac{1}{n}\sum_{i=1}^n \varsigma_{i,m}(\boldsymbol{\upsilon}_m^*) \varsigma_{i,m}(\boldsymbol{\upsilon}_m^*)^\top - \frac{1}{n}\mathcal{V}_m(\boldsymbol{\upsilon}_m^*).
\end{aligned}
$$

With Lemma 6.A.13 we infer that if $\mathtt{x} \le 9n/2 - \log(2m)$

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

Consequently with probability greater than $1-\mathrm{e}^{-\mathtt{x}}$ we find that for $n \in \mathbb{N}$ large enough

$$\varkappa_n \le \frac{\mathtt{C}_M\sqrt{8}m\Big(\mathtt{x}+\log(2m)\Big)^{1/2}}{\sqrt{n}\sigma^2 c_{\mathfrak{D}}^2} \le 1.$$

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

$$\begin{aligned} g &= \sqrt{n}c_{\mathfrak{D}}\widetilde{g}\Big((C_{\|\boldsymbol{\eta}^*\|}+1)\sqrt{34}s_{\mathbf{X}}\|\psi'\|_\infty + \sqrt{17}\|\psi\|_\infty\sqrt{m}\Big)^{-1}, \\ \nu_m^2 &= 2\widetilde{\nu}^2. \end{aligned}$$

Concerning $(\mathcal{E}\mathtt{r})$ we bound using the same arguments as in the proof of Lemma 6.A.18

$$\begin{aligned} \|\mathcal{V}_m(\boldsymbol{\upsilon})^{-1}\mathcal{V}_m(\boldsymbol{\upsilon}_m^*)\|^2 &\le 1+\|I-\mathcal{V}_m(\boldsymbol{\upsilon})^{-1}\mathcal{V}_m(\boldsymbol{\upsilon}_m^*)^2\mathcal{V}_m(\boldsymbol{\upsilon})^{-1}\| \\ &\le 1+\mathtt{C}\left(m^{3/2}+\mathtt{r}m^2/\sqrt{n}\right)\mathtt{r}/\sqrt{n}. \end{aligned}$$

Thus we get with the arguments from above $(\mathcal{E}\mathtt{r})$ using Lemma 6.A.14 with probability greater than $1-\mathrm{e}^{-\mathtt{x}}$ and

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

□

Lemma 6.A.17. *With probability greater than* $1-\mathrm{e}^{-\mathtt{x}}$ *we have* $(\mathcal{E}\mathfrak{D}_1)$ *with*

$$\begin{aligned} \mathtt{g} &\overset{\text{def}}{=} \sqrt{n}c_{\mathfrak{D}}\mathtt{r}m^{-3/2}C_{(\mathcal{E}\mathfrak{D}_1)}^{-1}, \\ \omega &\overset{\text{def}}{=} \frac{2}{\sqrt{n}c_{\mathfrak{D}}}, \\ \nu_{1,m}^2 &= \widetilde{\nu}^2\mathtt{C}_{(\mathcal{E}\mathfrak{D}_1)}m^2, \end{aligned}$$

where

$$\begin{aligned} C_{(\mathcal{E}\mathfrak{D}_1)} &= \sqrt{34}\Big(s_{\mathbf{X}}\|\psi'\|_\infty + 3(C_{\|\boldsymbol{f}\|}+1)\|\psi''\|_\infty s_{\mathbf{X}} + 3\|\psi'\|_\infty s_{\mathbf{X}} \\ &\qquad +\|\psi'\|_\infty C_{\|\boldsymbol{\eta}_m^*\|}\sqrt{2}L_{\nabla\Phi.}\Big). \end{aligned}$$

Proof. We get with Lemma 6.A.11, with Lemma 6.A.8 and with $\varsigma_{i,m}$ from (6.A.1)

$$\begin{aligned}
&\|\mathcal{D}_m^{-1}(\varsigma_{i,m}(\boldsymbol{\upsilon}) - \varsigma_{i,m}(\boldsymbol{\upsilon}'))\| \\
&\quad \le \frac{\sqrt{34}}{\sqrt{n}c_{\mathcal{D}}}\sum_{i=1}^{n}\varepsilon_i\Big(s_{\mathbf{X}}\|\psi'\|_{\infty}m^{3/2} + 3(C_{\|\boldsymbol{f}\|}+1)\sqrt{m}\|\psi''\|_{\infty}s_{\mathbf{X}} \\
&\qquad +3\|\psi'\|_{\infty}s_{\mathbf{X}}m^{3/2} + \|\psi'\|_{\infty}C_{\|\boldsymbol{\eta}_m^*\|}\sqrt{2}L_{\nabla\Phi.}\Big)\frac{2\mathtt{r}}{\sqrt{n}c_{\mathcal{D}}} \\
&\quad \overset{\text{def}}{=} C_{(\boldsymbol{\mathcal{E}\mathfrak{D}_1})}\frac{2m^{3/2}}{nc_{\mathcal{D}}^2}\|\mathcal{D}_m(\boldsymbol{\upsilon}-\boldsymbol{\upsilon}')\|,
\end{aligned}$$

We get with,

$$\begin{aligned}
\mu &\le \mathtt{g} \overset{\text{def}}{=} \sqrt{n}c_{\mathcal{D}}(\mathtt{r}m)^{-3/2}C_{(\boldsymbol{\mathcal{E}\mathfrak{D}_1})}^{-1} \\
\omega &\overset{\text{def}}{=} \frac{2}{\sqrt{n}c_{\mathcal{D}}},
\end{aligned}$$

and the same calculations as in (6.A.20) with some $\boldsymbol{\upsilon},\boldsymbol{\upsilon}' \in \varUpsilon_{\circ}(\mathtt{r})$, $\boldsymbol{\gamma}\in\mathbb{R}^{p^*}$ and $\|\boldsymbol{\gamma}\| = 1$

$$\begin{aligned}
&\log \mathbb{E}_{\varepsilon}[\exp\left\{\mu\frac{\boldsymbol{\gamma}^{\top}\mathcal{D}_m^{-1}\left(\nabla\zeta(\boldsymbol{\upsilon}) - \nabla\zeta(\boldsymbol{\upsilon}')\right)}{\omega\|\mathcal{D}_m(\boldsymbol{\upsilon}-\boldsymbol{\upsilon}')\|}\right\}] \\
&\quad \le \sum_{i=1}^{n}\log\mathbb{E}_{\varepsilon}[\exp\left\{\mu\varepsilon_i\frac{\gamma^{\top}\mathcal{D}_m^{-1}(\varsigma_{i,m}(\boldsymbol{\upsilon}) - \varsigma_{i,m}(\boldsymbol{\upsilon}'))}{\omega\|\mathcal{D}_m(\boldsymbol{\upsilon}-\boldsymbol{\upsilon}')\|}\right\}] \\
&\quad \le \frac{\mu^2\widetilde{\nu}^2}{2}(\omega\|\mathcal{D}_m(\boldsymbol{\upsilon}-\boldsymbol{\upsilon}')\|)^{-2} \\
&\qquad n\gamma^{\top}\mathcal{D}_m^{-1}\left(\sum_{i=1}^{n}(\varsigma_{i,m}(\boldsymbol{\upsilon}') - \varsigma_{i,m}(\boldsymbol{\upsilon}))(\varsigma_{i,m}(\boldsymbol{\upsilon}') - \varsigma_{i,m}(\boldsymbol{\upsilon}))^{\top}\right)\mathcal{D}_m^{-1}\gamma^{\top}.
\end{aligned}$$

We estimate

$$\begin{aligned}
&\widetilde{\gamma}^{\top}\mathcal{D}_m^{-1}\left(\sum_{i=1}^{n}(\varsigma_{i,m}(\boldsymbol{\upsilon}') - \varsigma_{i,m}(\boldsymbol{\upsilon}))(\varsigma_{i,m}(\boldsymbol{\upsilon}') - \varsigma_{i,m}(\boldsymbol{\upsilon}))^{\top}\right)\mathcal{D}_m^{-1}\widetilde{\gamma}^{\top} \\
&\quad \le \|\mathcal{D}_m^{-1}n\mathbb{E}\left[(\varsigma_{i,m}(\boldsymbol{\upsilon}') - \varsigma_{i,m}(\boldsymbol{\upsilon}))(\varsigma_{i,m}(\boldsymbol{\upsilon}') - \varsigma_{i,m}(\boldsymbol{\upsilon}))^{\top}\right]\mathcal{D}_m^{-1}\| \\
&\qquad +\varkappa_n \\
&\quad \le \mathbb{E}\|\left(n^{-1/2}\mathcal{D}_m\right)^{-1}(\varsigma_{i,m}(\boldsymbol{\upsilon}') - \varsigma_{i,m}(\boldsymbol{\upsilon}))\|^2 + \varkappa_n,
\end{aligned}$$

where

$$\begin{aligned}
\varkappa_n &= \| \left(n^{-1/2}\mathcal{D}_m\right)^{-1} \mathbf{S}_n \left(n^{-1/2}\mathcal{D}_m\right)^{-1} \|, \\
\mathbf{S}_n &= \frac{1}{n}\sum_{i=1}^{n}(\varsigma_{i,m}(\boldsymbol{\upsilon}') - \varsigma_{i,m}(\boldsymbol{\upsilon}))(\varsigma_{i,m}(\boldsymbol{\upsilon}') - \varsigma_{i,m}(\boldsymbol{\upsilon}))^\top \\
&\quad -I\!\!E(\varsigma_{i,m}(\boldsymbol{\upsilon}') - \varsigma_{i,m}(\boldsymbol{\upsilon}))(\varsigma_{i,m}(\boldsymbol{\upsilon}') - \varsigma_{i,m}(\boldsymbol{\upsilon}))^\top.
\end{aligned}$$

To controll $\varkappa_n > 0$ we apply Lemma 6.A.15 and we infer that with $t = C_M^2\|\mathcal{D}_m(\boldsymbol{\upsilon}-\boldsymbol{\upsilon}')\|^2\sqrt{5/n}m^3\Big(\mathtt{x}+\log(2m)\Big)^{1/2}$ and $\mathtt{x} \le 9n/2 - \log(2m)$ the set $\{\|\mathbf{S}_n\| \le n^{-1}t\}$ is of dominating probability and on this set we find

$$\begin{aligned}
\varkappa_n &\le \frac{C_M^2\|\mathcal{D}_m(\boldsymbol{\upsilon}-\boldsymbol{\upsilon}')\|^2\Big(\mathtt{x}+\log(2m)\Big)^{1/2}m^3\sqrt{5/n}}{nc_{\mathcal{D}}^2} \\
&\le \omega^2\|\mathcal{D}_m(\boldsymbol{\upsilon}-\boldsymbol{\upsilon}')\|^2\frac{C_M^2\sqrt{5}m^3\Big(\mathtt{x}+\log(2m)\Big)^{1/2}}{\sqrt{n}}.
\end{aligned}$$

For $\mathtt{r} \le \mathtt{r}_0 \le \mathtt{C}_{\mathtt{r}}\sqrt{p^*+\mathtt{x}}$ this gives because $m^{5/2}/\sqrt{n} \to 0$

$$\varkappa_n \le \mathtt{C}_{\varkappa}\sqrt{\Big(\mathtt{x}+\log(2m)\Big)p^*}.$$

We calculate with some $(\boldsymbol{\theta}^\circ, \boldsymbol{\eta}^\circ) \overset{\text{def}}{=} (\frac{1}{\sqrt{n}}\mathcal{D}_m)^{-1}\boldsymbol{\gamma}$

$$\begin{aligned}
&n\boldsymbol{\gamma}^\top\mathcal{D}_m^{-1}I\!\!E\left[(\varsigma_{1,m}(\boldsymbol{\upsilon}') - \varsigma_{1,m}(\boldsymbol{\upsilon}))(\varsigma_{1,m}(\boldsymbol{\upsilon}') - \varsigma_{1,m}(\boldsymbol{\upsilon}))^\top\right]\mathcal{D}_m^{-1}\boldsymbol{\gamma}^\top \\
&= I\!\!E\Bigg[\Big\{\left[\boldsymbol{f}_{\boldsymbol{\eta}}'(\mathbf{X}^\top\boldsymbol{\theta}) - \boldsymbol{f}_{\boldsymbol{\eta}'}'(\mathbf{X}^\top\boldsymbol{\theta}')\right](\mathbf{X}^\top\boldsymbol{\theta}^\circ)^2 \\
&\qquad + \boldsymbol{f}_{\boldsymbol{\eta}^\circ}(\mathbf{X}^\top\boldsymbol{\theta}) - \boldsymbol{f}_{\boldsymbol{\eta}^\circ}(\mathbf{X}^\top\boldsymbol{\theta}')\Big\}^2\Bigg] \\
&\le 3I\!\!E\left[\Big\{\left[\boldsymbol{f}_{\boldsymbol{\eta}}'(\mathbf{X}^\top\boldsymbol{\theta}) - \boldsymbol{f}_{\boldsymbol{\eta}'}'(\mathbf{X}^\top\boldsymbol{\theta}')\right](\mathbf{X}^\top\boldsymbol{\theta}^\circ)^2\Big\}^2\right] \\
&\quad +3I\!\!E\left[\Big\{\boldsymbol{f}_{\boldsymbol{\eta}^\circ}(\mathbf{X}^\top\boldsymbol{\theta}) - \boldsymbol{f}_{\boldsymbol{\eta}^\circ}(\mathbf{X}^\top\boldsymbol{\theta}')\Big\}^2\right].
\end{aligned}$$

We estimate separately

$$\begin{aligned}
&I\!\!E\left[\Big\{\left[\boldsymbol{f}_{\boldsymbol{\eta}}'(\mathbf{X}^\top\boldsymbol{\theta}) - \boldsymbol{f}_{\boldsymbol{\eta}'}'(\mathbf{X}^\top\boldsymbol{\theta}')\right](\mathbf{X}^\top\boldsymbol{\theta}^\circ)^2\Big\}^2\right] \\
&\quad \le 3s_{\mathbf{X}}^4\left(I\!\!E\left[\Big\{\boldsymbol{f}_{\boldsymbol{\eta}-\boldsymbol{\eta}'}'(\mathbf{X}^\top\boldsymbol{\theta})\Big\}^2\right] + I\!\!E\left[\Big\{\boldsymbol{f}_{\boldsymbol{\eta}'}'(\mathbf{X}^\top\boldsymbol{\theta}) - \boldsymbol{f}_{\boldsymbol{\eta}'}'(\mathbf{X}^\top\boldsymbol{\theta}')\Big\}^2\right]\right).
\end{aligned}$$

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 $l = 2^{j_l} + j_l 17 - 1 + r_l \in \mathbb{N}$ and $k = (2^{j_k} - 1)17 + r_k \in \mathbb{N}$ using (6.A.4)

$$\mathbb{E}[\boldsymbol{e}'_k\boldsymbol{e}'_l(\mathbf{X}^\top\boldsymbol{\theta})] \le 17\mathtt{C}_{p_\mathbf{X}}2^{j_k}\|\psi'\|_\infty^2 2^{j_l}1_{I_k\cap I_l\neq 0}. \tag{6.A.21}$$

This implies

$$\begin{aligned}
&\frac{1}{\|\boldsymbol{\eta} - \boldsymbol{\eta}'\|^2}\mathbb{E}\left[\left\{\boldsymbol{f}'_{\boldsymbol{\eta}-\boldsymbol{\eta}'}(\mathbf{X}^\top\boldsymbol{\theta})\right\}^2\right] \\
&\quad= \sum_{k=1}^{m}\sum_{l=k}^{m}(1 - 1_{k=l}/2)\gamma_l\gamma_k\mathbb{E}[\boldsymbol{e}'_l\boldsymbol{e}'_l(\mathbf{X}^\top\boldsymbol{\theta})] \\
&\quad\le 17\mathtt{C}_{p_\mathbf{X}}\|\psi'\|_\infty\sum_{k=0}^{m}\gamma_k 2^{j_k} \\
&\qquad\qquad\qquad\qquad \left(\sum_{j=j_k}^{j_m}\sum_{r=0}^{2^j+33} 2^{2j_l}1_{I_k\cap I_l\neq 0}(2^j - 2j17 - 1 + r, k)\right)^{1/2} \\
&\quad\le 17\mathtt{C}_{p_\mathbf{X}}\|\psi'\|_\infty\sum_{k=0}^{m}\gamma_k 2^{j_k}\left(\sum_{j=j_k}^{j_m}2^{2j_l}\ulcorner 2^{(j_l-j_k)}17\urcorner\right)^{1/2} \\
&\quad\le \sqrt{18}17\mathtt{C}_{p_\mathbf{X}}\|\psi'\|_\infty\sum_{k=0}^{m}\gamma_k 2^{j_k/2}\left(\sum_{j=j_k}^{j_m}2^{3j_l}\right)^{1/2} \\
&\quad\le 18^2\mathtt{C}_{p_\mathbf{X}}m^2.
\end{aligned} \tag{6.A.22}$$

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 (6.A.6) this gives

$$
\begin{aligned}
&I\!\!E\left[\left\{\boldsymbol{f}'_{\boldsymbol{\eta}'}(\mathbf{X}^\top\boldsymbol{\theta})-\boldsymbol{f}'_{\boldsymbol{\eta}'}(\mathbf{X}^\top\boldsymbol{\theta}')\right\}^2\right]\\
&\quad\le\|\boldsymbol{\theta}-\boldsymbol{\theta}'\|^2\mathtt{C}_{p_\mathbf{X}}\|\psi''\|_\infty^2 s_\mathbf{X}^4 17^2 2\sum_{k=1}^m\eta'_k 2^{2j_k}\sum_{l=k}^m\eta'_l 2^{2j_l}1_{\{I_k\cap I_l\neq\emptyset\}}\\
&\quad\le\|\boldsymbol{\theta}-\boldsymbol{\theta}'\|^2\mathtt{C}_{p_\mathbf{X}}\|\psi''\|_\infty^2 s_\mathbf{X}^4 17^2 2\\
&\quad\quad\sum_{k=1}^m\eta'_k 2^{2j_k}\left(\sum_{l=k}^m\eta_l'^2 k^4\right)^{1/2}\left(\sum_{l=k}^m 1_{\{I_k\cap I_l\neq\emptyset\}}\right)^{1/2}.
\end{aligned}
$$

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

$$
\begin{aligned}
\left(\sum_{l=k}^m\eta_l'^2 k^4\right)^{1/2}&\le\left(\sum_{l=k}^m\eta_{ml}^{*\,2}k^4\right)^{1/2}+\left(\sum_{l=k}^m|\eta'_l-\eta^*_{ml}|^2k^4\right)^{1/2}\\
&\le 2(1-\mathtt{C}_{\|\boldsymbol{\eta}^*\|}),
\end{aligned}
$$

which gives using (6.3.4)

$$
\begin{aligned}
&I\!\!E\left[\left\{\boldsymbol{f}'_{\boldsymbol{\eta}'}(\mathbf{X}^\top\boldsymbol{\theta})-\boldsymbol{f}'_{\boldsymbol{\eta}'}(\mathbf{X}^\top\boldsymbol{\theta}')\right\}^2\right]\\
&\quad\le\|\boldsymbol{\theta}-\boldsymbol{\theta}'\|^2(1-\mathtt{C}_{\|\boldsymbol{\eta}^*\|})\mathtt{C}_{p_\mathbf{X}}\|\psi''\|_\infty^2 s_\mathbf{X}^4 17^{5/2}4m\sum_{k=1}^m\eta'_k 2^{3j_k/2}.
\end{aligned}
$$

Repeating the same arguments gives

$$
\begin{aligned}
&I\!\!E\left[\left\{\boldsymbol{f}'_{\boldsymbol{\eta}'}(\mathbf{X}^\top\boldsymbol{\theta})-\boldsymbol{f}'_{\boldsymbol{\eta}'}(\mathbf{X}^\top\boldsymbol{\theta}')\right\}^2\right]\\
&\quad\le\|\boldsymbol{\theta}-\boldsymbol{\theta}'\|^2(1-\mathtt{C}_{\|\boldsymbol{\eta}^*\|})\mathtt{C}_{p_\mathbf{X}}\|\psi''\|_\infty^2 s_\mathbf{X}^4 17^3 4m^{3/2},
\end{aligned}
$$

such that

$$
I\!\!E\left[\left\{\left[\boldsymbol{f}'_{\boldsymbol{\eta}}(\mathbf{X}^\top\boldsymbol{\theta})-\boldsymbol{f}'_{\boldsymbol{\eta}'}(\mathbf{X}^\top\boldsymbol{\theta}')\right](\mathbf{X}^\top\boldsymbol{\theta}^\circ)^2\right\}^2\right]\le\mathtt{C}m^2\|\boldsymbol{\upsilon}-\boldsymbol{\upsilon}'\|^2\text{(6.A.23)}
$$

Finally we can estimate

$$\begin{aligned}
&I\!\!E\left[\left\{\boldsymbol{f}_{\boldsymbol{\eta}^\circ}(\mathbf{X}^\top\boldsymbol{\theta}) - \boldsymbol{f}_{\boldsymbol{\eta}^\circ}(\mathbf{X}^\top\boldsymbol{\theta}')\right\}^2\right] \\
&\quad = 2\sum_{k=1}^m\sum_{l=k}^m (1-1_{k=l}/2)\eta_k^\circ\eta_l^\circ \\
&\qquad\qquad\qquad I\!\!E\left[(\boldsymbol{e}_k(\mathbf{X}^\top\boldsymbol{\theta}) - \boldsymbol{e}_k(\mathbf{X}^\top\boldsymbol{\theta}'))(\boldsymbol{e}_l(\mathbf{X}^\top\boldsymbol{\theta}) - \boldsymbol{e}_l(\mathbf{X}^\top\boldsymbol{\theta}'))\right].
\end{aligned}$$

Using (6.A.5) and very similar arguments as before additionally using that $\|\boldsymbol{\eta}^\circ\| \le 1/c_{\mathfrak{D}}$

$$\begin{aligned}
&I\!\!E\left[\left\{\boldsymbol{f}_{\boldsymbol{\eta}^\circ}(\mathbf{X}^\top\boldsymbol{\theta}) - \boldsymbol{f}_{\boldsymbol{\eta}^\circ}(\mathbf{X}^\top\boldsymbol{\theta}')\right\}^2\right] \\
&\quad \le \|\boldsymbol{\theta}-\boldsymbol{\theta}'\|^2 \mathtt{C}_{p_{\mathbf{X}}}\|\psi'\|_\infty^2 s_{\mathbf{X}}^4 17^2 \sum_{k=1}^m \eta_k^\circ 2^{j_k}\sum_{l=k}^m \eta_l^\circ 2^{j_l} 1_{\{I_k\cap I_l\neq\emptyset\}} \\
&\quad \le \|\boldsymbol{\theta}-\boldsymbol{\theta}'\|^2 \mathtt{C}_{p_{\mathbf{X}}}\|\psi'\|_\infty^2 s_{\mathbf{X}}^4 17^{5/2} 4m^{3/2}\frac{1}{c_{\mathfrak{D}}}\sum_{k=1}^m 2^{j_k/2}\eta_k^\circ \\
&\quad \le \|\boldsymbol{\theta}-\boldsymbol{\theta}'\|^2 \mathtt{C}_{p_{\mathbf{X}}}\|\psi'\|_\infty^2 s_{\mathbf{X}}^4 17^{5/2} 4m^2\frac{1}{c_{\mathfrak{D}}^2}. \qquad (6.A.24)
\end{aligned}$$

Putting these bounds together gives

$$\begin{aligned}
&n\boldsymbol{\gamma}^\top \mathcal{D}_m^{-1} I\!\!E\left[(\varsigma_{1,m}(\boldsymbol{\upsilon}') - \varsigma_{1,m}(\boldsymbol{\upsilon}))(\varsigma_{1,m}(\boldsymbol{\upsilon}') - \varsigma_{1,m}(\boldsymbol{\upsilon}))^\top\right]\mathcal{D}_m^{-1}\boldsymbol{\gamma}^\top \\
&\quad \le \mathtt{C}_{(\boldsymbol{\mathcal{E}\mathfrak{D}_1})}^2 m^2\|\mathcal{D}_m(\boldsymbol{\upsilon}-\boldsymbol{\upsilon}')\|^2\omega^2.
\end{aligned}$$

This yields $(\boldsymbol{\mathcal{E}\mathfrak{D}_1})$ with

$$\nu_{1,m}^2 = \widetilde{\nu}^2\mathtt{C}_{(\boldsymbol{\mathcal{E}\mathfrak{D}_1})}^2 m^2.$$

□

Condition $(\boldsymbol{\mathcal{L}_0})$

Lemma 6.A.18. *The condition* $(\boldsymbol{\mathcal{L}_0})$ *is satisfied where*

$$\delta(\mathtt{r}) = \frac{\mathtt{C}_{(\boldsymbol{\mathcal{L}_0})}\left\{m^{3/2} + \mathtt{C}_{bias}m^{5/2}\right\}\mathtt{r}}{c_{\mathfrak{D}}\sqrt{n}},$$

where $\mathtt{C}_{\delta,1}, \mathtt{C}_{\delta,2} > 0$ *only depend on* $\|\psi\|_\infty, \|\psi'\|_\infty, \|\psi''\|_\infty, \mathtt{C}_{\|\boldsymbol{f}^*\|}, L_{\nabla\varPhi}, s_{\mathbf{X}}$.

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

$$\|I_{p^*} - \mathcal{D}_m^{-1}\nabla_{p^*}^2 I\!\!E[\mathcal{L}(\boldsymbol{\upsilon})]\mathcal{D}_m^{-1}\| \le \frac{1}{nc_{\mathcal{D}}^2}\|\mathcal{D}_m^2(\boldsymbol{\upsilon}) - \mathcal{D}_m^2(\boldsymbol{\upsilon}_m^*)\|.$$

We represent

$$\begin{aligned}
-\nabla_{p^*}^2 I\!\!E[\mathcal{L}_m(\boldsymbol{\upsilon})] &\overset{\text{def}}{=} \mathcal{D}_m^2(\boldsymbol{\upsilon}) = nd_m^2(\boldsymbol{\upsilon}) + nr_m^2(\boldsymbol{\upsilon}),\\
nd_m^2(\boldsymbol{\upsilon}) &= n\begin{pmatrix} d_{\boldsymbol{\theta}}^2(\boldsymbol{\upsilon}) & a_m(\boldsymbol{\upsilon}) \\ a_m^\top(\boldsymbol{\upsilon}) & h_m^2(\boldsymbol{\upsilon}) \end{pmatrix} \overset{\text{def}}{=} \begin{pmatrix} D(\boldsymbol{\upsilon})^2 & A_m^\top(\boldsymbol{\upsilon}) \\ A_m(\boldsymbol{\upsilon}) & \mathrm{H}_m^2(\boldsymbol{\upsilon}) \end{pmatrix},\\
r_m^2(\boldsymbol{\upsilon}) &= I\!\!E\Bigg[\Big(\boldsymbol{f}_{\boldsymbol{\eta}}(\mathbf{X}^\top\boldsymbol{\theta}) - g(\mathbf{X})\Big)\begin{pmatrix} v_{\boldsymbol{\theta}}^2(\boldsymbol{\upsilon}) & b_m(\boldsymbol{\upsilon}) \\ b_m^\top(\boldsymbol{\upsilon}) & 0 \end{pmatrix}\Bigg],\\
v_{\boldsymbol{\theta}}^2(\boldsymbol{\upsilon}) &= 2\boldsymbol{f}_{\boldsymbol{\eta}}''(\mathbf{X}^\top\boldsymbol{\theta})\nabla\varPhi_{\boldsymbol{\theta}}^\top\mathbf{X}(\mathbf{X})^\top\nabla\varPhi_{\boldsymbol{\theta}}\\
&\quad + |\boldsymbol{f}_{\boldsymbol{\eta}}'(\mathbf{X}^\top\boldsymbol{\theta})|^2\mathbf{X}^\top\nabla^2\varphi_{\boldsymbol{\theta}}^\top[\mathbf{X},\cdot,\cdot],\\
b_m(\boldsymbol{\upsilon}) &= \nabla\varPhi_{\boldsymbol{\theta}}\mathbf{X}^\top\boldsymbol{e}'^\top(\mathbf{X}^\top\boldsymbol{\theta}),
\end{aligned}$$

such that

$$\begin{aligned}
\frac{1}{n}\|\mathcal{D}_m^2(\boldsymbol{\upsilon}) - \mathcal{D}_m^2(\boldsymbol{\upsilon}_m^*)\| \le \frac{1}{n}\Big(&\|D^2(\boldsymbol{\upsilon}) - D^2(\boldsymbol{\upsilon}_m^*)\| + 2\|A_m(\boldsymbol{\upsilon}) - A_m(\boldsymbol{\upsilon}_m^*)\|\\
&+\|\mathrm{H}_m^2(\boldsymbol{\upsilon}) - \mathrm{H}_m^2(\boldsymbol{\upsilon}_m^*)\| + \|r_m^2(\boldsymbol{\upsilon}) - r_m^2(\boldsymbol{\upsilon}_m^*)\|\Big),
\end{aligned}$$

so that we can calculate separately

$$\begin{aligned}
&\frac{1}{n}\|D^2(\boldsymbol{\upsilon}) - D^2(\boldsymbol{\upsilon}_m^*)\|\\
&\quad\le I\!\!E[\|\mathbf{X}\|^2\Big\{|((\boldsymbol{f}_{\boldsymbol{\eta}}')^2 - (\boldsymbol{f}_{\boldsymbol{\eta}_m^*}')^2)(\mathbf{X}^\top\boldsymbol{\theta})|\\
&\qquad + |(\boldsymbol{f}_{\boldsymbol{\eta}_m^*}')^2(\mathbf{X}^\top\boldsymbol{\theta}) - (\boldsymbol{f}_{\boldsymbol{\eta}_m^*}')^2(\mathbf{X}^\top\boldsymbol{\theta}_m^*)|\\
&\qquad + 2|(\boldsymbol{f}_{\boldsymbol{\eta}_m^*}')^2(\mathbf{X}^\top\boldsymbol{\theta}_m^*)|\|\nabla\varPhi(\boldsymbol{\theta})^\top\mathbf{X} - \nabla\varPhi(\boldsymbol{\theta}_m^*)^\top\mathbf{X}\|\Big\}].
\end{aligned}$$

Using Lemma 6.A.10 we find

$$\begin{aligned}
&|(\boldsymbol{f}_{\boldsymbol{\eta}_m^*}')^2(\mathbf{X}^\top\boldsymbol{\theta}_m^*)|\|\nabla\varPhi(\mathbf{X}^\top\boldsymbol{\theta}) - \nabla\varPhi(\mathbf{X}^\top\boldsymbol{\theta}_m^*)\|\\
&\quad\le \|\psi'\|_\infty(C_{\|\boldsymbol{f}\|} + 1)\sqrt{2}L_{\nabla\varPhi}\|\boldsymbol{\theta} - \boldsymbol{\theta}_m^*\|.
\end{aligned}$$

Furthermore we have $M(j) \subset \{1, \dots, m\}$ in (6.A.10)

$$\begin{aligned}
I\!\!E|(\boldsymbol{f}'_{\boldsymbol{\eta}} - \boldsymbol{f}'_{\boldsymbol{\eta}^*_m})(\mathbf{X}^\top \boldsymbol{\theta})| &\le \sum_{k=1}^{m} |\eta_k - \eta^*_{mk}| I\!\!E|\boldsymbol{e}'_k(\mathbf{X}^\top \boldsymbol{\theta})| \qquad (6.A.25)\\
&\le \mathtt{C}_{p_{\mathbf{X}\boldsymbol{\theta}}} \|\psi'\| \|\boldsymbol{\eta} - \boldsymbol{\eta}^*_m\| \left(\sum_{k=1}^{j_m} 2^{j_k} |M(j)| \right)^{1/2}\\
&\le \mathtt{C} \|\boldsymbol{\eta} - \boldsymbol{\eta}^*_m\| \|\psi'\|_\infty m.
\end{aligned}$$

This implies using (6.A.14) , (6.A.25) and (6.A.22)

$$\begin{aligned}
&I\!\!E\left[|((\boldsymbol{f}'_{\boldsymbol{\eta}})^2 - (\boldsymbol{f}'_{\boldsymbol{\eta}^*_m})^2)(\mathbf{X}^\top \boldsymbol{\theta})| \right]\\
&\quad \le I\!\!E\left[|\boldsymbol{f}'_{\boldsymbol{\eta}}(\mathbf{X}^\top \boldsymbol{\theta})| + |\boldsymbol{f}'_{\boldsymbol{\eta}^*_m}(\mathbf{X}^\top \boldsymbol{\theta})|)|(\boldsymbol{f}'_{\boldsymbol{\eta}} - \boldsymbol{f}'_{\boldsymbol{\eta}^*_m})(\mathbf{X}^\top \boldsymbol{\theta})| \right]\\
&\quad \le I\!\!E\left[(|(\boldsymbol{f}'_{\boldsymbol{\eta}} - \boldsymbol{f}'_{\boldsymbol{\eta}^*_m})(\mathbf{X}^\top \boldsymbol{\theta})| + 2|\boldsymbol{f}'_{\boldsymbol{\eta}^*_m}(\mathbf{X}^\top \boldsymbol{\theta})|)|(\boldsymbol{f}'_{\boldsymbol{\eta}} - \boldsymbol{f}'_{\boldsymbol{\eta}^*_m})(\mathbf{X}^\top \boldsymbol{\theta})| \right]\\
&\quad \le I\!\!E\left[|(\boldsymbol{f}'_{\boldsymbol{\eta}} - \boldsymbol{f}'_{\boldsymbol{\eta}^*_m})(\mathbf{X}^\top \boldsymbol{\theta})|^2 \right]\\
&\qquad +2\|\psi'\|_\infty (C_{\|\boldsymbol{f}\|} + 1)\sqrt{2} I\!\!E[|(\boldsymbol{f}'_{\boldsymbol{\eta}} - \boldsymbol{f}'_{\boldsymbol{\eta}^*_m})(\mathbf{X}^\top \boldsymbol{\theta})|]\\
&\quad \le \|\psi'\|_\infty \left(2\|\psi'\|_\infty (C_{\|\boldsymbol{f}\|} + 1)\sqrt{2} + \mathtt{C}\frac{\mathtt{r}m}{\sqrt{n}} \right) m \|\boldsymbol{\eta} - \boldsymbol{\eta}^*_m\|\\
&\quad \le \mathtt{C} m \|\boldsymbol{\eta} - \boldsymbol{\eta}^*_m\|,
\end{aligned}$$

where we used $\frac{\mathtt{r}m}{\sqrt{n}} \to 0$ for $\mathtt{r}^2 \le \mathtt{C}m$. Finally we derive with (6.A.16) and (6.A.14)

$$\begin{aligned}
&\|(\boldsymbol{f}'_{\boldsymbol{\eta}^*_m})^2(\mathbf{X}^\top \boldsymbol{\theta}) - (\boldsymbol{f}'_{\boldsymbol{\eta}^*_m})^2(\mathbf{X}^\top \boldsymbol{\theta}^*_m)\|\\
&\quad \le (\|\boldsymbol{f}'_{\boldsymbol{\eta}^*_m}(\mathbf{X}^\top \boldsymbol{\theta}^*_m)\| + \|\boldsymbol{f}'_{\boldsymbol{\eta}^*_m}(\mathbf{X}^\top \boldsymbol{\theta})\|)\|\boldsymbol{f}'_{\boldsymbol{\eta}^*_m}(\mathbf{X}^\top \boldsymbol{\theta}) - \boldsymbol{f}'_{\boldsymbol{\eta}^*_m}(\mathbf{X}^\top \boldsymbol{\theta}^*_m)\|\\
&\quad \le 4\sqrt{2}\|\psi'\|_\infty (C_{\|\boldsymbol{f}\|} + 1)^2 \sqrt{m} \|\psi''\|_\infty s_{\mathbf{X}} \|\boldsymbol{\theta} - \boldsymbol{\theta}^*_m\|.
\end{aligned}$$

Collecting everything yields with some constant $\mathtt{C} > 0$

$$\frac{1}{n}\|D^2(\boldsymbol{\upsilon}) - D^2(\boldsymbol{\upsilon}^*_m)\| \le \mathtt{C}m\|\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^*_m\|.$$

Furthermore

$$
\begin{aligned}
&\frac{1}{n}\|\mathrm{H}_m^2(\boldsymbol{\upsilon}) - \mathrm{H}_m^2(\boldsymbol{\upsilon}_m^*)\| \\
&\quad = \sup_{\substack{\boldsymbol{\gamma}\in\mathbb{R}^m \\ \|\boldsymbol{\gamma}\|=1}} \sum_{k,l=1}^{m} \gamma_k\gamma_l \left(I\!E[\boldsymbol{e}_k\boldsymbol{e}_l(\mathbf{X}^\top\boldsymbol{\theta})] - I\!E[\boldsymbol{e}_k\boldsymbol{e}_l(\mathbf{X}^\top\boldsymbol{\theta}_m^*)]\right) 1_{I_l\cap I_k\neq\emptyset} \\
&\quad \le 2\sup_{\substack{\boldsymbol{\gamma}\in\mathbb{R}^m \\ \|\boldsymbol{\gamma}\|=1}} \sum_{k=1}^{m}\sum_{l=k}^{m} \gamma_k\gamma_l I\!E\left[\left(\boldsymbol{e}_k(\mathbf{X}^\top\boldsymbol{\theta}) - \boldsymbol{e}_k(\mathbf{X}^\top\boldsymbol{\theta}_m^*)\right)\boldsymbol{e}_l(\mathbf{X}^\top\boldsymbol{\theta})\right] 1_{I_l\cap I_k\neq\emptyset} \\
&\qquad +2\sup_{\substack{\boldsymbol{\gamma}\in\mathbb{R}^m \\ \|\boldsymbol{\gamma}\|=1}} \sum_{k=1}^{m}\sum_{l=k}^{m} \gamma_k\gamma_l \\
&\qquad\qquad I\!E\left[\boldsymbol{e}_k(\mathbf{X}^\top\boldsymbol{\theta}_m^*)\left(\boldsymbol{e}_l(\mathbf{X}^\top\boldsymbol{\theta}) - \boldsymbol{e}_l(\mathbf{X}^\top\boldsymbol{\theta}_m^*)\right)\right] 1_{I_l\cap I_k\neq\emptyset}.
\end{aligned}
$$

Using (6.A.7) and (6.3.4) this gives

$$
\begin{aligned}
&\sum_{k=1}^{m}\sum_{l=k}^{m} \gamma_k\gamma_l I\!E\left[\left(\boldsymbol{e}_k(\mathbf{X}^\top\boldsymbol{\theta}) - \boldsymbol{e}_k(\mathbf{X}^\top\boldsymbol{\theta}_m^*)\right)\boldsymbol{e}_l(\mathbf{X}^\top\boldsymbol{\theta})\right] 1_{I_l\cap I_k\neq\emptyset} \\
&\quad \le \|\boldsymbol{\theta}-\boldsymbol{\theta}_m^*\|\|\psi'\|s_{\mathbf{X}}^2 17\mathtt{C}_{p_{\mathbf{X}}} \sum_{k=1}^{m}\gamma_k 2^{j_k}\sum_{l=k}^{m}\gamma_l 1_{I_l\cap I_k\neq\emptyset} \\
&\quad \le \|\boldsymbol{\theta}-\boldsymbol{\theta}_m^*\|\|\psi'\|s_{\mathbf{X}}^2 17^{3/2}\mathtt{C}_{p_{\mathbf{X}}} \sum_{k=1}^{m}\gamma_k 2^{j_k/2}\left(\sum_{j=j_k}^{j_m} 2^{j}\right)^{1/2} \\
&\quad \le \|\boldsymbol{\theta}-\boldsymbol{\theta}_m^*\|\|\psi'\|s_{\mathbf{X}}^2 17^{3/2}\mathtt{C}_{p_{\mathbf{X}}}\sqrt{m}\left(\sum_{j=1}^{j_m} 2^{j_k}\right)^{1/2} \\
&\quad \le \|\boldsymbol{\theta}-\boldsymbol{\theta}_m^*\|\|\psi'\|s_{\mathbf{X}}^2 17^{3/2}\mathtt{C}_{p_{\mathbf{X}}} m,
\end{aligned}
$$

and

$$\begin{aligned}
&\sum_{k=1}^{m}\sum_{l=k}^{m}\gamma_k\gamma_l I\!\!E\left[\left(\boldsymbol{e}_l(\mathbf{X}^\top\boldsymbol{\theta})-\boldsymbol{e}_l(\mathbf{X}^\top\boldsymbol{\theta}^*_m)\right)\boldsymbol{e}_k(\mathbf{X}^\top\boldsymbol{\theta})\right]1_{I_l\cap I_k\neq\emptyset}\\
&\quad\le\|\boldsymbol{\theta}-\boldsymbol{\theta}^*_m\|\|\psi'\|s^2_{\mathbf{X}}17(\mathtt{C}_{p_{\mathbf{X}}}+\|\psi\|_\infty)\sum_{k=1}^{m}\gamma_k 2^{j_k/2}\sum_{l=k}^{m}\gamma_l 2^{j_l/2}1_{I_l\cap I_k\neq\emptyset}\\
&\quad\le\|\boldsymbol{\theta}-\boldsymbol{\theta}^*_m\|\|\psi'\|s^2_{\mathbf{X}}17^{3/2}(\mathtt{C}_{p_{\mathbf{X}}}+\|\psi\|_\infty)\sum_{k=1}^{m}\gamma_k\left(\sum_{j=j_k}^{j_m}2^{2j}\right)^{1/2}\\
&\quad\le\|\boldsymbol{\theta}-\boldsymbol{\theta}^*_m\|\|\psi'\|s^2_{\mathbf{X}}17^{3/2}(\mathtt{C}_{p_{\mathbf{X}}}+\|\psi\|_\infty)m.
\end{aligned}$$

Consequently with some constant $\mathtt{C}_{\mathrm{H}}\in\mathbb{R}$

$$\frac{1}{n}\|\mathrm{H}^2_m(\boldsymbol{\upsilon})-\mathrm{H}^2_m(\boldsymbol{\upsilon}^*_m)\|\le\frac{\mathtt{C}_{\mathrm{H}}m}{\sqrt{n}c_{\mathcal{D}}}\|\mathcal{D}(\boldsymbol{\upsilon}-\boldsymbol{\upsilon}^*_m)\|. \tag{6.A.26}$$

Again with some constant $\mathtt{C}>0$

$$\begin{aligned}
\frac{1}{n}\|A_m(\boldsymbol{\upsilon})-A_m(\boldsymbol{\upsilon}^*_m)\|\le\mathtt{C}\Bigg(&I\!\!E\Big[\Big\|\boldsymbol{f}'_{\boldsymbol{\eta}}(\mathbf{X}_1^\top\boldsymbol{\theta})-\boldsymbol{f}'_{\boldsymbol{\eta}^*_m}(\mathbf{X}_1^\top\boldsymbol{\theta}^*_m)\Big\|^2\Big]^{1/2}\\
&+I\!\!E\Big[\|\nabla\Phi(\boldsymbol{\theta})-\nabla\Phi(\boldsymbol{\theta}^*_m)\|^2\Big]^{1/2}\\
&+I\!\!E\Big[\Big\|\boldsymbol{e}(\mathbf{X}_1^\top\boldsymbol{\theta})-\boldsymbol{e}(\mathbf{X}_1^\top\boldsymbol{\theta}^*_m)\Big\|^2\Big]^{1/2}\Bigg).
\end{aligned}$$

Note that using (6.A.24)

$$\begin{aligned}
&I\!\!E\Big[\Big\|\boldsymbol{e}(\mathbf{X}_1^\top\boldsymbol{\theta})-\boldsymbol{e}(\mathbf{X}_1^\top\boldsymbol{\theta}^*_m)\Big\|^2\Big]\\
&\quad\le\sup_{\substack{\boldsymbol{\eta}^\circ\in\mathbb{R}^m\\\|\boldsymbol{\eta}^\circ\|=1}}I\!\!E\left[\Big\{\boldsymbol{f}_{\boldsymbol{\eta}^\circ}(\mathbf{X}^\top\boldsymbol{\theta})-\boldsymbol{f}_{\boldsymbol{\eta}^\circ}(\mathbf{X}^\top\boldsymbol{\theta}^*_m)\Big\}^2\right]\\
&\quad\le\|\boldsymbol{\theta}-\boldsymbol{\theta}^*_m\|^2\mathtt{C}_{p_{\mathbf{X}}}\|\psi'\|^2_\infty s^4_{\mathbf{X}}17^{5/2}4m^2.
\end{aligned}$$

Using (6.A.23) this yields

$$\frac{1}{n}\|A_m(\boldsymbol{\upsilon})-A_m(\boldsymbol{\upsilon}^*_m)\|\le\mathtt{C}m\|\boldsymbol{\upsilon}-\boldsymbol{\upsilon}'\|.$$

Finally we estimate the fourth term.

$$\begin{aligned}
\|r^2_m(\boldsymbol{\upsilon})-r^2_m(\boldsymbol{\upsilon}^*_m)\|\le\;&I\!\!E[|\boldsymbol{f}_{\boldsymbol{\eta}}(\mathbf{X}^\top\boldsymbol{\theta})-\boldsymbol{f}_{\boldsymbol{\eta}^*_m}(\mathbf{X}^\top\boldsymbol{\theta}^*_m)|\|\widetilde{\mathcal{V}}^2_m(\boldsymbol{\upsilon})\|] \qquad (6.A.27)\\
&+I\!\!E[|\boldsymbol{f}_{\boldsymbol{\eta}^*_m}(\mathbf{X}^\top\boldsymbol{\theta}^*_m)-g(\mathbf{X})|\|\widetilde{\mathcal{V}}^2_m(\boldsymbol{\upsilon}^*_m)-\widetilde{\mathcal{V}}^2_m(\boldsymbol{\upsilon})\|].
\end{aligned}$$

We estimate separately

$$\begin{aligned}
&I\!E[|\boldsymbol{f}_{\boldsymbol{\eta}}(\mathbf{X}^\top\boldsymbol{\theta}) - \boldsymbol{f}_{\boldsymbol{\eta}^*_m}(\mathbf{X}^\top\boldsymbol{\theta}^*_m)|\|(\widetilde{\mathcal{V}}^2_m)(\boldsymbol{\upsilon})\|]\\
&\quad\le I\!E[|\boldsymbol{f}_{\boldsymbol{\eta}}(\mathbf{X}^\top\boldsymbol{\theta}) - \boldsymbol{f}_{\boldsymbol{\eta}^*_m}(\mathbf{X}^\top\boldsymbol{\theta}^*_m)|\|v^2_{\boldsymbol{\theta}}(\boldsymbol{\upsilon})\|]\\
&\qquad+I\!E[|\boldsymbol{f}_{\boldsymbol{\eta}}(\mathbf{X}^\top\boldsymbol{\theta}) - \boldsymbol{f}_{\boldsymbol{\eta}^*_m}(\mathbf{X}^\top\boldsymbol{\theta}^*_m)|\|b_m(\boldsymbol{\upsilon})\|]
\end{aligned}$$

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 (6.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 $I\!E\|b_m(\boldsymbol{\upsilon})\|^2$ we use (6.A.21) to bound

$$\begin{aligned}
I\!E\|b_m(\boldsymbol{\upsilon})\|^2 &\le s^2_{\mathbf{X}}\sum_{k=1}^{m} I\!E\boldsymbol{e}'_k(\mathbf{X}^\top\boldsymbol{\theta})^2\\
&\le s^2_{\mathbf{X}}17^2\mathtt{C}^2_{p\mathbf{X}}\|\psi'\|^2_\infty\sum_{k=1}^{m}2^{2j_k}\\
&\le s^2_{\mathbf{X}}17^2\mathtt{C}^2_{p\mathbf{X}}\|\psi'\|^2_\infty m^3. \qquad (6.A.28)
\end{aligned}$$

This implies for the first summand in (6.A.27) with constants $\mathtt{C}, \mathtt{C}' > 0$ large enough

$$\begin{aligned}
&I\!E[|\boldsymbol{f}_{\boldsymbol{\eta}}(\mathbf{X}^\top\boldsymbol{\theta}) - \boldsymbol{f}_{\boldsymbol{\eta}^*_m}(\mathbf{X}^\top\boldsymbol{\theta}^*_m)|\|\widetilde{\mathcal{V}}^2_m(\boldsymbol{\upsilon})\|]\\
&\quad\le \Big(I\!E[|\boldsymbol{f}_{\boldsymbol{\eta}}(\mathbf{X}^\top\boldsymbol{\theta}) - \boldsymbol{f}_{\boldsymbol{\eta}^*_m}(\mathbf{X}^\top\boldsymbol{\theta})|^2]^{1/2}\\
&\qquad+I\!E[|\boldsymbol{f}_{\boldsymbol{\eta}^*_m}(\mathbf{X}^\top\boldsymbol{\theta}) - \boldsymbol{f}_{\boldsymbol{\eta}^*_m}(\mathbf{X}^\top\boldsymbol{\theta}^*_m)|^2]^{1/2}\Big)\mathtt{C}m^{3/2}\\
&\quad\le \mathtt{C}m^{3/2}\|\boldsymbol{\upsilon}-\boldsymbol{\upsilon}^*_m\| + \mathtt{C}m^{3/2}I\!E[|\boldsymbol{f}_{\boldsymbol{\eta}}(\mathbf{X}^\top\boldsymbol{\theta}) - \boldsymbol{f}_{\boldsymbol{\eta}^*_m}(\mathbf{X}^\top\boldsymbol{\theta})|^2]^{1/2}.
\end{aligned}$$

We estimate using (6.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

$$\begin{aligned}
&I\!\!E[|\boldsymbol{f}_{\boldsymbol{\eta}}(\mathbf{X}^\top\boldsymbol{\theta}) - \boldsymbol{f}_{\boldsymbol{\eta}^*_m}(\mathbf{X}^\top\boldsymbol{\theta})|^2]^{1/2} = \frac{1}{\sqrt{n}}\|\mathrm{H}_m(\boldsymbol{\upsilon})(\boldsymbol{\eta}-\boldsymbol{\eta}^*_m)\|\\
&\quad\le \frac{1}{\sqrt{n}}\|\mathrm{H}^2_m(\boldsymbol{\upsilon}) - \mathrm{H}^2_m(\boldsymbol{\upsilon}^*_m)\|^{1/2}\|(\boldsymbol{\eta}-\boldsymbol{\eta}^*_m)\| + \frac{1}{\sqrt{n}}\|\mathrm{H}_m(\boldsymbol{\upsilon}^*_m)(\boldsymbol{\eta}-\boldsymbol{\eta}^*_m)\|\\
&\quad\le \left(\frac{1}{\sqrt{n}}\|\mathrm{H}^2_m(\boldsymbol{\upsilon}) - \mathrm{H}^2_m(\boldsymbol{\upsilon}^*_m)\|^{1/2}\frac{1}{\sqrt{n}c_{\mathcal{D}}} + \frac{1}{\sqrt{n}}\right)\|\mathrm{H}_m(\boldsymbol{\upsilon}^*_m)(\boldsymbol{\eta}-\boldsymbol{\eta}^*_m)\|\\
&\quad\le \left\{\left(\mathtt{r}m^{3/2}/\sqrt{n}\right)^{1/2} + 1\right\}\frac{\mathtt{C}}{\sqrt{n}c_{\mathcal{D}}}\|\mathcal{D}_m(\boldsymbol{\upsilon}-\boldsymbol{\upsilon}^*_m)\|\\
&\quad\le \frac{\mathtt{C}'}{\sqrt{n}c_{\mathcal{D}}}\|\mathcal{D}_m(\boldsymbol{\upsilon}-\boldsymbol{\upsilon}^*_m)\|.
\end{aligned}$$

We also find

$$\begin{aligned}
&|\boldsymbol{f}_{\boldsymbol{\eta}^*_m}(\mathbf{X}^\top\boldsymbol{\theta}) - \boldsymbol{f}_{\boldsymbol{\eta}^*_m}(\mathbf{X}^\top\boldsymbol{\theta}^*_m)|\\
&\quad\le \left(\sum_{k=1}^m(\boldsymbol{\eta}^*_m)^2_k k^{2\alpha}\right)^{1/2}\left(\sum_{k=1}^m|\boldsymbol{e}'_k(\mathbf{X}^\top\boldsymbol{\theta}^*_m)|^2k^{-2\alpha}\right)^{1/2}L_{\nabla\Phi}\|\mathbf{X}\|\|\boldsymbol{\theta}-\boldsymbol{\theta}^*_m\|\\
&\quad\le 2\sqrt{34}C_{\|\boldsymbol{\eta}^*_m\|}\sqrt{2}L_{\nabla\Phi}s_{\mathbf{X}}\|\psi'\|_\infty\|\boldsymbol{\theta}-\boldsymbol{\theta}^*_m\|.
\end{aligned}$$

Consequently

$$I\!\!E[|\boldsymbol{f}_{\boldsymbol{\eta}}(\mathbf{X}^\top\boldsymbol{\theta}) - \boldsymbol{f}_{\boldsymbol{\eta}^*_m}(\mathbf{X}^\top\boldsymbol{\theta}^*_m)|\|\widetilde{\mathcal{V}}^2_m(\boldsymbol{\upsilon})\|] \le \frac{\mathtt{C}m^{3/2}}{\sqrt{n}c_{\mathcal{D}}}\|\mathcal{D}_m(\boldsymbol{\upsilon}-\boldsymbol{\upsilon}^*_m)\|.$$

Furthermore using that $|\boldsymbol{f}_{\boldsymbol{\eta}^*_m}(\mathbf{X}^\top\boldsymbol{\theta}^*_m) - g(\mathbf{X})| \le \mathtt{C}_{bias}$

$$\begin{aligned}
&I\!\!E[|\boldsymbol{f}_{\boldsymbol{\eta}^*_m}(\mathbf{X}^\top\boldsymbol{\theta}^*_m) - g(\mathbf{X})|\|\widetilde{\mathcal{V}}^2_m(\boldsymbol{\upsilon}^*_m) - \widetilde{\mathcal{V}}^2_m(\boldsymbol{\upsilon})\|]\\
&\quad\le \mathtt{C}_{bias}\left(I\!\!E[\|v^2_{\boldsymbol{\theta}}(\boldsymbol{\upsilon}^*_m) - v^2_{\boldsymbol{\theta}}(\boldsymbol{\upsilon})\|] + 2I\!\!E[\|b_m(\boldsymbol{\upsilon}^*_m) - b_m(\boldsymbol{\upsilon})\|]\right).
\end{aligned}$$

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.

$$\begin{aligned}
&\|v^2_{\boldsymbol{\theta}}(\boldsymbol{\upsilon}^*_m) - v^2_{\boldsymbol{\theta}}(\boldsymbol{\upsilon})\|\\
&\quad\le \|2\boldsymbol{f}''_{\boldsymbol{\eta}}(\mathbf{X}^\top\boldsymbol{\theta})\nabla\Phi^\top_{\boldsymbol{\theta}}\mathbf{X}(\mathbf{X})^\top\nabla\Phi_{\boldsymbol{\theta}} - 2\boldsymbol{f}''_{\boldsymbol{\eta}^*_m}(\mathbf{X}^\top\boldsymbol{\theta}^*_m)\nabla\Phi^\top_{\boldsymbol{\theta}^*_m}\mathbf{X}(\mathbf{X})^\top\nabla\Phi_{\boldsymbol{\theta}^*_m}\|\\
&\qquad+\||\boldsymbol{f}'_{\boldsymbol{\eta}^*_m}(\mathbf{X}^\top\boldsymbol{\theta}^*_m)|^2\mathbf{X}^\top\nabla^2\Phi^\top_{\boldsymbol{\theta}^*_m}[\mathbf{X},\cdot,\cdot] - |\boldsymbol{f}'_{\boldsymbol{\eta}}(\mathbf{X}^\top\boldsymbol{\theta})|^2\mathbf{X}^\top\nabla^2\varphi^\top_{\boldsymbol{\theta}}[\mathbf{X},\cdot,\cdot]\|\\
&\quad\le \mathtt{C}_1\left||\boldsymbol{f}'_{\boldsymbol{\eta}^*_m}(\mathbf{X}^\top\boldsymbol{\theta}^*_m)|^2 - |\boldsymbol{f}'_{\boldsymbol{\eta}}(\mathbf{X}^\top\boldsymbol{\theta})|^2\right| + \mathtt{C}_2|\boldsymbol{f}''_{\boldsymbol{\eta}^*_m}(\mathbf{X}^\top\boldsymbol{\theta}^*_m) - \boldsymbol{f}''_{\boldsymbol{\eta}}(\mathbf{X}^\top\boldsymbol{\theta})|\\
&\qquad+\mathtt{C}_3\|\boldsymbol{\theta}-\boldsymbol{\theta}^*_m\|.
\end{aligned}$$

With the same arguments as those used for the bound of $\frac{1}{n}\|D^2(\boldsymbol{\upsilon}) - D^2(\boldsymbol{\upsilon}_m^*)\|$ we find

$$I\!E\left||\boldsymbol{f}'_{\boldsymbol{\eta}_m^*}(\mathbf{X}^\top\boldsymbol{\theta}_m^*)|^2 - |\boldsymbol{f}'_{\boldsymbol{\eta}}(\mathbf{X}^\top\boldsymbol{\theta})|^2\right| \le \mathtt{C}m\|\boldsymbol{\upsilon} - \boldsymbol{\upsilon}_m^*\|.$$

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

$$\begin{aligned}|N(j)| \overset{\text{def}}{=} \Big|\Big\{ \quad & k \in \{2^j - 2j17 - 1, \ldots, 2^{j+1} - 2(j+1)17 - 1 - 1\} : \\ & |\boldsymbol{e}''_k(\mathbf{X}_i^\top\boldsymbol{\theta}') - \boldsymbol{e}''_k(\mathbf{X}_i^\top\boldsymbol{\theta})| > 0\Big\}\Big| \le 34.\end{aligned}$$

we estimate

$$\begin{aligned}&|\boldsymbol{f}''_{\boldsymbol{\eta}_m^*}(\mathbf{X}^\top\boldsymbol{\theta}_m^*) - \boldsymbol{f}''_{\boldsymbol{\eta}_m^*}(\mathbf{X}^\top\boldsymbol{\theta})| \\ &\quad \le \sqrt{34}\|\psi'''\|_\infty\|\boldsymbol{\theta} - \boldsymbol{\theta}_m^*\| \left(\sum_{k=1}^m {\eta_m^*}^2 k^{-2\alpha}\right)^{1/2} \left(\sum_{j=1}^{j_m} 2^{(7-2\alpha)j)}\right)^{1/2} \\ &\quad \le \mathtt{C}m^{3/2}\|\boldsymbol{\theta} - \boldsymbol{\theta}_m^*\|,\end{aligned}$$

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

$$\begin{aligned}I\!E[\|b_m(\boldsymbol{\upsilon}_m^*) - b_m(\boldsymbol{\upsilon})\|] &\le \mathtt{C}I\!E\|\boldsymbol{e}'(\mathbf{X}^\top\boldsymbol{\theta}) - \boldsymbol{e}'(\mathbf{X}^\top\boldsymbol{\theta}_m^*)\| \\ &\quad + \mathtt{C}I\!E[\|\boldsymbol{e}'(\mathbf{X}^\top\boldsymbol{\theta}_m^*)\|^2]^{1/2}\|\boldsymbol{\theta} - \boldsymbol{\theta}_m^*\|.\end{aligned}$$

By (6.A.28) we have

$$I\!E[\|\boldsymbol{e}'(\mathbf{X}^\top\boldsymbol{\theta}_m^*)\|^2]^{1/2} \le 17\mathtt{C}_{p_{\mathbf{X}}}\|\psi'\|_\infty m^{3/2}.$$

Furthermore

$$\begin{aligned}I\!E\|\boldsymbol{e}'(\mathbf{X}^\top\boldsymbol{\theta}) - \boldsymbol{e}'(\mathbf{X}^\top\boldsymbol{\theta}_m^*)\| &\le I\!E\left[\|\boldsymbol{e}'(\mathbf{X}^\top\boldsymbol{\theta}) - \boldsymbol{e}'(\mathbf{X}^\top\boldsymbol{\theta}_m^*)\|^2\right]^{1/2} \\ &= \left(\sum_{k=1}^m I\!E\left[(\boldsymbol{e}'_k(\mathbf{X}^\top\boldsymbol{\theta}) - \boldsymbol{e}'_k(\mathbf{X}^\top\boldsymbol{\theta}_m^*))^2\right]\right)^{1/2}\end{aligned}$$

With (6.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

$$\begin{aligned}&\mathbb{E}[|\boldsymbol{f}_{\boldsymbol{\eta}^*_m}(\mathbf{X}^\top\boldsymbol{\theta}^*_m)-g(\mathbf{X})|\|\widetilde{\mathcal{V}}^2_m(\boldsymbol{\upsilon}^*_m)-\widetilde{\mathcal{V}}^2_m(\boldsymbol{\upsilon})\|]\\ &\quad\le \mathtt{C}m^{3/2}\|\boldsymbol{\upsilon}-\mathit{\Pi}_{p^*}\boldsymbol{\upsilon}^*\|+\mathtt{C}_{bias}\mathtt{C}m^{5/2}.\end{aligned}$$

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

$$\delta(\mathtt{r})=\frac{\mathtt{C}_{(\mathcal{L}_0)}\left\{m^{3/2}+\mathtt{C}_{bias}m^{5/2}\right\}\mathtt{r}}{c_{\mathcal{D}}\sqrt{n}}.$$

□

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 [39] and is the key tool of this subsection.

Theorem 6.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* (6.A.29)

$$\mathbb{E}\left[\sup_{\boldsymbol{\upsilon}\in\varUpsilon_\circ(\mathtt{r})^c}(P_n-\mathbb{P})\chi_{\mathtt{b}}(\boldsymbol{\upsilon})\right]\le C_\chi,\quad \mathbb{P}(e>C_e)\le\tau_e,$$

$$\boldsymbol{Q}(\mathtt{b})\overset{\text{def}}{=}\inf_{\boldsymbol{\upsilon}\in\varUpsilon_\circ(\mathtt{r})^c}\mathbb{P}\left(f_i(\boldsymbol{\upsilon},\mathbf{X}_i)\ge\mathtt{b}\mathtt{r}^2/n\right)>0.$$

Choose

$$0 < \lambda \leq (\boldsymbol{Q}(2\mathtt{b}) - 2/n - 2C_{\chi})/4.$$

Then for $\mathtt{r}^2 \geq C_e/(\lambda\mathtt{b}) \vee \mathtt{r}_{\boldsymbol{Q}}^2$

$$I\!\!P\left(\inf_{\boldsymbol{\upsilon}\in\varUpsilon_{\circ}(\mathtt{r})^c} F(\boldsymbol{\upsilon}) \leq \lambda\mathtt{b}\mathtt{r}^2\right) \leq \exp\left\{-n\boldsymbol{Q}(2\mathtt{b})^2/4\right\} + \tau_e$$

The auxiliary function is defined as

$$\overline{\chi}_u(t) = \begin{cases} 0, & t \leq u; \\ t/u - 1, & t \in [u, 2u]; \\ 1, & t \geq 2u; \end{cases} \quad \chi_{\mathtt{b}}(\boldsymbol{\upsilon})_i \overset{\text{def}}{=} \overline{\chi}_{\mathtt{b}}(f_i(\boldsymbol{\upsilon})). \tag{6.A.29}$$

Remark 6.A.7. The proof is nearly the same as that of Theorem 4.3 of [39]. 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 proof. 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 6.A.20. *Assume the conditions* $(\mathcal{A})$ *. Then for* $n \in \mathbb{N}$ *large enough there exist* $c_{(\boldsymbol{Q})}, c_{(\boldsymbol{\mathcal{L}}\mathbf{r})}, \mathtt{C} > 0$ *such that with probability* $1 - \exp\left\{-m^3\mathtt{x}\right\} - \exp\left\{-nc_{(\boldsymbol{Q})}/4\right\}$

$$-\inf_{\boldsymbol{\upsilon}\in\varUpsilon_{\circ}(\mathtt{r})^c} I\!\!E[\mathcal{L}(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}_m^*)] > c_{(\boldsymbol{\mathcal{L}}\mathbf{r})}\mathtt{r}^2/2,$$

as soon as $\mathtt{r}^2 \geq \mathtt{C}(m + \mathtt{x})$ *.*

Proof. We will prove this claim using Theorem 6.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 (6.1.10)

$$\begin{aligned}
&-I\!\!E_{\varepsilon}[\mathcal{L}(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}_m^*)] = -I\!\!E[\mathcal{L}(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}_m^*)|(\mathbf{X})] \\
&\quad = \sum_{i=1}^n \left[|\boldsymbol{f}_{\boldsymbol{\eta}}(\mathbf{X}_i^\top\boldsymbol{\theta}) - \boldsymbol{f}_{\boldsymbol{\eta}^*}(\mathbf{X}^\top\boldsymbol{\theta}^*)|^2 - |\boldsymbol{f}_{\boldsymbol{\eta}_m^*}(\mathbf{X}_i^\top\boldsymbol{\theta}_m^*) - \boldsymbol{f}_{\boldsymbol{\eta}^*}(\mathbf{X}^\top\boldsymbol{\theta}^*)|^2\right] \\
&\quad \geq \sum_{i=1}^n \left[|\boldsymbol{f}_{\boldsymbol{\eta}}(\mathbf{X}_i^\top\boldsymbol{\theta}) - \boldsymbol{f}_{\boldsymbol{\eta}^*}(\mathbf{X}_i^\top\boldsymbol{\theta}^*)|^2\right] \\
&\quad\quad -nI\!\!E[|\boldsymbol{f}_{\boldsymbol{\eta}_m^*}(\mathbf{X}_i^\top\boldsymbol{\theta}_m^*) - \boldsymbol{f}_{\boldsymbol{\eta}^*}(\mathbf{X}_i^\top\boldsymbol{\theta}^*)|^2] \\
&\quad\quad -n\left|(\mathrm{P}_n - I\!\!P)|\boldsymbol{f}_{\boldsymbol{\eta}_m^*}(\mathbf{X}^\top\boldsymbol{\theta}_m^*) - \boldsymbol{f}_{\boldsymbol{\eta}^*}(\mathbf{X}_i^\top\boldsymbol{\theta}^*)|^2\right|.
\end{aligned}$$

We define

$$\begin{aligned} e \overset{\text{def}}{=} & n\mathbb{E}[|\boldsymbol{f}_{\boldsymbol{\eta}^*_m}(\mathbf{X}_i^\top \boldsymbol{\theta}^*_m) - \boldsymbol{f}_{\boldsymbol{\eta}^*}(\mathbf{X}^\top \boldsymbol{\theta}^*)|^2] \\ & + n\left|(\mathrm{P}_n - \mathbb{P})|\boldsymbol{f}_{\boldsymbol{\eta}^*_m}(\mathbf{X}^\top \boldsymbol{\theta}^*_m) - \boldsymbol{f}_{\boldsymbol{\eta}^*}(\mathbf{X}^\top \boldsymbol{\theta}^*)|^2\right|, \end{aligned}$$

such that

$$-\mathbb{E}[\mathcal{L}(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^*_m)|(\mathbf{X})] \ge \sum_{i=1}^n \Big(\boldsymbol{f}_{\boldsymbol{\eta}}(\mathbf{X}_i^\top \boldsymbol{\theta}) - \boldsymbol{f}_{\boldsymbol{\eta}^*}(\mathbf{X}_i^\top \boldsymbol{\theta}^*)\Big)^2 - e,$$

This hints that Theorem 6.A.19 gives the desired result. Consider the following list of assumptions:

(1) With some $\mathtt{C} > 0$

$$n\mathbb{E}[|\boldsymbol{f}_{\boldsymbol{\eta}^*_m}(\mathbf{X}^\top \boldsymbol{\theta}^*_m) - \boldsymbol{f}_{\boldsymbol{\eta}^*}(\mathbf{X}^\top \boldsymbol{\theta}^*)|^2] \le 3(2+\mathtt{C})\mathtt{r}^{*2},$$

(2) With probability $1 - \exp\left\{-m^3\mathtt{x}\right\}$ and a constant $\mathtt{C}_{\sum} > 0$

$$n\left|(\mathrm{P}_n - \mathbb{P})|\boldsymbol{f}_{\boldsymbol{\eta}^*_m}(\mathbf{X}^\top \boldsymbol{\theta}^*_m) - \boldsymbol{f}_{\boldsymbol{\eta}^*}(\mathbf{X}^\top \boldsymbol{\theta}^*)|^2]\right| \le \mathtt{C}_{\sum},$$

(3) For some $\mathtt{b} > 0$ and for $n \in \mathbb{N}$ large enough and $\mathtt{r} > \sqrt{m}$

$$\begin{aligned} & \boldsymbol{Q}(2\mathtt{b}) \\ & \overset{\text{def}}{=} \inf_{(\boldsymbol{\theta},\boldsymbol{\eta}) \in \Upsilon_\circ(\mathtt{r})^c} \mathbb{P}\left[\Big(\boldsymbol{f}_{\boldsymbol{\eta}}(\mathbf{X}_i^\top \boldsymbol{\theta}) - \boldsymbol{f}_{\boldsymbol{\eta}^*}(\mathbf{X}_i^\top \boldsymbol{\theta}^*)\Big)^2 \ge \mathtt{b}\mathtt{r}^2/n\right] > 0, \end{aligned} \tag{6.A.30}$$

This means that in terms of Theorem 6.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 6.A.22, 6.A.23 and 6.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 6.A.21 completes the proof. □

Lemma 6.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

$$0 < \lambda \overset{\mathrm{def}}{=} \left(\boldsymbol{Q}(2\mathtt{b}) - 2/n + \mathtt{C}\sqrt{\frac{\log(n)p^*}{n}}\right)/4,$$

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 6.A.19. It remains to bound using the proof of Theorem 8.15 of [34]

$$\begin{aligned} \mathbb{E}\left[\sup_{\boldsymbol{\upsilon}\in\varUpsilon_\circ(\mathtt{r})^c}(\mathrm{P}_n - \mathbb{P})\chi_{\mathtt{b}}(\boldsymbol{\upsilon})\right] &\le \mathbb{E}\left[\sup_{\boldsymbol{\upsilon}\in\varUpsilon}(\mathrm{P}_n - \mathbb{P})\chi_{\mathtt{b}}(\boldsymbol{\upsilon})\right] \qquad (6.A.31) \\ &\le 2\mathtt{C}^*\mathbb{E}\left[\sqrt{\frac{6\{1+\log N(\delta,\mathcal{F},L_1(\mathrm{P}_n))\}}{n}}\right] + \delta, \end{aligned}$$

where $N(\delta,\mathcal{F},L_1(\mathrm{P}_n))$ denotes the δ-ball covering number of $\mathcal{F} \overset{\mathrm{def}}{=} \{\chi_{\mathtt{b}}(\boldsymbol{\upsilon}) : \boldsymbol{\upsilon}\in\varUpsilon\}$ 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 [34] ($\mathtt{C}^* = K(\exp(x^2)-1)$). The function $\chi_{\mathtt{b}} : \varUpsilon_\circ \to \mathbb{R}$ is defined via

$$\overline{\chi}_u(t) = \begin{cases} 0, & t \le u; \\ t/u - 1, & t \in [u, 2u]; \\ 1, & t \ge 2u; \end{cases} \quad \chi_{\mathtt{b}}(\boldsymbol{\upsilon})_i \overset{\mathrm{def}}{=} \overline{\chi}_{\mathtt{b}}(|\boldsymbol{f_\eta}(\mathbf{X}_i^\top\boldsymbol{\theta}) - \boldsymbol{f}_{\boldsymbol{\eta}^*}(\mathbf{X}_i^\top\boldsymbol{\theta}^*)|^2).$$

We want to bound the right-hand side of (6.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 $\operatorname{diam}(\Upsilon_m) < \mathtt{C}\sqrt{m}$

$$
\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

$$
\begin{aligned}
&\sup_{\boldsymbol{\upsilon},\boldsymbol{\upsilon}^\circ} \frac{\|\chi_{\mathtt{b}}(\boldsymbol{\upsilon}) - \chi_{\mathtt{b}}(\boldsymbol{\upsilon}^\circ)\|_{L_1(\mathrm{P}_n)}}{\|\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^\circ\|_2} \\
&\qquad \le \sup_{\boldsymbol{\upsilon},\boldsymbol{\upsilon}^\circ} \left(\frac{2}{\|\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^\circ\|_2} \wedge \mathtt{C}_1 m^3 + \mathtt{C}_2 m^4 \|\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^\circ\|_2 \right) \\
&\qquad = \mathtt{C}_3 m^3.
\end{aligned}
$$

We infer setting $\delta = \sqrt{p^*/n}$

$$
\begin{aligned}
&\sqrt{\frac{6\{1 + \log N(\delta, \mathcal{F}, L_1(\mathrm{P}_n))\}}{n}} + \delta \\
&\qquad \le \sqrt{\frac{6\{1 + \log N(\delta/(L(\mathrm{P}_n) \vee 1), \Upsilon, \|\cdot\|_2)\}}{n}} + \delta \\
&\qquad \le \sqrt{\frac{6\{1 + \log(\mathtt{C} m^3) + \log(1/\delta) p^*\}}{n}} + \delta \\
&\qquad \le \mathtt{C}_1 \sqrt{\frac{\log(p^*) + \log(n/p^*) p^*/2}{n}} + \sqrt{p^*/n} \\
&\qquad \le \mathtt{C}_2 \sqrt{\frac{\log(n) p^*}{n}}.
\end{aligned}
$$

The claim follows with Theorem 6.A.19. □

It remains to prove the assumptions (1), (2) and (3) which we do in the following three lemmas.

Lemma 6.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 4.A.7 (which is applicable because it only needs $(\boldsymbol{\mathcal{L}}\mathtt{r})$ for the full model and with center $\boldsymbol{\upsilon}^* \in \varUpsilon$) and Lemma 6.A.6 with some $\boldsymbol{\theta}^\circ \in \mathbf{Conv}(\boldsymbol{\theta}^*_m, \boldsymbol{\theta}^*)$

$$\begin{aligned}
&n\mathbb{E}[\|\boldsymbol{f}_{\boldsymbol{\eta}^*_m}(\mathbf{X}^\top\boldsymbol{\theta}^*_m) - \boldsymbol{f}_{\boldsymbol{\eta}^*}(\mathbf{X}^\top\boldsymbol{\theta}^*)\|^2]\\
&\quad\le 3n\left(\mathbb{E}[\|\boldsymbol{f}_{\boldsymbol{\eta}^*}(\mathbf{X}^\top\boldsymbol{\theta}^*_m) - \boldsymbol{f}_{\boldsymbol{\eta}^*}(\mathbf{X}^\top\boldsymbol{\theta}^*)\|^2] + \mathbb{E}[\|\boldsymbol{f}_{\boldsymbol{\eta}^*_m-\boldsymbol{\eta}^*}(\mathbf{X}^\top\boldsymbol{\theta}^*_m)\|^2]\right)\\
&\quad\le 3\left(\|D(\boldsymbol{\theta}^\circ)(\boldsymbol{\theta}^*_m - \boldsymbol{\theta}^*)\|^2 + \|\mathcal{H}(\boldsymbol{\upsilon}^*_m)(\boldsymbol{\eta}^*_m - \boldsymbol{f}^*)\|^2\right)\\
&\quad\le 3\Big((1+\|I - D^{-1/2}nD(\boldsymbol{\xi})D^{-1/2}\|)\|D(\boldsymbol{\theta}^*_m - \boldsymbol{\theta}^*)\|^2\\
&\qquad\quad +(1+\|I - \mathcal{H}^{-1}n\widetilde{\mathcal{H}}(\boldsymbol{\upsilon}^*_m)\mathcal{H}^{-1}\|)\|\mathcal{H}(\boldsymbol{\eta}^*_m - \boldsymbol{f}^*)\|^2\Big)\\
&\quad\le 3\left[2 + \|I - D^{-1/2}nD(\boldsymbol{\theta}^\circ)D^{-1/2}\| + \|I - \mathcal{H}^{-1}n\mathcal{H}(\boldsymbol{\upsilon}^*_m)\mathcal{H}^{-1}\|\right]\\
&\qquad\quad \|\mathcal{D}(\boldsymbol{\upsilon}^*_m - \boldsymbol{\upsilon}^*)\|^2\\
&\quad\le 3(2+\mathtt{C})\mathtt{r}^{*2}.
\end{aligned}$$

□

Lemma 6.A.23. *We have for a constant* $\mathtt{C} > 0$ *that only depends on* $\|\psi\|_\infty$, $\|\psi'\|_\infty$ *and* $s_{\mathbf{X}^2}$ *that*

$$\mathbb{P}\left(n\left|(P_n - \mathbb{P})|\boldsymbol{f}_{\boldsymbol{\eta}^*_m}(\mathbf{X}^\top\boldsymbol{\theta}^*_m) - \boldsymbol{f}_{\boldsymbol{\eta}^*}(\mathbf{X}^\top\boldsymbol{\theta}^*)|^2\right| \ge \mathtt{C}\sqrt{\mathtt{x}}\right) \le \exp\left\{-m^3\mathtt{x}\right\}.$$

Proof. We want to use the finite difference inequality. As above define

$$f : \bigotimes_{i=1}^{n}\mathbb{R}^p \to \mathbb{R},\;\; f(\mathbf{X}_1,\ldots,\mathbf{X}_n) \overset{\text{def}}{=} \mathrm{P}_n|\boldsymbol{f}_{\boldsymbol{\eta}^*_m}(\mathbf{X}^\top\boldsymbol{\theta}^*_m) - \boldsymbol{f}_{\boldsymbol{\eta}^*}(\mathbf{X}^\top\boldsymbol{\theta}^*)|^2,$$

and note that for any $i = 1,\ldots,n$ and any alternative realization $\mathbf{X}_i' \in \mathbb{R}$

$$\begin{aligned}
&n|f(\mathbf{X}_1,\ldots,\mathbf{X}_i,\ldots,\mathbf{X}_n) - f(\mathbf{X}_1,\ldots,\mathbf{X}_i'\ldots,\mathbf{X}_n)|\\
&\quad\le |\boldsymbol{f}_{\boldsymbol{\eta}^*_m}(\mathbf{X}_i^\top\boldsymbol{\theta}^*_m) - \boldsymbol{f}_{\boldsymbol{\eta}^*}(\mathbf{X}_i^\top\boldsymbol{\theta}^*)|^2 + |\boldsymbol{f}_{\boldsymbol{\eta}^*_m}({\mathbf{X}'}_i^\top\boldsymbol{\theta}^*_m) - g({\mathbf{X}'}_i)|^2.
\end{aligned}$$

We have

$$
\begin{aligned}
|\boldsymbol{f}_{\boldsymbol{\eta}^*_m}(\mathbf{X}^\top\boldsymbol{\theta}^*_m) - \boldsymbol{f}_{\boldsymbol{\eta}^*}(\mathbf{X}^\top\boldsymbol{\theta}^*)|^2 \le{}& 3|\boldsymbol{f}_{\boldsymbol{\eta}^*-\boldsymbol{\eta}^*_m}(\mathbf{X}_i^\top\boldsymbol{\theta})|^2 \\
&+3|\boldsymbol{f}_{\boldsymbol{\eta}^*}(\mathbf{X}_i^\top\boldsymbol{\theta}^*) - \boldsymbol{f}_{\boldsymbol{\eta}^*}(\mathbf{X}_i^\top\boldsymbol{\theta}^*_m)|^2.
\end{aligned}
$$

As in Lemma 6.A.11 there are constants $\mathtt{C},\mathtt{C}'$ such that

$$
\begin{aligned}
&|\boldsymbol{f}_{\boldsymbol{\eta}^*-\boldsymbol{\eta}^*_m}(\mathbf{X}_i^\top\boldsymbol{\theta}^*_m)|^2 \\
&\quad\le 3\left|\sum_{k=1}^{m}(\eta^*_k - \eta^*_{k,m})\boldsymbol{e}_k(\mathbf{X}_i^\top\boldsymbol{\theta}^*_m)\right|^2 + 3\left|\sum_{k=m+1}^{\infty}\eta^*_k\boldsymbol{e}_k(\mathbf{X}_i^\top\boldsymbol{\theta}^*_m)\right|^2 \\
&\quad\le 3\left|\sum_{k=1}^{m}\boldsymbol{e}_k^2(\mathbf{X}_i^\top\boldsymbol{\theta}^*_m)\right|\|\mathit{\Pi}_m\boldsymbol{\eta}^* - \boldsymbol{\eta}^*_m\|^2 + \mathtt{C}(\boldsymbol{\varkappa}^*) \\
&\quad\le \mathtt{C}'\left|\sum_{j=0}^{j_m}2^j\right|\|\mathit{\Pi}_m\boldsymbol{\eta}^* - \boldsymbol{\eta}^*_m\|^2 + \mathtt{C}(\boldsymbol{\varkappa}^*) \\
&\quad\le \mathtt{C}m\|\mathit{\Pi}_m\boldsymbol{\eta}^* - \boldsymbol{\eta}^*_m\|^2 + \mathtt{C}(\boldsymbol{\varkappa}^*),
\end{aligned}
$$

where $\mathtt{C}(\boldsymbol{\varkappa}^*) \le \mathtt{C}m^{-2\alpha+1}$. Furthermore again as in Lemma 6.A.11 there are constants $\mathtt{C},\mathtt{C}'$ such that

$$
\begin{aligned}
|\boldsymbol{f}_{\boldsymbol{\eta}^*}(\mathbf{X}_i^\top\boldsymbol{\theta}^*) - \boldsymbol{f}_{\boldsymbol{\eta}^*}(\mathbf{X}_i^\top\boldsymbol{\theta}^*_m)|^2 &\le \left|\sum_{k=1}^{m}\eta^*_k\Big(\boldsymbol{e}_k(\mathbf{X}_i^\top\boldsymbol{\theta}^*) - \boldsymbol{e}_k(\mathbf{X}_i^\top\boldsymbol{\theta}^*_m)\Big)\right|^2 \\
&\le \mathtt{C}'\left|\sum_{j=0}^{j_m}2^{3j-2\alpha}\right|\|\boldsymbol{\theta}^* - \boldsymbol{\theta}^*_m\|^2 \\
&\le \mathtt{C}\|\boldsymbol{\theta}^* - \boldsymbol{\theta}^*_m\|^2.
\end{aligned}
$$

This implies with Lemma 6.3.5 and constants $\mathtt{C}_1,\mathtt{C}_2 > 0$

$$
|\boldsymbol{f}_{\boldsymbol{\eta}^*_m}(\mathbf{X}^\top\boldsymbol{\theta}^*_m) - \boldsymbol{f}_{\boldsymbol{\eta}^*}(\mathbf{X}^\top\boldsymbol{\theta}^*)|^2 \le \mathtt{C}_1\left(\frac{m}{nc^2_{\mathfrak{D}}}\mathtt{r}^{*2} + m^{-2\alpha+1}\right) \le \mathtt{C}m^{-3}.
$$

Note that $\mathtt{r}^{*2}m/n \to 0$. This gives with the bounded difference inequality (Theorem 6.A.1) that

$$
I\!\!P\left(n\left|(\mathrm{P}_n - I\!\!P)|\boldsymbol{f}_{\boldsymbol{\eta}^*_m}(\mathbf{X}^\top\boldsymbol{\theta}^*_m) - \boldsymbol{f}_{\boldsymbol{\eta}^*}(\mathbf{X}^\top\boldsymbol{\theta}^*)|^2\right| \ge t\mathtt{C}m^{-3}\right) \le \exp\left\{-t^2\right\}.
$$

From this we infer with $t = m^3\sqrt{\mathtt{x}} \to \infty$

$$
I\!\!P\left(n\left|(\mathrm{P}_n - I\!\!P)|\boldsymbol{f}_{\boldsymbol{\eta}^*_m}(\mathbf{X}^\top\boldsymbol{\theta}^*_m) - \boldsymbol{f}_{\boldsymbol{\eta}^*}(\mathbf{X}^\top\boldsymbol{\theta}^*)|^2\right| \ge \mathtt{C}_2\sqrt{\mathtt{x}}\right) \le \exp\left\{-m^3\mathtt{x}\right\}.
$$

□

For a set $A \subset \mathbb{R}^p$ we denote by $\lambda(A) \in \mathbb{R}_+$ its Lebesgue measure and define

$$\lambda_{\boldsymbol{e}} \qquad (6.A.32)$$

$$\overset{\text{def}}{=} \sup \left\{ \lambda > 0 : \inf_{\substack{\boldsymbol{v} \in \mathbb{R}^m, \|\boldsymbol{v}\|=1 \\ \boldsymbol{\theta} \in S_1^{p,+}}} I\!P\left(|\langle \boldsymbol{v}, \boldsymbol{e}(\mathbf{X}^\top \boldsymbol{\theta}) \rangle| > \lambda \right) > 3/4 \right\}.$$

Remark 6.A.8. $\lambda_{\boldsymbol{e}} \geq \mathbb{R}$ in (6.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 6.A.24. *Denote the cylinder*

$$C_{\rho,x,y}(x_0, y_0) \overset{\text{def}}{=} \{(x, y, z) \in \mathbb{R}^2 \times \mathbb{R}^{p-2};\ (x - x_0)^2 + (y - y_0)^2 \leq \rho^2\}.$$

There is a point $(x_0, y_0) \in \mathbb{R}^2$ *such that* $\boldsymbol{Q}(2\mathtt{b})$ *in* (6.A.30) *satisfies*

$$\boldsymbol{Q}(2\mathtt{b}) + 3\mathrm{e}^{-\mathtt{x}} \geq \frac{1}{2} \wedge c_{p_{\mathbf{X}}} \lambda \left(B_h(0) \cap C_{h,x,y}(0) \cap B_{s_{\mathbf{X}}}(x_0, y_0, 0) \right.$$
$$\left. \cap \left\{ (x, y) \in \mathbb{R}^2 : \operatorname{sign}(y_0) y \geq \operatorname{sign}(y_0) h/2 \right\} \right),$$

for $\tau = \lambda_{\boldsymbol{e}} / (8 \boldsymbol{L}_{\boldsymbol{\eta}^*} s_{\mathbf{X}})$ *and*

$$2\mathtt{b} = (1 - \nu^2) \left(\frac{\lambda_{\boldsymbol{e}}^2 c_{\mathfrak{D}}^2}{32} \wedge \frac{\tau c_{f'_{\boldsymbol{\eta}^*}}^2 h^2}{4 p \pi^2 s_{\mathbf{X}}^2 \|p_{\mathbf{X}}\|_\infty^2 C_{\|\boldsymbol{\eta}^*\|}} \right),$$

and for

$$\mathtt{r} \geq \sqrt{m} \frac{4 \mathtt{C}_{\boldsymbol{\varkappa}}}{\lambda_{\boldsymbol{e}} \sqrt{(1 - \nu)}}.$$

Remark 6.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

$$\inf_{\boldsymbol{v} \in \Upsilon_\circ(\mathtt{r})^c} I\!P \left[\left(\boldsymbol{f}_{\boldsymbol{\eta}}(\mathbf{X}_i^\top \boldsymbol{\theta}) - \boldsymbol{f}_{\boldsymbol{\eta}^*}(\mathbf{X}_i^\top \boldsymbol{\theta}^*) \right)^2 \geq \frac{\mathtt{br}^2}{n} \right] > 0. \qquad (6.A.33)$$

We carry out the proof in two steps.

1. Before we determine $\mathtt{b} > 0$ that allows to prove (6.A.33) note that

$$\begin{aligned}&\|\mathcal{D}_m(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}_m^*)\| - \|\mathcal{D}_m(\mathit{\Pi}_{p^*}\boldsymbol{\upsilon}^* - \boldsymbol{\upsilon}_m^*)\| \le \|\mathcal{D}_m(\boldsymbol{\upsilon} - \mathit{\Pi}_{p^*}\boldsymbol{\upsilon}^*)\| \\ &\quad \le \|\mathcal{D}_m(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}_m^*)\| + \|\mathcal{D}_m(\mathit{\Pi}_{p^*}\boldsymbol{\upsilon}^* - \boldsymbol{\upsilon}_m^*)\|.\end{aligned}$$

Slightly modifying Lemma 4.A.8 with $\boldsymbol{\theta} = \boldsymbol{\upsilon}$ gives

$$\|\mathcal{D}_m(\mathit{\Pi}_{p^*}\boldsymbol{\upsilon}^* - \boldsymbol{\upsilon}_m^*)\| \le \Bigl(\alpha(m) + \tau(m) + 2\delta(2\mathtt{r}^*)\mathtt{r}^*\Bigr) \stackrel{\text{def}}{=} \mathtt{r}_\epsilon^*(m),$$

where due to Lemma 6.A.6 and the definition of $\mathtt{r}^* > 0$ in Lemma 6.3.5

$$\begin{aligned}&\mathtt{r}^* \le \mathtt{C}\sqrt{m}, \quad \alpha(m) = \mathtt{C}\left(m^{-\alpha-1/2} + \mathtt{C}_{bias} m^{-(\alpha-1)}\right)\sqrt{n}, \\ &\tau(m) \le \mathtt{C}m^{-2\alpha+1/2}\sqrt{n}.\end{aligned}$$

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 $\epsilon > 0$

$$\begin{aligned}(1-\epsilon)\|\mathcal{D}_m(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}_m^*)\|^2 &\le \|\mathcal{D}_m(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^*)\|^2 \qquad (6.A.34)\\ &\le (1+\epsilon)\|\mathcal{D}_m(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}_m^*)\|^2.\end{aligned}$$

Assume that $n \in \mathbb{N}$ is large enough to ensure that $\epsilon < 1/2$. Then we find for $\boldsymbol{\upsilon} \in \varUpsilon_\circ(\mathtt{r})^c$ and with Lemma 4.A.6 and (6.A.34) that

$$\|D(\varphi_{\boldsymbol{\theta}} - \varphi_{\boldsymbol{\theta}^*})\|^2 + \|\mathrm{H}_m(\boldsymbol{\eta} - \boldsymbol{\eta}^*)\|^2 \ge (1-\nu)\|\mathcal{D}_m(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^*)\|^2 \ge (1-\nu)\mathtt{r}^2/2.$$

2. Now we show (6.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 $\|D(\varphi_{\boldsymbol{\theta}} - \varphi_{\boldsymbol{\theta}^*})\|^2 \le \frac{1}{4}(1-\nu)\mathtt{r}^2$. 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

$$\Bigl(\boldsymbol{f}_{\boldsymbol{\eta}}(\mathbf{X}^\top\boldsymbol{\theta}) - \boldsymbol{f}_{\boldsymbol{\eta}^*}(\mathbf{X}^\top\boldsymbol{\theta}^*)\Bigr)^2 > 0,$$

to obtain a good lower bound.

Case 1: $\|D(\varphi_{\boldsymbol{\theta}} - \varphi_{\boldsymbol{\theta}^*})\|^2 \le \frac{1}{2}\tau\mathtt{r}^2$. In this case we simply calculate and find

$$\begin{aligned}&|\boldsymbol{f}_{\boldsymbol{\eta}}(\mathbf{X}^\top\boldsymbol{\theta}) - \boldsymbol{f}_{\boldsymbol{\eta}^*}(\mathbf{X}^\top\boldsymbol{\theta}^*)|^2 \\ &\quad \ge |\boldsymbol{f}_{\boldsymbol{\eta}}(\mathbf{X}^\top\boldsymbol{\theta}) - \boldsymbol{f}_{\boldsymbol{\eta}^*}(\mathbf{X}^\top\boldsymbol{\theta})|^2 \\ &\qquad -2|\boldsymbol{f}_{\boldsymbol{\eta}}(\mathbf{X}^\top\boldsymbol{\theta}) - \boldsymbol{f}_{\boldsymbol{\eta}^*}(\mathbf{X}^\top\boldsymbol{\theta})||\boldsymbol{f}_{\boldsymbol{\eta}^*}(\mathbf{X}^\top\boldsymbol{\theta}) - \boldsymbol{f}_{\boldsymbol{\eta}^*}(\mathbf{X}^\top\boldsymbol{\theta}^*)| \\ &\quad \ge |\boldsymbol{f}_{\boldsymbol{\eta}}(\mathbf{X}^\top\boldsymbol{\theta}) - \boldsymbol{f}_{\boldsymbol{\eta}^*}(\mathbf{X}^\top\boldsymbol{\theta})|^2 \\ &\qquad -2|\boldsymbol{f}_{\boldsymbol{\eta}}(\mathbf{X}^\top\boldsymbol{\theta}) - \boldsymbol{f}_{\boldsymbol{\eta}^*}(\mathbf{X}^\top\boldsymbol{\theta})|\boldsymbol{L}_{\boldsymbol{\eta}^*}s_{\mathbf{X}}\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|.\end{aligned}$$

Now

$$|\boldsymbol{f}_{\boldsymbol{\eta}}(\mathbf{X}^\top\boldsymbol{\theta}) - \boldsymbol{f}_{\boldsymbol{\eta}^*}(\mathbf{X}^\top\boldsymbol{\theta})| \geq |\boldsymbol{f}_{\boldsymbol{\eta}-\boldsymbol{\eta}^*}(\mathbf{X}^\top\boldsymbol{\theta})| - |\boldsymbol{f}_{(0,\boldsymbol{\varkappa}^*)}(\mathbf{X}^\top\boldsymbol{\theta})|.$$

We find with probability greater than $3/4$

$$\begin{aligned}
|\boldsymbol{f}_{\boldsymbol{\eta}-\boldsymbol{\eta}^*}(\mathbf{X}^\top\boldsymbol{\theta})| &= |\langle \boldsymbol{\eta}-\boldsymbol{\eta}^*, \boldsymbol{e}(\mathbf{X}^\top\boldsymbol{\theta})\rangle| \\
&\geq \|\mathrm{H}_m(\boldsymbol{\eta}-\boldsymbol{\eta}^*)\|\lambda_{\boldsymbol{e}} \\
&\geq \mathtt{r}\lambda_{\boldsymbol{e}}\frac{1}{2}\sqrt{(1-\nu^2)},
\end{aligned}$$

where

$$\lambda_{\boldsymbol{e}} \overset{\mathrm{def}}{=} \sup\left\{\lambda > 0 : \inf_{\substack{\boldsymbol{\eta}\in\mathbb{R}^m, \|\boldsymbol{\eta}\|=1 \\ \boldsymbol{\theta}\in S_1^{p,+}}} I\!P\left(|\langle\boldsymbol{\eta}, \mathrm{H}_m^{-1}\boldsymbol{e}(\mathbf{X}^\top\boldsymbol{\theta})\rangle| > \lambda\right) > 3/4\right\},$$

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 6.A.6

$$\begin{aligned}
\|\mathcal{H}_m^{1/2}\boldsymbol{\varkappa}^*\|^2 \; &< \; \left(17\|p_{\mathbf{X}^\top\boldsymbol{\theta}^*}\|_\infty \mathtt{C}_{\|\boldsymbol{f}^*\|} + 17^2\sqrt{36}s_{\mathbf{X}}^{p+1}L_{p_{\mathbf{X}}}\|\psi\|_\infty \mathtt{C}_{\|\boldsymbol{f}^*\|}^2\right) nm^{-2\alpha} \\
&\overset{\mathrm{def}}{=} \mathtt{C}_{\boldsymbol{\varkappa}}^2 m.
\end{aligned}$$

We use the Markov inequality to obtain

$$I\!P\left(|\boldsymbol{f}_{(0,\boldsymbol{\varkappa}^*)}(\mathbf{X}^\top\boldsymbol{\theta})|^2 \geq 4\mathtt{C}_{\boldsymbol{\varkappa}}\frac{m}{n}\right) \leq \frac{\|\mathcal{H}_m^{1/2}\boldsymbol{\varkappa}^*\|^2}{4\mathtt{C}_{\boldsymbol{\varkappa}}^2 m} \leq 1/4.$$

This implies that with probability greater than $1/2 = 3/4 - 1/4$

$$\begin{aligned}
|\boldsymbol{f}_{\boldsymbol{\eta}}(\mathbf{X}^\top\boldsymbol{\theta}) - \boldsymbol{f}_{\boldsymbol{\eta}^*}(\mathbf{X}^\top\boldsymbol{\theta})| &\geq \mathtt{r}\lambda_{\boldsymbol{e}}\frac{1}{2}\sqrt{(1-\nu^2)} - 4\mathtt{C}_{\boldsymbol{\varkappa}}\sqrt{\frac{m}{n}} \\
&\geq \frac{\sqrt{(1-\nu^2)}\lambda_{\boldsymbol{e}}}{4\sqrt{n}}\mathtt{r},
\end{aligned}$$

for

$$\mathtt{r} \geq \sqrt{m}\frac{4\mathtt{C}_{\boldsymbol{\varkappa}}}{\lambda_{\boldsymbol{e}}\sqrt{(1-\nu^2)}}.$$

We still have to account for the summand $\boldsymbol{L}_{\boldsymbol{\eta}^*}s_{\mathbf{X}}\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|$ via

$$\boldsymbol{L}_{\boldsymbol{\eta}^*}s_{\mathbf{X}}\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\| \leq \frac{\boldsymbol{L}_{\boldsymbol{\eta}^*}s_{\mathbf{X}}\sqrt{\tau(1-\nu^2)}}{2c_{\mathfrak{D}}\sqrt{n}}\mathtt{r}.$$

This gives for the choice of $\tau = \lambda_{\boldsymbol{e}} c_{\mathfrak{D}}/(8\boldsymbol{L}_{\boldsymbol{\eta}^*} s_{\mathbf{X}})$

$$\begin{aligned}
&|\boldsymbol{f}_{\boldsymbol{\eta}}(\mathbf{X}^\top \boldsymbol{\theta}) - \boldsymbol{f}_{\boldsymbol{\eta}^*}(\mathbf{X}^\top \boldsymbol{\theta})| - 2\boldsymbol{L}_{\boldsymbol{\eta}^*} s_{\mathbf{X}} \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\| \\
&\geq \left(\frac{\lambda_{\boldsymbol{e}}}{4} - \frac{\boldsymbol{L}_{\boldsymbol{\eta}^*} s_{\mathbf{X}} \sqrt{\tau}}{c_{\mathfrak{D}}} \right) \frac{\sqrt{(1-\nu^2)}}{\sqrt{n}} \mathtt{r} \\
&= \frac{\lambda_{\boldsymbol{e}} c_{\mathfrak{D}} \sqrt{(1-\nu^2)}}{8\sqrt{n}} \mathtt{r}
\end{aligned}$$

We obtain in case 1 that $\boldsymbol{Q}(2\mathtt{b}) \geq 1/2$ for

$$2\mathtt{b}/n \overset{\text{def}}{=} \frac{(1-\nu^2)\lambda_{\boldsymbol{e}}^2 c_{\mathfrak{D}}^2}{32n}.$$

Case 2: $\frac{1}{2}\tau(1-\nu)\mathtt{r}^2 \leq \|D(\varphi_{\boldsymbol{\theta}} - \varphi_{\boldsymbol{\theta}^*})\|^2 \leq \sqrt{2}\lambda_{\max} D^2$.
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 $\lambda([-\sqrt{\rho^2 - x^2}, \sqrt{\rho^2 - x^2}] \setminus [-\rho/2, \rho/2])$. Addressing the way a centered cylinder intersects with a shifted ball this gives that

$$\begin{aligned}
V(\rho\beta c/2) &\geq \lambda\big(C_{\rho,x,y}(0) \cap B_{s_{\mathbf{X}}}(x_0, y_0, 0) \\
&\qquad \cap \{(x, y, z) \in \mathbb{R}^2 \times \mathbb{R}^{p-2}; \\
&\qquad (x, y) \in \mathbb{R}^2 : -\operatorname{sign}(y_0) y \geq -\operatorname{sign}(y_0)\rho/2\}\big) \\
&\geq \lambda(B_{\rho/4}(0)) > 0, \qquad (6.A.35)
\end{aligned}$$

for the ball $B_{h/4}(0) \subset \mathbb{R}^p$. Now we can prove the claim. For any $(\boldsymbol{\theta}, \boldsymbol{\eta}) = \boldsymbol{\upsilon} \in \varUpsilon$, 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 $(\mathbf{Cond_{X\boldsymbol{\theta}^*}})$ for any $(\boldsymbol{\theta}, \boldsymbol{\eta}) = \boldsymbol{\upsilon} \in \varUpsilon$, there exist constants $c_{f'}$, $c_{p_\mathbf{X}}, h > 0$ and a value $(x_0, y_0) \in \{x^2 + y^2 \le s_\mathbf{X}\} \subset \mathbb{R}^2$ such that for $(x, y) \in \{(x - x_0)^2 + (y - y_0)^2 \le h^2\}$ we have $|f'_{\boldsymbol{\eta}^*}(x)| > c_{f'}$ and $p_\mathbf{X} \ge c_{p_\mathbf{X}}$. We can estimate using (6.A.35)

$$
\begin{aligned}
&I\!\!P\Big\{ \Big(\boldsymbol{f}_{\boldsymbol{\eta}^*}(\mathbf{X}^\top\boldsymbol{\theta}^*) - \boldsymbol{f}_{\boldsymbol{\eta}}(\mathbf{X}^\top\boldsymbol{\theta})\Big)^2 \ge c_{f'}^2 h^2 \beta^2/4 \Big\} \\
&\quad \ge \inf_{\substack{f \in C^1(\mathbb{R}),\, f'>0,\\ g:\mathbb{R}\to\mathbb{R}}} I\!\!P\Big(\{\mathbf{X} \in B_{s_\mathbf{X}}(0)\} \cap \{\mathbf{X} \in C_{h,x,y}(x_0, y_0)\} \\
&\qquad\qquad \cap \{|f(\alpha x + \beta y) - g(x)| \ge c_{f'} h\beta/2\} \Big) \\
&\quad \ge c_{p_\mathbf{X}} \inf_{\substack{f \in C^1(\mathbb{R}),\, f'>0,\\ g:\mathbb{R}\to\mathbb{R}}} \lambda\Big(B_{s_\mathbf{X}}(-x_0, -y_0, 0) \cap C_{h,x,y}(0) \\
&\qquad\qquad \cap \{|f(\alpha x + \beta y) - g(x)| \ge c_{f'} h\beta/2\} \Big) \\
&\quad = c_{p_\mathbf{X}} V(h\beta c_{f'}/2) \ge \lambda(B_{h/4}(0)) > 0.
\end{aligned}
$$

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 (6.A.34) that

$$
\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|^2 \ge \frac{2}{p\pi^2}\|\varphi_{\boldsymbol{\theta}} - \varphi_{\boldsymbol{\theta}^*}\|^2 \ge \frac{2}{p\pi^2\|D^2\|}\|D(\varphi_{\boldsymbol{\theta}} - \varphi_{\boldsymbol{\theta}^*})\|^2 \ge \frac{\tau}{p\pi^2\|D^2\|}\mathtt{r}^2.
$$

With Lemma 6.A.5 this implies

$$\beta^2 \geq \frac{\tau}{2p\pi^2 s_{\mathbf{X}}^2 \|f_{\mathbf{X}}\|_\infty^2 C_{\|\boldsymbol{f}^*\|}} \mathtt{r}^2/n.$$

Combined this yields that with

$$2\mathtt{b}/n \overset{\text{def}}{=} \frac{\tau c_{f'}^2 h^2}{4np\pi^2 s_{\mathbf{X}}^2 \|p_{\mathbf{X}}\|_\infty^2 C_{\|\boldsymbol{\eta}^*\|}},$$

it holds

$$\begin{aligned} &I\!\!P\Big\{ \Big(\boldsymbol{f}_{\boldsymbol{\eta}}(\mathbf{X}^\top \boldsymbol{\theta}) - \boldsymbol{f}_{\boldsymbol{\eta}^*}(\mathbf{X}^\top \boldsymbol{\theta}^*)\Big)^2 \geq 9\mathtt{b}\mathtt{r}^2/n\Big\} \\ &\qquad \geq c_{p_{\mathbf{X}}} \lambda(B_1^{p-2}) \lambda\left(B_h(0) \cap \{(x,y) \in \mathbb{R}^2 : |y| \leq h/2\}\right). \end{aligned}$$

This gives the claim. □

Proof of Condition $(\mathcal{L}\mathtt{r})$ with modeling bias

We show the following Lemma

Lemma 6.A.25. *We have with some* $\mathtt{C} > 0$ *and with* $\mathtt{r}^\circ > 0$ *from* (6.3.6) *that*

$$\begin{aligned} &I\!\!P\Bigg(\sup_{\boldsymbol{\upsilon} \in \Upsilon_\circ(\sqrt{n}\mathtt{r}^\circ)} |I\!\!E_\epsilon \mathcal{L}(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^*) - I\!\!E \mathcal{L}(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^*)| \geq \sqrt{\mathtt{x} + p^*[\mathtt{C}\log(p^*) + \log(\mathtt{r})]} \Bigg) \\ &\quad \leq \mathrm{e}^{-\mathtt{x}}. \end{aligned}$$

Proof. We bound

$$\begin{aligned} &\sup_{\boldsymbol{\upsilon} \in \Upsilon_\circ(\sqrt{n}\mathtt{r}^\circ)} |I\!\!E_\epsilon \mathcal{L}(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^*) - I\!\!E \mathcal{L}(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^*)| \\ &\quad \leq n \sup_{\boldsymbol{\upsilon} \in \Upsilon_\circ(\sqrt{n}\mathtt{r}^\circ)} \Bigg|(P_n - I\!\!P)\Big\{ \Big(g(\mathbf{X}_i) - \boldsymbol{f}_{\boldsymbol{\eta}^*}(\mathbf{X}_i^\top \boldsymbol{\theta}^*)\Big)^2 \\ &\qquad\qquad - \Big(g(\mathbf{X}_i) - \boldsymbol{f}_{\boldsymbol{\eta}}(\mathbf{X}_i^\top \boldsymbol{\theta})\Big)^2 \Big\}\Bigg| \\ &\quad \leq n \sup_{\boldsymbol{\upsilon} \in \Upsilon_\circ(\sqrt{n}\mathtt{r}^\circ)} \left|(P_n - I\!\!P) \left\{\boldsymbol{f}_{\boldsymbol{\eta}}(\mathbf{X}_i^\top \boldsymbol{\theta}) - \boldsymbol{f}_{\boldsymbol{\eta}^*}(\mathbf{X}_i^\top \boldsymbol{\theta}^*)\right\}^2\right| \\ &\qquad + n\mathtt{C}_{bias} \sup_{\boldsymbol{\upsilon} \in \Upsilon_\circ(\sqrt{n}\mathtt{r}^\circ)} \left|(P_n - I\!\!P) \left|\boldsymbol{f}_{\boldsymbol{\eta}}(\mathbf{X}_i^\top \boldsymbol{\theta}) - \boldsymbol{f}_{\boldsymbol{\eta}^*}(\mathbf{X}_i^\top \boldsymbol{\theta}^*)\right|\right|. \end{aligned}$$

Furthermore

$$\Big\{\boldsymbol{f}_{\boldsymbol{\eta}}(\mathbf{X}_i^\top\boldsymbol{\theta}) - \boldsymbol{f}_{\boldsymbol{\eta}^*}(\mathbf{X}_i^\top\boldsymbol{\theta}^*)\Big\}^2 \le \Big|\boldsymbol{f}_{\boldsymbol{\eta}}(\mathbf{X}_i^\top\boldsymbol{\theta}) - \boldsymbol{f}_{\boldsymbol{\eta}^*}(\mathbf{X}_i^\top\boldsymbol{\theta}^*)\Big| \\ \big(\|\boldsymbol{f}_{\boldsymbol{\eta}^*}\|_\infty + \|\boldsymbol{f}_{\boldsymbol{\eta}^*_m}\|_\infty + \mathtt{C}\mathtt{r}\sqrt{m}/\sqrt{n}\big)\,.$$

Thus we have

$$\begin{aligned} &\sup_{\boldsymbol{\upsilon}\in\varUpsilon_\circ(\sqrt{n}\mathtt{r}^\circ)} |\mathbb{E}_\epsilon\mathcal{L}(\boldsymbol{\upsilon},\boldsymbol{\upsilon}^*) - \mathbb{E}\mathcal{L}(\boldsymbol{\upsilon},\boldsymbol{\upsilon}^*)| \\ &\quad\le n\left(\mathtt{C}_{bias} + \|\boldsymbol{f}_{\boldsymbol{\eta}^*}\|_\infty + \|\boldsymbol{f}_{\boldsymbol{\eta}^*_m}\|_\infty + \mathtt{C}\mathtt{r}^\circ\sqrt{m}\right) \\ &\qquad \sup_{\boldsymbol{\upsilon}\in\varUpsilon_\circ(\sqrt{n}\mathtt{r}^\circ)} \Big|(P_n - \mathbb{P})\Big|\boldsymbol{f}_{\boldsymbol{\eta}}(\mathbf{X}_i^\top\boldsymbol{\theta}) - \boldsymbol{f}_{\boldsymbol{\eta}^*}(\mathbf{X}_i^\top\boldsymbol{\theta}^*)\Big|\Big|\,. \end{aligned}$$

Define $\boldsymbol{\zeta}_{\mathbf{X}}(\boldsymbol{\upsilon}) \stackrel{\text{def}}{=} (P_n - \mathbb{P})|\boldsymbol{f}_{\boldsymbol{\eta}}(\mathbf{X}_i^\top\boldsymbol{\theta}) - \boldsymbol{f}_{\boldsymbol{\eta}^*}(\mathbf{X}_i^\top\boldsymbol{\theta}^*)|\,.$ Then we find using that $\mathtt{r}^\circ \le \mathtt{C}\sqrt{p^*\log(p^*) + \mathtt{x}}$

$$\begin{aligned} &\sup_{\boldsymbol{\upsilon}\in\varUpsilon_\circ(\sqrt{n}\mathtt{r}^\circ)} |\mathbb{E}_\epsilon\mathcal{L}(\boldsymbol{\upsilon},\boldsymbol{\upsilon}^*) - \mathbb{E}\mathcal{L}(\boldsymbol{\upsilon},\boldsymbol{\upsilon}^*)| \\ &\quad\le n\mathtt{C}m^{3/2} \sup_{\boldsymbol{\upsilon}\in\varUpsilon_\circ(\sqrt{n}\mathtt{r}^\circ)} |\boldsymbol{\zeta}_{\mathbf{X}}(\boldsymbol{\upsilon}) - \boldsymbol{\zeta}_{\mathbf{X}}(\boldsymbol{\upsilon}^*)|\,. \end{aligned}$$

We want to use Lemma 6.A.2. Define $\varUpsilon_0 = \{\boldsymbol{\upsilon}^*\}$ and with $\mathtt{r}_k = 2^{-k}\mathtt{r}$ with $\mathtt{r} > 0$ to be specified later the sequence of sets $\varUpsilon_k$ each with minimal cardinality such that

$$\varUpsilon_m \subset \bigcup_{\boldsymbol{\upsilon}\in\varUpsilon_k} B_{\mathtt{r}_k}(\boldsymbol{\upsilon}), \quad B_{\mathtt{r}}(\boldsymbol{\upsilon}) \stackrel{\text{def}}{=} \{\boldsymbol{\upsilon}^\circ \in \varUpsilon_m,\, \|\mathcal{D}(\boldsymbol{\upsilon}^\circ - \boldsymbol{\upsilon})\| \le \mathtt{r}\}.$$

We estimate for an application of the bounded differences inequality

$$\Big|\Big\{\boldsymbol{f}_{\boldsymbol{\eta}}(\mathbf{X}_i^\top\boldsymbol{\theta}) - \boldsymbol{f}_{\boldsymbol{\eta}'}(\mathbf{X}_i^\top\boldsymbol{\theta}')\Big\}\Big| \le \|\boldsymbol{f}_{\boldsymbol{\eta}-\boldsymbol{\eta}'}\|_\infty + \|\boldsymbol{f}'_{\boldsymbol{\eta}}\|_\infty\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|.$$

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 $\boldsymbol{\upsilon}_k \in \Upsilon_k$

$$I\!P\left(n \inf_{\Upsilon_{k-1}} |\boldsymbol{\zeta}_{\mathbf{X}}(\boldsymbol{\upsilon}_k) - \boldsymbol{\zeta}_{\mathbf{X}}(\boldsymbol{\upsilon}_{k-1})| \ge t\mathtt{C}_{\boldsymbol{\zeta}} m^{3/2}\frac{\mathtt{r}_{k-1}}{c_{\mathfrak{D}}}\right) \le \mathrm{e}^{-t^2}.$$

Define $\mathtt{r} \overset{\mathrm{def}}{=} \frac{(1-1/\sqrt{2})}{m^3}$ then we find

$$I\!P\left(n \inf_{\Upsilon_{k-1}} |\boldsymbol{\zeta}_{\mathbf{X}}(\boldsymbol{\upsilon}_k) - \boldsymbol{\zeta}_{\mathbf{X}}(\boldsymbol{\upsilon}_{k-1})| \ge \mathtt{C} m^{-3/2} t 2^{-(k-1)}(1 - 1/\sqrt{2})\right) \le \mathrm{e}^{-t^2},$$

and

$$\begin{aligned}|\Upsilon_k| \le \exp\Big\{\Big(&\log(2)k + \log(\mathtt{r}^{\circ}) + \log(n)/2 + 3\log(m)\\ &+ \log(1 - 1/\sqrt{2})\Big)p^*\Big\}.\end{aligned}$$

Set

$$\begin{aligned}\boldsymbol{T}(n,m) &\overset{\mathrm{def}}{=} \log(\mathtt{r}^{\circ}) + \log(n)/2 + 3\log(m) + \log(1 - 1/\sqrt{2},\\ t &\overset{\mathrm{def}}{=} \sqrt{\mathtt{x} + 1 + \log(2) + p^*\left(\log(2) + \boldsymbol{T}(n,m)\right)},\end{aligned}$$

then we infer with Lemma 6.A.2

$$\begin{aligned}&I\!P\left(\sup_{\boldsymbol{\upsilon}\in\Upsilon_{\circ}(\sqrt{n}\mathtt{r}^{\circ})} |I\!E_{\epsilon}\mathcal{L}(\boldsymbol{\upsilon},\boldsymbol{\upsilon}^*) - I\!E\mathcal{L}(\boldsymbol{\upsilon},\boldsymbol{\upsilon}^*)| \ge \mathtt{C}t\right)\\ &\le I\!P\left(n \sup_{\boldsymbol{\upsilon}\in\Upsilon_{\circ}(\sqrt{n}\mathtt{r}^{\circ})} |\boldsymbol{\zeta}_{\mathbf{X}}(\boldsymbol{\upsilon}) - \boldsymbol{\zeta}_{\mathbf{X}}(\boldsymbol{\upsilon}^*)| \ge \mathtt{C} m\log(m)t\right)\\ &\le \sum_{k=1}^{\infty} \exp\Big\{p^*\Big[(\log(2)k + \boldsymbol{T}(n,m)) - 2^{k-1}\left(\log(2) + \boldsymbol{T}(n,m)\right)\Big]\\ &\quad -2^{k-1}(\mathtt{x} + 1 + \log(2))\Big\}\\ &\le \mathrm{e}^{-\mathtt{x}}.\end{aligned}$$

□

We have as in the proof of Lemma 4.A.7

$$-I\!E\mathcal{L}(\boldsymbol{\upsilon}^*, \boldsymbol{\upsilon}^*_m) = I\!E\mathcal{L}(\boldsymbol{\upsilon}^*_m, \boldsymbol{\upsilon}^*) \ge I\!E\mathcal{L}(\Pi_{p^*}\boldsymbol{\upsilon}^*, \boldsymbol{\upsilon}^*) \ge -\mathtt{r}^{*2}. \quad (6.A.36)$$

Combining this lemma and Equation (6.A.36) with Lemma 6.A.7 and Lemma 6.3.6 we find for $\|\mathcal{D}_m(\boldsymbol{\upsilon}-\boldsymbol{\upsilon}_m^*)\|^2 = \mathtt{r}^2 \geq 2\mathtt{r}^{*2}$ that with probability greater than $1-2\mathrm{e}^{-\mathtt{x}}$

$$-I\!\!E_{\epsilon}\mathcal{L}(\boldsymbol{\upsilon},\boldsymbol{\upsilon}_m^*) \geq \mathtt{b}\mathtt{r}^2/2 - \sqrt{\mathtt{x}+\mathtt{C}p^*[\log(p^*)+\log(n)]} - \mathtt{r}^{*2}.$$

Consequently we get for $\mathtt{r}$ that additionally satisfies

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

that

$$-I\!\!E_{\epsilon}\mathcal{L}(\boldsymbol{\upsilon},\boldsymbol{\upsilon}_m^*) \geq \mathtt{b}\mathtt{r}^2/4 \overset{\mathrm{def}}{=} \mathtt{b}_{bias}\mathtt{r}^2.$$

Finally observe that by definition $\mathcal{L}(\boldsymbol{\upsilon},\boldsymbol{\upsilon}_m^*) = \mathcal{L}_m(\boldsymbol{\upsilon},\boldsymbol{\upsilon}_m^*)$.

6.A.9 Proof of Lemma 6.3.1

Proof. Note that with the definitions and with some $\boldsymbol{\upsilon} \in \varUpsilon_{m,0}(\mathtt{r})$, $\boldsymbol{\gamma}_0 \in \mathbb{R}^{p^*}$ with $\|\boldsymbol{\gamma}_0\| = 1$

$$\begin{aligned}
&\|\mathcal{D}_m^{-1}\nabla(I\!\!E - I\!\!E_{\varepsilon})[\mathcal{L}_m(\boldsymbol{\upsilon}_m^*) - \mathcal{L}_m(\boldsymbol{\upsilon})]\| \\
&\quad\leq \sup_{\boldsymbol{\upsilon}\in\varUpsilon_{m,0}(\mathtt{r})} \|\mathcal{D}_m^{-1}(I\!\!E - I\!\!E_{\varepsilon})\left[\nabla^2\mathcal{L}_m(\boldsymbol{\upsilon})\right]\mathcal{D}_m^{-1}\|\mathtt{r} \\
&\quad\leq \frac{1}{\sqrt{n}c_{\mathcal{D}}}\|(I\!\!E - I\!\!E_{\varepsilon})\left[\mathcal{D}_m^{-1}\nabla^2\mathcal{L}_m(\boldsymbol{\upsilon}_m^*)\right]\|\mathtt{r} \\
&\qquad + \sup_{\boldsymbol{\upsilon}\in\varUpsilon_{m,0}(\mathtt{r})} \left\|(I\!\!E - I\!\!E_{\varepsilon})\left[\mathcal{D}_m^{-1}\left(\left[\nabla^2\mathcal{L}_m(\boldsymbol{\upsilon})\right] - \left[\nabla^2\mathcal{L}_m(\boldsymbol{\upsilon}_m^*)\right]\right)\mathcal{D}_m^{-1}\right]\right\|\mathtt{r}.
\end{aligned}$$

For the first term we obtain with Lemma 6.A.31 and with some constant $\mathtt{C} > 0$

$$I\!\!P\left(\frac{1}{\sqrt{n}c_{\mathcal{D}}}\|(I\!\!E - I\!\!E_{\varepsilon})\left[\mathcal{D}^{-1}\nabla^2\mathcal{L}_m(\boldsymbol{\upsilon}_m^*)\right]\|\mathtt{r} \geq \mathtt{C}\sqrt{\log(p^*)+\mathtt{x}}\mathtt{r}/\sqrt{n}\right) \leq \mathrm{e}^{-\mathtt{x}}.$$

For the second term we can use similar arguments to those of Lemma 6.3.6 to find with some constant $\mathtt{C} > 0$ that

$$\begin{aligned}
I\!\!P\Bigg(&\sup_{\boldsymbol{\upsilon}\in\varUpsilon_{m,0}(\mathtt{r})} \left\|(I\!\!E - I\!\!E_{\varepsilon})\left[\mathcal{D}_m^{-1}\left(\left[\nabla^2\mathcal{L}_m(\boldsymbol{\upsilon})\right] - \left[\nabla^2\mathcal{L}_m(\boldsymbol{\upsilon}_m^*)\right]\right)\mathcal{D}_m^{-1}\right]\right\| \\
&\geq \mathtt{C}\sqrt{\mathtt{x}+p^*\log(p^*)}/\sqrt{n}\Bigg) \leq \mathrm{e}^{-\mathtt{x}}.
\end{aligned}$$

Adding $\log(2)$ to $\mathtt{x}$ in the above bounds we get the claim after increasing the constants appropriately. □

6.A.10 Condition $(bias'')$ is satisfied

Lemma 6.A.26. *Under the conditions of Proposition 6.2.2 condition* $(\boldsymbol{bias''})$ *is satisfied.*

Proof. It suffices to show that

$$\operatorname{Cov}(\nabla_{\boldsymbol{\theta}}\left(\ell_i(\boldsymbol{\upsilon}_m^*)-\ell_i(\boldsymbol{\upsilon}^*)\right))\to 0,\ \ \operatorname{Cov}(\nabla_{(\eta_1,\dots,\eta_m)}\left(\ell_i(\boldsymbol{\upsilon}_m^*)-\ell_i(\boldsymbol{\upsilon}^*)\right))\to 0.$$

We calculate

$$\begin{aligned}
&\|\operatorname{Cov}(\nabla_{\boldsymbol{\theta}}\left(\ell_i(\boldsymbol{\upsilon}_m^*)-\ell_i(\boldsymbol{\upsilon}^*)\right))\|\\
&\quad\le \mathbb{E}\|\left(\boldsymbol{f}'_{\boldsymbol{\eta}_m^*}(\mathbf{X}_i^\top\boldsymbol{\theta}_m^*)-\boldsymbol{f}'_{\boldsymbol{\eta}^*}(\mathbf{X}_i^\top\boldsymbol{\theta}^*)\right)\nabla\varPhi(\boldsymbol{\theta})^\top\mathbf{X}_i\|^2\\
&\quad\le s_{\mathbf{X}}^2\mathbb{E}\|\boldsymbol{f}'_{\boldsymbol{\eta}_m^*}(\mathbf{X}_i^\top\boldsymbol{\theta}_m^*)-\boldsymbol{f}'_{\boldsymbol{\eta}^*}(\mathbf{X}_i^\top\boldsymbol{\theta}^*)\|^2\\
&\quad\le 4s_{\mathbf{X}}^2\left(\mathbb{E}\|\boldsymbol{f}'_{\boldsymbol{\eta}_m^*-\boldsymbol{\eta}^*}(\mathbf{X}_i^\top\boldsymbol{\theta}^*)\|^2+\mathbb{E}\|\boldsymbol{f}'_{\boldsymbol{\eta}_m^*}(\mathbf{X}_i^\top\boldsymbol{\theta}_m^*)-\boldsymbol{f}'_{\boldsymbol{\eta}_m^*}(\mathbf{X}_i^\top\boldsymbol{\theta}^*)\|^2\right)\\
&\quad\le 4s_{\mathbf{X}}^2\left(\sum_{k=0}^{\infty}\|\boldsymbol{e}_k'\|_\infty(\eta_{mk}^*-\eta_k^*)\right)^2\\
&\qquad+4s_{\mathbf{X}}^4\left(\sum_{k=0}^{m-1}\|\boldsymbol{e}_k''\|_\infty\eta_{mk}^*\right)^2\|\boldsymbol{\theta}_m^*-\boldsymbol{\theta}^*\|^2.
\end{aligned}$$

We estimate separately

$$\begin{aligned}
\sum_{k=0}^{\infty}\|\boldsymbol{e}_k'\|_\infty(\eta_{mk}^*-\eta_k^*)&\le\mathtt{C}\|\psi'\|_\infty\left(\sum_{k=0}^{m-1}k^{3/2}(\eta_{mk}^*-\eta_k^*)+\sum_{k=m}^{\infty}k^{3/2}\eta_k^*\right)\\
&\le\mathtt{C}\|\psi'\|_\infty\left(m^2\|\boldsymbol{\eta}_m^*-\boldsymbol{\eta}^*\|+\left(\sum_{k=m}^{\infty}k^{-2\alpha-3}\right)^{1/2}\left(\sum_{k=m}^{\infty}2^\alpha\eta_k^{*2}\right)^{1/2}\right)\\
&\le\mathtt{C}\|\psi'\|_\infty\Bigg(m^2\frac{1}{\sqrt{n}c_{\mathcal{D}}}\|\mathcal{D}_m(\varPi_{p^*}\boldsymbol{\upsilon}^*-\boldsymbol{\upsilon}_m^*)\|\\
&\qquad\qquad+\sqrt{(2\alpha-3)/(2\alpha-4)}\left(\sum_{k=m}^{\infty}k^{2\alpha}\eta_k^{*2}\right)^{1/2}\Bigg)
\end{aligned}$$

The last term tends to 0 because of Lemma 6.3.5, 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

$$\begin{aligned}
&\left(\sum_{k=0}^{m-1}\|\boldsymbol{e}_k''\|_\infty \eta_{mk}^*\right)\|\boldsymbol{\theta}_m^* - \boldsymbol{\theta}^*\| \le \|\psi''\|_\infty \|\boldsymbol{\theta}_m^* - \boldsymbol{\theta}^*\| \left(\sum_{k=0}^{m-1} k^{5/2}\eta_{mk}^*\right)\\
&\quad\le \|\psi''\|_\infty \|\boldsymbol{\theta}_m^* - \boldsymbol{\theta}^*\| \left\{\left(\sum_{k=0}^{m-1} k^{2\alpha-5}\right)^{1/2}\left(\sum_{k=0}^{m-1} k^{2\alpha}\eta_k^*\right)^{1/2}\right.\\
&\qquad\left. + \frac{1}{\sqrt{n}c_{\mathcal{D}}}\left(\sum_{k=0}^{m-1} k^5\right)^2 \|\mathcal{D}(\boldsymbol{\upsilon}^* - \boldsymbol{\upsilon}_m^*)\|\right\}\\
&\quad\le \|\psi''\|_\infty \|\boldsymbol{\theta}_m^* - \boldsymbol{\theta}^*\| \left\{m\mathtt{C}_{\|\boldsymbol{\eta}^*\|} + \frac{1}{\sqrt{n}c_{\mathcal{D}}} m^3 \|\mathcal{D}_m(\Pi_{p^*}\boldsymbol{\upsilon}^* - \boldsymbol{\upsilon}_m^*)\|\right\}\\
&\quad\le \|\mathcal{D}_m(\Pi_{p^*}\boldsymbol{\upsilon}^* - \boldsymbol{\upsilon}_m^*)\| \frac{1}{\sqrt{n}c_{\mathcal{D}}} m\mathtt{C}_{\|\boldsymbol{\eta}^*\|}\|\psi''\|_\infty\\
&\qquad + \frac{1}{nc_{\mathcal{D}}^2} m^3 \|\mathcal{D}_m(\Pi_{p^*}\boldsymbol{\upsilon}^* - \boldsymbol{\upsilon}_m^*)\|^2 \|\psi''\|_\infty.
\end{aligned}$$

Again the last term tends to 0. Similarly we calculate

$$\begin{aligned}
&\mathrm{Cov}(\nabla_{(\eta_1,\ldots,\eta_m)}(\ell_i(\boldsymbol{\upsilon}_m^*) - \ell_i(\boldsymbol{\upsilon}^*))) \le I\!\!E\|\boldsymbol{e}(\mathbf{X}_i^\top\boldsymbol{\theta}_m^*) - \boldsymbol{e}(\mathbf{X}_i^\top\boldsymbol{\theta}^*)\|^2\\
&\quad\le s_{\mathbf{X}}^2\|\psi'\|_\infty^2 \|\boldsymbol{\theta}_m^* - \boldsymbol{\theta}^*\|^2 \left(\sum_{k=0}^{m-1} k^{3/2}\right)^2\\
&\quad\le s_{\mathbf{X}}^2\|\psi'\|_\infty \frac{1}{nc_{\mathcal{D}}} m^3 \|\mathcal{D}_m(\Pi_{p^*}\boldsymbol{\upsilon}^* - \boldsymbol{\upsilon}_m^*)\|^2,
\end{aligned}$$

which again is a zero sequence. This gives the claim. □

6.A.11 Proof of Lemma 6.3.11

Proof. Define

$$\boldsymbol{\theta}_{l^*} \overset{\text{def}}{=} \operatorname*{argmin}_{\boldsymbol{\theta}\in G_N} \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|.$$

Then by definition

$$\begin{aligned}
&\max_{\boldsymbol{\eta}} \mathcal{L}_m(\widetilde{\boldsymbol{\upsilon}}^{(0)}, \boldsymbol{\upsilon}_m^*) \\
&\quad \geq \mathcal{L}_m((\boldsymbol{\theta}_{l^*}, \widetilde{\boldsymbol{\eta}}_{l^*}^{(0)}), \boldsymbol{\upsilon}_m^*) \geq \mathcal{L}_m((\boldsymbol{\theta}_{l^*}, \boldsymbol{\eta}_m^*), \boldsymbol{\upsilon}_m^*) \\
&\quad = -\sum_{i=1}^n (\boldsymbol{f}_{\boldsymbol{\eta}_m^*}(\mathbf{X}_i^\top \boldsymbol{\theta}_m^*) - \boldsymbol{f}_{\boldsymbol{\eta}_m^*}(\mathbf{X}_i^\top \boldsymbol{\theta}_{l^*}))^2 \\
&\qquad + (g(\mathbf{X}_i) - \boldsymbol{f}_{\boldsymbol{\eta}_m^*}(\mathbf{X}_i^\top \boldsymbol{\theta}_m^*))(\boldsymbol{f}_{\boldsymbol{\eta}_m^*}(\mathbf{X}_i^\top \boldsymbol{\theta}_m^*) - \boldsymbol{f}_{\boldsymbol{\eta}_m^*}(\mathbf{X}_i^\top \boldsymbol{\theta}_{l^*})) \\
&\qquad - (\boldsymbol{f}_{\boldsymbol{\eta}_m^*}(\mathbf{X}_i^\top \boldsymbol{\theta}_m^*) - \boldsymbol{f}_{\boldsymbol{\eta}_m^*}(\mathbf{X}_i^\top \boldsymbol{\theta}_{l^*}))\varepsilon_i.
\end{aligned}$$

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 $\boldsymbol{\theta}^\circ \in {\boldsymbol{\theta}_m^*}^\perp$

$$I\!\!E\left[(g(\mathbf{X}_i) - \boldsymbol{f}_{\boldsymbol{\eta}_m^*}(\mathbf{X}_i^\top \boldsymbol{\theta}_m^*))\boldsymbol{f}'_{\boldsymbol{\eta}_m^*}(\mathbf{X}_i^\top \boldsymbol{\theta}_m^*)\mathbf{X}^\top \boldsymbol{\theta}^\circ\right] = 0,$$

We estimate with Taylor expansion

$$\begin{aligned}
&\left\|(\boldsymbol{f}_{\boldsymbol{\eta}_m^*}(\mathbf{X}_i^\top \boldsymbol{\theta}_m^*) - \boldsymbol{f}_{\boldsymbol{\eta}_m^*}(\mathbf{X}_i^\top \boldsymbol{\theta}_{l^*})) - \boldsymbol{f}'_{\boldsymbol{\eta}_m^*}(\mathbf{X}_i^\top \boldsymbol{\theta}_m^*)\mathbf{X}^\top \nabla \Phi_{\boldsymbol{\theta}_m^*}(\varphi_{\boldsymbol{\theta}_{l^*}} - \varphi_{\boldsymbol{\theta}_m^*})\right\| \\
&\quad \leq \mathtt{C}\sqrt{m}\|\boldsymbol{\theta}_{l^*} - \boldsymbol{\theta}_m^*\|.
\end{aligned}$$

Furthermore with the bounded differences inequality

$$\begin{aligned}
I\!\!P\Big(n\Big|(P_n - I\!\!P)(g(\mathbf{X}_i) - \boldsymbol{f}_{\boldsymbol{\eta}_m^*}(\mathbf{X}_i^\top \boldsymbol{\theta}_m^*))(\boldsymbol{f}_{\boldsymbol{\eta}_m^*}(\mathbf{X}_i^\top \boldsymbol{\theta}_m^*) - \boldsymbol{f}_{\boldsymbol{\eta}_m^*}(\mathbf{X}_i^\top \boldsymbol{\theta}_{l^*}))\Big| \\
\geq \sqrt{\mathtt{x}}\mathtt{C}_{bias}\mathtt{C} s_{\mathbf{X}}\tau\Big) \leq \mathrm{e}^{-\mathtt{x}}.
\end{aligned}$$

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 $\lambda \le \sqrt{n}\widetilde{\mathtt{g}}/(\mathtt{C}s_{\mathbf{X}}\tau)$

$$\begin{aligned}
&\mathbb{P}^{\varepsilon}\left(\sum_{i=1}^{n}(\boldsymbol{f}_{\boldsymbol{\eta}^*_m}(\mathbf{X}_i^{\top}\boldsymbol{\theta}^*_m) - \boldsymbol{f}_{\boldsymbol{\eta}^*_m}(\mathbf{X}_i^{\top}\boldsymbol{\theta}_{l^*}))\varepsilon_i \ge \sqrt{n}t\right)\\
&\qquad \le \exp\{-\lambda t\}\mathbb{E}^{\varepsilon}\left[\exp\{\lambda\sum_{i=1}^{n}(\boldsymbol{f}_{\boldsymbol{\eta}^*_m}(\mathbf{X}_i^{\top}\boldsymbol{\theta}^*_m) - \boldsymbol{f}_{\boldsymbol{\eta}^*_m}(\mathbf{X}_i^{\top}\boldsymbol{\theta}_{l^*}))\varepsilon_i/\sqrt{n}\}\right]\\
&\qquad \le \exp\{-\lambda t\}\prod_{i=1}^{n}\mathbb{E}^{\varepsilon}\left[\exp\{\lambda(\boldsymbol{f}_{\boldsymbol{\eta}^*_m}(\mathbf{X}_i^{\top}\boldsymbol{\theta}^*_m) - \boldsymbol{f}_{\boldsymbol{\eta}^*_m}(\mathbf{X}_i^{\top}\boldsymbol{\theta}_{l^*}))\varepsilon_i/\sqrt{n}\}\right]\\
&\qquad \le \exp\{-\lambda t + \widetilde{\nu}^2\mathtt{C}^2 s_{\mathbf{X}}^2\tau^2\lambda^2/2\}.
\end{aligned}$$

Setting $\lambda = \frac{t}{\widetilde{\nu}^2\mathtt{C}^2 s_{\mathbf{X}}^2\tau^2}$ we get

$$\mathbb{P}^{\varepsilon}\left(\sum_{i=1}^{n}(\boldsymbol{f}_{\boldsymbol{\eta}^*_m}(\mathbf{X}_i^{\top}\boldsymbol{\theta}^*_m) - \boldsymbol{f}_{\boldsymbol{\eta}^*_m}(\mathbf{X}_i^{\top}\boldsymbol{\theta}_{l^*}))\varepsilon_i \ge \sqrt{n}t\right) \le \exp\left\{-\frac{t^2}{2\widetilde{\nu}^2\mathtt{C}^2 s_{\mathbf{X}}^2\tau^2}\right\}.$$

With $t = \widetilde{\nu}\mathtt{C}s_{\mathbf{X}}\tau\sqrt{2\mathtt{x}}$ and $\mathtt{x} \le 2\widetilde{\nu}^2\widetilde{\mathtt{g}}^2 n/(\mathtt{C}^2 s_{\mathbf{X}}^2\tau^2)$ this gives

$$\mathbb{P}^{\varepsilon}\left(\sum_{i=1}^{n}(\boldsymbol{f}_{\boldsymbol{\eta}^*_m}(\mathbf{X}_i^{\top}\boldsymbol{\theta}^*_m) - \boldsymbol{f}_{\boldsymbol{\eta}^*_m}(\mathbf{X}_i^{\top}\boldsymbol{\theta}_{l^*}))\varepsilon_i \ge \widetilde{\nu}\mathtt{C}s_{\mathbf{X}}\tau\sqrt{2n\mathtt{x}}\right) \le \mathrm{e}^{-\mathtt{x}}.$$

Consequently

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

For the second claim note that by Lemma 6.3.7 the conditions $(\mathcal{E}\mathcal{D}_1)$ and $(\mathcal{L}_0)$ from Section 4.2.1 hold for all $\mathtt{r} \le \sqrt{n}\mathtt{r}^{\circ}$. We define

$$\mathrm{K}_0(\mathtt{x}) \overset{\text{def}}{=} \mathtt{C}\left\{(1+\mathtt{C}_{bias}\sqrt{m})n\tau^2 + (1+\mathtt{C}_{bias})\sqrt{\mathtt{x}}\tau\sqrt{n}\right\}.$$

This implies with Lemma 6.3.7 and Theorem 3.3.2 that

$$\begin{aligned}
\mathrm{R}_0(\mathtt{x}) &\le \mathtt{C}m^{3/2}\sqrt{p^*(1+\mathtt{C}_{bias}\log(n)) + \mathtt{x} + (1+\mathtt{C}_{bias}\sqrt{m})n\tau^2 + \sqrt{n}\tau\sqrt{\mathtt{x}}}\\
&\le \mathtt{C}m^{3/2}\sqrt{p^*(1+\mathtt{C}_{bias}\log(n)) + \mathtt{x}}\\
&\quad + \mathtt{C}m^{3/2}\sqrt{(1+\mathtt{C}_{bias}\sqrt{m})n\tau^2 + \sqrt{n}\tau\sqrt{\mathtt{x}}}.
\end{aligned}$$

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

$$\mathrm{R}_0(\mathtt{x}) \le \mathtt{C}m^{3/2}\sqrt{p^*(1+\mathtt{C}_{bias}\log(n)) + \mathtt{x}} + \mathtt{C}(\sqrt{n} + m^{1/2}n^{1/4}).$$

Repeating the same arguments as in Section 6.3.4 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

$$\mathrm{R}_0(\mathtt{x}) \le \mathtt{C}\sqrt{p^*(1+\mathtt{C}_{bias}\log(n))+\mathtt{x}+(1+\mathtt{C}_{bias}\sqrt{m})n\tau^2+\sqrt{n}\tau\sqrt{\mathtt{x}}}.$$

Furthermore with Lemma 6.3.7

$$\epsilon \overset{\mathrm{def}}{=} \delta(\mathtt{r})/\mathtt{r}+\omega = \mathtt{C}\frac{m^{3/2}+\mathtt{C}_{bias}m^{5/2}}{\sqrt{n}}.$$

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.

□

6.A.12 Proof of Lemma 6.3.12

Auxiliary results

First we need the following uniform bounds:

Lemma 6.A.27. *There is a generic constant* $\mathtt{C}>0$ *such that for any pair* $\boldsymbol{\upsilon},\boldsymbol{\upsilon}^{\circ}\in\varUpsilon_{\circ}(R_0)$ *with* $\varsigma_{i,m}$ *from* (6.A.1)

$$\|\nabla\varsigma_{i,m}(\boldsymbol{\upsilon}^*)\| \le \mathtt{C}(\|\boldsymbol{f}'_{\boldsymbol{\eta}^*}\|_\infty+\|\boldsymbol{f}''_{\boldsymbol{\eta}^*}\|_\infty), \tag{6.A.37}$$

$$\begin{aligned}&\|\mathcal{D}_m^{-1/2}\nabla\varsigma_{i,m}(\boldsymbol{\upsilon})-\mathcal{D}_m^{-1/2}\nabla\varsigma_{i,m}(\boldsymbol{\upsilon}^{\circ})\| \\ &\quad\le \frac{\mathtt{C}}{c_{\mathcal{D}}\sqrt{n}}m\left(m^{3/2}+\left(C_{\|\boldsymbol{f}\|}+\frac{m^2(R_0+\mathtt{r}^*)}{nc_{\mathcal{D}}^2}\right)m^{1/2}\right)\|\mathcal{D}_m(\boldsymbol{\upsilon}-\boldsymbol{\upsilon}^{\circ})\|.\end{aligned} \tag{6.A.38}$$

Proof. Since $\nabla^2_{\boldsymbol{\eta}}\zeta(\boldsymbol{\upsilon})=0$ we can estimate with help of Lemma 6.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

$$\begin{aligned}\|\nabla_{\boldsymbol{\theta}}\varsigma_{i,m}(\boldsymbol{\upsilon}^*)\| &\le \|\boldsymbol{f}''_{\boldsymbol{\eta}^*}(\mathbf{X}_i^\top\boldsymbol{\theta}^*)\nabla\varPhi(\boldsymbol{\theta}^*)^\top\mathbf{X}_i\mathbf{X}_i^\top\nabla\varPhi(\boldsymbol{\theta}^*)\|\\ &\quad+\|\boldsymbol{f}'_{\boldsymbol{\eta}^*}(\mathbf{X}_i^\top\boldsymbol{\theta}^*)\mathbf{X}_i\nabla^2\varPhi(\boldsymbol{\theta}^\top\mathbf{X}_i)[\mathbf{X}_i,\cdot,\cdot]\|\\ &\le \mathtt{C}_0 s_{\mathbf{X}}^2\left(|\boldsymbol{f}'_{\boldsymbol{\eta}^*}(\mathbf{X}_i^\top\boldsymbol{\theta})|+|\boldsymbol{f}''_{\boldsymbol{\eta}^*}(\mathbf{X}_i^\top\boldsymbol{\theta})|\right)\le \mathtt{C}(\|\boldsymbol{f}'_{\boldsymbol{\eta}^*}\|_\infty+\|\boldsymbol{f}''_{\boldsymbol{\eta}^*}\|_\infty),\end{aligned}$$

This gives (6.A.37). For the proof of (6.A.38) we again use $\nabla^2_{\boldsymbol{\eta}}\zeta(\boldsymbol{\upsilon})=0$ and estimate with help of Lemma 6.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

$$\begin{aligned}&\|\nabla_{\boldsymbol{\theta}}\varsigma_{i,m}(\boldsymbol{\upsilon})-\nabla_{\boldsymbol{\theta}}\varsigma_{i,m}(\boldsymbol{\upsilon}^\circ)\|\\ &\quad\le s_{\mathbf{X}}^2\|\boldsymbol{f}''_{\boldsymbol{\eta}}(\mathbf{X}_i^\top\boldsymbol{\theta})\nabla\varPhi(\boldsymbol{\theta})\nabla\varPhi(\boldsymbol{\theta})^\top\\ &\qquad\qquad-\boldsymbol{f}''_{\boldsymbol{\eta}^\circ}(\mathbf{X}_i^\top\boldsymbol{\theta}^\circ)\nabla\varPhi(\boldsymbol{\theta}^{\circ\top}\mathbf{X}_i)\nabla\varPhi(\boldsymbol{\theta}^{\circ\top}\mathbf{X}_i)^\top\|\\ &\qquad+s_{\mathbf{X}}^2\|\boldsymbol{f}'_{\boldsymbol{\eta}}(\mathbf{X}_i^\top\boldsymbol{\theta})\mathbf{X}_i^\top\nabla^2\varPhi(\boldsymbol{\theta}^\top\mathbf{X}_i)-\boldsymbol{f}'_{\boldsymbol{\eta}^\circ}(\mathbf{X}_i^\top\boldsymbol{\theta}^\circ)\mathbf{X}_i^\top\nabla^2\varPhi(\boldsymbol{\theta}^{\circ\top}\mathbf{X}_i)\|.\end{aligned}$$

We again separately estimate

$$\begin{aligned}&\|\boldsymbol{f}''_{\boldsymbol{\eta}}(\mathbf{X}_i^\top\boldsymbol{\theta})\nabla\varPhi(\boldsymbol{\theta})\nabla\varPhi(\boldsymbol{\theta})^\top-\boldsymbol{f}''_{\boldsymbol{\eta}^\circ}(\mathbf{X}_i^\top\boldsymbol{\theta}^\circ)\nabla\varPhi(\boldsymbol{\theta}^{\circ\top}\mathbf{X}_i)\nabla\varPhi(\boldsymbol{\theta}^{\circ\top}\mathbf{X}_i)^\top\|\\ &\quad\le\|[\boldsymbol{f}''_{\boldsymbol{\eta}}(\mathbf{X}_i^\top\boldsymbol{\theta})-\boldsymbol{f}''_{\boldsymbol{\eta}^\circ}(\mathbf{X}_i^\top\boldsymbol{\theta}^\circ)]\nabla\varPhi(\boldsymbol{\theta})\nabla\varPhi(\boldsymbol{\theta})^\top\|\\ &\qquad+\|\boldsymbol{f}''_{\boldsymbol{\eta}^\circ}(\mathbf{X}_i^\top\boldsymbol{\theta}^\circ)[\nabla\varPhi(\boldsymbol{\theta})-\nabla\varPhi(\boldsymbol{\theta}^{\circ\top}\mathbf{X}_i)]\nabla\varPhi(\boldsymbol{\theta})^\top\|\\ &\qquad+\|\boldsymbol{f}''_{\boldsymbol{\eta}^\circ}(\mathbf{X}_i^\top\boldsymbol{\theta}^\circ)\nabla\varPhi(\boldsymbol{\theta}^{\circ\top}\mathbf{X}_i)[\nabla\varPhi(\boldsymbol{\theta})^\top-\nabla\varPhi(\boldsymbol{\theta}^{\circ\top}\mathbf{X}_i)^\top\|.\end{aligned}$$

We estimate using that $\|\nabla\varPhi(\boldsymbol{\theta})\nabla\varPhi(\boldsymbol{\theta})^\top\|\le 1$

$$\begin{aligned}&\|[\boldsymbol{f}''_{\boldsymbol{\eta}}(\mathbf{X}_i^\top\boldsymbol{\theta})-\boldsymbol{f}''_{\boldsymbol{\eta}^\circ}(\mathbf{X}_i^\top\boldsymbol{\theta}^\circ)]\nabla\varPhi(\boldsymbol{\theta})\nabla\varPhi(\boldsymbol{\theta})^\top\|\\ &\quad\le\|\boldsymbol{f}''_{\boldsymbol{\eta}}(\mathbf{X}_i^\top\boldsymbol{\theta})-\boldsymbol{f}''_{\boldsymbol{\eta}^\circ}(\mathbf{X}_i^\top\boldsymbol{\theta}^\circ)\|\\ &\quad\le\|\boldsymbol{f}''_{\boldsymbol{\eta}}(\mathbf{X}_i^\top\boldsymbol{\theta})-\boldsymbol{f}''_{\boldsymbol{\eta}^\circ}(\mathbf{X}_i^\top\boldsymbol{\theta})\|+\|\boldsymbol{f}''_{\boldsymbol{\eta}^\circ}(\mathbf{X}_i^\top\boldsymbol{\theta})-\boldsymbol{f}''_{\boldsymbol{\eta}^\circ}(\mathbf{X}_i^\top\boldsymbol{\theta}^\circ)\|.\end{aligned}$$

Remember that due to the structure of the basis

$$|N(j)| \overset{\text{def}}{=} \Big|\Big\{k \in \{2^j - 2j17 - 1, \dots, 2^{j+1} - 2(j+1)17 - 1 - 1\} : \\ |\boldsymbol{e}'_k(\mathbf{X}_i^\top \boldsymbol{\theta}') - \boldsymbol{e}'_k(\mathbf{X}_i^\top \boldsymbol{\theta})| \vee |\boldsymbol{e}''_k(\mathbf{X}_i^\top \boldsymbol{\theta}') - \boldsymbol{e}''_k(\mathbf{X}_i^\top \boldsymbol{\theta})| \\ \vee |\boldsymbol{e}'_k(\mathbf{X}_i^\top \boldsymbol{\theta})| > 0\Big\}\Big| \le 34.$$

We get with the same arguments as in the proof of Lemma 6.A.11

$$\|[\boldsymbol{f}''_{\boldsymbol{\eta}}(\mathbf{X}_i^\top \boldsymbol{\theta}) - \boldsymbol{f}''_{\boldsymbol{\eta}^\circ}(\mathbf{X}_i^\top \boldsymbol{\theta}^\circ)]\nabla\Phi(\boldsymbol{\theta})\nabla\Phi(\boldsymbol{\theta})^\top\| \\ \le \frac{\sqrt{34}}{\sqrt{n}c_{\mathcal{D}}}\left(\|\psi''\|m^{5/2} + \|\psi'''\|\left(C_{\|\boldsymbol{f}\|} + \frac{m^2(\mathrm{R}_0 + \mathtt{r}^*)}{\sqrt{n}c_{\mathcal{D}}}\right)m^{3/2}\right) \\ \|\mathcal{D}_m(\boldsymbol{v} - \boldsymbol{v}^\circ)\|.$$

For the other two summands we estimate

$$\|\boldsymbol{f}''_{\boldsymbol{\eta}^\circ}(\mathbf{X}_i^\top \boldsymbol{\theta}^\circ)[\nabla\Phi(\boldsymbol{\theta}) - \nabla\Phi(\boldsymbol{\theta}^{\circ\top}\mathbf{X}_i)]\nabla\Phi(\boldsymbol{\theta})^\top\| \\ \le \|\boldsymbol{f}''_{\boldsymbol{\eta}^\circ}(\mathbf{X}_i^\top \boldsymbol{\theta}^\circ)\|\|\left\{\nabla\Phi(\boldsymbol{\theta}) - \nabla\Phi(\boldsymbol{\theta}^{\circ\top}\mathbf{X}_i)\right\}\nabla\Phi(\boldsymbol{\theta}^{\circ\top}\mathbf{X}_i)\|.$$

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

$$\|\boldsymbol{f}''_{\boldsymbol{\eta}^\circ}(\mathbf{X}_i^\top \boldsymbol{\theta}^\circ)[\nabla\Phi(\boldsymbol{\theta}) - \nabla\Phi(\boldsymbol{\theta}^{\circ\top}\mathbf{X}_i)]\nabla\Phi(\boldsymbol{\theta})^\top\| \\ \le \|\boldsymbol{f}''_{\boldsymbol{\eta}^\circ}(\mathbf{X}_i^\top \boldsymbol{\theta}^\circ)\|\mathtt{C}_2\|\boldsymbol{\theta} - \boldsymbol{\theta}^\circ\| \\ \le \mathtt{C}_1\|\boldsymbol{\theta} - \boldsymbol{\theta}^\circ\|\|\psi''\|\sum_{j=0}^{j_m-1}\sum_{k\in N(j)}\eta_k^\circ 2^{5j/2} \\ \le 17\mathtt{C}_1\|\boldsymbol{\theta} - \boldsymbol{\theta}^\circ\|\|\psi''\|\left(C_{\|\boldsymbol{f}\|} + \frac{m^2(\mathrm{R}_0 + \mathtt{r}^*)}{\sqrt{n}c_{\mathcal{D}}}\right)m^{1/2}.$$

We continue with

$$\|\boldsymbol{f}'_{\boldsymbol{\eta}}(\mathbf{X}_i^\top \boldsymbol{\theta})\mathbf{X}_i^\top\nabla^2\Phi(\boldsymbol{\theta}^\top\mathbf{X}_i) - \boldsymbol{f}'_{\boldsymbol{\eta}^\circ}(\mathbf{X}_i^\top \boldsymbol{\theta}^\circ)\mathbf{X}_i^\top\nabla^2\Phi(\boldsymbol{\theta}^{\circ\top}\mathbf{X}_i)\| \\ \le \|\boldsymbol{f}'_{\boldsymbol{\eta}}(\mathbf{X}_i^\top \boldsymbol{\theta}) - \boldsymbol{f}'_{\boldsymbol{\eta}^\circ}(\mathbf{X}_i^\top \boldsymbol{\theta}^\circ)\|\|\mathbf{X}_i^\top\nabla^2\Phi(\boldsymbol{\theta}^{\circ\top}\mathbf{X}_i)\| \\ +\|\boldsymbol{f}'_{\boldsymbol{\eta}}(\mathbf{X}_i^\top \boldsymbol{\theta})\|\mathbf{X}_i^\top\nabla^2\Phi(\boldsymbol{\theta}^\top\mathbf{X}_i) - \mathbf{X}_i^\top\nabla^2\Phi(\boldsymbol{\theta}^{\circ\top}\mathbf{X}_i)\|.$$

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 6.A.11

$$\begin{aligned}&\|\boldsymbol{f}'_{\boldsymbol{\eta}}(\mathbf{X}_i^\top\boldsymbol{\theta})\mathbf{X}_i^\top\nabla^2\Phi(\boldsymbol{\theta}^\top\mathbf{X}_i)-\boldsymbol{f}'_{\boldsymbol{\eta}^\circ}(\mathbf{X}_i^\top\boldsymbol{\theta}^\circ)\mathbf{X}_i^\top\nabla^2\Phi(\boldsymbol{\theta}^{\circ\top}\mathbf{X}_i)\|\\&\quad\le\frac{\sqrt{34}}{\sqrt{n}c_{\mathcal{D}}}m^{1/2}\left(\mathtt{C}_2 s_{\mathbf{X}}\|\psi''\|+\|\psi'\|+s_{\mathbf{X}}^2\mathtt{C}_3\right)\left(C_{\|\boldsymbol{f}\|}+\frac{m^2(\mathrm{R}_0+\mathtt{r}^*)}{\sqrt{n}c_{\mathcal{D}}}\right)\\&\qquad\|\mathcal{D}_m(\boldsymbol{\upsilon}-\boldsymbol{\upsilon}^\circ)\|.\end{aligned}$$

Finally

$$\begin{aligned}&\|\nabla_{\boldsymbol{\eta}}\varsigma_{i,m}(\boldsymbol{\upsilon})-\nabla_{\boldsymbol{\eta}}\varsigma_{i,m}(\boldsymbol{\upsilon}^\circ)\|\\&\quad\le\|\left(\nabla\Phi(\boldsymbol{\theta})^\top-\nabla\Phi(\boldsymbol{\theta}^{\circ\top}\mathbf{X}_i)^\top\right)\mathbf{X}_i\|\|\boldsymbol{e}'(\boldsymbol{\theta}^\top\mathbf{X}_i)^\top\|\\&\qquad+\|\nabla\Phi(\boldsymbol{\theta}^{\circ\top}\mathbf{X}_i)^\top\mathbf{X}_i\|\|\boldsymbol{e}'(\boldsymbol{\theta}^\top\mathbf{X}_i)-\boldsymbol{e}'(\boldsymbol{\theta}^{\circ\top}\mathbf{X}_i)\|.\end{aligned}$$

We estimate separately

$$\begin{aligned}\|\left(\nabla\Phi(\boldsymbol{\theta})^\top-\nabla\Phi(\boldsymbol{\theta}^\circ)^\top\right)\mathbf{X}_i\|&\le\mathtt{C}_4 s_{\mathbf{X}}^2\frac{1}{\sqrt{n}c_{\mathcal{D}}}\|\mathcal{D}_m(\boldsymbol{\upsilon}-\boldsymbol{\upsilon}^\circ)\|,\\\|\boldsymbol{e}'(\boldsymbol{\theta}^\top\mathbf{X}_i)^\top\|&\le\|\psi'\|_\infty\left(\sum_{j=0}^{j_m}2^{3j}|N(j)|\right)^{1/2}\\&\le\|\psi'\|_\infty\sqrt{34}m^{3/2}.\end{aligned}$$

Furthermore

$$\begin{aligned}\|\nabla\Phi(\boldsymbol{\theta}^{\circ\top}\mathbf{X}_i)^\top\mathbf{X}_i\|&\le\mathtt{C}_5 s_{\mathbf{X}},\\\|\boldsymbol{e}'(\boldsymbol{\theta}^\top\mathbf{X}_i)-\boldsymbol{e}'(\boldsymbol{\theta}^{\circ\top}\mathbf{X}_i)\|&\le\|\psi''\|_\infty\sqrt{34}m^{5/2}\frac{1}{\sqrt{n}c_{\mathcal{D}}}\|\mathcal{D}_m(\boldsymbol{\upsilon}-\boldsymbol{\upsilon}^\circ)\|.\end{aligned}$$

Putting all estimates together gives (6.A.38). □

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 6.A.28. *We have* $(\mathcal{E}\mathcal{D}_2)$ *with*

$$\omega_2=\frac{1}{\sqrt{n}c_{\mathcal{D}}},\ \mathtt{g}_2=\sqrt{n}\widetilde{\mathtt{g}}c_{\mathcal{D}}m^{-1}\mathtt{C}(R_0,p^*)^{-1},\ \nu_2^2=\frac{\widetilde{\nu}^2m^2\mathtt{C}(R_0,p^*)^2}{2c_{\mathcal{D}}},$$

where with $\mathtt{C}>0$ *in* (6.A.38)

$$\mathtt{C}(R_0,m)\overset{\mathrm{def}}{=}\mathtt{C}\left(m^{3/2}+\left(C_{\|\boldsymbol{f}\|}+\frac{m^2(R_0+\mathtt{r}^*)}{\sqrt{n}c_{\mathcal{D}}}\right)m^{1/2}\right).$$

Proof. Lemma 6.A.27 gives for any $\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^\circ \in \Upsilon(\mathtt{r})$ with $\varsigma_{i,m}$ from (6.A.1)

$$\begin{aligned}
&\|\mathcal{D}_m^{-1}\nabla\varsigma_{i,m}(\boldsymbol{\upsilon}) - \mathcal{D}^{-1}\nabla\varsigma_{i,m}(\boldsymbol{\upsilon}^\circ)\| \\
&\quad\le \frac{\mathtt{C}}{c_{\mathcal{D}}\sqrt{n}} m \left(m^{3/2} + \left(C_{\|\boldsymbol{f}\|} + \frac{m^2(\mathrm{R}_0 + \mathtt{r}^*)}{\sqrt{n} c_{\mathcal{D}}} \right) m^{1/2} \right) \|\mathcal{D}_m(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^\circ)\| \\
&\quad\overset{\text{def}}{=} \frac{1}{\sqrt{n} c_{\mathcal{D}}} m \mathtt{C}(\mathrm{R}_0, p^*) \|\mathcal{D}_m(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^\circ)\|. \qquad (6.A.39)
\end{aligned}$$

We get with $\mu \le \mathsf{g}_2$ and assumption $(\mathbf{Cond}_\varepsilon)$ for any pair $\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2 \in \{\|\boldsymbol{\gamma}\| = 1\}$

$$\begin{aligned}
&\mathbb{E}_\varepsilon \exp\left\{ \frac{\mu}{\omega_2 \|\mathcal{D}_m(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^\circ)\|} \boldsymbol{\gamma}_1^\top \left(\mathcal{D}_m^{-1} \nabla^2 \left\{ \zeta(\boldsymbol{\upsilon}) - \zeta(\boldsymbol{\upsilon}^\circ) \right\} \right) \boldsymbol{\gamma}_2 \right\} \\
&\quad = \mathbb{E}_\varepsilon \exp\left\{ \frac{\mu}{\omega_2 \|\mathcal{D}_m(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^\circ)\|} \right. \\
&\qquad\qquad \left. \sum_{i=1}^n \varepsilon_i \boldsymbol{\gamma}_1^\top \left(\mathcal{D}_m^{-1} \nabla \left\{ \varsigma_{i,m}(\boldsymbol{\upsilon}) - \varsigma_{i,m}(\boldsymbol{\upsilon}^\circ) \right\} \right) \boldsymbol{\gamma}_2 \right\} \\
&\quad = \prod_{i=1}^n \mathbb{E}_\varepsilon \exp\left\{ \frac{\mu}{\omega_2 \|\mathcal{D}_m(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^\circ)\|} \varepsilon_i \right. \\
&\qquad\qquad \left. \boldsymbol{\gamma}_1^\top \left(\mathcal{D}_m^{-1} \nabla \left\{ \varsigma_{i,m}(\boldsymbol{\upsilon}) - \varsigma_{i,m}(\boldsymbol{\upsilon}^\circ) \right\} \right) \boldsymbol{\gamma}_2 \right\} \\
&\quad \le \prod_{i=1}^n \exp\left\{ \frac{\widetilde{\nu}^2 \mu^2}{2\omega_2^2 \|\mathcal{D}_m(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^\circ)\|^2} \right. \\
&\qquad\qquad \left. \left(\boldsymbol{\gamma}_1^\top \left(\mathcal{D}_m^{-1} \nabla \left\{ \varsigma_{i,m}(\boldsymbol{\upsilon}) - \varsigma_{i,m}(\boldsymbol{\upsilon}^\circ) \right\} \right) \boldsymbol{\gamma}_2 \right)^2 \right\}.
\end{aligned}$$

With (6.A.39) this implies

$$\begin{aligned}
&\sup_{\substack{\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2 \in \mathbb{R}^{p^*} \\ \|\boldsymbol{\gamma}_i\| = 1}} \log \mathbb{E}_\varepsilon \exp\left\{ \frac{\mu}{\omega_2 \|\mathcal{D}_m(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^\circ)\|} \right. \\
&\qquad \left. \boldsymbol{\gamma}_1^\top \left(\mathcal{D}_m^{-1} \nabla^2 \zeta(\boldsymbol{\upsilon}) - \mathcal{D}^{-1} \nabla^2 \zeta(\boldsymbol{\upsilon}^\circ) \right) \boldsymbol{\gamma}_2 \right\} \le \frac{\widetilde{\nu}^2 \mu^2}{2 c_{\mathcal{D}}} m^2 \mathtt{C}(\mathrm{R}_0, p^*)^2.
\end{aligned}$$

□

Bound for Hessian

To control the deviation of $\mathcal{D}^{-1}\nabla\zeta(\boldsymbol{\upsilon}^*)$ we apply the following Theorem of [56]:

Theorem 6.A.29 (Corollary 3.7 of [56])**.** *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* $\mu>0$

$$I\!\!E \mathrm{e}^{\mu\boldsymbol{M}_i} \preceq \mathrm{e}^{g(\mu)\boldsymbol{A}_i}, \;\; \textit{where} \;\; \boldsymbol{M}\preceq\boldsymbol{M'} \Leftrightarrow \boldsymbol{\gamma}^\top\boldsymbol{M}\boldsymbol{\gamma}\le\boldsymbol{\gamma}^\top\boldsymbol{M}\boldsymbol{\gamma},\,\forall\boldsymbol{\gamma}\in\mathbb{R}^{p^*}.$$

Then for all $t\in\mathbb{R}$

$$I\!\!P\left(\left\|\sum_{i=1}^n \boldsymbol{M}_i\right\|\ge t\right)\le p^*\inf_{\mu}\exp\left\{-t\mu+g(\mu)\tau\right\}, \;\; \textit{where} \;\; \tau\stackrel{\mathrm{def}}{=}\left\|\sum_{i=1}^n\boldsymbol{A}_i\right\|.$$

Lemma 6.A.30. *We have for* $\mu\le\widetilde{\mathtt{g}}$

$$I\!\!E\exp\left\{\mu\mathcal{D}^{-1}\nabla^2\zeta(\boldsymbol{\upsilon}^*)\right\}\preceq\exp\left\{g(\mu)\operatorname{diag}(1,\ldots,1)\right\},$$

where

$$g(\mu)=\begin{cases}\frac{\widetilde{\nu}^2\mathtt{C}^2(\|\boldsymbol{f}'_{\boldsymbol{\eta}^*}\|_\infty+\|\boldsymbol{f}''_{\boldsymbol{\eta}^*}\|_\infty)^2\mu^2}{2}, & \textit{if } \mu\le\sqrt{n}\widetilde{\mathtt{g}}\mathtt{C}^{-1}(\|\boldsymbol{f}'_{\boldsymbol{\eta}^*}\|_\infty+\|\boldsymbol{f}''_{\boldsymbol{\eta}^*}\|_\infty)^{-1}\\ \infty, & \textit{otherwise.}\end{cases}$$

Proof. Due to Lemma 6.A.27

$$\begin{aligned}&\mathcal{D}^{-1}\nabla\varsigma_{i,m}(\boldsymbol{\upsilon}^*)\\ &\quad\preceq\operatorname{diag}\left(\frac{1}{\sqrt{n}}\mathtt{C}(\|\boldsymbol{f}'_{\boldsymbol{\eta}^*}\|_\infty+\|\boldsymbol{f}''_{\boldsymbol{\eta}^*}\|_\infty),\ldots,\frac{1}{\sqrt{n}}\mathtt{C}(\|\boldsymbol{f}'_{\boldsymbol{\eta}^*}\|_\infty+\|\boldsymbol{f}''_{\boldsymbol{\eta}^*}\|_\infty)\right).\end{aligned}$$

Thus denoting $\mathtt{C}_1\stackrel{\mathrm{def}}{=}\mathtt{C}(\|\boldsymbol{f}'_{\boldsymbol{\eta}^*}\|_\infty+\|\boldsymbol{f}''_{\boldsymbol{\eta}^*}\|_\infty)$

$$\begin{aligned}\exp\left\{\mu\mathcal{D}^{-1}\nabla^2\zeta(\boldsymbol{\upsilon}^*)\right\}&=\exp\left\{\mu\sum_{i=1}^n\mathcal{D}^{-1}\nabla\varsigma_{i,m}(\boldsymbol{\upsilon}^*)\varepsilon_i\right\}\\ &\preceq\exp\left\{\mu\sum_{i=1}^n\varepsilon_i\operatorname{diag}\left(\frac{1}{\sqrt{n}}\mathtt{C}_1,\ldots,\frac{1}{\sqrt{n}}\mathtt{C}_1\right)\right\}.\end{aligned}$$

Consequently we obtain due to the independence of the $\varsigma_{i,m}(\boldsymbol{\upsilon}^*)$ and assumption $(\mathbf{Cond}_{\varepsilon})$ for $\mu \leq \sqrt{n}\widetilde{\mathtt{g}}\mathtt{C}_1^{-1}$

$$\begin{aligned}
&\mathbb{E}\exp\left\{\mu\mathcal{D}^{-1}\nabla^2\zeta(\boldsymbol{\upsilon}^*)\right\}\\
&\quad\leq \prod_{i=1}^{n}\operatorname{diag}\left(\mathbb{E}\exp\left\{\frac{\mu}{\sqrt{n}}\varepsilon_i\mathtt{C}_1\right\},\ldots,\mathbb{E}\exp\left\{\frac{\mu}{\sqrt{n}}\varepsilon_i\mathtt{C}_1\right\}\right)\\
&\quad\leq \operatorname{diag}\left(\exp\left\{\frac{\widetilde{\nu}^2\mu^2}{2}\mathtt{C}_1^2\right\},\ldots,\exp\left\{\frac{\widetilde{\nu}^2\mu^2}{2}\mathtt{C}_1^2\right\}\right)\\
&\quad= \exp\left\{\frac{\widetilde{\nu}^2\mathtt{C}_1^2\mu^2}{2}\operatorname{diag}(1,\ldots,1)\right\}.
\end{aligned}$$

□

Lemma 6.A.31. *We have with* $\mathtt{C}(\|\boldsymbol{f}'_{\boldsymbol{\eta}^*}\|_\infty + \|\boldsymbol{f}''_{\boldsymbol{\eta}^*}\|_\infty)$ *and if* $\mathtt{x} \leq \frac{1}{2}(\widetilde{\nu}^2 n\widetilde{\mathtt{g}}^2 - \log(p^*))$

$$\mathbb{P}\left(\left\|\mathcal{D}^{-1}\nabla^2\zeta(\boldsymbol{\upsilon}^*)\right\| \geq \widetilde{\nu}\mathtt{C}(\|\boldsymbol{f}'_{\boldsymbol{\eta}^*}\|_\infty + \|\boldsymbol{f}''_{\boldsymbol{\eta}^*}\|_\infty)\sqrt{2\mathtt{x}+\log(p^*)}\right) \leq \mathrm{e}^{-2\mathtt{x}}.$$

Proof. With Lemma 6.A.30 and Theorem 6.A.29 we obtain for

$$t \leq \sqrt{n}\widetilde{\mathtt{g}}\mathtt{C}^{-1}(\|\boldsymbol{f}'_{\boldsymbol{\eta}^*}\|_\infty + \|\boldsymbol{f}''_{\boldsymbol{\eta}^*}\|_\infty)^{-1},$$

that

$$\begin{aligned}
&\mathbb{P}\left(\left\|\mathcal{D}^{-1}\nabla^2\zeta(\boldsymbol{\upsilon}^*)\right\| \geq t\right)\\
&\quad\leq p^*\inf_{\mu}\exp\left\{-t\mu + \frac{\widetilde{\nu}^2\mathtt{C}^2(\|\boldsymbol{f}'_{\boldsymbol{\eta}^*}\|_\infty + \|\boldsymbol{f}''_{\boldsymbol{\eta}^*}\|_\infty)^2\mu^2}{2}\right\}\\
&\quad= \inf_{\mu}\exp\left\{-t\mu + \widetilde{\nu}^2\mathtt{C}(\|\boldsymbol{f}'_{\boldsymbol{\eta}^*}\|_\infty + \|\boldsymbol{f}''_{\boldsymbol{\eta}^*}\|_\infty)^2\frac{\mu^2}{2}\right\}\\
&\quad= \exp\left\{-\frac{t^2}{2\widetilde{\nu}^2\mathtt{C}(\|\boldsymbol{f}'_{\boldsymbol{\eta}^*}\|_\infty + \|\boldsymbol{f}''_{\boldsymbol{\eta}^*}\|_\infty)^2}\right\}.
\end{aligned}$$

Defining $t(\mathtt{x})$ via

$$\mathbb{P}\left(\left\|\mathcal{D}^{-1}\nabla^2\zeta(\boldsymbol{\upsilon}^*)\right\| \geq t(\mathtt{x})\right) = \mathrm{e}^{-\mathtt{x}},$$

we find

$$t(\mathtt{x}) \leq \widetilde{\nu}\mathtt{C}(\|\boldsymbol{f}'_{\boldsymbol{\eta}^*}\|_\infty + \|\boldsymbol{f}''_{\boldsymbol{\eta}^*}\|_\infty)\sqrt{2\mathtt{x}+\log(p^*)},\ \text{ if }\ \mathtt{x} \leq \frac{1}{2}\left(\widetilde{\nu}^2 n\widetilde{\mathtt{g}}^2 - \log(p^*)\right).$$

□

Proof of Lemma

Lemma 6.A.31 together with Lemma 6.A.28 gives that in this setting

$$\begin{aligned}\frac{\sqrt{1-\nu}}{2\sqrt{2}(1+\sqrt{\nu})}\varkappa(\mathtt{x},\mathrm{R}_0) \leq{} & \mathtt{C}(\|\boldsymbol{f}'_{\boldsymbol{\eta}^*}\|_\infty + \|\boldsymbol{f}''_{\boldsymbol{\eta}^*}\|_\infty)^2\sqrt{2\mathtt{x}+\log(p^*)}/\sqrt{n} \\ & + \frac{\mathtt{C}\mathfrak{z}_1(\mathtt{x},3p^*)}{n}\left(m^{5/2}+\frac{m^{7/2}(\mathrm{R}_0+\mathtt{r}^*)}{\sqrt{n}c_{\mathcal{D}}}\right)\mathrm{R}_0 \\ & + \delta(\mathrm{R}_0+\mathtt{r}_0),\end{aligned}$$

if $\mathtt{x}$ is chosen moderately. As above

$$\begin{aligned}&\mathfrak{z}_1(\mathtt{x},3p^*) = O(\sqrt{\mathtt{x}+p^*}) = O(\mathtt{r}_0), \quad \|\mathcal{D}^{-1}\| \leq 1/(\sqrt{n}c_{\mathcal{D}}) \\ &\delta(\mathtt{r})/\mathtt{r} = O(p^{*3/2}+\mathtt{C}_{bias}m^{5/2})/\sqrt{n}.\end{aligned}$$

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 6.3.11 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 $\mathtt{x}$

$$\varkappa(\mathtt{x},\mathrm{R}_0) \leq \mathtt{C}\left(\frac{p^{*2}}{\sqrt{n}}+p^{*3/2}\tau+\frac{\sqrt{\tau}p^{*3/2}}{n^{1/4}}\right)(1+o(1)),$$

which tends to zero if $p^{*4}/n \to 0$ and $\tau = o(p^{*-3/2})$.

In case $\mathtt{C}_{bias} > 0$ we have

$$\mathtt{r}_0 = \mathtt{C}\sqrt{p^*\log(n)}, \quad \mathrm{R}_0 = \mathtt{C}\sqrt{p^*\log(n)+\sqrt{p^*}n\tau^2+\sqrt{\mathtt{x}n}\tau}.$$

Consequently

$$\begin{aligned}\varkappa(\mathtt{x},\mathrm{R}_0) \leq{} & \mathtt{C}\Bigg(p^{*3}\log(n)/\sqrt{n}+p^{*11/4}\tau \\ & +n^{-1/4}m^{5/2}\sqrt{\tau})\Bigg)(1+o(1)),\end{aligned}$$

which tends to 0 if $m^3\log(n)/n \to 0$ and $\tau = o(p^{*-11/4})$ since then

$$n^{-1/4}m^{5/2}\sqrt{\tau} = o(m^{-3/8}).$$

6.A.13 Proof of Proposition 6.2.5

Define the set

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

where

$$\begin{aligned}
\boldsymbol{\mathcal{N}}_l(\mathtt{x}) \overset{\text{def}}{=} & \left\{ \sup_{\boldsymbol{\upsilon}\in\varUpsilon_m\setminus\varUpsilon_\circ(\mathtt{r}_0^\circ)} \mathcal{L}(\boldsymbol{\upsilon},\boldsymbol{\upsilon}^*_m) < 0 \right\} \\
& \cap \{ \widetilde{\boldsymbol{\upsilon}}_{m(l)}, \widetilde{\boldsymbol{\upsilon}}_{m\boldsymbol{\theta}^*{}_{(l)}(l)}, \widetilde{\boldsymbol{\upsilon}}_{m\boldsymbol{\eta}^*{}_{(l)}(l)} \in \{ \|\mathcal{D}_{(l)}(\boldsymbol{\upsilon}-\boldsymbol{\upsilon}^*_{m(l)})\| \le \mathtt{r}_0 \} \} \\
& \cap \bigcap_{\mathtt{r}\le\mathtt{r}_0} \left\{ \sup_{\boldsymbol{\upsilon}\in\varUpsilon_\circ(\mathtt{r})} \left\| \mathcal{D}_{(l)}{}^{-1} \left(\nabla\zeta_{\varepsilon(l)}(\boldsymbol{\upsilon}) - \nabla\zeta_{\varepsilon(l)}(\boldsymbol{\upsilon}^*_{m(l)}) \right) \right\| - 2\mathtt{r}^2 \right. \\
& \qquad\qquad\qquad \left. \le \mathtt{C}\omega\nu_1(\mathtt{x}+p^*) \right\} \\
& \cap \left\{ \|\mathcal{D}_{(l)}{}^{-1}\nabla\mathcal{L}_{\varepsilon(l)}(\boldsymbol{\upsilon}^*_{m(l)})\| \le \mathtt{C}\sqrt{\mathtt{x}+p^*} \right\} \\
& \cap \left\{ \sup_{\boldsymbol{\upsilon}\in\varUpsilon_{\circ(l)}(\mathtt{r}^\infty)} \|\nabla(I\!\!E - I\!\!E_\varepsilon)[\mathcal{L}_{\varepsilon(l)}(\boldsymbol{\upsilon}^*_{m(l)}) - \mathcal{L}_{\varepsilon(l)}(\boldsymbol{\upsilon})]\| \right. \\
& \qquad\qquad\qquad \left. \le \mathtt{C}(\mathtt{x}+p^*)^2\mathtt{r}^\infty/\sqrt{n} \right\} \\
& \cap \left\{ \text{Conditions of Section 4.2.1 hold for } (\mathcal{L}_{\varepsilon(l)}, \varUpsilon_m, \mathcal{D}_{(l)}) \right\},
\end{aligned}$$

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 6.A.10. For $M=1$ this is the set on which Proposition 6.2.1 applies.

Lemma 6.A.32. *We have on the set* $\boldsymbol{\mathcal{M}}_M(\mathtt{x})$ *if* $p^{*5}/n < l$

$$\tau_{i(l)} \le \mathtt{C}l\sqrt{m}\left(\frac{p^{*7/2}+\mathtt{x}}{n} + \frac{\sqrt{p^*+\mathtt{x}}}{\sqrt{n}} \right). \tag{6.A.40}$$

Proof. We obtain with Proposition 4.2.3 that if $(\delta(\mathtt{r})/\mathtt{r}+6\nu_1\omega)\,\mathtt{r}_0 < 1$ 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 (6.2.9) 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 Lemma 4.A.3 and using Lemma 6.A.11 that on the set $\boldsymbol{\mathcal{M}}_M$ (we suppress $\cdot_{(l)}$)

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

Denote

$$\boldsymbol{B}_{(l-1)} \overset{\text{def}}{=} \max_i |\tau_i(l-1)|. \tag{6.A.41}$$

Then we find

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

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 $(g_{(l)})_{l=1,\ldots,M}$ we can bound as in Lemma 6.A.7

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

where $\mathtt{r} = \|\mathcal{D}_{(l)}(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^{*}{}_{(l)})\|$. With Lemma 4.A.7 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 (6.3.5) is a uniform upper bound for all $l \le M$. We derived that on the set $\boldsymbol{\mathcal{M}}_{M}$ using that $\mathtt{r}^{*} \le \mathtt{C}\sqrt{p^{*} + \mathtt{x}}$

$$\begin{aligned}&\left\|\mathcal{D}_{(l)}(\widetilde{\boldsymbol{\upsilon}}_{m(l)} - \boldsymbol{\upsilon}^{*}{}_{(l)})\right\| \\ &\quad\le \mathtt{C}\left(\sqrt{p^{*} + \mathtt{x}} + \mathtt{C}\frac{p^{*7/2} + \mathtt{x}}{\sqrt{n}} + p^{*2}\boldsymbol{B}_{(l-1)}\right) \\ &\quad\overset{\text{def}}{=} \mathtt{C}\boldsymbol{T}_{(l-1)}. \end{aligned} \tag{6.A.42}$$

Finally we bound

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

We estimate separately using (6.A.42)

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

Furthermore we find with (6.A.42)

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

Consequently

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

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 6.A.33. *We have*

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

Proof. With Lemma 6.3.6 we find

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

Due to the assumptions we find with Lemma 6.3.7 that

$$\begin{aligned}&I\!\!P\left(\text{The conditions of Section 4.2.1 are met for } (\mathcal{L}_{\varepsilon(l)}, \Upsilon_m, \mathcal{D}_{(l)})\right)\\&\quad\ge 1 - 4\mathrm{e}^{-\mathtt{x}} - \exp\left\{-m^3\mathtt{x}\right\} - \exp\left\{-nc_{(\boldsymbol{Q})}/4\right\}.\end{aligned}$$

On that set we find as in the proof of Proposition 4.2.3 for $\mathtt{C} > 0$ large enough

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

and

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

Furthermore by Lemma 6.3.1 We have that

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

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

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

Using (6.A.41) we can bound

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

As the conditions $(\mathcal{A})$ are satisfied for all $l=1,\ldots,M$ we can establish as in Lemma 6.3.7 for $\mathtt{r}^2 \geq \mathtt{C}_{\mathtt{b}}p^*\log(n)$

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

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

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

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

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

Plugging in (6.A.40) 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 6.3.4. This gives the claim. □

Bibliography

[1] Andresen, A. (2014a). Finite sample analysis of profile m-estimation in the single index model. *arXiv:1406.4052*.

[2] Andresen, A. (2014b). A note on critical dimensions in profile semiparametric estimation. *arXiv:1410.4709*.

[3] Andresen, A. (2014c). A note on the bias of sieve profile estimation. *arXiv:1406.4045*.

[4] Andresen, A. and Spokoiny, V. (2014a). Critical dimension in profile semiparametric estimation. *Electron. J. Statist.*, 8(2):3077–3125.

[5] Andresen, A. and Spokoiny, V. (2014b). Two convergence results for an alternation maximization procedure. *arXiv:1501.01525v1*.

[6] Balakrishnan, S., Wainwright, M. J., and Yu, B. (2014). Satistical guarantees for the em algorithm: From population to sample-based analysis. *arXiv: 1408.2156*.

[7] Beran, J., Feng, Y., Ghosh, S., and Kulik, R. (2013). *Long-Memory Processes*. Springer.

[8] Berry, A. (1941). The accuracy of the gaussian approximation to the sum of independent variates. *Transactions of the American Mathematical Society*, 49(1):122–136.

[9] Bickel, P. J., Klaassen, C. A. J., Ritov, Y., and Wellner, J. A. (1998). *Efficient and adaptive estimation for semiparametric models*. Springer.

[10] Boucheron, S. and Massart, P. (2011). A high-dimensional Wilks phenomenon. *Probability Theory and Related Fields*, 150:405–433. 10.1007/s00440-010-0278-7.

[11] Boyle's, R. A. (1983). On the convergence of the em algorithm. *Journal of the Royal Statistical Society,*, Series B, 45:47–50.

[12] Chen, X. (2007). Large sample sieve estimation of semi-nonparametric models. *Handbook of Econometrics*, 6:5549–5632.

[13] Cohen, A., Daubechies, I., and Vial, P. (1993). Wavelets on the interval and fast wavelet transforms. *Applied and computational harmonic analysis*, 1:54–81.

[14] Daubechies, I. (1992). *Ten Lectures on Wavelets.* Society for Industrial and Applied Mathematics.

[15] Delecroix., M., Haerdle, W., and Hristache, M. (1997). Efficient estimation in single-index regression. Technical report, SFB 373, Humboldt Univ. Berlin.

[16] Dempster, A., Laird, N., and Rubin, D. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society,*, Series B, 39:1–38.

[17] Dudley, R. M. (1967). The sizes of compact subsets of hilbert space and continuity of gaussian processes. *Journal of Functional Analysis*, 1:290–330.

[18] Fan, J. and Huang, T. (2005). Profile likelihood inferences on semiparametric varying-coefficient partially linear models. *Bernoulli*, 11(6):1031–1057.

[19] Fan, J., Zhang, C., and Zhang, J. (2001). Generalized likelihood ratio statistics and Wilks phenomenon. *Ann. Stat.*, 29(1):153–193.

[20] Friedman, J. H. and Stuetzle, W. (1981). Projection pursuit regression. *Journal of the American Statistical Association*, 76(376):817–823.

[21] Geman, S. and Hwang, C. (1982). Nonparametric maximum likelihood estimation by the method of sieves. *Ann. Statist.*, 10:401–414.

[22] Grenander, U. (1981). *Abstract Inference.* Whiley, New York.

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

[24] Hall, P. (1988). Estimating the direction in which a data set is most interesting. *Probability Theory and Related Fields*, 80:51–77.

[25] Hall, P. (1992). *The Bootstrap and Edgeworth Expansion.* Springer.

[26] Hristache, M., Juditski, A., Polzehl, J., and Spokoiny, V. (2001). Structure adaptive approach for dimension reduction. *Annals of Statistics*, 29:595–623.

[27] Huber, P. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. Proc. 5th Berkeley Symp. Math. Stat. Probab., Univ. Calif. 1965/66, 1, 221-233 (1967).

[28] Huber, P. J. (1985). Projection pursuit. *The Annals of Statistics*, 13(2):435–475.

[29] Ibragimov, I. and Khas'minskij, R. (1981). *Statistical estimation. Asymptotic theory. Transl. from the Russian by Samuel Kotz.* New York - Heidelberg -Berlin: Springer-Verlag .

[30] Ichimura, H. (1993). Semiparametric least squares (sls) and weighted sls estimation of single-index models. *J Econometrics*, 58:71–120.

[31] Jain, P., Netrapalli, P., and Sanghavi, S. (2013). Low-rank matrix completion using alternating minimization. *STOC*, pages 665–674.

[32] Jones, L. K. (1987). On a conjecture of huber concerning the convergence of projection pursuit regression. *Ann. Statist*, 15(2):880–882.

[33] Keshavan, R. H., Montanari, A., and Oh, S. (2010). Matrix completion from few entries. *IEEE Transactions onInformation Theory*, 56(6):2980–2998.

[34] Kosorok, M. (2005). *Introduction to Empirical Processes and Semiparametric Inference.* Springer in Statistics.

[35] Mammen, E. (1989). Asymptotics with increasing dimension for robust regression with applications to the bootstrap. *Ann. Stat.*, 17(1):382–400.

[36] Mammen, E. (1993). Bootstrap and wild bootstrap for high dimensional linear models. *Ann. Stat.*, 21(1):255–285.

[37] Mammen, E. (1996). Empirical process of residuals for high-dimensional linear models. *Ann. Stat.*, 24(1):307–335.

[38] McLachlan, G. and Krishnan, T. (1997). *The EM Algorithm and Extensions.* Wiley, New York.

[39] Mendelson, S. (2014). Learning without concentration. *arXiv:1401.0304.*

[40] Murphy, S. A. and Van der Vaart, A. W. (2000). On profile likelihood. *Journal of the American Statistical Association*, 95(450):449–465.

[41] Netrapalli, P., Jain, P., and Sanghavi, S. (2013). Phase retrieval using alternating minimization. *NIPS*, pages 2796–2804.

[42] Newey, W. K. (1997). Convergence rates and asymptotic normality for series estimators. *Journal of Econometrics*, 79(1):147–168.

[43] Portnoy, S. (1984). Asymptotic behavior of M-estimators of p regression parameters when p^2/n is large. I. Consistency. *Ann. Stat.*, 12:1298–1309.

[44] Portnoy, S. (1985). Asymptotic behavior of M estimators of p regression parameters when p^2/n is large. II: Normal approximation. *Ann. Stat.*, 13:1403–1417.

[45] Portnoy, S. (1986). Asymptotic behavior of the empiric distribution of M-estimated residuals from a regression model with many parameters. *Ann. Stat.*, 14:1152–1170.

[46] Portnoy, S. (1988). Asymptotic behavior of likelihood methods for exponential families when the number of parameters tends to infinity. *Ann. Statist.*, 16:356–366.

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

[48] Rosenblatt, M. (1956). Remarks on some nonparametric estimats of a density function. *Annals of Mathematical Statistics*, 27(3):832–837.

[49] Shen, X. Shi, J. (2005). Sieve likelihood ratio inference on general parameter space. *Science in China*, 48(1):67–78.

[50] Shen, X. (1997). On methods of sieves and penalization. *Ann. Stat.*, 25(6):2555–2591.

[51] Souza, G. and Gallant, R. (1979). Statistical inference based on m-estimators for the multivariate nonlinear regression model in implicit form. *Institute of Statistics mimeograph series*, 1229.

[52] Spokoiny, V. (2012). Parametric estimation. Finite sample theory. *Ann. Statist.*, 40(6):2877–2909. arXiv:1111.3029.

[53] Spokoiny, V. (2013). Bernstein - von Mises Theorem for growing parameter dimension. Manuscript. arXiv:1302.3430.

[54] Stone, C. J. (1980). Optimal rates of convergence for nonparametric estimators. *Ann. Statist.*, 8(6):1348–1360.

[55] Talagrand, M. (1996). Majorizing measures: the generic chaining. *Ann. Statist.*, 24(3):1049–1103.

[56] Tropp, J. A. (2012). User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12:389–434.

[57] van der Vaart, A. (1998). *Asymptotic Statistics.* Cambridge University Press 1998.

[58] Wilks, S. (1938). The large-sample distribution of the likelihood ratio for testing composite hypotheses. *Ann. Math. Statist.*, 9,(1):60–62.

[59] Wu, C. (1983). On the convergence properties of the em algorithm. *Annals of Statistics*, 11:95–103.

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

[61] Xia, Y., Tong, H., Li, W., and Zhu, L. (2002). An adaptive estimation of dimension reduction space. *Journal of the Royal Statistical Society*, pages 363–410.

[62] Yi, X., Caramanis, C., and Sanghavi, S. (2013). Alternating minimization for mixed linear regression. *arXiv: 1310.3745*.