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

Preprint ISSN 2198-5855

Bootstrap confidence sets under model misspecification

Vladimir Spokoiny [1,2] Mayya Zhilova [3]

submitted: November 18, 2014

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

[2] Moscow Institute of Physics and Technology
9 Institutskiy per., Dolgoprudny
Moscow Region, 141700
Russian Federation

[3] Weierstrass Institute
Mohrenstr. 39
10117 Berlin
Germany
E-Mail: mayya.zhilova@wias-berlin.de

No. 1992

Berlin 2014

2010 *Mathematics Subject Classification.* 62F25 (Primary) 62F40, 62E17 (Secondary).

Key words and phrases. likelihood-based bootstrap confidence set, misspecified model, finite sample size, multiplier bootstrap, weighted bootstrap, Gaussian approximation, Pinsker's inequality.

[1] 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 Research Unit 1735 is gratefully acknowledged.

[2] 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

A multiplier bootstrap procedure for construction of likelihood-based confidence sets is considered for finite samples and a possible model misspecification. Theoretical results justify the bootstrap consistency for a small or moderate sample size and allow to control the impact of the parameter dimension p: the bootstrap approximation works if p^3/n is small. The main result about bootstrap consistency continues to apply even if the underlying parametric model is misspecified under the so called Small Modeling Bias condition. In the case when the true model deviates significantly from the considered parametric family, the bootstrap procedure is still applicable but it becomes a bit conservative: the size of the constructed confidence sets is increased by the modeling bias. We illustrate the results with numerical examples for misspecified constant and logistic regressions.

1 Introduction

Since introducing in 1979 by Efron (1979) the bootstrap procedure became one of the most powerful and common tools in statistical confidence estimation and hypothesis testing. Many versions and extensions of the original bootstrap method have been proposed in the literature; see e.g. Wu (1986)Newton and Raftery (1994); Barbe and Bertail (1995); Horowitz (2001); Chatterjee and Bose (2005); Ma and Kosorok (2005); Chen and Pouzo (2009); Lavergne and Patilea (2013); Chen and Pouzo (2014) among many others. This paper focuses on the multiplier bootstrap procedure which attracted a lot of attention last time due to its nice theoretical properties and numerical performance. We mention the papers Chatterjee and Bose (2005), Arlot et al. (2010) and Chernozhukov et al. (2013) for the most advanced recent results. Chatterjee and Bose (2005) showed some results on asymptotic bootstrap consistency in a very general framework: for estimators obtained by solving estimating equations. Chernozhukov et al. (2013) presented a number of non-asymptotic results on bootstrap validity with applications to special problems like testing many moment restrictions or parameter choice for a LASSO procedure. Arlot et al. (2010) constructed a non-asymptotical confidence bound in ℓ_s norm ($s \in [1, \infty]$) for the mean of a sample of high dimensional i.i.d. Gaussian vectors (or with a symmetric and bounded distribution), using the generalized weighted bootstrap for resampling of the quantiles.

This paper makes a further step in studying the multiplier bootstrap method in the problem of confidence estimation by a quasi maximum likelihood method. For a rather general parametric model, we consider likelihood-based confidence sets with the radius determined by a multiplier bootstrap. The aim of the study is to check the validity of the bootstrap procedure in situations with a large parameter dimension, a limited sample size, and a possible misspecification of the parametric assumption. The main result of the paper explicitly describes the error term of the bootstrap approximation. This particularly allows to track the impact of the parameter

dimension p and of the sample size n in the quality of the bootstrap procedure. As one of the corollaries, we show bootstrap validity under the constraint "p^3/n-small". Chatterjee and Bose (2005) stated results under the condition "p/n-small" but their results only apply to low dimensional projections of the MLE vector. In the likelihood based approach, the construction involves the Euclidean norm of the MLE which leads to completely different tools and results. Chernozhukov et al. (2013) allowed a huge parameter dimension with "$\log(p)/n$ small" but they essentially work with a family of univariate tests which again differs essentially from the maximum likelihood approach.

Another interesting and important issue is the impact of the model misspecification on the accuracy of bootstrap approximation. A surprising corollary of our error bounds is that the bootstrap confidence set can be used even if the underlying parametric model is slightly misspecified under the so called *small modeling bias* $(\mathbf{SmB})$ condition. If the modeling bias becomes large, the bootstrap confidence sets are still applicable, but they become more and more conservative. $(\mathbf{SmB})$ condition is given in Section 4 and it is consistent with classical bias-variance relation in nonparametric estimation.

Our theoretical study uses the square-root Wilks (sq-Wilks) expansion from Spokoiny (2012a), Spokoiny (2013) which approximates the square root likelihood ratio statistic by the norm of the standardized score vector. Further we extend the sq-Wilks expansion to the bootstrap log-likelihood and adopt the Gaussian approximation theory (GAR) to the special case when the distribution of the Euclidean norm of a non-Gaussian vector is approximated by the distribution of the norm of a Gaussian one with the same first and second moments. The Gaussian comparison technique based on the Pinsker inequality completes the study and allows to bridge the real unknown coverage probability and the conditional bootstrap coverage probability under $(\mathbf{SmB})$ condition. In the case of a large modeling bias we state a one-sided bound: the bootstrap quantiles are uniformly larger than the real ones. This effect is nicely confirmed by our simulation study.

Now consider the problem and the approach in more detail. Let the data sample $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^\top$ consist of *independent* random observations and belong to the probability space $(\Omega, \mathcal{F}, I\!P)$. We do not assume that the observations Y_i are identically distributed, moreover, no specific parametric structure of $I\!P$ is being required. In order to explain the idea of the approach we start here with a parametric case, however the assumption (1.1) below is not required for the results. Let $I\!P$ belong to some known regular parametric family $\{I\!P_{\boldsymbol{\theta}}\} \stackrel{\text{def}}{=} \{I\!P_{\boldsymbol{\theta}} \ll \mu_0, \boldsymbol{\theta} \in \Theta \subset I\!R^p\}$. In this case the true parameter $\boldsymbol{\theta}^* \in \Theta$ is such that

$$I\!P \equiv I\!P_{\boldsymbol{\theta}^*} \in \{I\!P_{\boldsymbol{\theta}}\}, \tag{1.1}$$

and the initial problem of finding the properties of unknown distribution $I\!P$ is reduced to the equivalent problem for the finite-dimensional parameter $\boldsymbol{\theta}^*$. The parametric family $\{I\!P_{\boldsymbol{\theta}}\}$ induces the log-likelihood process $L(\boldsymbol{\theta})$ of the sample $\boldsymbol{Y}$:

$$L(\boldsymbol{\theta}) = L(\boldsymbol{Y}, \boldsymbol{\theta}) \stackrel{\text{def}}{=} \log \left\{ \frac{dI\!P_{\boldsymbol{\theta}}}{d\mu_0}(\boldsymbol{Y}) \right\}$$

and the maximum likelihood estimate (MLE) of $\boldsymbol{\theta}^*$:

$$\widetilde{\boldsymbol{\theta}} \stackrel{\text{def}}{=} \operatorname{argmax}_{\boldsymbol{\theta}\in\Theta} L(\boldsymbol{\theta}). \tag{1.2}$$

The asymptotic Wilks phenomenon Wilks (1938) states that for the case of i.i.d. observations with the sample size tending to the infinity the likelihood ratio statistic converges in distribution to $\chi^2_p/2$, where p is the parameter dimension:

$$2\bigl\{L(\widetilde{\boldsymbol{\theta}}) - L(\boldsymbol{\theta}^*)\bigr\} \xrightarrow{w} \chi^2_p,\ n \to \infty.$$

Define the likelihood-based confidence set as

$$\mathcal{E}(\mathfrak{z}) \stackrel{\text{def}}{=} \Bigl\{\boldsymbol{\theta} : L(\widetilde{\boldsymbol{\theta}}) - L(\boldsymbol{\theta}) \le \mathfrak{z}^2/2\Bigr\}, \tag{1.3}$$

then the Wilks phenomenon implies

$$I\!P\Bigl\{\boldsymbol{\theta}^* \in \mathcal{E}(\mathfrak{z}_{\alpha,\chi^2_p})\Bigr\} \to \alpha,\ n \to \infty,$$

where $\mathfrak{z}^2_{\alpha,\chi^2_p}$ is the $(1-\alpha)$ -quantile for the χ^2_p distribution. This result is very important and useful under the parametric assumption, i.e. when (1.1) holds. In this case the limit distribution of the likelihood ratio is independent of the model parameters or in other words it is *pivotal*. By this result a sufficiently large sample size allows to construct the confidence sets for $\boldsymbol{\theta}^*$ with a given coverage probability. However, a possibly low speed of convergence of the likelihood ratio statistic makes the asymptotic Wilks result hardly applicable to the case of small or moderate samples. Moreover, the asymptotical pivotality breaks down if the parametric assumption (1.1) does not hold (see Huber (1967)), and, therefore, the whole approach may be misleading if the model is considerably misspecified. If the assumption (1.1) does not hold, then the "true" parameter is defined by the projection of the true measure $I\!P$ on the parametric family $\{I\!P_{\boldsymbol{\theta}}\}$:

$$\boldsymbol{\theta}^* \stackrel{\text{def}}{=} \operatorname{argmax}_{\boldsymbol{\theta}\in\Theta} I\!E L(\boldsymbol{\theta}). \tag{1.4}$$

The recent results by Spokoiny (2012a), Spokoiny (2013) provide a non-asymptotic version of square-root Wilks phenomenon for the case of misspecified model. It holds with an exponentially high probability

$$\left|\sqrt{2\bigl\{L(\widetilde{\boldsymbol{\theta}}) - L(\boldsymbol{\theta}^*)\bigr\}} - \|\boldsymbol{\xi}\|\right| \le \Delta_{\mathrm{w}} \simeq \frac{p}{\sqrt{n}}, \tag{1.5}$$

where $\boldsymbol{\xi} \stackrel{\text{def}}{=} D_0^{-1}\nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta}^*)$, $D_0^2 \stackrel{\text{def}}{=} -\nabla^2_{\boldsymbol{\theta}} I\!E L(\boldsymbol{\theta}^*)$. The bound is non-asymptotical, the approximation error term Δ_{w} has an explicit form (the precise statement is given in Theorem A.2, Section A.1, and it depends on the parameter dimension p , sample size n , and the probability of the random set on which the result holds.

Due to this bound, the original problem of finding a quantile of the LR test statistic $L(\widetilde{\boldsymbol{\theta}}) - L(\boldsymbol{\theta}^*)$ is reduced to a similar question for the approximating quantity $\|\boldsymbol{\xi}\|$. The difficulty here is that in general $\|\boldsymbol{\xi}\|$ is non-pivotal, it depends on the unknown distribution $I\!P$ and the target parameter

$\boldsymbol{\theta}^*$. Another result by Spokoiny (2012b) gives the following non-asymptotical deviation bound for $\|\boldsymbol{\xi}\|^2$: for some explicit constant $C > 0$ it holds for $\mathtt{x} \geq \sqrt{p}$

$$I\!P\bigl(\|\boldsymbol{\xi}\|^2 \geq I\!E\|\boldsymbol{\xi}\|^2 + C\mathtt{x}\bigr) \leq 2\mathrm{e}^{-\mathtt{x}}$$

(the precise statement is given in Theorem A.3. This is a non-asymptotic deviation bound, sharp in leading approximating terms, however, the critical values yielded by it are too conservative for a valuable confidence set.

In the present work we study the *multiplier bootstrap* (or *weighted bootstrap*) procedure for estimation of the quantiles of the likelihood ratio statistic. The idea of the procedure is to mimic a distribution of the likelihood ratio statistic by reweighing its summands with random multipliers independent of the data:

$$L^\circ(\boldsymbol{\theta}) \stackrel{\mathrm{def}}{=} \sum\nolimits_{i=1}^n \log\left\{\frac{dI\!P_{\boldsymbol{\theta}}}{d\mu_0}(Y_i)\right\} u_i.$$

Here the probability distribution is taken conditionally on the data $\boldsymbol{Y}$, which is denoted by the sign $^\circ$. The random weights $u_1, \ldots, u_n$ are i.i.d. with continuos c.d.f., independent of $\boldsymbol{Y}$ and it holds for them: $I\!E(u_i) = 1$, $\mathrm{Var}(u_i) = 1$, $I\!E\exp(u_i) < \infty$. Therefore, the multiplier bootstrap induces the probability space conditional on the data $\boldsymbol{Y}$. A simple but important observation is that $I\!E^\circ L^\circ(\boldsymbol{\theta}) \equiv I\!E\bigl[L^\circ(\boldsymbol{\theta}) \bigm| \boldsymbol{Y}\bigr] = L(\boldsymbol{\theta})$, and hence,

$$\mathrm{argmax}_{\boldsymbol{\theta}}\, I\!E^\circ L^\circ(\boldsymbol{\theta}) = \mathrm{argmax}_{\boldsymbol{\theta}}\, L(\boldsymbol{\theta}) = \widetilde{\boldsymbol{\theta}}.$$

This means that the target parameter in the bootstrap world is precisely known and it coincides with the maximum likelihood estimator $\widetilde{\boldsymbol{\theta}}$ conditioned on $\boldsymbol{Y}$, therefore, the bootstrap likelihood ratio statistic $L^\circ(\widetilde{\boldsymbol{\theta}}^\circ) - L^\circ(\widetilde{\boldsymbol{\theta}}) \stackrel{\mathrm{def}}{=} \sup_{\boldsymbol{\theta}\in\Theta} L^\circ(\boldsymbol{\theta}) - L^\circ(\widetilde{\boldsymbol{\theta}})$ is fully computable and leads to a simple computational procedure for the approximation of the distribution of $L(\widetilde{\boldsymbol{\theta}}) - L(\boldsymbol{\theta}^*)$.

The goal of the present study is to show in a non-asymptotic way the consistency of the described multiplier bootstrap procedure and to obtain an explicit bound on the error of coverage probability. In other words, we are interested in non-asymptotic approximation of the distribution of $\bigl\{L(\widetilde{\boldsymbol{\theta}}) - L(\boldsymbol{\theta}^*)\bigr\}^{1/2}$ with the distribution of $\bigl\{L^\circ(\widetilde{\boldsymbol{\theta}}^\circ) - L^\circ(\widetilde{\boldsymbol{\theta}})\bigr\}^{1/2}$. So far there exist very few theoretical non-asymptotic results about bootstrap validity. Important contributions are given in the works by Chernozhukov et al. (2013) and Arlot et al. (2010). Finite sample methods for study of the bootstrap validity are essentially different from the asymptotic ones which are mainly based on weak convergence arguments. The main steps of our theoretical study are illustrated by the following scheme:

$$\begin{array}{lccccc}
 & \overset{\text{sq-Wilks theorem}}{} & & \overset{\text{Gauss. approx.}}{} & & \\
\boldsymbol{Y}\text{-world:} & \sqrt{2L(\widetilde{\boldsymbol{\theta}}) - 2L(\boldsymbol{\theta}^*)} & \approx & \|\boldsymbol{\xi}\| & \overset{w}{\approx} & \|\overline{\boldsymbol{\xi}}\| \\
 & & & & \wr\, w \ \text{Gauss. compar.} & \\
\text{Bootstrap world:} & \sqrt{2L^\circ(\widetilde{\boldsymbol{\theta}}^\circ) - 2L^\circ(\widetilde{\boldsymbol{\theta}})} & \approx & \|\boldsymbol{\xi}^\circ\| & \overset{w}{\approx} & \|\overline{\boldsymbol{\xi}}^\circ\|,
\end{array} \tag{1.6}$$

where $\boldsymbol{\xi}^\circ \stackrel{\text{def}}{=} D_0^{-1}\nabla_{\boldsymbol{\theta}}\left[L^\circ(\boldsymbol{\theta}^*) - I\!\!E\left\{L^\circ(\boldsymbol{\theta}^*)\,\middle|\,\boldsymbol{Y}\right\}\right]$; compare with the definition (1.5) of the vector $\boldsymbol{\xi}$ in the $\boldsymbol{Y}$ -world. The vectors $\overline{\boldsymbol{\xi}}$ and $\overline{\boldsymbol{\xi}}^\circ$ are zero mean Gaussian and they mimic the covariance structure of the vectors $\boldsymbol{\xi}$ and $\boldsymbol{\xi}^\circ$: $\overline{\boldsymbol{\xi}} \sim \mathcal{N}(0, \operatorname{Var}\boldsymbol{\xi})$, $\overline{\boldsymbol{\xi}}^\circ \sim \mathcal{N}\bigl(0, \operatorname{Var}\{\boldsymbol{\xi}^\circ \,\big|\, \boldsymbol{Y}\}\bigr)$.

The upper line of the scheme corresponds to the $\boldsymbol{Y}$ -world, the lower line - to the bootstrap world. In both lines we apply two steps for approximating the corresponding likelihood ratio statistics. The first approximating step is the non-asymptotic square-root Wilks theorem: the bound (1.5) for the $\boldsymbol{Y}$ case and a similar statement for the bootstrap case, which is obtained in Theorem A.4, Section A.2.

The next step is called *Gaussian approximation* (GAR) which means that the distribution of the Euclidean norm $\|\boldsymbol{\xi}\|$ of a centered random vector $\boldsymbol{\xi}$ is close to the distribution of the similar norm of a Gaussian vector $\|\overline{\boldsymbol{\xi}}\|$ with the same covariance matrix as $\boldsymbol{\xi}$. A similar statement holds for the vector $\boldsymbol{\xi}^\circ$. Thus, the initial problem of comparing the distributions of the likelihood ratio statistics is reduced to the comparison of the distributions of the Euclidean norms of two centered normal vectors $\overline{\boldsymbol{\xi}}$ and $\overline{\boldsymbol{\xi}}^\circ$ (Gaussian comparison). This last step links their distributions and encloses the approximating scheme. The Gaussian comparison step is done by computing the Kullback-Leibler divergence between two multivariate Gaussian distributions (i.e. by comparison of the covariance matrices of $\nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta}^*)$ and $\nabla_{\boldsymbol{\theta}} L^\circ(\boldsymbol{\theta}^*)$) and applying Pinsker's inequality (Lemma 5.7). At this point we need to introduce the "small modeling bias" condition $\mathbf{(SmB)}$ from Section 4.2. It is formulated in terms of the following nonnegative-definite $p \times p$ symmetric matrices:

$$H_0^2 \stackrel{\text{def}}{=} \sum\nolimits_{i=1}^n I\!\!E\left[\nabla_{\boldsymbol{\theta}}\ell_i(\boldsymbol{\theta}^*)\nabla_{\boldsymbol{\theta}}\ell_i(\boldsymbol{\theta}^*)^\top\right], \tag{1.7}$$

$$B_0^2 \stackrel{\text{def}}{=} \sum\nolimits_{i=1}^n I\!\!E\left[\nabla_{\boldsymbol{\theta}}\ell_i(\boldsymbol{\theta}^*)\right] I\!\!E\left[\nabla_{\boldsymbol{\theta}}\ell_i(\boldsymbol{\theta}^*)\right]^\top, \tag{1.8}$$

so that $\operatorname{Var}\{\nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta}^*)\} = H_0^2 - B_0^2$. If the parametric assumption (1.1) is true or if the data $\boldsymbol{Y}$ are i.i.d., then it holds $I\!\!E\left[\nabla_{\boldsymbol{\theta}}\ell_i(\boldsymbol{\theta}^*)\right] \equiv 0$ and $B_0^2 = 0$. The $\mathbf{(SmB)}$ condition roughly means that the bias term B_0^2 is small relative to H_0^2. Below we show that the Kullback-Leibler distance between the distributions of two Gaussian vectors $\overline{\boldsymbol{\xi}}$ and $\overline{\boldsymbol{\xi}}^\circ$ is bounded by $p\|H_0^{-1}B_0^2H_0^{-1}\|^2/2$. The $\mathbf{(SmB)}$ condition precisely means that this quantity is small. We consider two situations: when the condition $\mathbf{(SmB)}$ is fulfilled and when it is not. Theorem 2.1 in Section 2 deals with the first case, it provides the cumulative error term for the coverage probability of the confidence set (1.3), taken at the $(1-\alpha)$ -quantile computed with the multiplier bootstrap procedure. The proof of this result (see Section A.3) summarizes the steps of scheme (1.6). The biggest term in the full error is induced by Gaussian approximation and requires the ratio p^3/n to be small. In the case of a "large modelling bias" i.e., when $\mathbf{(SmB)}$ does not hold, the multiplier bootstrap procedure continues to apply. It turns out that the bootstrap quantile increases with the growing modelling bias, hence, the confidence set based on it remains valid, however, it may become conservative. This result is given in Theorem 2.4 of Section 2. The problems of Gaussian approximation and comparison for the Euclidean norm are considered in Sections 5.2 and 5.4 in general terms independently of the statistical setting of the paper, and might be interesting by themselves. Section 5.4 presents also an anti-concentration inequality for the Euclidean norm of a Gaussian vector. This inequality shows how the deviation probability changes with a threshold. The general results on GAR are summarized in Theorem 5.1 and

restated in Proposition A.9 for the setting of scheme (1.6). These results are also non-asymptotic with explicit errors and apply under the condition that the ratio p^3/n to be small.

In Theorem 2.3 we consider the case of a scalar parameter $p = 1$ with an improved error term. Furthermore in Section 2.1 we propose a modified version of a quantile function based on a smoothed probability distribution. In this case the obtained error term is also better, than in the general result.

Notations: $\|\cdot\|$ denotes Euclidean norm for vectors and spectral norm for matrices; $\mathtt{C}$ is a generic constant. The value $\mathtt{x} > 0$ describes our tolerance level: all the results will be valid on a random set of probability ($1 - C\mathrm{e}^{-\mathtt{x}}$) for an explicit constant C. Everywhere we give explicit error bounds and show how they depend on p and n for the case of the i.i.d. observations $Y_1, \ldots, Y_n$ and $\mathtt{x} \le \mathtt{C} \log n$. More details on it are given in Section 4.3.

The paper is organized as follows: the main results are stated in Section 2, their proofs are given in Sections A.3, A.4 and A.5; Section 3 contains numerical results for misspecified constant and logistic regressions. In Section 4 we give all the necessary conditions and provide an information about dependency of the involved terms on n and p. Section 5 collects some useful statements on Gaussian approximation and Gaussian comparison.

2 Multiplier bootstrap procedure

Let $\ell_i(\boldsymbol{\theta})$ denote the parametric log-density of the i-th observation:

$$\ell_i(\boldsymbol{\theta}) \stackrel{\text{def}}{=} \log\left\{\frac{d\mathbb{P}_{\boldsymbol{\theta}}}{d\mu_0}(Y_i)\right\},$$

then $L(\boldsymbol{\theta}) = \sum_{i=1}^n \ell_i(\boldsymbol{\theta})$. Consider i.i.d. scalar random variables u_i independent of $\boldsymbol{Y}$ with continuous c.d.f., $\mathbb{E}u_i = 1$, $\operatorname{Var} u_i = 1$, $\mathbb{E}\exp(u_i) < \infty$ for all $i = 1, \ldots, n$. Multiply the summands of the likelihood function $L(\boldsymbol{\theta})$ with the new random variables:

$$L^\circ(\boldsymbol{\theta}) \stackrel{\text{def}}{=} \sum_{i=1}^n \ell_i(\boldsymbol{\theta}) u_i,$$

then it holds $\mathbb{E}^\circ L^\circ(\boldsymbol{\theta}) = L(\boldsymbol{\theta})$, where $\mathbb{E}^\circ$ stands for the conditional expectation given $\boldsymbol{Y}$:

$$\mathbb{E}^\circ(\cdot) \stackrel{\text{def}}{=} \mathbb{E}(\cdot|\boldsymbol{Y}), \qquad \mathbb{P}^\circ(\cdot) \stackrel{\text{def}}{=} \mathbb{P}(\cdot|\boldsymbol{Y}).$$

Therefore, the quasi MLE for the $\boldsymbol{Y}$-world is a target parameter for the bootstrap world:

$$\operatorname{argmax}_{\boldsymbol{\theta}\in\Theta} \mathbb{E}^\circ L^\circ(\boldsymbol{\theta}) = \operatorname{argmax}_{\boldsymbol{\theta}\in\Theta} L(\boldsymbol{\theta}) = \widetilde{\boldsymbol{\theta}}.$$

The corresponding quasi MLE under the conditional measure $\mathbb{P}^\circ$ is defined

$$\widetilde{\boldsymbol{\theta}}^\circ \stackrel{\text{def}}{=} \operatorname{argmax}_{\boldsymbol{\theta}\in\Theta} L^\circ(\boldsymbol{\theta}).$$

The likelihood ratio statistic in the bootstrap world is equal to $L^\circ(\widetilde{\boldsymbol{\theta}}^\circ) - L^\circ(\widetilde{\boldsymbol{\theta}})$, where all the elements: the function $L^\circ(\boldsymbol{\theta})$ and the arguments $\widetilde{\boldsymbol{\theta}}^\circ$, $\widetilde{\boldsymbol{\theta}}$ are known and available for computation.

Let $1-\alpha \in (0,1)$ be a fixed desirable confidence level of the set $\mathcal{E}(\mathfrak{z})$:

$$I\!P\left(\boldsymbol{\theta}^* \in \mathcal{E}(\mathfrak{z})\right) \geq 1-\alpha. \tag{2.1}$$

Here the parameter $\mathfrak{z} \geq 0$ determines the size of the confidence set. Usually we are interested in finding a set of the smallest possible diameter satisfying this property. This leads to the problem of fixing the minimal possible value of $\mathfrak{z}$ such that (2.1) is fulfilled. Let $\mathfrak{z}_\alpha$ denote the upper α-quantile of the square-root likelihood ratio statistic:

$$\mathfrak{z}_\alpha \stackrel{\text{def}}{=} \min\left\{\mathfrak{z} \geq 0\colon I\!P\left(L(\widetilde{\boldsymbol{\theta}}) - L(\boldsymbol{\theta}^*) > \mathfrak{z}^2/2\right) \leq \alpha\right\}. \tag{2.2}$$

This means, that $\mathfrak{z}_\alpha$ is exactly the value of our interest. Estimation of $\mathfrak{z}_\alpha$ leads to recovering of the distribution of $L(\widetilde{\boldsymbol{\theta}}) - L(\boldsymbol{\theta}^*)$. The multiplier bootstrap procedure consists of generating a large number of independent samples $\{u_1, \ldots, u_n\}$ and computing from them the empirical distribution function of $L^\circ(\widetilde{\boldsymbol{\theta}}^\circ) - L^\circ(\widetilde{\boldsymbol{\theta}})$. By this procedure we can estimate $\mathfrak{z}^\circ_\alpha$, the upper α-quantile of $\sqrt{2L^\circ(\widetilde{\boldsymbol{\theta}}^\circ) - 2L^\circ(\widetilde{\boldsymbol{\theta}})}$:

$$\mathfrak{z}^\circ_\alpha \stackrel{\text{def}}{=} \min\left\{\mathfrak{z} \geq 0\colon I\!P^\circ\left(L^\circ(\widetilde{\boldsymbol{\theta}}^\circ) - L^\circ(\widetilde{\boldsymbol{\theta}}) > \mathfrak{z}^2/2\right) = \alpha\right\}. \tag{2.3}$$

Theorem 2.1 (Validity of the bootstrap under a small modeling bias)**.** *Let the conditions of Section 4 be fulfilled. It holds with probability* $\geq 1 - 12\mathrm{e}^{-\mathtt{x}}$ *for* $\mathfrak{z}^\circ_\alpha \geq \max\{2, \sqrt{p}\} + \mathtt{C}(p+\mathtt{x})/\sqrt{n}$*:*

$$\left| I\!P\left(L(\widetilde{\boldsymbol{\theta}}) - L(\boldsymbol{\theta}^*) > (\mathfrak{z}^\circ_\alpha)^2/2\right) - \alpha \right| \leq \Delta_{\mathrm{full}}, \tag{2.4}$$

where $\Delta_{\mathrm{full}} \leq \mathtt{C}\{(p+\mathtt{x})^3/n\}^{1/8}$ *in the case 4.3. An explicit definition of the error term* Δ_{full} *is given in the proof (see* (A.26)*,* (A.27) *in Section A.3).*

The term Δ_{full} can be viewed as a sum of the error terms corresponding to each step in the scheme (1.6). The largest error term equal to $\mathtt{C}\{(p+\mathtt{x})^3/n\}^{1/8}$ is induced by GAR. This error rate is not always optimal for GAR, e.g. in the case of $p=1$ or for the i.i.d. observations (see Remark 5.2). In Theorems 2.3 and 2.5 the rate is $\mathtt{C}\{(p+\mathtt{x})^3/n\}^{1/2}$.

In view of definition (1.3) of the likelihood-based confidence set Theorem 2.1 implies the following

Corollary 2.2 (Coverage probability error)**.** *Under the conditions of Theorem 2.1 it holds:*

$$\left| I\!P\left\{\boldsymbol{\theta}^* \in \mathcal{E}\left(\mathfrak{z}^\circ_\alpha\right)\right\} - (1-\alpha)\right| \leq \Delta_{\mathrm{full}}.$$

REMARK 2.1 (Critical dimension). The error term Δ_{full} depends on the ratio p^3/n. The bootstrap validity can be only stated if this ratio is small. The obtained error bound seems to be mainly of theoretical interest, because the condition "$(p^3/n)^{1/8}$ is small" may require a huge sample. However, it provides some qualitative information about the bootstrap behavior as the parameter dimension grows. Our numerical results show that the accuracy of bootstrap approximation is very reasonable in a variety of examples.

In the following theorem we consider the case of a scalar parameter $p = 1$. The obtained error rate is $1/\sqrt{n}$, which is sharper, than $1/n^{1/8}$. Instead of the GAR for the Euclidean norm from Section 5 we use here Berry-Esseen theorem (see also Remark 5.2).

Theorem 2.3 (The case of $p = 1$, using Berry-Esseen theorem)**.** *Let the conditions of Section 4 be fulfilled. It holds with probability* $\geq 1 - 12\mathrm{e}^{-\mathtt{x}}$ *for* $\mathfrak{z}_\alpha^\circ \geq 1 + \mathtt{C}(1+\mathtt{x})/\sqrt{n}$ *:*

$$\left| I\!P\left(L(\widetilde{\boldsymbol{\theta}}) - L(\boldsymbol{\theta}^*) > (\mathfrak{z}_\alpha^\circ)^2/2 \right) - \alpha \right| \leq \Delta_{\text{B.E.,full}}, \tag{2.5}$$

where $\Delta_{\text{B.E.,full}} \leq \mathtt{C}(1+\mathtt{x})/\sqrt{n}$ *in the case 4.3. An explicit definition of* $\Delta_{\text{B.E.,full}}$ *is given in* (A.28) *in Section A.3.*

REMARK 2.2 (Bootstrap validity and weak convergence). The standard way of proving the bootstrap validity is based on weak convergence arguments; see e.g. Mammen (1992), van ver Vaart and Wellner (1996), Janssen and Pauls (2003), Chatterjee and Bose (2005). If the statistic $L(\widetilde{\boldsymbol{\theta}}) - L(\boldsymbol{\theta}^*)$ weakly converges to a χ^2-type distribution, one can state an asymptotic version of the results (2.4), (2.5). Our way is based on a kind of non-asymptotic Gaussian approximation and Gaussian comparison for random vectors and allows to get explicit error terms.

REMARK 2.3 (Use of Edgeworth expansion). The classical results on confidence sets for the mean of population states the accuracy of order $1/n$ based on the second order Edgeworth expansion Hall (1992). Unfortunately, if the considered parametric model can be misspecified, even the leading term is affected by the modeling bias, and the use of Edgeworth expansion cannot help in improving the bootstrap accuracy.

REMARK 2.4 (Choice of the weights). In our construction, similarly to Chatterjee and Bose (2005), we apply a general distribution of the bootstrap weights u_i under some moment conditions. One particularly can use Gaussian multipliers as suggested by Chernozhukov et al. (2013). This leads to the exact Gaussian distribution of the vectors $\boldsymbol{\xi}^\circ$ and is helpful to avoid one step of Gaussian approximation for these vectors.

Now we discuss the impact of modeling bias, which comes from a possible misspecification of the parametric model. As explained by the approximating diagram (1.6), the distance between the distributions of the likelihood ratio statistics can be characterized via the distance between two multivariate normal distributions. To state the result let us recall the definition of the full Fisher information matrix $D_0^2 \stackrel{\text{def}}{=} -\nabla^2_{\boldsymbol{\theta}} I\!E L(\boldsymbol{\theta}^*)$. For the matrices H_0^2 and B_0^2, given in (1.7) and (1.8), it holds $H_0^2 > B_0^2 \geq 0$. If the parametric assumption (1.1) is true or in the case of an i.i.d. sample $\boldsymbol{Y}$, $B_0^2 = 0$. Under the condition $(\mathbf{SmB})$ $\|H_0^{-1} B_0^2 H_0^{-1}\|$ enters linearly in the error term Δ_{full} in Theorem 2.1.

The first statement in Theorem 2.4 below says that the effective coverage probability of the confidence set based on the multiplier bootstrap is *larger* than the nominal coverage probability up to the error term $\Delta_{\text{b,full}} \leq \mathtt{C}\{(p+\mathtt{x})^3/n\}^{1/8}$. The inequalities in the second part of Theorem 2.4 prove the *conservativeness of the bootstrap quantiles*: the quantity $\sqrt{\mathrm{tr}\{D_0^{-1} H_0^2 D_0^{-1}\}} - \sqrt{\mathrm{tr}\{D_0^{-1}(H_0^2 - B_0^2) D_0^{-1}\}} \geq 0$ increases with the growing modeling bias.

Theorem 2.4 (Performance of the bootstrap for a large modeling bias)**.** *Under the conditions of Section 4 except for* $(\mathbf{SmB})$ *it holds with probability* $\geq 1 - 14\mathrm{e}^{-\mathtt{x}}$ *for* $\mathfrak{z}, \mathfrak{z}_\alpha^\circ \geq \max\{2, \sqrt{p}\} + \mathtt{C}(p+\mathtt{x})/\sqrt{n}$

1. $$I\!P\left(L(\widetilde{\boldsymbol{\theta}}) - L(\boldsymbol{\theta}^*) > \mathfrak{z}^2/2\right) \le I\!P^\circ\left(L^\circ(\widetilde{\boldsymbol{\theta}}^\circ) - L^\circ(\widetilde{\boldsymbol{\theta}}) > \mathfrak{z}^2/2\right) + \Delta_{\mathrm{b,\,full}}.$$

2. $$\mathfrak{z}^\circ_\alpha \ge \mathfrak{z}_{(\alpha+\Delta_{\mathrm{b,\,full}})} + \sqrt{\mathrm{tr}\{D_0^{-1}H_0^2D_0^{-1}\}} - \sqrt{\mathrm{tr}\{D_0^{-1}(H_0^2 - B_0^2)D_0^{-1}\}} - \Delta_{\mathtt{qf},1},$$
$$\mathfrak{z}^\circ_\alpha \le \mathfrak{z}_{(\alpha-\Delta_{\mathrm{b,\,full}})} + \sqrt{\mathrm{tr}\{D_0^{-1}H_0^2D_0^{-1}\}} - \sqrt{\mathrm{tr}\{D_0^{-1}(H_0^2 - B_0^2)D_0^{-1}\}} + \Delta_{\mathtt{qf},2}.$$

The term $\Delta_{\mathrm{b,\,full}} \le \mathtt{C}\{(p+\mathtt{x})^3/n\}^{1/8}$ *is given in* (A.30) *in Section A.4. The positive values* $\Delta_{\mathtt{qf},1}, \Delta_{\mathtt{qf},2}$ *are given in* (A.34), (A.33) *in Section A.4, they are bounded from above with* $(\mathfrak{a}^2 + \mathfrak{a}_B^2)(\sqrt{8\mathtt{x}p} + 6\mathtt{x})$ *for the constants* $\mathfrak{a}^2, \mathfrak{a}_B^2 > 0$ *from conditions* $(\mathcal{I})$, $(\mathcal{I}_B)$.

REMARK 2.5. There exists some literature on robust (and heteroscedasticity robust) bootstrap procedures; see e.g. Mammen (1993), Aerts and Claeskens (2001), Kline and Santos (2012). However, up to our knowledge there are no robust bootstrap procedures for the likelihood ratio statistic, most of the results compare the distribution of the estimator obtained from estimating equations, or Wald / score test statistics with their bootstrap counterparts in the i.i.d. setup. In our context this would correspond to the noise misspecification in the log-likelihood function and it is addressed automatically by the multiplier bootstrap. Our notion of modeling bias includes the situation when the target value $\boldsymbol{\theta}^*$ from (1.4) only defines a projection (the best parametric fit) of the data distribution. In particularly, the quantities $I\!E\nabla_{\boldsymbol{\theta}}\ell_i(\boldsymbol{\theta}^*)$ for different i do not necessarily vanish yielding a significant modeling bias. Similar notion of misspecification is used in the literature on Generalized Method of Moments; see e.g. Hall (2005). Chapter 5 therein considers the hypothesis testing problem with two kinds of misspecification: local and non-local, which would correspond to our small and large modeling bias cases.

An interesting message of Theorem 2.4 is that the multiplier bootstrap procedure ensures a prescribed coverage level for this target value $\boldsymbol{\theta}^*$ even without small modeling bias restriction, however, in this case the method is somehow conservative because the modeling bias is transferred into the additional variance in the bootstrap world. The numerical experiments in Section 3 agree with this result.

2.1 Smoothed version of a quantile function

This section briefly discusses the use of a smoothed quantile function. The $(1-\alpha)$-quantile of $\sqrt{2L(\widetilde{\boldsymbol{\theta}}) - 2L(\boldsymbol{\theta}^*)}$ is defined as

$$\begin{aligned}\mathfrak{z}_\alpha &\stackrel{\mathrm{def}}{=} \min\left\{\mathfrak{z} \ge 0\colon I\!P\left(L(\widetilde{\boldsymbol{\theta}}) - L(\boldsymbol{\theta}^*) > \mathfrak{z}^2/2\right) \le \alpha\right\} \\ &= \min\left\{\mathfrak{z} \ge 0\colon I\!E\,\mathbb{1}\left\{L(\widetilde{\boldsymbol{\theta}}) - L(\boldsymbol{\theta}^*) > \mathfrak{z}^2/2\right\} \le \alpha\right\}.\end{aligned}$$

Introduce for $x \geq 0$ and $z, \Delta > 0$ the following function

$$g_{\Delta}(x, z) \stackrel{\text{def}}{=} g\left(\frac{1}{2\Delta z} \left(x^2 - z^2 \right) \right), \tag{2.6}$$

where $g(\cdot) \in C^2(\mathbb{R})$ is a non-negative function, which grows monotonously from 0 to 1, $g(x) = 0$ for $x \leq 0$ and $g(x) = 1$ for $x \geq 1$, therefore:

$$\mathbb{1}\{x \geq 1\} \leq g(x) \leq \mathbb{1}\{x \geq 0\} \leq g(x+1).$$

An example of such function is given in (5.9). In (5.10) it is shown

$$\mathbb{1}\{x - z \geq \Delta\} \leq g_{\Delta}(x, z) \leq \mathbb{1}(x - z \geq 0) \leq g_{\Delta}(x, z + \Delta).$$

This approximation is used in the proofs of Theorems 2.1 and 2.4 in the part of Gaussian approximation of Euclidean norm of a sum of independent vectors (see Section 5.2) yielding the error rate $(p^3/n)^{1/8}$ in the final bound (Theorems 2.1, 5.1). The next result shows that the use of a smoothed quantile function helps to improve the accuracy of bootstrap approximation: it becomes $(p^3/n)^{1/2}$ instead of $(p^3/n)^{1/8}$. The reason is that we do not need to account for the error induced by a smooth approximation of the indicator function.

Theorem 2.5 (Validity of the bootstrap in the smoothed case under $(\mathbf{SmB})$ condition)**.** *Let the conditions of Section 4 be fulfilled. It holds with probability* $\geq 1 - 12\mathrm{e}^{-\mathtt{x}}$ *for* $\mathfrak{z} \geq \max\{2, \sqrt{p}\} + \mathtt{C}(p + \mathtt{x})/\sqrt{n}$ *and* $\Delta \in (0, 0.22]$ *:*

$$\left| I\!\!E g_{\Delta}\left(\sqrt{2L(\widetilde{\boldsymbol{\theta}}) - 2L(\boldsymbol{\theta}^*)}, \mathfrak{z} \right) - I\!\!E^{\circ} g_{\Delta}\left(\sqrt{2L^{\circ}(\widetilde{\boldsymbol{\theta}}^{\circ}) - 2L^{\circ}(\widetilde{\boldsymbol{\theta}})}, \mathfrak{z} \right) \right| \leq \Delta_{\mathrm{sm}},$$

where $\Delta_{\mathrm{sm}} \leq \mathtt{C}\{(p + \mathtt{x})^3/n\}^{1/2} \Delta^{-3}$ *in the case 4.3. An explicit definition of* Δ_{sm} *is given in* (A.38)*,* (A.39) *in Section A.5.*

The modified bootstrap quantile function reads as

$$\mathfrak{z}^{\circ}_{\Delta,\alpha} \stackrel{\text{def}}{=} \min\left\{ \mathfrak{z} \geq 0 \colon I\!\!E^{\circ} g_{\Delta}\left(\sqrt{2L^{\circ}(\widetilde{\boldsymbol{\theta}}^{\circ}) - 2L^{\circ}(\widetilde{\boldsymbol{\theta}})}, \mathfrak{z} \right) = \alpha \right\}.$$

3 Numerical results

This section illustrates the performance of the multiplier bootstrap for some artificial examples. We especially aim to address the issues of noise misspecification and of increasing modeling bias. In all the experiments we took 10^4 data samples for estimation of empirical c.d.f. of $\sqrt{2L(\widetilde{\theta}) - 2L(\theta^*)}$, 10^4 $\{u_1, \ldots, u_n\}$ samples and 10^4 data samples for the estimation of the quantiles of $\sqrt{2L^{\circ}(\widetilde{\theta}^{\circ}) - 2L^{\circ}(\widetilde{\theta})}$. All sample sizes are $n = 50$. It should be mentioned that the obtained results are nicely consistent with the theoretical statements.

3.1 Computational error

Here we check numerically, how well the multiplier procedure works in the case of the correct model. Let the i.i.d. data follow the distribution $Y_i \sim \mathcal{N}(2,1)$, $i = 1, \ldots, n$. The true likelihood function is $L(\theta) = -\sum_{i=1}^{n}(Y_i - \theta)^2/2$.

Table 1 shows the effective coverage probabilities of the quantiles estimated using the multiplier bootstrap. The second line contains the range of the nominal confidence levels: $0.99, \ldots, 0.75$. The first left column describes the distribution of the bootstrap weights: $\mathcal{N}(1,1)$ or $exp(1)$. The 3-d and the 4-th lines show the frequency of the event: "the real likelihood ratio $\leq$ the quantile of the bootstrap likelihood ratio".

Table 1: Coverage probabilities for the correct i.i.d. model

	Confidence levels					
$\mathcal{L}(u_i)$	$\mathbf{0.99}$	$\mathbf{0.95}$	$\mathbf{0.90}$	$\mathbf{0.85}$	$\mathbf{0.80}$	$\mathbf{0.75}$
$exp(1)$	0.99	0.94	0.89	0.83	0.78	0.73
$\mathcal{N}(1,1)$	0.99	0.95	0.89	0.84	0.80	0.75

3.2 Constant regression with misspecified heteroscedastic errors

Here we show on a constant regression model that the quality of the confidence sets obtained by the multiplier bootstrap procedure is not significantly deteriorated by misspecified heteroscedastic errors. Let the data be defined as $Y_i = 2 + \sigma_i \varepsilon_i$, $i = 1, \ldots, n$. The i.i.d. random variables $\varepsilon_i \sim Lap(0, 2^{-1/2})$ are s.t. $I\!E(\varepsilon_i) = 0$, $\mathrm{Var}(\varepsilon_i) = 1$. The coefficients σ_i are deterministic: $\sigma_i \stackrel{\mathrm{def}}{=} 0.5\{4 - i \pmod 4\}$. The quasi-likelihood function is the same as in the previous experiment: $L(\theta) = -\sum_{i=1}^{n}(Y_i - \theta)^2/2$, i.e. it is misspecified, since it corresponds to the i.i.d. standard normal distribution. Table 2 describes the 2-nd experiment's results similarly to the Table 1.

Table 2: Coverage probabilities for the misspecified heteroscedastic noise

	Confidence levels					
$\mathcal{L}(u_i)$	$\mathbf{0.99}$	$\mathbf{0.95}$	$\mathbf{0.90}$	$\mathbf{0.85}$	$\mathbf{0.80}$	$\mathbf{0.75}$
$exp(1)$	0.98	0.93	0.87	0.82	0.77	0.72
$\mathcal{N}(1,1)$	0.98	0.94	0.88	0.83	0.78	0.73

3.3 Biased constant regression with misspecified errors

In the third experiment we consider biased regression with misspecified i.i.d. errors:

$$Y_i = \beta \sin(X_i) + \varepsilon_i, \quad \varepsilon_i \sim Lap(0, 2^{-1/2}), \text{ i.i.d},$$
$$X_i \text{ are equidistant in } [0, 2\pi].$$

Taking the likelihood function $L(\theta) = -\sum_{i=1}^{n}(Y_i - \theta)^2/2$ yields $\theta^* = 0$. Therefore, the larger is the deterministic amplitude $\beta > 0$, the bigger is bias of the mean constant regression. We consider two cases: $\beta = 0.25$ with fulfilled $\mathbf{(SmB)}$ condition and $\beta = 1.25$ when $\mathbf{(SmB)}$ does not hold. Table 3 shows that for the large bias quantiles yielded by the multiplier bootstrap are conservative. This conservative property of the multiplier bootstrap quantiles is

Table 3: Coverage probabilities for the misspecified biased regression

		Confidence levels					
$\mathcal{L}(u_i)$	β	$\mathbf{0.99}$	$\mathbf{0.95}$	$\mathbf{0.90}$	$\mathbf{0.85}$	$\mathbf{0.80}$	$\mathbf{0.75}$
$\mathcal{N}(1,1)$	0.25	0.98	0.94	0.89	0.84	0.79	0.74
	1.25	1.0	0.99	0.97	0.94	0.91	0.87

also illustrated with the graphs in Figure 3.1. They show the empirical distribution functions of the likelihood ratio statistics $L(\widetilde{\theta}) - L(\theta^*)$ and $L^\bullet(\widetilde{\theta}^\bullet) - L^\bullet(\widetilde{\theta})$ for $\beta = 0.25$ and $\beta = 1.25$. On the right graph for $\beta = 1.25$ the empirical distribution functions for the bootstrap case are smaller than the one for the $\boldsymbol{Y}$ case. It means that for the large bias the bootstrap quantiles are bigger than the $\boldsymbol{Y}$ quantiles, which increases the diameter of the confidence set based on the bootstrap quantiles. This confidence set remains valid, since it still contains the true parameter with a given confidence level.

Figure 3.2 shows the growth of the difference between the quantiles of $L^\bullet(\widetilde{\theta}^\bullet) - L^\bullet(\widetilde{\theta})$ and $L(\widetilde{\theta}) - L(\theta^*)$ with increasing β for the range of the confidence levels: $0.75, 0.8, \ldots, 0.99$.

3.4 Logistic regression with bias

In this example we consider logistic regression. Let the data come from the following distribution:

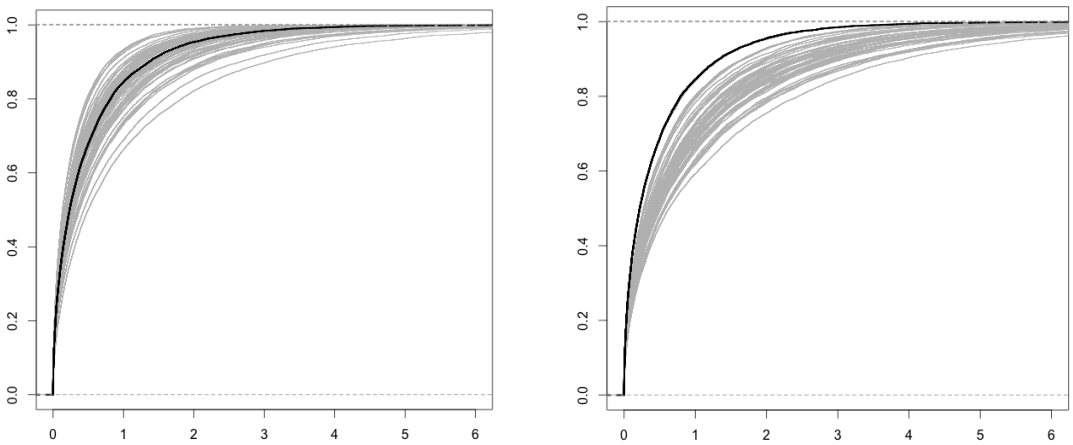
$$Y_i \sim Bernoulli(\beta X_i), \quad X_i \text{ are equidistant in } [0, 2], \ \beta \in (0, 1/2].$$

Consider the likelihood function corresponding to the i.i.d. observations:

$$L(\theta) = \sum_{i=1}^{n} \left\{ Y_i \theta - \log(1 + \mathrm{e}^\theta) \right\}.$$

By definition (1.4) $\theta^* = \log\{\beta/(1-\beta)\}$, bigger values of β induce larger modeling bias. The graphs below demonstrate the conservativeness of bootstrap quantiles. Here we consider two cases: $\beta = 0.1$ and $\beta = 0.5$. Similarly to the Example 3.3 in the case of the bigger β on

Figure 3.1: Empirical distribution functions of the likelihood ratios

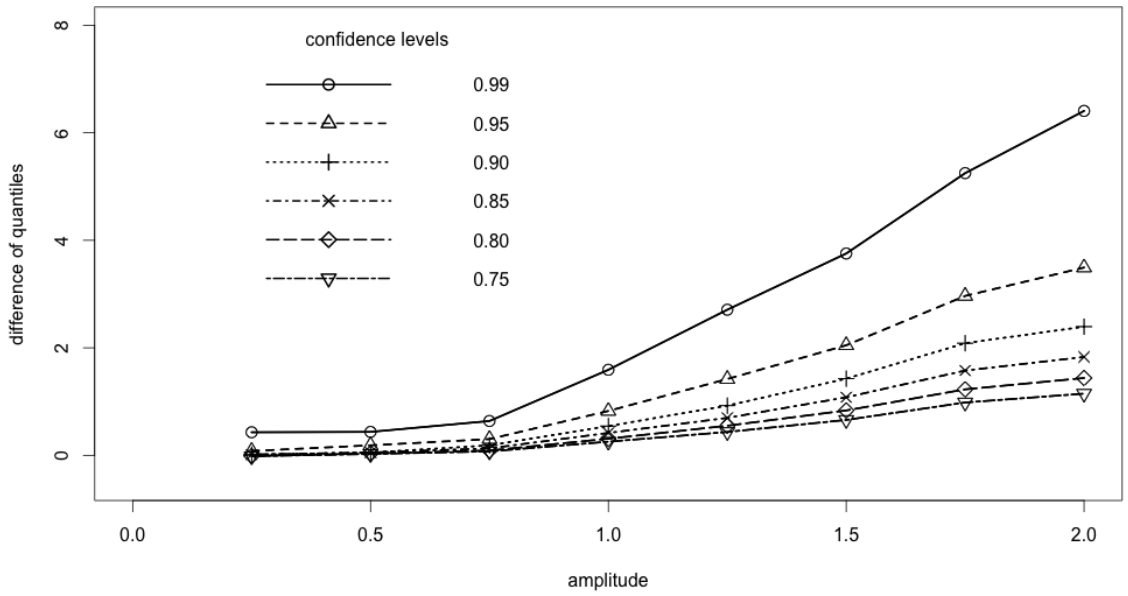


$Y_i = 0.25\sin(X_i) + Lap(0, 2^{-1/2})$, $n = 50$ $\quad$ $Y_i = 1.25\sin(X_i) + Lap(0, 2^{-1/2})$, $n = 50$

— empirical distribution function of $L(\widetilde{\theta}) - L(\theta^*)$ estimated with 10^4 $\boldsymbol{Y}$ samples

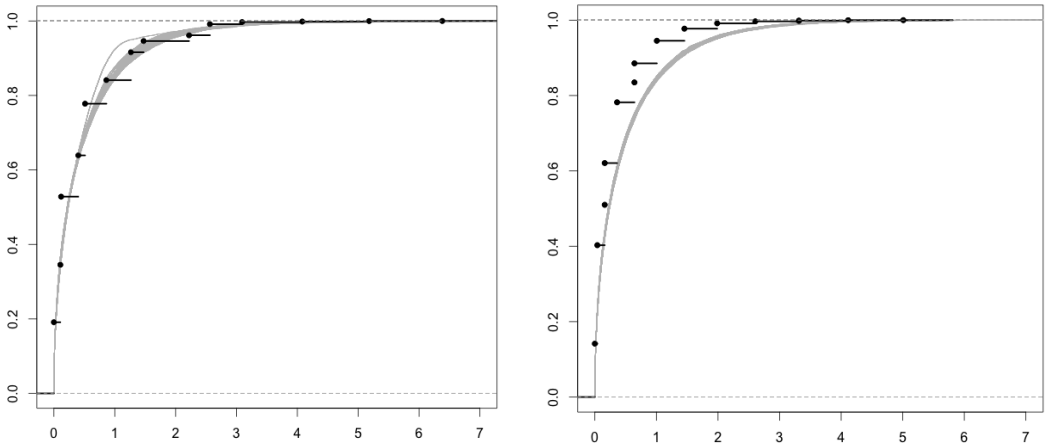
— 50 empirical distribution functions of $L^\bullet(\widetilde{\theta}^\bullet) - L^\bullet(\widetilde{\theta})$ estimated with 10^4 $\{u_i\} \sim exp(1)$ samples

Figure 3.2: The difference $\big($"Bootstrap quantile" $-$ "$\boldsymbol{Y}$-quantile"$\big)$ growing with modeling bias



the right graph in Figure 3.3 the empirical distribution functions of $L^\bullet(\widetilde{\theta}^\bullet) - L^\bullet(\widetilde{\theta})$ are smaller than the one for $L(\widetilde{\theta}) - L(\theta^*)$.

Figure 3.3:



$Y_i \sim Bernoulli(0.1X_i), n = 50$ $Y_i \sim Bernoulli(0.5X_i), n = 50$

— empirical distribution function of $L(\widetilde{\theta}) - L(\theta^*)$ estimated with 10^4 $\boldsymbol{Y}$ samples

— 50 empirical distribution functions of $L^{\circ}(\widetilde{\theta}^{\circ}) - L^{\circ}(\widetilde{\theta})$ estimated with 10^4 $\{u_i\} \sim exp(1)$ samples

4 Conditions

Here we state the conditions necessary for the main results. The conditions in Section 4.1 come from the general finite sample theory by Spokoiny (2012a), they are required for the results of Sections A.1 and A.2. Spokoiny (2012a) considers the examples of i.i.d. setup, generalized linear model and linear median regression providing a check of conditions from Section 4.1. The conditions in Section 4.2 are necessary to prove the results on multiplier bootstrap from Section 2.

4.1 Basic conditions

Introduce the stochastic part of the likelihood process: $\zeta(\boldsymbol{\theta}) \stackrel{\text{def}}{=} L(\boldsymbol{\theta}) - I\!\!E L(\boldsymbol{\theta})$, and its marginal summand: $\zeta_i(\boldsymbol{\theta}) \stackrel{\text{def}}{=} \ell_i(\boldsymbol{\theta}) - I\!\!E \ell_i(\boldsymbol{\theta})$.

$(\boldsymbol{ED_0})$ *There exist a positive-definite symmetric matrix* V_0^2 *and constants* $\mathtt{g} > 0, \nu_0 \geq 1$ *such that* $\operatorname{Var}\{\nabla_{\boldsymbol{\theta}} \zeta(\boldsymbol{\theta}^*)\} \leq V_0^2$ *and*

$$\sup_{\boldsymbol{\gamma} \in I\!R^p} \log I\!\!E \exp\left\{ \lambda \frac{\boldsymbol{\gamma}^\top \nabla_{\boldsymbol{\theta}} \zeta(\boldsymbol{\theta}^*)}{\|V_0 \boldsymbol{\gamma}\|} \right\} \leq \nu_0^2 \lambda^2 / 2, \qquad |\lambda| \leq \mathtt{g}.$$

$(\boldsymbol{ED_2})$ *There exists a constant* $\omega \geq 0$ *and for each* $\mathtt{r} > 0$ *a constant* $\mathtt{g}_2(\mathtt{r})$ *such that it*

holds for all $\boldsymbol{\theta} \in \Theta_0(\mathtt{r})$ *and for* $j = 1, 2$

$$\sup_{\substack{\boldsymbol{\gamma}_j \in I\!R^p \\ \|\boldsymbol{\gamma}_j\| \le 1}} \log I\!E \exp\left\{ \frac{\lambda}{\omega} \boldsymbol{\gamma}_1^\top D_0^{-1} \nabla^2_{\boldsymbol{\theta}} \zeta(\boldsymbol{\theta}) D_0^{-1} \boldsymbol{\gamma}_2 \right\} \le \nu_0^2 \lambda^2 / 2, \qquad |\lambda| \le \mathtt{g}_2(\mathtt{r}).$$

$(\boldsymbol{\mathcal{L}_0})$ *For each* $\mathtt{r} > 0$ *there exists a constant* $\delta(\mathtt{r}) \ge 0$ *such that for* $\mathtt{r} \le \mathtt{r}_0$ *(*$\mathtt{r}_0$ *comes from condition* (A.1) *of Theorem A.1 in Section A.1)* $\delta(\mathtt{r}) \le 1/2$ *, and for all* $\boldsymbol{\theta} \in \Theta_0(\mathtt{r})$ *it holds*

$$\| D_0^{-1} D^2(\boldsymbol{\theta}) D_0^{-1} - \boldsymbol{I}_p \| \le \delta(\mathtt{r}),$$

where $D^2(\boldsymbol{\theta}) \stackrel{\text{def}}{=} -\nabla^2_{\boldsymbol{\theta}} I\!E L(\boldsymbol{\theta})$ *,* $\Theta_0(\mathtt{r}) \stackrel{\text{def}}{=} \{\boldsymbol{\theta} : \|D_0(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\| \le \mathtt{r}\}$ *.*

$(\boldsymbol{\mathcal{I}})$ *There exists a constant* $\mathtt{a} > 0$ *s.t.* $\mathtt{a}^2 D_0^2 \ge V_0^2$ *.*

$(\boldsymbol{\mathcal{L}}\mathtt{r})$ *For each* $\mathtt{r} \ge \mathtt{r}_0$ *there exists a value* $\mathtt{b}(\mathtt{r}) > 0$ *s.t.* $\mathtt{r}\mathtt{b}(\mathtt{r}) \to \infty$ *for* $\mathtt{r} \to \infty$ *and* $\forall \boldsymbol{\theta} : \|D_0(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\| = \mathtt{r}$ *it holds*

$$-2\{I\!E L(\boldsymbol{\theta}) - I\!E L(\boldsymbol{\theta}^*)\} \ge \mathtt{r}^2 \mathtt{b}(\mathtt{r}).$$

4.2 Conditions required for the bootstrap validity

$(\mathbf{SmB})$ *There exists a constant* $\delta^2_{\mathrm{smb}} \in [0, 1/8]$ *such that it holds for all* $i = 1, \ldots, n$ *and the matrices* H_0^2 *,* B_0^2 *defined in* (1.7) *and* (1.8)*.*

$$\| H_0^{-1} B_0^2 H_0^{-1} \| \le \delta^2_{\mathrm{smb}} \le \mathtt{C} p n^{-1/2},$$

$(\boldsymbol{ED_{2m}})$ *For each* $\mathtt{r} > 0$ *,* $i = 1, \ldots, n$ *,* $j = 1, 2$ *and for all* $\boldsymbol{\theta} \in \Theta_0(\mathtt{r})$ *it holds for the values* $\omega \ge 0$ *and* $\mathtt{g}_2(\mathtt{r})$ *from the condition* $(\boldsymbol{ED_2})$ *:*

$$\sup_{\substack{\boldsymbol{\gamma}_j \in I\!R^p \\ \|\boldsymbol{\gamma}_j\| \le 1}} \log I\!E \exp\left\{ \frac{\lambda}{\omega} \boldsymbol{\gamma}_1^\top D_0^{-1} \nabla^2_{\boldsymbol{\theta}} \zeta_i(\boldsymbol{\theta}) D_0^{-1} \boldsymbol{\gamma}_2 \right\} \le \frac{\nu_0^2 \lambda^2}{2n}, \qquad |\lambda| \le \mathtt{g}_2(\mathtt{r}),$$

$(\boldsymbol{\mathcal{L}_{0m}})$ *For each* $\mathtt{r} > 0$ *,* $i = 1, \ldots, n$ *and for all* $\boldsymbol{\theta} \in \Theta_0(\mathtt{r})$ *there exists a constant* $\mathtt{C}_m(\mathtt{r}) \ge 0$ *such that*

$$\| D_0^{-1} \nabla^2_{\boldsymbol{\theta}} I\!E \ell_i(\boldsymbol{\theta}) D_0^{-1} \| \le \mathtt{C}_m(\mathtt{r}) n^{-1}.$$

$(\boldsymbol{\mathcal{L}_{3m}})$ *For all* $\boldsymbol{\theta} \in \Theta$ *and* $i = 1, \ldots, n$ *it holds* $\| D_0^{-1} \nabla^3_{\boldsymbol{\theta}} I\!E \ell_i(\boldsymbol{\theta}) D_0^{-1} \| \le \mathtt{C}$ *.*

$(\boldsymbol{\mathcal{I}_B})$ *There exists a constant* $\mathtt{a}_B^2 > 0$ *s.t.* $\mathtt{a}_B^2 D_0^2 \ge B_0^2$ *.*

$(\boldsymbol{SD_1})$ *There exists a constant* $0 \le \delta_v \le \mathtt{C}p/n$. *such that it holds for all* $i = 1, \ldots, n$ *with exponentially high probability*

$$\left\| H_0^{-1} \left\{ \nabla_{\boldsymbol{\theta}} \ell_i(\boldsymbol{\theta}^*) \nabla_{\boldsymbol{\theta}} \ell_i(\boldsymbol{\theta}^*)^\top - I\!\!E \left[\nabla_{\boldsymbol{\theta}} \ell_i(\boldsymbol{\theta}^*) \nabla_{\boldsymbol{\theta}} \ell_i(\boldsymbol{\theta}^*)^\top \right] \right\} H_0^{-1} \right\| \le \delta_v^2.$$

$(\boldsymbol{Eb})$ *The i.i.d. bootstrap weights* u_i *have continuous c.d.f., and it holds for all* $i = 1, \ldots, n$ *:* $I\!\!E^\circ u_i = 1$, $\mathrm{Var}^\circ u_i = 1$,

$$\log I\!\!E^\circ \exp\{\lambda(u_i - 1)\} \le \nu_0^2 \lambda^2 / 2, \qquad |\lambda| \le \mathtt{g}.$$

4.3 Dependence of the involved terms on the sample size and parameter dimension

Here we consider the case of the i.i.d. observations $Y_1, \ldots, Y_n$ and $\mathtt{x} = \mathtt{C} \log n$ in order to specify the dependence of the non-asymptotic bounds on n and p. Example 5.1 in Spokoiny (2012a) demonstrates that in this situation $\mathtt{g} = \mathtt{C}\sqrt{n}$ and $\omega = \mathtt{C}/\sqrt{n}$. then $\mathfrak{Z}(\mathtt{x}) = \mathtt{C}\sqrt{p + \mathtt{x}}$ for some constant $\mathtt{C} \ge 1.85$, for the function $\mathfrak{Z}(\mathtt{x})$ given in (A.3) in Section A.1. Similarly it can be checked that $\mathtt{g}_2(\mathtt{r})$ from condition $(\boldsymbol{ED_2})$ is proportional to $\sqrt{n}$: due to independency of the observations

$$\begin{aligned}
&\log I\!\!E \exp\left\{ \frac{\lambda}{\omega} \boldsymbol{\gamma}_1^\top D_0^{-1} \nabla^2_{\boldsymbol{\theta}} \zeta(\boldsymbol{\theta}) D_0^{-1} \boldsymbol{\gamma}_2 \right\} \\
&\quad = \sum\nolimits_{i=1}^n \log I\!\!E \exp\left\{ \frac{\lambda}{\sqrt{n}} \frac{1}{\omega/\sqrt{n}} \boldsymbol{\gamma}_1^\top d_0^{-1} \nabla^2_{\boldsymbol{\theta}} \zeta_i(\boldsymbol{\theta}) d_0^{-1} \boldsymbol{\gamma}_2 \right\} \\
&\quad \le n \frac{\lambda^2}{n} \mathtt{C} \quad \text{for } |\lambda| \le \overline{\mathtt{g}}_2(\mathtt{r}) \sqrt{n},
\end{aligned}$$

where $\zeta_i(\boldsymbol{\theta}) \stackrel{\mathrm{def}}{=} \ell_i(\boldsymbol{\theta}) - I\!\!E \ell_i(\boldsymbol{\theta})$, $d_0^2 \stackrel{\mathrm{def}}{=} -\nabla^2_{\boldsymbol{\theta}} I\!\!E \ell_i(\boldsymbol{\theta}^*)$ and $D_0^2 = n d_0^2$ in the i.i.d. case. Function $\overline{\mathtt{g}}_2(\mathtt{r})$ denotes the marginal analog of $\mathtt{g}_2(\mathtt{r})$.

Let us show, that for the value $\delta(\mathtt{r})$ from the condition $(\boldsymbol{\mathfrak{L}_0})$ it holds $\delta(\mathtt{r}) = \mathtt{C}\mathtt{r}/\sqrt{n}$. For some $\overline{\boldsymbol{\theta}}$

$$\begin{aligned}
\| D_0^{-1} D^2(\boldsymbol{\theta}) D_0^{-1} - \boldsymbol{I}_p \| &= \| D_0^{-1} (\boldsymbol{\theta}^* - \boldsymbol{\theta})^\top \nabla^3_{\boldsymbol{\theta}} I\!\!E L(\overline{\boldsymbol{\theta}}) D_0^{-1} \| \\
&= \| D_0^{-1} (\boldsymbol{\theta}^* - \boldsymbol{\theta})^\top D_0 D_0^{-1} \nabla^3_{\boldsymbol{\theta}} I\!\!E L(\overline{\boldsymbol{\theta}}) D_0^{-1} \| \\
&\le \mathtt{r} \| D_0^{-1} \| \| D_0^{-1} \nabla^3_{\boldsymbol{\theta}} I\!\!E L(\overline{\boldsymbol{\theta}}) D_0^{-1} \| \le \mathtt{C}\mathtt{r}/\sqrt{n} \quad \text{(by condition } (\boldsymbol{\mathfrak{L}_{3m}}) \text{)}.
\end{aligned}$$

Similarly $\mathtt{C}_m(\mathtt{r}) \le \mathtt{C}\mathtt{r}$ in condition $(\boldsymbol{\mathfrak{L}_{0m}})$.

If $\delta(\mathtt{r}) = \mathtt{C}\mathtt{r}/\sqrt{n}$ is sufficiently small, then the value $\mathtt{b}(\mathtt{r})$ from condition $(\boldsymbol{\mathfrak{L}\mathtt{r}})$ can be taken as $\mathtt{C}\{1 - \delta(\mathtt{r})\}^2$. Indeed, by $(\boldsymbol{\mathfrak{L}_0})$ and $(\boldsymbol{\mathfrak{L}\mathtt{r}})$ for $\boldsymbol{\theta}$: $\| D_0 (\boldsymbol{\theta} - \boldsymbol{\theta}^*) \| = \mathtt{r}$

$$-2 \{ I\!\!E L(\boldsymbol{\theta}) - I\!\!E L(\boldsymbol{\theta}^*) \} \ge \mathtt{r}^2 \left\{ 1 - \delta^2(\mathtt{r}) \right\}.$$

Therefore, if $\delta(\mathtt{r})$ is small, then $\mathtt{b}(\mathtt{r}) \stackrel{\text{def}}{=} \mathtt{C}\{1-\delta^2(\mathtt{r})\} \approx const$. Due to the obtained orders the conditions (A.1) and (A.17) of Theorems A.1 and A.6 on concentration of the MLEs $\widetilde{\boldsymbol{\theta}}$, $\widetilde{\boldsymbol{\theta}}^{\circ}$ require $\mathtt{r}_0 \geq \mathtt{C}\sqrt{p+\mathtt{x}}$.

5 Approximation of distributions of ℓ_2 norms of sums of independent vectors

Consider two samples $\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_n$ and $\boldsymbol{\psi}_1, \ldots, \boldsymbol{\psi}_n$, each consists of centered independent random vectors in $I\!R^p$ with nearly the same second moments. This section explains how one can quantify the closeness in distribution between the norms of $\boldsymbol{\phi} = \sum_i \boldsymbol{\phi}_i$ and of $\boldsymbol{\psi} = \sum_i \boldsymbol{\psi}_i$. Suppose that

$$I\!E\boldsymbol{\phi}_i = I\!E\boldsymbol{\psi}_i = 0, \quad \operatorname{Var}\boldsymbol{\phi}_i = \Sigma_i, \quad \operatorname{Var}\boldsymbol{\psi}_i = \breve{\Sigma}_i, \quad i = 1, \ldots, n.$$

Let also

$$\boldsymbol{\phi} \stackrel{\text{def}}{=} \sum\nolimits_{i=1}^n \boldsymbol{\phi}_i, \quad \boldsymbol{\psi} \stackrel{\text{def}}{=} \sum\nolimits_{i=1}^n \boldsymbol{\psi}_i, \tag{5.1}$$

$$\Sigma \stackrel{\text{def}}{=} \operatorname{Var}\boldsymbol{\phi} = \sum\nolimits_{i=1}^n \Sigma_i, \quad \breve{\Sigma} \stackrel{\text{def}}{=} \operatorname{Var}\boldsymbol{\psi} = \sum\nolimits_{i=1}^n \breve{\Sigma}_i. \tag{5.2}$$

Also introduce multivariate Gaussian vectors $\overline{\boldsymbol{\phi}}_i, \overline{\boldsymbol{\psi}}_i$ which are mutually independent for $i = 1, \ldots, n$ and

$$\begin{aligned} \overline{\boldsymbol{\phi}}_i \sim \mathcal{N}(0, \Sigma_i), \quad \overline{\boldsymbol{\psi}}_i \sim \mathcal{N}(0, \breve{\Sigma}_i), \\ \overline{\boldsymbol{\phi}} \stackrel{\text{def}}{=} \sum\nolimits_{i=1}^n \overline{\boldsymbol{\phi}}_i \sim \mathcal{N}(0, \Sigma), \quad \overline{\boldsymbol{\psi}} \stackrel{\text{def}}{=} \sum\nolimits_{i=1}^n \overline{\boldsymbol{\psi}}_i \sim \mathcal{N}(0, \breve{\Sigma}). \end{aligned} \tag{5.3}$$

The bar sign for a vector stands here for a normal distribution. The following theorem gives the conditions on Σ and $\breve{\Sigma}$ which ensure that $\|\boldsymbol{\phi}\|$ and $\|\boldsymbol{\psi}\|$ are close to each other in distribution. It also presents a general result on Gaussian approximation of $\|\boldsymbol{\phi}\|$ with $\|\overline{\boldsymbol{\phi}}\|$.

Introduce the following deterministic values, which are supposed to be finite:

$$\delta_n \stackrel{\text{def}}{=} \frac{1}{2}\sum_{i=1}^n I\!E\left(\|\boldsymbol{\phi}_i\|^3 + \|\overline{\boldsymbol{\phi}}_i\|^3\right), \quad \breve{\delta}_n \stackrel{\text{def}}{=} \frac{1}{2}\sum_{i=1}^n I\!E\left(\|\boldsymbol{\psi}_i\|^3 + \|\overline{\boldsymbol{\psi}}_i\|^3\right). \tag{5.4}$$

Theorem 5.1. *Assume for the covariance matrices defined in* (5.2) *that*

$$\left\|\breve{\Sigma}^{-1/2}\Sigma\breve{\Sigma}^{-1/2} - \boldsymbol{I}_p\right\| \leq 1/2, \ \text{and} \ \operatorname{tr}\left\{\left(\breve{\Sigma}^{-1/2}\Sigma\breve{\Sigma}^{-1/2} - \boldsymbol{I}_p\right)^2\right\} \leq \delta_\Sigma^2 \tag{5.5}$$

for some $\delta_\Sigma^2 \geq 0$. *The sign* $\|\cdot\|$ *for matrices denotes the spectral norm. Let also for* $z, \overline{z} \geq 2$ *and some* $\delta_z \geq 0$ $|z - \overline{z}| \leq \delta_z$, *then it holds for all* $0 < \Delta \leq 0.22$

1.1. $$\begin{aligned} \left|I\!P\left(\|\boldsymbol{\phi}\| \geq z\right) - I\!P\left(\|\overline{\boldsymbol{\psi}}\| \geq \overline{z}\right)\right| &\leq 16\delta_n\Delta^{-3} + \frac{\Delta+\delta_z}{\overline{z}}\sqrt{p/2} + \delta_\Sigma/2 \\ &\leq 16\delta_n\Delta^{-3} + (\Delta+\delta_z)/\sqrt{2} + \delta_\Sigma/2 \\ &\qquad \text{for } \overline{z} \geq \sqrt{p}, \end{aligned}$$

1.2. $$|I\!P(\|\boldsymbol{\phi}\| \geq z) - I\!P(\|\boldsymbol{\psi}\| \geq \overline{z})| \leq 16\Delta^{-3}\bigl(\delta_n + \breve{\delta}_n\bigr) + \frac{2\Delta + \delta_z}{\overline{z}}\sqrt{p/2} + \delta_\Sigma/2$$
$$\leq 16\Delta^{-3}\bigl(\delta_n + \breve{\delta}_n\bigr) + (2\Delta + \delta_z)/\sqrt{2} + \delta_\Sigma/2$$
for $\overline{z} \geq \sqrt{p}$.

Moreover, if $z, \overline{z} \geq \max\{2, \sqrt{p}\}$ *and* $\max\{\delta_n^{1/4}, \breve{\delta}_n^{1/4}\} \leq 0.11$ *, then*

2.1. $\bigl|I\!P(\|\boldsymbol{\phi}\| \geq z) - I\!P\bigl(\|\overline{\boldsymbol{\psi}}\| \geq \overline{z}\bigr)\bigr| \leq 1.55\delta_n^{1/4} + \delta_z/\sqrt{2} + \delta_\Sigma/2,$

2.2. $|I\!P(\|\boldsymbol{\phi}\| \geq z) - I\!P(\|\boldsymbol{\psi}\| \geq \overline{z})| \leq 1.55\bigl(\delta_n^{1/4} + \breve{\delta}_n^{1/4}\bigr) + \delta_z/\sqrt{2} + \delta_\Sigma/2.$

Proof of Theorem 5.1. The inequality *1.1* is based on the results of Lemmas 5.3, 5.6 and 5.7:

$$I\!P(\|\boldsymbol{\phi}\| \geq z) \overset{\text{by L. 5.3}}{\leq} I\!P\bigl(\|\overline{\boldsymbol{\phi}}\| \geq z - \Delta\bigr) + 16\Delta^{-3}\delta_n$$
$$\overset{\text{by L. 5.7}}{\leq} I\!P\bigl(\|\overline{\boldsymbol{\psi}}\| \geq z - \Delta\bigr) + 16\Delta^{-3}\delta_n + \delta_\Sigma/2$$
$$\overset{\text{by L. 5.6}}{\leq} I\!P\bigl(\|\overline{\boldsymbol{\psi}}\| \geq \overline{z}\bigr) + 16\Delta^{-3}\delta_n + \delta_\Sigma/2 + (\delta_z + \Delta)\overline{z}^{-1}\sqrt{p/2}.$$

The inequality *1.2* is implied by the triangle inequality and the sum of two bounds: the bound *1.1* for $\bigl|I\!P(\|\boldsymbol{\phi}\| \geq z) - I\!P\bigl(\|\overline{\boldsymbol{\psi}}\| \geq \overline{z}\bigr)\bigr|$ and the bound

$$\bigl|I\!P(\|\boldsymbol{\psi}\| \geq \overline{z}) - I\!P\bigl(\|\overline{\boldsymbol{\psi}}\| \geq \overline{z}\bigr)\bigr| \leq 16\breve{\delta}_n\Delta^{-3} + \Delta\overline{z}^{-1}\sqrt{p/2},$$

which also follows from *1.1* by taking $\boldsymbol{\phi} := \boldsymbol{\psi}$, $z := \overline{z}$. In this case $\Sigma = \breve{\Sigma}$ and $\delta_\Sigma = \delta_z = 0$.

The second part of the statement follows from the the first part by balancing the error term $16\delta_n\Delta^{-3} + \Delta/\sqrt{2}$ w.r.t. Δ. □

REMARK 5.1. The approximation error in the statements of Theorem 5.1 includes three terms, each of them is responsible for a step of derivation: Gaussian approximation, Gaussian comparison and anti-concentration. The value δ_Σ bounds the relation between covariance matrices, δ_z corresponds to the difference between quantiles. $\delta_n^{1/4}$ comes from the Gaussian approximation, under certain conditions this is the biggest term in the expressions *2.1, 2.2* (cf. the proof of Theorem 2.1).

REMARK 5.2. Here we briefly comment how our results can be compared with what is available in the literature. In the case of i.i.d. vectors $\boldsymbol{\phi}_i$ and $\operatorname{Var}\boldsymbol{\phi}_i \equiv \boldsymbol{I}_p$ Bentkus (2003) obtained the rate $I\!E\|\boldsymbol{\phi}_i\|^3/\sqrt{n}$ for the error of approximation $\sup_{A\in\mathcal{A}}\bigl|I\!P(\boldsymbol{\phi} \in A) - I\!P(\overline{\boldsymbol{\phi}} \in A)\bigr|$, where $\mathcal{A}$ is a class of all Euclidean balls in $I\!R^p$. Götze (1991) showed for independent vectors $\boldsymbol{\phi}_i$ and their standardized sum $\boldsymbol{\phi}$:

$$\delta_{GAR} \leq \begin{cases} C_1\sqrt{p}\sum_{i=1}^n I\!E\|\boldsymbol{\phi}_i\|^3/\sqrt{n}, & p \in [2,5], \\ C_2 p\sum_{i=1}^n I\!E\|\boldsymbol{\phi}_i\|^3/\sqrt{n}, & p \geq 6, \end{cases}$$

where $\delta_{GAR} \stackrel{\text{def}}{=} \sup_{B\in\mathcal{B}} \bigl|I\!P(\boldsymbol{\phi}\in B) - I\!P(\overline{\boldsymbol{\phi}}\in B)\bigr|$ and $\mathcal{B}$ is a class of all measurable convex sets in $I\!R^p$, the constants $C_1, C_2 > 150$. Bhattacharya and Holmes (2010) argued that the results by Götze (1991) might require more thorough derivation, they obtained the rate $p^{5/2}\sum_{i=1}^n I\!E\|\boldsymbol{\phi}_i\|^3$ for the previous bound (and $p^{5/2}I\!E\|\boldsymbol{\phi}_1\|^3/n^{1/2}$ in the i.i.d. case). Chen and Fang (2011) prove that $\delta_{GAR} \le 115\sqrt{p}\sum_{i=1}^n I\!E\|\boldsymbol{\phi}_i\|^3$ for independent vectors $\boldsymbol{\phi}_i$ with a standardized sum. Götze and Zaitsev (2014) obtained the rate $I\!E\|\boldsymbol{\phi}_i\|^4/n$ for i.i.d. vectors $\boldsymbol{\phi}_i$ with a standardized sum but only for $p \ge 5$. See also Prokhorov and Ulyanov (2013) for the review of the results about normal approximation of quadratic forms.

Our results ensure the error of the Gaussian approximation of order $1.55\delta_n^{1/4} \le 1.31\bigl\{\sum_{i=1}^n I\!E\left(\|\boldsymbol{\phi}_i\|^3 + \|\overline{\boldsymbol{\phi}}_i\|^3\right)\bigr\}^{1/4}$. The technique used here is much simpler than in the previous works, and the obtained bounding terms are explicit and only use independence of the $\boldsymbol{\phi}_i$ and $\boldsymbol{\psi}_i$. However, for some special cases, the use of more advanced results on Gaussian approximation may lead to sharper bounds. For instance, for an i.i.d. sample, the GAR error rate $\delta_{GAR} = \sqrt{p^3/n}$ by Bentkus (2003) is better then ours $(p^3/n)^{1/8}$, and in the one-dimensional case Berry-Esseen's theorem would also work better (see Section 5.1). In those cases one can improve the overall error bound of the bootstrap approximation by putting δ_{GAR} in place of the sum $16\delta_n\Delta^{-3} + \Delta/\sqrt{2}$. Section 5.3 comments how our results can be used to obtain the error rate $\sqrt{p^3/n}$ by using a smoothed quantile function.

5.1 The case of $p = 1$ using Berry-Esseen theorem

Let us consider how the results of Theorem 5.1 can be refined in the case $p = 1$ using Berry-Esseen theorem. Introduce similarly to δ_n and $\breve{\delta}_n$ from (5.4) the bounded values

$$\delta_{n,\text{B.E.}} \stackrel{\text{def}}{=} \sum\nolimits_{i=1}^n I\!E|\boldsymbol{\phi}_i|^3, \qquad \breve{\delta}_{n,\text{B.E.}} \stackrel{\text{def}}{=} \sum\nolimits_{i=1}^n I\!E|\boldsymbol{\psi}_i|^3. \tag{5.6}$$

Due to Berry-Esseen theorem by Berry (1941) and Esseen (1942) it holds

$$\begin{aligned}
\sup_{z\in I\!R}\bigl|I\!P\left(|\boldsymbol{\phi}|\ge z\right) - I\!P\left(|\overline{\boldsymbol{\phi}}|\ge z\right)\bigr| &\le 2C_0\frac{\delta_{n,\text{B.E.}}}{(\operatorname{Var}\boldsymbol{\phi})^{3/2}}, \\
\sup_{z\in I\!R}\bigl|I\!P\left(|\boldsymbol{\psi}|\ge z\right) - I\!P\left(|\overline{\boldsymbol{\psi}}|\ge z\right)\bigr| &\le 2C_0\frac{\breve{\delta}_{n,\text{B.E.}}}{(\operatorname{Var}\boldsymbol{\psi})^{3/2}},
\end{aligned} \tag{5.7}$$

for the constant $C_0 \in [0.4097, 0.560]$ by Esseen (1956) and Shevtsova (2010).

Lemma 5.2. *Under the conditions of Theorem 5.1 it holds*

1. $$\begin{aligned}
\bigl|I\!P\left(|\boldsymbol{\phi}|\ge z\right) - I\!P\left(|\overline{\boldsymbol{\psi}}|\ge \overline{z}\right)\bigr| &\le 2C_0\frac{\delta_{n,B.E.}}{(\operatorname{Var}\boldsymbol{\phi})^{3/2}} + \frac{\delta_\Sigma}{2} + \frac{\delta_z}{\sqrt{2}}\frac{1}{\overline{z}} \\
&\le 2C_0\frac{\delta_{n,B.E.}}{(\operatorname{Var}\boldsymbol{\phi})^{3/2}} + \frac{\delta_\Sigma}{2} + \frac{\delta_z}{\sqrt{2}} \\
&\qquad \textit{for } \overline{z}\ge 1,
\end{aligned}$$

2. $$\begin{aligned}&|I\!P(|\boldsymbol{\phi}| \geq z) - I\!P(|\boldsymbol{\psi}| \geq \overline{z})| \\ &\quad \leq 2C_0\Big\{\frac{\delta_{n,B.E.}}{(\operatorname{Var}\boldsymbol{\phi})^{3/2}} + \frac{\breve{\delta}_{n,B.E.}}{(\operatorname{Var}\boldsymbol{\psi})^{3/2}}\Big\} + \frac{\delta_\Sigma}{2} + \frac{\delta_z}{\sqrt{2}}\frac{1}{\overline{z}} \\ &\quad \leq 2C_0\Big\{\frac{\delta_{n,B.E.}}{(\operatorname{Var}\boldsymbol{\phi})^{3/2}} + \frac{\breve{\delta}_{n,B.E.}}{(\operatorname{Var}\boldsymbol{\psi})^{3/2}}\Big\} + \frac{\delta_\Sigma}{2} + \frac{\delta_z}{\sqrt{2}} \quad \text{for } \overline{z} \geq 1. \end{aligned} \tag{5.8}$$

Proof of Lemma 5.2. Similarly to the proof of Theorem 5.1:

$$\begin{aligned} I\!P(|\boldsymbol{\phi}| \geq z) &\overset{\text{by (5.7)}}{\leq} I\!P\left(|\overline{\boldsymbol{\phi}}| \geq z\right) + 2C_0(\operatorname{Var}\boldsymbol{\phi})^{-3/2}\delta_{n,\text{B.E.}} \\ &\overset{\text{by L. 5.7}}{\leq} I\!P\left(|\overline{\boldsymbol{\psi}}| \geq z\right) + 2C_0(\operatorname{Var}\boldsymbol{\phi})^{-3/2}\delta_{n,\text{B.E.}} + \delta_\Sigma/2 \\ &\overset{\text{by L. 5.6}}{\leq} I\!P\left(|\overline{\boldsymbol{\psi}}| \geq \overline{z}\right) + 2C_0(\operatorname{Var}\boldsymbol{\phi})^{-3/2}\delta_{n,\text{B.E.}} + \delta_\Sigma/2 + \delta_z\overline{z}^{-1}2^{-1/2}. \end{aligned}$$

The analogous chain in the inverse direction finishes the proof of the first part of the statement. The second part is implied by the triangle inequality applied to the first part and again to it with $\boldsymbol{\phi} := \boldsymbol{\psi}$ and $z := \overline{z}$. □

5.2 Gaussian approximation of ℓ_2 norm of a sum of independent vectors

Lemma 5.3 (GAR with equal covariance matrices)**.** *For the random vectors $\boldsymbol{\phi}$ and $\overline{\boldsymbol{\phi}}$ defined in (5.1), (5.3), s.t. $\operatorname{Var}\boldsymbol{\phi} = \operatorname{Var}\overline{\boldsymbol{\phi}}$, and for δ_n given in (5.4), it holds for all $z \geq 2$ and $\Delta \in (0, 0.22]$:*

$$\begin{aligned} I\!P(\|\boldsymbol{\phi}\| \geq z) &\leq I\!P\left(\|\overline{\boldsymbol{\phi}}\| \geq z - \Delta\right) + 16\Delta^{-3}\delta_n, \\ I\!P(\|\boldsymbol{\phi}\| \geq z) &\geq I\!P\left(\|\overline{\boldsymbol{\phi}}\| \geq z + \Delta\right) - 16\Delta^{-3}\delta_n. \end{aligned}$$

Proof of Lemma 5.3. It holds for $z \in I\!R$ $I\!P(\|\boldsymbol{\phi}\| \geq z) = I\!E\,\mathbb{1}\{\|\boldsymbol{\phi}\| \geq z\}$. The main idea of the proof is to approximate the discontinuous function $\mathbb{1}\{\|\boldsymbol{\phi}\| \geq z\}$ by a smooth function $f_\Delta(\boldsymbol{\phi}, z)$ and then to apply the Lindeberg's telescopic sum device. Let us introduce a non-negative function $g(\cdot) \in C^2(I\!R)$, which grows monotonously from 0 to 1:

$$g(x) \overset{\text{def}}{=} \begin{cases} 0, & x \leq 0, \\ 16x^3/3, & x \in [0, 1/4], \\ 0.5 + 2(x - 0.5) - 16(x - 0.5)^3/3, & x \in [1/4, 3/4], \\ 1 + 16(x-1)^3/3, & x \in [3/4, 1], \\ 1, & x \geq 1. \end{cases} \tag{5.9}$$

It holds for all $x \in I\!R$ $\mathbb{1}\{x \geq 1\} \leq g(x) \leq \mathbb{1}\{x \geq 0\}$. Hence, for the function $f_\Delta(\boldsymbol{\phi}, z) \overset{\text{def}}{=} g\left((\|\boldsymbol{\phi}\|^2 - z^2)/(2z\Delta)\right)$ with $z, \Delta > 0$, it holds due to $\mathbb{1}\{\|\boldsymbol{\phi}\| \geq z\} = \mathbb{1}\{(\|\boldsymbol{\phi}\|^2 - z^2)/2 \geq 0\}$:

$$\mathbb{1}\{\|\boldsymbol{\phi}\| \geq z + \Delta\} \leq \mathbb{1}\left\{\|\boldsymbol{\phi}\|^2 \geq z^2 + 2\Delta z\right\} \leq f_\Delta(\boldsymbol{\phi}, z) \leq \mathbb{1}\{\|\boldsymbol{\phi}\| \geq z\}. \tag{5.10}$$

Due to Lemma 5.4 one can apply the Lindeberg's telescopic sum device (see Lindeberg (1922)) in order to approximate $I\!E f_\Delta(\phi, z)$ with $I\!E f_\Delta(\overline{\phi}, z)$. Define for $k = 2, \ldots, n-1$ the following random sums

$$S_k \stackrel{\text{def}}{=} \sum_{i=1}^{k-1} \overline{\phi}_i + \sum_{i=k+1}^{n} \phi_i, \qquad S_1 \stackrel{\text{def}}{=} \sum_{i=2}^{n} \phi_i, \qquad S_n \stackrel{\text{def}}{=} \sum_{i=1}^{n-1} \overline{\phi}_i.$$

The difference $f_\Delta(\phi, z) - f_\Delta(\overline{\phi}, z)$ can be represented as the telescopic sum:

$$f_\Delta(\phi, z) - f_\Delta(\overline{\phi}, z) = \sum\nolimits_{k=1}^{n} \left\{ f_\Delta(S_k + \phi_k, z) - f_\Delta(S_k + \overline{\phi}_k, z) \right\}.$$

Due to Lemma 5.4 and the third order Taylor expansions of $f_\Delta(S_k + \phi_k, z)$ and $f_\Delta(S_k + \overline{\phi}_k, z)$ w.r.t. the first argument at S_k , it holds for each $k = 1, \ldots, n$:

$$\Big| f_\Delta(S_k + \phi_k, z) - f_\Delta(S_k + \overline{\phi}_k, z) - \nabla_\phi f_\Delta(S_k, z)^\top (\phi_k - \overline{\phi}_k) \\ - \frac{1}{2} (\phi_k - \overline{\phi}_k)^\top \nabla^2_\phi f_\Delta(S_k, z) (\phi_k + \overline{\phi}_k) \Big| \leq \mathtt{C}(\Delta, z) \left(\|\phi_k\|^3 + \|\overline{\phi}_k\|^3 \right) / 6,$$

where the value $\mathtt{C}(\Delta, z)$ is defined in (5.14). As S_k and $\phi_k - \overline{\phi}_k$ are independent, $I\!E \phi_k = I\!E \overline{\phi}_k = 0$ and $\operatorname{Var} \phi_k = \operatorname{Var} \overline{\phi}_k$, we derive

$$\begin{aligned} \left| I\!E f_\Delta(\phi, z) - I\!E f_\Delta(\overline{\phi}, z) \right| &= \left| \sum\nolimits_{k=1}^{n} \left\{ I\!E f_\Delta(S_k + \phi_k, z) - I\!E f_\Delta(S_k + \overline{\phi}_k, z) \right\} \right| \\ &\leq \mathtt{C}(\Delta, z) \sum\nolimits_{k=1}^{n} I\!E \left(\|\phi_k\|^3 + \|\overline{\phi}_k\|^3 \right) / 6 \\ \text{(by Def. (5.4))} \quad &= \mathtt{C}(\Delta, z) \delta_n / 3. \end{aligned} \tag{5.11}$$

Combining the derived bounds, we obtain:

$$I\!P \left(\|\phi\| \geq z + \Delta \right) \stackrel{\text{by (5.10)}}{\leq} I\!E f_\Delta(\phi, z) \stackrel{\text{by (5.11)}}{\leq} I\!E f_\Delta(\overline{\phi}, z) + \frac{\mathtt{C}(\Delta, z)}{3} \delta_n \\ \stackrel{\text{by (5.10)}}{\leq} I\!P \left(\|\overline{\phi}\| \geq z \right) + \frac{\mathtt{C}(\Delta, z)}{3} \delta_n,$$

or $I\!P \left(\|\phi\| \geq z \right) \leq I\!P \left(\|\overline{\phi}\| \geq z - \Delta \right) + \mathtt{C}(\Delta, z - \Delta) \delta_n / 3$. Interchanging the arguments ϕ and $\overline{\phi}$ implies the inequality in the inverse direction:

$$I\!P \left(\|\phi\| \geq z \right) \geq I\!P \left(\|\overline{\phi}\| \geq z + \Delta \right) - \mathtt{C}(\Delta, z) \delta_n / 3.$$

Let us bound the constants $\mathtt{C}(\Delta, z)$ and $\mathtt{C}(\Delta, z - \Delta)$ for the function $g(x)$ given above in (5.9). $|g''(x)| \leq 8$ and $|g'''(x)| \leq 32$ for all $x \in I\!R$. By definition (5.14) it holds for $0 < \Delta \leq 0.22$ and $z \geq 2$:

$$\mathtt{C}(\Delta, z) \leq \mathtt{C}(\Delta, z - \Delta) \leq \Delta^{-3} 48. \tag{5.12}$$

□

Lemma 5.4 (A property of the smooth approximant of the indicator)**.** *Let a function* $g(\cdot) \in C^2(I\!R)$ *be non-negative, monotonously increasing from* 0 *to* 1 *s.t.* $g(x) = 0$ *for* $x < 0, \quad g(x) = 1$ *for* $x \geq 1$*. It holds for all* $\boldsymbol{\phi}, \boldsymbol{\phi}_0 \in I\!R^p$*,* $z, \Delta \geq 0$*, for the Euclidean norm* $\|\cdot\|$ *and for the function*

$$f_\Delta(\boldsymbol{\phi}, z) \stackrel{\text{def}}{=} g\left(\frac{1}{2z\Delta}(\|\boldsymbol{\phi}\|^2 - z^2)\right) \tag{5.13}$$

$$\begin{aligned}\left|f_\Delta(\boldsymbol{\phi}_0 + \boldsymbol{\phi}, z) - f_\Delta(\boldsymbol{\phi}_0, z) - \boldsymbol{\phi}^\top \nabla_{\boldsymbol{\phi}} f_\Delta(\boldsymbol{\phi}_0, z) - \boldsymbol{\phi}^\top \nabla^2_{\boldsymbol{\phi}} f_\Delta(\boldsymbol{\phi}_0, z)\boldsymbol{\phi}/2\right| \\ \leq \mathtt{C}(\Delta, z)\|\boldsymbol{\phi}\|^3/3!,\end{aligned}$$

where

$$\mathtt{C}(\Delta, z) \stackrel{\text{def}}{=} \frac{1}{\Delta^3}\left(1 + 2\frac{\Delta}{z}\right)^{1/2}\left\{\left(1 + 2\frac{\Delta}{z}\right)\|g'''\|_\infty + 3\frac{\Delta}{z}\|g''\|_\infty\right\}. \tag{5.14}$$

Proof of Lemma 5.4. By the Taylor formula:

$$f_\Delta(\boldsymbol{\phi}_0 + \boldsymbol{\phi}, z) = f_\Delta(\boldsymbol{\phi}_0, z) + \boldsymbol{\phi}^\top \nabla_{\boldsymbol{\phi}} f_\Delta(\boldsymbol{\phi}_0, z) + \boldsymbol{\phi}^\top \nabla^2_{\boldsymbol{\phi}} f_\Delta(\boldsymbol{\phi}_0, z)\boldsymbol{\phi}/2 + R_3,$$

where R_3 is the 3-d order remainder term. Consider for $\boldsymbol{\gamma} \in I\!R^p : \|\boldsymbol{\gamma}\| = 1$ and $t \in R$ the function $f_\Delta(\boldsymbol{\phi}_0 + t\boldsymbol{\gamma}, z) = g\left(\frac{1}{2z\Delta}(\|\boldsymbol{\phi}_0 + t\boldsymbol{\gamma}\|^2 - z^2)\right)$. It holds

$$|R_3| \leq \frac{\|\boldsymbol{\phi}\|^3}{3!}\sup_{\boldsymbol{\gamma} \in I\!R^p,\, \|\boldsymbol{\gamma}\|=1}\sup_{t \in I\!R}\left|\frac{d^3 f_\Delta(\boldsymbol{\phi}_0 + t\boldsymbol{\gamma}, z)}{dt^3}\right|.$$

Now let us bound the third derivative $\frac{d^3}{dt^3} f_\Delta(\boldsymbol{\phi} + t\boldsymbol{\gamma}, z)$:

$$\begin{aligned}\frac{df_\Delta(\boldsymbol{\phi} + t\boldsymbol{\gamma}, z)}{dt} &= \frac{\boldsymbol{\gamma}^\top(\boldsymbol{\phi} + t\boldsymbol{\gamma})}{z\Delta} g'\left(\frac{1}{2z\Delta}(\|\boldsymbol{\phi} + t\boldsymbol{\gamma}\|^2 - z^2)\right), \\ \frac{d^2 f_\Delta(\boldsymbol{\phi} + t\boldsymbol{\gamma}, z)}{dt^2} &= \frac{\{\boldsymbol{\gamma}^\top(\boldsymbol{\phi} + t\boldsymbol{\gamma})\}^2}{(z\Delta)^2} g''\left(\frac{1}{2z\Delta}(\|\boldsymbol{\phi} + t\boldsymbol{\gamma}\|^2 - z^2)\right) \\ &\quad + \frac{1}{z\Delta} g'\left(\frac{1}{2z\Delta}(\|\boldsymbol{\phi} + t\boldsymbol{\gamma}\|^2 - z^2)\right), \\ \frac{d^3 f_\Delta(\boldsymbol{\phi} + t\boldsymbol{\gamma}, z)}{dt^3} &= \frac{\{\boldsymbol{\gamma}^\top(\boldsymbol{\phi} + t\boldsymbol{\gamma})\}^3}{(z\Delta)^3} g'''\left(\frac{1}{2z\Delta}(\|\boldsymbol{\phi} + t\boldsymbol{\gamma}\|^2 - z^2)\right) \\ &\quad + 3\frac{\boldsymbol{\gamma}^\top(\boldsymbol{\phi} + t\boldsymbol{\gamma})}{(z\Delta)^2} g''\left(\frac{1}{2z\Delta}(\|\boldsymbol{\phi} + t\boldsymbol{\gamma}\|^2 - z^2)\right).\end{aligned}$$

Now we use that $g''(x)$ and $g'''(x)$ vanish if $x < 0$ or $x \geq 1$. The inequality $\frac{1}{2z\Delta}(\|\boldsymbol{\phi} + t\boldsymbol{\gamma}\|^2 - z^2) \leq 1$ implies in view of $\|\boldsymbol{\gamma}\| = 1$ that

$$\boldsymbol{\gamma}^\top(\boldsymbol{\phi} + t\boldsymbol{\gamma}) \leq \|\boldsymbol{\phi} + t\boldsymbol{\gamma}\| \leq (2z\Delta + z^2)^{1/2}.$$

Therefore

$$\left| \frac{d^3 f_\Delta(\boldsymbol{\phi}_0 + t\boldsymbol{\gamma}, z)}{dt^3} \right| \le \frac{1}{\Delta^3} \left(1 + 2\frac{\Delta}{z}\right)^{1/2} \left\{ \left(1 + 2\frac{\Delta}{z}\right) \|g'''\|_\infty + 3\frac{\Delta}{z} \|g''\|_\infty \right\}.$$

□

5.3 Results for the smoothed indicator function

Theorem 5.5 (Theorem 5.1 for a smoothed indicator function)**.** *Under the conditions of Theorem 5.1 it holds for all* $\delta_z \in [0,1]$ *and the function* $f_\Delta(\boldsymbol{\phi}, z)$ *defined in* (5.13)*:*

1. $$\begin{aligned} \left| I\!\!E f_\Delta(\boldsymbol{\phi}, z) - I\!\!E f_\Delta(\overline{\boldsymbol{\psi}}, \overline{z}) \right| &\le \frac{16}{\Delta^3}\delta_n + 2\sqrt{p}\frac{\delta_z}{z} + \sqrt{p}\frac{\delta_z^2}{z^2} + \delta_\Sigma \\ &\le \frac{16}{\Delta^3}\delta_n + \sqrt{5}\delta_z + \delta_\Sigma \quad \textit{for } z \ge \sqrt{p}. \end{aligned}$$
2. $$\begin{aligned} |I\!\!E f_\Delta(\boldsymbol{\phi}, z) - I\!\!E f_\Delta(\boldsymbol{\psi}, \overline{z})| &\le \frac{16}{\Delta^3}\left(\delta_n + \breve{\delta}_n\right) + 2\sqrt{p}\frac{\delta_z}{z} + \sqrt{p}\frac{\delta_z^2}{z^2} + \delta_\Sigma \\ &\le \frac{16}{\Delta^3}\left(\delta_n + \breve{\delta}_n\right) + \sqrt{5}\delta_z + \delta_\Sigma \quad \textit{for } z \ge \sqrt{p}. \end{aligned}$$

REMARK 5.3. The approximating bounds above do not contain the term proportional to Δ unlike the bound in Theorem 5.1. This yields the smaller error terms for the case of the smoothed indicator.

Proof of Theorem 5.5. The following inequality is proved in Lemma 5.3 (see the expression (5.11)): $\left| I\!\!E f_\Delta(\boldsymbol{\phi}, z) - I\!\!E f_\Delta(\overline{\boldsymbol{\phi}}, z) \right| \le \mathtt{C}(\Delta, z)\delta_n/3$.

The function $f_\Delta(\boldsymbol{\phi}, z)$ is non-increasing in z:

$$\frac{df_\Delta(\boldsymbol{\phi}, z)}{dz} = -\frac{1}{2\Delta}\left(1 + \frac{\|\boldsymbol{\phi}\|^2}{z^2}\right) g'\left(\frac{1}{2\Delta z}\left(\|\boldsymbol{\phi}\|^2 - z^2\right)\right) \le 0.$$

The definition of $f_\Delta(\boldsymbol{\phi}, z)$ yields for $\overline{z} \ge z$, $a \stackrel{\mathrm{def}}{=} \overline{z}/z \ge 1$ and any $\boldsymbol{\phi}$

$$\begin{aligned} f_\Delta(\boldsymbol{\phi}, \overline{z}) \le f_\Delta(\boldsymbol{\phi}, z) \le f_\Delta(a\boldsymbol{\phi}, \overline{z}), \\ 0 \le f_\Delta(\boldsymbol{\phi}, z) - f_\Delta(\boldsymbol{\phi}, \overline{z}) \le f_\Delta(a\boldsymbol{\phi}, \overline{z}) - f_\Delta(\boldsymbol{\phi}, \overline{z}). \end{aligned} \tag{5.15}$$

Lemma 5.8 yields for $\delta_z \le z(\sqrt{3/2} - 1)$:

$$\begin{aligned} \left| I\!\!E f_\Delta(a\overline{\boldsymbol{\phi}}, \overline{z}) - I\!\!E f_\Delta(\overline{\boldsymbol{\phi}}, \overline{z}) \right| &\le \sqrt{p}\left(\frac{\overline{z}^2}{z^2} - 1\right) \le 2\sqrt{p}\frac{\delta_z}{z} + \sqrt{p}\frac{\delta_z^2}{z^2} \\ &\le (1 + \sqrt{3/2})\delta_z \le \sqrt{5}\delta_z \quad \text{for } z \ge \sqrt{p}. \end{aligned}$$

Inequalities similar to (5.15) hold for $\overline{z} \leq z$ and $a \stackrel{\text{def}}{=} z/\overline{z}$, therefore, by triangle inequality, bound (5.12) on $\mathtt{C}(\Delta, z)$ and Lemma 5.8:

$$\begin{aligned}\left|I\!E f_\Delta(\boldsymbol{\phi}, z) - I\!E f_\Delta(\overline{\boldsymbol{\psi}}, \overline{z})\right| &\leq \frac{16}{\Delta^3}\delta_n + 2\sqrt{p}\frac{\delta_z}{z} + \sqrt{p}\frac{\delta_z^2}{z^2} + \delta_\Sigma \\ &\leq \frac{16}{\Delta^3}\delta_n + \sqrt{5}\delta_z + \delta_\Sigma \quad \text{for } z \geq \sqrt{p}.\end{aligned}$$

The second part of the statement follows from triangle inequality applied to the first inequality and again to the same one with $\boldsymbol{\phi} := \boldsymbol{\psi}$ and $z := \overline{z}$. □

5.4 Gaussian anti-concentration and comparison by Pinsker's inequality

Lemma 5.6 (Anti-concentration bound for ℓ_2 norm of a Gaussian vector)**.** *Let $\overline{\boldsymbol{\phi}} \sim \mathcal{N}(0, \Sigma)$, $\overline{\boldsymbol{\phi}} \in I\!R^p$, then it holds for all $z > 0$ and $0 \leq \Delta \leq z$:*

$$\begin{aligned}\left|I\!P\left(\|\overline{\boldsymbol{\phi}}\| \geq z + \Delta\right) - I\!P\left(\|\overline{\boldsymbol{\phi}}\| \geq z\right)\right| &\leq \Delta\sqrt{p}/(z\sqrt{2}) \\ &\leq \Delta/\sqrt{2} \ \textit{ for } z \geq \sqrt{p}.\end{aligned}$$

Proof of Lemma 5.6. It holds $I\!P\left(\|\overline{\boldsymbol{\phi}}\| \geq z + \Delta\right) = I\!P\left(\|\overline{\boldsymbol{\phi}}_\Delta\| \geq z\right)$, where $\overline{\boldsymbol{\phi}}_\Delta \stackrel{\text{def}}{=} \overline{\boldsymbol{\phi}}_{\frac{z}{z+\Delta}}$. The Kullback-Leibler divergence between $I\!P_1 \stackrel{\text{def}}{=} \mathcal{N}(0, \Sigma)$ and $I\!P_2 \stackrel{\text{def}}{=} \mathcal{N}\bigl(0, \Sigma\frac{z^2}{(z+\Delta)^2}\bigr)$ is equal to

$$\begin{aligned}\mathrm{KL}(I\!P_1, I\!P_2) &= p\left\{(\Delta/z)^2 + 2(\Delta/z) - 2\log(1 + \Delta/z)\right\}/2 \\ &\leq p(\Delta/z)^2 \ \text{ for } 0 \leq \Delta \leq z.\end{aligned}$$

We use Pinsker's inequality in the following form (see the book by Tsybakov (2009), pp. 88, 132): for a measurable space $(\Omega, \mathcal{F})$ and two measures on it $I\!P_1, I\!P_2$:

$$\sup_{A \in \mathcal{F}} |I\!P_1(A) - I\!P_2(A)| \leq \sqrt{\mathrm{KL}(I\!P_1, I\!P_2)/2}. \tag{5.16}$$

Therefore, it holds:

$$\left|I\!P\left(\|\overline{\boldsymbol{\phi}}\| \geq z + \Delta\right) - I\!P\left(\|\overline{\boldsymbol{\phi}}\| \geq z\right)\right| \leq \sqrt{\mathrm{KL}(I\!P_1, I\!P_2)/2} \leq \Delta\sqrt{p}/(z\sqrt{2}).$$

□

Lemma 5.7 (Comparison of the Euclidian norms of Gaussian vectors)**.** *Let $\overline{\boldsymbol{\psi}}_1 \sim \mathcal{N}(0, \Sigma_1)$ and $\overline{\boldsymbol{\psi}}_2 \sim \mathcal{N}(0, \Sigma_2)$ belong to $I\!R^p$, and*

$$\left\|\Sigma_2^{-1/2}\Sigma_1\Sigma_2^{-1/2} - \boldsymbol{I}_p\right\| \leq 1/2, \ \textit{ and } \ \mathrm{tr}\left\{\left(\Sigma_2^{-1/2}\Sigma_1\Sigma_2^{-1/2} - \boldsymbol{I}_p\right)^2\right\} \leq \delta_\Sigma^2,$$

for some $\delta_\Sigma^2 \geq 0$. Then it holds

$$\sup_{z \in R}\left|I\!P\left(\|\overline{\boldsymbol{\psi}}_1\| \geq z\right) - I\!P\left(\|\overline{\boldsymbol{\psi}}_2\| \geq z\right)\right| \leq \delta_\Sigma/2.$$

Proof of Lemma 5.7. Let $I\!P_1 = \mathcal{N}(0, \Sigma_1)$ and $I\!P_2 = \mathcal{N}(0, \Sigma_2)$. Denote $G \stackrel{\text{def}}{=} \Sigma_2^{-1/2} \Sigma_1 \Sigma_2^{-1/2}$, then the Kullback-Leibler divergence between $I\!P_1$ and $I\!P_2$ is equal to

$$\begin{aligned} \mathrm{KL}(I\!P_1, I\!P_2) &= -0.5 \log\{\det(G)\} + 0.5 \operatorname{tr}\{G - \boldsymbol{I}_p\} \\ &= 0.5 \sum\nolimits_{j=1}^{p} \{\lambda_j - \log(\lambda_j + 1)\}, \end{aligned}$$

where $\lambda_p \leq \cdots \leq \lambda_1$ are the eigenvalues the matrix $G - \boldsymbol{I}_p$. By conditions of the lemma $|\lambda_1| \leq 1/2$, and it holds:

$$\mathrm{KL}(I\!P_1, I\!P_2) \leq 0.5 \sum\nolimits_{j=1}^{p} \lambda_j^2 = 0.5 \operatorname{tr}\{(G - \boldsymbol{I}_p)^2\} \leq \delta_\Sigma^2 / 2, \tag{5.17}$$

which finishes the proof due to the Pinsker's inequality (5.16). □

Lemma 5.8 (Gaussian comparison, smoothed version)**.** *Let* $\overline{\boldsymbol{\psi}}_1 \sim \mathcal{N}(0, \Sigma_1)$ *and* $\overline{\boldsymbol{\psi}}_2 \sim \mathcal{N}(0, \Sigma_2)$ *belong to* $I\!R^p$*, and for some* $\delta_\Sigma^2 \geq 0$*:*

$$\left\| \Sigma_2^{-1/2} \Sigma_1 \Sigma_2^{-1/2} - \boldsymbol{I}_p \right\| \leq 1/2, \quad \textit{and} \quad \operatorname{tr}\left\{ \left(\Sigma_2^{-1/2} \Sigma_1 \Sigma_2^{-1/2} - \boldsymbol{I}_p \right)^2 \right\} \leq \delta_\Sigma^2.$$

Then it holds for any function $f(\boldsymbol{x}) : I\!R^p \mapsto I\!R$ *s.t.* $|f(\boldsymbol{x})| \leq 1$*:*

$$\left| I\!E f(\overline{\boldsymbol{\psi}}_1) - I\!E f(\overline{\boldsymbol{\psi}}_2) \right| \leq \delta_\Sigma.$$

Proof of Lemma 5.8. Let $I\!P_1 = \mathcal{N}(0, \Sigma_1)$ and $I\!P_2 = \mathcal{N}(0, \Sigma_2)$. Due to $|f(\boldsymbol{x})| \leq 1$ and Pinsker's inequality (5.16) it holds:

$$\begin{aligned} \left| I\!E f(\overline{\boldsymbol{\psi}}_1) - I\!E f(\overline{\boldsymbol{\psi}}_2) \right| &\leq \int_{I\!R^p} |f(\boldsymbol{x})| \cdot |dI\!P_1(\boldsymbol{x}) - dI\!P_2(\boldsymbol{x})| \\ &\leq \int_{I\!R^p} |dI\!P_1(\boldsymbol{x}) - dI\!P_2(\boldsymbol{x})| \leq 2 \sqrt{\mathrm{KL}(I\!P_1, I\!P_2)/2}. \end{aligned}$$

Finally, as in (5.17), $2\sqrt{\mathrm{KL}(I\!P_1, I\!P_2)/2} \leq \delta_\Sigma$. □

A Appendix

This section contains proofs of the main results from Section 2. Due to the scheme (1.6) the key ingredients are:

- the square-root Wilks approximation for the $\boldsymbol{Y}$-world (Theorem A.2),
- the square-root Wilks approximation for the bootstrap world (Theorem A.4),
- the statement about closeness in distribution of the approximating terms $\|\boldsymbol{\xi}\|$ and $\|\boldsymbol{\xi}^\circ\|$ (Proposition A.9).

In Section A.1 we recall some results from the general finite sample theory by Spokoiny (2012a), Spokoiny (2012b) and Spokoiny (2013), including the square-root Wilks approximation in $\boldsymbol{Y}$ case. In Section A.2 we derive the necessary results from the finite sample theory for the bootstrap world (including the square-root Wilks approximation). In Section A.3 we adapt Theorem 5.1 (GAR for ℓ_2 norm of a sum of independent vectors) to the setting of maximum likelihood estimation (Proposition A.9). The proofs of the main results are given in Sections A.3, A.4 and A.5.

A.1 Finite sample theory

Let us use the notations given in the introduction: $L(\boldsymbol{\theta})$ is the log-likelihood process, which depends on the data $\boldsymbol{Y}$ and corresponds to the regular parametric family of probability distributions $\{I\!P_{\boldsymbol{\theta}}\}$. The general finite sample approach by Spokoiny (2012a) does not require the true measure $I\!P$ to belong to $\{I\!P_{\boldsymbol{\theta}}\}$. The target parameter $\boldsymbol{\theta}^*$ is defined as in (1.4) by projection of the true measure $I\!P$ on $\{I\!P_{\boldsymbol{\theta}}\}$. D_0^2 denotes the full Fisher information $p \times p$ matrix, which is deterministic, symmetric and positive-definite:

$$D_0^2 \stackrel{\text{def}}{=} -\nabla^2_{\boldsymbol{\theta}} I\!E L(\boldsymbol{\theta}^*).$$

A centered p-dimensional random vector $\boldsymbol{\xi}$ denotes the normalised score:

$$\boldsymbol{\xi} \stackrel{\text{def}}{=} D_0^{-1} \nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta}^*).$$

Introduce the following elliptic vicinity around the true point $\boldsymbol{\theta}^*$:

$$\Theta_0(\mathtt{r}) \stackrel{\text{def}}{=} \{\boldsymbol{\theta} : \|D_0(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\| \le \mathtt{r}\}.$$

The non-asymptotic Wilks approximating bound by Spokoiny (2012a), Spokoiny (2013) requires that the maximum likelihood estimate $\widetilde{\boldsymbol{\theta}}$ gets into the local vicinity $\Theta_0(\mathtt{r}_0)$ of some radius $\mathtt{r}_0 > 0$ with probability $\ge 1 - 3\mathrm{e}^{-\mathtt{x}}$, $\mathtt{x} > 0$. This is guaranteed by the following concentration result:

Theorem A.1 (Concentration of MLE, Spokoiny (2013)). *Let the conditions* (ED_0), (ED_2), $(\mathcal{L}_0)$, $(\mathcal{I})$ *and* $(\mathcal{L}\mathtt{r})$ *be fulfilled. If for the constant* $\mathtt{r}_0 > 0$ *and for the function* $\mathtt{b}(\mathtt{r})$ *from* $(\mathcal{L}\mathtt{r})$ *:*

$$\mathtt{b}(\mathtt{r})\mathtt{r} \ge 2\left\{\mathfrak{Z}_{\mathtt{qf}}(\mathtt{x}, I\!B) + 6\omega\nu_0\mathfrak{Z}(\mathtt{x} + \log(2\mathtt{r}/\mathtt{r}_0))\right\}, \quad \mathtt{r} > \mathtt{r}_0 \tag{A.1}$$

where the functions $\mathfrak{Z}(\mathtt{x})$ *and* $\mathfrak{Z}_{\mathtt{qf}}(\mathtt{x}, I\!B)$ *are defined respectively in* (A.3) *and* (A.4)*, then it holds*

$$I\!P\left(\widetilde{\boldsymbol{\theta}} \notin \Theta_0(\mathtt{r}_0)\right) \le 3\mathrm{e}^{-\mathtt{x}}.$$

The constants ω, ν_0 *and* $\mathfrak{a}$ *come from the imposed conditions* (ED_0) – $(\mathcal{I})$ *(from Section 4). In the case 4.3* $\mathtt{r}_0 \ge \mathtt{C}\sqrt{p + \mathtt{x}}$.

The following result is one of the central in the general finite sample theory and is crucial for the present study due to the scheme (1.6):

Theorem A.2 (Wilks approximation, Spokoiny (2013)). *Under the conditions of Theorem A.1 for some* $\mathtt{r}_0 > 0$ *s.t.* (A.1) *is fulfilled, it holds with probability* $\geq 1 - 5\mathrm{e}^{-\mathtt{x}}$

$$\left| 2\Big\{ L(\widetilde{\boldsymbol{\theta}}) - L(\boldsymbol{\theta}^*) \Big\} - \|\boldsymbol{\xi}\|^2 \right| \leq \Delta_{W^2}(\mathtt{r}_0, \mathtt{x}),$$

$$\left| \sqrt{2\Big\{ L(\widetilde{\boldsymbol{\theta}}) - L(\boldsymbol{\theta}^*) \Big\}} - \|\boldsymbol{\xi}\| \right| \leq \Delta_{W}(\mathtt{r}_0, \mathtt{x})$$

for

$$\Delta_{W}(\mathtt{r}, \mathtt{x}) \stackrel{\text{def}}{=} 3\mathtt{r} \left\{ \delta(\mathtt{r}) + 6\nu_0 \mathfrak{Z}(\mathtt{x})\omega \right\}, \tag{A.2}$$

$$\Delta_{W^2}(\mathtt{r}, \mathtt{x}) \stackrel{\text{def}}{=} \frac{2}{3} \left\{ 2\mathtt{r} + \mathfrak{Z}_{\mathtt{qf}}(\mathtt{x}, I\!B) \right\} \Delta_{W}(\mathtt{r}, \mathtt{x}),$$

$$\mathfrak{Z}(\mathtt{x}) \stackrel{\text{def}}{=} 2\sqrt{p} + \sqrt{2\mathtt{x}} + 4p(\mathtt{x}\mathtt{g}^{-2} + 1)\mathtt{g}^{-1}. \tag{A.3}$$

In the case 4.3 it holds for $\mathtt{r} \leq \mathtt{r}_0$ *:*

$$\Delta_{W}(\mathtt{r}, \mathtt{x}) = \mathtt{C} \frac{p + \mathtt{x}}{\sqrt{n}}, \quad \Delta_{W^2}(\mathtt{r}, \mathtt{x}) = \mathtt{C} \sqrt{\frac{(p + \mathtt{x})^3}{n}}.$$

The constants $\mathtt{g}$ *and* $\delta(\mathtt{r})$ *come from the imposed conditions* $(\boldsymbol{ED}_0)$, $(\boldsymbol{\mathcal{L}}_0)$ *(from Section 4), and the function* $\mathfrak{Z}_{\mathtt{qf}}(\mathtt{x}, I\!B)$ *, defined in* (A.4)*, corresponds to the quantile function of deviations of the random value* $\|\boldsymbol{\xi}\|$ *(see Theorem A.3 below).*

The following theorem characterizes the tail behaviour of the approximating term $\|\boldsymbol{\xi}\|^2$. It means that with a bounded exponential moment of the vector $\boldsymbol{\xi}$ (condition $(\boldsymbol{ED}_0)$) its squared Euclidean norm $\|\boldsymbol{\xi}\|^2$ has three regimes of deviations: sub-Gaussian, Poissonian and large-deviations' zone.

Theorem A.3 (Deviation bound for a random quadratic form, Spokoiny (2012b)). *Let condition* $(\boldsymbol{ED}_0)$ *be fulfilled, then for* $\mathtt{g} \geq \sqrt{2\,\mathrm{tr}(I\!B^2)}$ *it holds:*

$$I\!P\left(\|\boldsymbol{\xi}\|^2 \geq \mathfrak{Z}^2_{\mathtt{qf}}(\mathtt{x}, I\!B) \right) \leq 2\mathrm{e}^{-\mathtt{x}} + 8.4\mathrm{e}^{-\mathtt{x}_c},$$

where $I\!B^2 \stackrel{\text{def}}{=} D_0^{-1} V_0^2 D_0^{-1}$ *,* $\lambda(I\!B)$ *is a maximum eigenvalue of* $I\!B^2$ *,*

$$\mathfrak{Z}^2_{\mathtt{qf}}(\mathtt{x}, I\!B) \stackrel{\text{def}}{=} \begin{cases} \mathrm{tr}(I\!B^2) + \sqrt{8\,\mathrm{tr}(I\!B^4)\mathtt{x}}, & \mathtt{x} \leq \sqrt{2\,\mathrm{tr}(I\!B^4)}/\{18\lambda(I\!B)\}, \\ \mathrm{tr}(I\!B^2) + 6\mathtt{x}\lambda(I\!B), & \sqrt{2\,\mathrm{tr}(I\!B^4)}/\{18\lambda(I\!B)\} < \mathtt{x} \leq \mathtt{x}_c, \\ |\mathtt{z}_c + 2(\mathtt{x} - \mathtt{x}_c)/\mathtt{g}_c|^2 \, \lambda(I\!B), & \mathtt{x} > \mathtt{x}_c, \end{cases} \tag{A.4}$$

$$
\begin{aligned}
2\mathtt{x}_c &\stackrel{\text{def}}{=} 2\mathtt{x}_c(I\!B) \stackrel{\text{def}}{=} \mu_c \mathtt{z}_c{}^2 + \log\det\left(\boldsymbol{I}_p - \mu_c I\!B^2/\lambda(I\!B)\right), \qquad \text{(A.5)}\\
\mathtt{z}_c{}^2 &\stackrel{\text{def}}{=} \left\{\mathtt{g}^2/\mu_c^2 - \operatorname{tr}\left(I\!B^2\right)/\mu_c\right\}/\lambda(I\!B),\\
\mathtt{g}_c &\stackrel{\text{def}}{=} \sqrt{\mathtt{g}^2 - \mu_c \operatorname{tr}\left(I\!B^2\right)}/\sqrt{\lambda(I\!B)},\\
\mu_c &\stackrel{\text{def}}{=} 2/3.
\end{aligned}
$$

The matrix V_0^2 *comes from condition* $(\boldsymbol{ED_0})$ *and can be defined as*

$$
V_0^2 \stackrel{\text{def}}{=} \operatorname{Var}\left\{\nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta}^*)\right\}.
$$

By condition $(\boldsymbol{\mathcal{I}})$ $\operatorname{tr}(I\!B^2) \le \mathfrak{a}^2 p$, $\operatorname{tr}(I\!B^4) \le \mathfrak{a}^4 p$ *and* $\lambda(I\!B) \le \mathfrak{a}^2$. *In the case 4.3* $\mathtt{g} = \mathtt{C}\sqrt{n}$, *hence* $\mathtt{x}_c = \mathtt{C}n$, *and for* $\mathtt{x} \le \mathtt{x}_c$ *it holds:*

$$
\mathfrak{Z}^2_{\mathtt{qf}}(\mathtt{x}, I\!B) \le \mathfrak{a}^2(p + 6\mathtt{x}). \qquad \text{(A.6)}
$$

A.2 Finite sample theory for the bootstrap world

Let us introduce the bootstrap score vector at a point $\boldsymbol{\theta} \in \Theta$:

$$
\begin{aligned}
\boldsymbol{\xi}^\circ(\boldsymbol{\theta}) &\stackrel{\text{def}}{=} D_0^{-1} \nabla_{\boldsymbol{\theta}} \boldsymbol{\zeta}^\circ(\boldsymbol{\theta})\\
&= \sum_{i=1}^n D_0^{-1} \nabla_{\boldsymbol{\theta}} \ell_i(\boldsymbol{\theta})(u_i - 1).
\end{aligned}
$$

Theorem A.4 (Bootstrap Wilks approximation)**.** *Under the conditions of Theorems A.1 and A.6 for some* $\mathtt{r}_0{}^2 \ge 0$ *s.t.* (A.1) *and* (A.17) *are fulfilled, it holds with* $I\!P$*-probability* $\ge 1 - 5\mathrm{e}^{-\mathtt{x}}$

$$
\begin{aligned}
I\!P^\circ\left(\left|\sup_{\boldsymbol{\theta}\in\Theta} 2\left\{L^\circ(\boldsymbol{\theta}) - L^\circ(\widetilde{\boldsymbol{\theta}})\right\} - \|\boldsymbol{\xi}^\circ(\widetilde{\boldsymbol{\theta}})\|^2\right| \le \Delta^\circ_{W^2}(\mathtt{r}_0, \mathtt{x})\right) &\ge 1 - 4\mathrm{e}^{-\mathtt{x}},\\
I\!P^\circ\left(\left|\sqrt{\sup_{\boldsymbol{\theta}\in\Theta} 2\left\{L^\circ(\boldsymbol{\theta}) - L^\circ(\widetilde{\boldsymbol{\theta}})\right\}} - \|\boldsymbol{\xi}^\circ(\widetilde{\boldsymbol{\theta}})\|\right| \le \Delta^\circ_{W}(\mathtt{r}_0, \mathtt{x})\right) &\ge 1 - 4\mathrm{e}^{-\mathtt{x}}.
\end{aligned}
$$

where the error terms $\Delta^\circ_W(\mathtt{r}, \mathtt{x}), \Delta^\circ_{W^2}(\mathtt{r}, \mathtt{x})$ *are deterministic and*

$$
\begin{aligned}
\Delta^\circ_W(\mathtt{r}, \mathtt{x}) &\stackrel{\text{def}}{=} 2\Delta_W(\mathtt{r}, \mathtt{x}) + 36\nu_0 \mathtt{r}\omega_1(\mathtt{r})\mathfrak{Z}(\mathtt{x}),\\
\Delta^\circ_{W^2}(\mathtt{r}, \mathtt{x}) &\stackrel{\text{def}}{=} \frac{1}{18}\left\{12\mathtt{r}\Delta^\circ_W(\mathtt{r}, \mathtt{x}) + \Delta^\circ_W(\mathtt{r}, \mathtt{x})^2\right\}.
\end{aligned}
$$

$\Delta_W(\mathtt{r}, \mathtt{x})$ *and* $\mathfrak{Z}(\mathtt{x})$ *are defined respectively in* (A.2) *and* (A.3), *and* $\omega_1(\mathtt{r})$ *is given in* (A.12). *For the case 4.3 and* $\mathtt{r} \le \mathtt{r}_0$ *it holds:*

$$
\Delta^\circ_W(\mathtt{r}, \mathtt{x}) = \mathtt{C}\,\frac{p + \mathtt{x}}{\sqrt{n}}\sqrt{\mathtt{x}}, \quad \Delta^\circ_{W^2}(\mathtt{r}, \mathtt{x}) = \mathtt{C}\sqrt{\frac{(p + \mathtt{x})^3}{n}}\sqrt{\mathtt{x}}.
$$

Proof of Theorem A.4. Let us consider the following random process in the bootstrap world for $\boldsymbol{\theta}, \boldsymbol{\theta}_1 \in \Theta_0(\mathtt{r})$:

$$\mathcal{A}^\circ(\boldsymbol{\theta}, \boldsymbol{\theta}_1) \stackrel{\text{def}}{=} L^\circ(\boldsymbol{\theta}) - L^\circ(\boldsymbol{\theta}_1) - (\boldsymbol{\theta} - \boldsymbol{\theta}_1)^\top \nabla_{\boldsymbol{\theta}} L^\circ(\boldsymbol{\theta}_1) + \frac{1}{2} \|D_0(\boldsymbol{\theta} - \boldsymbol{\theta}_1)\|^2.$$

It holds $\mathcal{A}^\circ(\boldsymbol{\theta}_1, \boldsymbol{\theta}_1) = 0$. Taylor expansion w.r.t. $\boldsymbol{\theta}$ around $\boldsymbol{\theta}_1$ implies :

$$\mathcal{A}^\circ(\boldsymbol{\theta}, \boldsymbol{\theta}_1) = (\boldsymbol{\theta} - \boldsymbol{\theta}_1)^\top \nabla_{\boldsymbol{\theta}} \mathcal{A}^\circ(\overline{\boldsymbol{\theta}}_1, \boldsymbol{\theta}_1),$$

where $\overline{\boldsymbol{\theta}}_1$ is some convex combination of the vectors $\boldsymbol{\theta}$ and $\boldsymbol{\theta}_1$. Therefore,

$$|\mathcal{A}^\circ(\boldsymbol{\theta}, \boldsymbol{\theta}_1)| \leq \|D_0(\boldsymbol{\theta} - \boldsymbol{\theta}_1)\| \sup_{\boldsymbol{\theta} \in \Theta_0(\mathtt{r})} \left\| D_0^{-1} \nabla_{\boldsymbol{\theta}} \mathcal{A}^\circ(\boldsymbol{\theta}, \boldsymbol{\theta}_1) \right\| \tag{A.7}$$

$$\leq 2\mathtt{r} \sup_{\boldsymbol{\theta} \in \Theta_0(\mathtt{r})} \left\| D_0^{-1} \nabla_{\boldsymbol{\theta}} \mathcal{A}^\circ(\boldsymbol{\theta}, \boldsymbol{\theta}_1) \right\|. \tag{A.8}$$

Now let us consider the normalized gradient process:

$$D_0^{-1} \nabla_{\boldsymbol{\theta}} \mathcal{A}^\circ(\boldsymbol{\theta}, \boldsymbol{\theta}_1) = D_0^{-1} \left\{ \nabla_{\boldsymbol{\theta}} L^\circ(\boldsymbol{\theta}) - \nabla_{\boldsymbol{\theta}} L^\circ(\boldsymbol{\theta}_1) \right\} + D_0(\boldsymbol{\theta} - \boldsymbol{\theta}_1).$$

The deterministic part of it reads as:

$$D_0^{-1} \nabla_{\boldsymbol{\theta}} I\!\!E^\circ \mathcal{A}^\circ(\boldsymbol{\theta}, \boldsymbol{\theta}_1) = D_0^{-1} \left\{ \nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta}) - \nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta}_1) \right\} + D_0(\boldsymbol{\theta} - \boldsymbol{\theta}_1).$$

Proposition 3.1 in Spokoiny (2013) implies due to the conditions $(\mathcal{L}_0)$, (ED_2) , that the following random event holds with $I\!P$ -probability at least $1 - \mathrm{e}^{-\mathtt{x}}$ for all $\boldsymbol{\theta}, \boldsymbol{\theta}_1 \in \Theta_0(\mathtt{r})$ and $\mathtt{r} \leq \mathtt{r}_0$:

$$\begin{aligned} \|D_0^{-1} \nabla_{\boldsymbol{\theta}} I\!\!E^\circ \mathcal{A}^\circ(\boldsymbol{\theta}, \boldsymbol{\theta}_1)\| &= \left\| D_0^{-1} \left\{ \nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta}) - \nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta}_1) \right\} + D_0(\boldsymbol{\theta} - \boldsymbol{\theta}_1) \right\| \\ &\leq \frac{2}{3} \Delta_{\mathrm{w}}(\mathtt{r}, \mathtt{x}), \end{aligned} \tag{A.9}$$

where the deterministic error term $\Delta_{\mathrm{w}}(\mathtt{r}, \mathtt{x})$ is given in (A.2).

Denote the stochastic part of $D_0^{-1} \nabla_{\boldsymbol{\theta}} \mathcal{A}^\circ(\boldsymbol{\theta}, \boldsymbol{\theta}_1)$ as follows:

$$\begin{aligned} \mathcal{Y}^\circ(\boldsymbol{\theta}, \boldsymbol{\theta}_1) &\stackrel{\text{def}}{=} D_0^{-1} \left\{ \nabla_{\boldsymbol{\theta}} \mathcal{A}^\circ(\boldsymbol{\theta}, \boldsymbol{\theta}_1) - \nabla_{\boldsymbol{\theta}} I\!\!E^\circ \mathcal{A}^\circ(\boldsymbol{\theta}, \boldsymbol{\theta}_1) \right\} \\ &= \sum_{i=1}^n D_0^{-1} \left\{ \nabla_{\boldsymbol{\theta}} \ell_i(\boldsymbol{\theta}) - \nabla_{\boldsymbol{\theta}} \ell_i(\boldsymbol{\theta}_1) \right\} (u_i - 1). \end{aligned}$$

In order to bound its norm's supremum w.r.t. $\boldsymbol{\theta} \in \Theta_0(\mathtt{r})$ for $\mathtt{r} \leq \mathtt{r}_0$ we use the idea from the proof of Proposition 3.1 in Spokoiny (2013). Let us introduce the new parameters $\boldsymbol{\upsilon} \stackrel{\text{def}}{=} D_0(\boldsymbol{\theta} - \boldsymbol{\theta}^*)$ and $\boldsymbol{\upsilon}_1 \stackrel{\text{def}}{=} D_0(\boldsymbol{\theta}_1 - \boldsymbol{\theta}^*)$, then

$$\nabla_{\boldsymbol{\upsilon}} \mathcal{Y}^\circ(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}_1) = \sum_{i=1}^n D_0^{-1} \nabla^2_{\boldsymbol{\theta}} \ell_i(\boldsymbol{\theta}) D_0^{-1} (u_i - 1).$$

Thus, we obtain a proper normalisation for $\nabla_{\upsilon}\mathcal{Y}^{\circ}(\upsilon,\upsilon_1)$. Independency of $u_1,\ldots,u_n$ and Lemma A.5 imply with probability $\geq 1-\mathrm{e}^{-\mathtt{x}}$ for $j=1,2$ and $\omega_1(\mathtt{r})$ given in (A.12):

$$\sup_{\substack{\gamma_j\in I\!R^p\\ \|\gamma_j\|=1}} \log I\!E^{\circ}\exp\left\{\frac{\lambda}{\omega_1(\mathtt{r})}\gamma_1^{\top}\nabla_{\upsilon}\mathcal{Y}^{\circ}(\upsilon,\upsilon_1)\gamma_2\right\}\leq\frac{\lambda^2\nu_0^2}{2},\quad |\lambda|\leq \mathtt{g}_2(\mathtt{r}).$$

This allows to apply Theorem A.3 from Spokoiny (2013) on a uniform bound for the norm of stochastic process to $\omega_1^{-1}(\mathtt{r})\mathcal{Y}^{\circ}(\boldsymbol{\theta},\boldsymbol{\theta}_1)$. By the triangle inequality it holds for $\mathtt{r}\leq\mathtt{r}_0$:

$$I\!P^{\circ}\left(\sup_{\boldsymbol{\theta},\boldsymbol{\theta}_1\in\Theta_0(\mathtt{r})}\|\mathcal{Y}^{\circ}(\boldsymbol{\theta},\boldsymbol{\theta}_1)\|\leq 12\nu_0\mathtt{r}\omega_1(\mathtt{r})\mathfrak{Z}(\mathtt{x})\right)\geq 1-\mathrm{e}^{-\mathtt{x}}, \tag{A.10}$$

where $\mathfrak{Z}(\mathtt{x})$ is defined in (A.3). Collecting together the bounds (A.8), (A.9) and (A.10) we obtain that the following bound holds with $I\!P$-probability at least $1-\mathrm{e}^{-\mathtt{x}}$:

$$I\!P^{\circ}\left(\sup_{\boldsymbol{\theta},\boldsymbol{\theta}_1\in\Theta_0(\mathtt{r})}|\mathcal{A}^{\circ}(\boldsymbol{\theta},\boldsymbol{\theta}_1)|\leq 4\mathtt{r}\left\{\Delta_{\mathrm{w}}(\mathtt{r},\mathtt{x})/3+6\nu_0\mathtt{r}\omega_1(\mathtt{r})\mathfrak{Z}(\mathtt{x})\right\}\right)\geq 1-\mathrm{e}^{-\mathtt{x}}$$

for $\mathtt{r}\leq\mathtt{r}_0$.

Theorems A.6 and A.1 say that the maximum likelihood estimators $\widetilde{\boldsymbol{\theta}}^{\circ}$ and $\widetilde{\boldsymbol{\theta}}$ get into the local vicinity $\Theta_0(\mathtt{r}_0)$ with exponentially high $I\!P^{\circ}$- and $I\!P$-probabilities correspondingly. Therefore, taking $\boldsymbol{\theta}=\widetilde{\boldsymbol{\theta}}^{\circ}$ and $\boldsymbol{\theta}_1=\widetilde{\boldsymbol{\theta}}$ in the last bound, we obtain with dominating probability:

$$\begin{aligned}&\left|L^{\circ}(\widetilde{\boldsymbol{\theta}}^{\circ})-L^{\circ}(\widetilde{\boldsymbol{\theta}})-(\widetilde{\boldsymbol{\theta}}^{\circ}-\widetilde{\boldsymbol{\theta}})^{\top}\nabla_{\boldsymbol{\theta}}L^{\circ}(\widetilde{\boldsymbol{\theta}})+\frac{1}{2}\|D_0(\widetilde{\boldsymbol{\theta}}^{\circ}-\widetilde{\boldsymbol{\theta}})\|^2\right|\\ &\quad\leq 4\mathtt{r}\left\{\Delta_{\mathrm{w}}(\mathtt{r}_0,\mathtt{x})/3+6\nu_0\mathtt{r}_0\omega_1(\mathtt{r})\mathfrak{Z}(\mathtt{x})\right\}.\end{aligned}$$

Similarly bounds (A.9) and (A.10) imply:

$$\begin{aligned}&\frac{1}{2}\left|\|\boldsymbol{\xi}^{\circ}(\widetilde{\boldsymbol{\theta}})\|^2-2(\widetilde{\boldsymbol{\theta}}^{\circ}-\widetilde{\boldsymbol{\theta}})^{\top}\nabla_{\boldsymbol{\theta}}L^{\circ}(\widetilde{\boldsymbol{\theta}})+\|D_0(\widetilde{\boldsymbol{\theta}}^{\circ}-\widetilde{\boldsymbol{\theta}})\|^2\right|\\ &\quad=\frac{1}{2}\left\|D_0^{-1}\nabla_{\boldsymbol{\theta}}L^{\circ}(\widetilde{\boldsymbol{\theta}})-D_0(\widetilde{\boldsymbol{\theta}}^{\circ}-\widetilde{\boldsymbol{\theta}})\right\|^2\\ &\quad\leq 2\left\{\Delta_{\mathrm{w}}(\mathtt{r}_0,\mathtt{x})/3+6\nu_0\mathtt{r}_0\omega_1(\mathtt{r})\mathfrak{Z}(\mathtt{x})\right\}^2.\end{aligned} \tag{A.11}$$

Therefore it holds with $I\!P$-probablity at least $1-4\mathrm{e}^{-\mathtt{x}}$:

$$I\!P^{\circ}\left(\left|L^{\circ}(\widetilde{\boldsymbol{\theta}}^{\circ})-L^{\circ}(\widetilde{\boldsymbol{\theta}})-\frac{1}{2}\|\boldsymbol{\xi}^{\circ}(\widetilde{\boldsymbol{\theta}})\|^2\right|\leq\Delta^{\circ}_{\mathrm{w}\,2}(\mathtt{r}_0,\mathtt{x})\right)\geq 1-4\mathrm{e}^{-\mathtt{x}},$$

$$\begin{aligned}\Delta^{\circ}_{\mathrm{w}\,2}(\mathtt{r}_0,\mathtt{x})&\stackrel{\mathrm{def}}{=}4\mathtt{r}\left\{\Delta_{\mathrm{w}}(\mathtt{r}_0,\mathtt{x})/3+6\nu_0\mathtt{r}_0\omega_1(\mathtt{r})\mathfrak{Z}(\mathtt{x})\right\}\\ &\quad+2\left\{\Delta_{\mathrm{w}}(\mathtt{r}_0,\mathtt{x})/3+6\nu_0\mathtt{r}_0\omega_1(\mathtt{r})\mathfrak{Z}(\mathtt{x})\right\}^2.\end{aligned}$$

For the second bound of the statement we use the similar approach as in Theorem 2.3 in Spokoiny (2013).

$$\begin{aligned}
&\left| \sqrt{2\left\{ L^{\circ}(\widetilde{\boldsymbol{\theta}}^{\circ}) - L^{\circ}(\widetilde{\boldsymbol{\theta}}) \right\}} - \| D_0(\widetilde{\boldsymbol{\theta}}^{\circ} - \widetilde{\boldsymbol{\theta}}) \| \right| \\
&\quad \leq \frac{\left| 2\left\{ L^{\circ}(\widetilde{\boldsymbol{\theta}}^{\circ}) - L^{\circ}(\widetilde{\boldsymbol{\theta}}) \right\} - \| D_0(\widetilde{\boldsymbol{\theta}}^{\circ} - \widetilde{\boldsymbol{\theta}}) \|^2 \right|}{\| D_0(\widetilde{\boldsymbol{\theta}}^{\circ} - \widetilde{\boldsymbol{\theta}}) \|} \\
&\quad = \frac{\left| 2\mathcal{A}^{\circ}(\widetilde{\boldsymbol{\theta}}, \widetilde{\boldsymbol{\theta}}^{\circ}) \right|}{\| D_0(\widetilde{\boldsymbol{\theta}}^{\circ} - \widetilde{\boldsymbol{\theta}}) \|} \leq \sup_{\boldsymbol{\theta}, \boldsymbol{\theta}_1 \in \Theta_0(\mathtt{r}_0)} \frac{|2\mathcal{A}^{\circ}(\boldsymbol{\theta}, \boldsymbol{\theta}_1)|}{\| D_0(\boldsymbol{\theta} - \boldsymbol{\theta}_1) \|} \\
&\quad \overset{\text{by (A.7)}}{\leq} \sup_{\boldsymbol{\theta}, \boldsymbol{\theta}_1 \in \Theta_0(\mathtt{r}_0)} 2 \left\| D_0^{-1} \nabla_{\boldsymbol{\theta}} \mathcal{A}^{\circ}(\boldsymbol{\theta}, \boldsymbol{\theta}_1) \right\| \\
&\quad \overset{\text{by (A.9), (A.10)}}{\leq} 4\Delta_{\mathrm{w}}(\mathtt{r}_0, \mathtt{x})/3 + 24\nu_0 \mathtt{r}_0 \omega_1(\mathtt{r}) \mathfrak{Z}(\mathtt{x}).
\end{aligned}$$

This together with (A.11) imply the final statement. □

Lemma A.5 (Check of the bootstrap equivalent of $(\boldsymbol{ED_2})$). *Conditions* $(\boldsymbol{Eb})$, $(\boldsymbol{\mathfrak{L}_{0m}})$ *and* $(\boldsymbol{ED_{2m}})$ *imply for each* $\mathtt{r} > 0$, $\boldsymbol{\theta} \in \Theta_0(\mathtt{r})$, $\|\boldsymbol{\gamma}_j\| = 1$, $j = 1, 2$ *and all* $|\lambda| \leq \mathtt{g}_2(\mathtt{r})$ *with probability* $\geq 1 - \mathrm{e}^{-\mathtt{x}}$ *:*

$$\sup_{\substack{\boldsymbol{\gamma}_j \in I\!R^p \\ \|\boldsymbol{\gamma}_j\| = 1}} \sum_{i=1}^{n} \log I\!E^{\circ} \exp\left\{ \frac{\lambda}{\omega_1(\mathtt{r})} \boldsymbol{\gamma}_1^{\top} D_0^{-1} \nabla^2_{\boldsymbol{\theta}} \ell_i(\boldsymbol{\theta}) D_0^{-1} \boldsymbol{\gamma}_2 (u_i - 1) \right\} \leq \frac{\lambda^2 \nu_0^2}{2}.$$

where

$$\omega_1(\mathtt{r}) = \omega_1 \overset{\mathrm{def}}{=} \frac{\mathtt{C}_m(\mathtt{r})}{\sqrt{n}} + 2\omega\nu_0\sqrt{2\mathtt{x}} \tag{A.12}$$

In the case 4.3 it holds for $\mathtt{r} \leq \mathtt{r}_0$ $\omega_1(\mathtt{r}) = \mathtt{C}\mathtt{r}/n + \mathtt{C}\sqrt{\mathtt{x}/n}$ *.*

Proof of Lemma A.5. Introduce the independent random scalar values for $i = 1, \ldots, n$ and $j = 1, 2$:

$$\mu_i(\boldsymbol{\theta}, \boldsymbol{\gamma}_j) \overset{\mathrm{def}}{=} \boldsymbol{\gamma}_1^{\top} D_0^{-1} \nabla^2_{\boldsymbol{\theta}} \ell_i(\boldsymbol{\theta}) D_0^{-1} \boldsymbol{\gamma}_2.$$

It holds

$$\sum_{i=1}^{n} \log I\!E^{\circ} \exp\left\{ \frac{\lambda}{\omega_1} \boldsymbol{\gamma}_1^{\top} D_0^{-1} \nabla^2_{\boldsymbol{\theta}} \ell_i(\boldsymbol{\theta}) D_0^{-1} \boldsymbol{\gamma}_2 (u_i - 1) \right\}$$

$$= \sum_{i=1}^{n} \log I\!E^{\circ} \exp\left\{ \frac{\lambda}{\omega_1} \mu_i(\boldsymbol{\theta}, \boldsymbol{\gamma}_j)(u_i - 1) \right\}$$

$$\leq \frac{\lambda^2 \nu_0^2}{2\omega_1^2} \sum_{i=1}^{n} \mu_i^2(\boldsymbol{\theta}, \boldsymbol{\gamma}_j), \tag{A.13}$$

here the inequality (A.13) follows from condition $(\boldsymbol{Eb})$ if $\left|\lambda \mu_i(\boldsymbol{\theta}, \boldsymbol{\gamma}_j)\right| \leq \mathtt{g}\omega_1$ for all $i = 1, \ldots, n$, which is true due to the arguments below. Let us consider $\mu_i(\boldsymbol{\theta}, \boldsymbol{\gamma}_j)$, for each $\boldsymbol{\theta} \in \Theta_0(\mathtt{r})$, $i = 1, \ldots n$ it holds:

$$\begin{aligned} \left|\mu_i(\boldsymbol{\theta}, \boldsymbol{\gamma}_j)\right| \leq & \| D_0^{-1} \nabla^2_{\boldsymbol{\theta}} I\!E \ell_i(\boldsymbol{\theta}) D_0^{-1} \| \\ & + \| D_0^{-1} \left\{ \nabla^2_{\boldsymbol{\theta}} \ell_i(\boldsymbol{\theta}) - \nabla^2_{\boldsymbol{\theta}} I\!E \ell_i(\boldsymbol{\theta}) \right\} D_0^{-1} \|. \end{aligned} \tag{A.14}$$

Condition $(\boldsymbol{ED_{2m}})$, which is a stronger version of $(\boldsymbol{ED_2})$, implies that for all $i = 1, \ldots, n$, $\boldsymbol{\theta} \in \Theta_0(\mathtt{r})$ and each $\mathtt{r} > 0$ it holds with $I\!P$-probability $\geq 1 - \mathrm{e}^{-\mathtt{x}}$

$$\| D_0^{-1} \left\{ \nabla^2_{\boldsymbol{\theta}} \ell_i(\boldsymbol{\theta}) - \nabla^2_{\boldsymbol{\theta}} I\!E \ell_i(\boldsymbol{\theta}) \right\} D_0^{-1} \| \leq 2\omega\nu_0 \left(\frac{2\mathtt{x}}{n} \right)^{1/2}. \tag{A.15}$$

Indeed, by the exponential Chebyshev inequality for $\lambda > 0$

$$\begin{aligned} & I\!P\left(\omega^{-1} \| D_0^{-1} \left\{ \nabla^2_{\boldsymbol{\theta}} \ell_i(\boldsymbol{\theta}) - \nabla^2_{\boldsymbol{\theta}} I\!E \ell_i(\boldsymbol{\theta}) \right\} D_0^{-1} \| \geq t \right) \\ \leq & \quad I\!E \exp\left[-\lambda t + \omega^{-1} \lambda \| D_0^{-1} \left\{ \nabla^2_{\boldsymbol{\theta}} \ell_i(\boldsymbol{\theta}) - \nabla^2_{\boldsymbol{\theta}} I\!E \ell_i(\boldsymbol{\theta}) \right\} D_0^{-1} \| \right] \\ \overset{\text{by } (\boldsymbol{ED_{2m}})}{\leq} & \quad \exp\left\{ -\lambda t + \lambda^2 \nu_0^2 / (2n) \right\}, \quad 0 < \lambda < \mathtt{g}_2(\mathtt{r}) \\ \leq & \quad \exp\{-\mathtt{x}\}, \end{aligned}$$

here the last inequality holds under the assumption, that $\mathtt{g}_2(\mathtt{r})$ is large enough. In the case 4.3 it holds $\mathtt{g}_2(\mathtt{r}) = \mathtt{C} n^{1/2}$, $\omega = \mathtt{C} n^{-1/2}$ and $\mathtt{x} = \mathtt{C} \log(n)$; $t^2 := 8\nu_0^2 \mathtt{x}/n$ implies $\lambda t - \lambda^2 \nu_0^2/(2n) - \mathtt{x} \geq 0$ for $0 < \lambda < \mathtt{g}_2(\mathtt{r})$. For the deterministic term in (A.14) condition $(\boldsymbol{\mathcal{L}_{0m}})$ reads as:

$$\| D_0^{-1} \nabla^2_{\boldsymbol{\theta}} I\!E \ell_i(\boldsymbol{\theta}) D_0^{-1} \| \leq \frac{\mathtt{C}_m(\mathtt{r})}{n}. \tag{A.16}$$

Collecting the inequalities (A.13), (A.14), (A.15) and (A.16), we obtain:

$$\sum_{i=1}^{n} \log I\!E^{\circ} \exp\left\{ \frac{\lambda}{\omega_1} \boldsymbol{\gamma}_1^{\top} D_0^{-1} \nabla^2_{\boldsymbol{\theta}} \ell_i(\boldsymbol{\theta}) D_0^{-1} \boldsymbol{\gamma}_2 (u_i - 1) \right\}$$

$$\leq \frac{\lambda^2 \nu_0^2}{2} \frac{1}{{\omega_1}^2} \left\{ \frac{\mathtt{C}_m(\mathtt{r})}{\sqrt{n}} + 2\omega\nu_0 \sqrt{2\mathtt{x}} \right\}^2$$

Taking $\omega_1 = \omega_1(\mathtt{r})$ as in (A.12) implies the necessary statement. □

Theorem A.6 (Concentration of bootstrap MLE). *Let the conditions of Theorems A.1 and A.8, $(\mathcal{L}_{0m})$ and (ED_{2m}) be fulfilled. If the following holds for $\omega_1(\mathtt{r})$ defined in (A.12) and the $I\!P$ -random matrix $\mathcal{B}^2 \stackrel{\text{def}}{=} D_0^{-1} \operatorname{Var}^\circ \{\nabla_{\boldsymbol{\theta}} L^\circ(\boldsymbol{\theta}^*)\} D_0^{-1}$*

$$\begin{aligned} \mathtt{b}(\mathtt{r})\mathtt{r} \geq 2 \{ & \mathfrak{Z}_{\mathtt{qf}}(\mathtt{x}, I\!B) + \mathfrak{Z}_{\mathtt{qf}}(\mathtt{x}, \mathcal{B}) + 6\nu_0 \mathfrak{Z}(\mathtt{x}) \omega_1(\mathtt{r}_0) \mathtt{r}_0 \} \\ & + 12\nu_0(\omega + \omega_1(\mathtt{r}))\mathfrak{Z}(\mathtt{x} + \log(2\mathtt{r}/\mathtt{r}_0)) \quad \textit{for } \mathtt{r} > \mathtt{r}_0, \end{aligned} \tag{A.17}$$

then it holds with $I\!P$ -probability $\geq 1 - 3\mathrm{e}^{-\mathtt{x}}$

$$I\!P^\circ \left(\widetilde{\boldsymbol{\theta}}^\circ \notin \Theta_0(\mathtt{r}_0) \right) \leq 3\mathrm{e}^{-\mathtt{x}}.$$

Proof of Theorem A.6. We use the idea by Spokoiny (2013): if $\sup_{\boldsymbol{\theta} \in \Theta \setminus \Theta_0(\mathtt{r}_0)} \{ L(\boldsymbol{\theta}) - L(\boldsymbol{\theta}^*) \} < 0$, then $\widetilde{\boldsymbol{\theta}} \in \Theta_0(\mathtt{r}_0)$. We apply it here for the the bootstrap objects: $L^\circ(\boldsymbol{\theta}) - L^\circ(\widetilde{\boldsymbol{\theta}})$ and $\widetilde{\boldsymbol{\theta}}^\circ$. Denote the stochastic part of the bootstrap likelihood process as $\boldsymbol{\zeta}^\circ(\boldsymbol{\theta}) \stackrel{\text{def}}{=} L^\circ(\boldsymbol{\theta}) - I\!E^\circ L^\circ(\boldsymbol{\theta})$. It holds

$$\begin{aligned} L^\circ(\boldsymbol{\theta}) - L^\circ(\widetilde{\boldsymbol{\theta}}) &= \boldsymbol{\zeta}^\circ(\boldsymbol{\theta}) - \boldsymbol{\zeta}^\circ(\widetilde{\boldsymbol{\theta}}) + I\!E^\circ L^\circ(\boldsymbol{\theta}) - I\!E^\circ L^\circ(\widetilde{\boldsymbol{\theta}}) \\ &= \boldsymbol{\zeta}^\circ(\boldsymbol{\theta}) - \boldsymbol{\zeta}^\circ(\widetilde{\boldsymbol{\theta}}) + L(\boldsymbol{\theta}) - L(\widetilde{\boldsymbol{\theta}}) \\ &= \{ \boldsymbol{\zeta}^\circ(\boldsymbol{\theta}) - \boldsymbol{\zeta}^\circ(\widetilde{\boldsymbol{\theta}}) \} + \{ L(\boldsymbol{\theta}) - L(\boldsymbol{\theta}^*) \} + \{ L(\boldsymbol{\theta}^*) - L(\widetilde{\boldsymbol{\theta}}) \}. \end{aligned}$$

Here the last summand $\{ L(\boldsymbol{\theta}^*) - L(\widetilde{\boldsymbol{\theta}}) \}$ is non-positive by definition (1.2) of $\widetilde{\boldsymbol{\theta}}$. The following bound follows from the proof of Theorem 2.1 in Spokoiny (2013):

$$I\!P \left(\sup_{\boldsymbol{\theta} \in \Theta \setminus \Theta_0(\mathtt{r}_0)} \{ L(\boldsymbol{\theta}) - L(\boldsymbol{\theta}^*) \} < \varrho(\mathtt{r}, \mathtt{x})\mathtt{r} + \mathtt{r}\mathfrak{Z}_{\mathtt{qf}}(\mathtt{x}, I\!B) - \mathtt{r}^2 \mathtt{b}(\mathtt{r})/2 \right) \geq 1 - 3\mathrm{e}^{-\mathtt{x}},$$

$$\varrho(\mathtt{r}, \mathtt{x}) \stackrel{\text{def}}{=} 6\nu_0 \mathfrak{Z}(\mathtt{x} + \log(2\mathtt{r}/\mathtt{r}_0))\omega.$$

Due to Lemma A.5 the process $\boldsymbol{\zeta}^\circ(\boldsymbol{\theta}) - \boldsymbol{\zeta}^\circ(\widetilde{\boldsymbol{\theta}})$ satisfies the necessary conditions of Theorem A.1 in Spokoiny (2013), and it holds for $\mathtt{r} \geq \mathtt{r}_0$

$$I\!P^\circ \left(\sup_{\boldsymbol{\theta} \in \Theta_0(\mathtt{r})} \left| \boldsymbol{\zeta}^\circ(\boldsymbol{\theta}) - \boldsymbol{\zeta}^\circ(\widetilde{\boldsymbol{\theta}}) - (\boldsymbol{\theta} - \widetilde{\boldsymbol{\theta}})^\top \nabla_{\boldsymbol{\theta}} \boldsymbol{\zeta}^\circ(\widetilde{\boldsymbol{\theta}}) \right| \leq \varrho_1(\mathtt{r}, \mathtt{x})\mathtt{r} \right) \geq 1 - \mathrm{e}^{-\mathtt{x}},$$

$$\varrho_1(\mathtt{r}, \mathtt{x}) \stackrel{\text{def}}{=} 6\nu_0 \mathfrak{Z}(\mathtt{x} + \log(2\mathtt{r}/\mathtt{r}_0))\omega_1(\mathtt{r}).$$

By Lemma A.7 and Theorem A.8 it holds with dominating probability

$$\begin{aligned} \sup_{\boldsymbol{\theta} \in \Theta_0(\mathtt{r})} \left| (\boldsymbol{\theta} - \widetilde{\boldsymbol{\theta}})^\top \nabla_{\boldsymbol{\theta}} \boldsymbol{\zeta}^\circ(\widetilde{\boldsymbol{\theta}}) \right| &\leq \mathtt{r} \| \boldsymbol{\xi}^\circ(\widetilde{\boldsymbol{\theta}}) \| \\ &\leq \mathtt{r} \left\{ \| \boldsymbol{\xi}^\circ(\boldsymbol{\theta}^*) \| + \| \boldsymbol{\xi}^\circ(\widetilde{\boldsymbol{\theta}}) - \boldsymbol{\xi}^\circ(\boldsymbol{\theta}^*) \| \right\} \\ &\leq \mathtt{r} \{ \mathfrak{Z}_{\mathtt{qf}}(\mathtt{x}, \mathcal{B}) + 6\nu_0 \mathfrak{Z}(\mathtt{x}) \omega_1(\mathtt{r}_0) \mathtt{r}_0 \}. \end{aligned}$$

Finally we have:

$$\begin{aligned}\sup_{\boldsymbol{\theta}\in\Theta\setminus\Theta_0(\mathtt{r}_0)}\Big\{L^\circ(\boldsymbol{\theta})-L^\circ(\widetilde{\boldsymbol{\theta}})\Big\} &\le \sup_{\boldsymbol{\theta}\in\Theta\setminus\Theta_0(\mathtt{r}_0)}\Big\{L(\boldsymbol{\theta})-L(\boldsymbol{\theta}^*)\Big\}\\ &\quad+\sup_{\substack{\boldsymbol{\theta}\in\Theta_0(\mathtt{r}),\\ \mathtt{r}\ge\mathtt{r}_0}}\Big\{\boldsymbol{\zeta}^\circ(\boldsymbol{\theta})-\boldsymbol{\zeta}^\circ(\widetilde{\boldsymbol{\theta}})\Big\}\\ &\le \mathtt{r}\mathfrak{Z}_{\mathtt{qf}}(\mathtt{x},\mathcal{B})+\mathtt{r}\mathfrak{Z}_{\mathtt{qf}}(\mathtt{x},I\!\!B)+\varrho_1(\mathtt{r},\mathtt{x})\mathtt{r}\\ &\quad+\varrho(\mathtt{r},\mathtt{x})\mathtt{r}-\mathtt{r}^2\mathtt{b}(\mathtt{r})/2+6\nu_0\mathfrak{Z}(\mathtt{x})\omega_1(\mathtt{r}_0)\mathtt{r}\mathtt{r}_0,\end{aligned}$$

which implies the condition (A.17) in the statement. □

REMARK A.1. Condition (A.17) imposed for the bootstrap MLE concentration result is stronger, than condition (A.1) for the concentration of $\boldsymbol{Y}$ - MLE, and (A.17) implies the latter one.

The following lemma had already been derived in the proof of Theorem A.4: see the bound (A.10). We formulate it separately, since it is used again in another statements.

Lemma A.7. *Let the conditions of Lemma A.5 be fulfilled, then it holds with $I\!\!P$ -probability $\ge 1-\mathrm{e}^{-\mathtt{x}}$*

$$I\!\!P^\circ\left(\sup_{\boldsymbol{\theta}\in\Theta_0(\mathtt{r})}\|\boldsymbol{\xi}^\circ(\boldsymbol{\theta})-\boldsymbol{\xi}^\circ(\boldsymbol{\theta}^*)\|\le\Delta^\circ_\xi(\mathtt{r},\mathtt{x})\right)\ge 1-\mathrm{e}^{-\mathtt{x}},$$

where

$$\Delta^\circ_\xi(\mathtt{r},\mathtt{x})\stackrel{\mathrm{def}}{=}6\nu_0\mathfrak{Z}(\mathtt{x})\omega_1(\mathtt{r})\mathtt{r}$$

In the case 4.3 it holds for the bounding term.

$$\Delta^\circ_\xi(\mathtt{r}_0,\mathtt{x})\le \mathtt{C}\frac{p+\mathtt{x}}{\sqrt{n}}\sqrt{\mathtt{x}}.$$

Theorem A.8 (Deviation bound for the bootstrap quadratic form)**.** *Let conditions $(\boldsymbol{Eb})$, $(\boldsymbol{\mathcal{I}})$, $(\boldsymbol{SD_1})$, $(\boldsymbol{\mathcal{I}_B})$ be fulfilled, then for $\mathtt{g}\ge\sqrt{2\,\mathrm{tr}(\mathcal{B}^2)}$ it holds:*

$$I\!\!P^\circ\left(\|\boldsymbol{\xi}^\circ(\boldsymbol{\theta}^*)\|^2\le\mathfrak{Z}^2_{\mathtt{qf}}(\mathtt{x},\mathcal{B})\right)\ge 1-2\mathrm{e}^{-\mathtt{x}}-8.4\mathrm{e}^{-\mathtt{x}_c(\mathcal{B})},$$

where

$$\mathcal{B}^2\stackrel{\mathrm{def}}{=}D_0^{-1}\mathcal{V}^2(\boldsymbol{\theta}^*)D_0^{-1},\quad \mathcal{V}^2(\boldsymbol{\theta}^*)\stackrel{\mathrm{def}}{=}\mathrm{Var}^\circ\nabla_{\boldsymbol{\theta}}L^\circ(\boldsymbol{\theta}^*),\tag{A.18}$$

$\mathfrak{Z}_{\mathtt{qf}}(\mathtt{x},\cdot)$ and $\mathtt{x}_c(\cdot)$ are defined respectively in (A.4) and (A.5). Similarly to (A.6) it holds for $\mathtt{x}\le\mathtt{x}_c(\mathcal{B})$:

$$\mathfrak{Z}^2_{\mathtt{qf}}(\mathtt{x},\mathcal{B})\ \le\ \mathfrak{a}^{\circ 2}(p+6\mathtt{x})\tag{A.19}$$

$$\text{for }\mathfrak{a}^{\circ 2}\stackrel{\mathrm{def}}{=}(1+\delta^2_{\mathcal{V}})(\mathfrak{a}^2+\mathfrak{a}^2_B).$$

Proof of Theorem A.8. This result is the bootstrap equivalent of Theorem A.3. For the $\boldsymbol{Y}$ -world it demands condition $(\boldsymbol{ED_0})$ to be fulfilled. Let us check whether the bootstrap equivalent of $(\boldsymbol{ED_0})$ holds. It reads as follows: *there exist constants* $\mathrm{g}^\circ > 0$, $\nu_0^\circ \geq 1$ *such that for the positive-definite symmetric matrix* $\mathcal{V}^2(\boldsymbol{\theta}^*)$ *it holds for all* $|\lambda| \leq \mathrm{g}^\circ$

$$\sup_{\boldsymbol{\gamma} \in I\!R^p} \log I\!E^\circ \exp\left\{ \lambda \frac{\boldsymbol{\gamma}^\top \{\nabla_{\boldsymbol{\theta}} L^\circ(\boldsymbol{\theta}^*) - \nabla_{\boldsymbol{\theta}} I\!E^\circ L^\circ(\boldsymbol{\theta}^*)\}}{\|\mathcal{V}(\boldsymbol{\theta}^*)\boldsymbol{\gamma}\|} \right\} \leq \nu_0^{\circ 2} \lambda^2/2.$$

By definition $\mathcal{V}^2(\boldsymbol{\theta}^*) = \sum_{i=1}^n \nabla_{\boldsymbol{\theta}} \ell_i(\boldsymbol{\theta}^*) \nabla_{\boldsymbol{\theta}} \ell_i(\boldsymbol{\theta}^*)^\top$. Let us introduce the independent $I\!P$ -random variables $s_i(\boldsymbol{\gamma}) \stackrel{\mathrm{def}}{=} \boldsymbol{\gamma}^\top \nabla_{\boldsymbol{\theta}} \ell_i(\boldsymbol{\theta}^*) / \|\mathcal{V}(\boldsymbol{\theta}^*)\boldsymbol{\gamma}\|$ for $i = 1, \ldots, n$. It holds $\sum_{i=1}^n s_i^2(\boldsymbol{\gamma}) = 1$, hence $\max_{1 \leq i \leq n} |s_i| \leq 1$. Condition $(\boldsymbol{Eb})$ implies:

$$\begin{aligned}
&\log I\!E^\circ \exp\left\{ \lambda \frac{\boldsymbol{\gamma}^\top \{\nabla_{\boldsymbol{\theta}} L^\circ(\boldsymbol{\theta}^*) - \nabla_{\boldsymbol{\theta}} I\!E^\circ L^\circ(\boldsymbol{\theta}^*)\}}{\|\mathcal{V}(\boldsymbol{\theta}^*)\boldsymbol{\gamma}\|} \right\} \\
&\quad = \sum_{i=1}^n \log I\!E^\circ \exp\{\lambda s_i(\boldsymbol{\gamma})(u_i - 1)\} \\
&\quad \leq \frac{{\nu_0}^2 \lambda^2}{2} \sum_{i=1}^n s_i^2(\boldsymbol{\gamma}) = {\nu_0}^2 \lambda^2/2, \qquad |\lambda| \leq \mathrm{g}.
\end{aligned}$$

Thus the bootstrap equivalent for the condition $(\boldsymbol{ED_0})$ is fulfilled with the same constants ν_0, g , and the theorem's statements holds as well as for Theorem A.3.

The inequality (A.19) follows from conditions $(\boldsymbol{\mathcal{I}})$, $(\boldsymbol{\mathcal{I}_B})$, $(\boldsymbol{SD_1})$ and Bernstein matrix inequality by Tropp (2012) (see Section A.6):

$$\|D_0^{-1} \mathcal{V}_0^2(\boldsymbol{\theta}^*) D_0^{-1}\| \leq \|D_0^{-1} H_0\|^2 (1 + \delta_{\mathcal{V}}^2) \leq (1 + \delta_{\mathcal{V}}^2)(\mathfrak{a}^2 + \mathfrak{a}_B^2).$$

□

A.3 Proofs of Theorems 2.1 and 2.3

In order to justify theoretically the multiplier bootstrap procedure it has to be shown that the approximating terms $\|\boldsymbol{\xi}\|$ and $\|\boldsymbol{\xi}^\circ(\widetilde{\boldsymbol{\theta}})\|$ from the Wilks Theorems A.2 and A.4 have nearly the same distributions. By Lemma A.7 the random values $\|\boldsymbol{\xi}^\circ(\boldsymbol{\theta}^*)\|$ and $\|\boldsymbol{\xi}^\circ(\widetilde{\boldsymbol{\theta}})\|$ are close to each other within the error term $\leq \mathtt{C}(p + \mathtt{x})\sqrt{\mathtt{x}/n}$ with exponentially high probability, therefore, it is sufficient to compare the distributions of $\|\boldsymbol{\xi}^\circ(\boldsymbol{\theta}^*)\|$ and $\|\boldsymbol{\xi}\|$. This is done in Proposition A.9 using the results on Gaussian approximation for Euclidean norms from Section 5.

Let us introduce the multivariate normal vectors similarly to (5.3):

$$\overline{\boldsymbol{\xi}} \sim \mathcal{N}(0, \operatorname{Var} \boldsymbol{\xi}), \quad \overline{\boldsymbol{\xi}}^\circ(\boldsymbol{\theta}^*) \sim \mathcal{N}(0, \operatorname{Var}^\circ\{\boldsymbol{\xi}^\circ(\boldsymbol{\theta}^*)\}). \tag{A.20}$$

Let us also represent the vectors $\boldsymbol{\xi}$ and $\boldsymbol{\xi}^\circ(\boldsymbol{\theta}^*)$ as sums of the marginal score vectors $\boldsymbol{\xi}_i$ and

$\boldsymbol{\xi}_i^\circ(\boldsymbol{\theta}^*)$ s.t. $I\!\!E\boldsymbol{\xi}_i = I\!\!E^\circ\boldsymbol{\xi}_i^\circ = 0$:

$$\begin{aligned}\boldsymbol{\xi}_i &\stackrel{\text{def}}{=} D_0^{-1}\left\{\nabla_{\boldsymbol{\theta}}\ell_i(\boldsymbol{\theta}^*) - \nabla_{\boldsymbol{\theta}} I\!\!E\ell_i(\boldsymbol{\theta}^*)\right\},\\ \boldsymbol{\xi}_i^\circ(\boldsymbol{\theta}^*) &\stackrel{\text{def}}{=} D_0^{-1}\nabla_{\boldsymbol{\theta}}\ell_i(\boldsymbol{\theta}^*)\{u_i - 1\}.\end{aligned}$$

Their Gaussian analogs are

$$\overline{\boldsymbol{\xi}}_i \sim \mathcal{N}(0, \operatorname{Var}\boldsymbol{\xi}_i) \quad \text{and} \quad \overline{\boldsymbol{\xi}}_i^\circ \sim \mathcal{N}(0, \operatorname{Var}^\circ\{\boldsymbol{\xi}_i^\circ(\boldsymbol{\theta}^*)\}).$$

Similarly to (5.4) denote

$$\begin{aligned}\delta_n &\stackrel{\text{def}}{=} \frac{1}{2}\sum_{i=1}^n I\!\!E\left(\|\boldsymbol{\xi}_i\|^3 + \|\overline{\boldsymbol{\xi}}_i\|^3\right),\\ \breve{\delta}_n &\stackrel{\text{def}}{=} \frac{1}{2}\sum_{i=1}^n I\!\!E^\circ\left(\|\boldsymbol{\xi}_i^\circ(\boldsymbol{\theta}^*)\|^3 + \|\overline{\boldsymbol{\xi}}_i^\circ(\boldsymbol{\theta}^*)\|^3\right).\end{aligned} \tag{A.21}$$

Proposition A.9 (Closeness of the c.d.f. of $\|\boldsymbol{\xi}\|$ and $\|\boldsymbol{\xi}^\circ(\boldsymbol{\theta}^*)\|$). *If conditions* $(\mathbf{SmB})$ *and* $(\boldsymbol{SD_1})$ *are fulfilled, then it holds with probability* $\geq 1 - \mathrm{e}^{-\mathtt{x}}$ *for all* $0 < \Delta \leq 0.22$ *and for all* $z, \overline{z} > 2$ *s.t.* $|z - \overline{z}| \leq \delta_z$ *for some* $\delta_z \geq 0$*:*

$$\begin{aligned}&\left|I\!\!P\left(\|\boldsymbol{\xi}\| \geq z\right) - I\!\!P^\circ\left(\|\boldsymbol{\xi}^\circ(\boldsymbol{\theta}^*)\| \geq \overline{z}\right)\right|\\ &\qquad \leq 16\Delta^{-3}\left(\delta_n + \breve{\delta}_n\right) + \frac{2\Delta + \delta_z}{\overline{z}}\sqrt{\frac{p}{2}} + \frac{\sqrt{p}}{2}\frac{\delta_{\mathcal{V}}^2(\mathtt{x}) + \delta_{\mathrm{smb}}^2}{1 - \delta_{\mathcal{V}}^2(\mathtt{x})}\\ &\qquad \leq 16\Delta^{-3}\left(\delta_n + \breve{\delta}_n\right) + \frac{2\Delta + \delta_z}{\sqrt{2}} + \frac{2\sqrt{p}}{3}\left(\delta_{\mathcal{V}}^2(\mathtt{x}) + \delta_{\mathrm{smb}}^2\right)\\ &\qquad\quad \textit{for } \overline{z} \geq \sqrt{p},\ \delta_{\mathcal{V}}^2(\mathtt{x}) \leq 1/4.\end{aligned}$$

Moreover, if $z, \overline{z} \geq \max\{2, \sqrt{p}\}$ *and* $\max\{\delta_n^{1/4}, \breve{\delta}_n^{1/4}\} \leq 0.11$*, then*

$$\begin{aligned}&\left|I\!\!P\left(\|\boldsymbol{\xi}\| \geq z\right) - I\!\!P^\circ\left(\|\boldsymbol{\xi}^\circ(\boldsymbol{\theta}^*)\| \geq \overline{z}\right)\right|\\ &\qquad \leq 1.55\left(\delta_n^{1/4} + \breve{\delta}_n^{1/4}\right) + \frac{\delta_z}{\sqrt{2}} + \frac{2\sqrt{p}}{3}\left(\delta_{\mathcal{V}}^2(\mathtt{x}) + \delta_{\mathrm{smb}}^2\right).\end{aligned} \tag{A.22}$$

Proof of Proposition A.9. We use Theorem 5.1 taking $\boldsymbol{\phi} := \boldsymbol{\xi}$ and $\boldsymbol{\psi} := \boldsymbol{\xi}^\circ(\boldsymbol{\theta}^*)$. Let us check that the conditions (5.5) on the covariance matrices are fulfilled. By definitions (1.7), (1.8) and (A.18)

$$\begin{aligned}\operatorname{Var}\boldsymbol{\xi} &= D_0^{-1}H_0^2D_0^{-1} - D_0^{-1}B_0^2D_0^{-1},\\ \operatorname{Var}^\circ\left\{\boldsymbol{\xi}^\circ(\boldsymbol{\theta}^*)\right\} &= D_0^{-1}\mathcal{V}^2(\boldsymbol{\theta}^*)D_0^{-1}.\end{aligned}$$

Due to Theorem A.13 by Tropp (2012) (see Section A.6) it holds with probability $\geq 1 - \mathrm{e}^{-\mathtt{x}}$

$$\|H_0^{-1}\mathcal{V}^2(\boldsymbol{\theta}^*)H_0^{-1} - \boldsymbol{I}_p\| \leq \delta_{\mathcal{V}}^2(\mathtt{x}), \tag{A.23}$$

therefore, by Cauchy-Schwarz inequality

$$\|\mathcal{V}^{-1}(\boldsymbol{\theta}^*)H_0^2\mathcal{V}^{-1}(\boldsymbol{\theta}^*) - \boldsymbol{I}_p\| \le \delta_{\mathcal{V}}^2(\mathtt{x})(1-\delta_{\mathcal{V}}^2(\mathtt{x}))^{-1}.$$

Condition $\mathbf{(SmB)}$ says that $\|H_0^{-1}B_0^2H_0^{-1}\| \le \delta_{\mathrm{smb}}^2$, therefore, by the triangle inequality it holds:

$$\begin{aligned}\left\|\left[\mathrm{Var}^\circ\left\{\boldsymbol{\xi}^\circ(\boldsymbol{\theta}^*)\right\}\right]^{-1/2}\mathrm{Var}\{\boldsymbol{\xi}\}\left[\mathrm{Var}^\circ\left\{\boldsymbol{\xi}^\circ(\boldsymbol{\theta}^*)\right\}\right]^{-1/2} - \boldsymbol{I}_p\right\| &\le \frac{\delta_{\mathcal{V}}^2(\mathtt{x})+\delta_{\mathrm{smb}}^2}{1-\delta_{\mathcal{V}}^2(\mathtt{x})} \\ &\le 1/2 \\ \text{for } \delta_{\mathrm{smb}}^2 \le 1/8,\ \delta_{\mathcal{V}}^2(\mathtt{x}) &\le 1/4.\end{aligned}$$

□

Now we are ready to collect all the obtained bounds together for the following

Proof of Theorem 2.1. On a random set of probability $\ge 1-12\mathrm{e}^{-\mathtt{x}}$ it holds:

$$\begin{aligned}\alpha &\overset{\text{(Def. (2.3))}}{=} I\!\!P^\circ\left(\sqrt{2\left\{L^\circ(\widetilde{\boldsymbol{\theta}}^\circ) - L^\circ(\widetilde{\boldsymbol{\theta}})\right\}} > \mathfrak{z}_\alpha^\circ\right) \\ &\overset{\text{(Th. } A.4)}{\ge} I\!\!P^\circ\left(\|\boldsymbol{\xi}^\circ(\widetilde{\boldsymbol{\theta}})\| > \mathfrak{z}_\alpha^\circ + \Delta_{\mathrm{w}}^\circ(\mathtt{r}_0,\mathtt{x})\right) \\ &\overset{\text{(L. } A.7)}{\ge} I\!\!P^\circ\left(\|\boldsymbol{\xi}^\circ(\boldsymbol{\theta}^*)\| > \mathfrak{z}_\alpha^\circ + \Delta_{\mathrm{w}}^\circ(\mathtt{r}_0,\mathtt{x}) + \Delta_\xi^\circ(\mathtt{r}_0,\mathtt{x})\right) \qquad \text{(A.24)} \\ &\overset{\text{(Prop. } A.9)}{\ge} I\!\!P\left(\|\boldsymbol{\xi}\| > \mathfrak{z}_\alpha^\circ - \Delta_{\mathrm{w}}(\mathtt{r}_0,\mathtt{x})\right) - \Delta_{\mathrm{full}} \\ &\overset{\text{(Th. } A.2)}{\ge} I\!\!P\left(\sqrt{2\left\{L(\widetilde{\boldsymbol{\theta}}) - L(\boldsymbol{\theta}^*)\right\}} > \mathfrak{z}_\alpha^\circ\right) - \Delta_{\mathrm{full}}, \qquad \text{(A.25)}\end{aligned}$$

where the value Δ_{full} comes from the bound (A.22) with $\delta_z := \Delta_{\mathrm{w}}(\mathtt{r}_0,\mathtt{x}) + \Delta_{\mathrm{w}}^\circ(\mathtt{r}_0,\mathtt{x}) + \Delta_\xi^\circ(\mathtt{r}_0,\mathtt{x})$:

$$\begin{aligned}\Delta_{\mathrm{full}} \overset{\mathrm{def}}{=} 1.55\big(\delta_n^{1/4} + \breve{\delta}_n^{1/4}\big) + \frac{2\sqrt{p}}{3}\left(\delta_{\mathcal{V}}^2(\mathtt{x}) + \delta_{\mathrm{smb}}^2\right) \\ + \left\{\Delta_{\mathrm{w}}(\mathtt{r}_0,\mathtt{x}) + \Delta_{\mathrm{w}}^\circ(\mathtt{r}_0,\mathtt{x}) + \Delta_\xi^\circ(\mathtt{r}_0,\mathtt{x})\right\}/\sqrt{2}\end{aligned} \qquad \text{(A.26)}$$

By the similar arguments in the inverse direction we obtain the following inequality:

$$\left|I\!\!P\left(\sqrt{2\left\{L(\widetilde{\boldsymbol{\theta}}) - L(\boldsymbol{\theta}^*)\right\}} > \mathfrak{z}_\alpha^\circ\right) - \alpha\right| \le \Delta_{\mathrm{full}}.$$

Notice that inequality (A.22) from Proposition A.9, that we use here, requires $\max\{\delta_n^{1/4}, \breve{\delta}_n^{1/4}\} \le 0.11$.

Let us quantify, how the error term Δ_{full} depends on p and n. In the case 4.3 random vectors $\boldsymbol{\xi}_i$ and $\boldsymbol{\xi}_i^{\circ}(\boldsymbol{\theta}^*)$ satisfy the conditions of Theorems A.3 and A.8 correspondingly. Hence $\|\boldsymbol{\xi}_i\|, \|\boldsymbol{\xi}_i^{\circ}(\boldsymbol{\theta}^*)\| \leq \mathtt{C}\sqrt{(p+\mathtt{x})/n}$ and $\delta_n, \breve{\delta}_n \leq \mathtt{C}\sqrt{(p+\mathtt{x})^3/n}$. Finally we have in the case 4.3

$$\Delta_{\text{full}} = \mathtt{C}\left\{\frac{(p+\mathtt{x})^3}{n}\right\}^{1/8} + \mathtt{C}\,\frac{p+\mathtt{x}}{\sqrt{n}}\sqrt{\mathtt{x}} + \mathtt{C}\,\frac{p+\mathtt{x}}{\sqrt{n}}. \tag{A.27}$$

□

REMARK A.2. It is clear from expression (A.27), that the impact of the error term, induced by the Gaussian approximation, is the biggest. The requirement for the ratio $(p+\mathtt{x})^3/n$ to be small is imposed by our Gaussian approximation results (see also Remark 5.2 about the multivariate GAR).

Let us introduce for $p=1$ similarly to (5.6) and (A.21)

$$\delta_{n,\text{B.E.}} \stackrel{\text{def}}{=} \sum_{i=1}^{n} I\!\!E|\boldsymbol{\xi}_i|^3, \qquad \breve{\delta}_{n,\text{B.E.}} \stackrel{\text{def}}{=} \sum_{i=1}^{n} I\!\!E^{\circ}|\boldsymbol{\xi}_i^{\circ}(\boldsymbol{\theta}^*)|^3.$$

Proof of Theorem 2.3. On a random set of probability $\geq 1 - 12\mathrm{e}^{-\mathtt{x}}$ it holds:

$$\begin{aligned}
\alpha \quad &\stackrel{\text{(Def. (2.3))}}{=} \quad I\!\!P^{\circ}\left(\sqrt{2\Big\{L^{\circ}(\widetilde{\boldsymbol{\theta}}^{\circ}) - L^{\circ}(\widetilde{\boldsymbol{\theta}})\Big\}} > \mathfrak{z}_{\alpha}^{\circ}\right) \\
&\stackrel{\text{(Th. } A.4\text{)}}{\geq} \quad I\!\!P^{\circ}\left(\|\boldsymbol{\xi}^{\circ}(\widetilde{\boldsymbol{\theta}})\| > \mathfrak{z}_{\alpha}^{\circ} + \Delta_{\mathrm{W}}^{\circ}(\mathtt{r}_0, \mathtt{x})\right) \\
&\stackrel{\text{(L. } A.7\text{)}}{\geq} \quad I\!\!P^{\circ}\left(\|\boldsymbol{\xi}^{\circ}(\boldsymbol{\theta}^*)\| > \mathfrak{z}_{\alpha}^{\circ} + \Delta_{\mathrm{W}}^{\circ}(\mathtt{r}_0, \mathtt{x}) + \Delta_{\xi}^{\circ}(\mathtt{r}_0, \mathtt{x})\right) \\
&\stackrel{\text{(L. } 5.2\text{, Prop. } A.9\text{)}}{\geq} \quad I\!\!P\left(\|\boldsymbol{\xi}\| > \mathfrak{z}_{\alpha}^{\circ} - \Delta_{\mathrm{W}}(\mathtt{r}_0, \mathtt{x})\right) - \Delta_{\text{B.E., full}} \\
&\stackrel{\text{(Th. } A.2\text{)}}{\geq} \quad I\!\!P\left(\sqrt{2\Big\{L(\widetilde{\boldsymbol{\theta}}) - L(\boldsymbol{\theta}^*)\Big\}} > \mathfrak{z}_{\alpha}^{\circ}\right) - \Delta_{\text{B.E., full}},
\end{aligned}$$

where the value $\Delta_{\text{B.E., full}}$ comes from the bound (5.8) with $\delta_z := \Delta_{\mathrm{W}}(\mathtt{r}_0, \mathtt{x}) + \Delta_{\mathrm{W}}^{\circ}(\mathtt{r}_0, \mathtt{x}) + \Delta_{\xi}^{\circ}(\mathtt{r}_0, \mathtt{x})$, $C_0 \in [0.4097, 0.560]$ and

$$\begin{aligned}
\operatorname{Var}^{\circ}\{\boldsymbol{\xi}^{\circ}(\boldsymbol{\theta}^*)\} &\geq \{1 - \delta_{\mathcal{V}}^2(\mathtt{x})\} I\!\!E \operatorname{Var}^{\circ}\{\boldsymbol{\xi}^{\circ}(\boldsymbol{\theta}^*)\} \\
&\geq \frac{3}{4} D_0^{-1} H_0^2 D_0^{-1} \quad \text{for } \delta_{\mathcal{V}}^2(\mathtt{x}) \leq 1/4
\end{aligned}$$

with probability $\geq 1 - \mathrm{e}^{-\mathtt{x}}$ (due to the bound (A.23)):

$$\begin{aligned}\Delta_{\text{B.E., full}} &\stackrel{\text{def}}{=} 2C_0 \left\{ \frac{\delta_{n,\text{B.E.}}}{(\operatorname{Var}\boldsymbol{\xi})^{3/2}} + \frac{\breve{\delta}_{n,\text{B.E.}}}{(I\!\!E \operatorname{Var}^{\circ}\{\boldsymbol{\xi}^{\circ}(\boldsymbol{\theta}^*)\})^{3/2}} \left(\frac{2}{\sqrt{3}}\right)^3 \right\} \\ &+ \frac{1}{\sqrt{2}} \left\{ \Delta_{\mathrm{w}}(\mathtt{r}_0, \mathtt{x}) + \Delta^{\circ}_{\mathrm{w}}(\mathtt{r}_0, \mathtt{x}) + \Delta^{\circ}_{\xi}(\mathtt{r}_0, \mathtt{x}) \right\} + \frac{2}{3} \left\{ \delta^2_{\mathcal{V}}(\mathtt{x}) + \delta^2_{\xi} \right\} \\ &\leq \mathtt{C} \frac{1+\mathtt{x}}{\sqrt{n}} \quad \text{in the case 4.3.}\end{aligned} \tag{A.28}$$

The similar inequalities in the inverse direction finish the proof with the error term □

A.4 Proof of Theorem 2.4 (large modeling bias)

Lemma A.10 (Lower bound for deviations of a Gaussian quadratic form)**.** *Let $\boldsymbol{\phi} \sim \mathcal{N}(0, \boldsymbol{I}_p)$ and Σ is any symmetric non-negative definite matrix, then it holds for any $\mathtt{x} > 0$*

$$I\!\!P\left(\operatorname{tr}\Sigma - \|\Sigma^{1/2}\boldsymbol{\phi}\|^2 \geq 2\sqrt{\mathtt{x} \operatorname{tr}(\Sigma^2)} \right) \leq \exp(-\mathtt{x}).$$

Proof of Lemma A.10. It is sufficient to consider w.l.o.g. only the case of diagonal matrix Σ, since it can be represented as $\Sigma = U^\top \operatorname{diag}\{a_1, \ldots, a_p\} U$ for an orthogonal matrix U and the eigenvalues $a_1 \geq \cdots \geq a_p$; $U\boldsymbol{\phi} \sim \mathcal{N}(0, \boldsymbol{I}_p)$.

By the exponential Chebyshev inequality it holds for $\mu > 0$, $\Delta > 0$

$$I\!\!P\left(\operatorname{tr}\Sigma - \|\Sigma^{1/2}\boldsymbol{\phi}\|^2 \geq \Delta \right) \leq \exp(-\mu\Delta/2) I\!\!E \exp\left(\mu \left\{ \operatorname{tr}\Sigma - \|\Sigma^{1/2}\boldsymbol{\phi}\|^2 \right\} /2 \right).$$

$$\log I\!\!E \exp\left(\mu \left\{ \operatorname{tr}\Sigma - \|\Sigma^{1/2}\boldsymbol{\phi}\|^2 \right\} /2 \right) \leq \frac{1}{2} \sum_{j=1}^{p} \left\{ \mu a_j - \log(1 + a_j \mu) \right\},$$

therefore

$$\begin{aligned} I\!\!P\left(\operatorname{tr}\Sigma - \|\Sigma^{1/2}\boldsymbol{\phi}\|^2 \geq \Delta \right) &\leq \exp\left(-\frac{1}{2} \left[\mu\Delta + \sum_{j=1}^{p} \left\{ \log(1 + a_j\mu) - \mu a_j \right\} \right] \right) \\ &\leq \exp\left(-\frac{1}{2} \left[\mu\Delta - \mu^2 \sum_{j=1}^{p} a_j^2/2 \right] \right) \\ &\leq \exp\left(-\Delta^2 / \left\{ 4 \sum_{j=1}^{p} a_j^2 \right\} \right). \end{aligned}$$

If $\mathtt{x} := \Delta^2 / \left\{ 4 \sum_{i=1}^{p} a_j^2 \right\}$, then $\Delta = 2\sqrt{\mathtt{x} \sum_{j=1}^{p} a_j^2}$. □

Proof of Theorem 2.4. Due to the bound (A.24) it holds for $\mathfrak{z} \geq \max\{2, \sqrt{p}\} + \mathtt{C}(p+\mathtt{x})/\sqrt{n}$ with probability $\geq 1 - 5\mathrm{e}^{-\mathtt{x}}$

$$\begin{aligned} &I\!\!P^{\circ}\left(\sqrt{2\Big\{L^{\circ}(\widetilde{\boldsymbol{\theta}}^{\circ}) - L^{\circ}(\widetilde{\boldsymbol{\theta}})\Big\}} > \mathfrak{z}\right) \\ &\quad \geq I\!\!P^{\circ}\left(\|\boldsymbol{\xi}^{\circ}(\boldsymbol{\theta}^*)\| > \mathfrak{z} + \Delta^{\circ}_{\mathrm{w}}(\mathtt{r}_0, \mathtt{x}) + \Delta^{\circ}_{\xi}(\mathtt{r}_0, \mathtt{x})\right). \end{aligned}$$

Let us introduce the random vector $\boldsymbol{\xi}_0 \stackrel{\mathrm{def}}{=} (D_0^{-1}H_0^2D_0^{-1})^{1/2}(\operatorname{Var}\boldsymbol{\xi})^{-1/2}\boldsymbol{\xi}$. The bound (A.23) implies with probability $\geq 1 - \mathrm{e}^{-\mathtt{x}}$

$$\operatorname{tr}\left\{\left((\operatorname{Var}\boldsymbol{\xi}_0)^{-1/2}\operatorname{Var}^{\circ}\{\boldsymbol{\xi}^{\circ}(\boldsymbol{\theta}^*)\}(\operatorname{Var}\boldsymbol{\xi}_0)^{-1/2} - \boldsymbol{I}_p\right)^2\right\} \leq p\delta^4_{\mathcal{V}}(\mathtt{x}). \tag{A.29}$$

Applying statement *2.2* of Theorem 5.1 to the vectors $\boldsymbol{\xi}^{\circ}(\boldsymbol{\theta}^*)$ and $\boldsymbol{\xi}_0$, we have with probability $\geq 1 - \mathrm{e}^{-\mathtt{x}}$

$$\begin{aligned} &I\!\!P^{\circ}\left(\|\boldsymbol{\xi}^{\circ}(\boldsymbol{\theta}^*)\| > \mathfrak{z} + \Delta^{\circ}_{\mathrm{w}}(\mathtt{r}_0, \mathtt{x}) + \Delta^{\circ}_{\xi}(\mathtt{r}_0, \mathtt{x})\right) \\ &\quad \geq I\!\!P\left(\|\boldsymbol{\xi}_0\| > \mathfrak{z} - \Delta_{\mathrm{w}}(\mathtt{r}_0, \mathtt{x})\right) - \Delta_{\mathrm{b,\,full}} \end{aligned}$$

where

$$\begin{aligned} \Delta_{\mathrm{b,\,full}} \stackrel{\mathrm{def}}{=}\ & 1.55\left(\delta_n^{1/4} + \breve{\delta}_n^{1/4}\right) + \frac{\sqrt{p}}{2}\delta^2_{\mathcal{V}}(\mathtt{x}) \\ &+ \frac{\Delta_{\mathrm{w}}(\mathtt{r}_0, \mathtt{x}) + \Delta^{\circ}_{\mathrm{w}}(\mathtt{r}_0, \mathtt{x}) + \Delta^{\circ}_{\xi}(\mathtt{r}_0, \mathtt{x})}{\sqrt{2}}. \end{aligned} \tag{A.30}$$

By the definition of $\boldsymbol{\xi}_0$ it holds $\|\boldsymbol{\xi}_0\| \geq \|\boldsymbol{\xi}\|\|(\operatorname{Var}\boldsymbol{\xi})^{1/2}(D_0^{-1}H_0^2D_0^{-1})^{-1/2}\|^{-1}$. Consider the following matrix

$$\begin{aligned} \widetilde{V}^2 &\stackrel{\mathrm{def}}{=} (D_0^{-1}H_0^2D_0^{-1})^{-1/2}(\operatorname{Var}\boldsymbol{\xi})(D_0^{-1}H_0^2D_0^{-1})^{-1/2} \\ &= (D_0^{-1}H_0^2D_0^{-1})^{1/2}\left(D_0H_0^{-2}V_0^2H_0^{-2}D_0\right)(D_0^{-1}H_0^2D_0^{-1})^{1/2} \\ &\leq (D_0^{-1}H_0^2D_0^{-1})^{1/2}\left(D_0H_0^{-2}D_0\right)(D_0^{-1}H_0^2D_0^{-1})^{1/2} \\ &= \boldsymbol{I}_p, \end{aligned} \tag{A.31}$$

here $V_0^2 \stackrel{\mathrm{def}}{=} \operatorname{Var}\{\nabla_{\boldsymbol{\theta}}L(\boldsymbol{\theta}^*)\}$; the inequality (A.31) holds due to the definitions (1.7), (1.8) and $V_0^2 = H_0^2 - B_0^2 > 0$. Therefore $\|\widetilde{V}^2\| \leq 1$ and $\|\boldsymbol{\xi}_0\| \geq \|\boldsymbol{\xi}\|$. By (A.25)

$$\begin{aligned} &I\!\!P^{\circ}\left(\sqrt{2\Big\{L^{\circ}(\widetilde{\boldsymbol{\theta}}^{\circ}) - L^{\circ}(\widetilde{\boldsymbol{\theta}})\Big\}} > \mathfrak{z}\right) \\ &\quad \geq I\!\!P\left(\|\boldsymbol{\xi}\| > \mathfrak{z} - \Delta_{\mathrm{w}}(\mathtt{r}_0, \mathtt{x})\right) - \Delta_{\mathrm{b,\,full}} \\ &\quad \geq I\!\!P\left(\sqrt{2\Big\{L(\widetilde{\boldsymbol{\theta}}) - L(\boldsymbol{\theta}^*)\Big\}} > \mathfrak{z}\right) - \Delta_{\mathrm{b,\,full}} \end{aligned}$$

with probability $\geq 1 - 12\mathrm{e}^{-\mathtt{x}}$, which finishes the proof of the first part. For the second part let us introduce $\overline{\boldsymbol{\xi}}_0 \sim \mathcal{N}(0, D_0^{-1} H_0^2 D_0^{-1})$ s.t. $\operatorname{Var} \overline{\boldsymbol{\xi}}_0 = \operatorname{Var} \boldsymbol{\xi}_0$. Applying statement *2.1* of Theorem 5.1 to the vectors $\boldsymbol{\xi}^\circ(\boldsymbol{\theta}^*)$ and $\overline{\boldsymbol{\xi}}_0$, using the bound (A.29), we have with probability $\geq 1 - \mathrm{e}^{-\mathtt{x}}$

$$\begin{aligned} I\!P^\circ \left(\|\boldsymbol{\xi}^\circ(\boldsymbol{\theta}^*)\| > \mathfrak{z} + \Delta^\circ_{\mathrm{W}}(\mathtt{r}_0, \mathtt{x}) + \Delta^\circ_{\xi}(\mathtt{r}_0, \mathtt{x}) \right) \\ \geq I\!P \left(\|\overline{\boldsymbol{\xi}}_0\| > \mathfrak{z} \right) - \Delta_{\mathrm{G},1}, \end{aligned}$$

where

$$\Delta_{\mathrm{G},1} \stackrel{\mathrm{def}}{=} 1.55 \breve{\delta}_n^{1/4} + \frac{\Delta^\circ_{\mathrm{W}}(\mathtt{r}_0, \mathtt{x}) + \Delta^\circ_{\xi}(\mathtt{r}_0, \mathtt{x})}{\sqrt{2}} + \frac{\sqrt{p}}{2} \delta^2_{\mathcal{V}}(\mathtt{x}).$$

By definition (A.20) $\overline{\boldsymbol{\xi}} \sim \mathcal{N}(0, \operatorname{Var} \boldsymbol{\xi})$. Lemma A.10 and Theorem 1.2 by Spokoiny (2012b) imply

$$\begin{aligned} I\!P\Big(\|\overline{\boldsymbol{\xi}}\| - \|\overline{\boldsymbol{\xi}}_0\| \geq \sqrt{\operatorname{tr}(\operatorname{Var} \boldsymbol{\xi})} - \sqrt{\operatorname{tr}(\operatorname{Var} \overline{\boldsymbol{\xi}}_0)} + \Delta_{\mathtt{qf},1} \Big) \leq 2\mathrm{e}^{-\mathtt{x}}, \\ I\!P\Big(\|\overline{\boldsymbol{\xi}}\| - \|\overline{\boldsymbol{\xi}}_0\| \leq \sqrt{\operatorname{tr}(\operatorname{Var} \boldsymbol{\xi})} - \sqrt{\operatorname{tr}(\operatorname{Var} \overline{\boldsymbol{\xi}}_0)} - \Delta_{\mathtt{qf},2} \Big) \leq 2\mathrm{e}^{-\mathtt{x}}, \end{aligned} \tag{A.32}$$

where

$$\begin{aligned} \Delta_{\mathtt{qf},1} &\stackrel{\mathrm{def}}{=} \left[4\mathtt{x} \operatorname{tr}\{(\operatorname{Var} \overline{\boldsymbol{\xi}}_0)^2\} \right]^{1/4} \\ &\quad + \max \left[2\sqrt{2\mathtt{x} \operatorname{tr}\{(\operatorname{Var} \boldsymbol{\xi})^2\}}, 6\mathtt{x} \| \operatorname{Var} \boldsymbol{\xi} \| \right]^{1/2}, \\ \Delta_{\mathtt{qf},2} &\stackrel{\mathrm{def}}{=} \left[4\mathtt{x} \operatorname{tr}\{(\operatorname{Var} \boldsymbol{\xi})^2\} \right]^{1/4} \\ &\quad + \max \left[2\sqrt{2\mathtt{x} \operatorname{tr}\{(\operatorname{Var} \overline{\boldsymbol{\xi}}_0)^2\}}, 6\mathtt{x} \| \operatorname{Var} \overline{\boldsymbol{\xi}}_0 \| \right]^{1/2}. \end{aligned} \tag{A.33}$$

By conditions $(\mathcal{I})$, $(\mathcal{I}_B)$

$$\begin{aligned} \Delta_{\mathtt{qf},1} &\leq \left\{ \sqrt{4\mathtt{x}p} (\mathfrak{a}^2 + \mathfrak{a}_B^2) \right\}^{1/2} + \mathfrak{a} \max \left\{ \sqrt{8\mathtt{x}p}, 6\mathtt{x} \right\}^{1/2}, \\ \Delta_{\mathtt{qf},2} &\leq \left\{ 4\mathtt{x}p\mathfrak{a}^4 \right\}^{1/4} + \sqrt{\mathfrak{a}^2 + \mathfrak{a}_B^2} \max \left\{ \sqrt{8\mathtt{x}p}, 6\mathtt{x} \right\}^{1/2}. \end{aligned} \tag{A.34}$$

Further, it holds on a random set with probability $\geq 1 - 2\mathrm{e}^{-\mathtt{x}}$

$$
\begin{aligned}
& I\!P\left(\|\overline{\boldsymbol{\xi}}_0\| > \mathfrak{z}\right) - \Delta_{\mathrm{G},1} \\
= \quad & I\!P\left(\|\overline{\boldsymbol{\xi}}\| > \mathfrak{z} + \|\overline{\boldsymbol{\xi}}\| - \|\overline{\boldsymbol{\xi}}_0\|\right) - \Delta_{\mathrm{G},1} \\
\overset{(\text{by (A.32)})}{\geq} \quad & I\!P\left(\|\overline{\boldsymbol{\xi}}\| > \mathfrak{z} + \sqrt{\operatorname{tr}(\operatorname{Var}\boldsymbol{\xi})} - \sqrt{\operatorname{tr}(\operatorname{Var}\overline{\boldsymbol{\xi}}_0)} + \Delta_{\mathtt{qf},1}\right) - \Delta_{\mathrm{G},1} \\
\overset{(\text{Th. } 5.1)}{\geq} \quad & I\!P\left(\|\boldsymbol{\xi}\| > \mathfrak{z} - \Delta_{\mathrm{w}}(\mathtt{r}_0, \mathtt{x}) + \sqrt{\operatorname{tr}(\operatorname{Var}\boldsymbol{\xi})} - \sqrt{\operatorname{tr}(\operatorname{Var}\overline{\boldsymbol{\xi}}_0)} + \Delta_{\mathtt{qf},1}\right) \\
& - \Delta_{\mathrm{G},1} - \Delta_{\mathrm{G},2} \\
\overset{(\text{Th. } A.2)}{\geq} \quad & I\!P\left(\sqrt{2\Big\{L(\widetilde{\boldsymbol{\theta}}) - L(\boldsymbol{\theta}^*)\Big\}} > \mathfrak{z} + \sqrt{\operatorname{tr}(\operatorname{Var}\boldsymbol{\xi})} - \sqrt{\operatorname{tr}(\operatorname{Var}\overline{\boldsymbol{\xi}}_0)} + \Delta_{\mathtt{qf},1}\right) \\
& - \Delta_{\mathrm{b,\,full}},
\end{aligned}
$$

where

$$
\begin{aligned}
\Delta_{\mathrm{G},2} & \overset{\mathrm{def}}{=} 1.55\delta_n^{1/4} + \frac{\Delta_{\mathrm{w}}(\mathtt{r}_0, \mathtt{x})}{\sqrt{2}}, \\
\Delta_{\mathrm{b,\,full}} & = \Delta_{\mathrm{G},1} + \Delta_{\mathrm{G},2}.
\end{aligned}
$$

Hence, we obtain

$$
\begin{aligned}
& I\!P^\circ\left(\sqrt{2\Big\{L^\circ(\widetilde{\boldsymbol{\theta}}^\circ) - L^\circ(\widetilde{\boldsymbol{\theta}})\Big\}} > \mathfrak{z}\right) \\
& \geq I\!P\left(\sqrt{2\Big\{L(\widetilde{\boldsymbol{\theta}}) - L(\boldsymbol{\theta}^*)\Big\}} > \mathfrak{z} + \sqrt{\operatorname{tr}(\operatorname{Var}\boldsymbol{\xi})} - \sqrt{\operatorname{tr}(\operatorname{Var}\overline{\boldsymbol{\xi}}_0)} + \Delta_{\mathtt{qf},1}\right) \\
& \quad - \Delta_{\mathrm{b,\,full}}.
\end{aligned}
$$

By definition (2.2) of $(1-\alpha)$-quantile $\mathfrak{z}_\alpha$ it holds:

$$
\mathfrak{z}_{(\alpha + \Delta_{\mathrm{b,\,full}})} \leq \mathfrak{z}_\alpha^\circ + \sqrt{\operatorname{tr}(\operatorname{Var}\boldsymbol{\xi})} - \sqrt{\operatorname{tr}(\operatorname{Var}\overline{\boldsymbol{\xi}}_0)} + \Delta_{\mathtt{qf},1},
$$

and in addition

$$
\sqrt{\operatorname{tr}(\operatorname{Var}\boldsymbol{\xi})} - \sqrt{\operatorname{tr}(\operatorname{Var}\overline{\boldsymbol{\xi}}_0)} \leq -\frac{\operatorname{tr}(D_0^{-1}B_0^2D_0^{-1})}{2\sqrt{\operatorname{tr}(D_0^{-1}H_0^2D_0^{-1})}} \leq 0.
$$

The inverse inequalities are implied with the similar arguments:

$$\begin{aligned}
&I\!P^{\circ}\left(\sqrt{2\Big\{L^{\circ}(\widetilde{\boldsymbol{\theta}}^{\circ})-L^{\circ}(\widetilde{\boldsymbol{\theta}})\Big\}}>\mathfrak{z}\right)\\
&\quad\le I\!P\left(\sqrt{2\Big\{L(\widetilde{\boldsymbol{\theta}})-L(\boldsymbol{\theta}^{*})\Big\}}>\mathfrak{z}+\sqrt{\operatorname{tr}(\operatorname{Var}\boldsymbol{\xi})}-\sqrt{\operatorname{tr}(\operatorname{Var}\overline{\boldsymbol{\xi}}_{0})}-\Delta_{\mathtt{qf},2}\right)\\
&\qquad+\Delta_{\mathrm{b,\,full}}.
\end{aligned}$$

And

$$\mathfrak{z}_{(\alpha-\Delta_{\mathrm{b,\,full}})}\ge\mathfrak{z}^{\circ}_{\alpha}+\sqrt{\operatorname{tr}(\operatorname{Var}\boldsymbol{\xi})}-\sqrt{\operatorname{tr}(\operatorname{Var}\overline{\boldsymbol{\xi}}_{0})}-\Delta_{\mathtt{qf},2}.$$

□

A.5 Proof of Theorem 2.5 (the smoothed version)

Lemma A.11. *For the function* $g_{\Delta}(x,z)$ *defined in* (2.6)*, all* $\Delta_1\in[0,x]$ *and all* $C\ge 1$ *it holds*

$$g_{\Delta}(x-\Delta_1,z)\ge g_{\Delta}(x,z+\Delta_1 C)$$

Proof of Lemma A.11. By definition (5.9) of $g(x)$

$$\begin{aligned}
\max_{x\ge 0}\{g_{\Delta}(x-\Delta_1,z)=0\}&=z+\Delta_1,\\
\max_{x\ge 0}\{g_{\Delta}(x,z+\Delta_1 C)=0\}&=z+\Delta_1 C.
\end{aligned}$$

For $x\ge z+\Delta_1 C$ it holds

$$\begin{aligned}
g_{\Delta}(x-\Delta_1,z)&=g\left(\frac{1}{2\Delta z}\left\{(x-\Delta_1)^2-z^2\right\}\right)\\
&\ge g\left(\frac{1}{2\Delta(z+\Delta_1 C)}\left\{x^2-(z+\Delta_1 C)^2\right\}\right)\qquad\text{(A.35)}\\
&=g_{\Delta}(x,z+\Delta_1 C).
\end{aligned}$$

Indeed, the comparison in (A.35) reads as

$$\begin{aligned}
&(z+\Delta_1 C)(x-\Delta_1+z)(x-\Delta_1-z)\qquad\text{(A.36)}\\
&\qquad\vee\ z(x+z+\Delta_1 C)(x-z-\Delta_1 C).
\end{aligned}$$

Since $C\ge 1$, $(x-\Delta_1-z)\ge(x-\Delta_1 C-z)$ and it holds for the left side of (A.36):

$$\begin{aligned}
(z+\Delta_1 C)(x-\Delta_1+z)&=(zx+z^2+2\Delta_1 C)+\Delta_1(xC-\Delta_1 C-z)\\
&\ge(zx+z^2+2\Delta_1 C),
\end{aligned}$$

which is equal to the multiplier $z(x + \Delta_1 C + z)$ in right side. □

Proposition A.12 (Smooth analog of Proposition A.9)**.** *If conditions* $(\mathbf{SmB})$ *and* $(\boldsymbol{SD_1})$ *are fulfilled, then it holds for all* $0 < \Delta \le 0.22$ *and for all* $z, \overline{z} > 2$ *s.t.* $|z - \overline{z}| \le \delta_z$ *for some* $\delta_z \in [0, 1]$ *with probability* $\ge 1 - \mathrm{e}^{-\mathtt{x}}$ *:*

$$\begin{aligned}
&\Bigl| I\!\!E g_\Delta \left(\|\boldsymbol{\xi}\|, z \right) - I\!\!E^\circ g_\Delta \left(\|\boldsymbol{\xi}^\circ(\boldsymbol{\theta}^*)\|, \overline{z} \right) \Bigr| \\
&\qquad \le \frac{16}{\Delta^3} \left(\delta_n + \breve{\delta}_n \right) + 2\sqrt{p} \frac{\delta_z}{z} + \sqrt{p} \frac{\delta_z^2}{z^2} + \sqrt{p} \frac{\delta^2_{\mathcal{V}}(\mathtt{x}) + \delta^2_{\mathrm{smb}}}{1 - \delta^2_{\mathcal{V}}(\mathtt{x})} \\
&\qquad \le \frac{16}{\Delta^3} \left(\delta_n + \breve{\delta}_n \right) + \sqrt{5} \delta_z + \frac{4\sqrt{p}}{3} \bigl\{ \delta^2_{\mathcal{V}}(\mathtt{x}) + \delta^2_{\mathrm{smb}} \bigr\} \qquad \text{(A.37)} \\
&\qquad \text{for } \overline{z} \ge \sqrt{p},\ \delta^2_{\mathcal{V}}(\mathtt{x}) \le 1/4.
\end{aligned}$$

Proof of Proposition A.12. The conditions of Theorem 5.5 are fulfilled with the value $\delta_\Sigma = \sqrt{p} \bigl\{ \delta^2_{\mathcal{V}}(\mathtt{x}) + \delta^2_{\mathrm{smb}} \bigr\} / \bigl\{ 1 - \delta^2_{\mathcal{V}}(\mathtt{x}) \bigr\}$ due to the proof of Proposition A.9. □

Proof of Theorem 2.5. The following holds on a random set of probability $\ge 1 - 12\mathrm{e}^{-\mathtt{x}}$:

$$\begin{aligned}
& I\!\!E^\circ g_\Delta \left(\sqrt{2 \Bigl\{ L^\circ(\widetilde{\boldsymbol{\theta}}^\circ) - L^\circ(\widetilde{\boldsymbol{\theta}}) \Bigr\}}, \mathfrak{z} \right) \\
&\overset{\text{(Th. } A.4\text{)}}{\ge} I\!\!E^\circ g_\Delta \left(\|\boldsymbol{\xi}^\circ(\widetilde{\boldsymbol{\theta}})\| - \Delta^\circ_{\mathrm{w}}(\mathtt{r}_0, \mathtt{x}), \mathfrak{z} \right) \\
&\overset{\text{(L. } A.7\text{)}}{\ge} I\!\!E^\circ g_\Delta \left(\|\boldsymbol{\xi}^\circ(\boldsymbol{\theta}^*)\| - \Delta^\circ_{\mathrm{w}}(\mathtt{r}_0, \mathtt{x}) - \Delta^\circ_{\xi}(\mathtt{r}_0, \mathtt{x}), \mathfrak{z} \right) \\
&\overset{\text{(L. } A.11\text{)}}{\ge} I\!\!E^\circ g_\Delta \left(\|\boldsymbol{\xi}^\circ(\boldsymbol{\theta}^*)\|, \mathfrak{z} + \Delta^\circ_{\mathrm{w}}(\mathtt{r}_0, \mathtt{x}) + \Delta^\circ_{\xi}(\mathtt{r}_0, \mathtt{x}) \right) \\
&\overset{\text{(Prop. } A.12\text{)}}{\ge} I\!\!E g_\Delta \left(\|\boldsymbol{\xi}\|, \mathfrak{z} - \Delta_{\mathrm{w}}(\mathtt{r}_0, \mathtt{x}) \right) - \Delta_{\mathrm{sm}} \\
&\overset{\text{(Th. } A.2\text{, L. } A.11\text{)}}{\ge} I\!\!E g_\Delta \left(\sqrt{2 \Bigl\{ L(\widetilde{\boldsymbol{\theta}}) - L(\boldsymbol{\theta}^*) \Bigr\}}, \mathfrak{z} \right) - \Delta_{\mathrm{sm}},
\end{aligned}$$

where the term Δ_{sm} comes from (A.37) with $\delta_z := \Delta_{\mathrm{w}}(\mathtt{r}_0, \mathtt{x}) + \Delta^\circ_{\mathrm{w}}(\mathtt{r}_0, \mathtt{x}) + \Delta^\circ_{\xi}(\mathtt{r}_0, \mathtt{x})$:

$$\begin{aligned}
\Delta_{\mathrm{sm}} \overset{\mathrm{def}}{=} & \frac{16}{\Delta^3} \left(\delta_n + \breve{\delta}_n \right) + \frac{4\sqrt{p}}{3} \bigl\{ \delta^2_{\mathcal{V}}(\mathtt{x}) + \delta^2_{\mathrm{smb}} \bigr\} \\
& + \sqrt{5} \bigl\{ \Delta_{\mathrm{w}}(\mathtt{r}_0, \mathtt{x}) + \Delta^\circ_{\mathrm{w}}(\mathtt{r}_0, \mathtt{x}) + \Delta^\circ_{\xi}(\mathtt{r}_0, \mathtt{x}) \bigr\}.
\end{aligned} \qquad \text{(A.38)}$$

By the similar inequalities in the inverse direction we get the statement proved. Due to the arguments in the end of the proof of Theorem 2.1 it holds in the case 4.3

$$\Delta_{\mathrm{sm}} = \mathtt{C} \frac{1}{\Delta^3} \left\{ \frac{(p + \mathtt{x})^3}{n} \right\}^{1/2} + \mathtt{C} \frac{p + \mathtt{x}}{\sqrt{n}} \sqrt{\mathtt{x}} + \mathtt{C} \frac{p + \mathtt{x}}{\sqrt{n}}. \qquad \text{(A.39)}$$

□

A.6 Bernstein matrix inequality

Consider the following symmetric $p \times p$ $I\!P$-random matrix and its expected value:

$$\mathcal{V}^2(\boldsymbol{\theta}^*) \stackrel{\text{def}}{=} \operatorname{Var}^{\circ}(\nabla_{\boldsymbol{\theta}} L^{\circ}(\boldsymbol{\theta}^*)) = \sum_{i=1}^{n} \nabla_{\boldsymbol{\theta}} \ell_i(\boldsymbol{\theta}^*) \nabla_{\boldsymbol{\theta}} \ell_i(\boldsymbol{\theta}^*)^{\top},$$

$$H_0^2 \stackrel{\text{def}}{=} I\!E \mathcal{V}^2(\boldsymbol{\theta}^*) = \sum_{i=1}^{n} I\!E \left[\nabla_{\boldsymbol{\theta}} \ell_i(\boldsymbol{\theta}^*) \nabla_{\boldsymbol{\theta}} \ell_i(\boldsymbol{\theta}^*)^{\top} \right].$$

Matrix $\mathcal{V}^2(\boldsymbol{\theta}^*)$ equals to a sum of the independent random matrices $\nabla_{\boldsymbol{\theta}} \ell_i(\boldsymbol{\theta}^*) \nabla_{\boldsymbol{\theta}} \ell_i(\boldsymbol{\theta}^*)^{\top}$. Assuming the condition $(\boldsymbol{SD_1})$ to be fulfilled, we can refer to the result by Tropp (2012) in order to get the concentration bound below. Let us previously introduce some notations.

$$v_i^2(\boldsymbol{\theta}) \stackrel{\text{def}}{=} H_0^{-1} \left\{ \nabla_{\boldsymbol{\theta}} \ell_i(\boldsymbol{\theta}) \nabla_{\boldsymbol{\theta}} \ell_i(\boldsymbol{\theta})^{\top} - I\!E \left[\nabla_{\boldsymbol{\theta}} \ell_i(\boldsymbol{\theta}) \nabla_{\boldsymbol{\theta}} \ell_i(\boldsymbol{\theta})^{\top} \right] \right\} H_0^{-1},$$

then

$$H_0^{-1} \mathcal{V}^2(\boldsymbol{\theta}^*) H_0^{-1} = \sum_{i=1}^{n} v_i^2(\boldsymbol{\theta}^*).$$

Define also

$$\varkappa_v^2 \stackrel{\text{def}}{=} \Big\| \sum_{i=1}^{n} I\!E v_i^4(\boldsymbol{\theta}^*) \Big\|.$$

Theorem A.13 (Bernstein inequality for $\mathcal{V}^2(\boldsymbol{\theta}^*)$)**.** *Let the condition* $(\boldsymbol{SD_1})$ *be fulfilled, then it holds with probability* $\geq 1 - \mathrm{e}^{-\mathtt{x}}$ *:*

$$\| H_0^{-1} \mathcal{V}^2(\boldsymbol{\theta}^*) H_0^{-1} - \boldsymbol{I}_p \| \leq \delta_{\mathcal{V}}^2(\mathtt{x}),$$

where the error term is defined as

$$\delta_{\mathcal{V}}^2(\mathtt{x}) \stackrel{\text{def}}{=} \sqrt{2 \varkappa_v^2 \{ \log(p) + \mathtt{x} \}} + \frac{2}{3} \delta_v^2 \{ \log(p) + \mathtt{x} \}$$

and is proportional to $\sqrt{\{\log(p) + \mathtt{x}\}/n}$ *in the case 4.3.*

Proof. Due to Theorem 1.4 by Tropp (2012):

$$I\!P\big(\| H_0^{-1} \mathcal{V}^2(\boldsymbol{\theta}^*) H_0^{-1} - \boldsymbol{I}_p \| \geq t \big) \leq p \exp\left(\frac{-t^2}{2\varkappa_v^2 + 2\delta_v^2 t/3} \right).$$

For

$$\mathtt{x} = \frac{t^2}{2\varkappa_v^2 + 2\delta_v^2 t/3} - \log(p)$$

it holds:

$$I\!P\left(\|H_0^{-1}\mathcal{V}^2(\boldsymbol{\theta}^*)H_0^{-1} - \boldsymbol{I}_p\| \geq \delta_{\mathcal{V}}^2(\mathtt{x})\right) \leq \mathrm{e}^{-\mathtt{x}}.$$

□

References

Aerts, M. and Claeskens, G. (2001). Bootstrap tests for misspecified models, with application to clustered binary data. *Computational statistics & data analysis*, 36(3):383–401.

Arlot, S., Blanchard, G., and Roquain, E. (2010). Some nonasymptotic results on resampling in high dimension. I. Confidence regions. *Ann. Statist.*, 38(1):51–82.

Barbe, P. and Bertail, P. (1995). *The weighted bootstrap*, volume 98. Springer.

Bentkus, V. (2003). On the dependence of the Berry–Esseen bound on dimension. *Journal of Statistical Planning and Inference*, 113(2):385–402.

Berry, A. C. (1941). The accuracy of the Gaussian approximation to the sum of independent variates. *Transactions of the american mathematical society*, 49(1):122–136.

Bhattacharya, R. and Holmes, S. (2010). An exposition of Götze's estimation of the rate of convergence in the multivariate central limit theorem. *arXiv preprint arXiv:1003.4254*.

Chatterjee, S. and Bose, A. (2005). Generalized bootstrap for estimating equations. *The Annals of Statistics*, 33(1):414–436.

Chen, L. H. and Fang, X. (2011). Multivariate normal approximation by Stein's method: The concentration inequality approach. *arXiv preprint arXiv:1111.4073*.

Chen, X. and Pouzo, D. (2009). Efficient estimation of semiparametric conditional moment models with possibly nonsmooth residuals. *Journal of Econometrics*, 152(1):46–60.

Chen, X. and Pouzo, D. (2014). Sieve Wald and QLR Inferences on semi/nonparametric conditional moment models.

Chernozhukov, V., Chetverikov, D., and Kato, K. (2013). Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors. *The Annals of Statistics*, 41(6):2786–2819.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1):1–26.

Esseen, C.-G. (1942). *On the Liapounoff limit of error in the theory of probability*. Almqvist & Wiksell.

Esseen, C. G. (1956). A moment inequality with an application to the central limit theorem. *Scandinavian Actuarial Journal*, 1956(2):160–170.

Götze, F. (1991). On the rate of convergence in the multivariate CLT. *The Annals of Probability*, pages 724–739.

Götze, F. and Zaitsev, A. Y. (2014). Explicit rates of approximation in the CLT for quadratic forms. *The Annals of Probability*, 42(1):354–397.

Hall, A. R. (2005). *Generalized method of moments*. Oxford University Press Oxford.

Hall, P. (1992). *The bootstrap and Edgeworth expansion*. Springer.

Horowitz, J. L. (2001). The bootstrap. *Handbook of econometrics*, 5:3159–3228.

Huber, P. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. Proc. 5th Berkeley Symp. Math. Stat. Probab., Univ. Calif. 1965/66, 1, 221-233 (1967).

Janssen, A. and Pauls, T. (2003). How do bootstrap and permutation tests work? *Annals of statistics*, pages 768–806.

Kline, P. and Santos, A. (2012). Higher order properties of the wild bootstrap under misspecification. *Journal of Econometrics*, 171(1):54–70.

Lavergne, P. and Patilea, V. (2013). Smooth minimum distance estimation and testing with conditional estimating equations: uniform in bandwidth theory. *Journal of Econometrics*, 177(1):47–59.

Lindeberg, J. W. (1922). Eine neue Herleitung des Exponentialgesetzes in der Wahrscheinlichkeitsrechnung. *Mathematische Zeitschrift*, 15(1):211–225.

Ma, S. and Kosorok, M. R. (2005). Robust semiparametric M-estimation and the weighted bootstrap. *Journal of Multivariate Analysis*, 96(1):190–217.

Mammen, E. (1992). *When does bootstrap work?*, volume 77. Springer.

Mammen, E. (1993). Bootstrap and wild bootstrap for high dimensional linear models. *The Annals of Statistics*, pages 255–285.

Newton, M. A. and Raftery, A. E. (1994). Approximate bayesian inference with the weighted likelihood bootstrap. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 3–48.

Prokhorov, Y. V. and Ulyanov, V. V. (2013). Some approximation problems in statistics and probability. In *Limit Theorems in Probability, Statistics and Number Theory*, pages 235–249. Springer.

Shevtsova, I. (2010). An improvement of convergence rate estimates in the Lyapunov theorem. In *Doklady Mathematics*, volume 82, pages 862–864. Springer.

Spokoiny, V. (2012a). Parametric estimation. Finite sample theory. *The Annals of Statistics*, 40(6):2877–2909.

Spokoiny, V. (2012b). Supplement to “Parametric estimation. Finite sample theory”.

Spokoiny, V. (2013). Bernstein-von Mises Theorem for growing parameter dimension. *arXiv preprint arXiv:1302.3430*.

Tropp, J. A. (2012). User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12(4):389–434.

Tsybakov, A. B. (2009). *Introduction to Nonparametric Estimation*. Springer, New York.

van ver Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical processes*. Springer, New York.

Wilks, S. S. (1938). The large-sample distribution of the likelihood ratio for testing composite hypotheses. *The Annals of Mathematical Statistics*, 9(1):60–62.

Wu, C. F. J. (1986). Jackknife, bootstrap and other resampling methods in regression analysis. *The Annals of Statistics*, 14(4):1261–1295+.