Electronic Journal of Statistics
Vol. 8 (2014) 3077–3125
ISSN: 1935-7524
DOI: 10.1214/14-EJS982

Critical dimension in profile semiparametric estimation

Andreas Andresen*

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

and

Vladimir Spokoiny†

Weierstrass Institute and HU Berlin
Moscow Institute of Physics and Technology
Mohrenstr. 39
10117 Berlin, Germany
e-mail: spokoiny@wias-berlin.de

Abstract: This paper revisits the classical inference results for profile quasi maximum likelihood estimators (profile MLE) in semiparametric models. We mainly focus on two prominent theorems: the Wilks phenomenon and Fisher expansion for the profile MLE are stated in a new fashion allowing finite samples and model misspecification. The method of study is also essentially different from the usual analysis of the semiparametric problem based on the notion of the hardest parametric submodel. Instead we derive finite sample deviation bounds for the linear approximation error for the gradient of the loglikelihood. This novel approach particularly allows to address the impact of the effective target and nuisance dimension on the accuracy of the results. The obtained nonasymptotic results are surprisingly sharp and yield the classical asymptotic statements including the asymptotic normality and efficiency of the profile MLE. The general results are specified for the important special case of an i.i.d. sample and the analysis is exemplified with a single index model.

MSC 2010 subject classifications: Primary 62F10; secondary 62J12, 62F25, 62H12.
Keywords and phrases: Profile maximum likelihood, local linear approximation, spread, local concentration.

Received June 2014.

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

†The author is partially supported by Laboratory for Structural Methods of Data Analysis in Predictive Modeling, MIPT, RF government grant, ag. 11.G34.31.0073. Financial support by the German Research Foundation (DFG) through the CRC 649 "Economic Risk" is gratefully acknowledged.

Contents

1. Introduction

Many statistical tasks can be viewed as problems of semiparametric estimation when the unknown data distribution is described by a high or infinite dimensional parameter while the target is of low dimension. Typical examples are provided by functional estimation, estimation of a function at a point, or simply by estimating a given subvector of the parameter vector. The classical statistical theory provides a general solution to this problem: estimate the full parameter vector by the maximum likelihood method and project the obtained estimate onto the target subspace. This approach is known as *profile maximum likelihood* and it appears to be *semiparametrically efficient* under some mild regularity conditions. We refer to the papers [22, 23] and the book [18] for a detailed presentation of the modern state of the theory and further references. The famous Wilks result claims that the likelihood ratio test statistic in the semiparametric test problem is nearly chi-square with p degrees of freedom corresponding to the dimension of the target parameter. Various extensions of this result can be found e.g. in [9, 8, 6]; see also the references therein.

This study revisits the problem of profile semiparametric estimation and addresses some new issues. The most important difference between our approach and the classical theory is a nonasymptotic character of our study. A finite sample analysis is particularly challenging because most notions, methods and tools in the classical theory are formulated in the asymptotic setup with growing sample size. Only few general finite sample results are available; see e.g. the recent paper [6]. The results of this paper explicitly describes all "small" terms in the expansion of the log-likelihood. This helps to carefully treat the question of the applicability of the approach in different situations. A particularly

important question concerns the critical dimension of the target p and the full parameter dimension p^* for which the main results are still accurate. Another issue addressed in this paper is the model misspecification. In many practical problems, it is unrealistic to expect that the model assumptions are exactly fulfilled, even if some rich nonparametric models are used. This means that the true data distribution $\mathbb{P}$ does not belong to the considered parametric family. Applicability of the general semiparametric theory in such cases is questionable. An important feature of the presented approach is that it equally applies under a possible model misspecification.

Let $\boldsymbol{Y}$ denote the observed random data, and $\mathbb{P}$ denote the data distribution. The parametric statistical model assumes that the unknown data distribution $\mathbb{P}$ belongs to a given parametric family $(\mathbb{P}_{\boldsymbol{\upsilon}})$:

$$\boldsymbol{Y} \sim \mathbb{P} = \mathbb{P}_{\boldsymbol{\upsilon}^*} \in (\mathbb{P}_{\boldsymbol{\upsilon}},\, \boldsymbol{\upsilon} \in \Upsilon),$$

where Υ is some high dimensional or even infinite dimensional parameter space.

The maximum likelihood approach in the parametric estimation suggests to estimate the whole parameter vector $\boldsymbol{\upsilon} \in \Upsilon$ by maximizing the corresponding log-likelihood $\mathcal{L}(\boldsymbol{\upsilon}) = \log \frac{d\mathbb{P}_{\boldsymbol{\upsilon}}}{d\boldsymbol{\mu}_0}(\boldsymbol{Y})$ for some dominating measure $\boldsymbol{\mu}_0$:

$$\widetilde{\boldsymbol{\upsilon}} \stackrel{\text{def}}{=} \operatorname*{argmax}_{\boldsymbol{\upsilon} \in \Upsilon} \mathcal{L}(\boldsymbol{\upsilon}). \tag{1.1}$$

Our study admits a model misspecification $\mathbb{P} \notin (\mathbb{P}_{\boldsymbol{\upsilon}}, \boldsymbol{\upsilon} \in \Upsilon)$. Equivalently, one can say that $\mathcal{L}(\boldsymbol{\upsilon})$ is the *quasi log-likelihood function* on Υ. The "target" value $\boldsymbol{\upsilon}^*$ of the parameter $\boldsymbol{\upsilon}$ can be defined by

$$\boldsymbol{\upsilon}^* = \operatorname*{argmax}_{\boldsymbol{\upsilon} \in \Upsilon} \mathbb{E}\mathcal{L}(\boldsymbol{\upsilon}). \tag{1.2}$$

Under model misspecification, $\boldsymbol{\upsilon}^*$ defines the best parametric fit of the considered family to $\mathbb{P}$.

In the semiparametric framework, the target of analysis is only a low dimensional component $\boldsymbol{\theta} \in \mathbb{R}^p$ of the whole parameter $\boldsymbol{\upsilon}$. This means that the target of estimation is

$$\boldsymbol{\theta}^* = \Pi_{\boldsymbol{\theta}} \boldsymbol{\upsilon}^*,$$

for some mapping $\Pi_{\boldsymbol{\theta}} : \Upsilon \to \mathbb{R}^p$, and $p \in \mathbb{N}$ stands for the dimension of the target. Often the vector $\boldsymbol{\upsilon}$ is represented as $\boldsymbol{\upsilon} = (\boldsymbol{\theta}, \boldsymbol{\eta})$, where $\boldsymbol{\theta}$ is the target of analysis while $\boldsymbol{\eta}$ is the *nuisance parameter*. We refer to this situation as $(\boldsymbol{\theta}, \boldsymbol{\eta})$-setup and our presentation follows this setting.

Define the *profile likelihood*

$$\breve{L}(\boldsymbol{\theta}) \stackrel{\text{def}}{=} \max_{\substack{\boldsymbol{\upsilon} \in \Upsilon \\ \Pi_{\boldsymbol{\theta}} \boldsymbol{\upsilon} = \boldsymbol{\theta}}} \mathcal{L}(\boldsymbol{\upsilon}).$$

The *profile maximum likelihood* approach defines the estimator of $\boldsymbol{\theta}^*$ by projecting the obtained MLE $\widetilde{\boldsymbol{\upsilon}}$ on the target space:

$$\widetilde{\boldsymbol{\theta}} \stackrel{\text{def}}{=} \Pi_{\boldsymbol{\theta}} \widetilde{\boldsymbol{\upsilon}} = \Pi_{\boldsymbol{\theta}} \operatorname*{argmax}_{\boldsymbol{\upsilon} \in \Upsilon} \mathcal{L}(\boldsymbol{\upsilon}) = \operatorname*{argmax}_{\boldsymbol{\theta} \in \Theta} \max_{\substack{\boldsymbol{\upsilon} \in \Upsilon \\ \Pi_{\boldsymbol{\theta}} \boldsymbol{\upsilon} = \boldsymbol{\theta}}} \mathcal{L}(\boldsymbol{\upsilon}) \stackrel{\text{def}}{=} \operatorname*{argmax}_{\boldsymbol{\theta} \in \Theta} \breve{L}(\boldsymbol{\theta}). \tag{1.3}$$

The Gauss-Markov Theorem claims the efficiency of such procedures for linear Gaussian models and a linear mapping $\Pi_{\boldsymbol{\theta}}$, and the famous Fisher result extends it in the asymptotic sense to the general situation under some regularity conditions. The Wilks phenomenon describes the limiting distribution of the likelihood ratio test statistic T which is also called the *semiparametric excess*: $T \stackrel{\text{def}}{=} 2\bigl\{\breve{L}(\widetilde{\boldsymbol{\theta}}) - \breve{L}(\boldsymbol{\theta}^*)\bigr\}$. It appears that the distribution of this test statistic is nearly chi-square distributed with $p \in \mathbb{N}$ degrees if freedom as the samples size grows, [33]:

$$T \stackrel{\text{def}}{=} 2\bigl\{\max_{\boldsymbol{\upsilon} \in \Upsilon} \mathcal{L}(\boldsymbol{\upsilon}) - \max_{\substack{\boldsymbol{\upsilon} \in \Upsilon \\ \Pi_{\boldsymbol{\theta}} \boldsymbol{\upsilon} = \boldsymbol{\theta}^*}} \mathcal{L}(\boldsymbol{\upsilon})\bigr\} = 2\bigl\{\breve{L}(\widetilde{\boldsymbol{\theta}}) - \breve{L}(\boldsymbol{\theta}^*)\bigr\} \xrightarrow{w} \chi^2_p.$$

In particular, the limit distribution does not depend on the particular model structure and on the full dimension of the parameter $\boldsymbol{\upsilon}$, only the dimension of the target matters. The full parameter dimension can be even infinite under some upper bounds on its total entropy.

The *local asymptotic normality* (LAN) approach by Le Cam leads to the most general setup in which the Wilks and Fisher type results can be established. However, the classical theory of semiparametric estimation faces serious difficulties when the dimension of the nuisance parameter becomes large or infinite. The LAN property yields a strong local approximation of the log-likelihood of the full model by the log-likelihood of a linear Gaussian model, and this property is only validated in a root-n neighborhood of the true point. The non- and semiparametric cases require to consider larger neighborhoods where the LAN approach is not applicable any more. A proper extension of the Wilks and Fisher result to the case of a growing or infinite nuisance dimension is quite challenging and involves special constructions like a pilot consistent estimator of the target, a hardest parametric submodel as well as some power tools of the empirical process theory; see [22] or [18] for a comprehensive presentation.

The recent paper [29] offers a new look at the classical LAN theory. The key steps are a local quadratic bracketing for the log-likelihood process and some concentration results for its stochastic component. The results can be stated for finite samples and do not involve any asymptotic consideration. It is also shown that many corollaries of the LAN property like Fisher and Wilks expansions only rely on these two facts. The bracketing idea of [29] is to build two different quadratic processes such that the original log-likelihood can be sandwiched between them up to a small error. This paper offers another approach based on the local linear approximation of the gradient of the log-likelihood process. This allows to improve the error term of the Fisher and Wilks expansion by a factor $\sqrt{p^*}$.

For the further presentation we briefly outline the basic steps of the analysis. Introduce for $\boldsymbol{\upsilon} \in \Upsilon$ and $\boldsymbol{\upsilon}^* \in \Upsilon$ as defined in (1.2), the log-likelihood ratio process

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

An important step of our approach is a deviation bound for the MLE $\widetilde{\boldsymbol{\upsilon}} \in \Upsilon$ from (1.1). Given some $\mathtt{x} > 0$, we define a radius $\mathtt{r}_0 = \mathtt{r}_0(\mathtt{x}) > 0$ that ensures that

$$\mathbb{P}(\widetilde{\boldsymbol{\upsilon}} \in \Upsilon_\circ(\mathtt{r}_0)) \geq 1 - \mathrm{e}^{-\mathtt{x}}, \tag{1.4}$$

where $\Upsilon_\circ(\mathtt{r})$ is a ball of radius $\mathtt{r} > 0$ in the intrinsic semi-metric corresponding to the process $\mathcal{L}(\boldsymbol{\upsilon})$. We give conditions that ensure that the value $\mathtt{r}_0^2(\mathtt{x})$ grows almost linearly with $\mathtt{x}$. See Section 2.3 for a precise formulation. The second key step is to bound for $\mathtt{r} > 0$ the approximation error

$$\sup_{\boldsymbol{\upsilon} \in \Upsilon_\circ(\mathtt{r})} \left\| \breve{\mathcal{Y}}(\boldsymbol{\upsilon}) \right\| \stackrel{\text{def}}{=} \sup_{\boldsymbol{\upsilon} \in \Upsilon_\circ(\mathtt{r})} \left\| \breve{D}^{-1} \left\{ \breve{\nabla} \mathcal{L}(\boldsymbol{\upsilon}) - \breve{\nabla} \mathcal{L}(\boldsymbol{\upsilon}^*) + \breve{D}^2 \left(\boldsymbol{\theta} - \boldsymbol{\theta}^* \right) \right\} \right\|, \tag{1.5}$$

where $\breve{D}^{-2} = \boldsymbol{\Pi}_{\boldsymbol{\theta}} \mathcal{D}^{-2} \boldsymbol{\Pi}_{\boldsymbol{\theta}}^\top \in \mathbb{R}^{p \times p}$ with the full information matrix $\mathcal{D}^2 = -\nabla^2 \mathbb{E} \mathcal{L}(\boldsymbol{\upsilon}^*)$ and where the projected gradient $\breve{\nabla} \mathcal{L}$ is defined below in the next section. Section B.2.1 provides the following bound on a set of probability of at least $1 - \mathrm{e}^{-\mathtt{x}}$:

$$\sup_{\boldsymbol{\upsilon} \in \Upsilon_\circ(\mathtt{r})} \| \breve{\mathcal{Y}}(\boldsymbol{\upsilon}) \| \leq \breve{\diamondsuit}(\mathtt{r}, \mathtt{x}),$$

where $\breve{\diamondsuit}(\mathtt{r}, \mathtt{x})$ is a small error. In combination with the deviation bound (1.4) and the identity $\nabla \mathcal{L}(\widetilde{\boldsymbol{\upsilon}}) = 0$, this allows to derive the following Fisher and Wilks type expansions: with probability greater $1 - 2\mathrm{e}^{-\mathtt{x}}$

$$\left\| \breve{D} \left(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^* \right) - \breve{\boldsymbol{\xi}} \right\| \quad \leq \quad \breve{\diamondsuit}(\mathtt{r}_0, \mathtt{x}), \tag{1.6}$$

$$\left| \breve{L}(\widetilde{\boldsymbol{\theta}}) - \breve{L}(\boldsymbol{\theta}^*) - \| \breve{\boldsymbol{\xi}} \|^2 / 2 \right| \leq \mathtt{C} \sqrt{p + \mathtt{x}}\, \breve{\diamondsuit}(\mathtt{r}_0, \mathtt{x}). \tag{1.7}$$

In the case of correctly specified i.i.d models $\breve{D}^2$ is the covariance matrix of the efficient influence function; see [18]. The random vector $\breve{\boldsymbol{\xi}}$ satisfies $\mathbb{E} \breve{\boldsymbol{\xi}} = 0$ and $\mathbb{E} \| \breve{\boldsymbol{\xi}} \|^2 \asymp p$ and $\mathtt{C} > 0$ is a constant independent of $\mathtt{x} > 0$ and full dimension p^*. The precise definitions of the random p-vector $\breve{\boldsymbol{\xi}}$ is also given in the next section. Moreover, general deviation bounds for quadratic forms from [29] apply to $\| \breve{\boldsymbol{\xi}} \|^2$ (see Section A for details). In the case of a correct model specification the tails of $\| \breve{\boldsymbol{\xi}} \|^2$ behave like those of a chi-square random variable with p degrees of freedom, and the result (1.6) can be viewed as an extension of the Wilks phenomenon. Under general identifiability conditions, the radius $\mathtt{r}_0$ can be fixed by $\mathtt{r}_0^2 = \mathtt{C}_1 (p^* + \mathtt{x})$ for a fixed constant $\mathtt{C}_1$ to ensure the concentration property (1.4).

With this choice of $\mathtt{r}$, in the important i.i.d. case, the error term $\breve{\Diamond}(\mathtt{r}_0, \mathtt{x})$ can be bounded by $\mathtt{C}(p^* + \mathtt{x})/\sqrt{n}$. The results (1.6) and (1.7) are nonasymptotic and hold true even under model misspecification.

It is important to grasp the implications of (1.6) and (1.7). The central contribution of this work is to bound the term in (1.5). It appears in this or a similar form also in the asymptotic approaches (see [22]) but is shown to be a zero sequence in the sample size under certain complexity and smoothness assumptions on the set of scores $\{\breve{\nabla}\mathcal{L}(\boldsymbol{\upsilon}), \boldsymbol{\upsilon} \in \Upsilon\}$. We manage to quantify for finite samples upper bounds for this term as functions of the radius $\mathtt{r}_0$ and the full dimension p^*. This allows for example to address the error when constructing confidence sets. For this assume that the quantiles of $\|\breve{\boldsymbol{\xi}}\|$ are available or that they can be given up to small error based on the Berry Esseen theorem (see [4]) or Edgeworth expansions (see [13]). Then (1.6) and (1.7) allow the construction of "approximate" confidence sets that address the finite sample error term (1.5), see Remark 2.13. The obtained sets are more conservative, i.e. larger than the asymptotic ones, but guarantee that the desired confidence level is attained. The possible error made when neglecting the error term (1.5) is illustrated in an example in Remark 2.20. Note however that on this level the contribution is rather theoretical: as in the case of the asymptotic results in [22], crucial objects as the matrix $\breve{D}^2$ 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.

The proposed approach does not assume that the profile is consistent but gives conditions that ensure the right concentration behavior. Simply assuming that the profile is consistent can be even misleading in our setup because this would separate local and global considerations. This paper attempts to figure out a list of conditions ensuring global concentration and local expansion at the same time. 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. In the smooth semiparametric problem with a fixed dimension of the target parameter, both Fisher and Wilks results apply up to an error $p^*/n^{1/2}$. In particular, we obtain that the error term in the Fisher expansion can be by a factor $\sqrt{p}$ smaller than the similar error term in the Wilks Theorem. 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. It is of interest to compare our statements with the existing literature on the growing parameter asymptotics. We particularly mention [19, 20, 21] and a series of papers by S. Portnoy, see e.g. [26, 25, 27]. The typical dimensional asymptotic is $p^* = o(n^{1/2})$, which corresponds to our results. For some particular special problems and examples the condition on the parameter dimension can be relaxed to $p = o(n^{3/2})$; see [25]. However, the results are mainly limited to linear or generalized linear regression with independent observations and heavily use the model structure. To the contrary, our results apply in a rather general situation and deliver some useful information even in the case when the model is misspecified.

We begin by developing the results for the case that the full parameter space Υ is a subset of the Euclidean space of dimension $p^* \in \mathbb{N}$. In Section 2.6 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. [28]. The present paper combines the sieve approximation idea and the finite sample Fisher and Wilks results under a possibly misspecified model.

The paper is organized as follows. Section 2.1 contains the conditions that we impose for the approach. Section 2.2 introduces the objects and tools of the analysis and collects the main results including an extension of the Wilks Theorem, concentration properties of the profile estimator and the construction of confidence sets for the "true" parameter $\boldsymbol{\theta}^*$. Section 2.3 explains how to control the large deviations of $\widetilde{\boldsymbol{\upsilon}}$ from (1.1) and how to improve the accuracy of the main results. Section 2.4 explains how the results translate to the case of i.i.d. samples and how the approach allows to obtain asymptotic efficiency of the profile estimator in this setting. Section 2.5 presents 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 2.1. Section 2.6 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 $\nabla \mathcal{L}(\boldsymbol{\upsilon}^*) \in \mathcal{X}$ to also treat the bias. Section 2.9 briefly outlines how the approach can be employed to derive the main results in the context of single index modeling and which ratio of full dimension to sample size is sufficient in that context. The details of this section can be found in [1]. The appendix collects the proofs of the main results.

2. Main results

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

Most of the results are stated in a finite sample setup for just one fixed sample. As we are also interested in understanding what happens if *the full dimension* p^* becomes large we also consider a specification of the general finite sample results to an asymptotic setup with $p^* = p_n$, where n denotes the asymptotic parameter, e.g. the sample size with $n \to \infty$. Our results apply also if the target parameter $\boldsymbol{\theta} \in \mathbb{R}^p$ is also of growing dimension. The dimension p can be of order p^*. Even the case with a full dimensional target and low dimensional nuisance is included.

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 $\mathcal{D}^2 \in \mathbb{R}^{p^* \times p^*}$ 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 (1.2). The matrix $\mathcal{D}^2 \in \mathbb{R}^{p^* \times p^*}$ can be defined as follows:

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

It is worth mentioning that $-\nabla^2 \mathbb{E}\mathcal{L}(\boldsymbol{\upsilon}^\circ) = \mathrm{Cov}(\nabla \mathcal{L}(\boldsymbol{\upsilon}^*))$ if the model $\boldsymbol{Y} \sim \mathbb{P}_{\boldsymbol{\upsilon}^*} \in (\mathbb{P}_{\boldsymbol{\upsilon}})$ is correctly specified and sufficiently regular; see e.g. [15].

Remark 2.1. This is not the only possible choice for $\mathcal{D}^2$ and $\boldsymbol{\upsilon}^\circ$. Another candidate is to use $\mathcal{D}^2 = -\nabla^2 \mathbb{E}\mathcal{L}(\boldsymbol{\upsilon}^*)$ while $\boldsymbol{\upsilon}^\circ = \boldsymbol{\upsilon}^*_m$, where $\boldsymbol{\upsilon}^*_m$ approximates $\boldsymbol{\upsilon}^*$ with growing $m \in \mathbb{N}$, as in [1]. In general there is no restriction for the choice of $\mathcal{D}^2$, as long as the following list of conditions can be satisfied. The same holds for the matrix $\mathcal{V}^2 \in \mathbb{R}^{p^* \times p^*}$ that we introduce below.

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

$$\mathcal{D}^2 = \begin{pmatrix} D^2 & A \\ A^\top & H^2 \end{pmatrix}.$$

First we state an *identifiability condition*.

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

$$\|H^{-1} A^\top D^{-1}\| \leq \rho.$$

Remark 2.2. The condition ($\mathcal{I}$) allows to define the important $p \times p$ efficient information matrix $\breve{D}^2$ which is defined as the inverse of the $\boldsymbol{\theta}$-block of the inverse of the full dimensional matrix $\mathcal{D}^2$. The exact formula is given by

$$\breve{D}^2 \stackrel{\mathrm{def}}{=} \left(\Pi_{\boldsymbol{\theta}} \mathcal{D}^{-2} \Pi_{\boldsymbol{\theta}}^\top\right)^{-1} = D^2 - A H^{-2} A^\top,$$

and ($\mathcal{I}$) ensures that the matrix $\breve{D}^2$ is well posed, see for instance [5], Chapter 2.4.

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

$$\Upsilon_\circ(\mathtt{r}) \stackrel{\mathrm{def}}{=} \left\{\boldsymbol{\upsilon} = (\boldsymbol{\theta}, \boldsymbol{\eta}) \in \Upsilon \colon \|\mathcal{D}(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^\circ)\| \leq \mathtt{r}\right\}. \tag{2.1}$$

Remark 2.3. For readers familiar with [29] we remark that the use of $\mathcal{D}$ instead of $\mathcal{V}$ in the above definition has no deeper reason but is a choice of convenience.

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

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

and remeber the definition of the radius $\mathtt{r}_0 > 0$

$$\mathtt{r}_0(\mathtt{x}) \stackrel{\mathrm{def}}{=} \inf_{\mathtt{r} > 0} \left\{\mathbb{P}(\widetilde{\boldsymbol{\upsilon}}, \widetilde{\boldsymbol{\upsilon}}_{\boldsymbol{\theta}^*} \in \Upsilon_\circ(\mathtt{r})) \geq 1 - \mathrm{e}^{-\mathtt{x}}\right\}, \tag{2.2}$$

which we set to infinity if $\widetilde{\boldsymbol{\upsilon}} = \{\,\}$ or $\widetilde{\boldsymbol{\upsilon}}_{\boldsymbol{\theta}^*} = \{\,\}$. Under the conditions $(\mathcal{L}\mathtt{r})$ and $(\mathcal{E}\mathtt{r})$ Theorem 2.3 in Section 2.3 states that $\mathtt{r}_0 = \mathtt{r}_0(\mathtt{x}) \approx \mathtt{C}\sqrt{\mathtt{x} + p^*} > 0$.

Here and in what follows we implicitly assume that the log-likelihood function $\mathcal{L}(\boldsymbol{\upsilon})\colon \mathbb{R}^{p^*} \to \mathbb{R}$ is sufficiently smooth in $\boldsymbol{\upsilon} \in \mathbb{R}^{p^*}$, $\nabla\mathcal{L}(\boldsymbol{\upsilon}) \in \mathbb{R}^{p^*}$ stands for the gradient and $\nabla^2\mathbb{E}\mathcal{L}(\boldsymbol{\upsilon}) \in \mathbb{R}^{p^*\times p^*}$ for the Hessian of the expectation $\mathbb{E}\mathcal{L} : \mathbb{R}^{p^*} \to \mathbb{R}$ at $\boldsymbol{\upsilon} \in \mathbb{R}^{p^*}$. By smooth enough we mean that all appearing derivatives exist and that we can interchange $\nabla\mathbb{E}\mathcal{L}(\boldsymbol{\upsilon}) = \mathbb{E}\nabla\mathcal{L}(\boldsymbol{\upsilon})$ on $\Upsilon_\circ(\mathtt{r}_0)$, where $\mathtt{r}_0 > 0$ is defined in Equation (2.2) and $\Upsilon_\circ(\mathtt{r})$ in equation (2.1). The following two conditions further quantify the smoothness properties on $\Upsilon_\circ(\mathtt{r})$ of the expected log-likelihood $\mathbb{E}\mathcal{L}(\boldsymbol{\upsilon})$ and of the stochastic component $\zeta(\boldsymbol{\upsilon}) = \mathcal{L}(\boldsymbol{\upsilon}) - \mathbb{E}\mathcal{L}(\boldsymbol{\upsilon})$.

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

$$\begin{aligned}\|D^{-1}D^2(\boldsymbol{\upsilon})D^{-1} - I_p\| &\le \breve{\delta}(\mathtt{r}),\\ \|D^{-1}(A(\boldsymbol{\upsilon}) - A)H^{-1}\| &\le \breve{\delta}(\mathtt{r}),\\ \bigl\|D^{-1}AH^{-1}\bigl(I_m - H^{-1}H^2(\boldsymbol{\upsilon})H^{-1}\bigr)\bigr\| &\le \breve{\delta}(\mathtt{r}),\end{aligned}$$

where

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

Remark 2.4. This condition describes the local smoothness properties of the function $\mathbb{E}\mathcal{L}(\boldsymbol{\upsilon})$. In particular, it allows to bound the error of local linear approximation of the gradient $\breve{\nabla}_{\boldsymbol{\theta}}\mathbb{E}\mathcal{L}(\boldsymbol{\upsilon})$ where

$$\breve{\nabla}_{\boldsymbol{\theta}} = \nabla_{\boldsymbol{\theta}} - AH^{-2}\nabla_{\boldsymbol{\eta}}.$$

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 B.1)

$$\|\breve{D}^{-1}\Bigl(\breve{\nabla}\mathbb{E}\mathcal{L}(\boldsymbol{\upsilon}) - \breve{\nabla}\mathbb{E}\mathcal{L}(\boldsymbol{\upsilon}^*)\Bigr) + \breve{D}(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\| \le \mathtt{C}\breve{\delta}(\mathtt{r})\mathtt{r}. \tag{2.3}$$

In the proofs we actually only need the inequality (2.3) which in some cases can be weaker than $(\breve{\mathcal{L}}_0)$. For readers familiar with the classical theory (for instance [22]) we remark that $(\breve{\mathcal{L}}_0)$ is related to the condition that $t \mapsto l(t, \boldsymbol{\eta}_t(\boldsymbol{\theta}, \boldsymbol{\eta}))$ is twice continuously differentiable where $\breve{\delta}(\mathtt{r})$ 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 $\Upsilon_\circ(\mathtt{r})$.

The next condition concerns the regularity of the stochastic component $\zeta(\boldsymbol{\upsilon}) \stackrel{\mathrm{def}}{=} \mathcal{L}(\boldsymbol{\upsilon}) - \mathbb{E}\mathcal{L}(\boldsymbol{\upsilon})$.

$(\breve{\mathcal{E}}\mathcal{D}_1)$ $\zeta(\upsilon) \to \zeta(\upsilon')$ as $\upsilon \to \upsilon'$. Further for all $0 < \mathtt{r} < 4\mathtt{r}_0$, there exists a constant $\omega \le 1/2$ such that for all $|\mu| \le \breve{\mathtt{g}}$ and $\upsilon, \upsilon' \in \Upsilon_\circ(\mathtt{r})$

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

Remark 2.5. The above condition is strongly related to the assumption of Donsker- and and Glivenko-Cantelli properties in [22] in order to ensure that the error in the local linear approximation of $\breve{\nabla}\mathcal{L}(\upsilon) - \breve{\nabla}\mathcal{L}(\upsilon^*)$ disappears. We replace these conditions with the more specific assumption $(\breve{\mathcal{E}}\mathcal{D}_1)$, which in combination with the entropy of $\Upsilon_\circ(\mathtt{r})$ yields the desired error bounds. 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 $e_k : \mathbb{R} \to \mathbb{R}$ and a sub exponential moment bound on the additive noise $\varepsilon \in \mathbb{R}$, see condition $(\mathbf{Cond}_\varepsilon)$ in Section 2.9.

The above conditions suffice for our main results. But we include another condition that allows to control the deviation behavior of $\|\breve{D}^{-1} \breve{\nabla} \zeta(\upsilon^*)\|$.

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

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

Remark 2.6. One possible and natural choice for the matrices $\breve{V}^2 \in \mathbb{R}^{p \times p}$ and $\mathcal{V}^2 \in \mathbb{R}^{p^* \times p^*}$ (see $(\mathcal{E}\mathcal{D}_0)$ below) is

$$\mathcal{V}^2 \stackrel{\mathrm{def}}{=} \mathrm{Var}\{\nabla \mathcal{L}(\upsilon^\circ)\}, \quad \breve{V}^2 = \mathrm{Cov}(\breve{\nabla}_{\theta} \zeta(\upsilon^\circ)),$$

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.

In many situations the following, stronger conditions, are easier to check and allow a further improvement of the results of Theorem 2.2 with the help of Proposition 2.4:

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

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

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

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

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

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

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

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

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

Remark 2.7. 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 $(\nabla_{\boldsymbol{\theta}} + AH^{-1}\nabla_{\boldsymbol{\eta}})\mathcal{L}(\cdot)$. This can make a tremendous difference to $(\mathcal{L}_0)$, $(\mathcal{ED}_0)$ and $(\mathcal{ED}_1)$ if $A(\cdot) \in \mathbb{R}^{p \times m}$ is small while $\nabla_{\boldsymbol{\eta}}\mathcal{L}(\cdot)$ is rather rough or possesses bad moment properties. In that case $(\mathcal{ED}_0)$ and $(\mathcal{ED}_1)$ 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 $A(\cdot) \equiv 0$.

Finally we present two conditions that allow a specific approach to determine a radius $\mathtt{r}_0(\mathtt{x}) > 0$ such that $\mathbb{P}(\widetilde{\boldsymbol{\upsilon}} \in \Upsilon(\mathtt{r}_0)) \ge 1 - \mathrm{e}^{\mathtt{x}}$ (see Section 2.3). These conditions have to be satisfied on the whole set $\Upsilon \subseteq \mathbb{R}^{p^*}$. 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 $\mathbb{P}(\widetilde{\boldsymbol{\upsilon}} \in \Upsilon(\mathtt{r}_0)) \ge 1 - \mathrm{e}^{\mathtt{x}}$.

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

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

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

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

Remark 2.8. These two conditions serve a qualified apriori concentration result for the full estimator $\widetilde{\boldsymbol{\upsilon}}$, of the type $\mathbb{P}\{\widetilde{\boldsymbol{\upsilon}} \in \Upsilon_\circ(\mathtt{r}_0(\mathtt{x}))\} \ge 1 - \mathrm{e}^{-\mathtt{x}}$. 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 Section 2.9. In a more general setting it could be combined with yet another even rougher a priori consistency result $\mathbb{P}(\widetilde{\boldsymbol{\upsilon}} \in U(\boldsymbol{\upsilon}^*))$ for some open neighborhood $U(\boldsymbol{\upsilon}^*) \subset \Upsilon$. Then $(\mathcal{L}\mathtt{r})$ is automatically satisfied as smooth functions are quadratic around their maximum, in this case $\mathbb{E}\mathcal{L}$ around $\boldsymbol{\upsilon}^*$. Further the condition can be relaxed to $-\mathbb{E}\mathcal{L}(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^\circ)$ growing with

super linear speed in the distance $\|\mathcal{D}(\boldsymbol{\upsilon}-\boldsymbol{\upsilon}^*)\|$, see Theorem 2.1 in [30]. 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 sub Gaussian.

Remark 2.9. We briefly comment how restrictive the imposed conditions are. Our conditions on the regularity and smoothness of the log-likelihood process $\mathcal{L}(\boldsymbol{\upsilon})$ in terms of the second or even third derivative are stronger than usually required; cf. Chapters 1, 2 in [15]. But 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. A quantification seems unavoidable of "how much smoother than twice differentiable" the function $\mathbb{E}\mathcal{L}(\cdot)$ is (i.e. condition $(\mathcal{L}_0)$), and of "how much smoother than once differentiable and well bounded in exponential moment terms" is $\nabla\zeta(\cdot)$ (i.e. condition $(\mathcal{E}\mathcal{D}_1)$). Note further, that we do not require that $\mathcal{L}(\boldsymbol{\upsilon})$ is the true log-likelihood. It comes from a parametric family chosen by a statistician. For typical examples, such a family possesses the required regularity. In particular, [29], Section 5.1, considered in details the i.i.d. case and presented some mild sufficient conditions on the parametric family which imply the above general conditions.

Concerning moments the conditions $(\breve{\mathcal{E}}\mathcal{D}_1)$, $(\breve{\mathcal{E}}\mathcal{D}_0)$, $(\mathcal{E}\mathcal{D}_1)$, $(\mathcal{E}\mathcal{D}_0)$ and $(\mathcal{E}\mathtt{r})$ require sub exponential moments of the observations (errors). Usually one only assumes finite second or third moments of the errors; cf. [15], Chapter 2. Our condition is a bit more restrictive but it allows to obtain finite sample bounds of the kind that with some small $\epsilon > 0$

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

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

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

with $\upsilon^* \neq 0$, $\sqrt{x} \stackrel{\mathrm{def}}{=} \operatorname{sign}(x)\sqrt{|x|}$, $\mathbb{E}\varepsilon = 0$ 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

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

Now if $\log\mathbb{E}[\exp(\lambda\varepsilon)] < \lambda^2/2$ we can derive

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

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.

To list some settings in which the conditions can be satisfied we name the regression and generalized regression models; cf. [10, 11] or [17]. [29], Section 5.2, argued that $(\mathcal{E}D_1)$ is automatically fulfilled for a generalized linear model, while $(\mathcal{E}D_0)$ requires that regression errors have to fulfill some exponential moments conditions. If this condition is too restrictive and a more stable (robust) estimation procedure is desirable, one can apply the LAD-type contrast leading to median regression. [29], Section 5.3, showed for the case of linear median regression that all the required conditions are fulfilled automatically if the sample size n exceeds $\mathtt{C}p^*$ for a fixed constant $\mathtt{C}$. [31] applied this approach for local polynomial quantile regression. [34] applied the approach to the problem of regression with Gaussian process where the unknown parameters enter in the likelihood in a rather complicated way. Further in this work we show how to satisfy them in a general i.i.d. setting and in the single index model, see Sections 2.4 and 2.9.

Remark 2.10. Another indication that the conditions are not too strong is served by an example in [2], where the error term $\breve{\Diamond}$ in our main result 2.2 is increased by a factor $\sqrt{p^*}$ if the condition $(\mathcal{L}_0)$ is slightly relaxed to read

$(\mathcal{L}_0)'$ There exists a symmetric $p^* \times p^*$-matrix $\mathcal{D}^2$ such that such that it holds on the set $\Upsilon_\circ(\mathtt{r}_0)$ for all $\mathtt{r} \le \mathtt{r}_0$

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

which appears in [29] under the label $(\mathcal{L}_0)$.

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* $\mathcal{D}^2 \in \mathbb{R}^{p^* \times p^*}$ be from the condition in Section 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 $\mathcal{D}^2$

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

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

$$\breve{D}^2 = D^2 - AH^{-2}A^\top,$$

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

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

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 the case of correct specification.

Remark 2.11. It seems worthy to point our that $\breve{D}^{-2}\breve{\nabla}_{\boldsymbol{\theta}} = \Pi_{\boldsymbol{\theta}}\mathcal{D}^{-2}\nabla$, see again [5], Chapter 2.4.

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

$$\breve{\diamondsuit}(\mathtt{r},\mathtt{x}) \stackrel{\text{def}}{=} 4\left(\frac{4}{(1-\rho^2)^2}\breve{\delta}(4\mathtt{r}) + 6\nu_1\breve{\omega}\mathfrak{z}(\mathtt{x},2p^*+2p)\right)\mathtt{r}, \tag{2.4}$$

where $\breve{\delta}(\mathtt{r})$ is shown in the condition $(\breve{\mathcal{L}}_0)$ and the constants $\breve{\omega}$, ν_1 are from condition $(\breve{\mathcal{E}}\mathcal{D}_1)$ in Section 2.1. The value $\mathfrak{z}(\mathtt{x},2p^*+2p)$ is related to the entropy of the unit ball in a $\mathbb{R}^{p^*+p}$-dimensional Euclidean space with $\mathtt{g} > 0$ from $(\breve{\mathcal{E}}\mathcal{D}_1)$ it is defined as

$$\mathfrak{z}(\mathtt{x},\mathbb{Q}) \stackrel{\text{def}}{=} \begin{cases} \sqrt{2(\mathtt{x}+\mathbb{Q})} & \text{if } \sqrt{2(\mathtt{x}+\mathbb{Q})} \le \mathtt{g}, \\ \mathtt{g}^{-1}(\mathtt{x}+\mathbb{Q}) + \mathtt{g}/2 & \text{otherwise,} \end{cases} \tag{2.5}$$

and one can apply $\mathfrak{z}(\mathtt{x},p^*) \cong \sqrt{\mathtt{x}+p^*}$ for moderate choice of $\mathtt{x} > 0$; see Appendix C. 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 $\Upsilon_\circ(\mathtt{r}) = \{\boldsymbol{\upsilon}\in\Upsilon\colon \|\mathcal{D}(\boldsymbol{\upsilon}-\boldsymbol{\upsilon}^\circ)\| \le \mathtt{r}\}$. 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 2.4 below.

Theorem 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* $\mathcal{D}^2$. *Further assume that the sets of maximizers* $\widetilde{\boldsymbol{\upsilon}}$, $\widetilde{\boldsymbol{\upsilon}}_{\boldsymbol{\theta}^*}$ *are not empty. Then it holds on a set* $\Omega(\mathtt{x}) \subseteq \Omega$ *of probability greater* $1-2\mathrm{e}^{-\mathtt{x}}$ *for the profile MLE* $\widetilde{\boldsymbol{\theta}}$ *from* (1.3)

$$\bigl\|\breve{D}\bigl(\widetilde{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\bigr) - \breve{\boldsymbol{\xi}}\bigr\| \le \breve{\diamondsuit}(\mathtt{r}_0,\mathtt{x}), \tag{2.6}$$

$$\bigl|2\breve{L}(\widetilde{\boldsymbol{\theta}},\boldsymbol{\theta}^*) - \|\breve{\boldsymbol{\xi}}\|^2\bigr| \le 4\Bigl(\|\breve{\boldsymbol{\xi}}\| + \breve{\diamondsuit}(\mathtt{r}_0,\mathtt{x})\Bigr)\breve{\diamondsuit}(\mathtt{r}_0,\mathtt{x}) + \breve{\diamondsuit}(\mathtt{r}_0,\mathtt{x})^2, \tag{2.7}$$

where the spread $\breve{\diamondsuit}(\mathtt{r}_0,\mathtt{x})$ *is defined in* (2.4) *and where* $\mathtt{r}_0 > 0$ *is defined in* (2.2).

Remark 2.12. The Wilks expansion claims that the profile maximum likelihood $\breve{L}(\widetilde{\boldsymbol{\theta}},\boldsymbol{\theta}^*) \stackrel{\text{def}}{=} \breve{L}(\widetilde{\boldsymbol{\theta}}) - \breve{L}(\boldsymbol{\theta}^*)$ 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 $\|\breve{\boldsymbol{\xi}}\|^2 = \|\breve{D}^{-1}\breve{\nabla}_{\boldsymbol{\theta}}\|^2$ weakly converges 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 $\breve{B} \stackrel{\text{def}}{=} \breve{D}^{-1}\breve{V}^2\breve{D}^{-1}$ 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 $\mathfrak{z}(\mathtt{x}, \breve{B})$ of this quadratic form ensuring

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

see Proposition A.1. One can use the bound $\mathfrak{z}^2(\mathtt{x}, \breve{B}) \leq \mathtt{C}(p + \mathtt{x})$ in most situations. We call $\breve{B} \in \mathbb{R}^{p \times p}$ *semiparametric misspecification matrix* as it is related to the misspecification matrix introduced in [14]. $\breve{B}$ is equal to the identity matrix if a correctly specified log likelihood is used.

Remark 2.13. One can use the expansion (2.6) for the construction of elliptic confidence sets

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

for some $\mathfrak{z}(\mathtt{x}) > 0$. More precisely let $q_\alpha > 0$ be the $\alpha-$level quantile of $\|\breve{\boldsymbol{\xi}}\|$. Then we find with the triangular inequality and (2.6)

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

and

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

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

Remark 2.14. The profile maximum likelihood process $\breve{L}(\boldsymbol{\theta})$ can be used for defining the likelihood-based confidence sets of the form

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

The bound (2.7) helps to evaluate the coverage probability $\mathbb{P}\bigl(\boldsymbol{\theta}^* \notin \mathcal{E}(\mathfrak{z})\bigr)$ in terms of deviation probability for the quadratic form $\|\breve{\boldsymbol{\xi}}\|^2$ and in term; cf. Corollary 3.2 in [29].

Remark 2.15. In the classical finite dimensional case, a usual choice for the central point $\boldsymbol{\upsilon}^\circ$ is $\boldsymbol{\upsilon}^\circ = \boldsymbol{\upsilon}^* = \operatorname{argmax}_{\boldsymbol{\upsilon} \in \Upsilon} \mathbb{E}\mathcal{L}(\boldsymbol{\upsilon})$ and one can define the matrix $\mathcal{D}^2$ as $\mathcal{D}^2 = -\nabla^2 \mathbb{E}\mathcal{L}(\boldsymbol{\upsilon}^*)$. However, for the sieve semiparametric problem in Section 2.6, we use another definition related to the infinite dimensional model.

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 MLE $\widetilde{\boldsymbol{\upsilon}} \in \mathbb{R}^{p^*}$ belongs to the local vicinity $\Upsilon_\circ(\mathtt{r}_0) \subset \mathbb{R}^{p^*}$ with high probability. As a first step we adopt the upper function approach from [29]; cf. Theorem 4.2 therein. It is important to note that Corollary 4.4 is one particular approach which could be replaced by any other proper technique. For instance, in the model with i.i.d. observations, Theorem 5.3 of [15] 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 follow the general way of [29] because it allows to address possible model misspecification and finite samples.

A close look at the proof of Theorem 4.2 of [29] shows that it actually yields the following modified version:

Theorem 2.3 ([29], Theorem 4.2)**.** *Suppose* $(\mathcal{E}\mathtt{r})$ *and* $(\mathcal{L}\mathtt{r})$ *with* $\mathtt{b}(\mathtt{r}) \equiv \mathtt{b}$. *Further define the following random set*

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

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

$$\begin{aligned} 1 + \sqrt{\mathtt{x} + 2p^*} &\leq 3\nu_{\mathtt{r}}^2 \mathtt{g}(\mathtt{r})/\mathtt{b}, \\ 6\nu_{\mathtt{r}} \sqrt{\mathtt{x} + 2p^* + \frac{\mathtt{b}}{9\nu_{\mathtt{r}}^2} K} &\leq \mathtt{r}\mathtt{b}, \end{aligned} \tag{2.8}$$

then

$$\mathbb{P}(\Upsilon(K) \subseteq \Upsilon_\circ(\mathtt{r}_0)) \geq 1 - \mathrm{e}^{-\mathtt{x}}. \tag{2.9}$$

Remark 2.16. Note that this Theorem also ensures that the maximum of $\mathcal{L} : \mathbb{R}^{p^*} \to \mathbb{R}$ is actually attained. Clearly $\boldsymbol{\upsilon}^* \in \Upsilon(0)$ such that it is nonempty. Further

$$\mathbb{P}\left(\Upsilon(0) \subseteq \Upsilon_\circ(\mathtt{r}_0)\right) \geq 1 - \mathrm{e}^{-\mathtt{x}},$$

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

Remark 2.17. The condition (2.8) helps to understand which $\mathtt{r}_0 > 0$ ensures 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 $\Upsilon(0)$. Consequently, if $\mathtt{g}(\mathtt{r}) > 0$ is large enough, (2.8) follows from the bound

$$\mathtt{r}_0 \geq 6\mathtt{b}^{-1}\nu_{\mathtt{r}}\sqrt{\mathtt{x} + p^*}. \tag{2.10}$$

The upper function approach in Theorem 2.3 of showing the consistency for an M-estimator can be rather rough and the bound (2.10) could lead to quite large values of $\mathtt{r}_0 > 0$. As the obtained value $\mathtt{r}_0 > 0$ enters into the error term $\breve{\diamondsuit}(\mathtt{r}_0, \mathtt{x}) > 0$ of Theorem 2.2 it is desirable to obtain a general refined bound for $\mathtt{r}_1 \le \mathtt{r}_0$ that still ensures that $\mathbb{P}(\widetilde{\boldsymbol{\upsilon}} \in \Upsilon_\circ(\mathtt{r}_1)) \ge 1 - \mathtt{C}\mathrm{e}^{-\mathtt{x}}$ with a small constant $\mathtt{C} > 0$. Such an improvement is possible as the following proposition shows. Define the *parametric spread*:

$$\diamondsuit(\mathtt{r}, \mathtt{x}) \stackrel{\text{def}}{=} \{\delta(\mathtt{r}) + 6\,\nu_1\,\mathfrak{z}(\mathtt{x}, 4p^*)\,\omega\}\,\mathtt{r}, \tag{2.11}$$

where $\mathfrak{z}(\mathtt{x}, \mathbb{Q})$ is defined in (2.5). Further with $\mathcal{V}^2 \in \mathbb{R}^{p^* \times p^*}$ from condition $(\mathcal{ED}_0)$ introduce the *misspecification matrix* $B \in \mathbb{R}^{p^* \times p^*}$ given by the famous sandwich formula; see [14]:

$$B = \mathcal{D}^{-1}\mathcal{V}^2\mathcal{D}^{-1}.$$

In the case of correct model specification with $\mathcal{D}^2 = \mathcal{V}^2$, the *sandwich matrix* B becomes the identity: $B = \mathbb{I}_{p^*}$. Theorem A.1 tells us that

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

where $\mathfrak{z}(\mathtt{x}, B) \le \mathtt{C}\sqrt{\operatorname{tr}(B^2) + \mathtt{x}}$ for moderate choice of $\mathtt{x} > 0$, see (A.2).

Proposition 2.4. *Assume the conditions of Theorem 2.2 and additionally assume* $(\mathcal{ED}_1)$, $(\mathcal{L}_0)$ *and* $(\mathcal{ED}_0)$ *with* $\mathcal{V}^2 \in \mathbb{R}^{p^* \times p^*}$. *Let* $\mathtt{r}_0 > 0$ *be such that* (2.9) *holds and define the radius*

$$\mathtt{r}_1 \stackrel{\text{def}}{=} \mathfrak{z}(\mathtt{x}, B) + \diamondsuit(\mathtt{r}_0, \mathtt{x}) \wedge \mathtt{r}_0 \le \mathtt{r}_0.$$

Then the result of Theorem 2.2 applies with the error term $\breve{\diamondsuit}(\mathtt{r}_1, \mathtt{x})$ *in place of* $\breve{\diamondsuit}(\mathtt{r}_0, \mathtt{x})$ *and with probability greater* $1 - 5\mathrm{e}^{-\mathtt{x}}$.

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. This means that given i.i.d. $(\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n) \in \otimes_{i=1}^n \mathcal{Y}$ we use

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

where $\ell : \mathcal{Y} \times \Upsilon \to \mathbb{R}$ 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

$$\boldsymbol{\upsilon}^* \stackrel{\text{def}}{=} \operatorname*{argmax}_{\boldsymbol{\upsilon} \in \Upsilon} \mathbb{E}\mathcal{L}(\boldsymbol{\upsilon}) = \operatorname*{argmax}_{\boldsymbol{\upsilon} \in \Upsilon} \mathbb{E}\ell(\boldsymbol{\upsilon}),$$

$$\mathcal{D}^2 \stackrel{\mathrm{def}}{=} \nabla^2 \mathbb{E}\mathcal{L}(\boldsymbol{\upsilon}^*) = nd^2 \stackrel{\mathrm{def}}{=} n\nabla^2 \mathbb{E}\ell(\boldsymbol{\upsilon}^*),$$
$$\mathcal{V}^2 \stackrel{\mathrm{def}}{=} \operatorname{Cov}(\nabla \boldsymbol{\zeta}(\boldsymbol{\upsilon}^*) = nv^2 \stackrel{\mathrm{def}}{=} n \operatorname{Cov}(\nabla(\ell - \mathbb{E}\ell)(\boldsymbol{\upsilon}^*)).$$

To check the conditions of section 2.1 in principle we only have 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 $\mathtt{g} = \mathtt{g}_1$. Under these conditions, one can easily check the conditions in Section 2.1 for the full log-likelihood $\mathcal{L}(\boldsymbol{\upsilon}) = \sum_{i=1}^n \ell(y_i, \boldsymbol{\upsilon})$ 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 [29]. To gain a bit more intuition let us consider the following stronger sufficient list of assumptions:

(ℓ_0) The matrix valued function $\nabla^2 \mathbb{E}[\ell(\cdot)] : \Upsilon \to \mathbb{R}^{p^* \times p^*}$ is locally Lipschitz continuous with Lipschitz constant δ^* in an open neighborhood $U \ni \boldsymbol{\upsilon}^*$.

(ed_1) There are constants $\nu_0^*, \mathtt{g}^* > 0$ and an open neighborhood $U \ni \boldsymbol{\upsilon}^*$ such that for all $\boldsymbol{\upsilon} \in U$ the random matrix valued function $\nabla^2(\ell - \mathbb{E}\ell)(\cdot, \boldsymbol{Y})\Upsilon \to \mathbb{R}^{p^* \times p^*}$ satisfies for all $|\lambda| \le \mathtt{g}^*$

$$\sup_{\substack{\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2 \in \mathbb{R}^{p^*} \\ \|\boldsymbol{\gamma}_1\| = \|\boldsymbol{\gamma}_2\| = 1}} \log \mathbb{E} \exp\left\{ \lambda \boldsymbol{\gamma}_1^\top d^{-1} \nabla^2 (\ell - \mathbb{E}\ell)(\boldsymbol{\upsilon}) d^{-1} \boldsymbol{\gamma}_2 \right\} \le \nu_0^* \lambda^2 / 2.$$

(ed_0) The random vector valued function $\nabla(\ell - \mathbb{E}\ell)(\cdot, \boldsymbol{Y})\Upsilon \to \mathbb{R}^{p^* \times p^*}$ satisfies for all $|\lambda| \le \mathtt{g}^*$ and all $\boldsymbol{\upsilon} \in \Upsilon$

$$\sup_{\substack{\boldsymbol{\gamma} \mathbb{R}^{p^*} \\ \|\boldsymbol{\gamma}\| = 1}} \log \mathbb{E} \exp\left\{ \lambda \boldsymbol{\gamma}^\top d^{-1} \nabla (\ell - \mathbb{E}\ell)(\boldsymbol{\upsilon}) \right\} \le \nu_0^* \lambda^2 / 2.$$

($\ell_{\mathtt{r}}$) There is a constant $\mathtt{b}^* > 0$ such that

$$\mathbb{E}\left[\ell(\boldsymbol{\upsilon}) - \ell(\boldsymbol{\upsilon}^*)\right] \ge \mathtt{b}^* \|d(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^*)\|^2.$$

($\boldsymbol{\iota}$) There is a constant $c_d > 0$ such that the matrix $d^2 \stackrel{\mathrm{def}}{=} \nabla^2 \mathbb{E}\ell(\boldsymbol{\upsilon}^*)$ satisfies $\boldsymbol{\gamma}^\top d^2 \boldsymbol{\gamma} \ge c_d \|\boldsymbol{\gamma}\|^2$ for all $\boldsymbol{\gamma} \in \mathbb{R}^{p^*}$.

Lemma 2.5. *Assume that $n \in \mathbb{N}$ is large enough to ensure that the local neighborhood $U \subset \Upsilon$ of $\boldsymbol{\upsilon}^*$ from conditions (ℓ_0) and (ed_1) satisfies*

$$\begin{aligned} \Upsilon_\circ(\mathtt{r}^*) &\stackrel{\mathrm{def}}{=} \{\boldsymbol{\upsilon} \in \Upsilon : \|\mathcal{D}(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^*)\| \le \mathtt{r}^*\} \\ &= \frac{1}{\sqrt{n}} \{\boldsymbol{\upsilon} \in \Upsilon : \|d(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^*)\| \le \mathtt{r}^*\} \subseteq U. \end{aligned}$$

Then the conditions (ℓ_0), (ed_1), (ed_0), ($\ell_{\mathtt{r}}$) and (ι) imply ($\mathcal{L}_0$), ($\mathcal{ED}_1$), ($\mathcal{ED}_0$), ($\mathcal{ED}_{\mathtt{r}}$), ($\mathcal{L}_0$) 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 $\mathtt{r} \le \mathtt{r}^*$. Further $\rho^2 \ge 1 - \frac{c_d}{\|d_{\boldsymbol{\theta}}^2\| \vee \|h^2\|}$ where $d_{\boldsymbol{\theta}}^2 = \Pi_{\boldsymbol{\theta}}^\top d \Pi_{\boldsymbol{\theta}} \in \mathbb{R}^{p \times p}$ and $h^2 = \Pi_{\boldsymbol{\eta}}^\top d \Pi_{\boldsymbol{\eta}} \in \mathbb{R}^{m \times m}$.

Remark 2.18. 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 2.1.

Noting that $\mathcal{L}(\widetilde{\boldsymbol{\upsilon}}, \boldsymbol{\upsilon}^*) \geq 0$ and $\mathcal{L}(\widetilde{\boldsymbol{\upsilon}}_{\boldsymbol{\theta}^*}, \boldsymbol{\upsilon}^*) \geq 0$ Theorem 2.3 yields that

$$\mathbb{P}\left(\widetilde{\boldsymbol{\upsilon}}, \widetilde{\boldsymbol{\upsilon}}_{\boldsymbol{\theta}^*} \in \Upsilon_\circ(\mathtt{r}_0)\right) \geq 1 - \mathrm{e}^{-\mathtt{x}}, \quad \text{with} \quad \mathtt{r}_0(\mathtt{x}) = 6\frac{\nu_0^*}{\mathtt{b}^*}\sqrt{2p^* + \mathtt{x}}.$$

Theorem 2.2 applies with $\mathcal{D}^2 = n\nabla^2\mathbb{E}\ell(\boldsymbol{\upsilon}^*)$ and $\boldsymbol{\upsilon}^\circ = \boldsymbol{\upsilon}^*$. We immediately obtain the following result.

Corollary 2.6. *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 $\mathtt{r}_0(\mathtt{x}) = 6\frac{\nu_0^*}{\mathtt{b}^*}\sqrt{2p^* + \mathtt{x}} \leq \mathtt{r}^*$. Then we get the Fisher and Wilks results of Theorem 2.2 for $\mathtt{x} \ll \sqrt{n}\mathtt{g}^*$ with*

$$\breve{\diamondsuit}(\mathtt{r}_0, \mathtt{x}) \leq \frac{36\nu_0^*}{\sqrt{n}\mathtt{b}^*}\left(\frac{4}{(1-\rho^2)^2}\frac{\delta^*}{c_d^3}\frac{\nu_0}{\mathtt{b}^*}(\mathtt{x} + 2p^*) + \nu_0\mathfrak{z}(\mathtt{x}, 2p^* + 2p)\sqrt{\mathtt{x} + 2p^*}\right).$$

Remark 2.19. The definition of $\mathfrak{z}(\mathtt{x}, 2p^* + 2p)$ in (C.2) implies for moderate values of $\mathtt{x} > 0$ that

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

with some fixed constant $\mathtt{C}_\diamondsuit$. The Fisher result (2.6) 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 (2.7). However, if the target dimension p is of order p^*, the constraint for the Wilks theorem becomes $p^* = o(n^{1/3})$. See [2] for an example that shows, that this difference actually occurs in certain examples.

2.5. Critical dimension

This section discusses 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 2.2 refined by Proposition 2.4 are accurate if the spread function $\diamondsuit(\mathtt{r}, \mathtt{x})$ from (2.11) fulfills $\diamondsuit(\mathtt{r}_0, \mathtt{x}) \leq \mathfrak{z}(\mathtt{x}, B)$ and $\breve{\diamondsuit}(\mathtt{r}_1, \mathtt{x})$ is small, with $\mathtt{r}_1 = 2\mathfrak{z}(\mathtt{x}, B)$. Usually $\mathfrak{z}(\mathtt{x}, B) \leq \mathtt{C}\sqrt{\mathtt{x} + p^*}$ leading to

$$\breve{\diamondsuit}(\mathtt{r}_1, \mathtt{x}) \asymp \breve{\delta}(\mathtt{r}_1)\mathtt{r}_1 + \breve{\omega}\mathtt{r}_1^2 \quad \text{is small for } \mathtt{r}_1^2 \asymp p^*. \tag{2.12}$$

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 2.6) the condition (2.12) 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 series of papers on growing dimension in generalized linear models and for natural exponential families, see

e.g. [26, 25, 27]. Our results are non-asymptotic and apply to general statistical models under the conditions of Section 2.1. The following example shows that the constraint "p^{*2}/n is small" is critical.

Consider single observation model

$$\begin{aligned} \boldsymbol{Y} &= f(\boldsymbol{\upsilon}) + \boldsymbol{\varepsilon}, \\ f(\boldsymbol{\upsilon}) &= f(\theta, \boldsymbol{\eta}) \stackrel{\text{def}}{=} \begin{pmatrix} \theta \\ \boldsymbol{\eta}_1 \\ \vdots \\ \boldsymbol{\eta}_{p_n - 1} \end{pmatrix} + \begin{pmatrix} \|\boldsymbol{\eta}\|^2 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \in \mathbb{R}^{p_n}, \end{aligned}$$

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 2.7. *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 2.20. 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

$$\mathbb{P}\left(\theta^* \in \left\{|\widetilde{\theta} - \theta|^2 \le q_\alpha / n\right\}\right) \to \alpha. \tag{2.13}$$

But the proof of Proposition 2.7 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 median of a chi-square distribution converges to its number of degrees of freedom when the degrees of freedom tend to infinity. This means that for any $0 < \epsilon < 1$ the set

$$C \stackrel{\text{def}}{=} \left\{n\|\boldsymbol{\varepsilon}_{\boldsymbol{\eta}}\|^2 \ge (1 - \epsilon) p_n\right\},$$

is of probability greater $1/2$ for $n, p_n \in \mathbb{N}$ 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

$$\mathbb{P}\left(\theta^* \in \left\{|\widetilde{\theta} - \theta|^2 \le q_\alpha / n\right\}\right) = \int_0^\infty \mathbb{P}\left(|\varepsilon_\theta - z/n|^2 \le q_\alpha / n\right) f_{\chi^2_m}(z) dz$$

$$= \int_0^\infty \left[\Phi\left(\frac{z}{\sqrt{n}} + \sqrt{q_\alpha} \right) - \Phi\left(\frac{z}{\sqrt{n}} - \sqrt{q_\alpha} \right) \right] f_{\chi^2_m}(z) dz$$
$$< \frac{1}{2} \left[\alpha + \Phi\left(\frac{(1-\epsilon)p_n}{\sqrt{n}} + \sqrt{q_\alpha} \right) - \Phi\left((1-\epsilon)\frac{p_n}{\sqrt{n}} - \sqrt{q_\alpha} \right) \right],$$

where $\Phi : \mathbb{R} \to [0,1]$ denotes the distribution function of a standard normal random variable. If $p_n/\sqrt{n}$ is significantly larger than 0, the value

$$\Phi\left(\frac{(1-\epsilon)p_n}{\sqrt{n}} + \sqrt{q_\alpha} \right) - \Phi\left(\frac{(1-\epsilon)p_n}{\sqrt{n}} - \sqrt{q_\alpha} \right),$$

is distinctively smaller α. For example for $\alpha = 0.95$ and $(1-\epsilon)p_n/\sqrt{n} = 11/12$ we get

$$\mathbb{P}\left(\theta^* \in \left\{ |\widetilde{\theta} - \theta|^2 \le q_{0.95}/n \right\} \right) < 0.9.$$

In other words the asymptotic confidence statement in (2.13) is way off in the finite sample case because the error term in the local linear approximation is not addressed. This is exactly where a large full dimension has an impact on the behavior of the estimator. Our results in Theorem 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 address this effect. The price are more conservative sets, but their coverage property is ensured.

Remark 2.21. There is an interesting connection of the condition $p^*/\sqrt{n} \to 0$ with the general theory on semiparametric M-estimators. In the common asymptotic approach to semiparametric M estimators one assumes apriori consistency of the estimator $\widetilde{\upsilon} = (\widetilde{\boldsymbol{\theta}}, \widetilde{\boldsymbol{\eta}})$, more precisely in the case that the functional $\breve{\nabla}\mathcal{L}(\cdot)$ is smooth enough one assumes that $\|\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\| = o_{\mathbb{P}}(1)$ and $\|\widetilde{\boldsymbol{\eta}} - \boldsymbol{\eta}^*\| = O_{\mathbb{P}}(n^{-1/4})$, see [18] Section 21.1.4. On the other hand the results of Theorem 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 $\breve{\diamondsuit}(\mathtt{r}_0, \mathtt{x}) = o(1)$. Neglecting the contribution of $\|\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|$ to $\mathtt{r}_0$ this can be ensured if

$$\breve{\diamondsuit}(\mathtt{r}_0, \mathtt{x}) \le \mathtt{C}(p^* + \mathtt{r}_0^2)/\sqrt{n} \le o(1) + \mathtt{C}\sqrt{n}\|\widetilde{\boldsymbol{\eta}} - \boldsymbol{\eta}^*\|^2 \to 0,$$

i.e. if $\|\widetilde{\boldsymbol{\eta}} - \boldsymbol{\eta}^*\| = o(n^{-1/4})$. But consider the radius $\mathtt{r}_1 > 0$ from Proposition 2.4. It is of order $\sqrt{p^* + m}$ if $\diamondsuit(\mathtt{r}_0) = O(\sqrt{p^* + \mathtt{x}})$. In that case in the i.i.d. setting the constraint on the apriori deviation bound becomes $\diamondsuit(\mathtt{r}_0, \mathtt{x}) = O(\sqrt{p^* + \mathtt{x}})$ which can be ensured if

$$\diamondsuit(\mathtt{r}_0, \mathtt{x}) \le \mathtt{C}(p^* + \mathtt{r}_0^2)/\sqrt{n} \le o(1) + \mathtt{C}\sqrt{n}\|\widetilde{\boldsymbol{\eta}} - \boldsymbol{\eta}^*\|^2 = O(\sqrt{p^* + \mathtt{x}}),$$

which means if $p^* + \mathtt{x} = o(\sqrt{n})$ that $\|\widetilde{\boldsymbol{\eta}} - \boldsymbol{\eta}^*\| = o(n^{-1/8})$, which is a considerably weaker constraint. These bounds only concern the finite dimensional case. In the infinite dimensional setting, treated in Section 2.8 we have to impose conditions

that ensure that the bias induced by the sieve approach is small enough. [3] serve such conditions for the Hilbertspace setting. One of these conditions reads that $\|H(\boldsymbol{\eta}^* - \Pi_m \boldsymbol{\eta}^*)\|^2 \leq \mathtt{C} m$, i.e. the true nuisance component $\boldsymbol{\eta}^* \in \mathcal{X}$ is well approximated by its projection into the span of the first $m \in \mathbb{N}$ basis elements $(\boldsymbol{e}_k) \subset \mathcal{X}$. If we represent with some $\alpha > 0$

$$\boldsymbol{\eta}^* = \sum_{k=1}^{m} \eta_k^* \boldsymbol{e}_k, \quad \sum_{k=1}^{m} \eta_k^{*2} k^{2\alpha} < \infty,$$

we obtain the constraint $m \leq n^{1/(2\alpha+1)}$, which means that we need $\alpha > 1/2$ to get $m = o(n^{1/2})$ and in that case $\boldsymbol{\eta}^* \in \mathcal{X}$ is nonparametrically estimable with rate $n^{-1/4}$.

Remark 2.22. Concerning the difference between the critical dimensions in the Wilks (2.7) and the Fisher expansion (2.6) we remark that [2] presents an example where $p = p^*/2$, where all conditions of Section 2.1 are met with $\breve{\delta}(\mathtt{r})/\mathtt{r} \asymp \breve{\omega} \asymp 1/\sqrt{n}$ and where the Wilks phenomenon occurs iff ${p^*}^3/n \to 0$ while for the Fisher expansion ${p^*}^2/n$ suffices.

2.6. Infinite dimensional nuisance

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. Then 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. The ideas can be modified to treat the case when the nuisance parameter belongs to a Banach space.

2.7. Sieve approach

Consider the $(\boldsymbol{\theta}, \boldsymbol{f})$-setup with $\boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^p$ and $\boldsymbol{f} \in \mathcal{X}$, where $\mathcal{X}$ is an infinite dimensional separable Hilbert space. The target parameter $\boldsymbol{\theta}^*$ can be defined as

$$\boldsymbol{\theta}^* = \operatorname*{argmax}_{\boldsymbol{\theta}} \sup_{\boldsymbol{f} \in \mathcal{X}} \mathbb{E}\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{f}). \tag{2.14}$$

As the Hilbert space $\mathcal{X}$ is assumed to be separable it possesses a countable orthonormal basis $\{\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots\} \subset \mathcal{X}$. Any vector $\boldsymbol{f} \in \mathcal{X}$ admits a unique decomposition in the form

$$\boldsymbol{f} = \sum_{j=1}^{\infty} \eta_j \boldsymbol{e}_j,$$

where $\eta_j = \langle \boldsymbol{f}, \boldsymbol{e}_j \rangle$ 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 $\boldsymbol{f}$ can be approximated by a finite linear combination $\boldsymbol{f}_m(\boldsymbol{\eta})$ of the $\boldsymbol{e}_j$'s:

$$\boldsymbol{f}_m(\boldsymbol{\eta}) \stackrel{\mathrm{def}}{=} \sum_{j=1}^{m} \eta_j \boldsymbol{e}_j.$$

We denote the parameter by $\boldsymbol{\upsilon} = (\boldsymbol{\theta}, \boldsymbol{\eta}) \in \mathbb{R}^p \times l^2$. In the following we will need to quantify the accuracy of approximating $\boldsymbol{f}$ by $\boldsymbol{f}_m$ as m grows; see condition **(*bias*)** below.

Let $\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{f})$ be the log-likelihood in the original model. Define by abuse of notation

$$\mathcal{L}(\boldsymbol{\upsilon}) \stackrel{\mathrm{def}}{=} \mathcal{L}\left(\boldsymbol{\theta}, \sum_{j=1}^{\infty} \eta_j \boldsymbol{e}_j\right)$$

$$\boldsymbol{\upsilon}^* \stackrel{\mathrm{def}}{=} \operatorname*{argmax}_{(\boldsymbol{\theta},\boldsymbol{\eta}) \in l^2} \mathbb{E}\left[\mathcal{L}\left(\boldsymbol{\theta}, \sum_{k=1}^{\infty} \eta_j \boldsymbol{e}_j\right)\right],$$

and the m-dimensional sieve approximation $\mathcal{L}_m(\boldsymbol{\upsilon})$ of $\mathcal{L}(\boldsymbol{\upsilon})$ by

$$\mathcal{L}_m(\boldsymbol{\theta}, \boldsymbol{\eta}) \stackrel{\mathrm{def}}{=} \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{f}_m(\boldsymbol{\eta})),$$

$$(\boldsymbol{\theta}, \boldsymbol{\eta}) \in \Upsilon_m \stackrel{\mathrm{def}}{=} \{\boldsymbol{\upsilon} = (\boldsymbol{\theta}, \boldsymbol{\eta}) \in \mathbb{R}^{p^*} : (\boldsymbol{\theta}, \boldsymbol{f}_m(\boldsymbol{\eta})) \in \Upsilon\}.$$

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:

$$\widetilde{\boldsymbol{\theta}}_m \stackrel{\mathrm{def}}{=} \Pi_{\boldsymbol{\theta}} \widetilde{\boldsymbol{\upsilon}}_m \stackrel{\mathrm{def}}{=} \Pi_{\boldsymbol{\theta}} \operatorname*{argmax}_{\boldsymbol{\upsilon} \in \Upsilon_m} \mathcal{L}_m(\boldsymbol{\theta}, \boldsymbol{\eta}), \tag{2.15}$$

$$\boldsymbol{\theta}^*_m \stackrel{\mathrm{def}}{=} \Pi_{\boldsymbol{\theta}} \boldsymbol{\upsilon}^*_m \stackrel{\mathrm{def}}{=} \Pi_{\boldsymbol{\theta}} \operatorname*{argmax}_{\boldsymbol{\upsilon} \in \Upsilon_m} \mathbb{E}\mathcal{L}_m(\boldsymbol{\theta}, \boldsymbol{\eta}).$$

The question we are interested in can be formulated as follows: is $\widetilde{\boldsymbol{\theta}}_m$ a good (efficient) estimator of $\boldsymbol{\theta}^*$ from (2.14) under a proper choice of m?

2.8. Bias constraints and efficiency

The parametric results obtained in Section 2 claim that $\widetilde{\boldsymbol{\theta}}_m \in \mathbb{R}^p$ estimates well $\boldsymbol{\theta}^*_m \in \mathbb{R}^p$ if the spread $\ddot{\Diamond}(\mathtt{r}_0, \mathtt{x}) > 0$ is small. More precisely we have the following: Define for fixed $\mathtt{x} > 0$ the value $\mathtt{r}_0 > 0$ by

$$\mathtt{r}_0(\mathtt{x}) \stackrel{\mathrm{def}}{=} \inf_{\mathtt{r} \geq 0} \left\{\mathbb{P}\left\{\widetilde{\boldsymbol{\upsilon}}_m, \widetilde{\boldsymbol{\upsilon}}_{\boldsymbol{\theta}^*_m, m} \in \Upsilon_{0,m}(\mathtt{r})\right\} \geq 1 - \mathrm{e}^{-\mathtt{x}}\right\},$$

$$\Upsilon_{0,m}(\mathtt{r}) \stackrel{\mathrm{def}}{=} \left\{\boldsymbol{\upsilon} \in \Upsilon_m,\ \|\mathcal{D}_m(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^*_m)\| \leq \mathtt{r}\right\}.$$

where $\boldsymbol{\upsilon}^*_m = (\boldsymbol{\theta}^*_m, \boldsymbol{\eta}^*_m) = \operatorname{argmax}_{\boldsymbol{\upsilon}} \mathbb{E}\mathcal{L}_m(\boldsymbol{\upsilon})$ and

$$\widetilde{\boldsymbol{\upsilon}}_{\boldsymbol{\theta},m} \stackrel{\text{def}}{=} \operatorname*{argmax}_{\substack{\boldsymbol{\upsilon}\in\Upsilon_m \\ \Pi_{\boldsymbol{\theta}}\boldsymbol{\upsilon}=\boldsymbol{\theta}}} \mathcal{L}_m(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^*).$$

Further the matrix $\breve{D}^2_m$ is defined as

$$\breve{D}^2_m(\boldsymbol{\upsilon}^*_m) \stackrel{\text{def}}{=} \left(\Pi_{\boldsymbol{\theta}} \mathcal{D}_m^{-2} \Pi_{\boldsymbol{\theta}}^\top\right)^{-1} \in \mathbb{R}^{p\times p}, \quad \mathcal{D}^2_m \stackrel{\text{def}}{=} \nabla^2_{p+m}\mathbb{E}[\mathcal{L}(\boldsymbol{\upsilon}^*_m)] \in \mathbb{R}^{p^*\times p^*},$$

i.e. the derivatives of $\mathbb{E}[\mathcal{L}]$ 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 2.2 to $\widetilde{\boldsymbol{\theta}}_m$ from (2.15) we find that with probability greater $1 - 2\mathrm{e}^{-\mathtt{x}}$

$$\|\breve{D}_m\big(\widetilde{\boldsymbol{\theta}}_m - \boldsymbol{\theta}^*_m\big) - \breve{\boldsymbol{\xi}}_m(\boldsymbol{\upsilon}^*_m)\| \le \breve{\Diamond}(\mathtt{r}_0, \mathtt{x}), \tag{2.16}$$

The result (2.16) involves two kinds of bias, one that concerns the difference $\boldsymbol{\theta}^*_m - \boldsymbol{\theta}^*$ and the other the difference between $\breve{D}_m \in \mathbb{R}^{p\times p}$ and $\breve{D} \in \mathbb{R}^{p\times p}$ where

$$\breve{D}^2 \stackrel{\text{def}}{=} \left(\Pi_{\boldsymbol{\theta}} \nabla^2 \mathbb{E}[\mathcal{L}(\boldsymbol{\upsilon}^*)]^{-1} \Pi_{\boldsymbol{\theta}}^\top\right)^{-1} \in \mathbb{R}^{p\times p},$$

i.e. the derivatives of $\mathbb{E}[\mathcal{L}]$ are taken with respect to all coordinates of $\boldsymbol{\upsilon} \in l^2$ and the Hessian is calculated in the "true point" $\boldsymbol{\upsilon}^* \in l^2$. The second bias – i.e. bounds for $\|I - \breve{D}^{-1}_m(\boldsymbol{\upsilon}^*_m)\breve{D}^2(\boldsymbol{\upsilon}^*)\breve{D}^{-1}_m(\boldsymbol{\upsilon}^*_m)\|$ – will be neglected for now, as only the operator $\breve{D}^2_m(\boldsymbol{\upsilon}^*_m) \in \mathbb{R}^{p\times p}$ is available in practice. We will come back to it, when we derive efficiency for the sieve profile estimator $\widetilde{\boldsymbol{\theta}}_m \in \mathbb{R}^p$.

Remark 2.23. To be more precise we assume that $\mathbb{E}\mathcal{L} : \Upsilon \to \mathbb{R}$ is Fréchet differentiable and that each element of the gradient $\langle \nabla \mathbb{E}\mathcal{L}, \boldsymbol{e}_k\rangle$ again is Fréchet differentiable aswell. We denote the resulting operator by $\mathcal{D}^2 = \nabla^2\mathbb{E}[\mathcal{L}(\boldsymbol{\upsilon}^*)] : \overline{\operatorname{span}}\Upsilon \to \overline{\operatorname{span}}\Upsilon$.

For the first type of bias we impose the following condition:

(*bias*) There exists a function $\alpha : \mathbb{N} \to \mathbb{R}_+$ such that

$$\|\breve{D}_m(\boldsymbol{\upsilon}^*_m)(\boldsymbol{\theta}^*_m - \boldsymbol{\theta}^*)\| \le \alpha(m), \qquad \alpha(m) \to 0, \ \text{as } m \to \infty.$$

Remark 2.24. For now we focus on the result 2.2 and thus we do not elaborate on approximation theory. But [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

$$\mathcal{D}^2_m(\boldsymbol{\upsilon}^*_m) = \begin{pmatrix} D^2(\boldsymbol{\upsilon}^*_m) & A^\top_m(\boldsymbol{\upsilon}^*_m) \\ A_m(\boldsymbol{\upsilon}^*_m) & H^2_m(\boldsymbol{\upsilon}^*_m) \end{pmatrix} \in \mathbb{R}^{(p+m)\times(p+m)}.$$

With Theorem 2.2 and **(*bias*)** we directly get the following corollary:

Corollary 2.8. *Assume (bias) and that the conditions* $(\breve{\mathcal{E}}\mathcal{D})$, $(\breve{\mathcal{E}}\mathcal{D}_1)$ *and* $(\breve{\mathcal{L}}_0)$ *from Section 2.1 are satisfied for all* $m \geq m_0$ *for some* $m_0 \in \mathbb{N}$ *and with* $\mathcal{D}^2 = \nabla^2_{p+m}\mathbb{E}\mathcal{L}_m(\boldsymbol{\upsilon}^*_m) \in \mathbb{R}^{p^*\times p^*}$, $\mathcal{V}^2 = \mathrm{Cov}[\nabla_{p+m}\mathcal{L}_m(\boldsymbol{\upsilon}^*_m)] \in \mathbb{R}^{p^*\times p^*}$ *and* $\boldsymbol{\upsilon}^\circ = \boldsymbol{\upsilon}^*_m \in \mathbb{R}^{p^*}$. *Assume that* $\widetilde{\boldsymbol{\upsilon}}_m \neq \{\}$ *and* $\widetilde{\boldsymbol{\upsilon}}_{\boldsymbol{\theta}^*_m} \neq \{\}$. *Choose* $\mathtt{r}_0(\mathtt{x}) > 0$ *such that* $\mathbb{P}(\widetilde{\boldsymbol{\upsilon}}_m, \widetilde{\boldsymbol{\upsilon}}_{\boldsymbol{\theta}^*_m,m} \in \Upsilon_{0,m}(\mathtt{r}_0(\mathtt{x}))) \geq 1 - \mathrm{e}^{-\mathtt{x}}$. *Then it holds for any* $m \geq m_0$ *with probability greater* $1 - 2\mathrm{e}^{-\mathtt{x}_n}$

$$\bigl\|\breve{D}_m(\boldsymbol{\upsilon}^*_m)\bigl(\widetilde{\boldsymbol{\theta}}_m - \boldsymbol{\theta}^*\bigr) - \breve{\boldsymbol{\xi}}_m(\boldsymbol{\upsilon}^*_m)\bigr\| \leq \breve{\Diamond}(\mathtt{r}_0, \mathtt{x}) + \alpha(m),$$

where

$$\breve{\boldsymbol{\xi}}_m(\boldsymbol{\upsilon}^*_m) \stackrel{\mathrm{def}}{=} \breve{D}_m^{-1}(\nabla_{\boldsymbol{\theta}} - A_m H_m^{-1}\nabla_{\boldsymbol{\eta}})\mathcal{L}_m(\boldsymbol{\upsilon}^*_m).$$

Define

$$\breve{L}(\boldsymbol{\theta}) \stackrel{\mathrm{def}}{=} \max_{\boldsymbol{\eta}\in\mathbb{R}^m} \mathcal{L}_m(\boldsymbol{\theta}, \boldsymbol{\eta}),$$

where it is important to note that the maximization is restricted to the finite dimensional space $\mathbb{R}^m$. As above abbreviate $\breve{L}(\boldsymbol{\theta}, \boldsymbol{\theta}^*) \stackrel{\mathrm{def}}{=} \breve{L}(\boldsymbol{\theta}) - \breve{L}(\boldsymbol{\theta}^*)$. For the bias in the Wilks result a bit more work is needed. We can show the following:

Theorem 2.9. *Assume the same as in Corollary 2.8. Pick a radius* $0 < \mathtt{r}_0^\circ$ *such that*

$$\mathbb{P}\bigl(\{\widetilde{\boldsymbol{\upsilon}}_m, \widetilde{\boldsymbol{\upsilon}}_{\boldsymbol{\theta}^*_m,m}, \widetilde{\boldsymbol{\upsilon}}_{\boldsymbol{\theta}^*,m} \in \Upsilon_{0,m}(\mathtt{r}_0^\circ)\}\bigr) > 1 - \mathrm{e}^{-\mathtt{x}},$$

Then we get with probability greater $1 - 2\mathrm{e}^{-\mathtt{x}}$

$$\begin{aligned}
&\bigl|2\breve{L}(\widetilde{\boldsymbol{\theta}}_m, \boldsymbol{\theta}^*) - \|\breve{\boldsymbol{\xi}}_m(\boldsymbol{\upsilon}^*_m)\|^2\bigr| \\
&\quad\leq 8\Bigl(\|\breve{\boldsymbol{\xi}}_m(\boldsymbol{\upsilon}^*_m)\| + \breve{\Diamond}(\mathtt{r}_0^\circ, \mathtt{x})\Bigr)\breve{\Diamond}(2(1+\rho)\mathtt{r}_0^\circ, \mathtt{x}) + \breve{\Diamond}(\mathtt{r}_0^\circ, \mathtt{x})^2 \\
&\qquad + \alpha(m)\Bigl(2\|\breve{\boldsymbol{\xi}}_m(\boldsymbol{\upsilon}^*_m)\| + \alpha(m) + 2\breve{\Diamond}(2(1+\rho)\mathtt{r}_0^\circ, \mathtt{x})\Bigr)
\end{aligned}$$

Remark 2.25. With condition $(\breve{\mathcal{E}}\mathcal{D}_0)$ we can use Theorem A.1 to obtain

$$\mathbb{P}\Bigl(\|\breve{\boldsymbol{\xi}}_m(\boldsymbol{\upsilon}^*_m)\| \geq \mathfrak{z}(\mathtt{x}, \breve{B})\Bigr) \leq \mathrm{e}^{-\mathtt{x}}.$$

Remark 2.26. The radius $\mathtt{r}_0^\circ \in \mathbb{R}$ can be determined again using the tools of Section 2.3. Clearly Theorem 2.3 can be applied to find some $\mathtt{r}_0 \leq \mathtt{r}_0^\circ$ such that

$$\mathbb{P}\bigl(\widetilde{\boldsymbol{\upsilon}}_m, \widetilde{\boldsymbol{\upsilon}}_{\boldsymbol{\theta}^*_m,m} \in \Upsilon_{0,m}(\mathtt{r}_0)\bigr) > 1 - \mathrm{e}^{-\mathtt{x}}.$$

Further note that by the mean value theorem

$$\begin{aligned}
\mathcal{L}_m(\boldsymbol{\theta}^*, \widetilde{\boldsymbol{\eta}}_{\boldsymbol{\theta}^*,m}) - \mathcal{L}_m(\boldsymbol{\upsilon}^*_m) &\geq \mathcal{L}_m(\boldsymbol{\theta}^*, \boldsymbol{\eta}^*_m) - \mathcal{L}_m(\boldsymbol{\upsilon}^*_m) \\
&\geq -(1+\rho)\alpha(m) \sup_{\boldsymbol{\upsilon}\in\Upsilon_\circ((1+\rho)\alpha(m))} \|D^{-1}\nabla_{\boldsymbol{\theta}}\mathcal{L}_m(\boldsymbol{\upsilon})\|.
\end{aligned}$$

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 $1-2\mathrm{e}^{-\mathtt{x}}$ using the tools of Section 2.3. Combining this with Theorem 2.3 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

$$\begin{aligned}\mathbb{P}\left(\{\widetilde{\boldsymbol{\upsilon}}_m, \widetilde{\boldsymbol{\upsilon}}_{\boldsymbol{\theta}^*_m,m} \in \Upsilon_{0,m}(\mathtt{r}_0)\} \cap \{\widetilde{\boldsymbol{\upsilon}}_{\boldsymbol{\theta}^*,m} \in \Upsilon_{0,m}(\mathtt{r}_0^\circ)\}\right) &> 1-4\mathrm{e}^{-\mathtt{x}-\log(4)} \\ &= 1-\mathrm{e}^{-\mathtt{x}}.\end{aligned}$$

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 2.4 this gives that $\mathcal{D}_m^2 = nd_m$, $\breve{D}_m^2 = n\breve{d}_m$ and $\breve{D}^2 = n\breve{d}$. As the efficient covariance is derived for the score evaluated in the true full target $\boldsymbol{\upsilon}^* \in l^2$ we need further assumptions on the bias:

(*bias′*) With $\|\cdot\|$ denoting the spectral norm and with some function $\beta(m) \to 0$ as $m \to \infty$

$$\begin{aligned}\|I - \breve{D}_m(\boldsymbol{\upsilon}^*)^{-1}\breve{D}(\boldsymbol{\upsilon}^*)^2\breve{D}_m(\boldsymbol{\upsilon}^*)^{-1}\| &\le \beta(m), \\ \|I - \breve{D}_m(\boldsymbol{\upsilon}^*_m)^{-1}\breve{D}_m(\boldsymbol{\upsilon}^*)^2\breve{D}_m(\boldsymbol{\upsilon}^*_m)^{-1}\| &\le \beta(m).\end{aligned}$$

Remark 2.27. This paper focuses on the result 2.2 and thus we do not elaborate on approximation theory. But [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′*)**.

Further we need convergence of the covariance of the weighted score. For this define

$$\begin{aligned}\breve{v}^2_{m,\mathcal{D}}(\boldsymbol{\upsilon}^*_m) &\stackrel{\mathrm{def}}{=} \mathrm{Cov}\left(\nabla_{\boldsymbol{\theta}}\ell_1(\boldsymbol{\upsilon}^*_m) - A_m H_m^{-2}\nabla_{\boldsymbol{\eta}}\ell_1(\boldsymbol{\upsilon}^*_m)\right), \\ \breve{v}^2(\boldsymbol{\upsilon}^*) &\stackrel{\mathrm{def}}{=} \mathrm{Cov}\left(\nabla_{\boldsymbol{\theta}}\ell_1(\boldsymbol{\upsilon}^*) - AH^{-2}\nabla_{\boldsymbol{\eta}}\ell_1(\boldsymbol{\upsilon}^*)\right).\end{aligned}$$

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

$$\|\breve{D}_m^{-1}(\boldsymbol{\upsilon}^*_m)\breve{V}^2_{m,\mathcal{D}}(\boldsymbol{\upsilon}^*_m)\breve{D}_m^{-1}(\boldsymbol{\upsilon}^*_m) - \breve{d}^{-1}\breve{v}^2\breve{d}^{-1}\| \to 0.$$

Remark 2.28. 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 2.8 and Theorem 2.9 allow to derive the following corollary which yields the asymptotic efficiency of $\widetilde{\boldsymbol{\theta}}_m$ and the classical Wilks phenomenon.

Corollary 2.10. *Assume that we have iid observations from* $\mathbb{P} = \mathbb{P}_{\boldsymbol{\theta}^*,\boldsymbol{\eta}^*}$ *and that for some* $m_0 \in \mathbb{N}$ *any* $m \geq m_0$ *the conditions of Theorem 2.8 and the condition* $(\breve{\mathcal{E}}\mathcal{D}_0)$ *are satisfied with* $\boldsymbol{\upsilon}^\circ = \boldsymbol{\upsilon}^*_m$. *Further 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, that* $\breve{\omega}_n \to 0$ *and 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* $\mathbb{P}(\widetilde{\boldsymbol{\upsilon}}_m, \widetilde{\boldsymbol{\upsilon}}_{\boldsymbol{\theta}^*_m,m}, \widetilde{\boldsymbol{\upsilon}}_{\boldsymbol{\theta}^*,m} \in \Upsilon_\circ(\mathtt{r}_0)) \geq 1 - \mathrm{e}^{-\mathtt{x}}$. *Then there is a sequence* $m_n \to \infty$ *such that as* $n \to \infty$

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

Remark 2.29. On this level of generality we can not specify the right choice of $m_n \in \mathbb{N}$ that ensures the convergence. But in [1] it is shown that it equals the optimal choice for a series estimator of the nuisance component $\boldsymbol{\eta}^* \in l^2$ for know $\boldsymbol{\theta}^*$- as pointed out in [24] the best choice is $m = n^{1/(2\alpha+1)}$, with $\alpha > 1/2$ quantifying the "smoothness" of $\boldsymbol{\eta}^*$- is admissible.

Remark 2.30. 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 $\mathbb{F}^2_{\boldsymbol{\upsilon}^*} \stackrel{\mathrm{def}}{=} \mathrm{Cov}\big\{\nabla\ell(\boldsymbol{\upsilon}^*)\big\} : l^2 \to Im(\mathbb{F}^2_{\boldsymbol{\upsilon}^*})$ is invertible and that $\nabla\ell(\boldsymbol{\upsilon}^*) \in Im(\mathbb{F}^2_{\boldsymbol{\upsilon}^*})$. It is known from the convolution theorem (see [32], Theorem 3.11.2 p. 414, setting $\varkappa(\mathbb{P}_{\boldsymbol{\upsilon}}) = \boldsymbol{\theta}$) that the asymptotically optimal variance for regular estimators is given by the inverse of the partial information matrix

$$
\breve{\mathbb{F}}_{\boldsymbol{\upsilon}^*} = \Big(\Pi_{\boldsymbol{\theta}}\,\mathrm{Cov}\big\{\nabla\ell(\boldsymbol{\upsilon}^*)\big\}^{-1}\Pi^\top_{\boldsymbol{\theta}}\Big)^{-1},
$$

where as above $\Pi_{\boldsymbol{\theta}}$ is the orthogonal projection onto the $\boldsymbol{\theta}$-components, and $\Pi^\top_{\boldsymbol{\theta}}$ its adjoint operator. In the 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 2.10 yields the efficiency of the sieve profile MLE and we recover the Wilks phenomenon for that estimator.

2.9. Application to single index model

We illustrate how the results from Section 2 and the last statement can be derived for Single Index modeling. We focus on the complete set of assumptions

that allow to apply the results from above. For a detailed treatment of this model see [1]. Consider the following model

$$\boldsymbol{Y}_i = f(\boldsymbol{X}_i^\top \boldsymbol{\theta}^*) + \varepsilon_i, \qquad i = 1, \dots, n,$$

for some $f : \mathbb{R} \to \mathbb{R}$ and $\boldsymbol{\theta}^* \in S_1^{p,+} \subset \mathbb{R}^p$, i.i.d errors $\varepsilon_i \in \mathbb{R}$ with $\mathbb{E}\varepsilon_i = 0$ and $\mathrm{Var}(\varepsilon_i) = \sigma^2$ and i.i.d random variables $\boldsymbol{X}_i \in \mathbb{R}^p$ with distribution denoted by $\mathbb{P}^{\boldsymbol{X}}$. The single-index model is widely applied in statistics. For example in econometric studies it serves as a compromise between too restrictive parametric models and flexible but hardly estimable purely nonparametric models. Usually the statistical inference focuses on estimating the index vector $\boldsymbol{\theta}^*$. A lot of research has already been done in this field. For instance, [7] show the asymptotic efficiency of the general semiparametric maximum-likelihood estimator for particular examples and in [12] the right choice of bandwidth for the nonparametric estimation of the link function is analyzed.

To ensure identifiability of $\boldsymbol{\theta}^* \in \mathbb{R}^p$ we assume that it lies in the half sphere $S_1^{p,+} \stackrel{\mathrm{def}}{=} \{\boldsymbol{\theta} \in \mathbb{R}^p : \|\boldsymbol{\theta}\| = 1, \theta_1 > 0\} \subset \mathbb{R}^p$. For simplicity we assume that the support of the $\boldsymbol{X}_i \in \mathbb{R}^p$ is contained in the ball of radius $s_{\boldsymbol{X}} > 0$. This allows to approximate $f \in \{f : [-s_{\boldsymbol{X}}, s_{\boldsymbol{X}}] \mapsto \mathbb{R}\}$ by an orthonormal C^2-Daubechies-wavelet basis, i.e. for a suitable function $\boldsymbol{e}_0 \stackrel{\mathrm{def}}{=} \psi : [-s_{\boldsymbol{X}}, s_{\boldsymbol{X}}] \mapsto \mathbb{R}$ we set for $k = (2^{j_k} - 1)13 + r_k$ with $j_k \in \mathbb{N}_0$ and $r_k \in \{0, \dots, (2^{j_k})13 - 1\}$

$$\boldsymbol{e}_k(t) = 2^{j_k/2} \psi\left(2^{j_k}(t - 2r_k s_{\boldsymbol{X}})\right), \; k \in \mathbb{N}.$$

Our aim is to analyze the properties of the profile MLE

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

where

$$\mathcal{L}_m(\boldsymbol{\theta}, \boldsymbol{\eta}) \stackrel{\mathrm{def}}{=} -\frac{1}{2} \sum_{i=1}^n \Big| \boldsymbol{Y}_i - \sum_{k=0}^m \boldsymbol{\eta}_k \boldsymbol{e}_k(\boldsymbol{X}_i^\top \boldsymbol{\theta}) \Big|^2.$$

[16] analyzed a very similar estimator in a more general setting based on a kernel estimation of $\mathbb{E}\big[\boldsymbol{Y} \,\big|\, f(\boldsymbol{\theta}^\top \boldsymbol{X})\big]$ instead of using a parametric sieve approximation $\sum_{k=0}^m \boldsymbol{\eta}_k \boldsymbol{e}_k$. He showed $\sqrt{n}$-consistency and asymptotic normality of the proposed estimator.

To apply the technique presented above we need a list of assumptions denoted by $(\mathcal{A})$:

($\mathbf{Cond}_{\boldsymbol{X}}$) The measure $\mathbb{P}^{\boldsymbol{X}}$ is absolutely continuous with respect to the Lebesgue measure. The Lebesgue density $d_{\boldsymbol{X}} : \mathbb{R}^p \to \mathbb{R}$ of $\mathbb{P}^{\boldsymbol{X}}$ is only positive on the ball $B_{s_{\boldsymbol{x}}+h}(0) \subset \mathbb{R}^p$ with some small $h > 0$ and Lipschitz continuous on $B_{s_{\boldsymbol{x}}}(0) \subset \mathbb{R}^p$ with Lipschitz constant $L_{d_{\boldsymbol{X}}} \in \mathbb{R}_+$. Also the density $d_{\boldsymbol{X}} : \mathbb{R}^p \to \mathbb{R}$ of the regressors satisfies $c_{d_{\boldsymbol{X}}} \le d_{\boldsymbol{X}} \le C_{d_{\boldsymbol{X}}}$ on $B_{s_{\boldsymbol{x}}}(0) \subset \mathbb{R}^p$

for constants $0 < c_{d_{\boldsymbol{X}}} \leq C_{d_{\boldsymbol{X}}} < \infty$. Further we assume that for any $\boldsymbol{\theta} \perp \boldsymbol{\theta}^*$ with $\|\boldsymbol{\theta}\| = 1$ we have $\mathrm{Var}(\boldsymbol{X}^\top\boldsymbol{\theta}|\boldsymbol{X}^\top\boldsymbol{\theta}^*) > \sigma^2_{\boldsymbol{X}|\boldsymbol{\theta}^*}$ for some constant $\sigma^2_{\boldsymbol{X}|\boldsymbol{\theta}^*} > 0$ that does not depend on $\boldsymbol{X}^\top\boldsymbol{\theta}^* \in \mathbb{R}$.

($\mathbf{Cond}_f$) For some $\boldsymbol{\eta}^* \in l^2$

$$f = f_{\boldsymbol{\eta}^*} = \sum_{k=1}^{\infty} \eta^*_k \boldsymbol{e}_k,$$

where with some $\alpha > 2$ and a constant $C_{\|\boldsymbol{\eta}^*\|} > 0$

$$\sum_{l=0}^{\infty} l^{2\alpha} \eta^{*2}_l \leq C^2_{\|\boldsymbol{\eta}^*\|} < \infty.$$

($\mathbf{Cond}_{\boldsymbol{X}\boldsymbol{\theta}^*}$) It holds true that $\mathbb{P}(|f'_{\boldsymbol{\eta}^*}(\boldsymbol{X}^\top\boldsymbol{\theta}^*)| > c_{f'_{\boldsymbol{\eta}^*}}) > c_{\mathbb{P}f'}$ for some $c_{f'_{\boldsymbol{\eta}^*}}$, $c_{\mathbb{P}f'} > 0$.

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

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

($\mathbf{Cond}_\Upsilon$) $\Upsilon \subseteq \Upsilon_\circ(\sqrt{n}\mathtt{r}^\circ) \subset \mathbb{R}^{p+m}$ with $\mathtt{r}^\circ \in \mathbb{R}$, i.e. $d_\Upsilon \stackrel{\mathrm{def}}{=} diam(\Upsilon) < \infty$.

If these conditions denoted by $(\mathcal{A})$ are met we can proof the following results:

Proposition 2.11. *Assume* $(\mathcal{A})$ *with* $\alpha = 2 + \epsilon$ *for some* $\epsilon > 0$ *with* $p^{*5}/n \to 0$ *but* $p^{*5+2\epsilon}/n \to \infty$. *If* $n \in \mathbb{N}$ *is large enough it holds with probability greater* $1 - 4\mathrm{e}^{-\mathtt{x}} - \exp\{-m^3\} - \exp\{-nc_{(\boldsymbol{Q})}/4\}$

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

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

where $c_{(\boldsymbol{Q})} > 0$. *Further as* $n \to \infty$

$$\big\|\breve{D}\big(\widetilde{\boldsymbol{\theta}}_m - \boldsymbol{\theta}^*\big) - \breve{\boldsymbol{\xi}}_m(\boldsymbol{\upsilon}^*_m)\big\| \xrightarrow{\mathbb{P}} 0,$$

$$\breve{D}\big(\widetilde{\boldsymbol{\theta}}_m - \boldsymbol{\theta}^*\big) \xrightarrow{w} \mathcal{N}(0, \sigma^2\mathbb{I}_p),$$

$$2\breve{L}(\widetilde{\boldsymbol{\theta}}_m, \boldsymbol{\theta}^*) \xrightarrow{w} \chi^2_p.$$

Remark 2.31. The constant $\mathtt{C}_\diamondsuit > 0$ is a polynomial of $\|\psi\|_\infty$, $\|\psi'\|_\infty$, $\|\psi''\|_\infty$ and $s_{\boldsymbol{X}}$ that is independent of $\mathtt{x}$, n, p^*. The constant $c_{(\boldsymbol{Q})} > 0$ is related to $\mathtt{b}$ from $(\mathcal{L}\mathtt{r})$ and also does not depend on $\mathtt{x}$, n, p^*.

Remark 2.32. The necessary size of $n \in \mathbb{N}$ is determined by the size of $p^{*5/2}/\sqrt{n} \to 0$ and $m^{-2\alpha-1}n \to 0$. In the proof of Proposition 2.11 we impose conditions on $n \in \mathbb{N}$ of the kind

$$p^{*5/2}/\sqrt{n} \le \mathtt{C}_1^{-1}, \quad m^{-2\alpha-1}n \le \mathtt{C}_2^{-1},$$

for certain constants $\mathtt{C}_1, \mathtt{C}_2 > 0$ that are polynomials of $\|\psi\|_\infty$, $\|\psi'\|_\infty$, $\|\psi''\|_\infty$, $\mathtt{C}_{\|\boldsymbol{f}^*\|}$ and $s_{\boldsymbol{X}}$. These constants enter into the bound for $\mathtt{C}_\diamondsuit$.

For details see [1].

Appendix A: Deviation bounds for quadratic forms

The following general results from the supplement of [29] help to control the deviation for quadratic forms of type $\|B\boldsymbol{\xi}\|^2$ for a given positive matrix B – i.e. $BB^\top > 0$ – and a random vector $\boldsymbol{\xi}$. It will be used several times in our proofs. Suppose that

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

Remark A.1. In the setting of Section 2 we have either $\boldsymbol{\xi} = \breve{V}^{-1}\breve{\nabla}_{\boldsymbol{\theta}}\boldsymbol{\zeta}(\boldsymbol{\upsilon}^*)$ and $B = \breve{D}^{-1}\breve{V}$ or $\boldsymbol{\xi} = \mathcal{V}^{-1}\nabla\boldsymbol{\zeta}(\boldsymbol{\upsilon}^*)$ and $B = \mathcal{D}^{-1}\mathcal{V}$.

For a matrix B, define

$$\mathtt{p} = \operatorname{tr}(BB^\top), \qquad \mathtt{v}^2 = 2\operatorname{tr}(BB^\top BB^\top), \qquad \lambda_B \stackrel{\text{def}}{=} \|BB^\top\| \stackrel{\text{def}}{=} \lambda_{\max}(BB^\top).$$

For ease of presentation, suppose that $\mathtt{g}^2 \ge 2\mathtt{p}_B$. The other case only changes the constants in the inequalities. Define $\mu_c = 2/3$ and

$$\begin{aligned} \mathtt{g}_c &\stackrel{\text{def}}{=} \sqrt{\mathtt{g}^2 - \mu_c \mathtt{p}_B}, \\ 2(\mathtt{x}_c + 2) &\stackrel{\text{def}}{=} (\mathtt{g}^2/\mu_c - \mathtt{p}_B)/\lambda_B + \log\det\bigl(\mathbb{I}_p - \mu_c B/\lambda_B\bigr). \end{aligned} \tag{A.1}$$

Proposition A.1. *Let (ED_0) hold with $\nu_0 = 1$ and $\mathtt{g}^2 \ge 2\mathtt{p}_B$. Then for each $\mathtt{x} > 0$*

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

where $\mathfrak{z}(\mathtt{x}, BB^\top)$ is defined by

$$\mathfrak{z}^2(\mathtt{x}, BB^\top) \tag{A.2}$$
$$\stackrel{\text{def}}{=} \begin{cases} \mathtt{p}_B + 2\mathtt{v}_B(\mathtt{x}+1)^{1/2}, & \mathtt{x}+1 \le \mathtt{v}_B/(18\lambda_B), \\ \mathtt{p}_B + 6\lambda_B(\mathtt{x}+1), & \mathtt{v}_B/(18\lambda_B) < \mathtt{x}+1 \le \mathtt{x}_c+2, \\ \bigl|\mathtt{y}_c + 2\lambda_B(\mathtt{x}-\mathtt{x}_c+1)/\mathtt{g}_c\bigr|^2, & \mathtt{x} > \mathtt{x}_c+1, \end{cases}$$

with $\mathtt{y}_c^2 \le \mathtt{p}_B + 6\lambda_B(\mathtt{x}_c+2)$.

Depending on the value $\mathtt{x}$, we observe three types of tail behavior of the quadratic form $\|B\boldsymbol{\xi}\|^2$. The sub-Gaussian regime for $\mathtt{x}+1 \le \mathtt{v}_B/(18\lambda_B)$ and the Poissonian regime for $\mathtt{x} \le \mathtt{x}_c+1$ are similar to the case of a Gaussian quadratic form. The value $\mathtt{x}_c$ from (A.1) is of order $\mathtt{g}^2$. In all our results we suppose that $\mathtt{g}^2$ and hence, $\mathtt{x}_c$ is sufficiently large and the quadratic form $\|\boldsymbol{\xi}\|^2$ can be bounded with a dominating probability by $\mathtt{p}_B+6\lambda_B(\mathtt{x}+1)$ for a proper $\mathtt{x}$. We refer to the supplement of [29] for the proof of this and related results, further discussion and references.

Appendix B: Proofs

This section collects the proofs of the results in chronological order.

B.1. Proof of Lemma 2.1

Proof. Take any $\boldsymbol{\gamma} \in \mathbb{R}^p$ with $\|\boldsymbol{\gamma}\| = 1$ then

$$\begin{aligned}\boldsymbol{\gamma}^\top \breve{D}^{-1}\breve{\nabla}_{\boldsymbol{\theta}}\boldsymbol{\zeta}(\boldsymbol{\upsilon}) &= \boldsymbol{\gamma}^\top \begin{pmatrix} \breve{D}^{-1}D & \breve{D}^{-1}AH^{-1} \end{pmatrix} \begin{pmatrix} D^{-1} & 0 \\ 0 & H^{-1} \end{pmatrix} \nabla\boldsymbol{\zeta}(\boldsymbol{\upsilon}) \\ &\stackrel{\mathrm{def}}{=} \widehat{\boldsymbol{\gamma}}^\top \mathcal{D}^{-1}\nabla\boldsymbol{\zeta}(\boldsymbol{\upsilon}),\end{aligned}$$

where

$$\|\widehat{\boldsymbol{\gamma}}\| \le \left\| \begin{pmatrix} \breve{D}^{-1}D & \breve{D}^{-1}AH^{-1} \end{pmatrix} \right\| \left\| \begin{pmatrix} D^{-1} & 0 \\ 0 & H^{-1} \end{pmatrix} \mathcal{D} \right\| \le \frac{(1+\rho)\sqrt{1+\rho^2}}{\sqrt{1-\rho^2}}.$$

This gives that $(\boldsymbol{\mathcal{E}\mathcal{D}_1})$ implies $(\boldsymbol{\breve{\mathcal{E}}\mathcal{D}_1})$ and $(\boldsymbol{\mathcal{E}\mathcal{D}_0})$ implies $(\boldsymbol{\breve{\mathcal{E}}\mathcal{D}_0})$ with

$$\breve{\mathtt{g}} = \frac{\sqrt{1-\rho^2}}{(1+\rho)\sqrt{1+\rho^2}}\mathtt{g}, \quad \breve{\nu} = \frac{(1+\rho)\sqrt{1+\rho^2}}{\sqrt{1-\rho^2}}\nu.$$

Further for any $\boldsymbol{\upsilon} \in \Upsilon_\circ(\mathtt{r})$

$$\begin{aligned}\|I_p - D^{-1}D^2(\boldsymbol{\upsilon})D^{-1}\| &= \|D^{-1}(D^2 - D^2(\boldsymbol{\upsilon}))D^{-1}\| \\ &= \|D^{-1}\Pi_{\boldsymbol{\theta}}(\mathcal{D}^2 - \mathcal{D}^2(\boldsymbol{\upsilon}))\Pi_{\boldsymbol{\theta}}^\top D^{-1}\| \\ &= \|D^{-1}\Pi_{\boldsymbol{\theta}}\mathcal{D}(I_{p^*} - \mathcal{D}^{-1}\mathcal{D}^2(\boldsymbol{\upsilon})\mathcal{D}^{-1})\mathcal{D}\Pi_{\boldsymbol{\theta}}^\top D^{-1}\| \\ &\le \|D^{-1}\Pi_{\boldsymbol{\theta}}\mathcal{D}\|^2\|I_{p^*} - \mathcal{D}^{-1}\mathcal{D}^2(\boldsymbol{\upsilon})\mathcal{D}^{-1}\| = \delta(\mathtt{r}).\end{aligned}$$

Also

$$\begin{aligned}\|D^{-1}(A(\boldsymbol{\upsilon}) - A)H^{-1}\| &= \|D^{-1}\Pi_{\boldsymbol{\theta}}(\mathcal{D}^2(\boldsymbol{\upsilon}) - \mathcal{D}^2)\Pi_{\boldsymbol{\eta}}^\top H^{-1}\| \\ &= \|D^{-1}\Pi_{\boldsymbol{\theta}}\mathcal{D}(\mathcal{D}^{-1}\mathcal{D}^2(\boldsymbol{\upsilon})\mathcal{D}^{-1} -)\mathcal{D}\Pi_{\boldsymbol{\eta}}^\top H^{-1}\|\end{aligned}$$

$$\begin{aligned} &\leq \|D^{-1}\Pi_{\boldsymbol{\theta}}\mathcal{D}\|\|H^{-1}\Pi_{\boldsymbol{\eta}}\mathcal{D}\|\|I_{p^*} - \mathcal{D}^{-1}\mathcal{D}^2(\boldsymbol{\upsilon})\mathcal{D}^{-1}\| \\ &\leq \delta(\mathtt{r}). \end{aligned}$$

With the same arguments

$$\left\|D^{-1}AH^{-1}\left(I_m - H^{-1}H^2(\boldsymbol{\upsilon})H^{-1}\right)\right\| \leq \rho\delta(\mathtt{r}).$$

□

B.2. Proof of Theorem 2.2

Remember the semiparametric spread

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

For $\boldsymbol{\zeta}(\boldsymbol{\upsilon}) = \mathcal{L}(\boldsymbol{\upsilon}) - \mathbb{E}\mathcal{L}(\boldsymbol{\upsilon})$ define the semiparametric normalized stochastic gradient gap

$$\breve{\mathcal{Y}}(\boldsymbol{\upsilon}) = \breve{D}^{-1}\Big(\breve{\nabla}_{\boldsymbol{\theta}}\boldsymbol{\zeta}(\boldsymbol{\upsilon}) - \breve{\nabla}_{\boldsymbol{\theta}}\boldsymbol{\zeta}(\boldsymbol{\upsilon}^*)\Big). \tag{B.1}$$

Fix the radius $\mathtt{r}_0(\mathtt{x}) > 0$ that ensures $\mathbb{P}\{\widetilde{\boldsymbol{\upsilon}}, \widetilde{\boldsymbol{\upsilon}}_{\boldsymbol{\theta}^*} \in \Upsilon_\circ(\mathtt{r}_0)\} \geq 1 - \mathrm{e}^{-\mathtt{x}}$. Define $C(\mathtt{r}_0, \mathtt{x}) \subseteq \Omega$ as

$$C(\mathtt{r}_0,\mathtt{x}) \stackrel{\text{def}}{=} \left\{\widetilde{\boldsymbol{\upsilon}}, \widetilde{\boldsymbol{\upsilon}}_{\boldsymbol{\theta}^*} \in \Upsilon_\circ(\mathtt{r}_0)\right\} \cap \left\{\sup_{\boldsymbol{\upsilon}\in\Upsilon_\circ(4\mathtt{r}_0)} \|\breve{\mathcal{Y}}(\boldsymbol{\upsilon})\| \leq 6\nu_1\breve{\omega}_{\mathfrak{Z}}(\mathtt{x}, \mathbb{Q})4\mathtt{r}_0\right\}.$$

In the following we will derive statements that hold true on this set $C(\mathtt{r}_0, \mathtt{x}) \subseteq \Omega$ which is of probability greater $1 - 2\mathrm{e}^{-\mathtt{x}}$ because it follows right away from the definition of $\mathtt{r}_0 > 0$ that

$$\mathbb{P}\big\{\widetilde{\boldsymbol{\upsilon}}, \widetilde{\boldsymbol{\upsilon}}_{\boldsymbol{\theta}^*} \notin \Upsilon_\circ(\mathtt{r}_0)\big\} \leq \mathrm{e}^{-\mathtt{x}},$$

and by Theorem C.1 which is applicable because $(\breve{\mathcal{E}}\mathcal{D}_1)$ implies (C.1) with $\|\cdot\|_{\mathcal{Y}} = \|\mathcal{D}(\cdot)\|$

$$\mathbb{P}\left(\sup_{\boldsymbol{\upsilon}\in\Upsilon_\circ(\mathtt{r}_0)} \|\breve{\mathcal{Y}}(\boldsymbol{\upsilon})\| \leq 6\nu_1\breve{\omega}_{\mathfrak{Z}}(\mathtt{x}, 2p^* + 2p)\mathtt{r}_0\right) \geq 1 - \mathrm{e}^{-\mathtt{x}}.$$

B.2.1. Proof of claim on $C(\mathtt{r}_0, \mathtt{x}) \subseteq \Omega$

Before we prove the claim we prove the following useful lemma:

Lemma B.1. *Assume that the condition $(\breve{\mathcal{L}}_0)$ is fulfilled. Then*

$$\sup_{\boldsymbol{\upsilon}\in\Upsilon_\circ(\mathtt{r})} \left\|\breve{D}^{-1}\left(\breve{\nabla}\mathbb{E}\mathcal{L}(\boldsymbol{\upsilon}) - \breve{\nabla}\mathbb{E}\mathcal{L}(\boldsymbol{\upsilon}^*)\right) + \breve{D}(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\right\| \leq \frac{4}{(1-\rho^2)^2}\mathtt{r}\breve{\delta}(\mathtt{r}).$$

Proof. We have with Taylor expansion and some $\widehat{\upsilon} \in \Upsilon_\circ(\mathtt{r})$

$$\begin{aligned}\nabla \mathbb{E}\mathcal{L}(\upsilon) - \nabla \mathbb{E}\mathcal{L}(\upsilon^*) &= \nabla^2 \mathbb{E}\mathcal{L}(\widehat{\upsilon})(\upsilon - \upsilon^*) \\ &\stackrel{\mathrm{def}}{=} -\mathcal{D}^2(\widehat{\upsilon})(\upsilon - \upsilon^*) \\ &= -\begin{pmatrix} D^2(\widehat{\upsilon}) & A(\widehat{\upsilon}) \\ A^\top(\widehat{\upsilon}) & H^2(\widehat{\upsilon}) \end{pmatrix} (\upsilon - \upsilon^*).\end{aligned}$$

This gives

$$\begin{aligned}&-\breve{D}^{-1} \left(\breve{\nabla} \mathbb{E}\mathcal{L}(\upsilon) - \breve{\nabla} \mathbb{E}\mathcal{L}(\upsilon^*) \right) \\ &\quad = \breve{D}^{-1} \begin{pmatrix} D^2(\widehat{\upsilon}) - AH^{-2}A^\top(\widehat{\upsilon}) & A(\widehat{\upsilon}) - AH^{-2}H^2(\widehat{\upsilon}) \end{pmatrix} (\upsilon - \upsilon^*) \\ &\quad = \breve{D}^{-1} \left(D^2(\widehat{\upsilon}) - AH^{-2}A^\top(\widehat{\upsilon}) \right) \breve{D}^{-1} \breve{D} (\boldsymbol{\theta} - \boldsymbol{\theta}^*) \\ &\qquad + \left(\breve{D}^{-1} A(\widehat{\upsilon}) - \breve{D}^{-1} AH^{-2} H^2(\widehat{\upsilon}) \right) (\boldsymbol{\eta} - \boldsymbol{\eta}^*).\end{aligned}$$

We estimate separately using $(\breve{\mathcal{L}}_0)$ and $(\mathcal{I})$

$$\begin{aligned}&\| \breve{D}^{-1} \left(D^2(\widehat{\upsilon}) - AH^{-2}A^\top(\widehat{\upsilon}) \right) \breve{D}^{-1} - I_p \| \\ &\quad = \left\| \breve{D}^{-1} \left(D^2(\widehat{\upsilon}) - D^2 - \left\{ AH^{-2}(A^\top(\widehat{\upsilon}) - A^\top) \right\} \right) \breve{D}^{-1} \right\| \\ &\quad \le \| \breve{D}^{-1} D \|^2 \left(\| D^{-1} D^2(\widehat{\upsilon}) D^{-1} - I_p \| \right. \\ &\qquad \left. + \| D^{-1} A H^{-1} \| \| D^{-1} (A(\widehat{\upsilon}) - A) H^{-1} \| \right) \\ &\quad \le \frac{1+\rho}{1-\rho^2} \breve{\delta}_0(\mathtt{r}),\end{aligned}$$

and

$$\begin{aligned}&\left\| \left(\breve{D}^{-1} A(\widehat{\upsilon}) - \breve{D}^{-1} AH^{-2} H^2(\widehat{\upsilon}) \right) (\boldsymbol{\eta} - \boldsymbol{\eta}^*) \right\| \\ &\quad \le \left\| \breve{D}^{-1} A(\widehat{\upsilon}) H^{-1} - \breve{D}^{-1} AH^{-2} H^2(\widehat{\upsilon}) H^{-1} \right\| \| H(\boldsymbol{\eta} - \boldsymbol{\eta}^*) \| \\ &\quad \le \| \breve{D}^{-1} D \| \left\{ \| D^{-1} (A(\widehat{\upsilon}) - A) H^{-1} \| \right. \\ &\qquad \left. + \| D^{-1} A H^{-1} \left(I_m - H^{-1} H^2(\widehat{\upsilon}) H^{-1} \right) \| \right\} \| H(\boldsymbol{\eta} - \boldsymbol{\eta}^*) \| \\ &\quad \le \frac{2}{\sqrt{1-\rho^2}} \breve{\delta}_0(\mathtt{r}) \| H(\boldsymbol{\eta} - \boldsymbol{\eta}^*) \|.\end{aligned}$$

Further

$$\| \breve{D}(\boldsymbol{\theta} - \boldsymbol{\theta}^*) \| \vee \| H(\boldsymbol{\eta} - \boldsymbol{\eta}^*) \| \le \frac{1}{\sqrt{1-\rho^2}} \frac{1}{\sqrt{1-\rho^2}} \| \mathcal{D}(\upsilon - \upsilon^*) \| \le \frac{1}{1-\rho^2} \mathtt{r}.$$

Together this gives that

$$\begin{aligned}
&\sup_{\boldsymbol{\upsilon}\in\Upsilon_\circ(\mathtt{r})} \left\| \breve{D}^{-1} \left(\breve{\nabla}\mathbb{E}\mathcal{L}(\boldsymbol{\upsilon}) - \breve{\nabla}\mathbb{E}\mathcal{L}(\boldsymbol{\upsilon}^*) \right) + \breve{D}(\boldsymbol{\theta}-\boldsymbol{\theta}^*) \right\| \\
&\quad \le \left(\frac{1+\rho}{1-\rho^2} + \frac{2}{\sqrt{1-\rho^2}} \right) \frac{1}{1-\rho^2} \mathtt{r}\breve{\delta}(\mathtt{r}) \\
&\quad \le \frac{4}{(1-\rho^2)^2} \mathtt{r}\breve{\delta}(\mathtt{r}).
\end{aligned}$$

□

The next Lemma already completes the proof of (2.6) and (2.7) on $C(\mathtt{r}_0,\mathtt{x}) \subset \Omega$:

Lemma B.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* (2.6) *and* (2.7) *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

$$\begin{aligned}
\|\breve{D}(\widetilde{\boldsymbol{\theta}}-\boldsymbol{\theta}^*) - \breve{\boldsymbol{\xi}}\| &= \left\| \breve{D}^{-1} \left\{ \breve{\nabla}\mathcal{L}(\widetilde{\boldsymbol{\upsilon}}) - \breve{\nabla}\mathcal{L}(\boldsymbol{\upsilon}^*) \right\} + \breve{D}(\widetilde{\boldsymbol{\theta}}-\boldsymbol{\theta}^*) \right\| \\
&\le \left\| \breve{D}^{-1} \left(\breve{\nabla}\mathbb{E}\mathcal{L}(\widetilde{\boldsymbol{\upsilon}}) - \breve{\nabla}\mathbb{E}\mathcal{L}(\boldsymbol{\upsilon}^*) \right) + \breve{D}(\widetilde{\boldsymbol{\theta}}-\boldsymbol{\theta}^*) \right\| \\
&\quad + \left\| \breve{D}^{-1} \{ \breve{\nabla}_{\boldsymbol{\theta}}\boldsymbol{\zeta}(\widetilde{\boldsymbol{\upsilon}}) - \breve{\nabla}_{\boldsymbol{\theta}}\boldsymbol{\zeta}(\boldsymbol{\upsilon}^*) \} \right\|.
\end{aligned}$$

Note that by condition $(\breve{\mathcal{L}}_0)$ we get with Lemma B.1 as we assume that $\widetilde{\boldsymbol{\upsilon}} \in \Upsilon_\circ(\mathtt{r}_0)$

$$\left\| \breve{D}^{-1} \left(\breve{\nabla}\mathbb{E}\mathcal{L}(\widetilde{\boldsymbol{\upsilon}}) - \breve{\nabla}\mathbb{E}\mathcal{L}(\boldsymbol{\upsilon}^*) \right) + \breve{D}(\widetilde{\boldsymbol{\theta}}-\boldsymbol{\theta}^*) \right\| \le \frac{4}{(1-\rho^2)^2} \mathtt{r}_0\breve{\delta}(\mathtt{r}_0).$$

For the remainder we use that on $C(\mathtt{r}_0,\mathtt{x}) \subset \Omega$

$$\left\| \breve{D}^{-1} \{ \breve{\nabla}_{\boldsymbol{\theta}}\boldsymbol{\zeta}(\widetilde{\boldsymbol{\upsilon}}) - \breve{\nabla}_{\boldsymbol{\theta}}\boldsymbol{\zeta}(\boldsymbol{\upsilon}^*) \} \right\| \le \sup_{\boldsymbol{\upsilon}\in\Upsilon_\circ(4\mathtt{r}_0)} \|\breve{\mathcal{Y}}(\boldsymbol{\upsilon})\| \le 6\breve{\nu}_1\breve{\omega}_{\mathfrak{z}}(\mathtt{x},\mathbb{Q})4\mathtt{r}_0.$$

This gives (2.6) on $C(\mathtt{r}_0,\mathtt{x}) \subset \Omega$. For (2.7) we will first show that on $C(\mathtt{r}_0,\mathtt{x}) \subset \Omega$

$$\begin{aligned}
&\left| \breve{L}(\widetilde{\boldsymbol{\theta}}) - \breve{L}(\boldsymbol{\theta}^*) - \left(\breve{\nabla}\boldsymbol{\zeta}(\boldsymbol{\upsilon}^*)(\widetilde{\boldsymbol{\theta}}-\boldsymbol{\theta}^*) - \|\breve{D}(\widetilde{\boldsymbol{\theta}}-\boldsymbol{\theta}^*)\|^2/2 \right) \right| \qquad \text{(B.2)} \\
&\quad \le \left(\|\breve{D}^{-1}\breve{\nabla}\| + \breve{\Diamond}(\mathtt{r}_0,\mathtt{x}) \right) \breve{\Diamond}(\mathtt{r}_0,\mathtt{x}).
\end{aligned}$$

To show this we use some ideas of the proof of Theorem 1 of [22], that is we define

$$l : \mathbb{R}^p \times \Upsilon \to \mathbb{R}, \quad (\boldsymbol{\theta}_1,\boldsymbol{\theta}_2,\boldsymbol{\eta}) \mapsto \mathcal{L}(\boldsymbol{\theta}_1, \boldsymbol{\eta} + H^{-2}A^\top(\boldsymbol{\theta}_2-\boldsymbol{\theta}_1)). \qquad \text{(B.3)}$$

Note that

$$\begin{aligned}\nabla_{\boldsymbol{\theta}_1} l(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\eta}) &= \breve{\nabla}_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}_1, \boldsymbol{\eta} + H^{-2} A^\top (\boldsymbol{\theta}_2 - \boldsymbol{\theta}_1)),\\ \text{i.e. } \nabla_{\boldsymbol{\theta}_1} l(\boldsymbol{\theta}^*, \boldsymbol{\theta}^*, \boldsymbol{\eta}^*) &= \breve{\nabla} \boldsymbol{\zeta}(\boldsymbol{\upsilon}^*).\end{aligned}$$

Remark B.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 $\sum_{i=1}^n \widetilde{\psi}_{\mathbb{P}}(\boldsymbol{Y}_i)$, with $\widetilde{\psi}_{\mathbb{P}}$ the efficient influence function.

We can represent:

$$\breve{L}(\widetilde{\boldsymbol{\theta}}) - \breve{L}(\boldsymbol{\theta}^*) = l(\widetilde{\boldsymbol{\theta}}, \widetilde{\boldsymbol{\theta}}, \widetilde{\boldsymbol{\eta}}) - l(\boldsymbol{\theta}^*, \boldsymbol{\theta}^*, \widetilde{\boldsymbol{\eta}}_{\boldsymbol{\theta}^*}), \qquad \widetilde{\boldsymbol{\eta}}_{\boldsymbol{\theta}^*} \stackrel{\text{def}}{=} \Pi_{\boldsymbol{\eta}} \operatorname*{argmax}_{\substack{\boldsymbol{\upsilon} \in \Upsilon, \\ \Pi_{\boldsymbol{\theta}} \boldsymbol{\upsilon} = \boldsymbol{\theta}^*}} \mathcal{L}(\boldsymbol{\upsilon}).$$

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

$$\breve{\alpha}(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}^*) \le \left(\|\breve{D}^{-1} \breve{\nabla}\| + \breve{\Diamond}(\mathtt{r}_0, \mathtt{x}) \right) \breve{\Diamond}(\mathtt{r}_0, \mathtt{x}), \tag{B.4}$$

which gives the upper bound of (B.2). Note that $\breve{\alpha}(\boldsymbol{\theta}^*, \boldsymbol{\theta}^*) = 0$ such that we get with Taylor expansion

$$\breve{\alpha}(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}^*) \le \|\breve{D}(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\| \sup_{\boldsymbol{\theta} \in \Pi_{\boldsymbol{\theta}} \Upsilon_\circ(\mathtt{r}_0)} |\breve{D}^{-1} \nabla_{\boldsymbol{\theta}_1} \breve{\alpha}(\boldsymbol{\theta}, \boldsymbol{\theta}^*)|.$$

We find

$$\begin{aligned}\nabla_{\boldsymbol{\theta}_1} \breve{\alpha}(\boldsymbol{\theta}, \boldsymbol{\theta}^*) &= \nabla_{\boldsymbol{\theta}_1} l(\boldsymbol{\theta}, \widetilde{\boldsymbol{\theta}}, \widetilde{\boldsymbol{\eta}}) - \nabla_{\boldsymbol{\theta}_1} l(\boldsymbol{\theta}^*, \boldsymbol{\theta}^*, \boldsymbol{\eta}^*) + \breve{D}(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\\ &= \breve{\nabla} \boldsymbol{\zeta}(\boldsymbol{\upsilon}^\circ) - \breve{\nabla} \boldsymbol{\zeta}(\boldsymbol{\upsilon}^*) + \mathbb{E}\left[\breve{\nabla} \mathcal{L}(\boldsymbol{\upsilon}^\circ) - \breve{\nabla} \mathcal{L}(\boldsymbol{\upsilon}^*) \right] + \breve{D}(\boldsymbol{\theta} - \boldsymbol{\theta}^*),\end{aligned}$$

where

$$\begin{aligned}\boldsymbol{\upsilon}^\circ &\stackrel{\text{def}}{=} (\boldsymbol{\theta}, \widetilde{\boldsymbol{\eta}} + H^{-2} A^\top (\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta})),\\ \|\mathcal{D}(\boldsymbol{\upsilon}^\circ - \boldsymbol{\upsilon}^*)\| &\le \|D(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\| + \|H(\widetilde{\boldsymbol{\eta}} - \boldsymbol{\eta}^*)\| + \rho \|D(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta})\|\\ &\le 2(1 + \rho)\mathtt{r}_0 < 4\mathtt{r}_0.\end{aligned}$$

Using Lemma B.1 and the definition of $C(\mathtt{r}_0, \mathtt{x})$ we can bound

$$\sup_{\boldsymbol{\theta} \in \Pi_{\boldsymbol{\theta}} \Upsilon_\circ(\mathtt{r}_0)} |\check{D}^{-1} \nabla_{\boldsymbol{\theta}_1} \check{\alpha}(\boldsymbol{\theta}, \boldsymbol{\theta}^*)| \leq \check{\diamondsuit}(\mathtt{r}_0, \mathtt{x}).$$

Using (2.6) we find on $C(\mathtt{r}_0, \mathtt{x})$

$$\|\check{D}(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\| \leq \|\check{D}^{-1} \check{\nabla}\| + \check{\diamondsuit}(\mathtt{r}_0, \mathtt{x}).$$

This gives (B.4). Similarily we can bound from below:

$$\check{L}(\widetilde{\boldsymbol{\theta}}) - \check{L}(\boldsymbol{\theta}^*) \geq l(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}^*, \widetilde{\boldsymbol{\eta}}_{\boldsymbol{\theta}^*}) - l(\boldsymbol{\theta}^*, \boldsymbol{\theta}^*, \widetilde{\boldsymbol{\eta}}_{\boldsymbol{\theta}^*}),$$

and repeat the same arguments using that $\widetilde{\boldsymbol{\upsilon}}_{\boldsymbol{\theta}^*} \in \Upsilon_\circ(\mathtt{r}_0)$ on $C(\mathtt{r}_0, \mathtt{x}) \subset \Omega$ to obtain the lower bound of (B.2). Plugging (2.6) into (B.2) this gives

$$\begin{aligned} \left| 2\check{L}(\widetilde{\boldsymbol{\theta}}) - 2\check{L}(\boldsymbol{\theta}^*) - \|\check{D}^{-1} \check{\nabla} \boldsymbol{\zeta}(\boldsymbol{\upsilon}^*)\|^2 \right| &\leq 4 \left(\|\check{D}^{-1} \check{\nabla}\| + \check{\diamondsuit}(\mathtt{r}_0, \mathtt{x}) \right) \check{\diamondsuit}(\mathtt{r}_0, \mathtt{x}) \\ &\quad + \check{\diamondsuit}(\mathtt{r}_0, \mathtt{x})^2. \end{aligned}$$

□

B.3. Proof of Proposition 2.4

We start with an auxiliary result. Define the *parametric gradient gap*

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

Lemma B.3. *Assume that the condition* $(\mathcal{L}_0)$ *is fulfilled. Then for* $0 \leq \mathtt{r}$ *on the set*

$$\mathcal{M}(\mathtt{r}, \mathtt{x}) \stackrel{\text{def}}{=} \{\widetilde{\boldsymbol{\upsilon}}, \widetilde{\boldsymbol{\upsilon}}_{\boldsymbol{\theta}^*} \in \Upsilon_\circ(\mathtt{r})\} \cap \left\{ \sup_{\boldsymbol{\upsilon} \in \Upsilon_\circ(\mathtt{r})} \|\mathcal{Y}(\boldsymbol{\upsilon})\| \leq 6\nu_1 \omega_{\mathfrak{z}}(\mathtt{x}, 4p^*)\mathtt{r} \right\}, \quad \text{(B.5)}$$

we have

$$\begin{aligned} \|\mathcal{D}(\widetilde{\boldsymbol{\upsilon}} - \boldsymbol{\upsilon}^*) - \mathcal{D}^{-1} \nabla \mathcal{L}(\boldsymbol{\upsilon}^*)\| &\leq \diamondsuit(\mathtt{r}, \mathtt{x}), \\ \|H(\widetilde{\boldsymbol{\upsilon}} - \boldsymbol{\upsilon}^*) - H^{-1} \nabla_{\boldsymbol{\eta}} \mathcal{L}(\boldsymbol{\upsilon}^*)\| &\leq \diamondsuit(\mathtt{r}, \mathtt{x}). \end{aligned}$$

Proof. Since $\nabla \mathcal{L}(\widetilde{\boldsymbol{\upsilon}}) = 0$ we find with the triangular inequality

$$\begin{aligned} \|\mathcal{D}(\widetilde{\boldsymbol{\upsilon}} - \boldsymbol{\upsilon}^*) - \mathcal{D}^{-1} \nabla \mathcal{L}(\boldsymbol{\upsilon}^*)\| &\leq \|\mathcal{D}^{-1} \Big(\nabla \boldsymbol{\zeta}(\widetilde{\boldsymbol{\upsilon}}) - \nabla \boldsymbol{\zeta}(\boldsymbol{\upsilon}^*) \Big)\| \\ &\quad + \|\mathcal{D}^{-1} \mathbb{E} \nabla \mathcal{L}(\boldsymbol{\upsilon}) - \mathcal{D}^{-1} \mathbb{E} \nabla \mathcal{L}(\boldsymbol{\upsilon}^*) + \mathcal{D}\,(\widetilde{\boldsymbol{\upsilon}} - \boldsymbol{\upsilon}^*)\|. \end{aligned}$$

In section 2.1 we assume that $\mathcal{L} : \mathbb{R}^{p^*} \to \mathbb{R}$ is smooth enough such that we can interchange $\nabla \mathbb{E} \mathcal{L}(\boldsymbol{\upsilon}) = \mathbb{E} \nabla \mathcal{L}(\boldsymbol{\upsilon})$ on $\Upsilon_\circ(\mathtt{r}_0)$. This gives by condition $(\mathcal{L}_0)$ and

Taylor expansion

$$\sup_{\upsilon\in\Upsilon_\circ(\mathtt{r})} \|\mathcal{D}^{-1}\mathbb{E}\nabla\mathcal{L}(\upsilon) - \mathcal{D}^{-1}\mathbb{E}\nabla\mathcal{L}(\upsilon^*) + \mathcal{D}\,(\upsilon-\upsilon^*)\| \\ \leq \sup_{\upsilon\in\Upsilon_\circ(\mathtt{r})} \|\mathcal{D}^{-1}\nabla^2\mathbb{E}\mathcal{L}(\upsilon)\mathcal{D}^{-1} + \mathbb{I}_{p^*}\|\mathtt{r} \leq \delta(\mathtt{r})\mathtt{r}.$$

For the remainder we use the definition of $\mathcal{M}(\mathtt{r},\mathtt{x})$ in (B.5). This gives the the first claim. For the second claim we repeat the same arguments with the restriction to the set $\Upsilon_{\circ,\boldsymbol{\theta}^*}(\mathtt{r}) \stackrel{\text{def}}{=} \{(\boldsymbol{\theta},\boldsymbol{\eta}) \in \Upsilon_\circ(\mathtt{r}) : \boldsymbol{\theta} = \boldsymbol{\theta}^*\}$. We bound on $\Upsilon_{\circ,\boldsymbol{\theta}^*}(\mathtt{r})$

$$\|H^{-1}\{\nabla_{\boldsymbol{\eta}}\mathcal{L}(\upsilon) - \nabla_{\boldsymbol{\eta}}\mathcal{L}(\upsilon^*) + H^2\,(\boldsymbol{\eta}-\boldsymbol{\eta}^*)\}\| \leq \|H^{-1}\{\nabla_{\boldsymbol{\eta}}\mathcal{L}(\upsilon) - \nabla_{\boldsymbol{\eta}}\mathcal{L}(\upsilon^*)\}\| \\ + \|H^{-1}\{\nabla_{\boldsymbol{\eta}}\mathbb{E}\mathcal{L}(\upsilon) - \nabla_{\boldsymbol{\eta}}\mathbb{E}\mathcal{L}(\upsilon^*) + H^2\,(\boldsymbol{\eta}-\boldsymbol{\eta}^*)\}\|.$$

Take any $\boldsymbol{\gamma}\in\mathbb{R}^m$ with $\|\boldsymbol{\gamma}\|=1$ then

$$\boldsymbol{\gamma}^\top H^{-1}\nabla_{\boldsymbol{\eta}}\mathcal{L}(\upsilon) = (0, H^{-1}\boldsymbol{\gamma})^\top\nabla_\upsilon\mathcal{L}(\upsilon) = (0,H^{-1}\boldsymbol{\gamma})^\top\mathcal{D}\mathcal{D}^{-1}\nabla_\upsilon\mathcal{L}(\upsilon).$$

Now note that $\|\mathcal{D}(0,H^{-1}\boldsymbol{\gamma})\|^2 = \|\boldsymbol{\gamma}\|^2 = 1$ such that

$$\begin{aligned}\|H^{-1}\{\nabla_{\boldsymbol{\eta}}\boldsymbol{\zeta}(\upsilon) - \nabla_{\boldsymbol{\eta}}\boldsymbol{\zeta}(\upsilon^*)\}\| &= \sup_{\substack{\boldsymbol{\gamma}\in\mathbb{R}^m\\ \|\boldsymbol{\gamma}\|=1}} \boldsymbol{\gamma}^\top H^{-1}\{\nabla_{\boldsymbol{\eta}}\boldsymbol{\zeta}(\upsilon) - \nabla_{\boldsymbol{\eta}}\boldsymbol{\zeta}(\upsilon^*)\} \\ &\leq \sup_{\substack{\boldsymbol{\gamma}\in\mathbb{R}^{p^*}\\ \|\boldsymbol{\gamma}\|=1}} \boldsymbol{\gamma}^\top\mathcal{D}^{-1}\{\nabla_\upsilon\boldsymbol{\zeta}(\upsilon) - \nabla_\upsilon\boldsymbol{\zeta}(\upsilon^*)\} \\ &= \|\mathcal{D}^{-1}\{\nabla_\upsilon\boldsymbol{\zeta}(\upsilon) - \nabla_\upsilon\boldsymbol{\zeta}(\upsilon^*)\}\| \\ &\leq 6\nu_1\omega_{\mathfrak{z}}(\mathtt{x},4p^*)\mathtt{r}.\end{aligned}$$

As above we find with Taylor expansion

$$\sup_{\upsilon\in\Upsilon_{\circ,\boldsymbol{\theta}^*}(\mathtt{r})} \|H^{-1}\{\nabla_{\boldsymbol{\eta}}\mathbb{E}\mathcal{L}(\upsilon) - \nabla_{\boldsymbol{\eta}}\mathbb{E}\mathcal{L}(\upsilon^*) + H^2\,(\boldsymbol{\eta}-\boldsymbol{\eta}^*)\}\| \\ \leq \sup_{\upsilon\in\Upsilon_{\circ,\boldsymbol{\theta}^*}(\mathtt{r})} \|H^{-1}H^2(\upsilon)H^{-1} - \mathbb{I}_m\|\mathtt{r}.$$

We can bound using $\|\mathcal{D}(0,H^{-1}\boldsymbol{\gamma})\|^2 = \|\boldsymbol{\gamma}\|^2$ and $(\mathcal{L}_0)$

$$\begin{aligned}\|H^{-1}H^2(\upsilon)H^{-1} - \mathbb{I}_m\| &= \sup_{\substack{\boldsymbol{\gamma}\in\mathbb{R}^m\\ \|\boldsymbol{\gamma}\|=1}} \left(H^{-1}\boldsymbol{\gamma}\right)^\top\left\{H^2(\upsilon) - H^2\right\}H^{-1}\boldsymbol{\gamma} \\ &= \sup_{\substack{\boldsymbol{\gamma}\in\mathbb{R}^m\\ \|\boldsymbol{\gamma}\|=1}} \left(0,H^{-1}\boldsymbol{\gamma}\right)^\top\left\{\mathcal{D}^2(\upsilon) - \mathcal{D}^2\right\}\left(0,H^{-1}\boldsymbol{\gamma}\right)\end{aligned}$$

$$\leq \sup_{\substack{\gamma\in\mathbb{R}^p\\ \|\gamma\|=1}} \gamma\left\{\mathcal{D}^{-1}\mathcal{D}^2(\boldsymbol{\upsilon})\mathcal{D}^{-1}-\mathbb{I}_{p^*}\right\}\gamma \leq \delta(\mathtt{r}).$$

This gives the claim. □

Now we can proof Proposition 2.4. Define

$$\begin{aligned} C'(\mathtt{r}_0,\mathtt{x}) \stackrel{\text{def}}{=} \Big\{\ & \widetilde{\boldsymbol{\upsilon}},\widetilde{\boldsymbol{\upsilon}}_{\boldsymbol{\theta}^*}\in\Upsilon_\circ(\mathtt{r}_0), \\ & \|\mathcal{D}^{-1}\nabla\|\leq \mathfrak{z}(\mathtt{x},B),\ \|H^{-1}\nabla_{\boldsymbol{\eta}}\|\leq\mathfrak{z}(\mathtt{x},B),\ \Big\} \\ \cap & \left\{\sup_{\boldsymbol{\upsilon}\in\Upsilon_\circ(\mathtt{r}_0)}\|\mathcal{Y}(\boldsymbol{\upsilon})\|\leq 6\nu_1\omega\mathfrak{z}(\mathtt{x},4p^*)\mathtt{r}_0\right\} \\ \cap & \left\{\sup_{\boldsymbol{\upsilon}\in\Upsilon_\circ(4\mathtt{r}_1)}\|\breve{\mathcal{Y}}(\boldsymbol{\upsilon})\|\leq 6\nu_1\breve{\omega}\mathfrak{z}(\mathtt{x},2p^*+2p)4\mathtt{r}_1\right\}. \end{aligned} \tag{B.6}$$

The desired result occurs on this set. First we show that $\mathbb{P}(C'(\mathtt{r}_0,\mathtt{x}))\geq 1-5\mathrm{e}^{-\mathtt{x}}$. Lemma B.4 yields

$$\|H^{-1}\nabla_{\boldsymbol{\eta}}\|^2\leq\|\mathcal{D}^{-1}\nabla\|^2,$$

which implies that

$$\{\|\mathcal{D}^{-1}\nabla\|\leq\mathfrak{z}(\mathtt{x},B)\}\subseteq\{\|H^{-1}\nabla_{\boldsymbol{\eta}}\|\leq\mathfrak{z}(\mathtt{x},B)\}.$$

To control the probability $\mathbb{P}\big(\|\mathcal{D}^{-1}\nabla\|>\mathfrak{z}(\mathtt{x},B)\big)$ we apply Proposition A.1 with

$$B=\mathcal{D}^{-1}\mathcal{V}^2\mathcal{D}^{-1}.$$

We obtain

$$\mathbb{P}\left(\|\mathcal{D}^{-1}\nabla\|>\mathfrak{z}(\mathtt{x},B)\right)\leq 2\mathrm{e}^{-\mathtt{x}}.$$

By Theorem C.1 with $p=p^*$ we have

$$\mathbb{P}\left\{\sup_{\boldsymbol{\upsilon}\in\Upsilon_\circ(\mathtt{r}_0)}\|\mathcal{Y}(\boldsymbol{\upsilon})\|\leq 6\nu_1\omega\mathfrak{z}(\mathtt{x},4p^*)\mathtt{r}_0\right\}\geq 1-\mathrm{e}^{-\mathtt{x}}.$$

This gives that $\mathbb{P}(C'(\mathtt{r}_0,\mathtt{x}))\geq 1-5\mathrm{e}^{-\mathtt{x}}$. Lemma B.3 gives that on the set $C'(\mathtt{r}_0,\mathtt{x})$ from (B.6) we have

$$\|\mathcal{D}(\widetilde{\boldsymbol{\upsilon}}-\boldsymbol{\upsilon}^*)-\mathcal{D}^{-1}\nabla\|\leq\diamondsuit(\mathtt{r}_0,\mathtt{x}).$$

With the triangular inequality this gives

$$\|\mathcal{D}(\widetilde{\boldsymbol{\upsilon}}-\boldsymbol{\upsilon}^*)\|\leq\|\mathcal{D}^{-1}\nabla\|+\diamondsuit(\mathtt{r}_0,\mathtt{x}).$$

Now on $C'(\mathtt{r}_0, \mathtt{x})$ we have $\|\mathcal{D}^{-1}\nabla\| \le \mathfrak{z}(\mathtt{x}, B)$, which implies

$$\|\mathcal{D}(\widetilde{\boldsymbol{\upsilon}} - \boldsymbol{\upsilon}^*)\| \le \mathfrak{z}(\mathtt{x}, B) + \diamondsuit(\mathtt{r}_0, \mathtt{x}).$$

The same can be done for $\|\mathcal{D}(\widetilde{\boldsymbol{\upsilon}}_{\boldsymbol{\theta}^*} - \boldsymbol{\upsilon}^*)\|$ which gives

$$\begin{aligned} C'(\mathtt{r}_0, \mathtt{x}) &\subseteq \{\widetilde{\boldsymbol{\upsilon}}, \widetilde{\boldsymbol{\upsilon}}_{\boldsymbol{\theta}^*} \in \Upsilon_\circ(\mathtt{r}_1)\} \cap \Big\{ \sup_{\boldsymbol{\upsilon} \in \Upsilon_\circ(\mathtt{r}_1)} \|\breve{\mathtt{y}}(\boldsymbol{\upsilon})\| \le 6\nu_1 \breve{\omega}\mathfrak{z}(\mathtt{x}, p^*)\mathtt{r}_1 \Big\} \\ &= C(\mathtt{r}_1, \mathtt{x}) \subset \Omega. \end{aligned}$$

Now the claim follows as in the proof of Theorem 2.2 with $\mathtt{r}_0 > 0$ replaced with $\mathtt{r}_1 > 0$.

Lemma B.4. *Let $\mathcal{D} \in \mathbb{R}^{(p+p)\times(p+p)}$ be invertible and*

$$\mathcal{D}^2 = \begin{pmatrix} D^2 & A \\ A^\top & H^2 \end{pmatrix} \in \mathbb{R}^{(p+p)\times(p+p)}, \quad D \in \mathbb{R}^{p\times p},\ H \in \mathbb{R}^{m\times m} \text{ invertible.}$$

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 $\boldsymbol{\upsilon} = (\boldsymbol{\theta}, \boldsymbol{\eta}) \in \mathbb{R}^{p+m}$

$$\|D^{-1}\boldsymbol{\theta}\| = \|D^{-1}\Pi_{\boldsymbol{\theta}}\mathcal{D}\mathcal{D}^{-1}\boldsymbol{\upsilon}\| \le \|D^{-1}\Pi_{\boldsymbol{\theta}}\mathcal{D}\|\|\mathcal{D}^{-1}\boldsymbol{\upsilon}\| = \|\mathcal{D}^{-1}\boldsymbol{\upsilon}\|,$$

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 $\|H^{-1}\boldsymbol{\eta}\|$. □

B.4. Proof of Lemma 2.5

Proof. First note that due to (ι) we have

$$\|\mathcal{D}^{-1}\| = \frac{1}{\sqrt{n}}\|d^{-1}\| \le \frac{1}{\sqrt{n}c_d}. \tag{B.7}$$

Now we prove the implications.

$(\mathcal{L}_0)$ As by assumption $\Upsilon_\circ(\mathtt{r}^*) \subset U$ we simply estimate using (B.7) and (ℓ_0) for any $\boldsymbol{\upsilon} \in \Upsilon_\circ(\mathtt{r}^*)$

$$\begin{aligned} \|\mathbb{I} - \mathcal{D}^{-1}\nabla^2\mathbb{E}\mathcal{L}(\boldsymbol{\upsilon})\mathcal{D}^{-1}\| &\le \frac{1}{nc_d^2}\|\mathcal{D}^2 - \nabla^2\mathbb{E}\mathcal{L}(\boldsymbol{\upsilon})\| \\ &= \frac{1}{c_d^2}\|\nabla^2\mathbb{E}\ell(\boldsymbol{\upsilon}^*) - \nabla^2\mathbb{E}\ell(\boldsymbol{\upsilon})\| \le \frac{\delta^*}{\sqrt{n}c_d^3}\mathtt{r}. \end{aligned}$$

($\mathcal{ED}_1$) Abbreviate $\zeta_i = (\ell_i - \mathbb{E}\ell_i)$ and $\boldsymbol{\zeta} = (\mathcal{L} - \mathbb{E}\mathcal{L})$. Take any $\boldsymbol{\gamma} \in \mathbb{R}^{p^*}$ and $\boldsymbol{\upsilon}, \boldsymbol{\upsilon}' \in \Upsilon_\circ(\mathfrak{r}^*) \subset U$ and use the mean value theorem to find some $\widehat{\boldsymbol{\upsilon}} \in \operatorname{conv}(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}') \subset U$

$$\log \mathbb{E} \exp \left\{ \frac{\mu \boldsymbol{\gamma}^\top \mathcal{D}^{-1} \{\nabla \boldsymbol{\zeta}(\boldsymbol{\upsilon}) - \nabla \boldsymbol{\zeta}(\boldsymbol{\upsilon}')\}}{\omega \, \|\mathcal{D}(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}')\|} \right\}$$
$$= \log \mathbb{E} \exp \left\{ \frac{\mu}{\omega n} \boldsymbol{\gamma}^\top d^{-1} \left\{ \sum_{i=1}^n \nabla^2 \zeta_i(\widehat{\boldsymbol{\upsilon}}) \right\} d^{-1} \frac{d(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}')}{\|d(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}')\|} \right\}.$$

Using independence and (ed_1) this gives with $\omega = \frac{1}{\sqrt{n}}$ and $|\mu| \le \sqrt{n} \mathfrak{g}_0^*$

$$\log \mathbb{E} \exp \left\{ \frac{\mu \boldsymbol{\gamma}^\top \mathcal{D}^{-1} \{\nabla \boldsymbol{\zeta}(\boldsymbol{\upsilon}) - \nabla \boldsymbol{\zeta}(\boldsymbol{\upsilon}')\}}{\omega \, \|\mathcal{D}(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}')\|} \right\}$$
$$\le \sum_{i=1}^n \sup_{\substack{\boldsymbol{\gamma} \mathbb{R}^{p^*} \\ \|\boldsymbol{\gamma}\| = 1}} \log \mathbb{E} \exp \left\{ \frac{\mu}{\sqrt{n}} \boldsymbol{\gamma}_1^\top d^{-1} \nabla^2 \zeta_i(\widehat{\boldsymbol{\upsilon}}) d^{-1} \boldsymbol{\gamma}_2 \right\} \le \nu_0^* \mu^2 / 2.$$

Further ($\mathcal{I}$) is a consequence of Lemma B.5 and (ι). The other claims can be shown with the same argument or follow trivially from the setting. □

Lemma B.5. *For a positive definite symmetric matrix*

$$\mathcal{D}^2 = \begin{pmatrix} D^2 & A \\ A^\top & H^2 \end{pmatrix},$$

with $c_{\mathcal{D}} \|\boldsymbol{\upsilon}\|^2 \le \boldsymbol{\upsilon}^\top \mathcal{D} \upsilon$ for some $c_{\mathcal{D}} > 0$ we have that

$$\|D^{-1} A H^{-2} A^\top D^{-1}\| =: \rho^2 \le 1 - \frac{c_{\mathcal{D}}}{\|D\|^2 \wedge \|H\|^2}.$$

Proof. For any $v = (\boldsymbol{\theta}, \boldsymbol{\eta}) \in \mathbb{R}^{p+m}$ we have

$$\begin{aligned} \boldsymbol{\upsilon}^\top \mathcal{D}^2 \boldsymbol{\upsilon} &= (\boldsymbol{\theta}^\top, \boldsymbol{\eta}^\top) \begin{pmatrix} D^2 & A \\ A^\top & H^2 \end{pmatrix} \begin{pmatrix} \boldsymbol{\theta} \\ \boldsymbol{\eta} \end{pmatrix} \\ &= (\boldsymbol{\theta}^\top D^\top, \boldsymbol{\eta}^\top H^\top) \begin{pmatrix} I_p & D^{-1} A H^{-1} \\ H^{-1} A^\top D^{-1} & I_m \end{pmatrix} \begin{pmatrix} D\boldsymbol{\theta} \\ H\boldsymbol{\eta} \end{pmatrix} \\ &= \|D\boldsymbol{\theta}\|^2 + \|H\boldsymbol{\eta}\|^2 + 2\langle H\boldsymbol{\eta}, H^{-1} A^\top D^{-1} \boldsymbol{\theta} \rangle. \end{aligned}$$

Minimized with respect to $\boldsymbol{\eta}$, i.e. with $H\boldsymbol{\eta} = -H^{-1} A^\top D^{-1} D\boldsymbol{\theta}$ we find

$$\boldsymbol{\upsilon}^\top \mathcal{D}^2 \boldsymbol{\upsilon} = \|D\boldsymbol{\theta}\|^2 - \|H^{-1} A^\top D^{-1} D\boldsymbol{\theta}\|^2 = (D\boldsymbol{\theta})^\top (I_p - D^{-1} A H^{-2} A^\top D^{-1}) D\boldsymbol{\theta},$$

which gets minimal – i.e. equal to $(1-\rho^2)\|D\boldsymbol{\theta}\|$ - if

$$D^{-1}AH^{-2}A^\top D^{-1}D\boldsymbol{\theta} = \|D^{-1}AH^{-2}A^\top D^{-1}\|D\boldsymbol{\theta} = \rho^2 D\boldsymbol{\theta},$$

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 $c_{\mathcal{D}}\|\boldsymbol{\upsilon}\|^2 \leq \boldsymbol{\upsilon}^\top \mathcal{D}\boldsymbol{\upsilon}$ this gives

$$c_{\mathcal{D}}\|\boldsymbol{\upsilon}\|^2 \leq \boldsymbol{\upsilon}^\top \mathcal{D}^2\boldsymbol{\upsilon} = (1-\rho^2)\|D\boldsymbol{\theta}\|^2, \quad \|\boldsymbol{\upsilon}\|^2 = \|\boldsymbol{\theta}\|^2 + \|H^{-2}A^\top\boldsymbol{\theta}\|^2,$$

such that

$$\rho^2 \leq 1 - c_{\mathcal{D}}\frac{\|\boldsymbol{\theta}\|^2}{\|D\boldsymbol{\theta}\|^2} \leq 1 - \frac{c_{\mathcal{D}}}{\|D\|^2}.$$

With analogous arguments we can obtain

$$\rho^2 \leq 1 - c_{\mathcal{D}}\frac{\|\boldsymbol{\eta}\|^2}{\|H\boldsymbol{\eta}\|^2} \leq 1 - \frac{c_{\mathcal{D}}}{\|H\|^2},$$

which completes the proof. □

B.5. Proof of Proposition 2.7

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 2.1 are satisfied with $\mathcal{D}^2 = n\mathbb{E}[\nabla f \nabla f^\top(\boldsymbol{\upsilon}^*)] = Id_{p^*}$, $\breve{D}^2 = n$ 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

$$\begin{aligned}
-\max_{\boldsymbol{\eta}} \mathcal{L}(\boldsymbol{\theta}^*, \boldsymbol{\eta}) &= n\min_{\lambda\in\mathbb{R}}\left\{\left(y_{\boldsymbol{\theta}} - \lambda^2\|\boldsymbol{Y}_{\boldsymbol{\eta}}\|^2\right)^2 + (1-\lambda)^2\|\boldsymbol{Y}_{\boldsymbol{\eta}}\|^2\right\} \\
&= n\min_{\lambda\in\mathbb{R}}\left\{\left(\varepsilon_{\boldsymbol{\theta}} - \lambda^2\|\boldsymbol{\varepsilon}_{\boldsymbol{\eta}}\|^2\right)^2 + (1-\lambda)^2\|\boldsymbol{\varepsilon}_{\boldsymbol{\eta}}\|^2\right\} \\
&= n\min_{\lambda\in\mathbb{R}}\left\{\varepsilon^2_{\boldsymbol{\theta}} + \|\boldsymbol{\varepsilon}_{\boldsymbol{\eta}}\|^2\left(\lambda^4\|\boldsymbol{\varepsilon}_{\boldsymbol{\eta}}\|^2 - \lambda^2\varepsilon_{\boldsymbol{\theta}} + (1-\lambda)^2\right)\right\},
\end{aligned}$$

where $\boldsymbol{Y} = (y_{\boldsymbol{\theta}}, \boldsymbol{Y}_{\boldsymbol{\eta}}) \in \mathbb{R} \times \mathbb{R}^{p_n - 1}$ and $\boldsymbol{\varepsilon} = (\varepsilon_{\boldsymbol{\theta}}, \boldsymbol{\varepsilon}_{\boldsymbol{\eta}}) \in \mathbb{R} \times \mathbb{R}^{p_n - 1}$. Now 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 $n \geq n_\tau \in \mathbb{N}$ large enough

$$-\max_{\boldsymbol{\eta}} \mathcal{L}(\boldsymbol{\theta}^*, \boldsymbol{\eta}) \geq n\varepsilon_{\boldsymbol{\theta}}^2 + (1-\tau)n \|\boldsymbol{\varepsilon}_{\boldsymbol{\eta}}\|^4 - (1+\tau)n\varepsilon_{\boldsymbol{\theta}} \|\boldsymbol{\varepsilon}_{\boldsymbol{\eta}}\|^2 . \tag{B.8}$$

Further we get setting $\lambda = 1$

$$-\max_{\boldsymbol{\eta}} \mathcal{L}(\boldsymbol{\theta}^*, \boldsymbol{\eta}) \leq n\varepsilon_{\boldsymbol{\theta}}^2 + n \|\boldsymbol{\varepsilon}_{\boldsymbol{\eta}}\|^4 - n\varepsilon_{\boldsymbol{\theta}} \|\boldsymbol{\varepsilon}_{\boldsymbol{\eta}}\|^2 . \tag{B.9}$$

As $\breve{L}(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}^*) = -\max_{\boldsymbol{\eta}} \mathcal{L}(\boldsymbol{\theta}^*, \boldsymbol{\eta})$ the inequalities (B.8) and (B.9) combine to

$$\begin{aligned} n\varepsilon_{\boldsymbol{\theta}}^2 + n(1-\tau) \|\boldsymbol{\varepsilon}_{\boldsymbol{\eta}}\|^4 - (1+\tau)n\varepsilon_{\boldsymbol{\theta}} \|\boldsymbol{\varepsilon}_{\boldsymbol{\eta}}\|^2 &\leq \breve{L}(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}^*) \\ &\leq n\varepsilon_{\boldsymbol{\theta}}^2 + n \|\boldsymbol{\varepsilon}_{\boldsymbol{\eta}}\|^4 - n\varepsilon_{\boldsymbol{\theta}} \|\boldsymbol{\varepsilon}_{\boldsymbol{\eta}}\|^2 . \end{aligned}$$

This gives the Wilks phenomenon if $p_n^2/n \to 0$. Now if $p_n^2/n \to \infty$ the right hand side in (B.8) diverges since with $\tau = 1/2$

$$\begin{aligned} n\varepsilon_{\boldsymbol{\theta}}^2 + \frac{n}{2} \|\boldsymbol{\varepsilon}_{\boldsymbol{\eta}}\|^4 - 2n\varepsilon_{\boldsymbol{\theta}} \|\boldsymbol{\varepsilon}_{\boldsymbol{\eta}}\|^2 &\sim \chi_1^2 * (\chi_{p_n-1}^4/2n) * \left\{-\mathcal{N}(0,1)(2\chi_{p_n-1}^2/\sqrt{n})\right\} \\ &\xrightarrow{w} \delta_\infty . \end{aligned}$$

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.

B.6. Proof of Theorem 2.9

Remember the definition

$$\widetilde{\boldsymbol{\upsilon}}_{\boldsymbol{\theta}^*_m, m} = (\boldsymbol{\theta}^*_m, \widetilde{\boldsymbol{\eta}}_{\boldsymbol{\theta}^*_m}) \stackrel{\text{def}}{=} \operatorname*{argmax}_{\substack{\boldsymbol{\upsilon} \in \Upsilon \\ \Pi_{\boldsymbol{\theta}} \boldsymbol{\upsilon} = \boldsymbol{\theta}^*_m}} \mathcal{L}_m(\boldsymbol{\upsilon}), \quad \widetilde{\boldsymbol{\upsilon}}_{\boldsymbol{\theta}^*, m} = (\boldsymbol{\theta}^*_m, \widetilde{\boldsymbol{\eta}}_{\boldsymbol{\theta}^*}) \stackrel{\text{def}}{=} \operatorname*{argmax}_{\substack{\boldsymbol{\upsilon} \in \Upsilon \\ \Pi_{\boldsymbol{\theta}} \boldsymbol{\upsilon} = \boldsymbol{\theta}^*}} \mathcal{L}_m(\boldsymbol{\upsilon}).$$

Define for some $0 < \mathtt{r}_0^\circ$

$$\begin{aligned} \mathcal{A}(\mathtt{x}, \mathtt{r}_0^\circ) \stackrel{\text{def}}{=} &\left\{\widetilde{\boldsymbol{\upsilon}}_m, \widetilde{\boldsymbol{\upsilon}}_{\boldsymbol{\theta}^*_m, m}, \widetilde{\boldsymbol{\upsilon}}_{\boldsymbol{\theta}^*, m} \in \Upsilon_{0,m}(\mathtt{r}_0^\circ)\right\} \\ &\cap \left\{ \sup_{\boldsymbol{\upsilon} \in \Upsilon_\circ(4\mathtt{r}_0^\circ)} \|\breve{\mathcal{Y}}(\boldsymbol{\upsilon})\| \leq 6\nu_1 \breve{\omega}_{\mathfrak{z}}(\mathtt{x}, 2p^* + 2p) 4\mathtt{r}_0^\circ \right\} \subset \Omega, \end{aligned}$$

with $\breve{\mathcal{Y}}(\boldsymbol{\upsilon}) \in \mathbb{R}^{p^*}$ from (B.1).

We prove this claim in a similar fashion as in Section B.2.1. With the function $l_m : \mathbb{R}^p \times \Upsilon \to \mathbb{R}$ defined as in (B.3) with $\mathcal{L}$ replaced by $\mathcal{L}_m$ we can represent:

$$\breve{L}_m(\boldsymbol{\theta}^*_m) - \breve{L}_m(\boldsymbol{\theta}^*) = l_m(\boldsymbol{\theta}^*_m, \boldsymbol{\theta}^*_m, \widetilde{\boldsymbol{\eta}}_{\boldsymbol{\theta}^*_m}) - l_m(\boldsymbol{\theta}^*, \boldsymbol{\theta}^*, \widetilde{\boldsymbol{\eta}}_{\boldsymbol{\theta}^*}),$$

Repeating the same arguments as in Section B.2.1 we obtain

$$\begin{aligned}\check{L}_m(\boldsymbol{\theta}^*_m) - \check{L}_m(\boldsymbol{\theta}^*) &\le l_m(\boldsymbol{\theta}^*_m, \boldsymbol{\theta}^*_m, \widetilde{\boldsymbol{\eta}}_{\boldsymbol{\theta}^*_m}) - l_m(\boldsymbol{\theta}^*, \boldsymbol{\theta}^*_m, \widetilde{\boldsymbol{\eta}}_{\boldsymbol{\theta}^*_m}) \\ &= \check{\nabla}_{\boldsymbol{\theta}} \mathcal{L}_m(\boldsymbol{\upsilon}^*)(\boldsymbol{\theta}^*_m - \boldsymbol{\theta}^*) - \|\check{D}_m(\boldsymbol{\theta}^*_m - \boldsymbol{\theta}^*)\|^2/2 \\ &\quad + \check{\alpha}^*_m(\boldsymbol{\theta}^*_m, \boldsymbol{\theta}^*),\end{aligned}$$

where $\check{\alpha}^*_m(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) \in \mathbb{R}$ is defined as

$$\begin{aligned}\check{\alpha}^*_m(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) &\stackrel{\mathrm{def}}{=} l(\boldsymbol{\theta}_1, \boldsymbol{\theta}^*_m, \widetilde{\boldsymbol{\eta}}_{\boldsymbol{\theta}^*_m}) - l(\boldsymbol{\theta}_2, \boldsymbol{\theta}^*_m, \widetilde{\boldsymbol{\eta}}_{\boldsymbol{\theta}^*_m}) \\ &\quad - \nabla_{\boldsymbol{\theta}_1} l(\boldsymbol{\theta}^*, \boldsymbol{\theta}^*, \boldsymbol{\eta}^*)(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2) - \|\check{D}(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2)\|^2/2.\end{aligned}$$

and satisfies

$$\begin{aligned}\check{\alpha}^*_m(\boldsymbol{\theta}^*_m, \boldsymbol{\theta}^*) &\le \|\check{D}_m(\boldsymbol{\theta}^*_m - \boldsymbol{\theta}^*)\| \sup_{\boldsymbol{\theta} \in \Pi_{\boldsymbol{\theta}} \Upsilon_\circ(4\mathtt{r}^\circ_0)} |\check{D}_m^{-1} \nabla_{\boldsymbol{\theta}_1} \check{\alpha}_m(\boldsymbol{\theta}, \boldsymbol{\theta}^*)| \\ &\le \alpha(m) \check{\diamondsuit}(2(1+\rho)4\mathtt{r}^\circ_0, \mathtt{x}),\end{aligned}$$

since $\mathcal{A}(\mathtt{x}, \mathtt{r}^\circ_0) \subseteq \{\widetilde{\boldsymbol{\upsilon}}_{\boldsymbol{\theta}^*_m}, \widetilde{\boldsymbol{\upsilon}}_{\boldsymbol{\theta}^*} \in \Upsilon_\circ(\mathtt{r}^\circ_0)\}$. With similar arguments for the lower bound this gives

$$2|\check{L}_m(\boldsymbol{\theta}^*_m) - \check{L}_m(\boldsymbol{\theta}^*)| \le \alpha(m) \left(2\|\check{D}^{-1} \check{\nabla} \mathcal{L}_m(\boldsymbol{\upsilon}^*)\| + \alpha(m) + 2\check{\diamondsuit}(\mathtt{r}^\circ_0, \mathtt{x})\right).$$

The claim follows because the result (2.7) of Theorem 2.2 occurs on $\mathcal{A}(\mathtt{x}, \mathtt{r}^\circ_0) \subseteq C(\mathtt{x}, \mathtt{r}^\circ_0) \subset \Omega$. It remains to note that the set $\mathcal{A}(\mathtt{x}, \mathtt{r}^\circ_0) \subset \Omega$ is of probability greater $1 - 2\mathrm{e}^{-\mathtt{x}}$ by the choice of $\mathtt{r}^\circ_0 > 0$.

B.7. Proof of Corollary 2.10

We will only prove the asymptotic normality as the the proof the Wilks phenomenon is very similar. Define

$$\mathcal{V}^2_m(\boldsymbol{\upsilon}^*_m) = \mathrm{Cov}\big(\nabla_{p+m} \mathcal{L}_m(\boldsymbol{\upsilon}^*_m)\big),\ B_m = \mathcal{D}^{-1}_m \mathcal{V}^2_m \mathcal{D}^{-1}_m,$$

$$\check{\nabla}_{\boldsymbol{\theta},m} = \nabla_{\boldsymbol{\theta}} - A_m H_m^{-2} \nabla_{\boldsymbol{\eta}},\ \check{V}^2_m = \mathrm{Cov}(\check{\nabla}_{\boldsymbol{\theta}} \boldsymbol{\zeta}(\boldsymbol{\upsilon}^*_m)),\ \check{B}_m = \check{D}^{-1}_m \check{V}^2_m \check{D}^{-1}_m.$$

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 log-likelihood for the sieved functional models $\mathcal{L}_m$ and the operators $\mathcal{D}^2_m \in \mathbb{R}^{p^* \times p^*}$, $\check{D}^2_m \in \mathbb{R}^{p \times p}$ correspond to this point, i.e. we abbreviate $\mathcal{D}^2_m \stackrel{\mathrm{def}}{=} \mathcal{D}^2_m(\boldsymbol{\upsilon}^*_m)$, while $\mathcal{D}^2 = \mathcal{D}^2(\boldsymbol{\upsilon}^*)$ and $\check{D}^2_m = \check{D}^2_m(\boldsymbol{\upsilon}^*_m)$, $\check{D}^2 = \check{D}^2(\boldsymbol{\upsilon}^*)$, where $\boldsymbol{\upsilon}^* = \mathrm{argmax}_{\boldsymbol{\upsilon} \in \Upsilon} \mathbb{E}\mathcal{L}(\boldsymbol{\upsilon})$, i.e. the true full maximizer.

We get with Theorem 2.2 applied to $\widetilde{\boldsymbol{\theta}}_m$ from (2.15) that with probability greater $1 - 2\mathrm{e}^{-\mathtt{x}}$

$$\|\check{D}_m\big(\widetilde{\boldsymbol{\theta}}_m - \boldsymbol{\theta}^*_m\big) - \check{\boldsymbol{\xi}}_m(\boldsymbol{\upsilon}^*_m)\| \le \diamondsuit(\mathtt{r}_0, \mathtt{x}). \tag{B.10}$$

We write

$$\begin{aligned}&\breve{D}\big(\widetilde{\boldsymbol{\theta}}_m-\boldsymbol{\theta}^*\big)-\breve{\boldsymbol{\xi}}_m(\boldsymbol{\upsilon}_m^*)\\&\quad=\breve{D}_m\big(\widetilde{\boldsymbol{\theta}}_m-\boldsymbol{\theta}_m^*\big)-\breve{\boldsymbol{\xi}}_m(\boldsymbol{\upsilon}_m^*)+(\breve{D}_m-\breve{D})\big(\widetilde{\boldsymbol{\theta}}_m-\boldsymbol{\theta}_m^*\big)+\breve{D}_m\big(\boldsymbol{\theta}_m^*-\boldsymbol{\theta}^*\big).\end{aligned}$$

By (B.10) 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

$$\|\breve{D}_m(\boldsymbol{\theta}_m^*-\boldsymbol{\theta}^*)\|\le\alpha(m).$$

Further

$$\begin{aligned}&\|(\breve{D}_m-\breve{D})\big(\widetilde{\boldsymbol{\theta}}_m-\boldsymbol{\theta}_m^*\big)\|\\&\quad\le\|(\breve{D}_m-\breve{D}_m(\boldsymbol{\upsilon}^*))\big(\widetilde{\boldsymbol{\theta}}_m-\boldsymbol{\theta}_m^*\big)\|+\|(\breve{D}_m(\boldsymbol{\upsilon}^*)-\breve{D})\big(\widetilde{\boldsymbol{\theta}}_m-\boldsymbol{\theta}_m^*\big)\|\\&\quad\le\|\breve{D}_m\big(\widetilde{\boldsymbol{\theta}}_m-\boldsymbol{\theta}_m^*\big)\|\Big(\|\mathbb{I}-\breve{D}_m^{-1}\breve{D}_m^2(\boldsymbol{\upsilon}^*)\breve{D}_m^{-1}\|^{1/2}\\&\qquad+\|\mathbb{I}-\breve{D}_m(\boldsymbol{\upsilon}^*)^{-1}\breve{D}^2(\boldsymbol{\upsilon}^*)\breve{D}_m(\boldsymbol{\upsilon}^*)^{-1}\|^{1/2}\|\breve{D}_m(\boldsymbol{\upsilon}^*)\breve{D}_m^{-1}\|\Big).\end{aligned}$$

Condition $(\breve{\mathcal{ED}})$ yields that $\mathbb{P}(\|\breve{\boldsymbol{\xi}}_m(\boldsymbol{\upsilon}_m^*)\|\le\mathfrak{z}(\mathtt{x}_n,\breve{B}_m))\ge1-2\mathrm{e}^{\mathtt{x}_n}$ (see Section A). This gives with (B.10) that with probability greater $1-4\mathrm{e}^{\mathtt{x}_n}$

$$\|\breve{D}_m(\widetilde{\boldsymbol{\theta}}_m-\boldsymbol{\theta}_m^*)\|\le\|\breve{\boldsymbol{\xi}}_m(\boldsymbol{\upsilon}_m^*\|+\diamondsuit(\mathtt{r}_0,\mathtt{x})\le\mathfrak{z}(\mathtt{x},\breve{B}_m)+\diamondsuit(\mathtt{r}_0,\mathtt{x}).$$

where $\mathfrak{z}(\mathtt{x},\breve{B}_m)=O(\sqrt{p+\mathtt{x}})$ Combining these bounds gives with (*bias*′)

$$\|\breve{D}\big(\widetilde{\boldsymbol{\theta}}_m-\boldsymbol{\theta}^*\big)-\breve{\boldsymbol{\xi}}_m(\boldsymbol{\upsilon}_m^*)\|\le\diamondsuit(\mathtt{r}_0,\mathtt{x})+\beta(m)\Big(\mathfrak{z}(\mathtt{x},\breve{B}_m)+\diamondsuit(\mathtt{r}_0,\mathtt{x})\Big)+\alpha(m),$$

where $\mathtt{r}_0(\mathtt{x})$ is chosen such that $\mathbb{P}(\widetilde{\boldsymbol{\upsilon}}_n,\widetilde{\boldsymbol{\upsilon}}_{\boldsymbol{\theta}_m^*,m}\in\Upsilon_{0,m}(\mathtt{r}_0(\mathtt{x})))\ge1-\mathrm{e}^{-\mathtt{x}}$. By assumption $\mathtt{r}_0(\mathtt{x})<\infty$ for any $\mathtt{x}>0$, $m,n\in\mathbb{N}$. Remember that $\breve{\diamondsuit}\big(\mathtt{r}_0,\mathtt{x}_n\big)\approx\breve{\delta}_n(\mathtt{r}_0)\mathtt{r}_0+\breve{\omega}_n\sqrt{\mathtt{x}+p+m_n}\mathtt{r}_0$ where by assumption $\breve{\delta}_n(\mathtt{r})\to0$ for any $\mathtt{r}>0$ and $\omega_n\to0$. This implies that there exist sequences $(m_n)\subset\mathbb{N}$ with $m_n\to\infty$ and $\mathtt{x}_n\to\infty$ with

$$\diamondsuit(\mathtt{r}_0,\mathtt{x})+\beta(m)\Big(\mathfrak{z}(\mathtt{x},\breve{B}_{m_n})+\diamondsuit(\mathtt{r}_0,\mathtt{x})\Big)+\alpha(m_n)\to0\tag{B.11}$$

as $n\to\infty$. Fix such sequences $m_n\to\infty$ and $\mathtt{x}_n\to\infty$. Then we have due to (B.11) that for any $\epsilon>0$ there exists an $n\in\mathbb{N}$ such that

$$\mathbb{P}(\|\breve{D}(\widetilde{\boldsymbol{\theta}}_m-\boldsymbol{\theta}^*)-\breve{\boldsymbol{\xi}}_m(\boldsymbol{\upsilon}_m^*)\|\ge\epsilon)\le4\mathrm{e}^{\mathtt{x}_n}.$$

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

$$\begin{aligned}\breve{\boldsymbol{\xi}}_m(\boldsymbol{\upsilon}_m^*) &= \breve{D}_m^{-1}(\nabla_{\boldsymbol{\theta}} - A_m H_m^{-2}\nabla_{\boldsymbol{\eta}})\mathcal{L}(\boldsymbol{\upsilon}_m^*) \\ &= \frac{1}{\sqrt{n}}\sum_{i=1}^n (\frac{1}{\sqrt{n}}\breve{D}_m)^{-1}(\nabla_{\boldsymbol{\theta}}\ell_i(\boldsymbol{\upsilon}_m^*) - A_m H_m^{-2}\nabla_{\boldsymbol{\eta}}\ell_i(\boldsymbol{\upsilon}_m^*)) \\ &\stackrel{\text{def}}{=} \frac{1}{\sqrt{n}}\sum_{i=1}^n \boldsymbol{X}_i.\end{aligned}$$

Due to assumptions ***(bias'')*** we have $\mathrm{Cov}(\boldsymbol{X}_i) \to \breve{d}^{-1}\breve{v}^2\breve{d}^{-1} \in \mathbb{R}^{p\times p}$. Consequently

$$\breve{\boldsymbol{\xi}}_m(\boldsymbol{\upsilon}_m^*) = \frac{1}{\sqrt{n}}\sum_{i=1}^n \boldsymbol{X}_i,$$

where the random vectors $\boldsymbol{X}_i$ 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

$$\breve{\boldsymbol{\xi}}_m(\boldsymbol{\upsilon}_m^*) \xrightarrow{w} \mathcal{N}(0, \breve{d}^{-1}\breve{v}^2\breve{d}^{-1}).$$

Appendix C: A bound for the norm of a random process

We want to derive for a random process $\mathcal{Y}(\boldsymbol{\upsilon}) \in \mathbb{R}^p$ and $\boldsymbol{\upsilon} \in \Upsilon_\circ(\mathtt{r}) \subset \mathbb{R}^{p^*}$ a bound of the kind

$$\mathbb{P}\left(\sup_{\boldsymbol{\upsilon}\in\Upsilon_\circ(\mathtt{r})} \|\mathcal{Y}(\boldsymbol{\upsilon})\| \ge \mathtt{C}_\mathtt{3}(\mathtt{x}, 2p^* + 2p)\mathtt{r}\right) \le \mathrm{e}^{-\mathtt{x}}.$$

In the following we elaborate how to extend the results of the supplement of [29] on empirical processes to this situation without substantial changes to the bounds.

For this let $\mathcal{Y}(\boldsymbol{\upsilon})$ be a smooth centered random vector process with values in $\mathbb{R}^p$.We aim at bounding the maximum of the norm $\|\mathcal{Y}(\boldsymbol{\upsilon})\|$ over a vicinity $\Upsilon_\circ(\mathtt{r}) \stackrel{\text{def}}{=} \{\|\boldsymbol{\upsilon}-\boldsymbol{\upsilon}^*\|_\mathcal{Y} \le \mathtt{r}\}$ of $\boldsymbol{\upsilon}^*$ with some norm $\|\cdot\|_\mathcal{Y}$. Suppose that $\mathcal{Y}(\boldsymbol{\upsilon})$ satisfies for each $0 < \mathtt{r} < \mathtt{r}^*$ and for all pairs $\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^\circ \in \Upsilon_\circ(\mathtt{r}) = \{\boldsymbol{\upsilon} \in \Upsilon\colon \|\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^*\|_\mathcal{Y} \le \mathtt{r}\} \subset \mathbb{R}^{p^*}$ and $|\lambda| \le \mathtt{g}$

$$\sup_{\|\boldsymbol{u}\|\le 1} \log \mathbb{E}\exp\left\{\lambda \frac{\boldsymbol{u}^\top\big(\mathcal{Y}(\boldsymbol{\upsilon}) - \mathcal{Y}(\boldsymbol{\upsilon}^\circ)\big)}{\omega\|\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^\circ\|_\mathcal{Y}}\right\} \le \frac{\nu_0^2\lambda^2}{2}. \tag{C.1}$$

Remark C.1. In the setting of Theorem 2.2 and Proposition 2.4 we have

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

respectively and in both cases the norm becomes $\|\boldsymbol{v}-\boldsymbol{v}^\circ\|_{\mathcal{Y}} = \|\mathcal{D}(\boldsymbol{v}-\boldsymbol{v}^\circ)\|$ and condition (C.1) becomes $(\breve{\mathcal{E}}\mathcal{D}_1)$ from Section 2.1.

Theorem C.1. *Let a random p-vector process $\mathcal{Y}(\boldsymbol{v})$ fulfill $\mathcal{Y}(\boldsymbol{v}^*) = 0$ and the condition (C.1) be satisfied. Then for each $\mathtt{r} > 0$, on a set of probability greater $1 - \mathrm{e}^{-\mathtt{x}}$*

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

where with $\mathtt{g}_0 = \nu_0\mathtt{g}$ and for some $\mathbb{Q} > 0$

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

Remark C.2. Note that the entropy of the original set is increased by adding $p \in \mathbb{N}$ as the supremum is taken over $\Upsilon_\circ(\mathtt{r}) \times B_{\mathtt{r}}(0) \subset \mathbb{R}^{p^*} \times \mathbb{R}^p$.

Proof. In what follows, we use the representation

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

This implies

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

Due to Lemma C.2 the process $\mathcal{U}(\boldsymbol{v},\boldsymbol{u}) \stackrel{\mathrm{def}}{=} \frac{1}{2\mathtt{r}}\boldsymbol{u}^\top\mathcal{Y}(\boldsymbol{v})$ satisfies the condition (C.4) as process on $\mathbb{R}^{p^*} \times \mathbb{R}^p$. This allows to apply Corollary 2.2 of the supplement of [29] to obtain the desired result. We get on a set of probability greater $1 - \mathrm{e}^{-\mathtt{x}}$

$$\begin{aligned} \sup_{\boldsymbol{v}\in\Upsilon_\circ(\mathtt{r})} \|\mathcal{Y}(\boldsymbol{v})\| &\le 2 \sup_{\boldsymbol{v}\in\Upsilon_\circ(\mathtt{r})} \sup_{\|\boldsymbol{u}\|\le\mathtt{r}} \left\{\frac{1}{2\mathtt{r}}\boldsymbol{u}^\top\mathcal{Y}(\boldsymbol{v})\right\} \\ &\le 6\nu_1\mathtt{r}\mathfrak{z}\Big(\mathtt{x}, \mathbb{Q}\big(\Upsilon_\circ(\mathtt{r}) \times \mathcal{B}_{\mathtt{r}}(0)\big)\Big). \end{aligned}$$

The constant $\mathbb{Q}\big(\Upsilon_\circ(\mathtt{r}) \times \mathcal{B}_{\mathtt{r}}(0)\big) > 0$ quantifies the complexity of the set $\Upsilon_\circ(\mathtt{r}) \times \mathcal{B}_{\mathtt{r}}(0) \subset \mathbb{R}^{p^*} \times \mathbb{R}^p$. We point out that for compact $M \subset \mathbb{R}^{p^*}$ we have $\mathbb{Q}(M) = 2p^*$ (see Supplement of [29], Lemma 2.10). This gives $\mathbb{Q}\big(\Upsilon_\circ(\mathtt{r}) \times \mathcal{B}_{\mathtt{r}}(0)\big) = 2p^* + 2p$. □

Lemma C.2. *Suppose that $\mathcal{Y}(\boldsymbol{v})$ satisfies for each $\|\boldsymbol{u}\| \le 1$*

$$\sup_{\boldsymbol{v}\in\Upsilon_\circ} \log \mathbb{E}\exp\Big\{\frac{\lambda(\mathcal{Y}(\boldsymbol{v}) - \mathcal{Y}(\boldsymbol{v}^\circ))^\top\boldsymbol{u}}{\omega\|\boldsymbol{v}-\boldsymbol{v}^\circ\|_{\mathcal{Y}}}\Big\} \le \frac{\nu_0^2\lambda^2}{2}, \qquad |\lambda| \le \mathtt{g}. \tag{C.3}$$

Then for any $\boldsymbol{u}_1, \boldsymbol{u}_2 \in \mathbb{R}^p$ *with* $\|\boldsymbol{u}_i\|_{\mathcal{Y}} \le 2\mathtt{r}$ *and* $\|\boldsymbol{u}_i\| \le \mathtt{r}$

$$\log \mathbb{E} \exp\left\{ \frac{\lambda}{2\mathtt{r}} \frac{(\mathcal{Y}(\boldsymbol{\upsilon})^\top \boldsymbol{u}_1 - \mathcal{Y}(\boldsymbol{\upsilon}^\circ)^\top \boldsymbol{u}_2)}{\omega \sqrt{\|\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^\circ\|_{\mathcal{Y}}^2 + \|\boldsymbol{u}_1 - \boldsymbol{u}_2\|^2}} \right\} \le \frac{\nu_0^2 \lambda^2}{2}, \qquad |\lambda| \le \mathtt{g}. \quad \text{(C.4)}$$

Proof. We simply plug in the definition to find for $\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^\circ \in \Upsilon_\circ(\mathtt{r})$

$$\begin{aligned} &\log \mathbb{E} \exp\left\{ \frac{\lambda}{2\mathtt{r}} \frac{\boldsymbol{u}_1^\top \mathcal{Y}(\boldsymbol{\upsilon}) - \boldsymbol{u}_2^\top \mathcal{Y}(\boldsymbol{\upsilon}^\circ)}{\omega \sqrt{\|\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^\circ\|_{\mathcal{Y}}^2 + \|\boldsymbol{u}_1 - \boldsymbol{u}_2\|^2}} \right\} \\ &\quad = \log \mathbb{E} \exp\left\{ \frac{\lambda}{2\mathtt{r}} \frac{\boldsymbol{u}_1^\top \big(\mathcal{Y}(\boldsymbol{\upsilon}) - \mathcal{Y}(\boldsymbol{\upsilon}^\circ)\big) + (\boldsymbol{u}_1^\top - \boldsymbol{u}_2^\top)\mathcal{Y}(\boldsymbol{\upsilon}^\circ)}{\omega \sqrt{\|\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^\circ\|_{\mathcal{Y}}^2 + \|\boldsymbol{u}_1 - \boldsymbol{u}_2\|^2}} \right\}. \end{aligned}$$

By the Hölder inequality and (C.3) we infer

$$\begin{aligned} &\log \mathbb{E} \exp\left\{ \frac{\lambda}{2\mathtt{r}} \frac{\boldsymbol{u}_1^\top \big(\mathcal{Y}(\boldsymbol{\upsilon}) - \mathcal{Y}(\boldsymbol{\upsilon}^\circ)\big) + (\boldsymbol{u}_1^\top - \boldsymbol{u}_2^\top)\mathcal{Y}(\boldsymbol{\upsilon}^\circ)}{\omega \sqrt{\|\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^\circ\|_{\mathcal{Y}}^2 + \|\boldsymbol{u}_1 - \boldsymbol{u}_2\|^2}} \right\} \\ &\le \frac{1}{2} \log \mathbb{E} \exp\left\{ \frac{\lambda}{\mathtt{r}} \frac{\boldsymbol{u}_1^\top \big(\mathcal{Y}(\boldsymbol{\upsilon}) - \mathcal{Y}(\boldsymbol{\upsilon}^\circ)\big)}{\omega \|\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^\circ\|_{\mathcal{Y}}} \right\} \\ &\quad + \frac{1}{2} \log \mathbb{E} \exp\left\{ \frac{\lambda}{\mathtt{r}} \frac{(\boldsymbol{u}_1^\top - \boldsymbol{u}_2^\top)\mathcal{Y}(\boldsymbol{\upsilon}^\circ)}{\omega \|\boldsymbol{u}_1 - \boldsymbol{u}_2\|} \right\} \\ &\le \sup_{\|\boldsymbol{u}\| \le 1} \frac{1}{2} \log \mathbb{E} \exp\left\{ \lambda \frac{\boldsymbol{u}^\top \big(\mathcal{Y}(\boldsymbol{\upsilon}) - \mathcal{Y}(\boldsymbol{\upsilon}^\circ)\big)}{\omega \|\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^\circ\|_{\mathcal{Y}}} \right\} \\ &\quad + \sup_{\|\boldsymbol{u}\| \le 1} \frac{1}{2} \log \mathbb{E} \exp\left\{ \lambda \frac{\boldsymbol{u}^\top \big(\mathcal{Y}(\boldsymbol{\upsilon}^\circ) - \mathcal{Y}(\boldsymbol{\upsilon}^*)\big)}{\omega \|\boldsymbol{\upsilon}^\circ - \boldsymbol{\upsilon}^*\|_{\mathcal{Y}}} \right\} \le \frac{\nu_0^2 \lambda^2}{2}, \qquad \lambda \le \mathtt{g}. \end{aligned}$$

□

Aknowledgements

We are grateful to the referees and editors for very helpful remarks and comments.

References

[1] Andresen, A., Finite sample analysis of profile m-estimation in the single index model. ArXiv:1406.4052, 2014.
[2] Andresen, A., A note on critical dimensions in profile semiparametric estimation. ArXiv:1410.4709, 2014.
[3] Andresen, A., A note on the bias of sieve profile estimation. ArXiv: 1406.4045, 2014.

[4] Berry, A., The accuracy of the gaussian approximation to the sum of independent variates. *Transactions of the American Mathematical Society*, 49(1):122–136, 1941. MR0003498

[5] Bickel, P. J., Klaassen, C. A. J., Ritov, Y., and Wellner, J. A., *Efficient and Adaptive Estimation for Semiparametric Models*. Springer, 1998. MR1623559

[6] Boucheron, S. and Massart, P., A high-dimensional Wilks phenomenon. *Probability Theory and Related Fields*, 150:405–433, 2011. 10.1007/s00440-010-0278-7. MR2824862

[7] Delecroix, M., Haerdle, W., and Hristache, M., Efficient estimation in single-index regression. Technical report, SFB 373, Humboldt Univ. Berlin, 1997.

[8] Fan, J. and Huang, T., Profile likelihood inferences on semiparametric varying-coefficient partially linear models. *Bernoulli*, 11(6):1031–1057, 2005. MR2189080

[9] Fan, J., Zhang, C., and Zhang, J., Generalized likelihood ratio statistics and Wilks phenomenon. *Ann. Stat.*, 29(1):153–193, 2001. MR1833962

[10] Ghosal, S., Asymptotic normality of posterior distributions in high-dimensional linear models. *Bernoulli*, 5(2):315–331, 1999. MR1681701

[11] Ghosal, S., Asymptotic normality of posterior distributions for exponential families when the number of parameters tends to infinity. *J. Multivariate Anal.*, 74(1):49–68, 2000. MR1790613

[12] Haerdle, W., Hall, P., and Ichimura, H., Optimal smoothing in single-index models. *Ann. Statist.*, 21:157–178, 1993. MR1212171

[13] Hall, P., *The Bootstrap and Edgeworth Expansion*. Springer, 1992. MR1145237

[14] Huber, P. J., The behavior of maximum likelihood estimates under nonstandard conditions. Proc. 5th Berkeley Symp. Math. Stat. Probab., Univ. Calif. 1965/1966, 1, 221–233 (1967). MR0216620

[15] Ibragimov, I. A. and Khas'minskij, R. Z., *Statistical Estimation. Asymptotic Theory. Transl. from the Russian by Samuel Kotz*. Springer-Verlag, New York–Heidelberg–Berlin, 1981. MR0620321

[16] Ichimura, H., Semiparametric least squares (sls) and weighted sls estimation of single-index models. *J. Econometrics*, 58:71–120, 1993. MR1230981

[17] Kim, Y., The Bernstein-von Mises theorem for the proportional hazard model. *Ann. Statist.*, 34(4):1678–1700, 2006. MR2283713

[18] Kosorok, M. R., *Introduction to Empirical Processes and Semiparametric Inference*. Springer in Statistics, 2005. MR2724368

[19] Mammen, E., Asymptotics with increasing dimension for robust regression with applications to the bootstrap. *Ann. Stat.*, 17(1):382–400, 1989. MR0981457

[20] Mammen, E., Bootstrap and wild bootstrap for high dimensional linear models. *Ann. Stat.*, 21(1):255–285, 1993. MR1212176

[21] Mammen, E., Empirical process of residuals for high-dimensional linear models. *Ann. Stat.*, 24(1):307–335, 1996. MR1389892

[22] Murphy, S. A. and van der Vaart, A. W., On profile likelihood. *Journal of the American Statistical Association*, 95(450):449–465, 2000. MR1803168

[23] Murphy, S. A. and van der Vaart, A. W., Observed information in semi-parametric models. *Bernoulli*, 5(3):381–412, 1999. MR1693616

[24] Newey, W. K., Convergence rates and asymptotic normality for series estimators. *Journal of Econometrics*, 79(1):147–168, 1997.

[25] Portnoy, S., Asymptotic behavior of m estimators of p regression parameters when p^2/n is large: Ii normal approximation. *The Annals of Statistics*, 13(4):1403–1417, 1985. MR0811499

[26] Portnoy, S., Asymptotic behavior of M-estimators of p regression parameters when p^2/n is large. I. Consistency. *Ann. Stat.*, 12:1298–1309, 1984. MR0760690

[27] Portnoy, S., Asymptotic behavior of the empiric distribution of M-estimated residuals from a regression model with many parameters. *Ann. Stat.*, 14:1152–1170, 1986. MR0856812

[28] Shen, J., Shi, X., Sieve likelihood ratio inference on general parameter space. *Science in China*, 48(1):67–78, 2005. MR2156616

[29] Spokoiny, V., Parametric estimation. Finite sample theory. *Ann. Statist.*, 40(6):2877–2909, 2012. MR3097963

[30] Spokoiny, V., Bernstein–von Mises theorem for growing parameter dimension. Manuscript. ArXiv:1302.3430, 2013.

[31] Spokoiny, V., Wang, W., and Härdle, W., Local quantile regression (with rejoinder). *J. of Statistical Planing and Inference*, 143(7):1109–1129, 2013. ArXiv:1208.5384. MR3049611

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

[33] Wilks, S. S., The large-sample distribution of the likelihood ratio for testing composite hypotheses. *Ann. Math. Statist.*, 9(1):60–62, 1938.

[34] Zaitsev, A., Burnaev, E., and Spokoiny, V., Properties of the posterior distribution of a regression model based on gaussian random fields. *Automation and Remote Control*, 74(10):1645–1655, 2013. MR3219856