# Weierstraß-Institut
## für Angewandte Analysis und Stochastik
### Leibniz-Institut im Forschungsverbund Berlin e. V.

Preprint ISSN 0946 – 8633


# Critical dimension in profile semiparametric estimation


Andreas Andresen [1] Vladimir Spokoiny [2]



submitted: April 9, 2013



[1] Weierstrass Institute
Mohrenstr. 39
10117 Berlin
Germany
E-Mail: andreas.andresen@wias-berlin.de

[2] HU Berlin and Weierstrass Institute
Mohrenstr. 39
10117 Berlin
Germany
E-Mail: vladimir.spokoiny@wias-berlin.de



No. 1776
Berlin 2013



2010 *Mathematics Subject Classification.* Primary 62F10; secondary 62J12, 62F25, 62H12.

*Key words and phrases.* maximum likelihood, local quadratic bracketing, spread, local concentration.

The first author is supported by Research Units 1735 “Structural Inference in Statistics: Adaptation and Efficiency” Financial support by the German Research Foundation (DFG) through the Collaborative Research Center 649 “Economic Risk” is gratefully acknowledged .

Edited by
Weierstraß-Institut für Angewandte Analysis und Stochastik (WIAS)
Leibniz-Institut im Forschungsverbund Berlin e. V.
Mohrenstraße 39
10117 Berlin
Germany

Fax: +49 30 20372-303
E-Mail: preprint@wias-berlin.de
World Wide Web: http://www.wias-berlin.de/

**Abstract**

This paper revisits the classical inference results for profile quasi maximum likelihood estimators (profile MLE) in the semiparametric estimation problem. 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 apply the local bracketing and the upper function devices from Spokoiny (2012). This novel approach particularly allows to address the important issue of the effective target and nuisance dimension and it does not involve any pilot estimator of the target parameter. 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 to the important special cases of an i.i.d. sample.

# 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 Murphy and Van der Vaart (2000, 1999) and the book Kosorok (2005) 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 Fan et al. (2001); Fan and Huang (2005); Boucheron and Massart (2011); 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 of notions, methods and tools in the classical theory are formulated in the asymptotic setup with growing sample size. Only few finite sample general results are available; see e.g. the recent paper Boucheron and Massart (2011). 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 applicability of the approach in different situations. A particularly important question is about 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 $I\!P$ does not belong to the considered parametric family. Applicability of the general semiparametric theory in such cases is questionable. An important feature of the new approach of Spokoiny (2012) is that it equally applies under a possible model misspecification.

The mentioned issues, especially the non-asymptotic character of study dictate to change entirely the tools and methods of analysis. We apply the recent bracketing approach of Spokoiny (2012) and demonstrate its power on the considered case of semiparametric estimation. Let $\boldsymbol{Y}$ denote the observed random data, and $I\!P$ denote the data distribution. The parametric statistical model assumes that the unknown data distribution $I\!P$ belongs to a given parametric family $(I\!P_{\boldsymbol{\upsilon}})$:

$$\boldsymbol{Y} \sim I\!P = I\!P_{\boldsymbol{\upsilon}^*} \in (I\!P_{\boldsymbol{\upsilon}}, \boldsymbol{\upsilon} \in \Upsilon),$$

where $\Upsilon$ is some high dimensional or even infinite dimensional parameter space. This paper concentrates on a finite dimensional setting, however, an extension to a functional space is feasible and to be considered elsewhere. The maximum likelihood approach in the parametric estimation suggests to estimate the whole parameter vector $\boldsymbol{\upsilon}$ by maximizing the corresponding log-likelihood $\mathcal{L}(\boldsymbol{\upsilon}) = \log \frac{dI\!P_{\boldsymbol{\upsilon}}}{d\boldsymbol{\mu}_0}(\boldsymbol{Y})$ for some dominating measure $\boldsymbol{\mu}_0$:

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

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

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

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

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

$$\boldsymbol{\theta}^* = \mathit{\Pi}_0 \boldsymbol{\upsilon}^*,$$

for some mapping $\mathit{\Pi}_0 : \Upsilon \to I\!R^p$, and $p \in \mathbb{N}$ stands for the dimension of the target.

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

$$\widetilde{\boldsymbol{\theta}} = \mathit{\Pi}_0 \widetilde{\boldsymbol{\upsilon}}.$$

The Gauss-Markov Theorem claims the efficiency of such procedures for linear Gaussian models and linear mapping $\mathit{\Pi}_0$, 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 $T$:

$$T \stackrel{\text{def}}{=} \sup_{\boldsymbol{\upsilon} \in \Upsilon} \mathcal{L}(\boldsymbol{\upsilon}) - \sup_{\substack{\boldsymbol{\upsilon} \in \Upsilon \\ \Pi_0 \boldsymbol{\upsilon} = \boldsymbol{\theta}^*}} \mathcal{L}(\boldsymbol{\upsilon}). \tag{1.1}$$

It appears that the distribution of this test statistic is nearly chi-square $\chi^2_p$ as the samples size grows, Wilks (1938). In particular, this limiting behavior 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.

Below we consider a slightly different presentation of this estimator based on the partial optimization of the objective function $\mathcal{L}(\boldsymbol{\upsilon})$ for a fixed $\boldsymbol{\theta}$. Namely, define

$$\breve{L}(\boldsymbol{\theta}) \stackrel{\text{def}}{=} \max_{\substack{\boldsymbol{\upsilon} \in \Upsilon \\ \Pi_0 \boldsymbol{\upsilon} = \boldsymbol{\theta}}} \mathcal{L}(\boldsymbol{\upsilon}). \tag{1.2}$$

Then the profile MLE can be defined as the point of maximum of $\breve{L}(\boldsymbol{\theta})$:

$$\widetilde{\boldsymbol{\theta}} = \operatorname*{argmax}_{\boldsymbol{\theta} \in \Theta} \breve{L}(\boldsymbol{\theta}) = \operatorname*{argmax}_{\boldsymbol{\theta} \in \Theta} \max_{\substack{\boldsymbol{\upsilon} \in \Upsilon \\ \Pi_0 \boldsymbol{\upsilon} = \boldsymbol{\theta}}} \mathcal{L}(\boldsymbol{\upsilon}).$$

The test statistic $T$ from (1.1) is also called the *semiparametric excess* and it can be defined as

$$\breve{L}(\widetilde{\boldsymbol{\theta}}) - \breve{L}(\boldsymbol{\theta}^*) = \max_{\boldsymbol{\upsilon} \in \Upsilon} \mathcal{L}(\boldsymbol{\upsilon}) - \max_{\substack{\boldsymbol{\upsilon} \in \Upsilon \\ \Pi_0 \boldsymbol{\upsilon} = \boldsymbol{\theta}^*}} \mathcal{L}(\boldsymbol{\upsilon}).$$

The Wilks result can be rewritten as

$$2\bigl\{ \breve{L}(\widetilde{\boldsymbol{\theta}}) - \breve{L}(\boldsymbol{\theta}^*) \bigr\} \stackrel{w}{\longrightarrow} \chi^2_p.$$

The *local asymptotic normality* (LAN) approach by Le Cam leads to the most general setup in which the Wilks type results can be established. However, the classical theory of semiparametric estimation faces serious difficulties when the dimension of the nuisance parameter becomes large of infinite. The LAN property yields a 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 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 Murphy and Van der Vaart (2000) or Kosorok (2005) for a comprehensive presentation.

The recent paper Spokoiny (2012) offers a new look at the classical LAN theory. The basic idea is to replace the local approximation by *local bracketing*. Instead of one approximating Gaussian log-likelihood, one builds two different quadratic processes such that the original log-likelihood

can be sandwiched between them up to a small error. It appears that the bracketing device can be applied for much larger neighborhoods than in the LAN approach. In this paper we show that the local bracketing approach of Spokoiny (2012) can be used for obtaining a version of the Wilks Theorem in a quite general semiparametric setup avoiding any special construction like "the hardest parametric submodel".

Another important issue is that the new approach does not rely on any pilot estimator of the target. The usual assumption that a consistent pilot estimator is available can be even misleading in our setup because it separates local and global considerations. This paper attempts to figure out a list of condition ensuring global concentration and local expansion at the same time. This particularly allows to address the crucial question of the largest dimensionality or the nuisance parameter for which the Wilks result still holds. It appears that the profile semiparametric approach is validated under the constraint $p^{*3} \ll n$, where $p^*$ is the full parameter dimension. It applies even if the dimension $p$ of the target grows with the sample size under the mentioned constraint. The important *identifiability* issue is also addressed in a more careful way for the considered finite sample case.

For the further presentation we have to briefly outline the basic results from Spokoiny (2012). Introduce the log-likelihood ratio process

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

The key *bracketing* result of Spokoiny (2012) claims that $\mathcal{L}(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^*)$ can be sandwiched on a local elliptic set $\Upsilon_\circ(\mathtt{r})$ around $\boldsymbol{\upsilon}^*$ by two quadratic in $\boldsymbol{\upsilon}$ processes $\mathbb{L}_{\boldsymbol{\epsilon}}(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^*)$ and $\mathbb{L}_{\underline{\boldsymbol{\epsilon}}}(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^*)$:

$$\mathbb{L}_{\underline{\boldsymbol{\epsilon}}}(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^*) - \diamondsuit_{\underline{\boldsymbol{\epsilon}}}(\mathtt{r}) \leq \mathcal{L}(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^*) \leq \mathbb{L}_{\boldsymbol{\epsilon}}(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^*) + \diamondsuit_{\boldsymbol{\epsilon}}(\mathtt{r}), \quad \boldsymbol{\upsilon} \in \Upsilon_\circ(\mathtt{r}), \tag{1.3}$$

where $\diamondsuit_{\boldsymbol{\epsilon}}(\mathtt{r}) > 0$ and $\diamondsuit_{\underline{\boldsymbol{\epsilon}}}(\mathtt{r}) > 0$ are small terms. The value $\mathtt{r}$ here can be viewed as the radius of the set $\Upsilon_\circ(\mathtt{r})$ in the intrinsic semimetric corresponding to the process $L(\boldsymbol{\theta})$. See Section B for a precise formulation. This local result is accompanied with the deviation bound of the form

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

where $\mathtt{x}$ grows almost linearly with $\mathtt{r}$. The bracketing result (1.3) yields a number of important and informative corollaries. One of them shows that the excess $\mathcal{L}(\widetilde{\boldsymbol{\upsilon}}, \boldsymbol{\upsilon}^*)$ can be approximated by a quadratic form $\|\boldsymbol{\xi}\|^2/2$, where $\boldsymbol{\xi} \stackrel{\mathrm{def}}{=} \mathcal{D}_0^{-1} \nabla \mathcal{L}(\boldsymbol{\upsilon}^*)$ is the normalized score while $\mathcal{D}_0^2$ approximates the total Fisher information matrix. Another important corollary of (1.3) is an expansion of the quasi MLE $\widetilde{\boldsymbol{\upsilon}}$. The mentioned results can be written in the form

$$\begin{aligned} \bigl| 2\mathcal{L}(\widetilde{\boldsymbol{\upsilon}}, \boldsymbol{\upsilon}^*) - \|\boldsymbol{\xi}\|^2 \bigr| &\leq 2\Delta_{\boldsymbol{\epsilon}}, \\ \bigl\| \mathcal{D}_0 \bigl( \widetilde{\boldsymbol{\upsilon}} - \boldsymbol{\upsilon}^* \bigr) - \boldsymbol{\xi} \bigr\|^2 &\leq 2\Delta_{\boldsymbol{\epsilon}}, \end{aligned} \tag{1.4}$$

where $\Delta_{\boldsymbol{\epsilon}}$ is a random term called the *spread* which is small with a large probability. In a typical situation with a correctly specified model, $\boldsymbol{\xi}$ is nearly standard normal and hence, $2\mathcal{L}(\widetilde{\boldsymbol{\upsilon}}, \boldsymbol{\upsilon}^*)$ is nearly $\chi^2_{p^*}$, where $p^*$ is the full parameter dimension, while the MLE $\widetilde{\boldsymbol{\upsilon}}$ is asymptotically normal and efficient. The expansion (1.4) helps to build likelihood-based confidence sets for the

true parameter $\boldsymbol{\upsilon}^*$. Let $\chi_\alpha$ be the $(1-\alpha)$-quantile of the chi-square distribution with $p^*$ degrees of freedom. Set

$$\mathcal{E}(\alpha) \stackrel{\text{def}}{=} \{\boldsymbol{\upsilon} \in \Upsilon : 2\mathcal{L}(\widetilde{\boldsymbol{\upsilon}}, \boldsymbol{\upsilon}) \leq \chi_\alpha\}.$$

Then (1.4) ensures that the coverage probability $I\!P\big(\boldsymbol{\upsilon}^* \notin \mathcal{E}(\alpha)\big)$ is close to $\alpha$ provided that $\Delta_{\boldsymbol{\epsilon}}$ is sufficiently small.

This paper aims at establishing a similar statements for the process $\breve{L}(\boldsymbol{\theta})$ from (1.2). In particular, the Wilks result can be written as

$$\breve{L}(\widetilde{\boldsymbol{\theta}}) - \breve{L}(\boldsymbol{\theta}^*) \cong \|\breve{\boldsymbol{\xi}}\|^2/2,$$

where the random $p$-vector $\breve{\boldsymbol{\xi}}$ satisfies $I\!E\breve{\boldsymbol{\xi}} = 0$ and $I\!E\|\breve{\boldsymbol{\xi}}\|^2 \cong p$. The deviation properties of $\|\breve{\boldsymbol{\xi}}\|^2$ resemble the ones of a chi-square random variable with $p$ degrees of freedom just as in the Wilks phenomenon. The expansion of the profile MLE $\widetilde{\boldsymbol{\theta}}$ reads as

$$\breve{D}_0\big(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\big) \cong \breve{\boldsymbol{\xi}}.$$

The symmetric matrix $\breve{D}_0^2 \in I\!R^{p \times p}$ is usually called the influence matrix and it is the covariance of the efficient influence function; see Kosorok (2005).

Usually in the classical semiparametric setup, 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. An extension to the $\boldsymbol{\upsilon}$-setup with $\boldsymbol{\theta} = \Pi_0 \boldsymbol{\upsilon}$ is straightforward. Also for simplicity we only develop our results for the case that the full parameter space $\Upsilon$ is a subset of the Euclidean space of dimensionality $p^*$. An extension to an infinite dimensional parameter space is possible but involves a range of technical issues that have to be done elsewhere.

Section 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 $\boldsymbol{\theta}^*$. The concentration properties of the profile MLE are discussed in Section D.1. The appendix collects the conditions and 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 and the Fisher expansion of the profile MLE $\widetilde{\boldsymbol{\theta}}$. All the results are stated under the same list of conditions that can be found in Section A of the appendix. As already mentioned, our setup follows Spokoiny (2012). However, at one point there is an essential difference. The results of Spokoiny (2012) are stated for just one fixed finite sample. The same continues to hold for the results below. But we are also interested in understanding what happens if *the full dimension* $p^*$ becomes large. For this we consider below an asymptotic setup with $p^* = p_n$, where $n$ denotes the asymptotic parameter. It can

be viewed as the sample size with $n \to \infty$. We assume that all considered objects depend on $n$ including the likelihood function, the full parameter set $\Upsilon$ and its dimension $p^*$, as well as all the constants in our conditions. The primary goal of our study is to fix the necessary and sufficient conditions on growth of $p_n$ with $n$ which ensures the Wilks and Fisher results.

Our result apply even if the target parameter $\boldsymbol{\theta}$ is of growing dimension. The dimension $p$ can be of order $p^*$. The case with a full dimensional target and low dimensional nuisance is also included.

## 2.1 The Wilks and Fisher expansion

This section states the key results in the semiparametric framework which heavily use the local bracketing idea of Spokoiny (2012). First we introduce the main elements of the bracketing device. This includes two $p^* \times p^*$ matrices $\mathcal{V}_0^2$ and $\mathcal{D}_0^2$ and two constants $\boldsymbol{\epsilon} = (\delta, \varrho)$. The matrix $\mathcal{V}_0^2$ describes the variability of the process $\mathcal{L}(\boldsymbol{\upsilon})$ around the true point $\boldsymbol{\upsilon}^*$:

$$\mathcal{V}_0^2 \stackrel{\text{def}}{=} \operatorname{Var}\bigl\{\nabla \mathcal{L}(\boldsymbol{\upsilon}^*)\bigr\}. \tag{2.1}$$

The matrix $\mathcal{D}_0^2$ is defined similarly to the Fisher information matrix:

$$\mathcal{D}_0^2 \stackrel{\text{def}}{=} -\nabla^2 I\!\!E \mathcal{L}(\boldsymbol{\upsilon}^*). \tag{2.2}$$

Here and in what follows we implicitly assume that the log-likelihood function $\mathcal{L}(\boldsymbol{\upsilon})$ is sufficiently smooth in $\boldsymbol{\upsilon}$, $\nabla \mathcal{L}(\boldsymbol{\upsilon})$ stands for the gradient and $\nabla^2 I\!\!E \mathcal{L}(\boldsymbol{\upsilon})$ for the Hessian of the expectation $I\!\!E \mathcal{L}$ at $\boldsymbol{\upsilon}$. It is worth mentioning that the matrices $\mathcal{D}_0^2$ and $\mathcal{V}_0^2$ coincide if the model $\boldsymbol{Y} \sim I\!\!P_{\boldsymbol{\upsilon}^*} \in (I\!\!P_{\boldsymbol{\upsilon}})$ is correctly specified and sufficiently regular; see e.g. Ibragimov and Khas'minskij (1981).

Now we switch to the $(\boldsymbol{\theta}, \boldsymbol{\eta})$-setup. Consider the block representation of the vector $\nabla \stackrel{\text{def}}{=} \nabla \mathcal{L}(\boldsymbol{\upsilon}^*)$ and of the matrices $\mathcal{V}_0^2$ from (2.1) and $\mathcal{D}_0^2$ from (2.2):

$$\nabla = \begin{pmatrix} \nabla_{\boldsymbol{\theta}} \\ \nabla_{\boldsymbol{\eta}} \end{pmatrix}, \quad \mathcal{D}_0^2 = \begin{pmatrix} D_0^2 & A_0 \\ A_0^\top & H_0^2 \end{pmatrix}, \quad \mathcal{V}_0^2 = \begin{pmatrix} V_0^2 & B_0 \\ B_0^\top & Q_0^2 \end{pmatrix}.$$

Define also the $p \times p$ matrix $\breve{D}_0^2$ and $p$-vectors $\breve{\nabla}_{\boldsymbol{\theta}}$ and $\breve{\boldsymbol{\xi}}$ as

$$\begin{aligned} \breve{D}_0^2 &= D_0^2 - A_0 H_0^{-2} A_0^\top, \\ \breve{\nabla}_{\boldsymbol{\theta}} &= \nabla_{\boldsymbol{\theta}} - A_0 H_0^{-2} \nabla_{\boldsymbol{\eta}}, \\ \breve{\boldsymbol{\xi}} &= \breve{D}_0^{-1} \breve{\nabla}_{\boldsymbol{\theta}}. \end{aligned}$$

In what follows, by $\mathtt{C}$ we denote a generic fixed constant. For all results presented below we assume a sufficiently large value $\mathtt{x}$ to be fixed. It determines our level of overwhelming probability: a generic random set $\Omega(\mathtt{x})$ is of *dominating probability* if

$$I\!\!P\bigl(\Omega(\mathtt{x})\bigr) \geq 1 - \mathtt{C} \mathrm{e}^{-\mathtt{x}}.$$

Similarly to $p^*$, the value $\mathtt{x}$ may depend on the asymptotic parameter $n$ and grows to infinity with $n$. A particularly relevant choice is $\mathtt{x} = \mathtt{x}_n = \mathtt{C}\log n$ for a fixed $\mathtt{C} > 0$. We only require that $\mathtt{x}_n$ is not too large, more precisely, $\mathtt{x} \leq \mathtt{x}_c$; see (C.2) from Section C. In the i.i.d. setup $\mathtt{x}_c$ is of order $n^{1/2}$.

The other important value to be fixed is $\mathtt{r}_0$. This value determines the frontier between local and global consideration. In the local vicinity $\Upsilon_\circ(\mathtt{r}_0)$ of radius $\mathtt{r}_0$ we apply a very accurate local quadratic approximation of the log-likelihood process while outside of this vicinity a much more rough upper function device can be used; see Section B for more details. The general rule for the choice of $\mathtt{r}_0$ is given by the condition $\mathtt{r}_0^2 \geq \mathtt{C}_0(p^* + \mathtt{x})$ for some specific constant $\mathtt{C}_0$ that is independent of $p^*$ and $n \in \mathbb{N}$ if $\nu_0/\mathtt{b} > 0$ is independent of $p^*$ and $n \in \mathbb{N}$ (see Section A). The quality of local quadratic approximation is measured by two functions $\delta(\mathtt{r})$ and $\omega(\mathtt{r})$ shown in local conditions $(\mathcal{ED}_1)$, $(\mathcal{L}_0)$ of Section A. More exactly, it can be described by the quantities $\tau_{\boldsymbol{\epsilon}}$ defined as

$$\tau_{\boldsymbol{\epsilon}} \stackrel{\text{def}}{=} \delta(\mathtt{r}_0) + 3\nu_0\mathfrak{a}^2\omega(\mathtt{r}_0), \tag{2.3}$$

where the constants $\nu_0$ and $\mathfrak{a}$ are from conditions $(\mathcal{ED}_1)$ and $(\mathcal{I})$ in Section A. The sub-index $\boldsymbol{\epsilon}$ stands for the pair $\delta(\mathtt{r}_0), \omega(\mathtt{r}_0)$. Our results implicitly assume that $\tau_{\boldsymbol{\epsilon}}$ is small. We comment on typical behavior of $\tau_{\boldsymbol{\epsilon}}$ is Section 2.2 in context of i.i.d. models.

The first result can be viewed as an extension of the Wilks Theorem.

**Theorem 2.1.** *Assume* $(\mathcal{ED}_0)$, $(\mathcal{ED}_1)$, $(\mathcal{L}_0)$, $(\mathcal{I})$, $(\mathcal{E}\mathtt{r})$ *and* $(\mathcal{L}\mathtt{r})$ *with* $\mathtt{b}(\mathtt{r}) \equiv \mathtt{b}$*; see Section A. Let also* $\tau_{\boldsymbol{\epsilon}}$ *from* (2.3) *fulfill* $\tau_{\boldsymbol{\epsilon}} \leq 1/2$. *Then it holds on a random set* $\Omega(\mathtt{x})$ *of dominating probability*

$$\bigl|2\breve{L}(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}^*) - \|\breve{\boldsymbol{\xi}}\|^2\bigr| \leq \mathtt{C}\,\tau_{\boldsymbol{\epsilon}}\,(p^* + \mathtt{x}), \tag{2.4}$$

**Remark 2.1.** In the case of the correct model specification with $\mathcal{D}_0^2 = \mathcal{V}_0^2$, the deviation properties of the quadratic form $\|\breve{\boldsymbol{\xi}}\|^2 = \|\breve{D}_0^{-1}\breve{\nabla}_{\boldsymbol{\theta}}\|^2$ are essentially the same as of a chi-square random variable with $p$ degrees of freedom; see Theorem C.1 in the appendix. In the case of a possible model misspecification with $\mathcal{D}_0^2 \neq \mathcal{V}_0^2$, the behavior of the quadratic form $\|\breve{\boldsymbol{\xi}}\|^2$ will depend on the characteristics of the matrix $I\!B \stackrel{\text{def}}{=} \mathcal{D}_0^{-1}\mathcal{V}_0^2\mathcal{D}_0^{-1}$; see again Theorem C.1. Moreover, in the asymptotic setup the vector $\breve{\boldsymbol{\xi}}$ is asymptotically standard normal; see Section 2.2 for the i.i.d. case.

**Remark 2.2.** The partial maximum likelihood process $\breve{L}(\boldsymbol{\theta})$ can be used for defining the likelihood-based confidence sets of the form

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

for some $\mathfrak{z} > 0$. The bound (2.4) helps to evaluate the coverage probability $I\!P\bigl(\boldsymbol{\theta}^* \notin \mathcal{E}(\mathfrak{z})\bigr)$ in terms of deviation probability for the quadratic form $\|\breve{\boldsymbol{\xi}}\|^2$; cf. Corollary 3.2 in Spokoiny (2012).

The next result presents an expansion of the profile MLE $\widetilde{\boldsymbol{\theta}}$.

**Theorem 2.2.** *Under the conditions of Theorem 2.1, it holds on a random set* $\Omega(\mathtt{x})$ *of dominating probability*

$$\bigl\| \breve{D}_0 \bigl( \widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^* \bigr) - \breve{\boldsymbol{\xi}} \bigr\|^2 \leq \mathtt{C}\, \tau_{\boldsymbol{\epsilon}}\, (p^* + \mathtt{x}). \tag{2.5}$$

**Remark 2.3.** One can use the expansion (2.5) for describing the concentration probability for elliptic sets

$$\mathcal{A}(z) = \bigl\{ \boldsymbol{\theta} : \| \breve{D}_0 (\boldsymbol{\theta} - \boldsymbol{\theta}^*) \| \leq z \bigr\};$$

cf. Corollary 3.5 in Spokoiny (2012).

In the next section the result (2.5) is used to show asymptotic normality and efficiency of the profile estimator in the i.i.d. setting and under the correct model specification.

## 2.2 The i.i.d. case and asymptotic efficiency

Here we briefly discuss the implications of our general results to the case with $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^\top$ where observations $Y_i$ are i.i.d. from a measure $P$. The parametric assumption means $P = P_{\boldsymbol{\upsilon}^*} \in (P_{\boldsymbol{\upsilon}}, \boldsymbol{\upsilon} \in \Upsilon)$ for a given parametric family $(P_{\boldsymbol{\upsilon}})$, where $\Upsilon$ is a subset of the Euclidean space $\mathbb{R}^{p^*}$. We assume that $(P_{\boldsymbol{\upsilon}})$ obeys the regularity conditions listed in Section 5.1 of Spokoiny (2012). By $\ell(y, \boldsymbol{\upsilon})$ we denote the log-density of $P_{\boldsymbol{\upsilon}}$ w.r.t. some dominating measure $\mu_0$. For simplicity of comparison with the classical results we do not discuss the model misspecification issue, i.e. the parametric assumption is correct. However, an extension to the case of a misspecified model is straightforward. We utilize that $\mathcal{V}_0^2 = \mathcal{D}_0^2 = n\mathbb{F}$, $\omega(\mathtt{r}) = \omega^* \mathtt{r}/n^{1/2}$, $\delta(\mathtt{r}) = \delta^* \mathtt{r}/n^{1/2}$, and $\mathtt{g} = \mathtt{g}_1 \sqrt{n}$; see Lemma 5.1 in Spokoiny (2012). Here $\mathbb{F}$ is the Fisher information matrix of the family $(P_{\boldsymbol{\upsilon}})$ at the point $\boldsymbol{\upsilon}^*$, and $\omega^*$, $\delta^*$, and $\mathtt{g}_1$ are some positive constants.

It is shown in Spokoiny (2012) that the full parameter $\boldsymbol{\upsilon}^*$ can be well estimated provided that $p^*/n$ is sufficiently small. More precisely, the concentration property for the set $\Upsilon_\circ(\mathtt{r})$ requires $\mathtt{r}^2 \geq \mathtt{C} p^*$ for a fixed $\mathtt{C}$, while the local bracketing device is validated up to the spread $\Delta_{\boldsymbol{\epsilon}}(\mathtt{r})$ which is of order $p^* \delta(\mathtt{r}) \asymp p^* \mathtt{r}/n^{1/2} \asymp p^{*3/2}/n^{1/2}$. The range of applicability for the proposed approach can be informally defined by the rule "the spread is smaller than the value of the problem", where the value of the problem is understood as the expected excess. If the full parameter $\boldsymbol{\upsilon}$ is estimated, the value of the problem is of order $p^*$ leading to the constraint "$p^*/n$ is small". If the target parameter is of dimension $p$, then the value of the problem is also of order $p$ leading to the constraint "$p^{*3/2}/(n^{1/2} p)$ is small".

Now we specify the results in the $(\boldsymbol{\theta}, \boldsymbol{\eta})$ semiparametric setup. To state the result we only need a version of the identifiability condition $(\mathcal{I})$ on the marginal distribution. Let $\mathbb{F}$ be the Fisher information matrix of the family $(P_{\boldsymbol{\upsilon}})$ at the true point $\boldsymbol{\upsilon}^*$. Consider its block representation

$$\mathbb{F} = \begin{pmatrix} \mathbb{F}_{\boldsymbol{\theta}\boldsymbol{\theta}} & \mathbb{F}_{\boldsymbol{\theta}\boldsymbol{\eta}} \\ \mathbb{F}_{\boldsymbol{\theta}\boldsymbol{\eta}}^\top & \mathbb{F}_{\boldsymbol{\eta}\boldsymbol{\eta}} \end{pmatrix}.$$

The required identifiability condition reads as follows:

$(\boldsymbol{\iota})$ There is a constant $\rho < 1$ such that

$$\|\mathbb{F}_{\boldsymbol{\theta}\boldsymbol{\theta}}^{-1/2}\mathbb{F}_{\boldsymbol{\theta}\boldsymbol{\eta}}\mathbb{F}_{\boldsymbol{\eta}\boldsymbol{\eta}}^{-1}\mathbb{F}_{\boldsymbol{\theta}\boldsymbol{\eta}}^{\top}\mathbb{F}_{\boldsymbol{\theta}\boldsymbol{\theta}}^{-1/2}\|_{\infty} \leq \rho. \tag{2.6}$$

Also define

$$\breve{\mathbb{F}} \stackrel{\text{def}}{=} \mathbb{F}_{\boldsymbol{\theta}\boldsymbol{\theta}} - \mathbb{F}_{\boldsymbol{\theta}\boldsymbol{\eta}}\mathbb{F}_{\boldsymbol{\eta}\boldsymbol{\eta}}^{-1}\mathbb{F}_{\boldsymbol{\theta}\boldsymbol{\eta}}^{\top}.$$

The presented result admits that the full dimension $p^*$ grows with the sample size but slower than $n^{1/3}$. The result is applicable even in the case when the target dimension also depends on the sample size.

**Theorem 2.3.** *Let $Y_1, \ldots, Y_n$ be i.i.d. $I\!P_{\boldsymbol{\upsilon}^*}$ and let $(ed_0)$, $(ed_1)$, $(\ell_0)$, $(e\mathtt{u})$, and $(\ell\mathtt{u})$ with $\mathtt{b}(\mathtt{u}) \equiv \mathtt{b}$ of Spokoiny (2012) hold. In addition, assume $(\boldsymbol{\iota})$; see* (2.6)*. Define for $\mathtt{x} = \mathtt{x}_n \leq n^{1/3}$*

$$\beta_n \stackrel{\text{def}}{=} (p^* + \mathtt{x}_n)^{3/2}/n^{1/2}.$$

*It holds on the a set $\Omega(\mathtt{x}_n)$ of dominating probability:*

$$\begin{aligned} \big\|(n\breve{\mathbb{F}})^{1/2}\big(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\big) - \breve{\boldsymbol{\xi}}\big\|^2 &\leq \mathtt{C}\beta_n, \\ \big|2\breve{L}(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}^*) - \|\breve{\boldsymbol{\xi}}\|^2\big| &\leq \mathtt{C}\beta_n. \end{aligned}$$

*Moreover, the $p$-vector $\breve{\boldsymbol{\xi}} \stackrel{\text{def}}{=} \breve{\mathbb{F}}^{-1/2}\big(\nabla_{\boldsymbol{\theta}} - \mathbb{F}_{\boldsymbol{\theta}\boldsymbol{\eta}}\mathbb{F}_{\boldsymbol{\eta}\boldsymbol{\eta}}^{-1}\nabla_{\boldsymbol{\eta}}\big)$ is asymptotically standard normal as $n \to \infty$. This yields the asymptotic efficiency of the profile MLE $\widetilde{\boldsymbol{\theta}}$.*

## 2.3 Critical dimension

This section discusses the issue of a *critical dimension*. Namely we assume that the full dimension $p^*$ grows with the sample size $n$ and write $p^* = p_n$. Theorem 2.3 requires that $p_n = o(n^{1/3})$. Here we show that this condition is critical for the class of models satisfying the conditions of Section A. Namely, we present an example in which the behavior of the profile MLE $\widetilde{\boldsymbol{\theta}}$ heavily depends on the value $\beta_n = \sqrt{p_n^3/n} \geq \beta > 0$. The conditions of Section A are satisfied such that if $\beta_n \to 0$ we derive asymptotic efficiency of $\widetilde{\boldsymbol{\theta}}$. At the same time, if $\beta_n \geq \beta > 0$, then the MLE $\widetilde{\boldsymbol{\theta}}$ is not anymore root-n consistent. For technical reasons we assume that $p_n/\sqrt{n} \to 0$ which is no restriction as it is easy to find examples that do not admit an efficient profile if $p_n^2/n \nrightarrow 0$. Let a random vector $\boldsymbol{X} \in I\!R^{p_n}$ follow $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\upsilon}^*, n^{-1}\boldsymbol{I}_{p_n})$. Take for simplicity $\boldsymbol{\upsilon}^* = 0$ and let $I\!P = I\!P_0$ mean the distribution of $\boldsymbol{X}$. Introduce a special set $\mathcal{S} \subset I\!R^{p_n}$ with

$$\begin{aligned} \mathcal{S} \stackrel{\text{def}}{=} \Big\{\boldsymbol{\upsilon} = (\upsilon_1, \ldots, \upsilon_{p_n}) :\ \upsilon_1 = \frac{z}{2}\sqrt{\beta_n/n},\ z \in \mathbb{Z}\Big\} \\ \cap \Upsilon_{\circ}\left(\sqrt{2p_n/n} + \frac{1}{2}\sqrt{\beta_n/n}\right). \end{aligned} \tag{2.7}$$

We denote by $\mathcal{S}_\delta$ its $\delta$-vicinity:

$$\mathcal{S}_\delta \stackrel{\text{def}}{=} \{\boldsymbol{\upsilon} : \, d(\boldsymbol{\upsilon}, \mathcal{S}) < \delta\},$$

where $d(\boldsymbol{\upsilon}, \mathcal{S})$ is the Euclidean distance from the point $\boldsymbol{\upsilon}$ to the set $\mathcal{S}$. Also $\mathcal{S}_\delta^c$ stands for the complement of $\mathcal{S}_\delta$. Below we fix $\delta = 1/n$. Consider a special parametric quasi log-likelihood ratio $\mathcal{L}(\boldsymbol{\upsilon}, 0)$ defined as

$$\mathcal{L}(\boldsymbol{\upsilon}, 0) = n\boldsymbol{X}^\top \boldsymbol{\upsilon} - n\|\boldsymbol{\upsilon}\|^2/2 + nf(\boldsymbol{\upsilon})\|\boldsymbol{\upsilon}\|^3.$$

Here $f : I\!R \mapsto I\!R$ is a smooth function with

$$f(\boldsymbol{\upsilon}) = \begin{cases} 1 & \boldsymbol{\upsilon} \in \mathcal{S}, \\ 0 & \boldsymbol{\upsilon} \in \mathcal{S}_\delta^c. \end{cases}$$

Below we consider the problem of estimating the first component $\theta \stackrel{\text{def}}{=} \upsilon_1 \in I\!R$. Since by assumption $p_n/\sqrt{n} \to 0$ it holds for $n$ large enough and for any $\boldsymbol{\upsilon}$ with $\|\boldsymbol{\upsilon}\|^2 \leq 4p_n/n + \beta_n/n$ that $n\|\boldsymbol{\upsilon}\|^2/2 \geq nf(\boldsymbol{\upsilon})\|\boldsymbol{\upsilon}\|^3$ and thus

$$\operatorname*{argmax}_{\boldsymbol{\upsilon}} I\!E\mathcal{L}(\boldsymbol{\upsilon}) = \operatorname*{argmin}_{\boldsymbol{\upsilon}} \bigl\{ n\|\boldsymbol{\upsilon}\|^2/2 - nf(\boldsymbol{\upsilon})\|\boldsymbol{\upsilon}\|^3 \bigr\} = 0.$$

It is easy to see that all conditions from Section A are satisfied with $\tau_{\boldsymbol{\epsilon}} p_n \cong \beta_n^{1/2}$ and

$$\mathcal{D}_0^2 = \mathcal{V}_0^2 = nI_{p_n}.$$

Therefore, the results from Section 2.1 yield efficiency of the profile MLE $\widetilde{\theta}$ if $p_n^3/n \to 0$. Moreover, it is straightforward to see that

$$\breve{D}_0 = \sqrt{n}, \qquad \breve{\nabla}(\mathcal{L} - I\!E\mathcal{L}) = \nabla_{\boldsymbol{\theta}}(\mathcal{L} - I\!E\mathcal{L}) = nX_1, \text{ and } \breve{\boldsymbol{\xi}} = \sqrt{n}X_1.$$

It follows similarly to Theorem 2.1 that if $\beta_n^2 = p_n^3/n \to 0$ then

$$\|\breve{D}_0(\widetilde{\theta} - \theta^*) - \breve{\boldsymbol{\xi}}\| = \sqrt{n}|\widetilde{\upsilon}_1 - X_1| \to 0.$$

The next result shows that in the case when $\beta_n = \sqrt{p_n^3/n}$ is not small, the profile MLE $\widetilde{\theta}$ is not root-$n$ consistent.

**Theorem 2.4.** *Suppose that $\beta_n \to (6c)^2$ for some $c > 0$. Let also $n$ be large enough to ensure*

$$\frac{2^{1/3} - 1}{2^{1/6}} \sqrt{p_n/n} \geq \frac{1}{2} \left(p_n/n\right)^{3/4}.$$

*There exists a positive $\alpha > 0$ such that it holds with a probability exceeding $\alpha$*

$$\|\breve{D}_0(\widetilde{\theta} - \theta^*) - \breve{\boldsymbol{\xi}}\| \geq \frac{1}{6}\beta_n^{1/2} - \frac{1}{\sqrt{n}} \geq c - o_n(1).$$

*If* $\beta_n \to \infty$ *, then*

$$\|\breve{D}_0(\widetilde{\theta} - \theta^*) - \breve{\boldsymbol{\xi}}\| \xrightarrow{I\!P} +\infty,$$

*where* $\xrightarrow{I\!P}$ *means convergence in probability.*

# A Appendix

The appendix collects our conditions and proofs of the main results.

We adopt the conditions from Section 2 of Spokoiny (2012) with the obvious change of notations. The local conditions only describe the properties of the process $\mathcal{L}(\boldsymbol{\upsilon})$ for $\boldsymbol{\upsilon} \in \Upsilon_\circ(\mathtt{r}_0)$ with some fixed value $\mathtt{r}_0$. The global conditions have to be fulfilled on the whole $\Upsilon$. We start with the local conditions.

$(\boldsymbol{\mathcal{ED}_0})$ There exists a constant $\nu_0 > 0$, a positive symmetric $p^* \times p^*$ matrix $\mathcal{V}_0^2$ satisfying $\mathrm{Var}\{\nabla\zeta(\boldsymbol{\upsilon}^*)\} \le \mathcal{V}_0^2$, and a constant $\mathtt{g} > 0$ such that for all $|\mu| \le \mathtt{g}$

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

$(\boldsymbol{\mathcal{ED}_1})$ For all $0 < \mathtt{r} < \mathtt{r}_0$, there exists a constant $\omega(\mathtt{r}) \le 1/2$ such that for all $\boldsymbol{\upsilon} \in \Upsilon_\circ(\mathtt{r})$ and $|\mu| \le \mathtt{g}$

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

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

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

This condition together with the identity $\nabla I\!E\mathcal{L}(\boldsymbol{\upsilon}^*) = 0$ implies

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

The global conditions are:

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

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

$(\mathcal{E}\mathtt{r})$ For any $\mathtt{r} \geq \mathtt{r}_0$ there exists a constant $\nu_0 > 0$ and a constant $\mathtt{g}(\mathtt{r}) > 0$ such that

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

Our results are stated for $\mathtt{g}(\mathtt{r}) \equiv \mathtt{g} > 0$, however, an extension to the case $\mathtt{g}(\mathtt{r}) \to 0$ can be made similarly to Spokoiny (2012).

Finally we specify the regularity conditions. We begin by representing the information and the covariance matrices in block form:

$$\mathcal{D}_0^2 = \begin{pmatrix} D_0^2 & A_0 \\ A_0^\top & H_0^2 \end{pmatrix}, \qquad \mathcal{V}_0^2 = \begin{pmatrix} V_0^2 & B_0 \\ B_0^\top & Q_0^2 \end{pmatrix}.$$

The *identifiability conditions* in Spokoiny (2012) ensure that the matrix $\mathcal{D}_0$ is positive and satisfies $\mathfrak{a}^2 \mathcal{D}_0^2 \geq \mathcal{V}_0^2$ for some $\mathfrak{a} > 0$. Here we restate these conditions in the special block form which is specific for the $(\boldsymbol{\theta}, \boldsymbol{\eta})$ -setup.

$(\boldsymbol{\mathcal{I}})$ There are constants $\mathfrak{a} > 0$ and $\rho < 1$ such that

$$\mathfrak{a}^2 D_0^2 \geq V_0^2, \quad \mathfrak{a}^2 H_0^2 \geq Q_0^2, \quad \mathfrak{a}^2 \mathcal{D}_0^2 \geq \mathcal{V}_0^2. \tag{A.1}$$

and

$$\| D_0^{-1} A_0 H_0^{-2} A_0^\top D_0^{-1} \|_\infty \leq \rho. \tag{A.2}$$

The quantity $\rho$ bounds the angle between the target and nuisance subspaces in the tangent space. The regularity condition $(\mathcal{I})$ ensures that this angle is not too small and hence, the target and nuisance parameters are identifiable. In particular, the matrix $\breve{D}_0^2$ is well posed under $\mathcal{I}$.

The bounds in (A.1) are given with the same constant $\mathfrak{a}$ only for simplifying the notation. One can show that the last bound on $\mathcal{D}_0^2$ follows from the first two and (A.2) with another constant $\mathfrak{a}'$ depending on $\mathfrak{a}$ and $\rho$ only.

# B Bracketing and upper function devices

This section briefly overviews the main constructions of Spokoiny (2012) including the bracketing bound and the upper function results. The bracketing bound describes the quality of quadratic approximation of the log-likelihood process $\mathcal{L}(\boldsymbol{\upsilon})$ in a local vicinity of the point $\boldsymbol{\upsilon}^*$, while the upper function method is used to show that the full MLE $\widetilde{\boldsymbol{\upsilon}}$ belongs to this vicinity with a dominating probability. Given $\mathtt{r} > 0$, define the local set

$$\Upsilon_\circ(\mathtt{r}) \stackrel{\text{def}}{=} \bigl\{ \boldsymbol{\upsilon} : \, (\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^*)^\top \mathcal{V}_0^2 (\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^*) \leq \mathtt{r}^2 \bigr\}.$$

For $\boldsymbol{\epsilon} = (\delta, \varrho)$, define the bracketing quadratic processes $\mathbb{L}_{\boldsymbol{\epsilon}}(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^*)$ and $\mathbb{L}_{\underline{\boldsymbol{\epsilon}}}(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^*)$:

$$\begin{aligned} \mathbb{L}_{\boldsymbol{\epsilon}}(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^*) &\stackrel{\mathrm{def}}{=} (\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^*)^\top \nabla \mathcal{L}(\boldsymbol{\upsilon}^*) - \|\mathcal{D}_{\boldsymbol{\epsilon}}(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^*)\|^2/2, \\ \mathcal{D}^2_{\boldsymbol{\epsilon}} &\stackrel{\mathrm{def}}{=} \mathcal{D}_0^2(1-\delta) - \varrho \mathcal{V}_0^2, \end{aligned} \tag{B.1}$$

and accordingly for $\underline{\boldsymbol{\epsilon}} = -\boldsymbol{\epsilon} = (-\delta, -\varrho)$. The next result restates the local bracketing bound of Spokoiny (2012) in the semiparametric framework. The imposed conditions and the involved constants $\nu_0$, $\delta(\mathtt{r})$, and $\omega(\mathtt{r})$ are explained in Section A. The presented results implicitly assume that $p^*$ is large, $\mathtt{x}$ is large as well to ensure that $\mathrm{e}^{-\mathtt{x}}$ is negligible. A proper choice is $\mathtt{x} = \mathtt{C}p^*$ for a fixed $\mathtt{C}$.

**Theorem B.1** (Spokoiny (2012), Theorem 3.1). *Assume $(\mathcal{ED}_1)$ and $(\mathcal{L}_0)$. Let for some $\mathtt{r}$, the values $\varrho \geq 3\nu_0\,\omega(\mathtt{r})$ and $\delta \geq \delta(\mathtt{r})$ be such that $\mathcal{D}^2(1-\delta) - \varrho\mathcal{V}_0^2 \geq 0$. Then*

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

*where the random variables $\diamondsuit_{\boldsymbol{\epsilon}}(\mathtt{r}), \diamondsuit_{\underline{\boldsymbol{\epsilon}}}(\mathtt{r})$ fulfill on a random set $\Omega(\mathtt{x})$ of dominating probability*

$$\diamondsuit_{\boldsymbol{\epsilon}}(\mathtt{r}) \leq \mathtt{C}\,\varrho\,(p^* + \mathtt{x}), \qquad \diamondsuit_{\underline{\boldsymbol{\epsilon}}}(\mathtt{r}) \leq \mathtt{C}\,\varrho\,(p^* + \mathtt{x}). \tag{B.2}$$

In fact, Theorem 3.1 of Spokoiny (2012) states the following bound:

$$I\!P\big\{\varrho^{-1}\diamondsuit_{\boldsymbol{\epsilon}}(\mathtt{r}) \geq \mathfrak{z}(\mathbb{Q}, \mathtt{x})\big\} \leq \exp\big(-\mathtt{x}\big).$$

with $\mathbb{Q} = 2.4p^*$ and

$$\mathfrak{z}(\mathbb{Q}, \mathtt{x}) = \begin{cases} \big(1 + \sqrt{\mathtt{x} + \mathbb{Q}}\big)^2 & \text{if } 1 + \sqrt{\mathtt{x} + \mathbb{Q}} < \frac{\mathtt{g}}{\nu_0} \\ \big\{1 + \frac{\nu_0}{\mathtt{g}}(\mathtt{x} + \mathbb{Q}) + \frac{\mathtt{g}}{2\nu_0}\big\}^2 & \text{otherwise.} \end{cases}$$

Under the assumption that $\mathtt{g}$ is sufficiently large, that is, $\mathtt{g}/\nu_0 \gg p^*$, we can apply $\mathfrak{z}(\mathbb{Q}, \mathtt{x}) \approx \mathtt{x} + \mathbb{Q} \leq \mathtt{C}(p^* + \mathtt{x})$, and the result of Theorem B.1 follows.

The bracketing result of Theorem B.1 is local in the sense that it only applies for $\boldsymbol{\upsilon} \in \Upsilon_\circ(\mathtt{r})$. Following to the general approach of Spokoiny (2012) we accompany it with the large deviation bound on the concentration probability $I\!P\big(\widetilde{\boldsymbol{\upsilon}} \in \Upsilon_\circ(\mathtt{r})\big)$ when the local radius $\mathtt{r}$ exceeds some level $\mathtt{r}_0$ which has to be sufficiently large, namely $\mathtt{r}_0^2 \geq \mathtt{C}p^*$. We adopt the upper function approach from Spokoiny (2012); cf. Corollary 4.4 therein. Again the constants $\mathtt{g}(\mathtt{r})$ and $\mathtt{b}(\mathtt{r})$ are introduced in Section A.

**Theorem B.2** (Spokoiny (2012), Theorem 4.1). *Suppose $(\mathcal{E}\mathtt{r})$ and $(\mathcal{L}\mathtt{r})$ with $\mathtt{b}(\mathtt{r}) \equiv \mathtt{b}$. If for $\mathtt{r} \geq \mathtt{r}_0$, the following conditions are fulfilled:*

$$\begin{aligned} 1 + \sqrt{\mathtt{x} + \mathbb{Q}} &\leq 3\nu_0^2 \mathtt{g}(\mathtt{r})/\mathtt{b}, \\ 6\nu_0\sqrt{\mathtt{x} + \mathbb{Q}} &\leq \mathtt{r}\mathtt{b}, \end{aligned} \tag{B.3}$$

*then* $\widetilde{\boldsymbol{\upsilon}} \in \Upsilon_\circ(\mathtt{r}_0)$ *on a random set* $\Omega(\mathtt{x})$ *of dominating probability. The same bound holds for the probability* $\widetilde{\boldsymbol{\upsilon}}_{\boldsymbol{\theta}^*} \in \Upsilon_\circ(\mathtt{r}_0)$ *where* $\widetilde{\boldsymbol{\upsilon}}_{\boldsymbol{\theta}^*}$ *maximizes* $\mathcal{L}(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^*)$ *subject to* $\Pi_0 \boldsymbol{\upsilon} = \boldsymbol{\theta}^*$ *:*

$$\widetilde{\boldsymbol{\upsilon}}_{\boldsymbol{\theta}^*} \stackrel{\text{def}}{=} \operatorname*{argmax}_{\substack{\boldsymbol{\upsilon} \in \Upsilon \\ \Pi_0 \boldsymbol{\upsilon} = \boldsymbol{\theta}^*}} \mathcal{L}(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^*).$$

**Remark B.1.** The condition (B.3) helps to understand which $\mathtt{r}_0$ ensures prescribed concentration properties of $\widetilde{\boldsymbol{\upsilon}}$ and $\widetilde{\boldsymbol{\upsilon}}_{\boldsymbol{\theta}^*}$. Namely, if $\mathtt{g}(\mathtt{r})$ is large enough, then (B.3) follows from the bound

$$\mathtt{r}_0 \geq 6\mathtt{b}^{-1}\nu_0\sqrt{\mathtt{x} + \mathbb{Q}}.$$

# C Deviation bounds for quadratic forms

The following general result from Spokoiny (2013) helps to control the deviation for quadratic forms of type $\|I\!B\boldsymbol{\xi}\|^2$ for a given positive matrix $I\!B$ and a random vector $\boldsymbol{\xi}$. It will be used several times in our proofs. Suppose that

$$\log I\!E \exp\bigl(\boldsymbol{\gamma}^\top \boldsymbol{\xi}\bigr) \leq \|\boldsymbol{\gamma}\|^2/2, \qquad \boldsymbol{\gamma} \in I\!R^p,\ \|\boldsymbol{\gamma}\| \leq \mathtt{g}. \tag{C.1}$$

For a symmetric matrix $I\!B$, define

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

We suppose that $\lambda^* \leq 1$, otherwise we should replace everywhere $I\!B$ with $I\!B/\lambda^*$.

Let $\mathtt{g}$ be shown in (C.1). Define $\omega_c$ by the equation

$$\frac{\omega_c(1+\omega_c)}{(1+\omega_c^2)^{1/2}} = \mathtt{g}\mathtt{p}^{-1/2}.$$

Define also $\mu_c = \omega_c^2/(1+\omega_c^2) \wedge 2/3$. Note that $\omega_c^2 \geq 2$ implies $\mu_c = 2/3$. Further define

$$\mathtt{y}_c^2 = (1+\omega_c^2)\mathtt{p}, \qquad 2\mathtt{x}_c = \mu_c \mathtt{y}_c^2 + \log\det\{I_p - \mu_c I\!B^2\}. \tag{C.2}$$

**Theorem C.1** (Spokoiny (2013))**.** *Let* $\boldsymbol{\xi}$ *fulfill* (C.1) *with* $\mathtt{g}^2 \geq 2\mathtt{p}$*. Then we have for* $\mathtt{x} \leq \mathtt{x}_c$ *with* $\mathtt{x}_c$ *from* (C.2)*:*

$$I\!P\bigl(\|I\!B\boldsymbol{\xi}\|^2 \geq \mathfrak{z}(\mathtt{x}, I\!B)\bigr) \leq 2\mathrm{e}^{-\mathtt{x}} + 8.4\mathrm{e}^{-\mathtt{x}_c},$$

$$\mathfrak{z}(\mathtt{x}, I\!B) \stackrel{\text{def}}{=} \begin{cases} \mathtt{p} + 2\mathtt{v}\mathtt{x}^{1/2}, & \mathtt{x} \leq \mathtt{v}/18, \\ \mathtt{p} + 6\mathtt{x} & \mathtt{v}/18 < \mathtt{x} \leq \mathtt{x}_c. \end{cases}$$

*For* $\mathtt{x} > \mathtt{x}_c$

$$I\!P\bigl(\|I\!B\boldsymbol{\xi}\|^2 \geq \mathfrak{z}_c(\mathtt{x}, I\!B)\bigr) \leq 8.4\mathrm{e}^{-\mathtt{x}}, \qquad \mathfrak{z}_c(\mathtt{x}, I\!B) \stackrel{\text{def}}{=} \bigl|\mathtt{y}_c + 2(\mathtt{x} - \mathtt{x}_c)/\mathtt{g}_c\bigr|^2.$$

It appears that the bound is slightly different in two zones separated by some specific value $\mathtt{x}_c$ from (C.2). It is large in typical situations as $\mathtt{x}_c \cong \mathtt{g}$ (it is of order $\sqrt{n}$ in the i.i.d. case). For $\mathtt{x} \le \mathtt{x}_c$, we obtain the same type of bounds as in the Gaussian case, for $\mathtt{x} > \mathtt{x}_c$ they are a bit worse.

# D Proofs

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

## D.1 Proof of Theorem 2.1

Define the $m \times m$ matrices $H^2_{\boldsymbol{\epsilon}}$ and $H^2_{\underline{\boldsymbol{\epsilon}}}$ by

$$H^2_{\boldsymbol{\epsilon}} = H^2_0(1-\delta) - \varrho Q^2_0, \qquad H^2_{\underline{\boldsymbol{\epsilon}}} = H^2_0(1+\delta) + \varrho Q^2_0;$$

cf. (B.1). Below we fix some constant $\mathtt{r}$ which is assumed to be large enough for ensuring the dominating probability for the concentration event $C_{\boldsymbol{\epsilon}}(\mathtt{r})$ defined as

$$\begin{aligned} C_{\boldsymbol{\epsilon}}(\mathtt{r}) \stackrel{\text{def}}{=} \big\{ \, & \|\mathcal{V}_0(\widetilde{\boldsymbol{\upsilon}} - \boldsymbol{\upsilon}^*)\| \le \mathtt{r}, \, \|\mathcal{V}_0(\widetilde{\boldsymbol{\upsilon}}_{\boldsymbol{\theta}^*} - \boldsymbol{\upsilon}^*)\| \le \mathtt{r}, \\ & \|\mathcal{V}_0 \mathcal{D}^{-2}_{\underline{\boldsymbol{\epsilon}}} \nabla\| \le \mathtt{r}, \quad \|Q_0 H^{-2}_{\underline{\boldsymbol{\epsilon}}} \nabla_{\boldsymbol{\eta}}\| \le \mathtt{r} \big\}. \end{aligned} \tag{D.1}$$

Note that the conditions $\|\mathcal{V}_0(\widetilde{\boldsymbol{\upsilon}} - \boldsymbol{\upsilon}^*)\| \le \mathtt{r}$ and $\|\mathcal{V}_0(\widetilde{\boldsymbol{\upsilon}}_{\boldsymbol{\theta}^*} - \boldsymbol{\upsilon}^*)\| \le \mathtt{r}$ can be represented as $\{\widetilde{\boldsymbol{\upsilon}} \in \Upsilon_\circ(\mathtt{r})\}$ and $\{\widetilde{\boldsymbol{\upsilon}}_{\boldsymbol{\theta}^*} \in \Upsilon_\circ(\mathtt{r})\}$. Similar representation holds for

$$\widetilde{\boldsymbol{\upsilon}}_{\underline{\boldsymbol{\epsilon}}} \stackrel{\text{def}}{=} \mathcal{D}^{-2}_{\underline{\boldsymbol{\epsilon}}} \nabla = \operatorname*{argmin}_{\boldsymbol{\upsilon}} \mathbb{L}_{\underline{\boldsymbol{\epsilon}}}(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^*),$$

$$\widetilde{\boldsymbol{\eta}}_{\underline{\boldsymbol{\epsilon}}} \stackrel{\text{def}}{=} H^{-2}_{\underline{\boldsymbol{\epsilon}}} \nabla_{\boldsymbol{\eta}} = \operatorname*{argmin}_{\substack{\boldsymbol{\upsilon} \in \Upsilon \\ \Pi_0 \boldsymbol{\upsilon} = \boldsymbol{\theta}^*}} \mathbb{L}_{\underline{\boldsymbol{\epsilon}}}(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^*).$$

For instance, $\big\{\|\mathcal{V}_0 \mathcal{D}^{-2}_{\underline{\boldsymbol{\epsilon}}} \nabla\| \le \mathtt{r}\big\} = \big\{\widetilde{\boldsymbol{\upsilon}}_{\underline{\boldsymbol{\epsilon}}} \in \Upsilon_\circ(\mathtt{r})\big\}$. Later we show that a proper choice of $\mathtt{r}$ ensures a dominating probability of the random set $C_{\boldsymbol{\epsilon}}(\mathtt{r})$; see Section D.1.

We first show that the bound (2.4) is fulfilled on the set $C_{\boldsymbol{\epsilon}}(\mathtt{r})$ from (D.1) with

$$\Delta^+_{\boldsymbol{\epsilon}}(\mathtt{r}) = \diamondsuit_{\boldsymbol{\epsilon}}(\mathtt{r}) + \diamondsuit_{\underline{\boldsymbol{\epsilon}}}(\mathtt{r}) + \frac{\tau_{\boldsymbol{\epsilon}}}{1-\tau_{\boldsymbol{\epsilon}}} \big\|\mathcal{D}^{-1}_0 \nabla\big\|^2 + \frac{\tau_{\boldsymbol{\epsilon}}}{1+\tau_{\boldsymbol{\epsilon}}} \big\|H^{-1}_0 \nabla_{\boldsymbol{\eta}}\big\|^2, \tag{D.2}$$

$$\Delta^-_{\boldsymbol{\epsilon}}(\mathtt{r}) = \diamondsuit_{\boldsymbol{\epsilon}}(\mathtt{r}) + \diamondsuit_{\underline{\boldsymbol{\epsilon}}}(\mathtt{r}) + \frac{\tau_{\boldsymbol{\epsilon}}}{1+\tau_{\boldsymbol{\epsilon}}} \big\|\mathcal{D}^{-1}_0 \nabla\big\|^2 + \frac{\tau_{\boldsymbol{\epsilon}}}{1-\tau_{\boldsymbol{\epsilon}}} \big\|H^{-1}_0 \nabla_{\boldsymbol{\eta}}\big\|^2. \tag{D.3}$$

In analogy with Spokoiny (2012), the quantity $\Delta_{\boldsymbol{\epsilon}}(\mathtt{r})$ with

$$\begin{aligned} \Delta_{\boldsymbol{\epsilon}}(\mathtt{r}) &= \Delta^+_{\boldsymbol{\epsilon}}(\mathtt{r}) + \Delta^-_{\boldsymbol{\epsilon}}(\mathtt{r}) \\ &= 2\diamondsuit_{\boldsymbol{\epsilon}}(\mathtt{r}) + 2\diamondsuit_{\underline{\boldsymbol{\epsilon}}}(\mathtt{r}) + \frac{2\tau_{\boldsymbol{\epsilon}}}{1-\tau^2_{\boldsymbol{\epsilon}}} \left( \big\|\mathcal{D}^{-1}_0 \nabla\big\|^2 + \big\|H^{-1}_0 \nabla_{\boldsymbol{\eta}}\big\|^2 \right), \end{aligned} \tag{D.4}$$

can be called the *semiparametric spread*. It can be seen as a payment for the bracketing device. Below we show that $\Delta_{\boldsymbol{\epsilon}}(\mathtt{r}) \leq \mathtt{C}\,\tau_{\boldsymbol{\epsilon}}\,(p^* + \mathtt{x})$ with a dominating probability.

We start with some technical results about the maximum of the upper and lower quadratic processes $\mathbb{L}_{\boldsymbol{\epsilon}}(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^*)$ and $\mathbb{L}_{\underline{\boldsymbol{\epsilon}}}(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^*)$. Remind the notation $\nabla \stackrel{\text{def}}{=} \nabla \mathcal{L}(\boldsymbol{\upsilon}^*)$.

**Lemma D.1.** *It holds*

$$\sup_{\boldsymbol{\upsilon}} \mathbb{L}_{\boldsymbol{\epsilon}}(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^*) = \frac{1}{2}\bigl\|\mathcal{D}_{\boldsymbol{\epsilon}}^{-1}\nabla\bigr\|^2, \quad \sup_{\boldsymbol{\upsilon}} \mathbb{L}_{\underline{\boldsymbol{\epsilon}}}(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^*) = \frac{1}{2}\bigl\|\mathcal{D}_{\underline{\boldsymbol{\epsilon}}}^{-1}\nabla\bigr\|^2, \tag{D.5}$$

*where* $\sup_{\boldsymbol{\upsilon}}$ *means the maximum over all* $\boldsymbol{\upsilon} \in I\!R^{p^*}$. *Moreover, on the random set* $\bigl\{\|\mathcal{V}_0\mathcal{D}_0^{-2}\nabla\| \leq \mathtt{r}\bigr\}$ *it holds*

$$\sup_{\boldsymbol{\upsilon} \in \Upsilon_\circ(\mathtt{r})} \mathbb{L}_{\underline{\boldsymbol{\epsilon}}}(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^*) = \sup_{\boldsymbol{\upsilon}} \mathbb{L}_{\underline{\boldsymbol{\epsilon}}}(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^*) = \frac{1}{2}\bigl\|\mathcal{D}_{\underline{\boldsymbol{\epsilon}}}^{-1}\nabla\bigr\|^2.$$

*Proof.* The identity (D.5) directly follows by maximizing the quadratic expression

$$\mathbb{L}_{\boldsymbol{\epsilon}}(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^*) = (\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^*)^\top \nabla - \|\mathcal{D}_{\boldsymbol{\epsilon}}(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^*)\|^2/2,$$

with the maximum at $\boldsymbol{\upsilon} = \boldsymbol{\upsilon}^* + \mathcal{D}_{\boldsymbol{\epsilon}}^{-2}\nabla$. Similarly, the maximum of $\mathbb{L}_{\underline{\boldsymbol{\epsilon}}}(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^*)$ is achieved at $\boldsymbol{\upsilon} = \boldsymbol{\upsilon}^* + \mathcal{D}_{\underline{\boldsymbol{\epsilon}}}^{-2}\nabla \in C_{\boldsymbol{\epsilon}}(\mathtt{r})$ which is within $\Upsilon_\circ(\mathtt{r})$ under the condition

$$\|\mathcal{V}_0\mathcal{D}_{\underline{\boldsymbol{\epsilon}}}^{-2}\nabla\| \leq \|\mathcal{V}_0\mathcal{D}_0^{-2}\nabla\| \leq \mathtt{r}.$$

This yields the claim. □

The next lemma states similar results for the constrained maximum of $\mathbb{L}_{\boldsymbol{\epsilon}}$ and $\mathbb{L}_{\underline{\boldsymbol{\epsilon}}}$ subject to $\Pi_0\boldsymbol{\upsilon} = \boldsymbol{\theta}^*$. The proof is the same as for Lemma D.1. Remember the notation $\nabla_{\boldsymbol{\theta}} \stackrel{\text{def}}{=} \nabla_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\upsilon}^*)$, $\nabla_{\boldsymbol{\eta}} \stackrel{\text{def}}{=} \nabla_{\boldsymbol{\eta}}\mathcal{L}(\boldsymbol{\upsilon}^*)$. We also use the block representation of $\mathcal{D}_0^2$:

$$\mathcal{D}_0^2 = \begin{pmatrix} D_0^2 & A_0 \\ A_0^\top & H_0^2 \end{pmatrix}.$$

**Lemma D.2.** *It holds*

$$\sup_{\boldsymbol{\upsilon}:\, \Pi_0\boldsymbol{\upsilon} = \boldsymbol{\theta}^*} \mathbb{L}_{\boldsymbol{\epsilon}}(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^*) = \frac{1}{2}\bigl\|H_{\boldsymbol{\epsilon}}^{-1}\nabla_{\boldsymbol{\eta}}\bigr\|^2, \tag{D.7}$$

*Moreover, it holds on the random set* $\bigl\{\|Q_0H_0^{-2}\nabla_{\boldsymbol{\eta}}\| \leq \mathtt{r}\bigr\}$

$$\sup_{\boldsymbol{\upsilon} \in \Upsilon_\circ(\mathtt{r}):\, \Pi_0\boldsymbol{\upsilon} = \boldsymbol{\theta}^*} \mathbb{L}_{\underline{\boldsymbol{\epsilon}}}(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^*) = \sup_{\boldsymbol{\upsilon}:\, \Pi_0\boldsymbol{\upsilon} = \boldsymbol{\theta}^*} \mathbb{L}_{\underline{\boldsymbol{\epsilon}}}(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^*) = \frac{1}{2}\bigl\|H_{\underline{\boldsymbol{\epsilon}}}^{-1}\nabla_{\boldsymbol{\eta}}\bigr\|^2.$$

Further, define the process

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

Remember the definition of $\breve{\nabla}_{\boldsymbol{\theta}}$ and $\breve{D}_0^2$:

$$\breve{\nabla}_{\boldsymbol{\theta}} \stackrel{\mathrm{def}}{=} \nabla_{\boldsymbol{\theta}} - A_0 H_0^{-2} \nabla_{\boldsymbol{\eta}},$$
$$\breve{D}_0^2 \stackrel{\mathrm{def}}{=} D_0^2 - A_0 H_0^{-2} A_0^\top.$$

**Lemma D.3.** *It holds on the random set* $\bigl\{ \|\mathcal{V}_0 \mathcal{D}_0^{-2} \nabla\| \le \mathtt{r},\ \|Q_0 H_0^{-2} \nabla_{\boldsymbol{\eta}}\| \le \mathtt{r} \bigr\}$

$$\sup_{\boldsymbol{\upsilon}} \mathbb{L}(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^*) = \sup_{\boldsymbol{\upsilon} \in \Upsilon_\circ(\mathtt{r})} \mathbb{L}(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^*) = \frac{1}{2} \bigl\| \mathcal{D}_0^{-1} \nabla \bigr\|^2,$$
$$\sup_{\boldsymbol{\upsilon} \in \Upsilon_\circ(\mathtt{r}):\, \Pi_0 \boldsymbol{\upsilon} = \boldsymbol{\theta}^*} \mathbb{L}(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^*) = \sup_{\boldsymbol{\upsilon}:\, \Pi_0 \boldsymbol{\upsilon} = \boldsymbol{\theta}^*} \mathbb{L}(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^*) = \frac{1}{2} \bigl\| H_0^{-1} \nabla_{\boldsymbol{\eta}} \bigr\|^2,$$
$$\sup_{\boldsymbol{\upsilon}} \mathbb{L}(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^*) - \sup_{\boldsymbol{\upsilon}:\, \Pi_0 \boldsymbol{\upsilon} = \boldsymbol{\theta}^*} \mathbb{L}(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^*) = \frac{1}{2} \bigl\| \breve{D}_0^{-1} \breve{\nabla}_{\boldsymbol{\theta}} \bigr\|^2. \tag{D.8}$$

*Proof.* First consider the adaptive cases with $A_0 = 0$ yielding $\breve{D}_0^2 = D_0^2$ and $\breve{\nabla}_{\boldsymbol{\theta}} = \nabla_{\boldsymbol{\theta}}$. Then the process $\mathbb{L}(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^*)$ can be decomposed as

$$\mathbb{L}(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^*) = (\boldsymbol{\theta} - \boldsymbol{\theta}^*)^\top \nabla_{\boldsymbol{\theta}} - \frac{1}{2} \|D_0 (\boldsymbol{\theta} - \boldsymbol{\theta}^*)\|^2$$
$$+ (\boldsymbol{\eta} - \boldsymbol{\eta}^*)^\top \nabla_{\boldsymbol{\eta}} - \frac{1}{2} \|H_0 (\boldsymbol{\eta} - \boldsymbol{\eta}^*)\|^2,$$

and the partial optimization subject to $\boldsymbol{\theta} = \boldsymbol{\theta}^*$ yields the results (D.6) and (D.7). Note that the constrained maximum is attained at $\boldsymbol{\eta} = \boldsymbol{\eta}^* + H_0^{-2} \nabla_{\boldsymbol{\eta}}$.

The general case can be reduced to the adaptive one by the change of variable. With $\boldsymbol{\gamma} \stackrel{\mathrm{def}}{=} \boldsymbol{\eta} - \boldsymbol{\eta}^* + H_0^{-2} A_0^\top (\boldsymbol{\theta} - \boldsymbol{\theta}^*)$, one can represent $\mathbb{L}(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^*)$ in the form

$$\mathbb{L}(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^*) = (\boldsymbol{\theta} - \boldsymbol{\theta}^*)^\top \breve{\nabla}_{\boldsymbol{\theta}} - \|\breve{D}_0 (\boldsymbol{\theta} - \boldsymbol{\theta}^*)\|^2 / 2 + \boldsymbol{\gamma}^\top \nabla_{\boldsymbol{\eta}} - \|H_0 \boldsymbol{\gamma}\|^2 / 2,$$

which corresponds to the decomposition in the adaptive case. □

On the random set $\{\widetilde{\boldsymbol{\upsilon}} \in \Upsilon_\circ(\mathtt{r}), \widetilde{\boldsymbol{\upsilon}}_{\boldsymbol{\theta}^*} \in \Upsilon_\circ(\mathtt{r})\}$, it holds

$$\breve{L}(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}^*) = \breve{L}(\widetilde{\boldsymbol{\theta}}) - \breve{L}(\boldsymbol{\theta}^*) = \sup_{\boldsymbol{\upsilon} \in \Upsilon_\circ(\mathtt{r})} \mathcal{L}(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^*) - \sup_{\boldsymbol{\upsilon} \in \Upsilon_\circ(\mathtt{r}):\, \Pi_0 \boldsymbol{\upsilon} = \boldsymbol{\theta}^*} \mathcal{L}(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^*).$$

Theorem B.1 implies

$$\sup_{\boldsymbol{\upsilon} \in \Upsilon_\circ(\mathtt{r})} \mathbb{L}_{\underline{\boldsymbol{\epsilon}}}(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^*) - \diamondsuit_{\underline{\boldsymbol{\epsilon}}}(\mathtt{r}) \le \sup_{\boldsymbol{\upsilon} \in \Upsilon_\circ(\mathtt{r})} \mathcal{L}(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^*) \le \sup_{\boldsymbol{\upsilon} \in \Upsilon_\circ(\mathtt{r})} \mathbb{L}_{\boldsymbol{\epsilon}}(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^*) + \diamondsuit_{\boldsymbol{\epsilon}}(\mathtt{r}). \tag{D.9}$$

The same bound applies to the maximum taken over the subset $\{\boldsymbol{\upsilon} \in \Upsilon_\circ(\mathtt{r}) : \Pi_0 \boldsymbol{\upsilon} = \boldsymbol{\theta}^*\}$. By Lemmas D.1 and D.2, on the random set $C_{\boldsymbol{\epsilon}}(\mathtt{r})$, one can replace the sup of $\mathbb{L}_{\underline{\boldsymbol{\epsilon}}}(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^*)$ over

$\Upsilon_\circ(\mathtt{r})$ by the sup over the whole vector space $I\!R^{p^*}$. Putting all the obtained bounds together yields

$$\begin{aligned}
\breve{L}(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}^*) &\geq \sup_{\boldsymbol{\upsilon}} \mathbb{L}_{\underline{\boldsymbol{\epsilon}}}(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^*) - \sup_{\boldsymbol{\upsilon}:\, \Pi_0 \boldsymbol{\upsilon} = \boldsymbol{\theta}^*} \mathbb{L}_{\boldsymbol{\epsilon}}(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^*) - \diamondsuit_{\boldsymbol{\epsilon}}(\mathtt{r}) - \diamondsuit_{\underline{\boldsymbol{\epsilon}}}(\mathtt{r}), \\
\breve{L}(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}^*) &\leq \sup_{\boldsymbol{\upsilon}} \mathbb{L}_{\boldsymbol{\epsilon}}(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^*) - \sup_{\boldsymbol{\upsilon}:\, \Pi_0 \boldsymbol{\upsilon} = \boldsymbol{\theta}^*} \mathbb{L}_{\underline{\boldsymbol{\epsilon}}}(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^*) + \diamondsuit_{\boldsymbol{\epsilon}}(\mathtt{r}) + \diamondsuit_{\underline{\boldsymbol{\epsilon}}}(\mathtt{r}).
\end{aligned} \tag{D.10}$$

Define

$$\begin{aligned}
\square_{\boldsymbol{\epsilon}} &\stackrel{\text{def}}{=} \sup_{\boldsymbol{\upsilon}} \mathbb{L}_{\boldsymbol{\epsilon}}(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^*) - \sup_{\boldsymbol{\upsilon}} \mathbb{L}(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^*), \\
\square_{\underline{\boldsymbol{\epsilon}}} &\stackrel{\text{def}}{=} \sup_{\boldsymbol{\upsilon}} \mathbb{L}(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^*) - \sup_{\boldsymbol{\upsilon}} \mathbb{L}_{\underline{\boldsymbol{\epsilon}}}(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^*).
\end{aligned}$$

Lemmas D.1 implies

$$\begin{aligned}
2\,\square_{\boldsymbol{\epsilon}} &= \bigl\| \mathcal{D}_{\boldsymbol{\epsilon}}^{-1} \nabla \bigr\|^2 - \bigl\| \mathcal{D}_0^{-1} \nabla \bigr\|^2, \\
2\,\square_{\underline{\boldsymbol{\epsilon}}} &= \bigl\| \mathcal{D}_0^{-1} \nabla \bigr\|^2 - \bigl\| \mathcal{D}_{\underline{\boldsymbol{\epsilon}}}^{-1} \nabla \bigr\|^2.
\end{aligned}$$

Define now

$$\begin{aligned}
\alpha_{\boldsymbol{\epsilon}} &\stackrel{\text{def}}{=} \bigl\| \mathcal{D}_0 \mathcal{D}_{\boldsymbol{\epsilon}}^{-2} \mathcal{D}_0 - I_{p^*} \bigr\|_\infty, \\
\alpha_{\underline{\boldsymbol{\epsilon}}} &\stackrel{\text{def}}{=} \bigl\| I_{p^*} - \mathcal{D}_0 \mathcal{D}_{\underline{\boldsymbol{\epsilon}}}^{-2} \mathcal{D}_0 \bigr\|_\infty.
\end{aligned}$$

The regularity conditions $(\mathcal{I})$ $\mathfrak{a}^2 \mathcal{D}_0^2 \geq \mathcal{V}_0^2$ implies for $\mathcal{D}_{\boldsymbol{\epsilon}}^2 = \mathcal{D}_0^2(1-\delta) - \varrho \mathcal{V}_0^2$

$$\begin{aligned}
\mathcal{D}_0^2(1 - \tau_{\boldsymbol{\epsilon}}) \leq \mathcal{D}_{\boldsymbol{\epsilon}}^2 &\leq \mathcal{D}_0^2, \\
\mathcal{D}_0^2 \leq \mathcal{D}_{\underline{\boldsymbol{\epsilon}}}^2 &\leq \mathcal{D}_0^2(1 + \tau_{\boldsymbol{\epsilon}}).
\end{aligned}$$

with $\tau_{\boldsymbol{\epsilon}} = \delta + \varrho \mathfrak{a}^{-2}$ so that the quantities $\alpha_{\boldsymbol{\epsilon}}$ and $\alpha_{\underline{\boldsymbol{\epsilon}}}$ satisfy

$$\alpha_{\boldsymbol{\epsilon}} \leq \frac{1}{1 - \tau_{\boldsymbol{\epsilon}}} - 1 = \frac{\tau_{\boldsymbol{\epsilon}}}{1 - \tau_{\boldsymbol{\epsilon}}}, \qquad \alpha_{\underline{\boldsymbol{\epsilon}}} \leq 1 - \frac{1}{1 + \tau_{\boldsymbol{\epsilon}}} = \frac{\tau_{\boldsymbol{\epsilon}}}{1 + \tau_{\boldsymbol{\epsilon}}}.$$

This yields

$$2\,\square_{\boldsymbol{\epsilon}} \leq \alpha_{\boldsymbol{\epsilon}} \bigl\| \mathcal{D}_0^{-1} \nabla \bigr\|^2, \qquad 2\,\square_{\underline{\boldsymbol{\epsilon}}} \leq \alpha_{\underline{\boldsymbol{\epsilon}}} \bigl\| \mathcal{D}_0^{-1} \nabla \bigr\|^2.$$

Similarly by using the result of Lemma D.2

$$\begin{aligned}
\square_{\boldsymbol{\epsilon},1} &\stackrel{\text{def}}{=} \sup_{\boldsymbol{\upsilon}:\, \Pi_0 \boldsymbol{\upsilon} = \boldsymbol{\theta}^*} \mathbb{L}_{\boldsymbol{\epsilon}}(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^*) - \sup_{\boldsymbol{\upsilon}:\, \Pi_0 \boldsymbol{\upsilon} = \boldsymbol{\theta}^*} \mathbb{L}(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^*) \\
&= \frac{1}{2} \bigl( \bigl\| H_{\boldsymbol{\epsilon}}^{-1} \nabla_{\boldsymbol{\eta}} \bigr\|^2 - \bigl\| H_0^{-1} \nabla_{\boldsymbol{\eta}} \bigr\|^2 \bigr) \leq \frac{\alpha_{\boldsymbol{\epsilon}}}{2} \bigl\| H_0^{-1} \nabla_{\boldsymbol{\eta}} \bigr\|^2, \\
\square_{\underline{\boldsymbol{\epsilon}},1} &\stackrel{\text{def}}{=} \sup_{\boldsymbol{\upsilon}:\, \Pi_0 \boldsymbol{\upsilon} = \boldsymbol{\theta}^*} \mathbb{L}(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^*) - \sup_{\boldsymbol{\upsilon}:\, \Pi_0 \boldsymbol{\upsilon} = \boldsymbol{\theta}^*} \mathbb{L}_{\underline{\boldsymbol{\epsilon}}}(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^*) \\
&= \frac{1}{2} \bigl( \bigl\| H_0^{-1} \nabla_{\boldsymbol{\eta}} \bigr\|^2 - \bigl\| H_{\underline{\boldsymbol{\epsilon}}}^{-1} \nabla_{\boldsymbol{\eta}} \bigr\|^2 \bigr) \leq \frac{\alpha_{\underline{\boldsymbol{\epsilon}}}}{2} \bigl\| H_0^{-1} \nabla_{\boldsymbol{\eta}} \bigr\|^2.
\end{aligned}$$

Further, (D.10) and (D.8) yield

$$\breve{L}(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}^*) \geq \frac{1}{2}\bigl\|\breve{D}_0^{-1}\breve{\nabla}_{\boldsymbol{\theta}}\bigr\|^2 - \square_{\underline{\boldsymbol{\epsilon}}} - \square_{\boldsymbol{\epsilon},1} - \diamondsuit_{\boldsymbol{\epsilon}}(\mathtt{r}) - \diamondsuit_{\underline{\boldsymbol{\epsilon}}}(\mathtt{r}),$$

$$\breve{L}(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}^*) \leq \frac{1}{2}\bigl\|\breve{D}_0^{-1}\breve{\nabla}_{\boldsymbol{\theta}}\bigr\|^2 + \square_{\boldsymbol{\epsilon}} + \square_{\underline{\boldsymbol{\epsilon}},1} + \diamondsuit_{\boldsymbol{\epsilon}}(\mathtt{r}) + \diamondsuit_{\underline{\boldsymbol{\epsilon}}}(\mathtt{r}).$$

The proof of (D.2) and (D.3) is completed.

The next step is to bound the spread $\Delta_{\boldsymbol{\epsilon}}(\mathtt{r})$ from (D.4). The error terms $\diamondsuit_{\boldsymbol{\epsilon}}(\mathtt{r})$ and $\diamondsuit_{\boldsymbol{\epsilon}}(\mathtt{r})$ follow the bound (B.2) of Theorem B.1 and they are of order $\varrho(p^* + \mathtt{x})$. Further we have to show that $\tau_{\boldsymbol{\epsilon}}\|\mathcal{D}_0^{-1}\nabla\|^2$ is small relative to $\|\breve{\boldsymbol{\xi}}\|^2$ and similarly for $\tau_{\boldsymbol{\epsilon}}\|H_0^{-1}\nabla_{\boldsymbol{\eta}}\|^2$. Theorem C.1 provides a general deviation probability bound for such quadratic forms. In particular, for $I\!B \stackrel{\mathrm{def}}{=} \mathcal{D}_0^{-1}\mathcal{V}_0^2\mathcal{D}_0^{-1}$ and $\mathtt{x} \leq \mathtt{x}_c$

$$I\!P\bigl(\|\mathcal{D}_0^{-1}\nabla\|^2 > \mathfrak{z}(\mathtt{x}, I\!B)\bigr) \leq 2\mathrm{e}^{-\mathtt{x}} + 8.4\mathrm{e}^{-\mathtt{x}_c},$$

where $\mathfrak{z}(\mathtt{x}, I\!B) \leq \mathrm{tr}(I\!B) + 6\mathtt{x}$ and the constant $\mathtt{x}_c$ is large; see Section C for a precise formulation. Under the regularity condition $(\mathcal{I})$ it holds $\mathrm{tr}(I\!B) \leq \mathfrak{a}^2 p^*$. A similar bound holds for $\|H_0^{-1}\nabla_{\boldsymbol{\eta}}\|^2$. We conclude that the spread $\Delta_{\boldsymbol{\epsilon}}(\mathtt{r})$ can be bounded with a probability of order $1 - \mathrm{e}^{-\mathtt{x}}$ by $\mathtt{C}\,\tau_{\boldsymbol{\epsilon}}(p^* + \mathtt{x})$ for a fixed constant $\mathtt{C}$.

Further we have to show that the random set $C_{\boldsymbol{\epsilon}}(\mathtt{r})$ from (D.1) is of dominating probability if $\mathtt{r}^2 = \mathtt{C}(p^* + \mathtt{x})$ for a proper constant $\mathtt{C}$. By definition

$$C_{\boldsymbol{\epsilon}}(\mathtt{r}) = \bigl\{\widetilde{\boldsymbol{\upsilon}} \in \Upsilon_\circ(\mathtt{r}),\, \widetilde{\boldsymbol{\upsilon}}_{\boldsymbol{\theta}^*} \in \Upsilon_\circ(\mathtt{r}),\, \|\mathcal{V}_0\mathcal{D}_{\underline{\boldsymbol{\epsilon}}}^{-2}\nabla\| \leq \mathtt{r},\, \|Q_0 H_{\underline{\boldsymbol{\epsilon}}}^{-2}\nabla_{\boldsymbol{\eta}}\| \leq \mathtt{r}\bigr\}.$$

Theorem B.2 yields

$$I\!P\bigl\{\widetilde{\boldsymbol{\upsilon}} \notin \Upsilon_\circ(\mathtt{r})\bigr\} + I\!P\bigl\{\widetilde{\boldsymbol{\upsilon}}_{\boldsymbol{\theta}^*} \notin \Upsilon_\circ(\mathtt{r})\bigr\} \leq 2\mathrm{e}^{-\mathtt{x}}.$$

To control the probability $I\!P\bigl(\|\mathcal{V}_0\mathcal{D}_{\underline{\boldsymbol{\epsilon}}}^{-2}\nabla\| > \mathtt{r}\bigr)$ we apply Corollary C.1 with

$$I\!B = \mathcal{D}_0^{-1}\mathcal{V}_0^2\mathcal{D}_0^{-1}, \quad .$$

With the definitions from Section C

$$\begin{aligned}
I\!P\bigl(\|\mathcal{V}_0\mathcal{D}_{\underline{\boldsymbol{\epsilon}}}^{-2}\nabla\| > \mathtt{r}\bigr) &\leq I\!P\bigl(\|\mathcal{D}_{\underline{\boldsymbol{\epsilon}}}^{-1}\nabla\| \times \|\mathcal{V}_0\mathcal{D}_{\underline{\boldsymbol{\epsilon}}}^{-1}\|_\infty > \mathtt{r}\bigr) \\
&\leq I\!P\bigl(\|\mathcal{D}_0^{-1}\nabla\| \times \|\mathcal{V}_0\mathcal{D}_0^{-1}\|_\infty \geq (1 - \tau_{\boldsymbol{\epsilon}})\mathtt{r}\bigr) \\
&\leq I\!P\bigl(\|\mathcal{D}_0^{-1}\nabla\| \geq (1 - \tau_{\boldsymbol{\epsilon}})\mathtt{r}/\mathfrak{a}\bigr) \\
&\leq I\!P\bigl\{\|\mathcal{D}_0^{-1}\nabla\|^2 \geq \mathfrak{z}(\mathtt{x}, I\!B)\bigr\} \\
&< \mathrm{e}^{-\mathtt{x}} + 8.4\mathrm{e}^{-\mathtt{x}_c},
\end{aligned}$$

provided that $\mathtt{r}^2 > \mathfrak{a}^4(1 - \tau_{\boldsymbol{\epsilon}})^{-2}(p^* + 6\mathtt{x})$ and $\mathtt{x} \leq \mathtt{x}_c$. By similar arguments with $I\!B_{\boldsymbol{\eta}} = H_0^{-1}Q_0^2H_0^{-1}$ in place of $I\!B$

$$I\!P\bigl(\|Q_0 H_{\underline{\boldsymbol{\epsilon}}}^{-2}\nabla_{\underline{\boldsymbol{\epsilon}}}\| > \mathtt{r}\bigr) < \mathrm{e}^{-\mathtt{x}} + 8.4\mathrm{e}^{-\mathtt{x}_c^{\boldsymbol{\eta}}}.$$

Putting the obtained bounds together shows that for $\mathtt{x} \le \mathtt{x}_c$ and $\mathtt{r}_0^2 \ge \mathtt{C}_1(p^* + \mathtt{x})$, it holds

$$1 - I\!P\big(C_{\boldsymbol{\epsilon}}(\mathtt{r}_0)\big) \le \mathtt{C}_2 \mathrm{e}^{-\mathtt{x}},$$

for some fixed constants $\mathtt{C}_1$ and $\mathtt{C}_2$ depending on $\tau_{\boldsymbol{\epsilon}}$ and $\mathfrak{a}$ only. This completes the proof.

## D.2 Proof of Theorem 2.2

We show that

$$2\Delta^*_{\boldsymbol{\epsilon}}(\mathtt{r}) \stackrel{\mathrm{def}}{=} \frac{2\Delta_{\boldsymbol{\epsilon}}(\mathtt{r})}{1-\tau_{\boldsymbol{\epsilon}}} + \frac{\tau_{\boldsymbol{\epsilon}}}{1-\tau_{\boldsymbol{\epsilon}}} \big\| \mathcal{D}_0^{-1} \nabla \big\|. \tag{D.11}$$

First we derive the expansion for the whole parameter vector $\boldsymbol{\upsilon}$. On the set $C_{\boldsymbol{\epsilon}}(\mathtt{r})$, the bracketing bound (D.9) and (D.5) imply

$$\begin{aligned} \mathcal{L}(\widetilde{\boldsymbol{\upsilon}}, \boldsymbol{\upsilon}^*) &\ge \sup_{\boldsymbol{\upsilon}} \mathbb{L}_{\underline{\boldsymbol{\epsilon}}}(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^*) - \diamondsuit_{\underline{\boldsymbol{\epsilon}}}(\mathtt{r}) \\ &= \|\mathcal{D}_{\underline{\boldsymbol{\epsilon}}}^{-1} \nabla\|^2/2 - \diamondsuit_{\underline{\boldsymbol{\epsilon}}}(\mathtt{r}) \\ &\ge \|\mathcal{D}_{\boldsymbol{\epsilon}}^{-1} \nabla\|^2/2 - \diamondsuit_{\underline{\boldsymbol{\epsilon}}}(\mathtt{r}) - \square_{\boldsymbol{\epsilon}} - \square_{\underline{\boldsymbol{\epsilon}}}\,. \end{aligned}$$

The bracketing bound (D.9) applied at $\widetilde{\boldsymbol{\upsilon}}$ implies

$$\mathcal{L}(\widetilde{\boldsymbol{\upsilon}}, \boldsymbol{\upsilon}^*) \le \mathbb{L}_{\boldsymbol{\epsilon}}(\widetilde{\boldsymbol{\upsilon}}, \boldsymbol{\upsilon}^*) + \diamondsuit_{\boldsymbol{\epsilon}}(\mathtt{r}).$$

These two bounds together yield by the definition of $\mathbb{L}_{\boldsymbol{\epsilon}}(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^*)$

$$(\widetilde{\boldsymbol{\upsilon}} - \boldsymbol{\upsilon}^*)^\top \nabla - \frac{1}{2} \big\| \mathcal{D}_{\boldsymbol{\epsilon}}(\widetilde{\boldsymbol{\upsilon}} - \boldsymbol{\upsilon}^*) \big\|^2 \ge \frac{1}{2} \|\mathcal{D}_{\boldsymbol{\epsilon}}^{-1} \nabla\|^2 - \diamondsuit_{\boldsymbol{\epsilon}}(\mathtt{r}) - \diamondsuit_{\underline{\boldsymbol{\epsilon}}}(\mathtt{r}) - \square_{\boldsymbol{\epsilon}} - \square_{\underline{\boldsymbol{\epsilon}}},$$

and thus

$$\big\| \mathcal{D}_{\boldsymbol{\epsilon}}(\widetilde{\boldsymbol{\upsilon}} - \boldsymbol{\upsilon}^*) - \mathcal{D}_{\boldsymbol{\epsilon}}^{-1} \nabla \big\|^2 \le 2\big\{ \square_{\boldsymbol{\epsilon}} + \square_{\underline{\boldsymbol{\epsilon}}} + \diamondsuit_{\boldsymbol{\epsilon}}(\mathtt{r}) + \diamondsuit_{\underline{\boldsymbol{\epsilon}}}(\mathtt{r}) \big\} \le 2\Delta_{\boldsymbol{\epsilon}}\,. \tag{D.12}$$

The condition $(\mathcal{I})$ implies the inequality $\big\| \mathcal{D}_0^{-1} \mathcal{D}_{\boldsymbol{\epsilon}}^2 \mathcal{D}_0^{-1} \big\|_\infty \ge 1 - \tau_{\boldsymbol{\epsilon}}$ and hence,

$$\big\| \mathcal{D}_0 \mathcal{D}_{\boldsymbol{\epsilon}}^{-2} \mathcal{D}_0 \big\|_\infty \le (1 - \tau_{\boldsymbol{\epsilon}})^{-1}.$$

This and (D.12) provide

$$\big\| \mathcal{D}_0(\widetilde{\boldsymbol{\upsilon}} - \boldsymbol{\upsilon}^*) - \mathcal{D}_0 \mathcal{D}_{\boldsymbol{\epsilon}}^{-2} \nabla \big\|^2 \le \frac{2\Delta_{\boldsymbol{\epsilon}}}{1 - \tau_{\boldsymbol{\epsilon}}}\,.$$

Similarly

$$\big\| \mathcal{D}_0 \mathcal{D}_{\boldsymbol{\epsilon}}^{-2} \nabla - \mathcal{D}_0^{-1} \nabla \big\| = \big\| \big( \mathcal{D}_0 \mathcal{D}_{\boldsymbol{\epsilon}}^{-2} \mathcal{D}_0 - I_{p^*} \big) \mathcal{D}_0^{-1} \nabla \big\| \le \frac{\tau_{\boldsymbol{\epsilon}}}{1 - \tau_{\boldsymbol{\epsilon}}} \big\| \mathcal{D}_0^{-1} \nabla \big\|.$$

Putting together the last two bounds yields

$$\bigl\| \mathcal{D}_0(\widetilde{\boldsymbol{\upsilon}} - \boldsymbol{\upsilon}^*) - \mathcal{D}_0^{-1}\nabla \bigr\| \le \sqrt{\frac{2\Delta_{\boldsymbol{\epsilon}}}{1-\tau_{\boldsymbol{\epsilon}}}} + \frac{\tau_{\boldsymbol{\epsilon}}}{1-\tau_{\boldsymbol{\epsilon}}} \bigl\| \mathcal{D}_0^{-1}\nabla \bigr\|.$$

It remains to note that for any $\boldsymbol{u} \in I\!R^p$, $\boldsymbol{\eta} \in I\!R^m$, and $\boldsymbol{w} = (\boldsymbol{u}, \boldsymbol{\eta}) \in I\!R^{p^*}$, it holds with $\boldsymbol{\gamma} \stackrel{\text{def}}{=} \boldsymbol{\eta} + H_0^{-2} A_0^\top \boldsymbol{u} \in I\!R^m$

$$\bigl\| \mathcal{D}_0 \boldsymbol{w} \bigr\|^2 = \bigl\| \breve{D}_0 \boldsymbol{u} \bigr\|^2 + \bigl\| H_0 \boldsymbol{\gamma} \bigr\|^2 \ge \bigl\| \breve{D}_0 \boldsymbol{u} \bigr\|^2. \tag{D.13}$$

Also we use $I\!I_0 \mathcal{D}_0^{-2} \nabla = \breve{D}_0^{-2} \breve{\nabla}$ .This implies for $\boldsymbol{w} = \widetilde{\boldsymbol{\upsilon}} - \boldsymbol{\upsilon}^*$ by (D.13)

$$\bigl\| \breve{D}_0 (\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) - \breve{D}_0^{-1} \breve{\nabla} \bigr\| = \bigl\| \breve{D}_0 (\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^* - \breve{D}_0^{-2} \breve{\nabla}) \bigr\| \le \bigl\| \mathcal{D}_0 (\boldsymbol{w} - \mathcal{D}_0^{-2} \nabla) \bigr\|,$$

and the assertion (D.11) follows.

## D.3 Proof of Theorem 2.3

Choose $\mathtt{x}_n \to \infty$ and $\mathtt{x}_n = o(n^{1/3})$, e.g. $\mathtt{x}_n = \mathtt{C} \log(n)$. Then $\beta_n \to 0$ and $I\!P(\Omega(\mathtt{x}_n)) \to 1$. Moreover, in the i.i.d. setting $\mathtt{x}_c \cong \mathtt{g} \cong \sqrt{n}$ and thus $\mathtt{x}_n \le \mathtt{x}_c$. Similarly for $n$ large enough with $\mathtt{r}_0^2 = \mathtt{r}_0^2(\mathtt{x}_n) = \mathtt{C}(p^* + \mathtt{x}_n)$

$$\tau_{\boldsymbol{\epsilon}} \cong \mathtt{r}_0 / \sqrt{n} \cong \sqrt{(p^* + \mathtt{x}_n)/n} < n^{-1/3} \le 1/2.$$

Also the i.i.d. structure of the data yields

$$\breve{D}_0^2 = n \breve{\mathbb{F}}.$$

Now Theorems 2.1 and 2.2 can be applied yielding the first statement of the theorem. It remains to check asymptotic standard normality of the sum

$$\breve{\boldsymbol{\xi}} = (n\breve{\mathbb{F}})^{-1/2} \breve{\nabla} = (n\breve{\mathbb{F}})^{-1/2} \sum_{i=1}^n \breve{\nabla}_i,$$

with

$$\breve{\nabla}_i \stackrel{\text{def}}{=} \nabla_{\boldsymbol{\theta}} \ell(Y_i, \boldsymbol{\upsilon}) - \mathbb{F}_{\boldsymbol{\theta}\boldsymbol{\eta}} \mathbb{F}_{\boldsymbol{\eta}\boldsymbol{\eta}}^{-1} \nabla_{\boldsymbol{\eta}} \ell(Y_i, \boldsymbol{\upsilon}).$$

The result follows from the central limit theorem because $\mathrm{Cov}(\breve{\nabla}_i) = \breve{\mathbb{F}}$ for all $i$.

## D.4 Proof of Theorem 2.4

The first step of the proof shows that for $n$ large enough, the MLE $\widetilde{\boldsymbol{\upsilon}} \in I\!R^{p_n}$ belongs with probability close to one to the $\delta = 1/n$ vicinity $\mathcal{S}_\delta$ of the set $\mathcal{S}$ from (2.7). The second step is to show that with a probability exceeding a fixed constant $\alpha > 0$, the profile MLE $\widetilde{\theta}$ differs

significantly from $X_1$ which is the profile MLE in the linear Gaussian model. The third step focuses on the case $\beta_n \to \infty$.

1. First we show that for $n$ large enough, the MLE $\widetilde{\boldsymbol{\upsilon}} \in I\!R^{p_n}$ lies in $\mathcal{S}_\delta$ with probability close to one. For this we check that the maximum of $\mathcal{L}(\boldsymbol{\upsilon})$ on $\mathcal{S}_\delta^c$ is smaller than a similar maximum on $\mathcal{S}$ for "typical" values of $\boldsymbol{X}$ and $n$ large enough. Indeed, for any point $\boldsymbol{\upsilon} \in \mathcal{S}_\delta^c$

$$\begin{aligned}\mathcal{L}(\boldsymbol{\upsilon}, 0) &\le \max_{\boldsymbol{\upsilon} \in \mathcal{S}_\delta^c} \mathcal{L}(\boldsymbol{\upsilon}, 0) = \max_{\boldsymbol{\upsilon} \in \mathcal{S}_\delta^c} \bigl\{ n \boldsymbol{X}^\top \boldsymbol{\upsilon} - n\|\boldsymbol{\upsilon}\|^2/2 \bigr\} \\ &\le \max_{\boldsymbol{\upsilon} \in I\!R^{p_n}} \bigl\{ n \boldsymbol{X}^\top \boldsymbol{\upsilon} - n\|\boldsymbol{\upsilon}\|^2/2 \bigr\} = \frac{n}{2} \|\boldsymbol{X}\|^2.\end{aligned}$$

Further, introduce a random set of "typical" values $\boldsymbol{X}$ :

$$C_1 \stackrel{\text{def}}{=} \left\{ \boldsymbol{X} : \ \frac{1}{2} \left( \frac{p_n}{n} \right)^{3/2} < \|\boldsymbol{X}\|^3 < \left( \frac{2p_n}{n} \right)^{3/2}, \text{ and } |X_1| \le 1 \right\}.$$

It is straightforward to see that $I\!P\bigl(\boldsymbol{X} \in C_1\bigr)$ is exponentially close to one for $n$ large. Below we assume that $\boldsymbol{X} \in C_1$ and study the value $\mathcal{L}(\boldsymbol{\upsilon}, 0)$ for $\boldsymbol{\upsilon} \in \mathcal{S}$. Let also $n$ be large enough to ensure that

$$\frac{2^{1/3} - 1}{2^{1/6}} \left( \frac{p_n}{n} \right)^{1/2} \ge \frac{1}{2} \left( \frac{p_n}{n} \right)^{3/4} = \frac{1}{2} \sqrt{\beta_n / n}. \tag{D.14}$$

Introduce $\boldsymbol{X}_{\mathcal{S}}$ as the closest point in $\mathcal{S}$ to $\boldsymbol{X}$ with $|\upsilon_1| \ge |X_1|$. This point always exists by the definition of $\mathcal{S}$. Denote

$$\delta(\boldsymbol{X}) = \|\boldsymbol{X} - \boldsymbol{X}_{\mathcal{S}}\| = |X_1 - \upsilon_1|.$$

By construction of $\mathcal{S}$, it holds $\delta(\boldsymbol{X}) \le 0.5 \sqrt{\beta_n / n}$ for $\boldsymbol{X} \in C_1$. For $n$ satisfying (D.14) this also yields $\bigl[ \|\boldsymbol{X}\| - \delta(\boldsymbol{X}) \bigr]^3 \ge 1/2 \|\boldsymbol{X}\|^3$. Now we have for $\boldsymbol{X} \in C_1$

$$\begin{aligned}\max_{\boldsymbol{\upsilon} \in \mathcal{S}} \mathcal{L}(\boldsymbol{\upsilon}, 0) &\ge \mathcal{L}(\boldsymbol{X}_{\mathcal{S}}, 0) \\ &\ge n\|\boldsymbol{X}\|^2 - n|X_1|\delta(\boldsymbol{X}) - \frac{n}{2} \bigl\{ \|\boldsymbol{X}\|^2 - 2|X_1|\delta(\boldsymbol{X}) + \delta^2(\boldsymbol{X}) \bigr\} \\ &\qquad + n \bigl\{ \|\boldsymbol{X}\|^2 - 2|X_1|\delta(\boldsymbol{X}) + \delta^2(\boldsymbol{X}) \bigr\}^{3/2} \\ &\ge \frac{n}{2} \|\boldsymbol{X}\|^2 - n\delta^2(\boldsymbol{X}) + n \bigl\{ \|\boldsymbol{X}\| - \delta(\boldsymbol{X}) \bigr\}^3 \\ &> \frac{n}{2} \|\boldsymbol{X}\|^2 - \frac{\beta_n}{4} + \frac{n}{2} \|\boldsymbol{X}\|^3 > \frac{n}{2} \|\boldsymbol{X}\|^2 = \max_{\boldsymbol{\upsilon} \in \mathcal{S}_\delta^c} \mathcal{L}(\boldsymbol{\upsilon}, 0).\end{aligned}$$

This implies $\widetilde{\boldsymbol{\upsilon}} \in \mathcal{S}_\delta$.

2. Now we discuss the case when $\beta_n^2 = p_n^3 / n \to (6c)^4$ for some $c \ge 0$ and show that the profile MLE $\widetilde{\theta}$ deviates significantly from $X_1$ on a random set of positive probability. Define for each $n \in \mathbb{N}$

$$C_n \stackrel{\text{def}}{=} C_1 \cap \left\{ \|\boldsymbol{X} - \boldsymbol{X}_{\mathcal{S}}\| \ge \frac{1}{6} \sqrt{\beta_n / n} \right\} = C_1 \cap \left\{ |X_1 - X_{\mathcal{S},1}| \ge \frac{1}{6} \sqrt{\beta_n / n} \right\}.$$

It is easy to see that $I\!P(C_n) \geq \alpha$ for some fixed $\alpha > 0$ and all $n$. It remains to note that on the set $C_n$ it holds under (D.14)

$$\begin{aligned}
\|\breve{D}_0(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) - \breve{\boldsymbol{\xi}}\| &= \sqrt{n}|\widetilde{\upsilon}_1 - X_1| \\
&\geq \sqrt{n}|X_1 - X_{\mathbb{S},1}| - \sqrt{n}/n \\
&\geq \frac{1}{6}\beta_n^{1/2} - \frac{1}{\sqrt{n}} \to \begin{cases} \infty & p_n^3/n \to \infty, \\ c & p_n^3/n \to (6c)^4, \end{cases}
\end{aligned}$$

which yields the claim.

3. Finally consider the case when $\beta_n \to \infty$. Fix any sequence $c_n$ such that $c_n \to 0$ and $c_n^2\beta_n \to \infty$, e.g. $c_n = \beta_n^{-1/4}$. Consider the random set

$$C_n \stackrel{\text{def}}{=} C_1 \cap \left\{\|\boldsymbol{X} - \boldsymbol{X}_{\mathbb{S}}\| \geq \frac{c_n}{6}\sqrt{\beta_n/n}\right\} = C_1 \cap \left\{|X_1 - X_{\mathbb{S},1}| \geq \frac{c_n}{6}\sqrt{\beta_n/n}\right\}.$$

Then $I\!P(C_n) \to 1$ and on $C_n$

$$\|\breve{D}_0(\widetilde{\theta} - \theta^*) - \breve{\boldsymbol{\xi}}\| \geq \frac{c_n}{6}\beta_n^{1/2} - \frac{1}{\sqrt{n}} \to \infty,$$

as required.

# References


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

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

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

Ibragimov, I. and Khas'minskij, R. (1981). *Statistical estimation. Asymptotic theory. Transl. from the Russian by Samuel Kotz.* New York - Heidelberg -Berlin: Springer-Verlag .

Kosorok, M. (2005). *Introduction to Empirical Processes and Semiparametric Inference*. Springer in Statistics.

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

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

Spokoiny, V. (2012). Parametric estimation. Finite sample theory. *Ann. Statist.*, 40(6):2877–2909. arXiv:1111.3029.

Spokoiny, V. (2013). Sharp deviation bounds for quadratic forms. Manuscript. arXiv:1302.1699.

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