Submitted to the Annals of Statistics

PARAMETRIC ESTIMATION. FINITE SAMPLE THEORY*

By Vladimir Spokoiny

Weierstrass-Institute, Humboldt University Berlin, and Moscow Institute of Physics and Technology

The paper aims at reconsidering the famous Le Cam LAN theory. The main features of the approach which make it different from the classical one are: (1) the study is non-asymptotic, that is, the sample size is fixed and does not tend to infinity; (2) the parametric assumption is possibly misspecified and the underlying data distribution can lie beyond the given parametric family. These two features enable to bridge the gap between parametric and nonparametric theory and to build a unified framework for statistical estimation. The main results include a large deviation bounds for the (quasi) maximum likelihood and the local quadratic bracketing of the log-likelihood process. The latter yields a number of important corollaries for statistical inference: concentration, confidence and risk bounds, expansion of the maximum likelihood estimate, etc. All these corollaries are stated in a non-classical way admitting a model misspecification and finite samples. However, the classical asymptotic results including the efficiency bounds can be easily derived as corollaries of the obtained non-asymptotic statements. At the same time, the new bracketing device works well in the situations with large or growing parameter dimension in which the classical parametric theory fails. The general results are illustrated for the i.i.d. set-up as well as for generalized linear modeling and median estimation. The results apply for any dimension of the parameter space and provide a quantitative lower bound on the sample size yielding the root-n accuracy.

1. Introduction. One of the most popular approaches in statistics is based on the parametric assumption (PA) that the distribution $I\!P$ of the observed data $\boldsymbol{Y}$ belongs to a given parametric family $(I\!P_{\boldsymbol{\theta}}, \boldsymbol{\theta} \in \Theta \subseteq I\!R^p)$, where p stands for the number of parameters. This assumption allows to reduce the problem of statistical inference about $I\!P$ to recovering the parameter $\boldsymbol{\theta}$. The theory of parameter estimation and inference is nicely developed

*The author is supported by Predictive Modeling Laboratory, MIPT, RF government grant, ag. 11.G34.31.0073. Financial support by the German Research Foundation (DFG) through the Collaborative Research Center 649 "Economic Risk" is gratefully acknowledged. Critics and suggestions of two anonymous referees helped a lot in improving the paper

AMS 2000 subject classifications: Primary 62F10; secondary 62J12,62F25,62H12

Keywords and phrases: maximum likelihood, local quadratic bracketing, deficiency, concentration

in a quite general set-up. There is a vast literature on this issue. We only mention the book by Ibragimov and Khas'minskij (1981), which provides a comprehensive study of asymptotic properties of maximum likelihood and Bayesian estimators. The theory is essentially based on two major assumptions: (1) the underlying data distribution follows the PA; (2) the sample size or the amount of available information is large relative to the number of parameters.

In many practical applications, both assumptions can be very restrictive and limit the scope of applicability for the whole approach. Indeed, the PA is usually only an approximation of real data distribution and in most statistical problems it is too restrictive to assume that the PA is exactly fulfilled. Many modern statistical problems deal with very complex high dimensional data where a huge number of parameters are involved. In such situations, the applicability of large sample asymptotics is questionable. These two issues partially explain why the parametric and nonparametric theory are almost isolated from each other. Relaxing these restrictive assumptions can be viewed as an important challenge of the modern statistical theory. The present paper attempts at developing a unified approach which does not require the restrictive parametric assumptions but still enjoys the main benefits of the parametric theory. The main steps of the approach are similar to the classical local asymptotic normality (LAN) theory; see e.g. Chapters 1–3 in the monograph Ibragimov and Khas'minskij (1981): first one localizes the problem to a neighborhood of the target parameter. Then one uses a local quadratic expansion of the log-likelihood to solve the corresponding estimation problem. There is, however, one feature of the proposed approach which makes it essentially different from classical scheme. Namely, the use of the bracketing device instead of classical Taylor expansion allows to consider much larger local neighborhoods than in the LAN theory. More specifically, the classical LAN theory effectively requires a strict localization to a root-n vicinity of the true point. At this point, the LAN theory fails in extending to the nonparametric situation. Our approach works for any local vicinity of the true point. This opens the door to building a unified theory including most of classical parametric and nonparametric results.

Let $\boldsymbol{Y}$ stand for the available data. Everywhere below we assume that the observed data $\boldsymbol{Y}$ follow the distribution $I\!P$ on a metric space $\mathcal{Y}$. We do not specify any particular structure of $\boldsymbol{Y}$. In particular, no assumption like independence or weak dependence of individual observations is imposed. The basic parametric assumption is that $I\!P$ can be approximated by a parametric distribution $I\!P_{\boldsymbol{\theta}}$ from a given parametric family $(I\!P_{\boldsymbol{\theta}}, \boldsymbol{\theta} \in \Theta \subseteq I\!R^p)$.

Our approach allows that the PA can be misspecified, that is, in general, $I\!P \notin \bigl(I\!P_{\boldsymbol{\theta}}\bigr)$.

Let $L(\boldsymbol{Y},\boldsymbol{\theta})$ be the log-likelihood for the considered parametric model: $L(\boldsymbol{Y},\boldsymbol{\theta}) = \log \frac{dI\!P_{\boldsymbol{\theta}}}{d\boldsymbol{\mu}_0}(\boldsymbol{Y})$, where $\boldsymbol{\mu}_0$ is any dominating measure for the family $(I\!P_{\boldsymbol{\theta}})$. We focus on the properties of the process $L(\boldsymbol{Y},\boldsymbol{\theta})$ as a function of the parameter $\boldsymbol{\theta}$. Therefore, we suppress the argument $\boldsymbol{Y}$ there and write $L(\boldsymbol{\theta})$ instead of $L(\boldsymbol{Y},\boldsymbol{\theta})$. One has to keep in mind that $L(\boldsymbol{\theta})$ is random and depends on the observed data $\boldsymbol{Y}$. By $L(\boldsymbol{\theta},\boldsymbol{\theta}^*) \stackrel{\text{def}}{=} L(\boldsymbol{\theta}) - L(\boldsymbol{\theta}^*)$ we denote the log-likelihood ratio. The classical likelihood principle suggests to estimate $\boldsymbol{\theta}$ by maximizing the corresponding log-likelihood function $L(\boldsymbol{\theta})$:

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

Our ultimate goal is to study the properties of the *quasi maximum likelihood estimator* (MLE) $\widetilde{\boldsymbol{\theta}}$. It turns out that such properties can be naturally described in terms of the maximum of the process $L(\boldsymbol{\theta})$ rather than the point of maximum $\widetilde{\boldsymbol{\theta}}$. To avoid technical burdens it is assumed that the maximum is attained leading to the identity $\max_{\boldsymbol{\theta}} L(\boldsymbol{\theta}) = L(\widetilde{\boldsymbol{\theta}})$. However, the point of maximum needs not to be unique. If there are many such points we take $\widetilde{\boldsymbol{\theta}}$ as any of them. Basically, the notation $\widetilde{\boldsymbol{\theta}}$ is used for the identity $L(\widetilde{\boldsymbol{\theta}}) = \sup_{\boldsymbol{\theta}\in\Theta} L(\boldsymbol{\theta})$.

If $I\!P \notin \bigl(I\!P_{\boldsymbol{\theta}}\bigr)$, then the (quasi) MLE $\widetilde{\boldsymbol{\theta}}$ from (1.1) is still meaningful and it appears to be an estimator of the value $\boldsymbol{\theta}^*$ defined by maximizing the expected value of $L(\boldsymbol{\theta})$:

$$\boldsymbol{\theta}^* \stackrel{\text{def}}{=} \operatorname*{argmax}_{\boldsymbol{\theta}\in\Theta} I\!E L(\boldsymbol{\theta}) \tag{1.2}$$

which is the true value in the parametric situation and can be viewed as the parameter of the best parametric fit in the general case.

The results below show that the main properties of the quasi MLE $\widetilde{\boldsymbol{\theta}}$ like concentration or coverage probability can be described in terms of the *excess* which is the difference between the maximum of the process $L(\boldsymbol{\theta})$ and its value at the "true" point $\boldsymbol{\theta}^*$:

$$L(\widetilde{\boldsymbol{\theta}},\boldsymbol{\theta}^*) \stackrel{\text{def}}{=} L(\widetilde{\boldsymbol{\theta}}) - L(\boldsymbol{\theta}^*) = \max_{\boldsymbol{\theta}\in\Theta} L(\boldsymbol{\theta}) - L(\boldsymbol{\theta}^*),$$

The established results can be split into two big groups. A large deviation bound states some concentration properties of the estimator $\widetilde{\boldsymbol{\theta}}$. For specific local sets $\Theta_0(\mathtt{r})$ with elliptic shape, the deviation probability $I\!P\bigl(\widetilde{\boldsymbol{\theta}} \notin \Theta_0(\mathtt{r})\bigr)$

is exponentially small in $\mathtt{r}$. This concentration bound allows to restrict the parameter space to a properly selected vicinity $\Theta_0(\mathtt{r})$. Our main results concern the local properties of the process $L(\boldsymbol{\theta})$ within $\Theta_0(\mathtt{r})$ including a bracketing bound and its corollaries.

The paper is organized as follows. Section 2 presents the list of conditions which are systematically used in the text. The conditions only concern the properties of the quasi log-likelihood process $L(\boldsymbol{\theta})$.

Section 3 appears to be central in the whole approach and it focuses on local properties of the process $L(\boldsymbol{\theta})$ within $\Theta_0(\mathtt{r})$. The idea is to sandwich the underlying (quasi) log-likelihood process $L(\boldsymbol{\theta})$ for $\boldsymbol{\theta} \in \Theta_0(\mathtt{r})$ between two quadratic (in parameter) expressions. Then the maximum of $L(\boldsymbol{\theta})$ over $\Theta_0(\mathtt{r})$ will be sandwiched as well by the maxima of the lower and upper processes. The quadratic structure of these processes help to compute these maxima explicitly yielding the bounds for the value of the original problem. This approximation result is used to derive a number of corollaries including the concentration and coverage probability, expansion of the estimator $\widetilde{\boldsymbol{\theta}}$, polynomial risk bounds, etc. In contrary to the classical theory, all the results are non-asymptotic and do not involve any small values of the form $o(1)$, all the terms are specified explicitly. Also the results are stated under possible model misspecification.

Section 4 accomplishes the local results with the concentration property which bounds the probability that $\widetilde{\boldsymbol{\theta}}$ deviates from the local set $\Theta_0(\mathtt{r})$. In the modern statistical literature there is a number of studies considering maximum likelihood or more generally minimum contrast estimators in a general i.i.d. situation, when the parameter set Θ is a subset of some functional space. We mention the papers Van de Geer (1993), Birgé and Massart (1993), Birgé and Massart (1998), Birgé (2006) and references therein. The established results are based on deep probabilistic facts from empirical process theory; see e.g. Talagrand (1996, 2001, 2005), van der Vaart and Wellner (1996), Boucheron et al. (2003). The general result presented in Section **??** follows the generic chaining idea due to Talagrand (2005); cf. Bednorz (2006). However, we do not assume any specific structure of the model. In particular, we do not assume independent observations and thus, cannot apply the most developed concentration bounds from the empirical process theory.

Section 5 illustrates the applicability of the general results to the classical case of an i.i.d. sample. The previously established general results apply under rather mild conditions. Basically we assume some smoothness of the log-likelihood process and some minimal number of observations pro parameter: the sample size should be at least of order of the dimensionality p of

the parameter space. We also consider the examples of generalized linear modeling and of median regression.

It is important to mention that the non-asymptotic character of our study yields an almost complete change of the mathematical tools: the notions of convergence and tightness become meaningless, the arguments based on compactness of the parameter space do not apply, etc. Instead we utilize the tools of the empirical process theory based on the ideas of concentration of measures and nonasymptotic entropy bounds. Section **??** in the Appendix presents an exponential bound for a general quadratic form which is very essential for getting the sharp risk bounds for the quasi MLE. This bound is an important step in the concentration results for the quasi MLE. Section **??** explains how generic chaining and majorizing measure device by Talagrand (2005) refined in Bednorz (2006) can be used for obtaining a general exponential bound for the log-likelihood process.

The proposed approach seems to be very promising and can be useful in many further research directions. Here we briefly outline some of them. The results of Sections 3 and 4 indicate the important role of the dimensionality of the parameter space. Namely, in the regular case, the entropy of the parametric family and hence, all the established bounds are proportional to the number of parameters to be estimated. This makes the results almost uninformative when the dimensionality becomes large relative to the sample size. The nonparametric theory operates with parameter sets of infinite dimension under some entropy restriction. An important advantage of the developed technique is that it is extendable to the case of the penalized likelihood. A proper choice of penalization allows to reduce the entropy bounds and to get sharp results even in the case of a large dimensional parameter space in terms of the so called effective dimension; Spokoiny (2012).

This paper focuses on the maximum likelihood estimation. However, the bracketing device appears to be very helpful and efficient for analysis of the posterior measure in the Bayes approach as well. The subsequent paper Spokoiny (2012) shows how the prominent Bernstein – von Mises result for quasi posteriors can be obtained using the new methodology. It also explains an interesting relation between the penalized maximum likelihood estimation and the Bayes method with Gaussian priors.

The proposed methodology has been successfully applied to the problem of bandwidth selection in local quantile regression, Spokoiny et al. (2012). An extension of the theory to the case of a semiparametric estimation is another important issue. The modern semiparametric theory is quite involved and requires special tools like the hardest parametric subspace. The new

approach can be directly applied to the situation when the target of analysis is a projection (mapping) of the parameter vector onto some subspace; Andresen and Spokoiny (2012). Under some restriction on the dimension or on the entropy of the nuisance parameter, the bracketing device continues to work. The established results include the semiparametric efficiency of the MLE and the Wilks theorem.

2. Conditions. Below we collect the list of conditions which are systematically used in the text. It seems to be an advantage of the whole approach that all the results are stated in a unified way under the same conditions. Once checked, one obtains automatically all the established results. We do not try to formulate the conditions and the results in the most general form. In some cases we sacrifice generality in favor of readability and ease of presentation. It is important to stress that all the conditions only concern the properties of the quasi likelihood process $L(\boldsymbol{\theta})$. Even if the process $L(\cdot)$ is not a sufficient statistic, the whole analysis is entirely based on its geometric structure and probabilistic properties. The conditions are not restrictive and can be effectively checked in many particular situations. Some examples are given in Section 5 for i.i.d setup, generalized linear models, and for median regression.

The imposed conditions can be classified into the following groups by their meaning:

- smoothness conditions on $L(\boldsymbol{\theta})$ allowing the second order Taylor expansion;
- exponential moment conditions;
- identifiability and regularity conditions;

We also distinguish between local and global conditions. The global conditions concern the global behavior of the process $L(\boldsymbol{\theta})$ while the local conditions focus on its behavior in the vicinity of the central point $\boldsymbol{\theta}^*$. Below we suppose that degree of locality is described by a number $\mathtt{r}$. The local zone corresponds to $\mathtt{r} \leq \mathtt{r}_0$ for a fixed $\mathtt{r}_0$. The global conditions concern $\mathtt{r} > 0$.

2.1. *Local conditions.* Local conditions describe the properties of $L(\boldsymbol{\theta})$ in a vicinity of the central point $\boldsymbol{\theta}^*$ from (1.2).

To bound local fluctuations of the process $L(\boldsymbol{\theta})$, we introduce an exponential moment condition on the stochastic component $\zeta(\boldsymbol{\theta})$:

$$\zeta(\boldsymbol{\theta}) \stackrel{\text{def}}{=} L(\boldsymbol{\theta}) - \mathbb{E}L(\boldsymbol{\theta}).$$

Below we suppose that the random function $\zeta(\boldsymbol{\theta})$ is differentiable in $\boldsymbol{\theta}$ and its gradient $\nabla\zeta(\boldsymbol{\theta}) = \partial\zeta(\boldsymbol{\theta})/\partial\boldsymbol{\theta} \in \mathbb{R}^p$ has some exponential moments. Our

first condition describes the property of the gradient $\nabla\zeta(\boldsymbol{\theta}^*)$ at the central point $\boldsymbol{\theta}^*$.

($\boldsymbol{ED_0}$) *There exist a positive symmetric matrix V_0^2, and constants $\mathtt{g} > 0$, $\nu_0 \geq 1$ such that $\operatorname{Var}\bigl\{\nabla\zeta(\boldsymbol{\theta}^*)\bigr\} \leq V_0^2$ and for all $|\lambda| \leq \mathtt{g}$*

$$\sup_{\boldsymbol{\gamma}\in I\!R^p} \log I\!E \exp\biggl\{\lambda \frac{\boldsymbol{\gamma}^\top \nabla\zeta(\boldsymbol{\theta}^*)}{\|V_0\boldsymbol{\gamma}\|}\biggr\} \leq \nu_0^2\lambda^2/2.$$

In typical situation, the matrix V_0^2 can be defined as the covariance matrix of the gradient vector $\nabla\boldsymbol{\zeta}(\boldsymbol{\theta}^*)$: $V_0^2 = \operatorname{Var}\bigl(\nabla\boldsymbol{\zeta}(\boldsymbol{\theta}^*)\bigr) = \operatorname{Var}\bigl(\nabla L(\boldsymbol{\theta}^*)\bigr)$. If $L(\boldsymbol{\theta})$ is the log-likelihood for a correctly specified model, then $\boldsymbol{\theta}^*$ is the true parameter value and V_0^2 coincides with the corresponding Fisher information matrix. The matrix V_0 shown in this condition determines the local geometry in the vicinity of $\boldsymbol{\theta}^*$. In particular, define the local elliptic neighborhoods of $\boldsymbol{\theta}^*$ as

$$\Theta_0(\mathtt{r}) \stackrel{\text{def}}{=} \{\boldsymbol{\theta} \in \Theta : \|V_0(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\| \leq \mathtt{r}\}. \tag{2.1}$$

The further conditions are restricted to such defined neighborhoods $\Theta_0(\mathtt{r})$.

($\boldsymbol{ED_1}$) *For each $\mathtt{r} \leq \mathtt{r}_0$, there exist a constant $\omega(\mathtt{r}) \leq 1/2$ such that it holds for all $\boldsymbol{\theta} \in \Theta_0(\mathtt{r})$*

$$\sup_{\boldsymbol{\gamma}\in I\!R^p} \log I\!E \exp\biggl\{\lambda \frac{\boldsymbol{\gamma}^\top \{\nabla\zeta(\boldsymbol{\theta}) - \nabla\zeta(\boldsymbol{\theta}^*)\}}{\omega(\mathtt{r})\|V_0\boldsymbol{\gamma}\|}\biggr\} \leq \nu_0^2\lambda^2/2, \qquad |\lambda| \leq \mathtt{g}.$$

The main bracketing result also requires second order smoothness of the expected log-likelihood $I\!E L(\boldsymbol{\theta})$. By definition, $L(\boldsymbol{\theta}^*, \boldsymbol{\theta}^*) \equiv 0$ and $\nabla I\!E L(\boldsymbol{\theta}^*) = 0$ because $\boldsymbol{\theta}^*$ is the extreme point of $I\!E L(\boldsymbol{\theta})$. Therefore, $-I\!E L(\boldsymbol{\theta}, \boldsymbol{\theta}^*)$ can be approximated by a quadratic function of $\boldsymbol{\theta} - \boldsymbol{\theta}^*$ in the neighborhood of $\boldsymbol{\theta}^*$. The *local identifiability* condition quantifies this quadratic approximation from above and from below on the set $\Theta_0(\mathtt{r})$ from (2.1).

($\boldsymbol{\mathcal{L}_0}$) *There are a symmetric strictly positive-definite matrix D_0^2 and for each $\mathtt{r} \leq \mathtt{r}_0$ and a constant $\delta(\mathtt{r}) \leq 1/2$, such that it holds on the set $\Theta_0(\mathtt{r}) = \{\boldsymbol{\theta} : \|V_0(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\| \leq \mathtt{r}\}$*

$$\biggl|\frac{-2I\!E L(\boldsymbol{\theta}, \boldsymbol{\theta}^*)}{\|D_0(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\|^2} - 1\biggr| \leq \delta(\mathtt{r}).$$

Usually D_0^2 is defined as the negative Hessian of $I\!E L(\boldsymbol{\theta}^*)$: $D_0^2 = -\nabla^2 I\!E L(\boldsymbol{\theta}^*)$. If $L(\boldsymbol{\theta}, \boldsymbol{\theta}^*)$ is the log-likelihood ratio and $I\!P = I\!P_{\boldsymbol{\theta}^*}$ then $-I\!E L(\boldsymbol{\theta}, \boldsymbol{\theta}^*) =$

$\mathbb{E}_{\boldsymbol{\theta}^*} \log\bigl(d\mathbb{P}_{\boldsymbol{\theta}^*}/d\mathbb{P}_{\boldsymbol{\theta}}\bigr) = \mathcal{K}(\mathbb{P}_{\boldsymbol{\theta}^*}, \mathbb{P}_{\boldsymbol{\theta}})$, the Kullback-Leibler divergence between $\mathbb{P}_{\boldsymbol{\theta}^*}$ and $\mathbb{P}_{\boldsymbol{\theta}}$. Then condition $(\mathcal{L}_0)$ with $D_0 = V_0$ follows from the usual regularity conditions on the family $(\mathbb{P}_{\boldsymbol{\theta}})$; cf. Ibragimov and Khas'minskij (1981). If the log-likelihood process $L(\boldsymbol{\theta})$ is sufficiently smooth in $\boldsymbol{\theta}$, e.g. three times stochastically differentiable, then the quantities $\omega(\mathtt{r})$ and $\delta(\mathtt{r})$ can be taken proportional to the value $\varrho(\mathtt{r})$ defined as

$$\varrho(\mathtt{r}) \stackrel{\text{def}}{=} \max_{\boldsymbol{\theta} \in \Theta_0(\mathtt{r})} \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|.$$

In the important special case of an i.i.d. model one can take $\omega(\mathtt{r}) = \omega^* \mathtt{r}/n^{1/2}$ and $\delta(\mathtt{r}) = \delta^* \mathtt{r}/n^{1/2}$ for some constants ω^*, δ^*; see Section 5.1.

The *identifiability condition* relates the matrices D_0^2 and V_0^2.

$(\boldsymbol{\mathcal{I}})$ There is a constant $\mathtt{a} > 0$ such that $\mathtt{a}^2 D_0^2 \geq V_0^2$.

2.2. *Global conditions.* The global conditions have to be fulfilled for all $\boldsymbol{\theta}$ lying beyond $\Theta_0(\mathtt{r}_0)$. We only impose one condition on the smoothness of the stochastic component of the process $L(\boldsymbol{\theta})$ in term of its gradient, and one identifiability condition in terms of the expectation $\mathbb{E}L(\boldsymbol{\theta}, \boldsymbol{\theta}^*)$.

The first condition is similar to the local condition (ED_0) and it requires some exponential moment of the gradient $\nabla\zeta(\boldsymbol{\theta})$ for all $\boldsymbol{\theta} \in \Theta$. However, the constant $\mathtt{g}$ may be dependent of the radius $\mathtt{r} = \|V_0(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\|$.

($\boldsymbol{E}\mathtt{r}$) *For any* $\mathtt{r}$, *there exists a value* $\mathtt{g}(\mathtt{r}) > 0$ *such that for all* $\lambda \leq \mathtt{g}(\mathtt{r})$

$$\sup_{\boldsymbol{\theta} \in \Theta_0(\mathtt{r})} \sup_{\boldsymbol{\gamma} \in \mathbb{R}^p} \log \mathbb{E} \exp\biggl\{ \lambda \frac{\boldsymbol{\gamma}^\top \nabla\zeta(\boldsymbol{\theta})}{\|V_0 \boldsymbol{\gamma}\|} \biggr\} \leq \nu_0^2 \lambda^2/2.$$

The global identification property means that the deterministic component $\mathbb{E}L(\boldsymbol{\theta}, \boldsymbol{\theta}^*)$ of the log-likelihood is competitive with its variance $\operatorname{Var} L(\boldsymbol{\theta}, \boldsymbol{\theta}^*)$.

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

$$\inf_{\boldsymbol{\theta}:\, \|V_0(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\| = \mathtt{r}} \bigl| \mathbb{E}L(\boldsymbol{\theta}, \boldsymbol{\theta}^*) \bigr| \geq \mathtt{b}(\mathtt{r}) \mathtt{r}^2.$$

3. Local inference. The *Local Asymptotic Normality* (LAN) condition since introduced in Le Cam (1960) became one of the central notions in the statistical theory. It postulates a kind of local approximation of the log-likelihood of the original model by the log-likelihood of a Gaussian shift

experiment. The LAN property being once checked yields a number of important corollaries for statistical inference. In words, if you can solve a statistical problem for the Gaussian shift model, the result can be translated under the LAN condition to the original setup. We refer to Ibragimov and Khas'minskij (1981) for a nice presentation of the LAN theory including asymptotic efficiency of MLE and Bayes estimators. The LAN property was extended to *mixed LAN* or *Local Asymptotic Quadraticity* (LAQ); see e.g. Le Cam and Yang (2000). All these notions are very much asymptotic and very much local. The LAN theory also requires that $L(\boldsymbol{\theta})$ is the correctly specified log-likelihood. The strict localization does not allow for considering a growing or infinite parameter dimension and limits applications of the LAN theory to nonparametric estimation.

Our approach tries to avoid asymptotic constructions and attempts to include a possible model misspecification and a large dimension of the parameter space. The presentation below shows that such an extension of the LAN theory can be made essentially by no price: all the major asymptotic results like Fisher and Cramér-Rao information bounds, as well as the Wilks phenomenon can be derived as corollaries of the obtained non-asymptotic statements simply by letting the sample size to infinity. At the same time, it applies to a high dimensional parameter space.

The LAN property states that the considered process $L(\boldsymbol{\theta})$ can be approximated by a quadratic in $\boldsymbol{\theta}$ expression in a vicinity of the central point $\boldsymbol{\theta}^*$. This property is usually checked using the second order Taylor expansion. The main problem arising here is that the error of the approximation grows too fast with the local size of the neighborhood. Section 3.1 presents the non-asymptotic version of the LAN property in which the local quadratic approximation of $L(\boldsymbol{\theta})$ is replaced by bounding this process from above and from below by two different quadratic in $\boldsymbol{\theta}$ processes. More precisely, we apply the *bracketing* idea: the difference $L(\boldsymbol{\theta},\boldsymbol{\theta}^*) = L(\boldsymbol{\theta}) - L(\boldsymbol{\theta}^*)$ is put between two quadratic processes $\mathbb{L}_{\underline{\boldsymbol{\epsilon}}}(\boldsymbol{\theta},\boldsymbol{\theta}^*)$ and $\mathbb{L}_{\boldsymbol{\epsilon}}(\boldsymbol{\theta},\boldsymbol{\theta}^*)$:

$$\mathbb{L}_{\underline{\boldsymbol{\epsilon}}}(\boldsymbol{\theta},\boldsymbol{\theta}^*) - \diamondsuit_{\underline{\boldsymbol{\epsilon}}} \leq L(\boldsymbol{\theta},\boldsymbol{\theta}^*) \leq \mathbb{L}_{\boldsymbol{\epsilon}}(\boldsymbol{\theta},\boldsymbol{\theta}^*) + \diamondsuit_{\boldsymbol{\epsilon}}, \qquad \boldsymbol{\theta} \in \Theta_0(\mathtt{r}), \tag{3.1}$$

where $\boldsymbol{\epsilon}$ is a numerical parameter, $\underline{\boldsymbol{\epsilon}} = -\boldsymbol{\epsilon}$, and $\diamondsuit_{\underline{\boldsymbol{\epsilon}}}$ and $\diamondsuit_{\boldsymbol{\epsilon}}$ are stochastic errors which only depend on the selected vicinity $\Theta_0(\mathtt{r})$. The upper process $\mathbb{L}_{\boldsymbol{\epsilon}}(\boldsymbol{\theta},\boldsymbol{\theta}^*)$ and the lower process $\mathbb{L}_{\underline{\boldsymbol{\epsilon}}}(\boldsymbol{\theta},\boldsymbol{\theta}^*)$ can deviate substantially from each other, however, the errors $\diamondsuit_{\boldsymbol{\epsilon}}, \diamondsuit_{\underline{\boldsymbol{\epsilon}}}$ remain small even if the value $\mathtt{r}$ describing the size of the local neighborhood $\Theta_0(\mathtt{r})$ is large.

The sandwiching result (3.1) naturally leads to two important notions: the value of the problem and the spread. It turns out that most of the statements like confidence and concentration probability rely upon the maximum of

$L(\boldsymbol{\theta},\boldsymbol{\theta}^*)$ over $\boldsymbol{\theta}$ which we call *the excess*. Its expectation will be referred to as *the value of the problem*. Due to (3.1) the excess can be bounded from above and from below using the similar quantities $\max_{\boldsymbol{\theta}} \mathbb{L}_{\underline{\boldsymbol{\epsilon}}}(\boldsymbol{\theta},\boldsymbol{\theta}^*)$ and $\max_{\boldsymbol{\theta}} \mathbb{L}_{\boldsymbol{\epsilon}}(\boldsymbol{\theta},\boldsymbol{\theta}^*)$ which can be called the *lower and upper excess*, while their expectations are *the values of the lower and upper problems*. Note that $\max_{\boldsymbol{\theta}}\{\mathbb{L}_{\boldsymbol{\epsilon}}(\boldsymbol{\theta},\boldsymbol{\theta}^*) - \mathbb{L}_{\underline{\boldsymbol{\epsilon}}}(\boldsymbol{\theta},\boldsymbol{\theta}^*)\} = \infty$. However, this is not crucial. What really matters is the difference between the upper and the lower excess. The *spread* $\Delta_{\boldsymbol{\epsilon}}$ can be defined as the width of the interval bounding the excess due to (3.1), that is, as the sum of the approximation errors and of the difference between the upper and the lower excess:

$$\Delta_{\boldsymbol{\epsilon}} \stackrel{\text{def}}{=} \diamondsuit_{\boldsymbol{\epsilon}} + \diamondsuit_{\underline{\boldsymbol{\epsilon}}} + \bigl\{\max_{\boldsymbol{\theta}} \mathbb{L}_{\boldsymbol{\epsilon}}(\boldsymbol{\theta},\boldsymbol{\theta}^*) - \max_{\boldsymbol{\theta}} \mathbb{L}_{\underline{\boldsymbol{\epsilon}}}(\boldsymbol{\theta},\boldsymbol{\theta}^*)\bigr\}.$$

The range of applicability of this approach can be described by the following mnemonic rule: "The value of the upper problem is larger in order than the spread." The further sections explain in details the meaning and content of this rule. Section 3.1 presents the key bound (3.1) and derives it from the general results on empirical processes. Section 3.2 presents some straightforward corollaries of the bound (3.1) including the coverage and concentration probabilities, expansion of the MLE and the risk bounds. It also indicates how the classical results on asymptotic efficiency of the MLE follow from the obtained non-asymptotic bounds.

3.1. *Local quadratic bracketing.* This section presents the key result about local quadratic approximation of the quasi log-likelihood process given by Theorem 3.1 below.

Let the radius $\mathtt{r}$ of the local neighborhood $\Theta_0(\mathtt{r})$ be fixed in a way that the deviation probability $\mathbb{P}\bigl(\widetilde{\boldsymbol{\theta}} \notin \Theta_0(\mathtt{r})\bigr)$ is sufficiently small. Precise results about the choice of $\mathtt{r}$ which ensures this property are postponed until Section 4. In this neighborhood $\Theta_0(\mathtt{r})$ we aim at building some quadratic lower and upper bounds for the process $L(\boldsymbol{\theta})$. The first step is the usual decomposition of this process into deterministic and stochastic components:

$$L(\boldsymbol{\theta}) = \mathbb{E}L(\boldsymbol{\theta}) + \zeta(\boldsymbol{\theta}),$$

where $\zeta(\boldsymbol{\theta}) = L(\boldsymbol{\theta}) - \mathbb{E}L(\boldsymbol{\theta})$. Condition $(\mathcal{L}_0)$ allows to approximate the smooth deterministic function $\mathbb{E}L(\boldsymbol{\theta}) - \mathbb{E}L(\boldsymbol{\theta}^*)$ around the point of maximum $\boldsymbol{\theta}^*$ by the quadratic form $-\|D_0(\boldsymbol{\theta}-\boldsymbol{\theta}^*)\|^2/2$. The smoothness properties of the stochastic component $\zeta(\boldsymbol{\theta})$ given by conditions (ED_0) and (ED_1) leads to linear approximation $\zeta(\boldsymbol{\theta}) - \zeta(\boldsymbol{\theta}^*) \approx (\boldsymbol{\theta}-\boldsymbol{\theta}^*)^\top \nabla\zeta(\boldsymbol{\theta}^*)$.

Putting these two approximations together yields the following approximation of the process $L(\boldsymbol{\theta})$ on $\Theta_0(\mathtt{r})$:

$$L(\boldsymbol{\theta},\boldsymbol{\theta}^*) \approx \mathbb{L}(\boldsymbol{\theta},\boldsymbol{\theta}^*) \stackrel{\text{def}}{=} (\boldsymbol{\theta}-\boldsymbol{\theta}^*)^\top \nabla\zeta(\boldsymbol{\theta}^*) - \|D_0(\boldsymbol{\theta}-\boldsymbol{\theta}^*)\|^2/2. \quad (3.2)$$

This expansion is used in most of statistical calculus. However, it does not suit our purposes because the error of approximation grows quadratically with the radius $\mathtt{r}$ and starts to dominate at some critical value of $\mathtt{r}$. We slightly modify the construction by introducing two different approximating processes. They only differ in the deterministic quadratic term which is either shrunk or stretched relative to the term $\|D_0(\boldsymbol{\theta}-\boldsymbol{\theta}^*)\|^2/2$ in $\mathbb{L}(\boldsymbol{\theta},\boldsymbol{\theta}^*)$.

Let δ,ϱ be nonnegative constants. Introduce for a vector $\boldsymbol{\epsilon}=(\delta,\varrho)$ the following notation:

$$\begin{aligned}\mathbb{L}_{\boldsymbol{\epsilon}}(\boldsymbol{\theta},\boldsymbol{\theta}^*) &\stackrel{\text{def}}{=} (\boldsymbol{\theta}-\boldsymbol{\theta}^*)^\top \nabla L(\boldsymbol{\theta}^*) - \|D_{\boldsymbol{\epsilon}}(\boldsymbol{\theta}-\boldsymbol{\theta}^*)\|^2/2 \\ &= \boldsymbol{\xi}_{\boldsymbol{\epsilon}}^\top D_{\boldsymbol{\epsilon}}(\boldsymbol{\theta}-\boldsymbol{\theta}^*) - \|D_{\boldsymbol{\epsilon}}(\boldsymbol{\theta}-\boldsymbol{\theta}^*)\|^2/2, \qquad (3.3)\end{aligned}$$

where

$$D_{\boldsymbol{\epsilon}}^2 = D_0^2(1-\delta) - \varrho V_0^2, \qquad \boldsymbol{\xi}_{\boldsymbol{\epsilon}} \stackrel{\text{def}}{=} D_{\boldsymbol{\epsilon}}^{-1}\nabla L(\boldsymbol{\theta}^*).$$

Here we implicitly assume that with the proposed choice of the constants δ and ϱ, the matrix $D_{\boldsymbol{\epsilon}}^2$ is non-negative: $D_{\boldsymbol{\epsilon}}^2 \geq 0$. The representation (3.3) indicates that the process $\mathbb{L}_{\boldsymbol{\epsilon}}(\boldsymbol{\theta},\boldsymbol{\theta}^*)$ has the geometric structure of log-likelihood of a linear Gaussian model. We do not require that the vector $\boldsymbol{\xi}_{\boldsymbol{\epsilon}}$ is Gaussian and hence, it is not the Gaussian log-likelihood. However, the geometric structure of this process appears to be more important than its distributional properties.

One can see that if δ,ϱ are positive, the quadratic drift component of the process $\mathbb{L}_{\boldsymbol{\epsilon}}(\boldsymbol{\theta},\boldsymbol{\theta}^*)$ is shrunk relative to $\mathbb{L}(\boldsymbol{\theta},\boldsymbol{\theta}^*)$ in (3.2) for $\boldsymbol{\epsilon}$ positive and it is stretched if δ,ϱ are negative. Now, given $\mathtt{r}$, define $\delta=\delta(\mathtt{r})$, $\varrho = 3\nu_0\omega(\mathtt{r})$ with the value $\delta(\mathtt{r})$ from condition $(\mathcal{L}_0)$ and $\omega(\mathtt{r})$ from condition (ED_1). Finally set $\underline{\boldsymbol{\epsilon}}=-\boldsymbol{\epsilon}$, so that $D_{\underline{\boldsymbol{\epsilon}}}^2 = D_0^2(1+\delta)+\varrho V_0^2$.

THEOREM 3.1. *Assume* (ED_1) *and* $(\mathcal{L}_0)$. *Let for some* $\mathtt{r}$, *the values* $\varrho \geq 3\nu_0\,\omega(\mathtt{r})$ *and* $\delta \geq \delta(\mathtt{r})$ *be such that* $D_0^2(1-\delta)-\varrho V_0^2 \geq 0$. *Then*

$$\mathbb{L}_{\underline{\boldsymbol{\epsilon}}}(\boldsymbol{\theta},\boldsymbol{\theta}^*) - \diamondsuit_{\underline{\boldsymbol{\epsilon}}}(\mathtt{r}) \leq L(\boldsymbol{\theta},\boldsymbol{\theta}^*) \leq \mathbb{L}_{\boldsymbol{\epsilon}}(\boldsymbol{\theta},\boldsymbol{\theta}^*) + \diamondsuit_{\boldsymbol{\epsilon}}(\mathtt{r}), \quad \boldsymbol{\theta}\in\Theta_0(\mathtt{r}), \qquad (3.4)$$

with $\mathbb{L}_{\boldsymbol{\epsilon}}(\boldsymbol{\theta},\boldsymbol{\theta}^*), \mathbb{L}_{\underline{\boldsymbol{\epsilon}}}(\boldsymbol{\theta},\boldsymbol{\theta}^*)$ *defined by* (3.3). *The error terms* $\diamondsuit_{\boldsymbol{\epsilon}}(\mathtt{r})$ *and* $\diamondsuit_{\underline{\boldsymbol{\epsilon}}}(\mathtt{r})$ *satisfy the bound* (3.11) *from Proposition 3.7.*

REMARK 3.1. This bracketing bound (3.4) describes some properties of the log-likelihood process and the estimator $\widetilde{\boldsymbol{\theta}}$ is not shown there. However, it directly implies most of our inference results. We therefore formulate it as a separate statement. Section 3.3 below presents some exponential bounds on the error terms $\diamondsuit_{\boldsymbol{\epsilon}}(\mathtt{r})$ and $\diamondsuit_{\underline{\boldsymbol{\epsilon}}}(\mathtt{r})$. The main message is that under rather broad conditions, these errors are small and have only minor impact on the inference for the quasi MLE $\widetilde{\boldsymbol{\theta}}$.

3.2. *Local inference.* This section presents a list of corollaries from the basic approximation bounds of Theorem 3.1. The idea is to replace the original problem by a similar one for the approximating upper and lower models. It is important to stress once again that all the corollaries only rely on the *bracketing result* (3.1) and the *geometric structure* of the processes $\mathbb{L}_{\boldsymbol{\epsilon}}$ and $\mathbb{L}_{\underline{\boldsymbol{\epsilon}}}$. Define the *spread* $\Delta_{\boldsymbol{\epsilon}}(\mathtt{r})$ by

$$\Delta_{\boldsymbol{\epsilon}}(\mathtt{r}) \stackrel{\text{def}}{=} \diamondsuit_{\boldsymbol{\epsilon}}(\mathtt{r}) + \diamondsuit_{\underline{\boldsymbol{\epsilon}}}(\mathtt{r}) + \bigl(\|\boldsymbol{\xi}_{\boldsymbol{\epsilon}}\|^2 - \|\boldsymbol{\xi}_{\underline{\boldsymbol{\epsilon}}}\|^2\bigr)/2. \tag{3.5}$$

Here $\boldsymbol{\xi}_{\boldsymbol{\epsilon}} = D_{\boldsymbol{\epsilon}}^{-1}\nabla L(\boldsymbol{\theta}^*)$ and $\boldsymbol{\xi}_{\underline{\boldsymbol{\epsilon}}} = D_{\underline{\boldsymbol{\epsilon}}}^{-1}\nabla L(\boldsymbol{\theta}^*)$. The quantity $\Delta_{\boldsymbol{\epsilon}}(\mathtt{r})$ appears to be the price induced by our bracketing device. Section 3.3 below presents some probabilistic bounds on the spread showing that it is small relative to the other terms. All our corollaries below are stated under conditions of Theorem 3.1 and implicitly assume that the spread can be nearly ignored.

3.2.1. *Local coverage probability.* Our first result describes the probability of covering $\boldsymbol{\theta}^*$ by the random set

$$\mathcal{E}(\mathfrak{z}) = \{\boldsymbol{\theta} : 2L(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}) \le \mathfrak{z}\}. \tag{3.6}$$

COROLLARY 3.2. *For any* $\mathfrak{z} > 0$

$$I\!P\bigl\{\mathcal{E}(\mathfrak{z}) \not\ni \boldsymbol{\theta}^*, \widetilde{\boldsymbol{\theta}} \in \Theta_0(\mathtt{r})\bigr\} \le I\!P\bigl\{\|\boldsymbol{\xi}_{\boldsymbol{\epsilon}}\|^2 \ge \mathfrak{z} - \diamondsuit_{\boldsymbol{\epsilon}}(\mathtt{r})\bigr\}. \tag{3.7}$$

PROOF. The bound (3.7) follows from the upper bound of Theorem 3.1 and the statement (3.12) of Lemma 3.8 below. □

Below; see (3.14), we also present an exponential bound which helps to answer a very important question about a proper choice of the critical value $\mathfrak{z}$ ensuring a prescribed covering probability.

3.2.2. *Local expansion, Wilks theorem, and local concentration.* Now we show how the bound (3.4) can be used for obtaining a local expansion of the

quasi MLE $\widetilde{\boldsymbol{\theta}}$. All our results will be conditioned to the random set $C_{\boldsymbol{\epsilon}}(\mathtt{r})$ defined as

$$C_{\boldsymbol{\epsilon}}(\mathtt{r}) \stackrel{\text{def}}{=} \{\widetilde{\boldsymbol{\theta}} \in \Theta_0(\mathtt{r}),\ \|V_0 D_{\underline{\boldsymbol{\epsilon}}}^{-1} \boldsymbol{\xi}_{\underline{\boldsymbol{\epsilon}}}\| \le \mathtt{r}\}. \tag{3.8}$$

Below in Section 3.3 we present some upper bounds on the value $\mathtt{r}$ ensuring a dominating probability of this random set.

The first result can be viewed as a finite sample version of the famous Wilks theorem.

Corollary 3.3. *On the random set $C_{\boldsymbol{\epsilon}}(\mathtt{r})$ from* (3.8)*, it holds*

$$\|\boldsymbol{\xi}_{\underline{\boldsymbol{\epsilon}}}\|^2/2 - \diamondsuit_{\underline{\boldsymbol{\epsilon}}}(\mathtt{r}) \le L(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}^*) \le \|\boldsymbol{\xi}_{\boldsymbol{\epsilon}}\|^2/2 + \diamondsuit_{\boldsymbol{\epsilon}}(\mathtt{r}). \tag{3.9}$$

The next result is an extension of another prominent asymptotic result, namely, the Fisher expansion of the MLE.

Corollary 3.4. *On the random set $C_{\boldsymbol{\epsilon}}(\mathtt{r})$ from* (3.8)*, it holds*

$$\bigl\| D_{\boldsymbol{\epsilon}}\bigl(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\bigr) - \boldsymbol{\xi}_{\boldsymbol{\epsilon}} \bigr\|^2 \le 2\Delta_{\boldsymbol{\epsilon}}(\mathtt{r}). \tag{3.10}$$

The proof of Corollaries 3.3 and 3.4 relies on the solution of the upper and lower problems and it is given below at the end of this section.

Now we describe concentration properties of $\widetilde{\boldsymbol{\theta}}$ assuming that $\widetilde{\boldsymbol{\theta}}$ is restricted to $\Theta_0(\mathtt{r})$. More precisely, we bound the probability that $\|D_{\boldsymbol{\epsilon}}(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\| > z$ for a given $z > 0$.

Corollary 3.5. *For any $z > 0$, it holds*

$$I\!P\bigl\{\|D_{\boldsymbol{\epsilon}}(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\| > z,\, C_{\boldsymbol{\epsilon}}(\mathtt{r})\bigr\} \le I\!P\bigl\{\|\boldsymbol{\xi}_{\boldsymbol{\epsilon}}\| > z - \sqrt{2\Delta_{\boldsymbol{\epsilon}}(\mathtt{r})}\bigr\}$$

An interesting and important question is for which $\mathfrak{z}$ in (3.6) the coverage probability of the event $\bigl\{\mathcal{E}(\mathfrak{z}) \ni \boldsymbol{\theta}^*\bigr\}$ or for which z, the concentration probability of the event $\{\|D_{\boldsymbol{\epsilon}}(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\| \le z\}$ becomes close to one. It will be addressed in Section 3.3.

3.2.3. *A local risk bound.* Below we also bound the moments of the excess $L(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}^*)$ and of the normalized loss $D_{\boldsymbol{\epsilon}}(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)$ when $\widetilde{\boldsymbol{\theta}}$ is restricted to $\Theta_0(\mathtt{r})$. The result follows directly from Corollaries 3.3 and 3.4.

Corollary 3.6. *For $u > 0$*

$$I\!E\bigl\{L^u(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}^*)\, 1\!\!1\bigl(\widetilde{\boldsymbol{\theta}} \in \Theta_0(\mathtt{r})\bigr)\bigr\} \le I\!E\bigl[\bigl\{\|\boldsymbol{\xi}_{\boldsymbol{\epsilon}}\|^2/2 + \diamondsuit_{\boldsymbol{\epsilon}}(\mathtt{r})\bigr\}^u\bigr].$$

Moreover, it holds

$$\mathbb{E}\bigl\{\|D_{\boldsymbol{\epsilon}}(\widetilde{\boldsymbol{\theta}}-\boldsymbol{\theta}^*)\|^u\,\mathbb{1}\bigl(C_{\boldsymbol{\epsilon}}(\mathtt{r})\bigr)\bigr\} \le \mathbb{E}\bigl[\bigl\{\|\boldsymbol{\xi}_{\boldsymbol{\epsilon}}\|+\sqrt{2\Delta_{\boldsymbol{\epsilon}}(\mathtt{r})}\bigr\}^u\bigr].$$

3.2.4. *Comparing with the asymptotic theory.* This section briefly discusses the relation between the established non-asymptotic bounds and the classical asymptotic results in parametric estimation. This comparison is not straightforward because the asymptotic theory involves the sample size or noise level as the asymptotic parameter, while our setup is very general and works even for a "single" observation. Here we simply treat $\boldsymbol{\epsilon}=(\delta,\varrho)$ as a small parameter. This is well justified by the i.i.d. case with n observations, where it holds $\delta=\delta(\mathtt{r})\asymp\sqrt{\mathtt{r}/n}$ and similarly for ϱ; see Section 5 for more details. The bounds below in Section 3.3 show that the spread $\Delta_{\boldsymbol{\epsilon}}(\mathtt{r})$ from (3.5) is small and can be ignored in the asymptotic calculations. The results of Corollary 3.2 through 3.6 represent the desired bounds in terms of deviation bounds for the quadratic form $\|\boldsymbol{\xi}_{\boldsymbol{\epsilon}}\|^2$.

For better understanding the essence of the presented results, consider first the "true" parametric model with the correctly specified log-likelihood $L(\boldsymbol{\theta})$. Then $D_0^2=V_0^2$ is the total Fisher information matrix. In the i.i.d. case it becomes $n\mathbf{f}_0$ where $\mathbf{f}_0$ is the usual Fisher information matrix of the considered parametric family at $\boldsymbol{\theta}^*$. In particular, $\operatorname{Var}\bigl\{\nabla L(\boldsymbol{\theta}^*)\bigr\}=n\mathbf{f}_0$. So, if $D_{\boldsymbol{\epsilon}}$ is close to D_0, then $\boldsymbol{\xi}_{\boldsymbol{\epsilon}}$ can be treated as the normalized score. Under usual assumptions, $\boldsymbol{\xi}\stackrel{\text{def}}{=}D_0^{-1}\nabla L(\boldsymbol{\theta}^*)$ is asymptotically standard normal p-vector. The same applies to $\boldsymbol{\xi}_{\boldsymbol{\epsilon}}$. Now one can observe that Corollary 3.2 through 3.6 directly imply most of classical asymptotic statements. In particular, Corollary 3.3 shows that the twice excess $2L(\widetilde{\boldsymbol{\theta}},\boldsymbol{\theta}^*)$ is nearly $\|\boldsymbol{\xi}_{\boldsymbol{\epsilon}}\|^2$ and thus nearly χ^2_p (Wilks Theorem). Corollary 3.4 yields the expansion $D_{\boldsymbol{\epsilon}}\bigl(\widetilde{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\bigr)\approx\boldsymbol{\xi}_{\boldsymbol{\epsilon}}$ (the Fisher expansion) and hence, $D_{\boldsymbol{\epsilon}}\bigl(\widetilde{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\bigr)$ is asymptotically standard normal. Asymptotic variance of $D_{\boldsymbol{\epsilon}}\bigl(\widetilde{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\bigr)$ is nearly one, so $\widetilde{\boldsymbol{\theta}}$ achieves the Cramér-Rao efficiency bound in the asymptotic set-up.

3.3. *Spread.* This section presents some bounds on the value

$$\Delta_{\boldsymbol{\epsilon}}(\mathtt{r})\stackrel{\text{def}}{=}\diamondsuit_{\boldsymbol{\epsilon}}(\mathtt{r})+\diamondsuit_{\underline{\boldsymbol{\epsilon}}}(\mathtt{r})+\bigl(\|\boldsymbol{\xi}_{\boldsymbol{\epsilon}}\|^2-\|\boldsymbol{\xi}_{\underline{\boldsymbol{\epsilon}}}\|^2\bigr)/2.$$

This quantity is random but it can be easily evaluated under the conditions made. We present two different results: one bounds the errors $\diamondsuit_{\boldsymbol{\epsilon}}(\mathtt{r}),\diamondsuit_{\underline{\boldsymbol{\epsilon}}}(\mathtt{r})$, while the other presents a deviation bound on quadratic forms like $\|\boldsymbol{\xi}_{\boldsymbol{\epsilon}}\|^2$. The results are stated under conditions (ED_0) and (ED_1) in a non-asymptotic way, so the formulation is quite technical. An informal discussion at the end

of this section explains the typical behavior of the spread. The first result accomplishes the bracketing bound (3.4).

PROPOSITION 3.7. *Assume* (ED_1). *The error term* $\diamondsuit_{\boldsymbol{\epsilon}}(\mathtt{r})$ *in* (3.4) *fulfills*

$$\mathbb{P}\bigl\{\varrho^{-1}\diamondsuit_{\boldsymbol{\epsilon}}(\mathtt{r}) \geq \mathfrak{z}_0(\mathtt{x},\mathbb{Q})\bigr\} \leq \exp\bigl(-\mathtt{x}\bigr) \tag{3.11}$$

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

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

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

REMARK 3.2. The bound (3.11) essentially depends on the value $\mathtt{g}$ from condition (ED_1). The result requires that $\mathtt{g}\nu_0 \geq 3$. However, this constant can usually be taken of order $n^{1/2}$; see Section 5 for examples. If $\mathtt{g}^2$ is larger in order than $p+\mathtt{x}$, then $\mathfrak{z}_0(\mathtt{x},\mathbb{Q}) \approx \mathfrak{c}_1 p + \mathtt{x}$.

PROOF. Consider for fixed $\mathtt{r}$ and $\boldsymbol{\epsilon} = (\delta,\varrho)$ the quantity

$$\diamondsuit_{\boldsymbol{\epsilon}}(\mathtt{r}) \stackrel{\text{def}}{=} \sup_{\boldsymbol{\theta}\in\Theta_0(\mathtt{r})} \bigl\{L(\boldsymbol{\theta},\boldsymbol{\theta}^*) - \mathbb{E}L(\boldsymbol{\theta},\boldsymbol{\theta}^*) - (\boldsymbol{\theta}-\boldsymbol{\theta}^*)^\top\nabla L(\boldsymbol{\theta}^*) - \frac{\varrho}{2}\|V_0(\boldsymbol{\theta}-\boldsymbol{\theta}^*)\|^2\bigr\}.$$

As $\delta \geq \delta(\mathtt{r})$, it holds $-\mathbb{E}L(\boldsymbol{\theta},\boldsymbol{\theta}^*) \geq (1-\delta)D_0^2$ and $L(\boldsymbol{\theta},\boldsymbol{\theta}^*) - \mathbb{L}_{\boldsymbol{\epsilon}}(\boldsymbol{\theta},\boldsymbol{\theta}^*) \leq \diamondsuit_{\boldsymbol{\epsilon}}(\mathtt{r})$. Moreover, in view of $\nabla\mathbb{E}L(\boldsymbol{\theta}^*) = 0$, the definition of $\diamondsuit_{\boldsymbol{\epsilon}}(\mathtt{r})$ can be rewritten in the form

$$\diamondsuit_{\boldsymbol{\epsilon}}(\mathtt{r}) \stackrel{\text{def}}{=} \sup_{\boldsymbol{\theta}\in\Theta_0(\mathtt{r})} \bigl\{\zeta(\boldsymbol{\theta},\boldsymbol{\theta}^*) - (\boldsymbol{\theta}-\boldsymbol{\theta}^*)^\top\nabla\zeta(\boldsymbol{\theta}^*) - \frac{\varrho}{2}\|V_0(\boldsymbol{\theta}-\boldsymbol{\theta}^*)\|^2\bigr\}.$$

Now the claim of the theorem can be easily reduced to an exponential bound for the quantity $\diamondsuit_{\boldsymbol{\epsilon}}(\mathtt{r})$. We apply Theorem **??** to the process

$$\mathcal{U}(\boldsymbol{\theta},\boldsymbol{\theta}^*) = \frac{1}{\omega(\mathtt{r})}\bigl\{\zeta(\boldsymbol{\theta},\boldsymbol{\theta}^*) - (\boldsymbol{\theta}-\boldsymbol{\theta}^*)^\top\nabla\zeta(\boldsymbol{\theta}^*)\bigr\}, \qquad \boldsymbol{\theta}\in\Theta_0(\mathtt{r}),$$

and $H_0 = V_0$. Condition $(\mathcal{E}D)$ follows from (ED_1) with the same ν_0 and $\mathtt{g}$ in view of $\nabla\mathcal{U}(\boldsymbol{\theta},\boldsymbol{\theta}^*) = \bigl\{\nabla\zeta(\boldsymbol{\theta}) - \nabla\zeta(\boldsymbol{\theta}^*)\bigr\}/\omega(\mathtt{r})$. So, the conditions of Theorem **??** are fulfilled yielding (3.11) in view of $\varrho \geq 3\nu_0\,\omega(\mathtt{r})$. □

Due to the main bracketing result, the local excess $\sup_{\boldsymbol{\theta}\in\Theta_0(\mathtt{r})}$ can be put between similar quantities for the upper and lower approximating processes up to the error term $\diamondsuit^*_{\boldsymbol{\epsilon}}(\mathtt{r})$. The random quantity $\sup_{\boldsymbol{\theta}} \mathbb{L}_{\boldsymbol{\epsilon}}(\boldsymbol{\theta},\boldsymbol{\theta}^*)$ can be called the *upper excess* while $\sup_{\boldsymbol{\theta}} \mathbb{L}_{\underline{\boldsymbol{\epsilon}}}(\boldsymbol{\theta},\boldsymbol{\theta}^*)$ is the *lower excess*. The quadratic (in $\boldsymbol{\theta}$) structure of the functions $\mathbb{L}_{\boldsymbol{\epsilon}}(\boldsymbol{\theta},\boldsymbol{\theta}^*)$ and $\mathbb{L}_{\underline{\boldsymbol{\epsilon}}}(\boldsymbol{\theta},\boldsymbol{\theta}^*)$ enables us to explicitly solve the problem of maximizing the corresponding function w.r.t. $\boldsymbol{\theta}$. Next result describes the *upper and lower excesses* $\sup_{\boldsymbol{\theta}} \mathbb{L}_{\underline{\boldsymbol{\epsilon}}}(\boldsymbol{\theta},\boldsymbol{\theta}^*)$.

Lemma 3.8. *It holds*

$$\sup_{\boldsymbol{\theta}\in I\!R^p} \mathbb{L}_{\boldsymbol{\epsilon}}(\boldsymbol{\theta},\boldsymbol{\theta}^*) = \|\boldsymbol{\xi}_{\boldsymbol{\epsilon}}\|^2/2. \tag{3.12}$$

On the random set $\{\|V_0 D_{\underline{\boldsymbol{\epsilon}}}^{-1}\boldsymbol{\xi}_{\underline{\boldsymbol{\epsilon}}}\| \le \mathtt{r}\}$, *it also holds*

$$\sup_{\boldsymbol{\theta}\in\Theta_0(\mathtt{r})} \mathbb{L}_{\underline{\boldsymbol{\epsilon}}}(\boldsymbol{\theta},\boldsymbol{\theta}) = \|\boldsymbol{\xi}_{\underline{\boldsymbol{\epsilon}}}\|^2/2.$$

Proof. The unconstrained maximum of the quadratic form $\mathbb{L}_{\boldsymbol{\epsilon}}(\boldsymbol{\theta},\boldsymbol{\theta}^*)$ w.r.t. $\boldsymbol{\theta}$ is attained at $\widetilde{\boldsymbol{\theta}}_{\boldsymbol{\epsilon}} = D_{\boldsymbol{\epsilon}}^{-1}\boldsymbol{\xi}_{\boldsymbol{\epsilon}} = D_{\boldsymbol{\epsilon}}^{-2}\nabla L(\boldsymbol{\theta}^*)$ yielding the expression (3.12). The lower excess is computed similarly. □

Our next step is in bounding the difference $\|\boldsymbol{\xi}_{\boldsymbol{\epsilon}}\|^2 - \|\boldsymbol{\xi}_{\underline{\boldsymbol{\epsilon}}}\|^2$. It can be decomposed as

$$\|\boldsymbol{\xi}_{\boldsymbol{\epsilon}}\|^2 - \|\boldsymbol{\xi}_{\underline{\boldsymbol{\epsilon}}}\|^2 = \|\boldsymbol{\xi}_{\boldsymbol{\epsilon}}\|^2 - \|\boldsymbol{\xi}\|^2 + \|\boldsymbol{\xi}\|^2 - \|\boldsymbol{\xi}_{\underline{\boldsymbol{\epsilon}}}\|^2$$

with $\boldsymbol{\xi} = D_0^{-1}\nabla L(\boldsymbol{\theta}^*)$. If the values δ, ϱ are small then the difference $\|\boldsymbol{\xi}_{\boldsymbol{\epsilon}}\|^2 - \|\boldsymbol{\xi}_{\underline{\boldsymbol{\epsilon}}}\|^2$ is automatically smaller than $\|\boldsymbol{\xi}\|^2$.

Lemma 3.9. *Suppose* $(\mathcal{I})$ *and let* $\tau_{\boldsymbol{\epsilon}} \stackrel{\text{def}}{=} \delta + \varrho\mathfrak{a}^2 < 1$. *Then*

$$D_{\boldsymbol{\epsilon}}^2 \ge (1-\tau_{\boldsymbol{\epsilon}})D_0^2, \qquad D_{\underline{\boldsymbol{\epsilon}}}^2 \le (1+\tau_{\boldsymbol{\epsilon}})D_0^2,$$
$$\|\boldsymbol{I}_p - D_{\boldsymbol{\epsilon}} D_{\underline{\boldsymbol{\epsilon}}}^{-2} D_{\boldsymbol{\epsilon}}\|_\infty \le \alpha_{\boldsymbol{\epsilon}} \stackrel{\text{def}}{=} \frac{2\tau_{\boldsymbol{\epsilon}}}{1-\tau_{\boldsymbol{\epsilon}}^2}. \tag{3.13}$$

Moreover,

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

Our final step is in showing that under (ED_0), the norm $\|\boldsymbol{\xi}\|$ behaves essentially as a norm of a Gaussian vector with the same covariance matrix. Define for $\mathbb{B} \stackrel{\text{def}}{=} D_0^{-1} V_0^2 D_0^{-1}$

$$\mathtt{p}_0 \stackrel{\text{def}}{=} \operatorname{tr}(\mathbb{B}), \qquad \mathtt{v}^2 \stackrel{\text{def}}{=} 2\operatorname{tr}(\mathbb{B}^2), \qquad \lambda_0 \stackrel{\text{def}}{=} \|\mathbb{B}\|_\infty = \lambda_{\max}(\mathbb{B}).$$

Under the identifiability condition $(\mathcal{I})$, one can bound

$$\mathbb{B}^2 \le \mathfrak{a}^2 I_p, \qquad \mathtt{p}_0 \le \mathfrak{a}^2 p, \qquad \mathtt{v}^2 \le 2\mathfrak{a}^4 p, \qquad \lambda_0 \le \mathfrak{a}^2.$$

Similarly to the previous result, we assume that the constant $\mathtt{g}$ from condition (ED_0) is sufficiently large, namely $\mathtt{g}^2 \ge 2\mathtt{p}_0$. Define $\mu_c = 2/3$ and

$$\begin{aligned}
\mathtt{y}_c^2 &\stackrel{\text{def}}{=} \mathtt{g}^2/\mu_c^2 - \mathtt{p}_0/\mu_c, \\
\mathtt{g}_c &\stackrel{\text{def}}{=} \mu_c \mathtt{y}_c = \sqrt{\mathtt{g}^2 - \mu_c \mathtt{p}_0}, \\
2\mathtt{x}_c &\stackrel{\text{def}}{=} \mu_c \mathtt{y}_c^2 + \log\det\bigl(I_p - \mu_c \mathbb{B}^2/\lambda_0\bigr).
\end{aligned}$$

It is easy to see that $\mathtt{y}_c^2 \ge 3\mathtt{g}^2/2$ and $\mathtt{g}_c \ge \sqrt{2/3}\,\mathtt{g}$.

THEOREM 3.10. *Let (ED_0) hold with $\nu_0 = 1$ and $\mathtt{g}^2 \ge 2\mathtt{p}_0$. Then $\mathbb{E}\|\boldsymbol{\xi}\|^2 \le \mathtt{p}_0$, and for each $\mathtt{x} \le \mathtt{x}_c$*

$$\mathbb{P}\bigl(\|\boldsymbol{\xi}\|^2/\lambda_0 \ge \mathfrak{z}(\mathtt{x}, \mathbb{B})\bigr) \le 2\mathrm{e}^{-\mathtt{x}} + 8.4\mathrm{e}^{-\mathtt{x}_c}, \tag{3.14}$$

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

$$\mathfrak{z}(\mathtt{x}, \mathbb{B}) \stackrel{\text{def}}{=} \begin{cases} \mathtt{p}_0 + 2\mathtt{v}\mathtt{x}^{1/2}, & \mathtt{x} \le \mathtt{v}/18, \\ \mathtt{p}_0 + 6\mathtt{x} & \mathtt{v}/18 < \mathtt{x} \le \mathtt{x}_c, \end{cases}$$

Moreover, for $\mathtt{x} > \mathtt{x}_c$, it holds with $\mathfrak{z}(\mathtt{x}, \mathbb{B}) = \bigl|\mathtt{y}_c + 2(\mathtt{x} - \mathtt{x}_c)/\mathtt{g}_c\bigr|^2$

$$\mathbb{P}\bigl(\|\boldsymbol{\xi}\|^2/\lambda_0 \ge \mathfrak{z}(\mathtt{x}, \mathbb{B})\bigr) \le 8.4\mathrm{e}^{-\mathtt{x}}.$$

PROOF. It follows from condition (ED_0) that

$$\begin{aligned}
\mathbb{E}\|\boldsymbol{\xi}\|^2 &= \mathbb{E}\operatorname{tr}\boldsymbol{\xi}\boldsymbol{\xi}^\top \\
&= \operatorname{tr} D_0^{-1}\bigl[\mathbb{E}\nabla L(\boldsymbol{\theta}^*)\{\nabla L(\boldsymbol{\theta}^*)\}^\top\bigr] D_0^{-1} = \operatorname{tr}\bigl[D_0^{-2}\operatorname{Var}\bigl\{\nabla L(\boldsymbol{\theta}^*)\bigr\}\bigr]
\end{aligned}$$

and (ED_0) implies $\boldsymbol{\gamma}^\top \operatorname{Var}\bigl\{\nabla L(\boldsymbol{\theta}^*)\bigr\}\boldsymbol{\gamma} \le \boldsymbol{\gamma}^\top V_0^2 \boldsymbol{\gamma}$ and thus, $\mathbb{E}\|\boldsymbol{\xi}\|^2 \le \mathtt{p}_0$. The deviation bound (3.14) is proved in Corollary **??**. □

REMARK 3.3. This small remark concerns the term $8.4\mathrm{e}^{-\mathtt{x}_c}$ in the probability bound (3.14). As already mentioned, this bound implicitly assumes that the constant $\mathtt{g}$ is large (usually $\mathtt{g} \asymp n^{1/2}$). Then $\mathtt{x}_c \asymp \mathtt{g}^2 \asymp n$ is large as well. So, $\mathrm{e}^{-\mathtt{x}_c}$ is very small and asymptotically negligible. Below we often ignore this term. For $\mathtt{x} \le \mathtt{x}_c$, we can use $\mathfrak{z}(\mathtt{x}, I\!\!B) = \mathtt{p}_0 + 6\mathtt{x}$.

REMARK 3.4. The exponential bound of Theorem 3.10 helps to describe the critical value of $\mathfrak{z}$ ensuring a prescribed deviation probability $I\!\!P\bigl(\|\boldsymbol{\xi}\|^2 \ge \mathfrak{z}\bigr)$. Namely, this probability starts to gradually decrease when $\mathfrak{z}$ grows over $\lambda_0 \mathtt{p}_0$. In particular, this helps to answer a very important question about a proper choice of the critical value $\mathfrak{z}$ providing the prescribed covering probability, or of the value z ensuring the dominating concentration probability $I\!\!P\bigl(\|D_{\boldsymbol{\epsilon}}(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\| \le z\bigr)$.

The definition of the set $C_{\boldsymbol{\epsilon}}(\mathtt{r})$ from (3.8) involves the event $\{\|V_0 D_{\underline{\boldsymbol{\epsilon}}}^{-1} \boldsymbol{\xi}_{\underline{\boldsymbol{\epsilon}}}\| > \mathtt{r}\}$. Under $(\mathcal{I})$, it is included in the set $\{\|\boldsymbol{\xi}_{\underline{\boldsymbol{\epsilon}}}\| > (1+\alpha_{\boldsymbol{\epsilon}})^{-1} \mathtt{a}^{-1} \mathtt{r}\}$, see (3.13), and its probability is of order $\mathrm{e}^{-\mathtt{x}}$ for $\mathtt{r}^2 \ge C(\mathtt{x} + p)$ with a fixed $C > 0$.

By Theorem 3.7, one can use $\max\bigl\{\diamondsuit_{\boldsymbol{\epsilon}}(\mathtt{r}), \diamondsuit_{\underline{\boldsymbol{\epsilon}}}(\mathtt{r})\bigr\} \le \varrho\, \mathfrak{z}_0(\mathtt{x}, \mathbb{Q})$ on a set of probability at least $1 - 2\mathrm{e}^{-\mathtt{x}}$. Further, $\|\boldsymbol{\xi}\|^2/\lambda_0 \le \mathfrak{z}(\mathtt{x}, I\!\!B)$ with a probability of order $1 - 2\mathrm{e}^{-\mathtt{x}}$; see (3.14). Putting together the obtained bounds yields for the spread $\Delta_{\boldsymbol{\epsilon}}(\mathtt{r})$ with a probability about $1 - 4\mathrm{e}^{-\mathtt{x}}$

$$\Delta_{\boldsymbol{\epsilon}}(\mathtt{r}) \le 2\varrho\, \mathfrak{z}_0(\mathtt{x}, \mathbb{Q}) + \alpha_{\boldsymbol{\epsilon}}\, \lambda_0\, \mathfrak{z}(\mathtt{x}, I\!\!B).$$

The results obtained in Section 3.2 are sharp and meaningful if the spread $\Delta_{\boldsymbol{\epsilon}}(\mathtt{r})$ is smaller in order than the value $I\!\!E\|\boldsymbol{\xi}\|^2$. Theorem 3.10 states that $\|\boldsymbol{\xi}\|^2$ does not significantly deviate over its expected value $\mathtt{p}_0 \stackrel{\mathrm{def}}{=} I\!\!E\|\boldsymbol{\xi}\|^2$ which is our leading term. We know that $\mathfrak{z}_0(\mathtt{x}, \mathbb{Q}) \approx \mathbb{Q} + \mathtt{x} = \mathfrak{c}_1 p + \mathtt{x}$ if $\mathtt{x}$ is not too large. Also $\mathfrak{z}(\mathtt{x}, I\!\!B) \le \mathtt{p}_0 + 6\mathtt{x}$, where $\mathtt{p}_0$ is of order p due to $(\mathcal{I})$. Summarizing the above discussion yields that the local results apply if the regularity condition $(\mathcal{I})$ holds and the values ϱ and $\alpha_{\boldsymbol{\epsilon}}$, or equivalently, $\omega(\mathtt{r}), \delta(\mathtt{r})$ are small. In Section 5 we show for the i.i.d. example that $\omega(\mathtt{r}) \asymp \sqrt{\mathtt{r}^2/n}$ and similarly for $\delta(\mathtt{r})$.

3.4. *Proof of Corollaries 3.3 and 3.4.* The bound (3.4) together with Lemma 3.8 yield on $C_{\boldsymbol{\epsilon}}(\mathtt{r})$

$$\begin{aligned} L(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}^*) &= \sup_{\boldsymbol{\theta} \in \Theta_0(\mathtt{r})} L(\boldsymbol{\theta}, \boldsymbol{\theta}^*) \\ &\ge \sup_{\boldsymbol{\theta} \in \Theta_0(\mathtt{r})} \mathbb{L}_{\underline{\boldsymbol{\epsilon}}}(\boldsymbol{\theta}, \boldsymbol{\theta}^*) - \diamondsuit_{\underline{\boldsymbol{\epsilon}}}(\mathtt{r}) = \|\boldsymbol{\xi}_{\underline{\boldsymbol{\epsilon}}}\|^2/2 - \diamondsuit_{\underline{\boldsymbol{\epsilon}}}(\mathtt{r}). \qquad (3.15) \end{aligned}$$

Similarly

$$L(\widetilde{\boldsymbol{\theta}},\boldsymbol{\theta}^*) \le \sup_{\boldsymbol{\theta}\in\Theta_0(\mathtt{r})} \mathbb{L}_{\boldsymbol{\epsilon}}(\boldsymbol{\theta},\boldsymbol{\theta}^*) + \diamondsuit_{\boldsymbol{\epsilon}}(\mathtt{r}) \le \|\boldsymbol{\xi}_{\boldsymbol{\epsilon}}\|^2/2 + \diamondsuit_{\boldsymbol{\epsilon}}(\mathtt{r})$$

yielding (3.9). For getting (3.10), we again apply the inequality $L(\boldsymbol{\theta},\boldsymbol{\theta}^*) \le \mathbb{L}_{\boldsymbol{\epsilon}}(\boldsymbol{\theta},\boldsymbol{\theta}^*)+\diamondsuit_{\boldsymbol{\epsilon}}(\mathtt{r})$ from Theorem 3.1 for $\boldsymbol{\theta}$ equal to $\widetilde{\boldsymbol{\theta}}$. With $\boldsymbol{\xi}_{\boldsymbol{\epsilon}} = D_{\boldsymbol{\epsilon}}^{-1}\nabla L(\boldsymbol{\theta}^*)$ and $\boldsymbol{u}_{\boldsymbol{\epsilon}} \stackrel{\text{def}}{=} D_{\boldsymbol{\epsilon}}(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)$, this gives

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

Therefore, by (3.15)

$$\|\boldsymbol{\xi}_{\underline{\boldsymbol{\epsilon}}}\|^2/2 - \diamondsuit_{\underline{\boldsymbol{\epsilon}}}(\mathtt{r}) - \boldsymbol{\xi}_{\boldsymbol{\epsilon}}^{\top}\boldsymbol{u}_{\boldsymbol{\epsilon}} + \|\boldsymbol{u}_{\boldsymbol{\epsilon}}\|^2/2 \le \diamondsuit_{\boldsymbol{\epsilon}}(\mathtt{r})$$

or, equivalently

$$\|\boldsymbol{\xi}_{\boldsymbol{\epsilon}}\|^2/2 - \boldsymbol{\xi}_{\boldsymbol{\epsilon}}^{\top}\boldsymbol{u}_{\boldsymbol{\epsilon}} + \|\boldsymbol{u}_{\boldsymbol{\epsilon}}\|^2/2 \le \diamondsuit_{\boldsymbol{\epsilon}}(\mathtt{r}) + \diamondsuit_{\underline{\boldsymbol{\epsilon}}}(\mathtt{r}) + \bigl(\|\boldsymbol{\xi}_{\boldsymbol{\epsilon}}\|^2 - \|\boldsymbol{\xi}_{\underline{\boldsymbol{\epsilon}}}\|^2\bigr)/2$$

and the definition of $\Delta_{\boldsymbol{\epsilon}}(\mathtt{r})$ implies $\bigl\|\boldsymbol{u}_{\boldsymbol{\epsilon}} - \boldsymbol{\xi}_{\boldsymbol{\epsilon}}\bigr\|^2 \le 2\Delta_{\boldsymbol{\epsilon}}(\mathtt{r})$.

4. Upper function approach and concentration of the qMLE. A very important step in the analysis of the qMLE $\widetilde{\boldsymbol{\theta}}$ is *localization*. This property means that $\widetilde{\boldsymbol{\theta}}$ concentrates in a small vicinity of the central point $\boldsymbol{\theta}^*$. This section states such a concentration bound under the global conditions of Section 2. Given $\mathtt{r}_0$, the deviation bound describes the probability $\mathbb{P}\bigl(\widetilde{\boldsymbol{\theta}} \notin \Theta_0(\mathtt{r}_0)\bigr)$ that $\widetilde{\boldsymbol{\theta}}$ does not belong to the local vicinity $\Theta_0(\mathtt{r}_0)$ of Θ. The question of interest is to check a possibility of selecting $\mathtt{r}_0$ in a way that the local bracketing result and the deviation bound apply simultaneously; see the discussion at the end of the section.

Below we suppose that a sufficiently large constant $\mathtt{x}$ is fixed to specify the accepted level be of order $\mathrm{e}^{-\mathtt{x}}$ for this deviation probability. All the constructions below depend upon this constant. We do not indicate it explicitly for ease of notation.

The key step in this large deviation bound is made in terms of an *upper function* for the process $L(\boldsymbol{\theta},\boldsymbol{\theta}^*) \stackrel{\text{def}}{=} L(\boldsymbol{\theta}) - L(\boldsymbol{\theta}^*)$. Namely, $\mathfrak{u}(\boldsymbol{\theta})$ is a deterministic *upper function* if it holds with a high probability:

$$\sup_{\boldsymbol{\theta}\in\Theta}\bigl\{L(\boldsymbol{\theta},\boldsymbol{\theta}^*) + \mathfrak{u}(\boldsymbol{\theta})\bigr\} \le 0 \tag{4.1}$$

Such bounds are usually called for in the analysis of the posterior measure in the Bayes approach. Below we present sufficient conditions ensuring (4.1).

Now we explain how the concentration bounds can be derived from (4.1). Let $\mathfrak{u}(\boldsymbol{\theta})$ be an upper function. It can be used for describing the *concentration sets* for $\widetilde{\boldsymbol{\theta}}$.

LEMMA 4.1. *Let* $\mathfrak{u}(\boldsymbol{\theta})$ *be an upper function in the sense*

$$I\!P\Bigl(\sup_{\boldsymbol{\theta}\in\Theta}\bigl\{L(\boldsymbol{\theta},\boldsymbol{\theta}^*)+\mathfrak{u}(\boldsymbol{\theta})\bigr\}\ge 0\Bigr)\le \mathrm{e}^{-\mathtt{x}} \tag{4.2}$$

for $\mathtt{x}>0$. *Given a subset* $\Theta_0\subset\Theta$ *with* $\boldsymbol{\theta}^*\in\Theta_0$, *the condition* $\mathfrak{u}(\boldsymbol{\theta})\ge 0$ *for* $\boldsymbol{\theta}\notin\Theta_0$ *ensures*

$$I\!P\bigl(\widetilde{\boldsymbol{\theta}}\notin\Theta_0\bigr)\le \mathrm{e}^{-\mathtt{x}}.$$

PROOF. If Θ° is a subset of Θ not containing $\boldsymbol{\theta}^*$, then the event $\widetilde{\boldsymbol{\theta}}\in\Theta^\circ$ is only possible if $\sup_{\boldsymbol{\theta}\in\Theta^\circ}L(\boldsymbol{\theta},\boldsymbol{\theta}^*)\ge 0$, because $L(\boldsymbol{\theta}^*,\boldsymbol{\theta}^*)\equiv 0$. This yields the result. □

A possible way of checking the condition (4.2) is based on a lower quadratic bound for the negative expectation $-I\!E L(\boldsymbol{\theta},\boldsymbol{\theta}^*)\ge \mathtt{b}(\mathtt{r})\|V_0(\boldsymbol{\theta}-\boldsymbol{\theta}^*)\|^2/2$ in the sense of condition $(\mathcal{L}\mathtt{r})$ from Section 2.2. We present two different results. The first one assumes that the values $\mathtt{b}(\mathtt{r})$ can be fixed universally for all $\mathtt{r}\ge\mathtt{r}_0$.

THEOREM 4.2. *Suppose* $(E\mathtt{r})$ *and* $(\mathcal{L}\mathtt{r})$ *with* $\mathtt{b}(\mathtt{r})\equiv\mathtt{b}$. *Let, for* $\mathtt{r}\ge\mathtt{r}_0$,

$$1+\sqrt{\mathtt{x}+\mathbb{Q}}\le 3\nu_0^2\mathtt{g}(\mathtt{r})/\mathtt{b}, \tag{4.3}$$

$$6\nu_0\sqrt{\mathtt{x}+\mathbb{Q}}\le \mathtt{r}\mathtt{b}, \tag{4.4}$$

with $\mathtt{x}+\mathbb{Q}\ge 2.5$ *and* $\mathbb{Q}=\mathfrak{c}_1 p$. *Then*

$$I\!P\bigl(\widetilde{\boldsymbol{\theta}}\notin\Theta_0(\mathtt{r}_0)\bigr)\le \mathrm{e}^{-\mathtt{x}}. \tag{4.5}$$

PROOF. The result follows from Theorem **??** with $\mu=\frac{\mathtt{b}}{3\nu_0}$, $\mathfrak{t}(\mu)\equiv 0$, $\mathcal{U}(\boldsymbol{\theta})=L(\boldsymbol{\theta})-I\!E L(\boldsymbol{\theta})$ and $M(\boldsymbol{\theta},\boldsymbol{\theta}^*)=-I\!E L(\boldsymbol{\theta},\boldsymbol{\theta}^*)\ge\frac{\mathtt{b}}{2}\|V_0(\boldsymbol{\theta}-\boldsymbol{\theta}^*)\|^2$. □

REMARK 4.1. The bound (4.5) requires only two conditions. Condition (4.3) means that the value $\mathtt{g}(\mathtt{r})$ from condition $(E\mathtt{r})$ fulfills $\mathtt{g}^2(\mathtt{r})\ge C(\mathtt{x}+p)$, that is, we need a qualified rate in the exponential moment conditions. This is similar to requiring finite polynomial moments for the score function. Condition (4.4) requires that $\mathtt{r}$ exceeds some fixed value, namely, $\mathtt{r}^2\ge C(\mathtt{x}+p)$. This bound is helpful for fixing the value $\mathtt{r}_0$ providing a sensible deviation probability bound.

If $\mathtt{b}(\mathtt{r})$ decreases with $\mathtt{r}$, the result is a bit more involved. The key requirement is that $\mathtt{b}(\mathtt{r})$ decreases not too fast, so that the product $\mathtt{r}\mathtt{b}(\mathtt{r})$ grows to infinity with $\mathtt{r}$. The idea is to include the complement of the central set Θ_0 in Θ in the union of the growing sets $\Theta_0(\mathtt{r}_k)$ with $\mathtt{b}(\mathtt{r}_k) \geq \mathtt{b}(\mathtt{r}_0)2^{-k}$, and then apply Theorem 4.2 for each $\Theta_0(\mathtt{r}_k)$.

THEOREM 4.3. *Suppose* $(E\mathtt{r})$ *and* $(\mathcal{L}\mathtt{r})$. *Let* $\mathtt{r}_k$ *be such that* $\mathtt{b}(\mathtt{r}_k) \geq \mathtt{b}(\mathtt{r}_0)2^{-k}$ *for* $k \geq 1$. *If the conditions*

$$1 + \sqrt{\mathtt{x} + \mathbb{Q} + ck} \leq 3\nu_0^2 \mathtt{g}(\mathtt{r}_k)/\mathtt{b}(\mathtt{r}_k),$$
$$6\nu_0 \sqrt{\mathtt{x} + \mathbb{Q} + ck} \leq \mathtt{r}_k \mathtt{b}(\mathtt{r}_k),$$

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

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

PROOF. The result (4.5) is applied to each set $\Theta_0(\mathtt{r}_k)$ and $\mathtt{x}_k = \mathtt{x} + ck$. This yields

$$I\!P\bigl(\widetilde{\boldsymbol{\theta}} \notin \Theta_0(\mathtt{r}_0)\bigr) \leq \sum_{k \geq 1} I\!P\bigl(\widetilde{\boldsymbol{\theta}} \notin \Theta_0(\mathtt{r}_k)\bigr) \leq \sum_{k \geq 1} \mathrm{e}^{-\mathtt{x}+ck} = \mathrm{e}^{-\mathtt{x}}$$

as required. □

REMARK 4.2. Here we briefly discuss the very important question: how one can fix the value $\mathtt{r}_0$ ensuring the bracketing result in the local set $\Theta_0(\mathtt{r}_0)$ and a small probability of the related set $C_{\boldsymbol{\epsilon}}(\mathtt{r})$ from (3.8)? The event $\{\|V_0 D_{\underline{\boldsymbol{\epsilon}}}^{-1} \boldsymbol{\xi}_{\underline{\boldsymbol{\epsilon}}}\| > \mathtt{r}\}$ requires $\mathtt{r} \geq C(\mathtt{x} + p)$. Further we inspect the deviation bound for the complement $\Theta \backslash \Theta_0(\mathtt{r}_0)$. For simplicity, assume $(\mathcal{L}\mathtt{r})$ with $\mathtt{b}(\mathtt{r}) \equiv \mathtt{b}$. Then the most important condition (4.4) of Theorem 4.2 requires that $\mathtt{r}^2 \geq C\mathtt{b}^{-2}(\mathtt{x} + p)$. In words, the squared radius $\mathtt{r}^2$ should be of order p. The other condition (4.3) of Theorem 4.2 is technical and only requires that $\mathtt{g}(\mathtt{r})$ is sufficiently large while the local results only require that $\delta(\mathtt{r})$ and $\varrho(\mathtt{r})$ are small for such $\mathtt{r}$. In the asymptotic setup one can typically bring these conditions together. Section 5 provides further discussion for the i.i.d. setup.

5. Examples. The model with independent identically distributed (i.i.d.) observations is one of the most popular setups in statistical literature and in statistical applications. The essential and the most developed part of the

statistical theory is designed for the i.i.d. modeling. Especially, the classical asymptotic parametric theory is almost complete including asymptotic root-n normality and efficiency of the MLE and Bayes estimators under rather mild assumptions; see e.g. Chapter 2 and 3 in Ibragimov and Khas'minskij (1981). So, the i.i.d. model can naturally serve as a benchmark for any extension of the statistical theory: being applied to the i.i.d. setup, the new approach should lead to essentially the same conclusions as in the classical theory. Similar reasons apply to the regression model and its extensions. Below we try demonstrate that the proposed non-asymptotic viewpoint is able to reproduce the existing brilliant and well established results of the classical parametric theory. Surprisingly, the majority of classical efficiency results can be easily derived from the obtained general non-asymptotic bounds.

The next question is whether there is any added value or benefits of the new approach being restricted to the i.i.d. situation relative to the classical one. Two important issues have been already mentioned: the new approach applies to the situation with finite samples and survives under model misspecification. One more important question is whether the obtained results remain applicable and informative if the dimension of the parameter space is high – this is one of the main challenge in the modern statistics. We show that the dimensionality p naturally appears in the risk bounds and the results apply as long as the sample size exceeds in order this value p. All these questions are addressed in Section 5.1 for the i.i.d. setup, Section 5.2 focuses on generalized linear modeling, while Section 5.3 discusses linear median regression.

5.1. *Quasi MLE in an i.i.d. model.* The basic i.i.d. parametric model means that the observations $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ are independent identically distributed from a distribution P which belongs to a given parametric family $(P_{\boldsymbol{\theta}}, \boldsymbol{\theta} \in \Theta)$ on the observation space $\mathcal{Y}_1$. Each $\boldsymbol{\theta} \in \Theta$ clearly yields the product data distribution $I\!P_{\boldsymbol{\theta}} = P_{\boldsymbol{\theta}}^{\otimes n}$ on the product space $\mathcal{Y} = \mathcal{Y}_1^n$. This section illustrates how the obtained general results can be applied to this type of modeling under possible model misspecification. Different types of misspecification can be considered. Each of the assumptions, namely, data independence, identical distribution, parametric form of the marginal distribution can be violated. To be specific, we assume the observations Y_i independent and identically distributed. However, we admit that the distribution of each Y_i does not necessarily belong to the parametric family $(P_{\boldsymbol{\theta}})$. The case of non-identically distributed observations can be done similarly at cost of more complicated notation.

In what follows the parametric family $(P_{\boldsymbol{\theta}})$ is supposed to be dominated

by a measure μ_0, and each density $p(y,\boldsymbol{\theta}) = dP_{\boldsymbol{\theta}}/d\mu_0(y)$ is two times continuously differentiable in $\boldsymbol{\theta}$ for all y. Denote $\ell(y,\boldsymbol{\theta}) = \log p(y,\boldsymbol{\theta})$. The parametric assumption $Y_i \sim P_{\boldsymbol{\theta}^*} \in (P_{\boldsymbol{\theta}})$ leads to the log-likelihood

$$L(\boldsymbol{\theta}) = \sum \ell(Y_i,\boldsymbol{\theta}),$$

where the summation is taken over $i = 1,\ldots,n$. The quasi MLE $\widetilde{\boldsymbol{\theta}}$ maximizes this sum over $\boldsymbol{\theta} \in \Theta$:

$$\widetilde{\boldsymbol{\theta}} \stackrel{\text{def}}{=} \operatorname*{argmax}_{\boldsymbol{\theta}\in\Theta} L(\boldsymbol{\theta}) = \operatorname*{argmax}_{\boldsymbol{\theta}\in\Theta} \sum \ell(Y_i,\boldsymbol{\theta}).$$

The target of estimation $\boldsymbol{\theta}^*$ maximizes the expectation of $L(\boldsymbol{\theta})$:

$$\boldsymbol{\theta}^* \stackrel{\text{def}}{=} \operatorname*{argmax}_{\boldsymbol{\theta}\in\Theta} I\!E L(\boldsymbol{\theta}) = \operatorname*{argmax}_{\boldsymbol{\theta}\in\Theta} I\!E \ell(Y_1,\boldsymbol{\theta}).$$

Let $\zeta_i(\boldsymbol{\theta}) \stackrel{\text{def}}{=} \ell(Y_i,\boldsymbol{\theta}) - I\!E\ell(Y_i,\boldsymbol{\theta})$. Then $\zeta(\boldsymbol{\theta}) = \sum \zeta_i(\boldsymbol{\theta})$. The equation $\nabla I\!E L(\boldsymbol{\theta}^*) = 0$ implies

$$\nabla\zeta(\boldsymbol{\theta}^*) = \sum \nabla\zeta_i(\boldsymbol{\theta}^*) = \sum \nabla\ell_i(\boldsymbol{\theta}^*). \tag{5.1}$$

I.i.d. structure of the Y_i's allows to rewrite the local conditions $(E\mathtt{r})$, (ED_0), (ED_1), and $(\mathcal{L}_0)$, and $(\mathcal{I})$ in terms of the marginal distribution.

(ed_0) *There exists a positively definite symmetric matrix $\mathbf{v}_0$, such that for all $|\lambda| \le \mathtt{g}_1$*

$$\sup_{\boldsymbol{\gamma}\in I\!R^p} \log I\!E \exp\biggl\{\lambda \frac{\boldsymbol{\gamma}^\top \nabla\zeta_1(\boldsymbol{\theta}^*)}{\|\mathbf{v}_0\boldsymbol{\gamma}\|}\biggr\} \le \nu_0^2\lambda^2/2.$$

A natural candidate on $\mathbf{v}_0^2$ is given by the variance of the gradient $\nabla\ell(Y_1,\boldsymbol{\theta}^*)$, that is, $\mathbf{v}_0^2 = \operatorname{Var}\bigl\{\nabla\ell(Y_1,\boldsymbol{\theta}^*)\bigr\} = \operatorname{Var}\bigl\{\nabla\zeta_1(\boldsymbol{\theta}^*)\bigr\}$.

Next consider the local sets

$$\Theta_{\text{loc}}(\mathtt{u}) = \{\boldsymbol{\theta} : \|\mathbf{v}_0(\boldsymbol{\theta}-\boldsymbol{\theta}^*)\| \le \mathtt{u}\}.$$

In view of $V_0^2 = n\mathbf{v}_0^2$, it holds $\Theta_0(\mathtt{r}) = \Theta_{\text{loc}}(\mathtt{u})$ with $\mathtt{r}^2 = n\mathtt{u}^2$.

Below we distinguish between local conditions for $\mathtt{u} \le \mathtt{u}_0$ and the global conditions for all $\mathtt{u} > 0$, where $\mathtt{u}_0$ is some fixed value.

The local smoothness conditions (ED_1) and $(\mathcal{L}_0)$ require to specify the functions $\delta(\mathtt{r})$ and $\varrho(\mathtt{r})$ for $\mathtt{r} \le \mathtt{r}_0$ where $\mathtt{r}_0^2 = n\mathtt{u}_0^2$. If the log-likelihood function $\ell(y,\boldsymbol{\theta})$ is sufficiently smooth in $\boldsymbol{\theta}$, these functions can be selected proportional to $\mathtt{u} = \mathtt{r}/n^{1/2}$.

(ed_1) *There are constants* $\omega^* > 0$ *and* $\mathtt{g}_1 > 0$ *such that for each* $\mathtt{u} \le \mathtt{u}_0$ *and* $|\lambda| \le \mathtt{g}_1$

$$\sup_{\gamma \in I\!R^p} \sup_{\boldsymbol{\theta} \in \Theta_{\mathrm{loc}}(\mathtt{u})} \log I\!E \exp\Bigl\{ \lambda \frac{\boldsymbol{\gamma}^\top \bigl[\nabla \zeta_1(\boldsymbol{\theta}) - \nabla \zeta_1(\boldsymbol{\theta}^*) \bigr]}{\omega^* \, \mathtt{u} \, \| \mathbf{v}_0 \boldsymbol{\gamma} \|} \Bigr\} \le \nu_0^2 \lambda^2 / 2.$$

Further we restate the local identifiability condition $(\mathcal{L}_0)$ in terms of the expected value $\Bbbk(\boldsymbol{\theta}, \boldsymbol{\theta}^*) \stackrel{\mathrm{def}}{=} -I\!E\bigl\{ \ell(Y_i, \boldsymbol{\theta}) - \ell(Y_i, \boldsymbol{\theta}^*) \bigr\}$ for each i. We suppose that $\Bbbk(\boldsymbol{\theta}, \boldsymbol{\theta}^*)$ is two times differentiable w.r.t. $\boldsymbol{\theta}$. The definition of $\boldsymbol{\theta}^*$ implies $\nabla I\!E \ell(Y_i, \boldsymbol{\theta}^*) = 0$. Define also the matrix $\mathbf{f}_0 = -\nabla^2 I\!E \ell(Y_i, \boldsymbol{\theta}^*)$. In the parametric case $P = P_{\boldsymbol{\theta}^*}$, $\Bbbk(\boldsymbol{\theta}, \boldsymbol{\theta}^*)$ is the *Kullback-Leibler divergence* between $P_{\boldsymbol{\theta}^*}$ and $P_{\boldsymbol{\theta}}$ while the matrices $\mathbf{v}_0^2 = \mathbf{f}_0$ are equal to each other and coincide with the *Fisher information matrix* of the family $(P_{\boldsymbol{\theta}})$ at $\boldsymbol{\theta}^*$.

(ℓ_0) *There is a constant* δ^* *such that it holds for each* $\mathtt{u} \le \mathtt{u}_0$

$$\sup_{\boldsymbol{\theta} \in \Theta_{\mathrm{loc}}(\mathtt{u})} \Bigl| \frac{2 \Bbbk(\boldsymbol{\theta}, \boldsymbol{\theta}^*)}{(\boldsymbol{\theta} - \boldsymbol{\theta}^*)^\top \mathbf{f}_0 \, (\boldsymbol{\theta} - \boldsymbol{\theta}^*)} - 1 \Bigr| \le \delta^* \mathtt{u}.$$

(ι) There is a constant $\mathtt{a} > 0$ such that $\mathtt{a}^2 \mathbf{f}_0^2 \ge \mathbf{v}_0^2$.

($e\mathtt{u}$) *For each* $\mathtt{u} > 0$, *there exists* $\mathtt{g}_1(\mathtt{u}) > 0$, *such that for all* $|\lambda| \le \mathtt{g}_1(\mathtt{u})$

$$\sup_{\gamma \in I\!R^p} \sup_{\boldsymbol{\theta} \in \Theta_{\mathrm{loc}}(\mathtt{u})} \log I\!E \exp\Bigl\{ \lambda \frac{\boldsymbol{\gamma}^\top \nabla \zeta_1(\boldsymbol{\theta})}{\| \mathbf{v}_0 \boldsymbol{\gamma} \|} \Bigr\} \le \nu_0^2 \lambda^2 / 2.$$

($\ell\mathtt{u}$) *For each* $\mathtt{u} > 0$, *there exists* $\mathtt{b}(\mathtt{u}) > 0$ *such that*

$$\sup_{\boldsymbol{\theta} \in \Theta : \| \mathbf{v}_0 (\boldsymbol{\theta} - \boldsymbol{\theta}^*) \| = \mathtt{u}} \frac{\Bbbk(\boldsymbol{\theta}, \boldsymbol{\theta}^*)}{\| \mathbf{v}_0 (\boldsymbol{\theta} - \boldsymbol{\theta}^*) \|^2} \ge \mathtt{b}(\mathtt{u}),$$

LEMMA 5.1. *Let* $Y_1, \ldots, Y_n$ *be i.i.d. Then* $(e\mathtt{u})$, (ed_0), (ed_1), (ι), *and* (ℓ_0) *imply* $(E\mathtt{r})$, (ED_0), (ED_1), $(\mathcal{I})$, *and* $(\mathcal{L}_0)$ *with* $V_0^2 = n \mathbf{v}_0^2$, $D_0^2 = n \mathbf{f}_0$, $\omega(\mathtt{r}) = \omega^* \mathtt{r} / n^{1/2}$, $\delta(\mathtt{r}) = \delta^* \mathtt{r} / n^{1/2}$, *and* $\mathtt{g} = \mathtt{g}_1 \sqrt{n}$.

PROOF. The identities $V_0^2 = n \mathbf{v}_0^2$, $D_0^2 = n \mathbf{f}_0$ follow from the i.i.d. structure of the observations Y_i. We briefly comment on condition $(E\mathtt{r})$. The use once again the i.i.d. structure yields by (5.1) in view of $V_0^2 = n \mathbf{v}_0^2$

$$\log I\!E \exp\Bigl\{ \lambda \frac{\boldsymbol{\gamma}^\top \nabla \zeta(\boldsymbol{\theta})}{\| V_0 \boldsymbol{\gamma} \|} \Bigr\} = n I\!E \exp\Bigl\{ \frac{\lambda}{n^{1/2}} \frac{\boldsymbol{\gamma}^\top \nabla \zeta_1(\boldsymbol{\theta})}{\| \mathbf{v}_0 \boldsymbol{\gamma} \|} \Bigr\} \le \nu_0^2 \lambda^2 / 2$$

as long as $\lambda \le n^{1/2} \mathtt{g}_1(\mathtt{u}) \le \mathtt{g}(\mathtt{r})$. Similarly for (ED_0) and (ED_1). $\square$

REMARK 5.1. This remark discusses how the presented conditions relate to what is usually assumed in statistical literature. One general remarks concern the choice of the parametric family $(P_{\boldsymbol{\theta}})$. The point of the classical theory is that the true measure is in this family, so the conditions should be as weak as possible. The viewpoint of this paper is slightly different: whatever family $(P_{\boldsymbol{\theta}})$ is taken, the true measure is never included, any model is only an approximation of reality. From the other side, the choice of the parametric model $(P_{\boldsymbol{\theta}})$ is always done by a statistician. Sometime some special stylized features of the model force to include an irregularity in this family. Otherwise any smoothness condition on the density $\ell(y, \boldsymbol{\theta})$ can be secured by a proper choice of the family $(P_{\boldsymbol{\theta}})$.

The presented list also includes the exponential moment conditions (ed_0) and (ed_1) on the gradient $\nabla \ell(Y_1, \boldsymbol{\theta})$. We need exponential moments for establishing some nonasymptotic risk bounds, the classical concentration bounds require even stronger conditions that the considered random variables are bounded.

The identifiability condition $(\ell \mathtt{u})$ is very easy to check in the usual asymptotic setup. Indeed, if the parameter set Θ is compact, the Kullback-Leibler divergence $\mathbb{k}(\boldsymbol{\theta}, \boldsymbol{\theta}^*)$ is continuous and positive for all $\boldsymbol{\theta} \neq \boldsymbol{\theta}^*$, then $(\ell \mathtt{u})$ is fulfilled automatically with a universal constant $\mathtt{b}$. If Θ is not compact, the condition is still fulfilled but the function $\mathtt{b}(\mathtt{u})$ may depend on $\mathtt{u}$.

Below we specify the general results of Section 3 and 4 to the i.i.d. setup.

5.1.1. *A large deviation bound.* This section presents some sufficient conditions ensuring a small deviation probability for the event $\{\widetilde{\boldsymbol{\theta}} \notin \Theta_{\mathrm{loc}}(\mathtt{u}_0)\}$ for a fixed $\mathtt{u}_0$. Below $\mathbb{Q} = \mathfrak{c}_1 p$. We only discuss the case $\mathtt{b}(\mathtt{u}) \equiv \mathtt{b}$. The general case only requires more complicated notations. The next result follows from Theorem 4.2 with the obvious changes.

THEOREM 5.2. *Suppose* $(e\mathtt{u})$ *and* $(\ell \mathtt{u})$ *with* $\mathtt{b}(\mathtt{u}) \equiv \mathtt{b}$. *If, for* $\mathtt{u}_0 > 0$,

$$n^{1/2} \mathtt{u}_0 \mathtt{b} \geq 6 \nu_0 \sqrt{\mathtt{x} + \mathbb{Q}}, \tag{5.2}$$
$$1 + \sqrt{\mathtt{x} + \mathbb{Q}} \leq 3 \mathtt{b}^{-1} \nu_0^2 \mathtt{g}_1(\mathtt{u}_0)\, n^{1/2},$$

then

$$I\!P\bigl(\widetilde{\boldsymbol{\theta}} \notin \Theta_{\mathrm{loc}}(\mathtt{u}_0)\bigr) \leq \mathrm{e}^{-\mathtt{x}}.$$

REMARK 5.2. The presented result helps to qualify two important values $\mathtt{u}_0$ and n providing a sensible deviation probability bound. For simplicity

suppose that $\mathtt{g}_1(\mathtt{u}) \equiv \mathtt{g}_1 > 0$. Then the condition (5.2) can be written as $n\mathtt{u}_0^2 \gg \mathtt{x} + \mathbb{Q}$. In other words, the result of the theorem claims a large deviation bound for the vicinity $\Theta_{\mathrm{loc}}(\mathtt{u}_0)$ with $\mathtt{u}_0^2$ of order p/n. In classical asymptotic statistics this result is usually referred to as *root-n consistency.* Our approach yields this result in a very strong form and for finite samples.

5.1.2. *Local inference.* Now we restate the general local bounds of Section 3 for the i.i.d. case. First we describe the approximating linear models. The matrices $\mathbf{v}_0^2$ and $\mathbf{f}_0$ from conditions (ed_0), (ed_1), and (ℓ_0) determine their drift and variance components. Define

$$\mathbf{f}_{\boldsymbol{\epsilon}} \stackrel{\mathrm{def}}{=} \mathbf{f}_0(1-\delta) - \varrho\mathbf{v}_0^2.$$

If $\tau_{\boldsymbol{\epsilon}} \stackrel{\mathrm{def}}{=} \delta + \mathfrak{a}^2\varrho < 1$, then

$$\mathbf{f}_{\boldsymbol{\epsilon}} \geq (1-\tau_{\boldsymbol{\epsilon}})\mathbf{f}_0 > 0.$$

Further, $D_{\boldsymbol{\epsilon}}^2 = n\mathbf{f}_{\boldsymbol{\epsilon}}$ and

$$\boldsymbol{\xi}_{\boldsymbol{\epsilon}} \stackrel{\mathrm{def}}{=} D_{\boldsymbol{\epsilon}}^{-1}\nabla\zeta(\boldsymbol{\theta}^*) = \bigl(n\mathbf{f}_{\boldsymbol{\epsilon}}\bigr)^{-1/2}\sum\nabla\ell(Y_i,\boldsymbol{\theta}^*).$$

The upper bracketing process reads as

$$\mathbb{L}_{\boldsymbol{\epsilon}}(\boldsymbol{\theta},\boldsymbol{\theta}^*) = (\boldsymbol{\theta}-\boldsymbol{\theta}^*)^\top D_{\boldsymbol{\epsilon}}\boldsymbol{\xi}_{\boldsymbol{\epsilon}} - \|D_{\boldsymbol{\epsilon}}(\boldsymbol{\theta}-\boldsymbol{\theta}^*)\|^2/2.$$

This expression can be viewed as log-likelihood for the linear model $\boldsymbol{\xi}_{\boldsymbol{\epsilon}} = D_{\boldsymbol{\epsilon}}\boldsymbol{\theta} + \boldsymbol{\varepsilon}$ for a standard normal error $\boldsymbol{\varepsilon}$. The (quasi) MLE $\widetilde{\boldsymbol{\theta}}_{\boldsymbol{\epsilon}}$ for this model is of the form $\widetilde{\boldsymbol{\theta}}_{\boldsymbol{\epsilon}} = D_{\boldsymbol{\epsilon}}^{-1}\boldsymbol{\xi}_{\boldsymbol{\epsilon}}$.

THEOREM 5.3. *Suppose (ed_0). Given $\mathtt{u}_0$, assume (ed_1), (ℓ_0), and (ι) on $\Theta_{\mathrm{loc}}(\mathtt{u}_0)$, and let $\varrho = 3\nu_0\,\omega^*\mathtt{u}_0$, $\delta = \delta^*\mathtt{u}_0$, and $\tau_{\boldsymbol{\epsilon}} \stackrel{\mathrm{def}}{=} \delta + \mathfrak{a}^2\varrho < 1$. Then the results of Theorem 3.1 and all its corollaries apply to the case of i.i.d. modeling with $\mathtt{r}_0^2 = n\mathtt{u}_0^2$. In particular, on the random set $C_{\boldsymbol{\epsilon}}(\mathtt{r}_0) = \bigl\{\widetilde{\boldsymbol{\theta}} \in \Theta_{\mathrm{loc}}(\mathtt{u}_0), \|\boldsymbol{\xi}_{\underline{\boldsymbol{\epsilon}}}\| \leq \mathtt{r}_0\bigr\}$, it holds*

$$\|\boldsymbol{\xi}_{\underline{\boldsymbol{\epsilon}}}\|^2/2 - \diamondsuit_{\underline{\boldsymbol{\epsilon}}}(\mathtt{r}_0) \leq L(\widetilde{\boldsymbol{\theta}},\boldsymbol{\theta}^*) \leq \|\boldsymbol{\xi}_{\boldsymbol{\epsilon}}\|^2/2 + \diamondsuit_{\boldsymbol{\epsilon}}(\mathtt{r}_0),$$
$$\bigl\|\sqrt{n\mathbf{f}_{\boldsymbol{\epsilon}}}\bigl(\widetilde{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\bigr) - \boldsymbol{\xi}_{\boldsymbol{\epsilon}}\bigr\|^2 \leq 2\Delta_{\boldsymbol{\epsilon}}(\mathtt{r}_0).$$

The random quantities $\diamondsuit_{\boldsymbol{\epsilon}}(\mathtt{r}_0)$ and $\Delta_{\boldsymbol{\epsilon}}(\mathtt{r}_0)$ follow the probability bounds of Theorem 3.7 and 3.10.

Now we briefly discuss the implications of Theorem 5.2 and 5.3 to the classical asymptotic setup with $n \to \infty$. We fix $\mathtt{u}_0^2 = Cp/n$ for a constant C ensuring the deviation bound of Theorem 5.2. Then δ is of order $\mathtt{u}_0$ and the same for ϱ. For a sufficiently large n, both quantities are small and thus, the spread $\Delta_{\boldsymbol{\epsilon}}(\mathtt{r}_0)$ is small as well; see Section 3.3.

Further, under (ed_0) condition, the normalized score

$$\boldsymbol{\xi} \stackrel{\text{def}}{=} \bigl(n\mathbf{f}_0\bigr)^{-1/2} \sum \nabla \ell(Y_i, \boldsymbol{\theta}^*)$$

is zero mean asymptotically normal by the central limit theorem. Moreover, if $\mathbf{f}_0 = \mathbf{v}_0^2$, then $\boldsymbol{\xi}$ is asymptotically standard normal. The same holds for $\boldsymbol{\xi}_{\boldsymbol{\epsilon}}$. This immediately yields all classical asymptotic results like Wilks theorem or the Fisher expansion for MLE in the i.i.d. setup as well as the asymptotic efficiency of the MLE. Moreover, our results bounds yield the asymptotic result for the case when the parameter dimension $p = p_n$ grows linearly with n. Below $u_n = o_n(p_n)$ means that $u_n/p_n \to 0$ with n.

THEOREM 5.4. *Let $Y_1, \ldots, Y_n$ be i.i.d. $I\!P_{\boldsymbol{\theta}^*}$ and let (ed_0), (ed_1), (ℓ_0), (ι), $(e\mathtt{u})$, and $(\ell\mathtt{u})$ with $\mathtt{b}(\mathtt{u}) \equiv \mathtt{b}$ hold. If $n > Cp_n$ for a fixed constant C depending on constants in the above conditions only, then*

$$\| \sqrt{n\mathbf{f}_0}\bigl(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\bigr) - \boldsymbol{\xi}\|^2 = o_n(p_n), \qquad 2L(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}^*) - \|\boldsymbol{\xi}\|^2 = o_n(p_n).$$

This result particularly yields that $\sqrt{n\mathbf{f}_0}\bigl(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\bigr)$ is nearly standard normal and $2L(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}^*)$ is nearly χ^2_p.

5.2. *Generalized linear modeling.* Now we consider a generalized linear modeling (GLM) which is often used for describing some categorical data. Let $\mathcal{P} = (P_w, w \in \Upsilon)$ be an exponential family with a canonical parametrization; see e.g. McCullagh and Nelder (1989). The corresponding log-density can be represented as $\ell(y, w) = yw - d(w)$ for a convex function $d(w)$. The popular examples are given by the binomial (binary response, logistic) model with $d(w) = \log\bigl(e^w + 1\bigr)$, the Poisson model with $d(w) = e^w$, the exponential model with $d(w) = -\log(w)$. Note that linear Gaussian regression is a special case with $d(w) = w^2/2$.

A GLM specification means that every observation Y_i has a distribution from the family $\mathcal{P}$ with the parameter w_i which linearly depends on the regressor $\Psi_i \in I\!R^p$:

$$Y_i \sim P_{\Psi_i^\top \boldsymbol{\theta}^*}. \tag{5.3}$$

The corresponding log-density of a GLM reads as

$$L(\boldsymbol{\theta}) = \sum \bigl\{ Y_i \Psi_i^\top \boldsymbol{\theta} - d(\Psi_i^\top \boldsymbol{\theta}) \bigr\}.$$

Under $I\!P_{\boldsymbol{\theta}^*}$ each observation Y_i follows (5.3), in particular, $I\!E Y_i = d'(\Psi_i^\top \boldsymbol{\theta}^*)$. However, similarly to the previous sections, it is accepted that the parametric model (5.3) is misspecified. Response misspecification means that the vector $\boldsymbol{f} \stackrel{\text{def}}{=} I\!E \boldsymbol{Y}$ cannot be represented in the form $d'(\Psi^\top \boldsymbol{\theta})$ whatever $\boldsymbol{\theta}$ is. The other sort of misspecification concerns the data distribution. The model (5.3) assumes that the Y_i's are independent and the marginal distribution belongs to the given parametric family $\mathcal{P}$. In what follows, we only assume independent data having certain exponential moments. The target of estimation $\boldsymbol{\theta}^*$ is defined by

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

The quasi MLE $\widetilde{\boldsymbol{\theta}}$ is defined by maximization of $L(\boldsymbol{\theta})$:

$$\widetilde{\boldsymbol{\theta}} = \operatorname*{argmax}_{\boldsymbol{\theta}} L(\boldsymbol{\theta}) = \operatorname*{argmax}_{\boldsymbol{\theta}} \sum \bigl\{ Y_i \Psi_i^\top \boldsymbol{\theta} - d(\Psi_i^\top \boldsymbol{\theta}) \bigr\}.$$

Convexity of $d(\cdot)$ implies that $L(\boldsymbol{\theta})$ is a concave function of $\boldsymbol{\theta}$, so that the optimization problem has a unique solution and can be effectively solved. However, a closed form solution is only available for the constant regression or for the linear Gaussian regression. The corresponding target $\boldsymbol{\theta}^*$ is the maximizer of the expected log-likelihood:

$$\boldsymbol{\theta}^* = \operatorname*{argmax}_{\boldsymbol{\theta}} I\!E L(\boldsymbol{\theta}) = \operatorname*{argmax}_{\boldsymbol{\theta}} \sum \bigl\{ f_i \Psi_i^\top \boldsymbol{\theta} - d(\Psi_i^\top \boldsymbol{\theta}) \bigr\}$$

with $f_i = I\!E Y_i$. The function $I\!E L(\boldsymbol{\theta})$ is concave as well and the vector $\boldsymbol{\theta}^*$ is also well defined.

Define the individual errors (residuals) $\varepsilon_i = Y_i - I\!E Y_i$. Below we assume that these errors fulfill some exponential moment conditions.

(e_1) *There exist some constants ν_0 and $\mathtt{g}_1 > 0$, and for every i a constant $\mathfrak{s}_i$ such that $I\!E\bigl(\varepsilon_i/\mathfrak{s}_i\bigr)^2 \le 1$ and*

$$\log I\!E \exp\bigl(\lambda \varepsilon_i/\mathfrak{s}_i\bigr) \le \nu_0^2 \lambda^2/2, \qquad |\lambda| \le \mathtt{g}_1. \tag{5.4}$$

A natural candidate for $\mathfrak{s}_i$ is σ_i where $\sigma_i^2 = I\!E \varepsilon_i^2$ is the variance of ε_i; see Lemma **??**. Under (5.4), introduce a $p \times p$ matrix V_0 defined by

$$V_0^2 \stackrel{\text{def}}{=} \sum \mathfrak{s}_i^2 \Psi_i \Psi_i^\top. \tag{5.5}$$

Condition (e_1) effectively means that each error term $\varepsilon_i = Y_i - \mathbb{E}Y_i$ has some bounded exponential moments: for $|\lambda| \leq \mathtt{g}_1$, it holds $f(\lambda) \stackrel{\text{def}}{=} \log \mathbb{E} \exp\bigl(\lambda \varepsilon_i / \mathfrak{s}_i\bigr) < \infty$. This implies the quadratic upper bound for the function $f(\lambda)$ for $|\lambda| \leq \mathtt{g}_1$; see Lemma **??**. In words, condition (e_1) requires light (exponentially decreasing) tail for the marginal distribution of each ε_i.

Define also

$$N^{-1/2} \stackrel{\text{def}}{=} \max_i \sup_{\boldsymbol{\gamma} \in \mathbb{R}^p} \frac{\mathfrak{s}_i |\Psi_i^\top \boldsymbol{\gamma}|}{\|V_0 \boldsymbol{\gamma}\|} . \tag{5.6}$$

Lemma 5.5. *Assume* (e_1) *and let* V_0 *be defined by* (5.5) *and* N *by* (5.6). *Then conditions* (ED_0) *and* $(E\mathtt{r})$ *follow from* (e_1) *with the matrix* V_0 *due to* (5.5) *and* $\mathtt{g} = \mathtt{g}_1 N^{1/2}$. *Moreover, the stochastic component* $\zeta(\boldsymbol{\theta})$ *is linear in* $\boldsymbol{\theta}$ *and the condition* (ED_1) *is fulfilled with* $\omega(\mathtt{r}) \equiv 0$.

Proof. The gradient of the stochastic component $\zeta(\boldsymbol{\theta})$ of $L(\boldsymbol{\theta})$ does not depend on $\boldsymbol{\theta}$, namely, $\nabla \zeta(\boldsymbol{\theta}) = \sum \Psi_i \varepsilon_i$ with $\varepsilon_i = Y_i - \mathbb{E}Y_i$. Now, for any unit vector $\boldsymbol{\gamma} \in \mathbb{R}^p$ and $\lambda \leq \mathtt{g}$, independence of the ε_i's implies that

$$\log \mathbb{E} \exp\Bigl\{ \frac{\lambda}{\|V_0 \boldsymbol{\gamma}\|} \boldsymbol{\gamma}^\top \sum \Psi_i \varepsilon_i \Bigr\} = \sum \log \mathbb{E} \exp\Bigl\{ \frac{\lambda \mathfrak{s}_i \Psi_i^\top \boldsymbol{\gamma}}{\|V_0 \boldsymbol{\gamma}\|} \varepsilon_i / \mathfrak{s}_i \Bigr\}.$$

By definition $\mathfrak{s}_i |\Psi_i^\top \boldsymbol{\gamma}| / \|V_0 \boldsymbol{\gamma}\| \leq N^{-1/2}$ and therefore, $\lambda \mathfrak{s}_i |\Psi_i^\top \boldsymbol{\gamma}| / \|V_0 \boldsymbol{\gamma}\| \leq \mathtt{g}_1$. Hence, (5.4) implies

$$\log \mathbb{E} \exp\Bigl\{ \frac{\lambda}{\|V_0 \boldsymbol{\gamma}\|} \boldsymbol{\gamma}^\top \sum \Psi_i \varepsilon_i \Bigr\} \leq \frac{\nu_0^2 \lambda^2}{2 \|V_0 \boldsymbol{\gamma}\|^2} \sum \mathfrak{s}_i^2 |\Psi_i^\top \boldsymbol{\gamma}|^2 = \frac{\nu_0^2 \lambda^2}{2} , \tag{5.7}$$

and (ED_0) follows. □

It only remains to bound the quality of quadratic approximation for the mean of the process $L(\boldsymbol{\theta}, \boldsymbol{\theta}^*)$ in a vicinity of $\boldsymbol{\theta}^*$. An interesting feature of the GLM is that the effect of model misspecification disappears in the expectation of $L(\boldsymbol{\theta}, \boldsymbol{\theta}^*)$.

Lemma 5.6. *It holds*

$$\begin{aligned} -\mathbb{E}L(\boldsymbol{\theta}, \boldsymbol{\theta}^*) &= \sum \bigl\{ d(\Psi_i^\top \boldsymbol{\theta}) - d(\Psi_i^\top \boldsymbol{\theta}^*) - d'(\Psi_i^\top \boldsymbol{\theta}^*) \Psi_i^\top (\boldsymbol{\theta} - \boldsymbol{\theta}^*) \bigr\} \\ &= \mathcal{K}\bigl(\mathbb{P}_{\boldsymbol{\theta}^*}, \mathbb{P}_{\boldsymbol{\theta}} \bigr), \end{aligned} \tag{5.8}$$

where $\mathcal{K}\bigl(I\!P_{\boldsymbol{\theta}^*}, I\!P_{\boldsymbol{\theta}}\bigr)$ *is the Kullback-Leibler divergence between measures* $I\!P_{\boldsymbol{\theta}^*}$ *and* $I\!P_{\boldsymbol{\theta}}$ *. Moreover,*

$$-I\!E L(\boldsymbol{\theta}, \boldsymbol{\theta}^*) = \|D(\boldsymbol{\theta}^\circ)(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\|^2/2, \tag{5.9}$$

where $\boldsymbol{\theta}^\circ \in [\boldsymbol{\theta}^*, \boldsymbol{\theta}]$ *and*

$$D^2(\boldsymbol{\theta}^\circ) = \sum d''(\Psi_i^\top \boldsymbol{\theta}^\circ) \Psi_i \Psi_i^\top .$$

Proof. The definition implies

$$I\!E L(\boldsymbol{\theta}, \boldsymbol{\theta}^*) = \sum \bigl\{ f_i \Psi_i^\top (\boldsymbol{\theta} - \boldsymbol{\theta}^*) - d(\Psi_i^\top \boldsymbol{\theta}) + d(\Psi_i^\top \boldsymbol{\theta}^*) \bigr\}.$$

As $\boldsymbol{\theta}^*$ is the extreme point of $I\!E L(\boldsymbol{\theta})$, it holds $\nabla I\!E L(\boldsymbol{\theta}^*) = \sum \bigl[f_i - d'(\Psi_i^\top \boldsymbol{\theta}^*) \bigr] \Psi_i = 0$ and (5.8) follows. The Taylor expansion of the second order around $\boldsymbol{\theta}^*$ yields the expansion (5.9). □

Define now the matrix D_0 by

$$D_0^2 \stackrel{\text{def}}{=} D^2(\boldsymbol{\theta}^*) = \sum d''(\Psi_i^\top \boldsymbol{\theta}^*) \Psi_i \Psi_i^\top .$$

Let also V_0 be defined by (5.5). Note that the matrices D_0 and V_0 coincide if the model $Y_i \sim P_{\Psi_i^\top \boldsymbol{\theta}^*}$ is correctly specified and $\mathfrak{s}_i^2 = d''(\Psi_i^\top \boldsymbol{\theta}^*)$. The matrix V_0 describes a local elliptic neighborhood of the central point $\boldsymbol{\theta}^*$ in the form $\Theta_0(\mathtt{r}) = \{\boldsymbol{\theta} : \|V_0(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\| \le \mathtt{r}\}$. If the matrix function $D^2(\boldsymbol{\theta})$ is continuous in this vicinity $\Theta_0(\mathtt{r})$ then the value $\delta(\mathtt{r})$ measuring the approximation quality of $-I\!E L(\boldsymbol{\theta}, \boldsymbol{\theta}^*)$ by the quadratic function $\|D_0(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\|^2/2$ is small and the identifiability condition $(\mathcal{L}_0)$ is fulfilled on $\Theta_0(\mathtt{r})$.

Lemma 5.7. *Suppose that*

$$\|\boldsymbol{I}_p - D_0^{-1} D^2(\boldsymbol{\theta}) D_0^{-1}\|_\infty \le \delta(\mathtt{r}), \qquad \boldsymbol{\theta} \in \Theta_0(\mathtt{r}). \tag{5.10}$$

Then $(\mathcal{L}_0)$ *holds with this* $\delta(\mathtt{r})$ *. Moreover, as the quantities* $\omega(\mathtt{r}), \diamondsuit_{\boldsymbol{\epsilon}}(\mathtt{r}), \diamondsuit_{\underline{\boldsymbol{\epsilon}}}(\mathtt{r})$ *vanish, one can take* $\varrho = 0$ *leading to the following representation for* $D_{\boldsymbol{\epsilon}}$ *and* $\boldsymbol{\xi}_{\boldsymbol{\epsilon}}$ *:*

$$\begin{aligned} D_{\boldsymbol{\epsilon}}^2 &= (1-\delta) D_0^2, \qquad & \boldsymbol{\xi}_{\boldsymbol{\epsilon}} &= (1+\delta)^{1/2} \boldsymbol{\xi} \\ D_{\underline{\boldsymbol{\epsilon}}}^2 &= (1+\delta) D_0^2, \qquad & \boldsymbol{\xi}_{\underline{\boldsymbol{\epsilon}}} &= (1-\delta)^{1/2} \boldsymbol{\xi} \end{aligned}$$

with

$$\boldsymbol{\xi} \stackrel{\text{def}}{=} D_0^{-1} \nabla \zeta = D_0^{-1} \sum \Psi_i (Y_i - I\!E Y_i).$$

Linearity of the stochastic component $\zeta(\boldsymbol{\theta})$ in the considered GLM implies the important fact that the quantities $\diamondsuit_{\boldsymbol{\epsilon}}(\mathtt{r}), \diamondsuit_{\underline{\boldsymbol{\epsilon}}}(\mathtt{r})$ in the general bracketing bound (3.4) vanish for any $\mathtt{r}$. Therefore, in the GLM case, the deficiency can be defined as the difference between upper and lower excess and it can be easily evaluated:

$$\Delta(\mathtt{r}) = \|\boldsymbol{\xi}_{\boldsymbol{\epsilon}}\|^2/2 - \|\boldsymbol{\xi}_{\underline{\boldsymbol{\epsilon}}}\|^2/2 = \delta\|\boldsymbol{\xi}\|^2.$$

Our result assumes some concentration properties of the squared norm $\|\boldsymbol{\xi}\|^2$ of the vector $\boldsymbol{\xi}$. These properties can be established by general results of Section **??** under the regularity condition: for some $\mathfrak{a}$

$$V_0 \leq \mathfrak{a} D_0. \tag{5.11}$$

Now we are prepared to state the local results for the GLM estimation.

THEOREM 5.8. *Let* (e_1) *hold. Then for* $\delta \geq \delta(\mathtt{r})$ *any* $z > 0$ *and* $\mathfrak{z} > 0$, *it holds*

$$\begin{aligned} I\!P\bigl(\|D_0(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\| > z,\, \|V_0(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\| \leq \mathtt{r}\bigr) &\leq I\!P\bigl\{\|\boldsymbol{\xi}\|^2 > (1-\delta)z^2\bigr\} \\ I\!P\bigl(L(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}^*) > \mathfrak{z},\, \|V_0(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\| \leq \mathtt{r}\bigr) &\leq I\!P\bigl\{\|\boldsymbol{\xi}\|^2/2 > (1-\delta)\mathfrak{z}\bigr\}. \end{aligned}$$

Moreover, on the set $C_{\boldsymbol{\epsilon}}(\mathtt{r}) = \{\|V_0(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\| \leq \mathtt{r},\, \|\boldsymbol{\xi}_{\underline{\boldsymbol{\epsilon}}}\| \leq \mathtt{r}\}$, *it holds*

$$\|D_0(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) - \boldsymbol{\xi}\|^2 \leq \frac{2\delta}{1-\delta^2}\|\boldsymbol{\xi}\|^2. \tag{5.12}$$

If the function $d(w)$ is quadratic then the approximation error δ vanishes as well and the expansion (5.12) becomes equality which is also fulfilled globally, a localization step in not required. However, if $d(w)$ is not quadratic, the result applies only locally and it has to be accomplished with a large deviation bound. The GLM structure is helpful in the large deviation zone as well. Indeed, the gradient $\nabla\zeta(\boldsymbol{\theta})$ does not depend on $\boldsymbol{\theta}$ and hence, the most delicate condition $(E\mathtt{r})$ is fulfilled automatically with $\mathtt{g} = \mathtt{g}_1 N^{1/2}$ for all local sets $\Theta_0(\mathtt{r})$. Further, the identifiability condition $(\mathcal{L}\mathtt{r})$ easily follows from Lemma 5.6: it suffices to bound from below the matrix $D(\boldsymbol{\theta})$ for $\boldsymbol{\theta} \in \Theta_0(\mathtt{r})$:

$$D(\boldsymbol{\theta}) \geq \mathtt{b}(\mathtt{r}) V_0, \qquad \boldsymbol{\theta} \in \Theta_0(\mathtt{r}).$$

An interesting question, similarly to the i.i.d. case, is the minimal radius $\mathtt{r}_0$ of the local vicinity $\Theta_0(\mathtt{r}_0)$ ensuring the desirable concentration property. Suppose for the moment that the constants $\mathtt{b}(\mathtt{r})$ are all the same for

different $\mathtt{r}$: $\mathtt{b}(\mathtt{r}) \equiv \mathtt{b}$. Under the regularity condition (5.11), a sufficient lower bound for $\mathtt{r}_0$ can be based on Corollary 4.3. The required condition can be restated as

$$1 + \sqrt{\mathtt{x} + \mathbb{Q}} \leq 3\nu_0^2 \mathtt{g}/\mathtt{b}, \qquad 6\nu_0 \sqrt{\mathtt{x} + \mathbb{Q}} \leq \mathtt{rb}.$$

It remains to note that $\mathbb{Q} = \mathfrak{c}_1 p$ and $\mathtt{g} = \mathtt{g}_1 N^{1/2}$. So, the required conditions are fulfilled for $\mathtt{r}^2 \geq \mathtt{r}_0^2 = C(\mathtt{x} + p)$, where C only depends on $\nu_0, \mathtt{b}$, and $\mathtt{g}$.

5.3. *Linear median estimation.* This section illustrates how the proposed approach applies to robust estimation in linear models. The target of analysis is the linear dependence of the observed data $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ on the set of features $\Psi_i \in I\!R^p$:

$$Y_i = \Psi_i^\top \boldsymbol{\theta} + \varepsilon_i \,, \tag{5.13}$$

where ε_i means the i th individual error. As usual, the true data distribution can deviate from the linear model. In addition, we admit contaminated data which naturally leads to the idea of robust estimation. This section offers a qMLE view on the robust estimation problem. Our parametric family assumes the linear dependence (5.13) with i.i.d. errors ε_i which follow the double exponential (Laplace) distribution with the density $(1/2)e^{-|y|}$. Then the corresponding log-likelihood reads as

$$L(\boldsymbol{\theta}) = -\frac{1}{2} \sum |Y_i - \Psi_i^\top \boldsymbol{\theta}|$$

and $\widetilde{\boldsymbol{\theta}} \stackrel{\text{def}}{=} \operatorname{argmax}_{\boldsymbol{\theta}} L(\boldsymbol{\theta})$ is called the *least absolute deviation* (LAD) estimate. In the context of linear regression, it is also called the *linear median* estimate. The target of estimation $\boldsymbol{\theta}^*$ is defined as usually by the equation $\boldsymbol{\theta}^* = \operatorname{argmax}_{\boldsymbol{\theta}} I\!E L(\boldsymbol{\theta})$.

It is useful to define the residuals $\widetilde{\varepsilon}_i = Y_i - \Psi_i^\top \boldsymbol{\theta}^*$ and their distributions

$$P_i(A) = I\!P\bigl(\widetilde{\varepsilon}_i \in A\bigr) = I\!P\bigl(Y_i - \Psi_i^\top \boldsymbol{\theta}^* \in A\bigr)$$

for any Borel set A on the real line. If $Y_i = \Psi_i^\top \boldsymbol{\theta}^* + \varepsilon_i$ is the true model then P_i coincides with the distribution of each ε_i. Below we suppose that each P_i has a positive density $f_i(y)$.

Note that the difference $L(\boldsymbol{\theta}) - L(\boldsymbol{\theta}^*)$ is bounded by $\frac{1}{2} \sum |\Psi_i^\top (\boldsymbol{\theta} - \boldsymbol{\theta}^*)|$. Next we check conditions (ED_0) and (ED_1). Denote $\xi_i(\boldsymbol{\theta}) = 1\!\!1(Y_i - \Psi_i^\top \boldsymbol{\theta} \leq$

$0) - q_i(\boldsymbol{\theta})$ for $q_i(\boldsymbol{\theta}) = I\!\!P(Y_i - \Psi_i^\top \boldsymbol{\theta} \le 0)$. This is a centered Bernoulli random variable, and it is easy to check that

$$\nabla \zeta(\boldsymbol{\theta}) = -\sum \xi_i(\boldsymbol{\theta}) \Psi_i. \tag{5.14}$$

This expression differs from the similar ones from the linear and generalized linear regression because the stochastic terms ξ_i now depend on $\boldsymbol{\theta}$. First we check the global condition $(E\mathtt{r})$. Fix any $\mathtt{g}_1 < 1$. Then it holds for a Bernoulli r.v. Z with $I\!\!P(Z = 1) = q$, $\xi = Z - q$, and $|\lambda| \le \mathtt{g}_1$

$$\begin{aligned} \log I\!\!E \exp(\lambda \xi) &= \log\bigl[q \exp\{\lambda(1-q)\} + (1-q)\exp(-\lambda q)\bigr] \\ &\le \nu_0^2 q(1-q)\lambda^2/2, \end{aligned} \tag{5.15}$$

where $\nu_0 \ge 1$ depends on $\mathtt{g}_1$ only. Let now a vector $\boldsymbol{\gamma} \in I\!\!R^p$ and $\rho > 0$ be such that $\rho |\Psi_i^\top \boldsymbol{\gamma}| \le \mathtt{g}_1$ for all $i = 1, \ldots, n$. Then

$$\begin{aligned} \log I\!\!E \exp\{\rho \boldsymbol{\gamma}^\top \nabla \zeta(\boldsymbol{\theta})\} &\le \frac{\nu_0^2 \rho^2}{2} \sum q_i(\boldsymbol{\theta}) \bigl\{1 - q_i(\boldsymbol{\theta})\bigr\} |\Psi_i^\top \boldsymbol{\gamma}|^2 \\ &\le \frac{\nu_0^2 \rho^2}{2} \|V(\boldsymbol{\theta}) \boldsymbol{\gamma}\|^2, \end{aligned} \tag{5.16}$$

where

$$V^2(\boldsymbol{\theta}) = \sum q_i(\boldsymbol{\theta}) \bigl\{1 - q_i(\boldsymbol{\theta})\bigr\} \Psi_i \Psi_i^\top.$$

Denote also

$$V_0^2 = \frac{1}{4} \sum \Psi_i \Psi_i^\top. \tag{5.17}$$

Clearly $V(\boldsymbol{\theta}) \le V_0$ for all $\boldsymbol{\theta}$ and condition $(E\mathtt{r})$ is fulfilled with the matrix V_0 and $\mathtt{g}(\mathtt{r}) \equiv \mathtt{g} = \mathtt{g}_1 N^{1/2}$ for N defined by

$$N^{-1/2} \stackrel{\text{def}}{=} \max_i \sup_{\boldsymbol{\gamma} \in I\!\!R^p} \frac{\Psi_i^\top \boldsymbol{\gamma}}{2 \|V_0 \boldsymbol{\gamma}\|}; \tag{5.18}$$

cf. (5.7).

Let some $\mathtt{r}_0 > 0$ be fixed. We will specify this choice later. Now we check the local conditions within the elliptic vicinity $\Theta_0(\mathtt{r}_0) = \{\boldsymbol{\theta} : \|V_0(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\| \le \mathtt{r}_0\}$ of the central point $\boldsymbol{\theta}^*$ for V_0 from (5.17). Then condition (ED_0) with the matrix V_0 and $\mathtt{g} = N^{1/2} \mathtt{g}_1$ is fulfilled on $\Theta_0(\mathtt{r}_0)$ due to (5.16). Next,

in view of (5.18), it holds $|\Psi_i^\top \boldsymbol{\gamma}| \le 2N^{-1/2}\|V_0\boldsymbol{\gamma}\|$ for any vector $\boldsymbol{\gamma} \in I\!R^p$. By (5.14)

$$\nabla\zeta(\boldsymbol{\theta}) - \nabla\zeta(\boldsymbol{\theta}^*) = \sum \Psi_i \bigl\{\xi_i(\boldsymbol{\theta}) - \xi_i(\boldsymbol{\theta}^*)\bigr\}.$$

If $\Psi_i^\top \boldsymbol{\theta} \ge \Psi_i^\top \boldsymbol{\theta}^*$, then

$$\xi_i(\boldsymbol{\theta}) - \xi_i(\boldsymbol{\theta}^*) = 1\!\!1(\Psi_i^\top \boldsymbol{\theta}^* \le Y_i < \Psi_i^\top \boldsymbol{\theta}) - I\!P\bigl(\Psi_i^\top \boldsymbol{\theta}^* \le Y_i < \Psi_i^\top \boldsymbol{\theta}\bigr).$$

Similarly for $\Psi_i^\top \boldsymbol{\theta} < \Psi_i^\top \boldsymbol{\theta}^*$

$$\xi_i(\boldsymbol{\theta}) - \xi_i(\boldsymbol{\theta}^*) = -\,1\!\!1(\Psi_i^\top \boldsymbol{\theta} \le Y_i < \Psi_i^\top \boldsymbol{\theta}^*) + I\!P\bigl(\Psi_i^\top \boldsymbol{\theta} \le Y_i < \Psi_i^\top \boldsymbol{\theta}^*\bigr).$$

Define $q_i(\boldsymbol{\theta}, \boldsymbol{\theta}^*) \stackrel{\text{def}}{=} \bigl|q_i(\boldsymbol{\theta}) - q_i(\boldsymbol{\theta}^*)\bigr|$. Now (5.15) yields similarly to (5.16)

$$\begin{aligned}
\log I\!E \exp\bigl\{\rho\boldsymbol{\gamma}^\top \bigl\{\nabla\zeta(\boldsymbol{\theta}) - \nabla\zeta(\boldsymbol{\theta}^*)\bigr\}\bigr\} &\le \frac{\nu_0^2\rho^2}{2} \sum q_i(\boldsymbol{\theta}, \boldsymbol{\theta}^*) |\Psi_i^\top \boldsymbol{\gamma}|^2 \\
&\le 2\nu_0^2\rho^2 \max_{i \le n} q_i(\boldsymbol{\theta}, \boldsymbol{\theta}^*)\, \|V_0\boldsymbol{\gamma}\|^2 \le \omega(\mathtt{r})\nu_0^2\rho^2\|V_0\boldsymbol{\gamma}\|^2/2,
\end{aligned}$$

with

$$\omega(\mathtt{r}) \stackrel{\text{def}}{=} 4 \max_{i \le n} \sup_{\boldsymbol{\theta} \in \Theta_0(\mathtt{r})} q_i(\boldsymbol{\theta}, \boldsymbol{\theta}^*).$$

If each density function p_i is uniformly bounded by a constant C then

$$|q_i(\boldsymbol{\theta}) - q_i(\boldsymbol{\theta}^*)| \le C\bigl|\Psi_i^\top(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\bigr| \le CN^{-1/2}\|V_0(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\| \le CN^{-1/2}\mathtt{r}.$$

Next we check the local identifiability condition. We use the following technical lemma.

Lemma 5.9. *It holds for any* $\boldsymbol{\theta}$

$$-\frac{\partial^2}{\partial^2\boldsymbol{\theta}} I\!\!E L(\boldsymbol{\theta}) = D^2(\boldsymbol{\theta}) \stackrel{\text{def}}{=} \sum p_i\bigl(\Psi_i^\top(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\bigr)\Psi_i\Psi_i^\top, \tag{5.19}$$

where $f_i(\cdot)$ *is the density of* $\widetilde{\varepsilon}_i = Y_i - \Psi_i^\top \boldsymbol{\theta}^*$. *Moreover, there is* $\boldsymbol{\theta}^\circ \in [\boldsymbol{\theta}, \boldsymbol{\theta}^*]$ *such that*

$$\begin{aligned}
-I\!\!E L(\boldsymbol{\theta}, \boldsymbol{\theta}^*) &= \frac{1}{2}\sum |\Psi_i^\top(\boldsymbol{\theta} - \boldsymbol{\theta}^*)|^2 f_i(\Psi_i^\top(\boldsymbol{\theta}^\circ - \boldsymbol{\theta}^*)) \\
&= (\boldsymbol{\theta} - \boldsymbol{\theta}^*)^\top D^2(\boldsymbol{\theta}^\circ)(\boldsymbol{\theta} - \boldsymbol{\theta}^*)/2.
\end{aligned} \tag{5.20}$$

Proof. Obviously

$$\frac{\partial \mathbb{E}L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \sum \bigl\{ \mathbb{P}(Y_i \le \Psi_i^\top \boldsymbol{\theta}) - 1/2 \bigr\} \Psi_i .$$

The identity (5.19) is obtained by one more differentiation. By definition, $\boldsymbol{\theta}^*$ is the extreme point of $\mathbb{E}L(\boldsymbol{\theta})$. The equality $\nabla \mathbb{E}L(\boldsymbol{\theta}^*) = 0$ yields

$$\sum \bigl\{ \mathbb{P}(Y_i \le \Psi_i^\top \boldsymbol{\theta}^*) - 1/2 \bigr\} \Psi_i = 0 .$$

Now (5.20) follows by the Taylor expansion of the second order at $\boldsymbol{\theta}^*$. □

Define

$$D_0^2 \stackrel{\text{def}}{=} \sum |\Psi_i^\top (\boldsymbol{\theta} - \boldsymbol{\theta}^*)|^2 f_i(0). \tag{5.21}$$

Due to this lemma, condition $(\mathcal{L}_0)$ is fulfilled in $\Theta_0(\mathtt{r})$ with this choice D_0 for $\delta(\mathtt{r})$ from (5.10); see Lemma 5.7. Moreover, if $f_i(0) \ge \mathfrak{a}^2/4$ for $\mathfrak{a} > 0$, then the identifiability condition $(\mathcal{I})$ is also satisfied. Now all the local conditions are fulfilled yielding the general bracketing bound of Theorem 3.1 and all its corollaries.

It only remains to accomplish them by a large deviation bound, that is, to specify the local vicinity $\Theta_0(\mathtt{r}_0)$ providing the prescribed deviation bound. A sufficient condition for the concentration property is that the expectation $\mathbb{E}L(\boldsymbol{\theta}, \boldsymbol{\theta}^*)$ grows in absolute value with the distance $\|V_0(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\|$. We use the representation (5.19). Suppose that for some fixed $\delta < 1/2$ and $\rho > 0$

$$\bigl| f_i(u)/f_i(0) - 1 \bigr| \le \delta, \qquad |u| \le \rho . \tag{5.22}$$

For any $\boldsymbol{\theta}$ with $\|V_0(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\| = \mathtt{r} \ge \mathtt{r}_0$, and for any $i = 1, \ldots, n$, it holds

$$|\Psi_i^\top (\boldsymbol{\theta} - \boldsymbol{\theta}^*)| \le N^{-1/2} \|V_0(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\| = N^{-1/2} \mathtt{r} .$$

Therefore, for $\mathtt{r} \le \rho N^{1/2}$ and any $\boldsymbol{\theta} \in \Theta_0(\mathtt{r})$ with $\|V_0(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\| = \mathtt{r}$, it holds $f_i\bigl(\Psi_i^\top (\boldsymbol{\theta}^\circ - \boldsymbol{\theta}^*)\bigr) \ge (1 - \delta) f_i(0)$. Now Lemma 5.9 implies

$$-\mathbb{E}L(\boldsymbol{\theta}, \boldsymbol{\theta}^*) \ge \frac{1-\delta}{2} \|D_0(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\|^2 \ge \frac{1-\delta}{2\mathfrak{a}^2} \|V_0(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\|^2 = \frac{1-\delta}{2\mathfrak{a}^2} \mathtt{r}^2 .$$

By Lemma 5.9 the function $-\mathbb{E}L(\boldsymbol{\theta}, \boldsymbol{\theta}^*)$ is convex. This easily yields

$$-\mathbb{E}L(\boldsymbol{\theta}, \boldsymbol{\theta}^*) \ge \frac{1-\delta}{2\mathfrak{a}^2} \rho N^{1/2} \mathtt{r}$$

for all $\mathtt{r} \geq \rho N^{1/2}$. Thus,

$$\mathtt{r}\mathtt{b}(\mathtt{r}) \geq \begin{cases} (1-\delta)(2\mathfrak{a}^2)^{-1}\mathtt{r} & \text{if } \mathtt{r} \leq \rho N^{1/2}, \\ (1-\delta)(2\mathfrak{a}^2)^{-1}\rho N^{1/2} & \text{if } \mathtt{r} > \rho N^{1/2}. \end{cases}$$

So, the global identifiability condition $(\mathcal{L}_1)$ is fulfilled if $\mathtt{r}_0^2 \geq C_1\mathfrak{a}^2(\mathtt{x}+\mathbb{Q})$ and if $\rho^2 N \geq C_2\mathfrak{a}^2(\mathtt{x}+\mathbb{Q})$ for some fixed constants C_1 and C_2.

Putting all together yields the following result.

THEOREM 5.10. *Let Y_i be independent, $\boldsymbol{\theta}^* = \operatorname{argmax}_{\boldsymbol{\theta}} \mathbb{E}L(\boldsymbol{\theta})$, D_0^2 be given by* (5.21)*, and V_0^2 by* (5.17)*. Let also the densities $f_i(\cdot)$ of $Y_i - \Psi_i^\top \boldsymbol{\theta}^*$ be uniformly bounded by a constant C, fulfill* (5.22) *for some $\rho > 0$ and $\delta > 0$, and $f_i(0) \geq \mathfrak{a}^2/4$ for all i. Finally, let $N \geq C_2\rho^{-2}\mathfrak{a}^2(\mathtt{x}+p)$ for some fixed $\mathtt{x} > 0$ and C_2. Then on the random set of probability at least $1 - \mathrm{e}^{-\mathtt{x}}$, one obtains for $\boldsymbol{\xi} \stackrel{\text{def}}{=} D_0^{-1}\nabla L(\boldsymbol{\theta}^*)$ the bounds*

$$\|\sqrt{D_0}\bigl(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\bigr) - \boldsymbol{\xi}\|^2 = o(p), \qquad 2L(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}^*) - \|\boldsymbol{\xi}\|^2 = o(p).$$

SUPPLEMENTARY MATERIAL

Some results from the theory of empirical processes (doi: ???http://lib.stat.cmu.edu/aos/???/???). This part collects some general deviation bounds for non-Gaussian quadratic forms (Section A) and for general centered random processes (Section B) used in the text.

REFERENCES

Andresen, A. and Spokoiny, V. (2012). Wilks theorem for a quasi profile maximum likelihood.

Bednorz, W. (2006). A theorem on majorizing measures. *Ann. Probab.*, 34(5):1771–1781.

Birgé, L. (2006). Model selection via testing: an alternative to (penalized) maximum likelihood estimators. *Annales de l'Institut Henri Poincare (B) Probability and Statistics.*

Birgé, L. and Massart, P. (1993). Rates of convergence for minimum contrast estimators. *Probab. Theory Relat. Fields*, 97(1-2):113–150.

Birgé, L. and Massart, P. (1998). Minimum contrast estimators on sieves: Exponential bounds and rates of convergence. *Bernoulli*, 4(3):329–375.

Boucheron, S., Lugosi, G., and Massart, P. (2003). Concentration inequalities using the entropy method. *Ann. Probab.*, 31(3):1583–1614.

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

Le Cam, L. (1960). Locally asymptotically normal families of distributions. Certain approximations to families of distributions and their use in the theory of estimation and testing hypotheses. *Univ. California Publ. Stat.*, 3:37–98.

Le Cam, L. and Yang, G. L. (2000). *Asymptotics in Statistics: Some Basic Concepts.* Springer-Verlag, New York.

McCullagh, P. and Nelder, J. (1989). *Generalized linear models. 2nd ed.* Monographs on Statistics and Applied Probability. 37. London etc.: Chapman and Hall. xix, 511 p. .

Spokoiny, V. (2012). Roughness penalty, wilks phenomenon, and bernstein - von mises theorem.

Spokoiny, V., Wang, W., and Härdle, W. (2012). Local quantile regression.

Talagrand, M. (1996). Majorizing measures: The generic chaining. *Ann. Probab.*, 24(3):1049–1103.

Talagrand, M. (2001). Majorizing measures without measures. *Ann. Probab.*, 29(1):411–417.

Talagrand, M. (2005). *The generic chaining.* Springer Monographs in Mathematics. Springer-Verlag, Berlin. Upper and lower bounds of stochastic processes.

Van de Geer, S. (1993). Hellinger-consistency of certain nonparametric maximum likelihood estimators. *Ann. Stat.*, 21(1):14–44.

van der Vaart, A. and Wellner, J. A. (1996). *Weak convergence and empirical processes. With applications to statistics.* Springer Series in Statistics. New York, Springer.

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