Spokoiny, Vladimir

Working Paper

Parametric estimation: Finite sample theory

SFB 649 Discussion Paper, No. 2011-081

Provided in Cooperation with:
Collaborative Research Center 649: Economic Risk, Humboldt University Berlin

Suggested Citation: Spokoiny, Vladimir (2011) : Parametric estimation: Finite sample theory, SFB 649 Discussion Paper, No. 2011-081, Humboldt University of Berlin, Collaborative Research Center 649 - Economic Risk, Berlin

This Version is available at:
https://hdl.handle.net/10419/56754

SFB 649 Discussion Paper 2011-081

Parametric estimation. Finite sample theory

Vladimir Spokoiny*

* Weierstrass Institute (WIAS) Berlin, Germany

This research was supported by the Deutsche Forschungsgemeinschaft through the SFB 649 "Economic Risk".

http://sfb649.wiwi.hu-berlin.de
ISSN 1860-5664

SFB 649, Humboldt-Universität zu Berlin
Spandauer Straße 1, D-10178 Berlin

Parametric estimation. Finite sample theory

Vladimir Spokoiny *

Weierstrass-Institute,
Mohrenstr. 39, 10117 Berlin, Germany
spokoiny@wias-berlin.de

Abstract

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.

The main results include a large deviation bounds for the (quasi) maximum likelihood and the local quadratic majorization 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. The general results are illustrated for the i.i.d. set-up as well as for generalized linear 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.

AMS 2000 Subject Classification: Primary 62F10. Secondary 62J12,62F25,62H12

Keywords: maximum likelihood, local quadratic approximation, concentration, coverage, deficiency

JEL-Classification: C13, C14

*Financial support by the German Research Foundation (DFG) through the Collaborative Research Center 649 "Economic Risk" is gratefully acknowledged

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 states 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 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 limiting the scope of applicability for the whole approach. Indeed, the PA is usually only an approximation of real data distribution and in the most of 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 feature of the presentation is the *non-asymptotic* framework. The notions like asymptotic normality, convergence or tightness are meaningless in the non-asymptotic setup, the arguments based on compactness of the parameter space are not really helpful. Instead some exact exponential bounds and concentration results are systematically used. 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 we establish a kind of large deviation bound allowing to localize the problem into a neighborhood of the target parameter. Then we use a local quadratic expansion of the log-likelihood to solve the corresponding estimation problem.

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, $I\!P \notin (I\!P_{\boldsymbol{\theta}})$.

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 $(I\!P_{\boldsymbol{\theta}})$. The classical likelihood principle suggests to estimate $\boldsymbol{\theta}$ by maximizing the corresponding log-likelihood function $L(\boldsymbol{Y}, \boldsymbol{\theta})$:

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

Our ultimate goal is to study the properties of the quasi 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 then 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 $I\!P \notin (I\!P_{\boldsymbol{\theta}})$, then the (quasi) MLE estimate $\widetilde{\boldsymbol{\theta}}$ from (1.1) is still meaningful and it appears to be an estimate of the value $\boldsymbol{\theta}^*$ defined by maximizing the expected value of $L(\boldsymbol{Y}, \boldsymbol{\theta})$:

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

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.

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 $\boldsymbol{Y}$. We also define the excess or maximum log-likelihood $L(\boldsymbol{\theta}, \boldsymbol{\theta}^*) = L(\boldsymbol{\theta}) - L(\boldsymbol{\theta}^*)$. 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 quasi maximum likelihood $L(\widetilde{\boldsymbol{\theta}}) - L(\boldsymbol{\theta}^*) = \max_{\boldsymbol{\theta} \in \Theta} L(\boldsymbol{\theta}) - L(\boldsymbol{\theta}^*)$, which is the difference between the maximum of the process $L(\boldsymbol{\theta})$ and its value at the "true" point $\boldsymbol{\theta}^*$.

The established results can be split into two big groups. A large deviation bound states some concentration properties of the estimate $\widetilde{\boldsymbol{\theta}}$. 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 for restricting the parameter space to a properly selected vicinity $\Theta_0(\mathtt{r})$. Our main results describe local properties of the process $L(\boldsymbol{\theta})$ within $\Theta_0(\mathtt{r})$. They can be viewed as a non-asymptotic version of the Le Cam LAN

theory.

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 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 estimate $\widetilde{\boldsymbol{\theta}}$, polynomial risk bounds, etc. In the 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 the 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 7 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 6 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 7 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.

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 which are quite general and not very much restrictive. 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. In some cases we indicate possible extensions of the results under more general conditions. It is important to stress that all the conditions only concern the properties of the quasi likelihood process $L(\boldsymbol{\theta})$.

The imposed conditions on the process 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 $\boldsymbol{\theta}^*$.

2.1 Global conditions

The first global condition (E) assumes some exponential moments for the quasi log-likelihood $L(\boldsymbol{\theta})$ for each $\boldsymbol{\theta} \in \Theta$. The second condition (ED) assumes some smoothness of $L(\boldsymbol{\theta})$ and requires exponential moments of its gradient. The important identifiability condition is stated later; see Section 4.2.

The formulation involves a subset $\mathbb{M}$ of $\mathbb{R}_+$ describing all possible exponents in the moment conditions.

$(\boldsymbol{E})$ *For each $\boldsymbol{\theta} \in \Theta$, there exists a positive value $\mu \in \mathbb{M}$ such that*

$$\mathbb{E} \exp\bigl\{\mu L(\boldsymbol{\theta}, \boldsymbol{\theta}^*)\bigr\} < \infty.$$

Note that this condition is automatically fulfilled if $I\!P = I\!P_{\boldsymbol{\theta}^*}$ and all the $I\!P_{\boldsymbol{\theta}}$'s are absolutely continuous w.r.t. $I\!P_{\boldsymbol{\theta}^*}$ with $\mu \leq 1$. Indeed, $L(\boldsymbol{\theta}, \boldsymbol{\theta}^*) = \log\bigl(dI\!P_{\boldsymbol{\theta}}/dI\!P_{\boldsymbol{\theta}^*}\bigr)$ and $\log I\!E_{\boldsymbol{\theta}^*}(dI\!P_{\boldsymbol{\theta}}/dI\!P_{\boldsymbol{\theta}^*}) = 0$. For $\mu < 1$, it holds by the Jensen inequality that $-\log I\!E_{\boldsymbol{\theta}^*}(dI\!P_{\boldsymbol{\theta}}/dI\!P_{\boldsymbol{\theta}^*})^{\mu} \geq 0$.

Condition (E) enables us to define the function

$$\mathfrak{M}(\mu, \boldsymbol{\theta}, \boldsymbol{\theta}^*) \stackrel{\text{def}}{=} -\log I\!E \exp\bigl\{\mu L(\boldsymbol{\theta}, \boldsymbol{\theta}^*)\bigr\} < \infty. \tag{2.1}$$

This definition can be extended to all $\mu \in \mathbb{M}$ by letting $\mathfrak{M}(\mu, \boldsymbol{\theta}, \boldsymbol{\theta}^*) = -\infty$ when the exponential moment of $\mu L(\boldsymbol{\theta}, \boldsymbol{\theta}^*)$ does not exists. The main observation behind condition (E) is that

$$I\!E \exp\bigl\{\mu L(\boldsymbol{\theta}, \boldsymbol{\theta}^*) + \mathfrak{M}(\mu, \boldsymbol{\theta}, \boldsymbol{\theta}^*)\bigr\} = 1$$

provided that $\mathfrak{M}(\mu, \boldsymbol{\theta}, \boldsymbol{\theta}^*)$ is finite. Note that $\mathfrak{M}(0, \boldsymbol{\theta}, \boldsymbol{\theta}^*) = 0$. The concentration results established in Section 4 require some *identification* properties. A pointwise *identifiability* means that one can separate the target measure $I\!P_{\boldsymbol{\theta}^*}$ and the measure $I\!P_{\boldsymbol{\theta}}$ corresponding to another point $\boldsymbol{\theta}$ in the parameter space. This condition can be expressed in terms of the function $\mathfrak{M}(\mu, \boldsymbol{\theta}, \boldsymbol{\theta}^*)$ from (2.1). Namely, this value has to be significantly positive for some $\mu \in \mathbb{M}$:

$$\mathfrak{M}^*(\boldsymbol{\theta}, \boldsymbol{\theta}^*) \stackrel{\text{def}}{=} \sup_{\mu \in \mathbb{M}} \mathfrak{M}(\mu, \boldsymbol{\theta}, \boldsymbol{\theta}^*) > 0.$$

This condition, however, only ensures a pointwise separation between $\boldsymbol{\theta}^*$ and $\boldsymbol{\theta}$. A *strict* identification requires a quantitative lower bound on the value $\mathfrak{M}^*(\boldsymbol{\theta}, \boldsymbol{\theta}^*)$: it has to grow at least logarithmically with the norm $\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|$. A precise formulation involves some more notation and it is given in Corollary 4.3 below; see condition (4.9).

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

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

Suppose that the random function $\zeta(\boldsymbol{\theta})$ is differentiable in $\boldsymbol{\theta}$ and its gradient $\nabla\zeta(\boldsymbol{\theta}) = \partial\zeta(\boldsymbol{\theta})/\partial\boldsymbol{\theta} \in I\!R^p$ fulfills the following condition:

($\boldsymbol{ED}$) *There exist some constant ν_0, a positive symmetric matrix V^2, and constant $\mathsf{g} > 0$ such that for all $|\lambda| \leq \mathsf{g}$*

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

This condition effectively means that the gradient $\nabla\zeta(\boldsymbol{\theta})$ normalized by the matrix V has bounded exponential moments. It can be relaxed by allowing the matrix V^2 and/or the value $\mathtt{g}$ to be dependent of $\boldsymbol{\theta}$ in a uniformly continuous way.

2.2 Local conditions

Local conditions describe the properties of $L(\boldsymbol{\theta})$ in a vicinity of the central point $\boldsymbol{\theta}^*$ from (1.2). First we refine condition (ED). It is fulfilled for all $\boldsymbol{\theta}$ including $\boldsymbol{\theta} = \boldsymbol{\theta}^*$. However, the matrix V can be improved if the only point $\boldsymbol{\theta}^*$ is concerned.

(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

$$\Theta_0(\mathtt{r}) \stackrel{\text{def}}{=} \{\boldsymbol{\theta} \in \Theta : \|V_0(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\| \leq \mathtt{r}\}. \tag{2.2}$$

The further conditions are restricted to such defined neighborhoods $\Theta_0(\mathtt{r})$. In fact, they quantify local smoothness properties of the log-likelihood function $L(\boldsymbol{\theta})$.

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

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

The main majorization result also requires second order smoothness of the expected log-likelihood $I\!EL(\boldsymbol{\theta})$. By definition, $L(\boldsymbol{\theta}^*, \boldsymbol{\theta}^*) \equiv 0$ and $\nabla I\!EL(\boldsymbol{\theta}^*) = 0$ because $\boldsymbol{\theta}^*$ is the extreme point of $I\!EL(\boldsymbol{\theta})$. Therefore, $-I\!EL(\boldsymbol{\theta}, \boldsymbol{\theta}^*)$ can be approximated by a quadratic function of $\boldsymbol{\theta}-\boldsymbol{\theta}^*$ in the neighborhood of $\boldsymbol{\theta}^*$. The *local identifiability* condition qualifies this quadratic approximation from above and from below on the set $\Theta_0(\mathtt{r})$ from (2.2).

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

$$\left| \frac{-2\mathbb{E}L(\boldsymbol{\theta}, \boldsymbol{\theta}^*)}{\|D_0(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\|^2} - 1 \right| \le \delta(\mathtt{r}).$$

Note that if $L(\boldsymbol{\theta}, \boldsymbol{\theta}^*)$ is the log-likelihood ratio and $\mathbb{P} = \mathbb{P}_{\boldsymbol{\theta}^*}$ then $-\mathbb{E}L(\boldsymbol{\theta}, \boldsymbol{\theta}^*) = \mathbb{E}_{\boldsymbol{\theta}^*} \log\bigl(d\mathbb{P}_{\boldsymbol{\theta}^*}/d\mathbb{P}_{\boldsymbol{\theta}}\bigr) = \mathcal{K}(\mathbb{P}_{\boldsymbol{\theta}^*}, \mathbb{P}_{\boldsymbol{\theta}})$, 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 $\delta(\mathtt{r})$ are proportional to the radius $\varrho(\mathtt{r})$ of the set $\Theta_0(\mathtt{r})$ 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.

3 Local non-asymptotic quadraticity

The *Local Asymptotic Normality* (LAN) condition since introduced by L.Le Cam 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 properties 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 a local quadratic majorization of this process from above and from below by two different quadratic in $\boldsymbol{\theta}$ processes. More precisely, we apply the *sandwiching* 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 $\mathbb{L}_{\boldsymbol{\epsilon}}(\boldsymbol{\theta},\boldsymbol{\theta}^*)$:

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

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 depends 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: value of the problem and deficiency. It turns out that the most of statements like confidence and concentration probability rely upon the maximum of $L(\boldsymbol{\theta},\boldsymbol{\theta}^*)$ over $\boldsymbol{\theta}$ which we call *the value of the problem*. Due to (3.1) this value 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 *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 values. The *deficiency* $\Delta_{\boldsymbol{\epsilon}}$ can be defined as the width of the interval bounding the value of the problem due to (3.1), that is, as the sum of the approximation errors and of this difference:

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

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 deficiency." 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 majorization

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 $I\!P\bigl(\widetilde{\boldsymbol{\theta}} \notin \Theta_0(\mathtt{r})\bigr)$ is sufficiently small. Precise results about the choice of $\mathtt{r}$ which ensures this property are postponed until Section 4.2. In this neighborhood $\Theta_0(\mathtt{r})$ we aim to build a quadratic majorization of the process $L(\boldsymbol{\theta})$. The first step is the usual decomposition of this process into deterministic and stochastic components:

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

where $\zeta(\boldsymbol{\theta}) = L(\boldsymbol{\theta}) - I\!\!E L(\boldsymbol{\theta})$. Condition $(\mathcal{L}_0)$ allows for approximating the smooth deterministic function $I\!\!E L(\boldsymbol{\theta}) - I\!\!E L(\boldsymbol{\theta}^*)$ 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 $\Theta_0(\mathtt{r})$:

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

This expansion is used in the most of asymptotic 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 terms which is either shrunk or stretched relative to the term $\|D_0(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\|^2/2$ in $\mathbb{L}(\boldsymbol{\theta}, \boldsymbol{\theta}^*)$.

Introduce for a vector $\boldsymbol{\epsilon} = (\delta, \varrho)$ the following notation:

$$\begin{aligned} \mathbb{L}_{\boldsymbol{\epsilon}}(\boldsymbol{\theta}, \boldsymbol{\theta}^*) &\stackrel{\text{def}}{=} (\boldsymbol{\theta} - \boldsymbol{\theta}^*)^\top \nabla L(\boldsymbol{\theta}^*) - \|D_{\boldsymbol{\epsilon}}(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\|^2/2 \\ &= \boldsymbol{\xi}_{\boldsymbol{\epsilon}}^\top D_{\boldsymbol{\epsilon}}(\boldsymbol{\theta} - \boldsymbol{\theta}^*) - \|D_{\boldsymbol{\epsilon}}(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\|^2/2, \end{aligned} \tag{3.2}$$

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 (3.2) 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 $\mathbb{L}(\boldsymbol{\theta},\boldsymbol{\theta}^*)$ and stretched if δ, ϱ are negative. Now, given $\mathtt{r}$, define $\delta = \delta(\mathtt{r})$, $\varrho = \omega(\mathtt{r})/(3\nu_0)$ 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

$$D^2_{\underline{\boldsymbol{\epsilon}}} = D_0^2(1+\delta) + \varrho V_0^2.$$

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 $D_0^2(1-\delta) - \varrho V_0^2 \geq 0$. Set $\boldsymbol{\epsilon} = (\delta, \varrho)$, $\underline{\boldsymbol{\epsilon}} = -\boldsymbol{\epsilon} = (-\delta, -\varrho)$. Then*

$$\mathbb{L}_{\underline{\boldsymbol{\epsilon}}}(\boldsymbol{\theta},\boldsymbol{\theta}^*) - \diamondsuit_{\underline{\boldsymbol{\epsilon}}}(\mathtt{r}) \leq L(\boldsymbol{\theta},\boldsymbol{\theta}^*) \leq \mathbb{L}_{\boldsymbol{\epsilon}}(\boldsymbol{\theta},\boldsymbol{\theta}^*) + \diamondsuit_{\boldsymbol{\epsilon}}(\mathtt{r}), \quad \boldsymbol{\theta} \in \Theta_0(\mathtt{r}), \tag{3.3}$$

with $\mathbb{L}_{\boldsymbol{\epsilon}}(\boldsymbol{\theta},\boldsymbol{\theta}^*), \mathbb{L}_{\underline{\boldsymbol{\epsilon}}}(\boldsymbol{\theta},\boldsymbol{\theta}^*)$ defined by (3.2)*. Moreover, the random variable $\diamondsuit_{\boldsymbol{\epsilon}}(\mathtt{r})$ fulfills*

$$I\!P\bigl\{\varrho^{-1}\diamondsuit_{\boldsymbol{\epsilon}}(\mathtt{r}) \geq \mathfrak{z}_0(\mathtt{x},p)\bigr\} \leq \exp\bigl(-\mathtt{x}\bigr) \tag{3.4}$$

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

$$\mathfrak{z}_0(\mathtt{x},p) \stackrel{\text{def}}{=} \begin{cases} \bigl(1+\sqrt{\mathtt{x}+\mathfrak{c}_1 p}\bigr)^2 & \text{if } 1+\sqrt{\mathtt{x}+\mathfrak{c}_1 p} \leq \mathtt{g}_0, \\ 1 + (1+2\mathtt{g}_0^{-1})^2\bigl(\mathtt{g}_0^{-1}(\mathtt{x}+\mathfrak{c}_1 p) + \mathtt{g}_0/2\bigr)^2 & \text{otherwise.} \end{cases}$$

where $\mathfrak{c}_1 = 2$ for $p \geq 2$ and $\mathfrak{c}_1 = 2.4$ for $p = 1$. Similarly for $\diamondsuit_{\underline{\boldsymbol{\epsilon}}}(\mathtt{r})$.

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

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

In view of $I\!E\nabla L(\boldsymbol{\theta}^*) = 0$, this definition can be rewritten in the form

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

Similarly define

$$\begin{aligned} \diamondsuit_{\underline{\boldsymbol{\epsilon}}}(\mathtt{r}) &\stackrel{\text{def}}{=} \sup_{\boldsymbol{\theta} \in \Theta_0(\mathtt{r})} \bigl\{ L(\boldsymbol{\theta}^*,\boldsymbol{\theta}) - I\!E L(\boldsymbol{\theta}^*,\boldsymbol{\theta}) - (\boldsymbol{\theta}^*-\boldsymbol{\theta})^\top \nabla L(\boldsymbol{\theta}^*) - \frac{\varrho}{2}\|V_0(\boldsymbol{\theta}-\boldsymbol{\theta}^*)\|^2 \bigr\} \\ &= \sup_{\boldsymbol{\theta} \in \Theta_0(\mathtt{r})} \bigl\{ \zeta(\boldsymbol{\theta}^*,\boldsymbol{\theta}) - (\boldsymbol{\theta}^*-\boldsymbol{\theta})^\top \nabla \zeta(\boldsymbol{\theta}^*) - \frac{\varrho}{2}\|V_0(\boldsymbol{\theta}-\boldsymbol{\theta}^*)\|^2 \bigr\}. \end{aligned}$$

Now the claim of the theorem can be easily reduced to an exponential bound for the quantities $\diamondsuit_{\boldsymbol{\epsilon}}(\mathtt{r}), \diamondsuit_{\underline{\boldsymbol{\epsilon}}}(\mathtt{r})$. We apply Theorem 7.9 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 7.9 are fulfilled yielding (3.4) in view of $\varrho \geq 3 \nu_0 \, \omega(\mathtt{r})$. □

Remark 3.1. If $\mathtt{x}$ is not too big, then the value $\mathfrak{z}_0(\mathtt{x}, p)$ is close to $\mathtt{x} + \mathfrak{c}_1 p$; cf. (7.7). The bound (3.4) tells us that the errors $\diamondsuit_{\boldsymbol{\epsilon}}(\mathtt{r})$ and $\diamondsuit_{\underline{\boldsymbol{\epsilon}}}(\mathtt{r})$ are of order $\omega(\mathtt{r}) \, p$.

3.2 Local inference. Deficiency

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 *majorization* (3.1) and the *geometric structure* of the processes $\mathbb{L}_{\boldsymbol{\epsilon}}$ and $\mathbb{L}_{\underline{\boldsymbol{\epsilon}}}$.

The random quantity $\sup_{\boldsymbol{\theta}} \mathbb{L}_{\boldsymbol{\epsilon}}(\boldsymbol{\theta}, \boldsymbol{\theta}^*)$ can be called the *value of the upper problem*, while $\sup_{\boldsymbol{\theta}} \mathbb{L}_{\underline{\boldsymbol{\epsilon}}}(\boldsymbol{\theta}, \boldsymbol{\theta}^*)$ is the *value of the lower problem*. 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. $\boldsymbol{\theta}$. The value of the original problem which is the maximum of the original process $L(\boldsymbol{\theta}, \boldsymbol{\theta}^*)$ over $\boldsymbol{\theta}$ is sandwiched between the similar expressions for the two approximating processes in view of the approximation bound (3.3). This suggests to measure the quality of approximation by the difference between values of the upper and lower approximating problems. The approximating quality is sufficiently good if this difference is smaller in order than the value itself.

3.2.1 Upper and lower values

First consider the maximization problem for the upper approximating processes $\mathbb{L}_{\boldsymbol{\epsilon}}(\boldsymbol{\theta}, \boldsymbol{\theta}^*)$. This is a quadratic optimization with the closed form solution $\|\boldsymbol{\xi}_{\boldsymbol{\epsilon}}\|^2 / 2$. Moreover, the Euclidean norm of the random vector $\boldsymbol{\xi}_{\boldsymbol{\epsilon}}$ behaves nearly as a chi-squared random variable. The results below make these statements more concise. Define for

$$I\!B_{\boldsymbol{\epsilon}} \stackrel{\text{def}}{=} \frac{D_{\boldsymbol{\epsilon}}^{-1} V_0^2 D_{\boldsymbol{\epsilon}}^{-1}}{\lambda_{\max}\bigl(D_{\boldsymbol{\epsilon}}^{-1} V_0^2 D_{\boldsymbol{\epsilon}}^{-1}\bigr)}, \qquad \mathtt{p}_{\boldsymbol{\epsilon}} \stackrel{\text{def}}{=} \operatorname{tr}\bigl(I\!B_{\boldsymbol{\epsilon}}\bigr), \qquad \mathtt{v}_{\boldsymbol{\epsilon}}^2 \stackrel{\text{def}}{=} 2 \operatorname{tr}(I\!B_{\boldsymbol{\epsilon}}^2). \tag{3.5}$$

Moreover, with the constant $\mathtt{g}$ from (ED_0), define also $\mu_c = \mathtt{g}^2/(\mathtt{p}_\epsilon + \mathtt{g}^2)$, and

$$\begin{aligned} \mathtt{y}_c^2 &\stackrel{\text{def}}{=} \mathtt{g}^2/\mu_c^2 - \mathtt{p}_\epsilon/\mu_c, \\ \mathtt{g}_c &\stackrel{\text{def}}{=} \mu_c \mathtt{y}_c = \sqrt{\mathtt{g}^2 - \mu_c \mathtt{p}_\epsilon}, \\ 2\mathtt{x}_c &\stackrel{\text{def}}{=} \mathtt{g}_c \mathtt{y}_c + \log\det\bigl(I_p - \mu_c I\!B_\epsilon^2\bigr). \end{aligned}$$

To gain some feeling of these quantities consider a special case with $\mathtt{g}^2 = \mathtt{p}_\epsilon$. Then $\mu_c = 1/2$, and the inequality $x + \log(1-x) \geq x/3$ for $0 \leq x \leq 1/2$ implies

$$\begin{aligned} \mathtt{y}_c &= 4\mathtt{g}^2 - 2\mathtt{p}_\epsilon = 2\mathtt{p}_\epsilon, \\ \mathtt{g}_c^2 &= (\mathtt{y}_c/2)^2 = \mathtt{p}_\epsilon/2, \\ 2\mathtt{x}_c &= \sqrt{\mathtt{p}_\epsilon/2 \cdot 2\mathtt{p}_\epsilon} + \log\det\bigl(I_p - I\!B_\epsilon/2\bigr) \geq \mathtt{p}_\epsilon/3. \end{aligned}$$

Theorem 3.2. *Assume* (ED_0) *with* $\nu_0 = 1$ *and* $\mathtt{g}^2 \geq 2\mathtt{p}_\epsilon$. *It holds*

$$\sup_{\boldsymbol{\theta} \in I\!R^p} \mathbb{L}_\epsilon(\boldsymbol{\theta}, \boldsymbol{\theta}^*) = \|\boldsymbol{\xi}_\epsilon\|^2/2. \tag{3.6}$$

Moreover, $I\!E\|\boldsymbol{\xi}_\epsilon\|^2 \leq \mathtt{p}_\epsilon$, *and for each* $\mathtt{x} > 0$

$$I\!P\bigl(\|\boldsymbol{\xi}_\epsilon\|^2 \geq \mathfrak{z}(\mathtt{x}, I\!B_\epsilon)\bigr) \leq 2\mathrm{e}^{-\mathtt{x}} + 8.4\mathrm{e}^{-\mathtt{x}_c}, \tag{3.7}$$

where $\mathfrak{z}(\mathtt{x}, I\!B_\epsilon)$ *is defined by*

$$\mathfrak{z}(\mathtt{x}, I\!B_\epsilon) \stackrel{\text{def}}{=} \begin{cases} \mathtt{p}_\epsilon + \sqrt{2\mathtt{x}\mathtt{v}_\epsilon}, & \mathtt{x} \leq \mathtt{v}_\epsilon/18, \\ \mathtt{p}_\epsilon + 6\mathtt{x} & \mathtt{v}_\epsilon/18 < \mathtt{x} \leq \mathtt{x}_c, \\ \bigl|\mathtt{y}_c + 2(\mathtt{x} - \mathtt{x}_c)/\mathtt{g}_c\bigr|^2, & \mathtt{x} > \mathtt{x}_c. \end{cases}$$

Proof. The unconstrained maximum of the quadratic form $\mathbb{L}_\epsilon(\boldsymbol{\theta}, \boldsymbol{\theta}^*)$ w.r.t. $\boldsymbol{\theta}$ is attained at $\widetilde{\boldsymbol{\theta}}$ yielding the expression (3.6). The second moment of $\boldsymbol{\xi}_\epsilon$ can be bounded from condition (ED_0). Indeed,

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

and (ED_0) implies $\boldsymbol{\gamma}^\top \operatorname{Var}\bigl\{\nabla L(\boldsymbol{\theta}^*)\bigr\}\boldsymbol{\gamma} \leq \boldsymbol{\gamma}^\top V_0^2 \boldsymbol{\gamma}$ and thus, $I\!E\|\boldsymbol{\xi}_\epsilon\|^2 \leq \mathtt{p}_\epsilon$. The deviation bound (3.7) is proved in Corollary 6.12. □

Next result describes the *lower value* $\sup_{\boldsymbol{\theta}} \mathbb{L}_{\underline{\epsilon}}(\boldsymbol{\theta}, \boldsymbol{\theta}^*)$ and the *difference* between upper and lower values. The proof is straightforward.

Theorem 3.3. *On the random set* $\{\|\boldsymbol{\xi}_{\underline{\boldsymbol{\epsilon}}}\| \leq \mathtt{r}\}$, *it holds*

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

Moreover, $\|\boldsymbol{\xi}_{\underline{\boldsymbol{\epsilon}}}\| \leq \|\boldsymbol{\xi}_{\boldsymbol{\epsilon}}\|$ *and*

$$\|\boldsymbol{\xi}_{\boldsymbol{\epsilon}}\|^2 - \|\boldsymbol{\xi}_{\underline{\boldsymbol{\epsilon}}}\|^2 = \boldsymbol{\xi}_{\boldsymbol{\epsilon}}^\top \bigl(I_p - D_{\boldsymbol{\epsilon}} D_{\underline{\boldsymbol{\epsilon}}}^{-2} D_{\boldsymbol{\epsilon}}\bigr) \boldsymbol{\xi}_{\boldsymbol{\epsilon}} \leq \alpha_{\boldsymbol{\epsilon}} \|\boldsymbol{\xi}_{\boldsymbol{\epsilon}}\|^2,$$

with

$$\alpha_{\boldsymbol{\epsilon}} \stackrel{\text{def}}{=} \|I_p - D_{\boldsymbol{\epsilon}} D_{\underline{\boldsymbol{\epsilon}}}^{-2} D_{\boldsymbol{\epsilon}}\|_\infty = \lambda_{\max}\bigl(I_p - D_{\boldsymbol{\epsilon}} D_{\underline{\boldsymbol{\epsilon}}}^{-2} D_{\boldsymbol{\epsilon}}\bigr). \tag{3.8}$$

If the value $\alpha_{\boldsymbol{\epsilon}}$ is small then the difference $\|\boldsymbol{\xi}_{\boldsymbol{\epsilon}}\|^2 - \|\boldsymbol{\xi}_{\underline{\boldsymbol{\epsilon}}}\|^2$ is automatically smaller than the upper value $\|\boldsymbol{\xi}_{\boldsymbol{\epsilon}}\|^2$.

3.2.2 Deficiency

In view of the results of Theorems 3.2 and 3.3, the sandwiching approach (3.3) bounds the value $L(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}^*) = \sup_{\boldsymbol{\theta} \in \Theta} L(\boldsymbol{\theta}, \boldsymbol{\theta}^*)$ in the interval $\bigl[\|\boldsymbol{\xi}_{\underline{\boldsymbol{\epsilon}}}\|^2 - \diamondsuit_{\underline{\boldsymbol{\epsilon}}}(\mathtt{r}), \|\boldsymbol{\xi}_{\boldsymbol{\epsilon}}\|^2 + \diamondsuit_{\boldsymbol{\epsilon}}(\mathtt{r})\bigr]$. The width of this interval describes the accuracy of the approach. By analogy to the general Le Cam theory of statistical experiments, this value will be called the *deficiency*. Define the value $\Delta_{\boldsymbol{\epsilon}}(\mathtt{r})$ by

$$\Delta_{\boldsymbol{\epsilon}}(\mathtt{r}) \stackrel{\text{def}}{=} \diamondsuit_{\boldsymbol{\epsilon}}(\mathtt{r}) + \diamondsuit_{\underline{\boldsymbol{\epsilon}}}(\mathtt{r}) + \bigl(\|\boldsymbol{\xi}_{\boldsymbol{\epsilon}}\|^2 - \|\boldsymbol{\xi}_{\underline{\boldsymbol{\epsilon}}}\|^2\bigr)/2. \tag{3.9}$$

This quantity is random but it can be easily evaluated under the considered conditions. Indeed, the approximation errors $\diamondsuit_{\boldsymbol{\epsilon}}(\mathtt{r}), \diamondsuit_{\underline{\boldsymbol{\epsilon}}}(\mathtt{r})$ can be bounded by $\varrho\, \mathfrak{z}_0(\mathtt{x}, p)$ with the probability at least $1 - 2\mathrm{e}^{-\mathtt{x}}$; see (3.4). Also $\|\boldsymbol{\xi}_{\boldsymbol{\epsilon}}\|^2 \leq \mathfrak{z}(\mathtt{x}, I\!\!B_{\boldsymbol{\epsilon}})$ with a probability of order $1 - 2\mathrm{e}^{-\mathtt{x}}$; see (3.7). This yields for the deficiency $\Delta_{\boldsymbol{\epsilon}}(\mathtt{r})$ with a probability about $1 - 4\mathrm{e}^{-\mathtt{x}}$

$$\Delta_{\boldsymbol{\epsilon}}(\mathtt{r}) \leq 2\varrho\, \mathfrak{z}_0(\mathtt{x}, p) + \alpha_{\boldsymbol{\epsilon}}\, \mathfrak{z}(\mathtt{x}, I\!\!B_{\boldsymbol{\epsilon}}).$$

3.2.3 The regular case

The bound (3.9) can be further specified in the so called *regular case* under the condition

$$V_0 \leq \mathfrak{a} D_0\,. \tag{3.10}$$

If the parametric assumption is correct, that is, $I\!\!P = I\!\!P_{\boldsymbol{\theta}^*}$ for a regular parametric family, then the both matrices coincide with the total Fisher information matrix, and the regularity condition is fulfilled automatically with $\mathfrak{a} = 1$. Otherwise, the regularity

means that the local variability of the process $L(\boldsymbol{\theta})$ measured by the matrix V_0 is not significantly larger than the local information measured by the matrix D_0.

Theorem 3.4. *Suppose* (3.10). *Then*

$$D_{\boldsymbol{\epsilon}}^2 \geq (1-\delta-\varrho\mathfrak{a}^2)D_0^2, \qquad \alpha_{\boldsymbol{\epsilon}} = \|I_p - D_{\boldsymbol{\epsilon}} D_{\underline{\boldsymbol{\epsilon}}}^{-2} D_{\boldsymbol{\epsilon}}\|_\infty \leq \frac{2(\delta+\varrho\mathfrak{a}^2)}{1-\delta-\varrho\mathfrak{a}^2}.$$

Moreover, $I\!B_{\boldsymbol{\epsilon}}^2 = D_{\boldsymbol{\epsilon}}^{-1} V_0^2 D_{\boldsymbol{\epsilon}}^{-1}$ *satisfies with* $\mathfrak{a}_{\boldsymbol{\epsilon}} \stackrel{\text{def}}{=} \mathfrak{a}(1-\delta-\varrho\mathfrak{a}^2)$ *:*

$$I\!B_{\boldsymbol{\epsilon}}^2 \leq \mathfrak{a}_{\boldsymbol{\epsilon}}^{-2} I_p, \qquad \mathtt{p}_{\boldsymbol{\epsilon}} \leq \mathfrak{a}_{\boldsymbol{\epsilon}}^{-2} p, \qquad \mathtt{v}_{\boldsymbol{\epsilon}}^2 \leq 2\mathfrak{a}_{\boldsymbol{\epsilon}}^{-4} p, \qquad \lambda_{\boldsymbol{\epsilon}} \leq \mathfrak{a}_{\boldsymbol{\epsilon}}^{-2}. \tag{3.11}$$

Proof. The results follow directly from the definition of $D_{\boldsymbol{\epsilon}}$ and $D_{\underline{\boldsymbol{\epsilon}}}$ and (3.5) by making use of (3.10). □

In particular, the matrices $D_{\boldsymbol{\epsilon}}$ and $D_{\underline{\boldsymbol{\epsilon}}}$ are close to each other if δ and $\varrho\mathfrak{a}^2$ are small. So, all we need in the regular case, is a large deviation bound for the probability $I\!P\bigl(\widetilde{\boldsymbol{\theta}} \notin \Theta_0(\mathtt{r})\bigr)$ and that the quantities $\omega(\mathtt{r})$ and $\delta(\mathtt{r})$ are small.

3.2.4 Local coverage probability

Now we state some immediate corollaries of the exponential bound from Theorem 3.1. First we study the probability of covering $\boldsymbol{\theta}^*$ by the random set $\mathcal{E}(\mathfrak{z}) = \{\boldsymbol{\theta} : 2L(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}) \leq \mathfrak{z}\}$.

Theorem 3.5. *Suppose* (ED_0), (ED_1), *and* $(\mathcal{L}_0)$ *on* $\Theta_0(\mathtt{r})$, *and let* $\varrho \geq 3\nu_0\,\omega(\mathtt{r})$, $\delta \geq \delta(\mathtt{r})$, *and* $D_0^2(1-\delta) - \varrho V_0^2 \geq 0$. *Then for any* $\mathfrak{z} > 0$, *it holds*

$$\begin{aligned} I\!P\bigl\{\mathcal{E}(\mathfrak{z}) \not\ni \boldsymbol{\theta}^*, \widetilde{\boldsymbol{\theta}} \in \Theta_0(\mathtt{r})\bigr\} &= I\!P\bigl\{2L(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}^*) > \mathfrak{z}, \widetilde{\boldsymbol{\theta}} \in \Theta_0(\mathtt{r})\bigr\} \\ &\leq I\!P\bigl\{\|\boldsymbol{\xi}_{\boldsymbol{\epsilon}}\|^2 \geq \mathfrak{z} - \diamondsuit_{\boldsymbol{\epsilon}}(\mathtt{r})\bigr\}. \end{aligned} \tag{3.12}$$

Proof. The bound (3.12) follows from the upper bound of Theorem 3.1 and the statement (3.6) of Lemma 3.2. □

The exponential bound (3.7) helps to answer a very important question about a proper choice of the critical value $\mathfrak{z}$ providing the prescribed covering probability. Namely, this probability starts to decrease gradually when $\mathfrak{z}$ grows over $\mathfrak{z}(\mathtt{x}, I\!B_{\boldsymbol{\epsilon}})$.

3.2.5 Local concentration

Now we describe local 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 $\widetilde{\boldsymbol{\theta}}$ does not belong to the set

$\mathcal{A}_{\boldsymbol{\epsilon}}(z)$ of the form

$$\mathcal{A}_{\boldsymbol{\epsilon}}(z) = \bigl\{\boldsymbol{\theta} : \|D_{\boldsymbol{\epsilon}}(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\| \le z\bigr\}.$$

It is obvious that

$$\bigl\{\widetilde{\boldsymbol{\theta}} \notin \mathcal{A}_{\boldsymbol{\epsilon}}(z)\bigr\} = \bigl\{\|D_{\boldsymbol{\epsilon}}(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\| > z\bigr\}.$$

Theorem 3.6. *Assume* (ED_0), (ED_1), *and* $(\mathcal{L}_0)$ *on* $\Theta_0(\mathtt{r})$, *and let* $\varrho \ge 3\nu_0\,\omega(\mathtt{r})$, $\delta \ge \delta(\mathtt{r})$, *and* $D_0^2(1-\delta) - \varrho V_0^2 \ge 0$. *Then for any* $z > 0$, *it holds with* $\alpha_{\boldsymbol{\epsilon}}$ *from* (3.8)

$$\mathbb{P}\bigl\{\|D_{\boldsymbol{\epsilon}}(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\| > z,\, \widetilde{\boldsymbol{\theta}} \in \Theta_0(\mathtt{r})\bigr\} \le \mathbb{P}\bigl\{\|\boldsymbol{\xi}_{\boldsymbol{\epsilon}}\| > z - \sqrt{2\Delta_{\boldsymbol{\epsilon}}(\mathtt{r})}\bigr\} \tag{3.13}$$

$$\le \mathbb{P}\Bigl\{\bigl(1 - \sqrt{\alpha_{\boldsymbol{\epsilon}}}\bigr)\|\boldsymbol{\xi}_{\boldsymbol{\epsilon}}\| > z - \sqrt{2\diamondsuit_{\boldsymbol{\epsilon}}(\mathtt{r}) + 2\diamondsuit_{\underline{\boldsymbol{\epsilon}}}(\mathtt{r})}\Bigr\}. \tag{3.14}$$

Proof. It obviously holds on the set $\{\widetilde{\boldsymbol{\theta}} \in \Theta_0(\mathtt{r})\}$

$$\begin{aligned}\{\widetilde{\boldsymbol{\theta}} \notin \mathcal{A}_{\boldsymbol{\epsilon}}(z)\} &= \Bigl\{\sup_{\boldsymbol{\theta} \notin \mathcal{A}_{\boldsymbol{\epsilon}}(z)} L(\boldsymbol{\theta}, \boldsymbol{\theta}^*) = \sup_{\boldsymbol{\theta}} L(\boldsymbol{\theta}, \boldsymbol{\theta}^*)\Bigr\} \\ &\subseteq \Bigl\{\sup_{\boldsymbol{\theta} \notin \mathcal{A}_{\boldsymbol{\epsilon}}(z)} \mathbb{L}_{\boldsymbol{\epsilon}}(\boldsymbol{\theta}, \boldsymbol{\theta}^*) + \diamondsuit_{\boldsymbol{\epsilon}}(\mathtt{r}) \ge \sup_{\boldsymbol{\theta}} \mathbb{L}_{\underline{\boldsymbol{\epsilon}}}(\boldsymbol{\theta}, \boldsymbol{\theta}^*) - \diamondsuit_{\underline{\boldsymbol{\epsilon}}}(\mathtt{r})\Bigr\}.\end{aligned}$$

As $\mathbb{L}_{\boldsymbol{\epsilon}}(\boldsymbol{\theta}, \boldsymbol{\theta}^*)$ is a quadratic function of $D_{\boldsymbol{\epsilon}}(\boldsymbol{\theta} - \boldsymbol{\theta}^*)$; cf. (3.2), its maximum on the complement $\mathcal{A}_{\boldsymbol{\epsilon}}^{\mathrm{c}}(z)$ of the set $\mathcal{A}_{\boldsymbol{\epsilon}}(z)$ is attained at the point $\boldsymbol{\theta}$ satisfying $D_{\boldsymbol{\epsilon}}(\boldsymbol{\theta} - \boldsymbol{\theta}^*) = \gamma\boldsymbol{\xi}_{\boldsymbol{\epsilon}}$ with $\gamma = z/\|\boldsymbol{\xi}_{\boldsymbol{\epsilon}}\|$. This implies for all $\boldsymbol{\theta} \notin \mathcal{A}_{\boldsymbol{\epsilon}}(z)$

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

By Lemma 3.2 $\sup_{\boldsymbol{\theta}} \mathbb{L}_{\underline{\boldsymbol{\epsilon}}}(\boldsymbol{\theta}, \boldsymbol{\theta}^*) = \|\boldsymbol{\xi}_{\underline{\boldsymbol{\epsilon}}}\|^2/2$. Therefore,

$$\begin{aligned}\{\widetilde{\boldsymbol{\theta}} \notin \mathcal{A}_{\boldsymbol{\epsilon}}(z)\} &\subseteq \bigl\{z\|\boldsymbol{\xi}_{\boldsymbol{\epsilon}}\| - z^2/2 \ge \|\boldsymbol{\xi}_{\underline{\boldsymbol{\epsilon}}}\|^2/2 - \diamondsuit_{\boldsymbol{\epsilon}}(\mathtt{r}) - \diamondsuit_{\underline{\boldsymbol{\epsilon}}}(\mathtt{r})\bigr\} \\ &= \bigl\{z^2/2 - z\|\boldsymbol{\xi}_{\boldsymbol{\epsilon}}\| + \|\boldsymbol{\xi}_{\boldsymbol{\epsilon}}\|^2/2 \le \Delta_{\boldsymbol{\epsilon}}(\mathtt{r})\bigr\}\end{aligned}$$

and (3.13) follows. Further, the bound $\|\boldsymbol{\xi}_{\boldsymbol{\epsilon}}\|^2 - \|\boldsymbol{\xi}_{\underline{\boldsymbol{\epsilon}}}\|^2 \le \alpha_{\boldsymbol{\epsilon}}\|\boldsymbol{\xi}_{\boldsymbol{\epsilon}}\|^2$ implies

$$\sqrt{2\Delta_{\boldsymbol{\epsilon}}(\mathtt{r})} \le \sqrt{2\diamondsuit_{\boldsymbol{\epsilon}}(\mathtt{r}) + 2\diamondsuit_{\underline{\boldsymbol{\epsilon}}}(\mathtt{r})} + \sqrt{\alpha_{\boldsymbol{\epsilon}}}\|\boldsymbol{\xi}_{\boldsymbol{\epsilon}}\|$$

which yields (3.14). □

An interesting and important question is for which z the probability of the event $\{\|D_{\boldsymbol{\epsilon}}(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\| > z\}$ becomes small. We use that $\max\bigl\{\diamondsuit_{\boldsymbol{\epsilon}}(\mathtt{r}), \diamondsuit_{\underline{\boldsymbol{\epsilon}}}(\mathtt{r})\bigr\} \le \varrho\,\mathfrak{z}_0(\mathtt{x}, p)$ and on a set of probability at least $1 - 2\mathrm{e}^{-\mathtt{x}}$. This and the bound (3.14) imply

$$\sqrt{2\Delta_{\boldsymbol{\epsilon}}(\mathtt{r})} \le 2\sqrt{\varrho\,\mathfrak{z}_0(\mathtt{x}, p)} + \sqrt{\alpha_{\boldsymbol{\epsilon}}}\|\boldsymbol{\xi}_{\boldsymbol{\epsilon}}\|.$$

Now it follows from the local concentration result of Theorem 3.6:

$$\begin{aligned} &I\!P\bigl\{\|D_{\boldsymbol{\epsilon}}(\widetilde{\boldsymbol{\theta}}-\boldsymbol{\theta}^*)\| > z,\, \widetilde{\boldsymbol{\theta}} \in \Theta_0(\mathtt{r})\bigr\} \\ &\quad \le I\!P\Bigl\{\bigl(1-\sqrt{\alpha_{\boldsymbol{\epsilon}}}\bigr)\|\boldsymbol{\xi}_{\boldsymbol{\epsilon}}\| > z - 2\sqrt{\varrho\, \mathfrak{z}_0(\mathtt{x},p)}\Bigr\} + 2\mathrm{e}^{-\mathtt{x}}. \end{aligned}$$

The probability $I\!P\bigl(\|\boldsymbol{\xi}_{\boldsymbol{\epsilon}}\| > z\bigr)$ starts to vanish when z^2 significantly exceeds $\mathtt{p}_{\boldsymbol{\epsilon}} = I\!E\|\boldsymbol{\xi}_{\boldsymbol{\epsilon}}\|^2$. Under the regularity conditions, the value $\mathtt{p}_{\boldsymbol{\epsilon}}$ is of order p. Moreover, $\mathfrak{z}_0(\mathtt{x},p)$ is also of order p for moderate $\mathtt{x}$. If ϱ and $\alpha_{\boldsymbol{\epsilon}}$ are small then the latter deviation probability can be bounded by $I\!P\bigl(\|\boldsymbol{\xi}_{\boldsymbol{\epsilon}}\| > z'\bigr)$ with $z/z' \approx 1$ which can be evaluated by (3.7).

3.2.6 Local expansions

Now we show how the bound (3.3) can be used for obtaining a local expansion of the quasi MLE $\widetilde{\boldsymbol{\theta}}$. The basic idea is to plug $\widetilde{\boldsymbol{\theta}}$ in place of $\boldsymbol{\theta}$ in the definition of $\diamondsuit_{\boldsymbol{\epsilon}}(\mathtt{r})$.

Theorem 3.7. *Assume* (ED_0), (ED_1), *and* $(\mathcal{L}_0)$ *on* $\Theta_0(\mathtt{r})$ *and let* $\varrho \ge 3\nu_0\,\omega(\mathtt{r})$, $\delta \ge \delta(\mathtt{r})$, *and* $D_0^2(1-\delta) - \varrho V_0^2 \ge 0$. *Then the following approximation holds on the random set* $\mathcal{C}_{\boldsymbol{\epsilon}}(\mathtt{r}) = \{\widetilde{\boldsymbol{\theta}} \in \Theta_0(\mathtt{r}), \|\boldsymbol{\xi}_{\underline{\boldsymbol{\epsilon}}}\| \le \mathtt{r}\}$*:*

$$\|\boldsymbol{\xi}_{\underline{\boldsymbol{\epsilon}}}\|^2/2 - \diamondsuit_{\underline{\boldsymbol{\epsilon}}}(\mathtt{r}) \le L(\widetilde{\boldsymbol{\theta}},\boldsymbol{\theta}^*) \le \|\boldsymbol{\xi}_{\boldsymbol{\epsilon}}\|^2/2 + \diamondsuit_{\boldsymbol{\epsilon}}(\mathtt{r}). \tag{3.15}$$

Moreover, it holds on the same random set $\mathcal{C}_{\boldsymbol{\epsilon}}(\mathtt{r})$

$$\bigl\|D_{\boldsymbol{\epsilon}}\bigl(\widetilde{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\bigr) - \boldsymbol{\xi}_{\boldsymbol{\epsilon}}\bigr\|^2 \le 2\Delta_{\boldsymbol{\epsilon}}(\mathtt{r}). \tag{3.16}$$

Proof. The bound (3.3) together with Lemma 3.2 yield on $\mathcal{C}_{\boldsymbol{\epsilon}}(\mathtt{r})$

$$\begin{aligned} L(\widetilde{\boldsymbol{\theta}},\boldsymbol{\theta}^*) &= \sup_{\boldsymbol{\theta}\in\Theta_0(\mathtt{r})} L(\boldsymbol{\theta},\boldsymbol{\theta}^*) \\ &\ge \sup_{\boldsymbol{\theta}\in\Theta_0(\mathtt{r})} \mathbb{L}_{\underline{\boldsymbol{\epsilon}}}(\boldsymbol{\theta},\boldsymbol{\theta}^*) - \diamondsuit_{\underline{\boldsymbol{\epsilon}}}(\mathtt{r}) = \|\boldsymbol{\xi}_{\underline{\boldsymbol{\epsilon}}}\|^2/2 - \diamondsuit_{\underline{\boldsymbol{\epsilon}}}(\mathtt{r}). \end{aligned} \tag{3.17}$$

Similarly

$$L(\widetilde{\boldsymbol{\theta}},\boldsymbol{\theta}^*) \le \sup_{\boldsymbol{\theta}\in\Theta_0(\mathtt{r})} \mathbb{L}_{\boldsymbol{\epsilon}}(\boldsymbol{\theta},\boldsymbol{\theta}^*) + \diamondsuit_{\boldsymbol{\epsilon}}(\mathtt{r}) \le \|\boldsymbol{\xi}_{\boldsymbol{\epsilon}}\|^2/2 + \diamondsuit_{\boldsymbol{\epsilon}}(\mathtt{r}) \tag{3.18}$$

yielding (3.15). For getting (3.16), 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 (3.17)

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

or, equivalently

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

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

3.2.7 A local risk bound

Below we also bound the moments of the excess $L(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}^*)$ when $\widetilde{\boldsymbol{\theta}}$ is restricted to the local vicinity $\Theta_0(\mathtt{r})$ of $\boldsymbol{\theta}^*$.

Theorem 3.8. *Assume* (ED_0), (ED_1), *and* $(\mathcal{L}_0)$ *on* $\Theta_0(\mathtt{r})$ *and let* $\varrho \geq 3\nu_0\,\omega(\mathtt{r})$, $\delta \geq \delta(\mathtt{r})$, *and* $D_0^2(1-\delta) - \varrho V_0^2 \geq 0$. *Then for* $u > 0$

$$I\!\!E L^u(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}^*)\, 1\!\!1(\widetilde{\boldsymbol{\theta}} \in \Theta_0(\mathtt{r})) \leq I\!\!E\bigl[\|\boldsymbol{\xi}_{\boldsymbol{\epsilon}}\|^2/2 + \diamondsuit_{\boldsymbol{\epsilon}}(\mathtt{r})\bigr]^u. \tag{3.19}$$

Moreover, for $\mathcal{C}_{\boldsymbol{\epsilon}}(\mathtt{r}) = \{\widetilde{\boldsymbol{\theta}} \in \Theta_0(\mathtt{r}), \|\boldsymbol{\xi}_{\boldsymbol{\epsilon}}\| \leq \mathtt{r}\}$, *it holds*

$$I\!\!E\|D_{\boldsymbol{\epsilon}}(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\|^u\, 1\!\!1(\mathcal{C}_{\boldsymbol{\epsilon}}(\mathtt{r})) \leq I\!\!E\bigl\{\|\boldsymbol{\xi}_{\boldsymbol{\epsilon}}\| + \sqrt{2\Delta_{\boldsymbol{\epsilon}}(\mathtt{r})}\bigr\}^u. \tag{3.20}$$

Proof. The bound (3.19) follows from (3.18). Next, the expansion (3.16) yields on $\mathcal{C}_{\boldsymbol{\epsilon}}(\mathtt{r})$

$$\|D_{\boldsymbol{\epsilon}}(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\| \leq \|\boldsymbol{\xi}_{\boldsymbol{\epsilon}}\| + \sqrt{2\Delta_{\boldsymbol{\epsilon}}(\mathtt{r})}$$

and (3.20) follows. □

3.2.8 Range of applicability

The whole proposed approach relies implicitly on the two groups of assumptions: global and local. These assumptions are linked to each other by the value $\mathtt{r}$. From one side, the *global* assumptions listed in Theorem 4.1 and its corollaries should ensure a sensitive bound for the deviation probability $I\!\!P\bigl\{\widetilde{\boldsymbol{\theta}} \notin \Theta_0(\mathtt{r})\bigr\}$. This particularly requires that $\mathtt{r}$ is sufficiently large. In the contrary, the *local* conditions are based on the assumption that the local set $\Theta_0(\mathtt{r})$ is sufficiently small to guarantee that errors of approximation $\diamondsuit_{\boldsymbol{\epsilon}}(\mathtt{r})$ and $\diamondsuit_{\underline{\boldsymbol{\epsilon}}}(\mathtt{r})$ and the deficiency $\Delta_{\boldsymbol{\epsilon}}(\mathtt{r})$ from (3.9) are small as well in a probabilistic sense.

More precisely, the obtained results are sharp and meaningful if the deficiency $\Delta_{\boldsymbol{\epsilon}}(\mathtt{r})$ is smaller in order than the value of the upper problem $\|\boldsymbol{\xi}_{\boldsymbol{\epsilon}}\|^2$. The general results of Section 6 state that $\|\boldsymbol{\xi}_{\boldsymbol{\epsilon}}\|^2$ concentrates around its expected value $\mathtt{p}_{\boldsymbol{\epsilon}} \stackrel{\text{def}}{=} I\!\!E\|\boldsymbol{\xi}_{\boldsymbol{\epsilon}}\|^2$. Therefore, the latter condition about the deficiency can be decomposed into two others:

- the values $\diamondsuit_{\boldsymbol{\epsilon}}(\mathtt{r}), \diamondsuit_{\underline{\boldsymbol{\epsilon}}}(\mathtt{r})$ are smaller in order than $\mathtt{p}_{\boldsymbol{\epsilon}}$;
- the difference $\|\boldsymbol{\xi}_{\boldsymbol{\epsilon}}\|^2 - \|\boldsymbol{\xi}_{\underline{\boldsymbol{\epsilon}}}\|^2$ is smaller in order than $\mathtt{p}_{\boldsymbol{\epsilon}}$;

Due to the result of Theorem 3.1, the random quantity $\diamondsuit_{\boldsymbol{\epsilon}}(\mathtt{r})$ can be bounded with a probability larger than $1 - \mathrm{e}^{-\mathtt{x}}$ by $\varrho\, \mathfrak{z}_0(\mathtt{x}, p)$, where $\mathfrak{z}_0(\mathtt{x}, p) \approx p + \mathtt{x}$ if $\mathtt{x}$ is not too large. So, the first conditions requires that $\varrho\, p$ is smaller in order than $\mathtt{p}_{\boldsymbol{\epsilon}}$. If $\mathtt{p}_{\boldsymbol{\epsilon}}$ is of order p (see the regular case in the next section) then the condition "$\diamondsuit_{\boldsymbol{\epsilon}}(\mathtt{r})$ is small" only requires that ϱ, or equivalently, $\omega(\mathtt{r})$ is small. The same holds for $\diamondsuit_{\underline{\boldsymbol{\epsilon}}}(\mathtt{r})$.

Summarizing the above discussion yields that the local results apply if, for a fixed $\mathtt{r}$:

- $\omega(\mathtt{r})p/\mathtt{p}_{\boldsymbol{\epsilon}}$ is small;
- $\alpha_{\boldsymbol{\epsilon}}$ is small.

In the regular case studied in Section 3.2.3, these two conditions simplify to "$\omega(\mathtt{r}), \delta(\mathtt{r})$ are small".

3.2.9 Non-asymptotic efficiency

This section discusses the efficiency issues. The famous Cramér-Rao result describes the lower bound for the estimation risk of an unbiased estimate. For linear models this result implies that the true MLE is efficient while a quasi MLE for a misspecified noise covariance is not. The Le Cam LAN theory transfers this result on the general statistical model under the LAN condition; see e.g. Chapter 3 in Ibragimov and Khas'minskij (1981). The results obtained in the previous sections provide a non-asymptotic version of the LAN approach. As already mentioned in Section 2.2, if $L(\boldsymbol{\theta})$ is the true log-likelihood function of a regular parametric family and $I\!P = I\!P_{\boldsymbol{\theta}^*}$, then both matrices D_0^2 and V_0^2 are equal to the Fisher information matrix of this family. This implies the regularity condition with $\mathfrak{a} = 1$; see (3.10). Suppose in addition that the values $\omega(\mathtt{r}), \delta(\mathtt{r})$ are small, so that

$$\mathfrak{a}_{\boldsymbol{\epsilon}} = 1 - \delta - \varrho = 1 - \delta(\mathtt{r}) - 3\nu_0\omega(\mathtt{r})$$

is close to one. By (3.11) this implies that $D_{\boldsymbol{\epsilon}} \approx D_0$, the matrix $I\!B_{\boldsymbol{\epsilon}}$ is close to the identity matrix I_p and $\mathtt{p}_{\boldsymbol{\epsilon}} \approx p$, $\mathtt{v}_{\boldsymbol{\epsilon}}^2 \approx 2p$, $\lambda_{\boldsymbol{\epsilon}} \approx 1$. This in turn implies that the twice upper value $\|\boldsymbol{\xi}_{\boldsymbol{\epsilon}}\|^2$ behaves as a χ^2_p random variable and the deficiency $\Delta_{\boldsymbol{\epsilon}}(\mathtt{r})$ is small in probability. In particular, all the corollaries about confidence and coverage probability for $\widetilde{\boldsymbol{\theta}}$ reproduce the similar statements for the correct linear Gaussian model. Moreover, the decomposition (3.16) can be rewritten as

$$D_0\bigl(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\bigr) \approx D_0^{-1}\nabla L(\boldsymbol{\theta}^*)$$

and this is the famous expansion of the MLE in the LAN situation yielding the asymptotic normality and all other asymptotic properties of $\widetilde{\boldsymbol{\theta}}$ including its asymptotic efficiency. We present a precise statement in Section 5.1 when studying the i.i.d. case.

4 Deviation bounds 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 a local vicinity Θ_0 of Θ, the concentration result describes the deviation probability $I\!P\bigl(\widetilde{\boldsymbol{\theta}} \notin \Theta_0\bigr)$. The key step in this large deviation bound is made in terms of a *multiscale upper function* for the process $L(\boldsymbol{\theta}, \boldsymbol{\theta}^*) \stackrel{\text{def}}{=} L(\boldsymbol{\theta}) - L(\boldsymbol{\theta}^*)$. Namely, we build a deterministic function $\mathfrak{C}(\mu, \boldsymbol{\theta}, \boldsymbol{\theta}^*)$ such that the probability

$$I\!P\Bigl(\sup_{\boldsymbol{\theta} \notin \Theta_0} \sup_{\mu \in \mathbb{M}} \bigl[\mu L(\boldsymbol{\theta}, \boldsymbol{\theta}^*) + \mathfrak{C}(\mu, \boldsymbol{\theta}, \boldsymbol{\theta}^*)\bigr] \geq \mathtt{x}\Bigr) \tag{4.1}$$

is exponentially small in $\mathtt{x}$. Here μ is a positive *scale* parameter and $\mathbb{M}$ is the discrete set of considered scale values. Concentration sets for $\widetilde{\boldsymbol{\theta}}$ can be naturally defined via level sets of the function $\mathfrak{C}^*(\boldsymbol{\theta}, \boldsymbol{\theta}^*) = \max_{\mu \in \mathbb{M}} \mathfrak{C}(\mu, \boldsymbol{\theta}, \boldsymbol{\theta}^*)$. The bound (4.1) is established by some rather crude methods of the theory of empirical processes; see Section 7. Once established, the concentration properties of $\widetilde{\boldsymbol{\theta}}$ can be refined in the local vicinity Θ_0 using local majorization technique; see Section 3.2.

The other important result describes an *upper function* $\mathfrak{b}(\boldsymbol{\theta}, \alpha)$ for the non-scaled process $L(\boldsymbol{\theta}, \boldsymbol{\theta}^*)$ providing that the probability

$$I\!P\Bigl(\sup_{\boldsymbol{\theta} \notin \Theta_0} \bigl\{L(\boldsymbol{\theta}, \boldsymbol{\theta}^*) + \mathfrak{b}(\boldsymbol{\theta}, \mathtt{x})\bigr\} > 0\Bigr)$$

is exponentially small in $\mathtt{x}$. Such bounds are usually called for in the analysis of the posterior measure in the Bayes approach.

4.1 A multiscale upper function for the log-likelihood

This section presents a construction of the multiscale upper function $\mathfrak{C}(\mu, \boldsymbol{\theta}, \boldsymbol{\theta}^*)$. Then it will be used for controlling the deviation probability of the estimate $\widetilde{\boldsymbol{\theta}}$. Below we suppose that the scaling factor μ runs over some discrete set $\mathbb{M}$ and consider the process $\mu L(\boldsymbol{\theta}, \boldsymbol{\theta}^*)$ for $\boldsymbol{\theta} \in \Theta$ and $\mu \in \mathbb{M}$. Assume that for each $\boldsymbol{\theta}$, there is some $\mu \in \mathbb{M}$ such that the exponential moment of $\mu L(\boldsymbol{\theta}, \boldsymbol{\theta}^*)$ is finite. This enables us to define for each $\boldsymbol{\theta}$ the function $\mathfrak{M}(\mu, \boldsymbol{\theta}, \boldsymbol{\theta}^*)$ ensuring the identity

$$I\!E \exp\bigl\{\mu L(\boldsymbol{\theta}, \boldsymbol{\theta}^*) + \mathfrak{M}(\mu, \boldsymbol{\theta}, \boldsymbol{\theta}^*)\bigr\} = 1.$$

This means that the process $\mu L(\boldsymbol{\theta},\boldsymbol{\theta}^*)+\mathfrak{M}(\mu,\boldsymbol{\theta},\boldsymbol{\theta}^*)$ is pointwise stochastically bounded in a rather strict sense. For each $\mu>0$ and $\mathfrak{z}\geq 0$ by the Markov inequality

$$\begin{aligned}\mathbb{P}\bigl(L(\boldsymbol{\theta},\boldsymbol{\theta}^*)\geq \mathfrak{z}\bigr) &\leq \exp\bigl\{-\mu\mathfrak{z}\bigr\}\mathbb{E}\exp\bigl\{\mu L(\boldsymbol{\theta},\boldsymbol{\theta}^*)\bigr\}\\ &= \exp\bigl\{-\mu\mathfrak{z}-\mathfrak{M}(\mu,\boldsymbol{\theta},\boldsymbol{\theta}^*)\bigr\}. \end{aligned} \tag{4.2}$$

In particular, with $\mathfrak{z}=0$,

$$\mathbb{P}\bigl(L(\boldsymbol{\theta},\boldsymbol{\theta}^*)\geq 0\bigr)\leq \exp\bigl\{-\max_{\mu\in\mathbb{M}}\mathfrak{M}(\mu,\boldsymbol{\theta},\boldsymbol{\theta}^*)\bigr\}.$$

So, a reasonable choice of μ can be made via maximization of $\mathfrak{M}(\mu,\boldsymbol{\theta},\boldsymbol{\theta}^*)$ w.r.t. μ. The famous Chernoff result describes the asymptotic separation rate between these two measures $\mathbb{P}_{\boldsymbol{\theta}^*}$ and $\mathbb{P}_{\boldsymbol{\theta}}$ in terms of the value $\mathfrak{M}^*(\boldsymbol{\theta},\boldsymbol{\theta}^*)$ with

$$\mathfrak{M}^*(\boldsymbol{\theta},\boldsymbol{\theta}^*)=\sup_{\mu\in\mathbb{M}}\mathfrak{M}(\mu,\boldsymbol{\theta},\boldsymbol{\theta}^*). \tag{4.3}$$

The larger $\mathfrak{M}^*(\boldsymbol{\theta},\boldsymbol{\theta}^*)$ is, the stronger is the pointwise identification.

If Θ° is a subset of Θ not containing $\boldsymbol{\theta}^*$, then the event $\widetilde{\boldsymbol{\theta}}\in\Theta^\circ$ is only possible if $\sup_{\boldsymbol{\theta}\in\Theta^\circ}L(\boldsymbol{\theta},\boldsymbol{\theta}^*)\geq 0$, because $L(\boldsymbol{\theta}^*,\boldsymbol{\theta}^*)\equiv 0$. It is intuitively clear that the uniform identification over Θ° requires that the rate function $\mathfrak{M}^*(\boldsymbol{\theta},\boldsymbol{\theta}^*)$ is bounded away from zero on Θ°. The result of Theorem 4.1 below quantifies this condition.

We present a result for the smooth case which assumes the conditions (E), (ED), and (ED_0) to be fulfilled with the corresponding matrices V and V_0. The set Θ° is taken as the complement of the local set $\Theta_0(\mathtt{r})$ with a sufficiently large $\mathtt{r}$. The result also involves a value $\mathtt{r}_1$ entering into the pilling device; see Section 7.3 for details. The choice of $\mathtt{r}_1$ is done by the equality $\det(\mathtt{r}_1^{-1}V)=\det(V_0)$ relating two matrices V_0 and V to each others.

For stating the results, some further notations have to be introduced. Define $\zeta(\boldsymbol{\theta},\boldsymbol{\theta}^*)=\zeta(\boldsymbol{\theta})-\zeta(\boldsymbol{\theta}^*)$ and

$$\mathfrak{N}(\mu,\boldsymbol{\theta},\boldsymbol{\theta}^*)\stackrel{\text{def}}{=}\log\mathbb{E}\exp\bigl\{\mu\zeta(\boldsymbol{\theta},\boldsymbol{\theta}^*)\bigr\}.$$

Then it holds for the function $\mathfrak{M}(\mu,\boldsymbol{\theta},\boldsymbol{\theta}^*)$:

$$\mathfrak{M}(\mu,\boldsymbol{\theta},\boldsymbol{\theta}^*)=-\log\mathbb{E}\exp\bigl\{\mu L(\boldsymbol{\theta},\boldsymbol{\theta}^*)\bigr\}=-\mu\mathbb{E}L(\boldsymbol{\theta},\boldsymbol{\theta}^*)-\mathfrak{N}(\mu,\boldsymbol{\theta},\boldsymbol{\theta}^*).$$

Further, consider for each $\boldsymbol{\theta}^\circ\in\Theta$ a local ball $\mathcal{B}_\mu(\boldsymbol{\theta}^\circ)\stackrel{\text{def}}{=}\bigl\{\boldsymbol{\theta}\in\Theta:\|V(\boldsymbol{\theta}-\boldsymbol{\theta}^*)\|\leq\mathtt{r}_1/\mu\bigr\}$. Next, for the Lebesgue measure π on $\mathbb{R}^p$, define the smoothing operator $\mathbb{S}_\mu$ by

$$\mathbb{S}_\mu f(\boldsymbol{\theta}^\circ)\stackrel{\text{def}}{=}\frac{1}{\pi_\mu(\boldsymbol{\theta}^\circ)}\int_{\mathcal{B}_\mu(\boldsymbol{\theta}^\circ)}f(\boldsymbol{\theta})\pi(d\boldsymbol{\theta}).$$

Let ν_0 and $\mathtt{g}$ be the constants from (ED). Define $\mathfrak{c}_1$ by $\mathfrak{c}_1 = 2$ for $p \geq 2$ and $\mathfrak{c}_1 = 2.4$ for $p = 1$. Let a constant s be selected under the condition $3\nu_0 \mathtt{r}_1/s \leq \mathtt{g} \wedge \sqrt{2\mathfrak{c}_1 p}$. The value $s = 1$ is a proper candidate in typical situations.

For shortening the notation assume that the discrete set $\mathbb{M}$ is fixed under the condition $\mathrm{M} \equiv \sum_{\mu \in \mathbb{M}} \mu^{p+2} \leq \mathrm{e}/2$.

Theorem 4.1. *Suppose that (E), (ED), and (ED_0) hold with some $\mathtt{g}, \nu_0$ and with matrices V and V_0. Let $\mathtt{r}_1$ be such that $\det(\mathtt{r}_1^{-1} V) \leq \det(V_0)$, and s be such that $3\nu_0 \mathtt{r}_1/s \leq \mathtt{g} \wedge \sqrt{2\mathfrak{c}_1 p}$. For $\mathtt{r} \geq \sqrt{p/2}$ and $\mathtt{x} > 0$, it holds*

$$\mathbb{P}\Big\{ \sup_{\boldsymbol{\theta} \notin \Theta_0(\mathtt{r})} \sup_{\mu \in \mathbb{M}} \big[\mu L(\boldsymbol{\theta}, \boldsymbol{\theta}^*) + \mathcal{C}(\mu, \boldsymbol{\theta}, \boldsymbol{\theta}^*) \big] > \mathfrak{z}_1(\mathtt{x}) \Big\} \leq \mathrm{e}^{-\mathtt{x}+1}, \tag{4.4}$$

with $\mathfrak{z}_1(\mathtt{x}) \stackrel{\text{def}}{=} 2s\mathfrak{c}_1 p + (1+s)\mathtt{x}$ and

$$\begin{aligned} \mathcal{C}(\mu, \boldsymbol{\theta}, \boldsymbol{\theta}^*) &\stackrel{\text{def}}{=} -\mu \mathbb{E} L(\boldsymbol{\theta}, \boldsymbol{\theta}^*) - \mathbb{S}_\mu \mathfrak{N}(\mu, \boldsymbol{\theta}, \boldsymbol{\theta}^*) - (1+s)\mathfrak{t}(\boldsymbol{\theta}) \\ \mathfrak{t}(\boldsymbol{\theta}) &\stackrel{\text{def}}{=} (p+2) \log\big(\| V_0(\boldsymbol{\theta} - \boldsymbol{\theta}^*) \| \big). \end{aligned} \tag{4.5}$$

Proof. First fix $\mu \in \mathbb{M}$ and apply the general results of Theorem 7.10 to the process $\mathcal{U}(\boldsymbol{\theta}) = \zeta(\boldsymbol{\theta}, \boldsymbol{\theta}^*) = L(\boldsymbol{\theta}, \boldsymbol{\theta}^*) - \mathbb{E} L(\boldsymbol{\theta}, \boldsymbol{\theta}^*)$, $M(\boldsymbol{\theta}) = -\mathbb{E} L(\boldsymbol{\theta}, \boldsymbol{\theta}^*)$, and $H_0 = V_0$ on $\Upsilon^\circ = \Theta_0^c(\mathtt{r})$. The condition $(\mathcal{E})$ is fulfilled, condition (ED) implies $(\mathcal{E}D)$ with $H(\boldsymbol{\theta}) = V$, and Theorem 7.10 yields (4.4) in view of $\| \mathtt{r}_1^{-1} V(\boldsymbol{\theta} - \boldsymbol{\theta}^*) \| \leq \| V_0(\boldsymbol{\theta} - \boldsymbol{\theta}^*) \|$. □

Remark 4.1. The construction of the multiscale upper function $\mathcal{C}(\mu, \boldsymbol{\theta}, \boldsymbol{\theta}^*)$ in (4.5) deserves some discussion. More precisely, it is interesting to compare this construction with the pointwise upper function $\mathfrak{M}(\mu, \boldsymbol{\theta}, \boldsymbol{\theta}^*) = -\mu \mathbb{E} L(\boldsymbol{\theta}, \boldsymbol{\theta}^*) - \mathfrak{N}(\mu, \boldsymbol{\theta}, \boldsymbol{\theta}^*)$. The smoothing operator in the term $\mathbb{S}_\mu \mathfrak{N}(\mu, \boldsymbol{\theta}, \boldsymbol{\theta}^*)$ from (4.5) is only used for technical reasons and it can be handled by simple rescaling arguments. The other term $(1+s)\mathfrak{t}(\boldsymbol{\theta})$ in (4.5) is really important and it can be viewed as the price for uniform concentration. Simple white noise examples show that this penalty term is nearly sharp if s is taken small and p large. However, our aim is only to obtain a rough exponential bound because really sharp results are stated by mean of local quadratic approximation; see Section 3.2.5.

4.2 Concentration sets and deviation probability

This section describes so called *concentration sets* for the estimate $\widetilde{\boldsymbol{\theta}}$. Any such set is deterministic ensuring that $\widetilde{\boldsymbol{\theta}}$ deviates from this set with a small probability. Such concentration sets are usually used in theoretical studies and their construction typically depends on the unknown quantities the moment generation function $\mathfrak{M}(\mu, \boldsymbol{\theta}, \boldsymbol{\theta}^*)$. Let

$\mathcal{C}(\mu, \boldsymbol{\theta}, \boldsymbol{\theta}^*)$ be the upper multiscale function from (4.5). Define

$$\mu(\boldsymbol{\theta}, \boldsymbol{\theta}^*) \stackrel{\text{def}}{=} \operatorname*{argmax}_{\mu \in \mathbb{M}} \mathcal{C}(\mu, \boldsymbol{\theta}, \boldsymbol{\theta}^*), \tag{4.6}$$

$$\mathcal{C}^*(\boldsymbol{\theta}, \boldsymbol{\theta}^*) \stackrel{\text{def}}{=} \max_{\mu \in \mathbb{M}} \mathcal{C}(\mu, \boldsymbol{\theta}, \boldsymbol{\theta}^*) = \mathcal{C}(\mu(\boldsymbol{\theta}, \boldsymbol{\theta}^*), \boldsymbol{\theta}, \boldsymbol{\theta}^*). \tag{4.7}$$

For simplicity we assume that the maximum in (4.7) is attained and $\mu(\boldsymbol{\theta}, \boldsymbol{\theta}^*)$ is well defined. In the considered case of a discrete set $\mathbb{M}$ this is always fulfilled. The value $\mathcal{C}^*(\boldsymbol{\theta}, \boldsymbol{\theta}^*)$ shown in (4.7) replaces $\mathfrak{M}^*(\boldsymbol{\theta}, \boldsymbol{\theta}^*)$ from (4.3) if one is concerned with uniform non-asymptotic bounds. For any fixed subset $\mathcal{A} \subset \Theta$ define the value $\mathfrak{g}(\mathcal{A})$ by

$$\mathfrak{g}(\mathcal{A}) \stackrel{\text{def}}{=} \inf_{\boldsymbol{\theta} \notin \mathcal{A}} \mathcal{C}^*(\boldsymbol{\theta}, \boldsymbol{\theta}^*). \tag{4.8}$$

A particular choice of the set $\mathcal{A}$ is given by the level set of $\mathcal{C}^*(\boldsymbol{\theta}, \boldsymbol{\theta}^*)$: for every $\mathtt{x} > 0$, define $\mathcal{A}(\mathtt{x}, \boldsymbol{\theta}^*)$ with

$$\mathcal{A}(\mathtt{x}, \boldsymbol{\theta}^*) = \{\boldsymbol{\theta} : \mathcal{C}^*(\boldsymbol{\theta}, \boldsymbol{\theta}^*) \le \mathfrak{z}_1(\mathtt{x})\} = \Big\{\boldsymbol{\theta} : \sup_{\mu \in \mathbb{M}} \mathcal{C}(\mu, \boldsymbol{\theta}, \boldsymbol{\theta}^*) \le \mathfrak{z}_1(\mathtt{x})\Big\},$$

where $\mathfrak{z}_1(\mathtt{x}) = 2s\mathfrak{c}_1 p + (1+s)\mathtt{x}$. Obviously $\mathfrak{g}\big(\mathcal{A}(\mathtt{x}, \boldsymbol{\theta}^*)\big) \ge \mathfrak{z}_1(\mathtt{x})$.

Below we show that the bound (4.4) yields some concentration properties of the estimate $\widetilde{\boldsymbol{\theta}}$ in terms of the function $\mathfrak{g}(\cdot)$ from (4.8).

Corollary 4.2. *Under* (4.4)*, it holds for any set* $\mathcal{A}$ *with* $\mathfrak{g}(\mathcal{A}) \ge \mathfrak{z}_1(\mathtt{x})$

$$I\!P\big(\widetilde{\boldsymbol{\theta}} \notin \mathcal{A}\big) \le \mathrm{e}^{-\mathtt{x}+1}.$$

In particular,

$$\log I\!P\big(\widetilde{\boldsymbol{\theta}} \notin \mathcal{A}(\mathtt{x}, \boldsymbol{\theta}^*)\big) \le -\mathtt{x} + 1.$$

Proof. Denote $\widetilde{\mu} = \mu(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}^*)$ with $\mu(\boldsymbol{\theta}, \boldsymbol{\theta}^*)$ from (4.6). Then in view of $L(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}^*) \ge 0$ and $\mathcal{C}\big(\widetilde{\mu}, \boldsymbol{\theta}, \boldsymbol{\theta}^*\big) \ge \mathfrak{g}(\mathcal{A})$ for $\widetilde{\boldsymbol{\theta}} \notin \mathcal{A}$

$$\begin{aligned}\big\{\widetilde{\boldsymbol{\theta}} \notin \mathcal{A}\big\} &\subseteq \big\{\mathcal{C}\big(\widetilde{\mu}, \widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}^*\big) \ge \mathfrak{g}(\mathcal{A})\big\} \\ &\subseteq \big\{\widetilde{\mu} L(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}^*) + \mathcal{C}\big(\widetilde{\mu}, \widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}^*\big) > \mathfrak{g}(A)\big\},\end{aligned}$$

and the result follows from (4.4). □

Corollary 4.2 presents an upper bound for the probability that $\widetilde{\boldsymbol{\theta}}$ deviates from a set $\mathcal{A}$ defined via the function $\mathcal{C}(\mu, \boldsymbol{\theta}, \boldsymbol{\theta}^*)$. The local approach of Section 3.1 requires also to bound the deviation probability for a local set $\Theta_0(\mathtt{r}) \stackrel{\text{def}}{=} \{\boldsymbol{\theta} : \|V_0(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\| > \mathtt{r}\}$ for

$\mathtt{r}$ sufficiently large. It suffices to evaluate the quantity $\mathfrak{g}(\mathcal{A})$ for $\mathcal{A} = \Theta_0(\mathtt{r})$. This in turns requires to bound from below the value $\mathfrak{C}^*(\boldsymbol{\theta}, \boldsymbol{\theta}^*)$ for $\boldsymbol{\theta} \notin \mathcal{A}$.

The next result presents sufficient conditions ensuring a sensible large deviation probability bound in terms of the rate function $\mathfrak{M}^*(\boldsymbol{\theta}, \boldsymbol{\theta}^*) = \max_\mu \bigl\{ -\log I\!\!E \exp \mu L(\boldsymbol{\theta}, \boldsymbol{\theta}^*) \bigr\}$. Define

$$\delta_\mu(\boldsymbol{\theta}) \stackrel{\text{def}}{=} \bigl| \mathfrak{N}(\mu, \boldsymbol{\theta}, \boldsymbol{\theta}^*) - \mathbb{S}_\mu \mathfrak{N}(\mu, \boldsymbol{\theta}, \boldsymbol{\theta}^*) \bigr|.$$

Corollary 4.3. *Suppose the conditions of Theorem 4.1. Let, given $\mathtt{x} > 0$, there be a value $\mathtt{r} = \mathtt{r}(\mathtt{x})$, such that it holds*

$$\mathfrak{M}^*(\boldsymbol{\theta}, \boldsymbol{\theta}^*) \geq (1+s)\mathfrak{t}(\boldsymbol{\theta}) + \delta_\mu(\boldsymbol{\theta}) + \mathfrak{z}_1(\mathtt{x}), \qquad \boldsymbol{\theta} \notin \Theta_0(\mathtt{r}), \tag{4.9}$$

with $\mathfrak{t}(\boldsymbol{\theta}) = (p+2) \log\bigl(\| V_0(\boldsymbol{\theta} - \boldsymbol{\theta}^*) \| \bigr)$ and $\mathfrak{z}_1(\mathtt{x}) \stackrel{\text{def}}{=} 2s\mathfrak{c}_1 p + (1+s)\mathtt{x}$. Then

$$I\!\!P\bigl(\| V_0(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) \| > \mathtt{r}(\mathtt{x}) \bigr) \leq \mathrm{e}^{-\mathtt{x}+1}.$$

Proof. The result follows from Corollary 4.2 by making use of the inequality

$$-\mu I\!\!E L(\boldsymbol{\theta}, \boldsymbol{\theta}^*) - \mathbb{S}_\mu \mathfrak{N}(\mu, \boldsymbol{\theta}, \boldsymbol{\theta}^*) \geq \mathfrak{M}(\mu, \boldsymbol{\theta}, \boldsymbol{\theta}^*) - \delta_\mu(\boldsymbol{\theta})$$

for any $\mu \in \mathbb{M}$. □

Remark 4.2. Due to this result, a logarithmic growth of $\mathfrak{M}^*(\boldsymbol{\theta}, \boldsymbol{\theta}^*)$ as function of the distance $\| \boldsymbol{\theta} - \boldsymbol{\theta}^* \|$ ensures a sensitive large deviation bound for the process $L(\boldsymbol{\theta})$.

4.3 Probability bounds for the quasi log-likelihood

This section presents a uniform upper bound for process $L(\boldsymbol{\theta}, \boldsymbol{\theta}^*)$ without scaling. Our starting point is again a pointwise bound on $L(\boldsymbol{\theta}, \boldsymbol{\theta}^*)$ for a fixed $\boldsymbol{\theta} \neq \boldsymbol{\theta}^*$. Namely, given $\mathtt{x}$, we first try to find $b(\boldsymbol{\theta}, \mathtt{x})$ providing

$$I\!\!P\bigl(L(\boldsymbol{\theta}, \boldsymbol{\theta}^*) + b(\boldsymbol{\theta}, \mathtt{x}) \geq 0 \bigr) \leq \mathrm{e}^{-\mathtt{x}}. \tag{4.10}$$

Define

$$b(\boldsymbol{\theta}, \mathtt{x}) \stackrel{\text{def}}{=} \max_{\mu \in \mathbb{M}} \mu^{-1} \bigl\{ -\mathtt{x} + \mathfrak{M}(\mu, \boldsymbol{\theta}, \boldsymbol{\theta}^*) \bigr\}.$$

Then applying (4.2) with the corresponding value μ yields (4.10). Now we aim to establish an extension of this pointwise bound to a uniform bound on a subset Θ° of Θ, that is, to build a function $\mathfrak{b}(\boldsymbol{\theta}, \mathtt{x})$ such that

$$I\!\!P\Bigl(\sup_{\boldsymbol{\theta} \in \Theta^\circ} \bigl\{ L(\boldsymbol{\theta}, \boldsymbol{\theta}^*) + \mathfrak{b}(\boldsymbol{\theta}, \mathtt{x}) \bigr\} > 0 \Bigr) \leq \mathrm{e}^{-\mathtt{x}+1}. \tag{4.11}$$

Such bounds are naturally called for in the analysis of the posterior measure in the Bayes approach. The function $\mathfrak{b}(\boldsymbol{\theta},\mathtt{x})$ can be described via the multiscale upper function $\mathcal{C}(\mu,\boldsymbol{\theta},\boldsymbol{\theta}^*)$. Namely, for each $\boldsymbol{\theta}$ and $\mathtt{x}>0$ define

$$\mathfrak{b}(\boldsymbol{\theta},\mathtt{x}) \stackrel{\text{def}}{=} \max_{\mu\in\mathbb{M}}\bigl\{-\mu^{-1}\mathfrak{z}_1(\mathtt{x})+\mu^{-1}\mathcal{C}(\mu,\boldsymbol{\theta},\boldsymbol{\theta}^*)\bigr\}. \tag{4.12}$$

Corollary 4.4. *Assume the bound* (4.4) *for a function* $\mathcal{C}(\mu,\boldsymbol{\theta},\boldsymbol{\theta}^*)$. *Then* (4.11) *holds with* $\mathfrak{b}(\boldsymbol{\theta},\mathtt{x})$ *from* (4.12).

Proof. The bound (4.4) implies

$$I\!\!P\Bigl(\sup_{\boldsymbol{\theta}\in\Theta^\circ}\min_{\mu\in\mathbb{M}}\bigl\{\mu L(\boldsymbol{\theta},\boldsymbol{\theta}^*)+\mathcal{C}(\mu,\boldsymbol{\theta},\boldsymbol{\theta}^*)\bigr\}>\mathfrak{z}_1(\mathtt{x})\Bigr)\le \mathrm{e}^{-\mathtt{x}+1}.$$

This yields that there exists a random set $\Omega(\mathtt{x})$ with $I\!\!P\bigl(\Omega(\mathtt{x})\bigr)\ge 1-\mathrm{e}^{-\mathtt{x}+1}$ such that for any $\mu\in\mathbb{M}$ and any $\boldsymbol{\theta}\in\Theta^\circ$, it holds $\mu L(\boldsymbol{\theta},\boldsymbol{\theta}^*)+\mathcal{C}(\mu,\boldsymbol{\theta},\boldsymbol{\theta}^*)\le\mathfrak{z}_1(\mathtt{x})$ on $\Omega(\mathtt{x})$. Let $\mu(\boldsymbol{\theta},\mathtt{x})$ fulfill

$$\mu(\boldsymbol{\theta},\mathtt{x}) \stackrel{\text{def}}{=} \operatorname*{argmax}_{\mu\in\mathbb{M}}\bigl\{-\mu^{-1}\mathfrak{z}_1(\mathtt{x})+\mu^{-1}\mathcal{C}(\mu,\boldsymbol{\theta},\boldsymbol{\theta}^*)\bigr\}.$$

Then particular choice $\mu=\mu(\boldsymbol{\theta},\mathtt{x})$ yields on $\Omega(\mathtt{x})$

$$L(\boldsymbol{\theta},\boldsymbol{\theta}^*)+\frac{1}{\mu(\boldsymbol{\theta},\mathtt{x})}\bigl\{\mathcal{C}(\mu(\boldsymbol{\theta},\mathtt{x}),\boldsymbol{\theta},\boldsymbol{\theta}^*)-\mathfrak{z}_1(\mathtt{x})\bigr\}\le 0$$

or equivalently $L(\boldsymbol{\theta},\boldsymbol{\theta}^*)+\mathfrak{b}(\boldsymbol{\theta},\mathtt{x})\le 0$. □

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. With some surprise, 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 from 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 $L(\boldsymbol{\theta})$:

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

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 $I\!\!E\nabla L(\boldsymbol{\theta}^*) = 0$ 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 for rewriting the conditions (E), (ED), (ED_0), (ED_1), and $(\mathcal{L}_0)$ in terms of the marginal distribution. In the following conditions the index i runs from 1 to n.

(e) For each $\boldsymbol{\theta} \in \Theta$, there exists a positive value $\mu \in \mathbb{M}$ such that

$$\mathfrak{m}(\mu,\boldsymbol{\theta},\boldsymbol{\theta}^*) \stackrel{\text{def}}{=} -\log I\!\!E \exp\bigl\{\mu\bigl[\ell(Y_i,\boldsymbol{\theta}) - \ell(Y_i,\boldsymbol{\theta}^*)\bigr]\bigr\}$$

is finite.

(ed) *There exist some constants ν_0, and $\mathtt{g}_1 > 0$, and a positive symmetric $p \times p$ matrix $\mathbf{v}$, such that for all $|\lambda| \le \mathtt{g}_1$*

$$\sup_{\boldsymbol{\gamma}\in\mathcal{S}^p} \sup_{\boldsymbol{\theta}\in\Theta} \log I\!\!E \exp\left\{\lambda \frac{\boldsymbol{\gamma}^\top \nabla\zeta_i(\boldsymbol{\theta})}{\|\mathbf{v}\boldsymbol{\gamma}\|}\right\} \le \nu_0^2\lambda^2/2.$$

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

$$\sup_{\boldsymbol{\gamma}\in\mathcal{S}^p} \log I\!\!E \exp\left\{\lambda \frac{\boldsymbol{\gamma}^\top \nabla\zeta_i(\boldsymbol{\theta}^*)}{\|\mathbf{v}_0\boldsymbol{\gamma}\|}\right\} \le \nu_0^2\lambda^2/2.$$

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

Next consider the local sets

$$\Theta_0(\mathtt{r}) = \{\boldsymbol{\theta} : \|\mathbf{v}_0(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\| \le \mathtt{r}/n^{1/2}\}.$$

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

(ed_1) *For each $\mathtt{r} \le R$, there exists a constant ω^* such that for all $i = 1,\ldots,n$ and $|\lambda| \le \mathtt{g}_1$*

$$\sup_{\boldsymbol{\gamma}\in\mathcal{S}^p} \sup_{\boldsymbol{\theta}\in\Theta_0(\mathtt{r})} \log I\!\!E \exp\left\{\lambda \frac{\boldsymbol{\gamma}^\top\bigl[\nabla\zeta_i(\boldsymbol{\theta}) - \nabla\zeta_i(\boldsymbol{\theta}^*)\bigr]}{\omega^*\,\mathtt{r}\,\|\mathbf{v}_0\boldsymbol{\gamma}\|}\right\} \le \nu_0^2\lambda^2/2.$$

Further we restate the local identifiability condition $(\mathcal{L}_0)$ in terms of the expected value $\overline{\ell}(\boldsymbol{\theta}) \stackrel{\text{def}}{=} I\!\!E\ell(Y_i,\boldsymbol{\theta})$ of each $\ell(Y_i,\boldsymbol{\theta})$. We suppose that $\overline{\ell}(\boldsymbol{\theta})$ is two times differentiable w.r.t. $\boldsymbol{\theta}$ and define the matrix $I\!\!F_0 = -\nabla^2\overline{\ell}(\boldsymbol{\theta}^*)$.

(ℓ_0) *For each $\mathtt{r} \le R$, there is a constant δ^*, such that it holds on $\Theta_0(\mathtt{r})$*

$$\left| \frac{\overline{\ell}(\boldsymbol{\theta}) - \overline{\ell}(\boldsymbol{\theta}^*) - (\boldsymbol{\theta} - \boldsymbol{\theta}^*)^\top \nabla \overline{\ell}(\boldsymbol{\theta}^*)}{(\boldsymbol{\theta} - \boldsymbol{\theta}^*)^\top I\!\!F_0 \, (\boldsymbol{\theta} - \boldsymbol{\theta}^*)/2} - 1 \right| \le \delta^* \mathtt{r}.$$

In the regular parametric case with $I\!\!P \in (P_{\boldsymbol{\theta}})$, the matrices $\mathbf{v}_0^2$ and $I\!\!F_0$ coincide with the Fisher information matrix $I\!\!F(\boldsymbol{\theta}^*)$ of the family $(P_{\boldsymbol{\theta}})$ at the point $\boldsymbol{\theta}^*$.

Lemma 5.1. *Let $Y_1, \ldots, Y_n$ be i.i.d. Then (e), (ed), (ed_0), $(ed)_1$, and (ℓ_0) imply (E), (ED), (ED_0), $(ED)_1$, and $(\mathcal{L}_0)$ with $\mathfrak{M}(\mu, \boldsymbol{\theta}, \boldsymbol{\theta}^*) = n\mathfrak{m}(\mu, \boldsymbol{\theta}, \boldsymbol{\theta}^*)$, $V^2 = n\mathbf{v}^2$, $V_0^2 = n\mathbf{v}_0^2$, $D_0^2 = nI\!\!F_0$, $\omega(\mathtt{r}) = \omega^* \mathtt{r}$, $\delta(\mathtt{r}) = \delta^* \mathtt{r}$, the same constant ν_0, and*

$$\mathtt{g} \stackrel{\text{def}}{=} \mathtt{g}_1 \sqrt{n}.$$

Proof. The identities $\mathfrak{M}(\mu, \boldsymbol{\theta}, \boldsymbol{\theta}^*) = n\,\mathfrak{m}(\mu, \boldsymbol{\theta}, \boldsymbol{\theta}^*)$, $V^2 = n\mathbf{v}^2$, $V_0^2 = n\mathbf{v}_0^2$, $D_0^2 = nI\!\!F_0$ follow from the i.i.d. structure of the observations Y_i. We briefly comment on condition (ED). The use once again the i.i.d. structure yields by (5.1) in view of $V^2 = n\mathbf{v}^2$

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

as long as $\lambda \le n^{1/2} \mathtt{g}_1 \le \mathtt{g}$. Similarly one can check (ED_0) and (ED_1). □

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

5.1.1 A large deviation bound

First we describe the large deviation probability for the event $\{\widetilde{\boldsymbol{\theta}} \notin \Theta_0(\mathtt{r})\}$ for a fixed $\mathtt{r}$. Corollary 4.3 provides a sufficient condition for such a bound: the rate function $\mathfrak{M}^*(\boldsymbol{\theta}, \boldsymbol{\theta}^*)$ should grow at least logarithmic with the distance $\|V(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\|$. Define

$$\mathfrak{m}^*(\boldsymbol{\theta}, \boldsymbol{\theta}^*) = \max_{\mu} \mathfrak{m}(\mu, \boldsymbol{\theta}, \boldsymbol{\theta}^*).$$

(ld_1) *There exist constants $b > 0$, such that it holds on the set $\Theta_0^c(\mathtt{r})$*

$$\mathfrak{m}^*(\boldsymbol{\theta}, \boldsymbol{\theta}^*) \ge b \log\bigl(1 + \|\mathbf{v}_0(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\|^2\bigr).$$

This condition implies by $\mathfrak{M}^*(\boldsymbol{\theta}, \boldsymbol{\theta}^*) = n\mathfrak{m}^*(\boldsymbol{\theta}, \boldsymbol{\theta}^*)$

$$\mathfrak{M}^*(\boldsymbol{\theta}, \boldsymbol{\theta}^*) \ge nb \log\bigl(1 + \|\mathbf{v}_0(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\|^2\bigr), \qquad \boldsymbol{\theta} \notin \Theta_0(\mathtt{r}).$$

For the next result we assume that the value $\mathtt{r}_1 \ge 1$ is fixed by $\det(\mathtt{r}_1^{-1}\mathbf{v}) \le \det(\mathbf{v}_0)$. Further, the constant s has to be fixed such that $3\nu_0 \mathtt{r}_1 / s \le \mathtt{g} \wedge \sqrt{2\mathfrak{c}_1 p}$.

Theorem 5.2. *Suppose* (e), (ed), (ld_1). *For each* $\mathtt{x}$ *and* $\mathtt{r} = \mathtt{r}(\mathtt{x})$ *such that*

$$(2/3)b\mathtt{r}^2 \geq (1+s)(p/2+1)\log(1+\mathtt{r}^2) + \mathfrak{z}_1(\mathtt{x}), \tag{5.2}$$

it holds for $n \geq \mathtt{r}^2$

$$I\!\!P\bigl(\sqrt{n}\|\mathbf{v}_0(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\| \geq \mathtt{r}\bigr) \leq \mathrm{e}^{-\mathtt{x}+1}.$$

Proof. Given $\mathtt{r}$ and $\mathtt{x}$, (5.2) implies for all $u \geq \mathtt{r}$

$$nb\log(1+u^2/n) \geq (1+s)(p/2+1)\log(1+u^2) + \mathfrak{z}_1(\mathtt{x}).$$

because $n\log(1+r^2/n) \geq 2\mathtt{r}^2/3$ for $\mathtt{r}^2 \leq n$. Now it follows from (ld_1) for any $\boldsymbol{\theta} \notin \Theta_0(\mathtt{r})$ that

$$\begin{aligned}\mathfrak{M}^*(\boldsymbol{\theta},\boldsymbol{\theta}^*) &\geq nb\log(1+\|\mathbf{v}_0(\boldsymbol{\theta}-\boldsymbol{\theta}^*)\|^2) \\ &\geq (1+s)(p/2+1)\log\bigl(1+\|V_0(\boldsymbol{\theta}-\boldsymbol{\theta}^*)\|^2\bigr) + \mathfrak{z}_1(\mathtt{x}).\end{aligned}$$

Now the result follows from Corollary 4.3.

(to be done) treatment of δ_μ. □

Remark 5.1. The presented result helps to quantify two important values $\mathtt{r}$ and n providing a sensitive deviation probability bound: the radius $\mathtt{r}$ of the local neighborhood should be large enough to ensure (5.2), while the sample size n should be larger than $\mathtt{r}^2$. It is straightforward to see that (5.2) starts to hold for $\mathtt{r}^2 \geq \text{Const.}\, p\log p$ for some fixed constant Const. Therefore, for any $\mathtt{r}$ exceeding this value, a deviation probability $I\!\!P\bigl(\widetilde{\boldsymbol{\theta}} \notin \Theta_0(\mathtt{r})\bigr)$ is negligible when $n \geq \mathtt{r}^2 \geq \text{Const.}\, p\log p$.

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 $I\!\!F_0$ from conditions (ed_0), (ed_1), and (ℓ_0) determine their drift and variance components. Define

$$I\!\!F_{\boldsymbol{\epsilon}} \stackrel{\text{def}}{=} I\!\!F_0(1-\delta) - \varrho\mathbf{v}_0^2.$$

Then $D_{\boldsymbol{\epsilon}}^2 = nI\!\!F_{\boldsymbol{\epsilon}}$ and

$$\boldsymbol{\xi}_{\boldsymbol{\epsilon}} \stackrel{\text{def}}{=} D_{\boldsymbol{\epsilon}}^{-1}\nabla\zeta(\boldsymbol{\theta}^*) = \bigl(nI\!\!F_{\boldsymbol{\epsilon}}\bigr)^{-1/2}\sum\nabla\ell(Y_i,\boldsymbol{\theta}^*). \tag{5.3}$$

The upper approximating 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 appears 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* $\mathtt{r}$, *assume* (ed_1), *and* (ℓ_0) *on* $\Theta_0(\mathtt{r})$, *and let* $\varrho = 3\nu_0\,\omega^*\mathtt{r}/n^{1/2}$, $\delta = \delta(\mathtt{r}) = \delta^*\mathtt{r}/n^{1/2}$, *and* $I\!\!F_{\boldsymbol{\epsilon}} \stackrel{\text{def}}{=} I\!\!F_0(1-\delta) - \varrho\mathbf{v}_0^2 \geq 0$. *Then the results of Theorem 3.1 through 3.8 apply to the case of i.i.d. modeling. In particular, for any* $z > 0$, *it holds*

$$\begin{aligned} &I\!\!P\bigl\{\bigl\|\sqrt{nI\!\!F_{\boldsymbol{\epsilon}}}(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\bigr\| > z, \widetilde{\boldsymbol{\theta}} \in \Theta_0(\mathtt{r})\bigr\} \\ &\quad \leq I\!\!P\Bigl\{\bigl(1 - \sqrt{\alpha_{\boldsymbol{\epsilon}}}\bigr)\|\boldsymbol{\xi}_{\boldsymbol{\epsilon}}\| > z - \sqrt{2\diamondsuit_{\boldsymbol{\epsilon}}(\mathtt{r}) + 2\diamondsuit_{\underline{\boldsymbol{\epsilon}}}(\mathtt{r})}\Bigr\}, \end{aligned}$$

where $\diamondsuit_{\boldsymbol{\epsilon}}(\mathtt{r}), \diamondsuit_{\underline{\boldsymbol{\epsilon}}}(\mathtt{r})$ *follow the bound* (3.4) *and* $\alpha_{\boldsymbol{\epsilon}}$ *is defined by*

$$\alpha_{\boldsymbol{\epsilon}} \stackrel{\text{def}}{=} \|I_p - I\!\!F_{\boldsymbol{\epsilon}}^{1/2} I\!\!F_{\underline{\boldsymbol{\epsilon}}}^{-1} I\!\!F_{\boldsymbol{\epsilon}}^{1/2}\|_\infty = \lambda_{\max}\bigl(I_p - I\!\!F_{\boldsymbol{\epsilon}}^{1/2} I\!\!F_{\underline{\boldsymbol{\epsilon}}}^{-1} I\!\!F_{\boldsymbol{\epsilon}}^{1/2}\bigr).$$

Moreover, on the random set $\mathcal{C}_{\boldsymbol{\epsilon}}(\mathtt{r}) = \bigl\{\widetilde{\boldsymbol{\theta}} \in \Theta_0(\mathtt{r}), \|\boldsymbol{\xi}_{\underline{\boldsymbol{\epsilon}}}\| \leq \mathtt{r}\bigr\}$, *it holds*

$$\bigl\|\sqrt{nI\!\!F_{\boldsymbol{\epsilon}}}\bigl(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\bigr) - \boldsymbol{\xi}_{\boldsymbol{\epsilon}}\bigr\|^2 \leq 2\Delta_{\boldsymbol{\epsilon}}(\mathtt{r}).$$

The presented results are stated via the probability bound for the squared norm of the vector $\boldsymbol{\xi}_{\boldsymbol{\epsilon}}$ from (5.3). One can apply the general results of Section 6.4. For ease of notation we consider the vector $\boldsymbol{\xi}$ instead of $\boldsymbol{\xi}_{\boldsymbol{\epsilon}}$:

$$\boldsymbol{\xi} \stackrel{\text{def}}{=} (nI\!\!F_0)^{-1/2}\nabla\zeta(\boldsymbol{\theta}^*) = \bigl(nI\!\!F_0\bigr)^{-1/2}\sum \nabla\ell(Y_i, \boldsymbol{\theta}^*). \tag{5.4}$$

The necessary condition (6.18) coincides with (ED_0). Now Corollary 6.13 implies that the probability $I\!\!P\bigl(\|\boldsymbol{\xi}\|^2 > \mathfrak{z}(\mathtt{x}, I\!\!B)\bigr)$ is of order $2\mathrm{e}^{-\mathtt{x}}$ for all moderate $\mathtt{x}$ and $I\!\!B = I\!\!F_0^{-1/2}\mathbf{v}_0^2 I\!\!F_0^{-1/2}$. In particular, this probability starts to degenerate when $\mathfrak{z}$ significantly exceeds the value $\mathtt{p} = \mathrm{tr}(I\!\!B)$.

5.1.3 The regular parametric case and asymptotic efficiency

The conditions and the results become even more transparent in the regular situation when $I\!\!F_0 \geq \mathfrak{a}^2\mathbf{v}_0^2$. An important special case corresponds to the correct parametric specification with $\mathfrak{a} = 1$ and $I\!\!F_0 = \mathbf{v}_0^2$. Under regularity one can use $I\!\!F_{\boldsymbol{\epsilon}} = (1 - \delta - \varrho\mathfrak{a}^2)I\!\!F_0$. Here we briefly discuss the corollaries of Theorem 5.3 for the classical asymptotic setup when n tends to infinity.

Under the imposed conditions, the quantities δ and ϱ can be taken of order $\mathtt{r}/n^{1/2}$, where $\mathtt{r}^2 \gg p\log(p)$. If n grows then δ and ϱ decreases to zero, and the matrix $I\!\!F_{\boldsymbol{\epsilon}}$ is close to $I\!\!F_0$. The value $\Delta_{\boldsymbol{\epsilon}}(\mathtt{r})$ is close to zero in probability. Moreover, the random vector $\boldsymbol{\xi}$ from (5.4) fulfills $\mathrm{Var}(\boldsymbol{\xi}) \leq I\!\!F_0^{-1/2}\mathbf{v}_0^2 I\!\!F_0^{-1/2} \stackrel{\text{def}}{=} I\!\!B^2$ and by the central limit theorem $\boldsymbol{\xi}$

is asymptotically normal $\mathcal{N}(0, I\!B^2)$. This yields by Theorem 5.3 that $\sqrt{n I\!F_0}\big(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\big)$ is asymptotically normal $\mathcal{N}(0, I\!B^2)$ as well. The correct model specification implies $I\!B \equiv \boldsymbol{I}_p$ and hence $\widetilde{\boldsymbol{\theta}}$ is asymptotically efficient; see Ibragimov and Khas'minskij (1981). Also $2L(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}^*) \approx \|\boldsymbol{\xi}\|^2$ which is nearly χ^2 r.v. with p degrees of freedom. This result is known as asymptotic Wilks theorem.

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 $d(w) = \log\big(e^w + 1\big)$, 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 $\Psi_i \in I\!R^p$:

$$Y_i \sim P_{\Psi_i^\top \boldsymbol{\theta}^*}. \tag{5.5}$$

The corresponding log-density of a GLM reads as

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

First we specify the data distribution allowing that the parametric model (5.5) is misspecified. Misspecification of the first kind means that the vector $\boldsymbol{f} \stackrel{\text{def}}{=} I\!E\boldsymbol{Y}$ cannot be represented in the form $\Psi^\top \boldsymbol{\theta}$ whatever $\boldsymbol{\theta}$ is. In this situation, the target of estimation $\boldsymbol{\theta}^*$ is defined by

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

The other sort of misspecification concerns the data distribution. The model (5.5) 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 quasi MLE $\widetilde{\boldsymbol{\theta}}$ is defined by maximization of $L(\boldsymbol{\theta})$:

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

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 $\mathfrak{n}_i$ such that $I\!\!E\bigl(\varepsilon_i/\mathfrak{n}_i\bigr)^2 \le 1$ and for all $|\lambda| \le \mathtt{g}_1$*

$$\log I\!\!E \exp\bigl(\lambda \varepsilon_i / \mathfrak{n}_i\bigr) \le \nu_0^2 \lambda^2 / 2, \qquad |\lambda| \le \mathtt{g}_1. \tag{5.6}$$

A natural candidate for $\mathfrak{n}_i$ is σ_i where $\sigma_i^2 = I\!\!E \varepsilon_i^2$ is the variance of ε_i; see Lemma 7.12. Under (5.6), introduce a $p \times p$ matrix V_0 defined by

$$V_0^2 \stackrel{\text{def}}{=} \sum \mathfrak{n}_i^2 \Psi_i \Psi_i^\top. \tag{5.7}$$

Condition (e_1) effectively means that each error term $\varepsilon_i = Y_i - I\!\!E Y_i$ has some bounded exponential moments: for $\lambda = \mathtt{g}_1$, it holds $f(\lambda) \stackrel{\text{def}}{=} \log I\!\!E \exp\bigl(\lambda \varepsilon_i / \mathfrak{n}_i\bigr) < \infty$. This implies the quadratic upper bound for the function $f(\lambda)$ for $|\lambda| \le \mathtt{g}_1$; see Lemma 7.12. In words, condition (e_1) requires light (exponentially decreasing) tail for the marginal distribution of each ε_i.

Define also

$$N^{-1/2} \stackrel{\text{def}}{=} \max_i \sup_{\boldsymbol{\gamma} \in I\!R^p} \frac{\mathfrak{n}_i |\Psi_i^\top \boldsymbol{\gamma}|}{\|V_0 \boldsymbol{\gamma}\|}. \tag{5.8}$$

Lemma 5.4. *Assume (e_1) and let V_0 be defined by (5.7) and N by (5.8). Then conditions (ED_0) and (ED) follow with $V = V_0$, $\mathtt{g} = \mathtt{g}_1 N^{1/2}$, and with the constant ν_0 from (e_1). 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 $\boldsymbol{\theta}$:

$$\nabla \zeta(\boldsymbol{\theta}) = \sum \Psi_i \varepsilon_i$$

with $\varepsilon_i = Y_i - I\!\!E Y_i$. Now, for any unit vector $\boldsymbol{\gamma} \in I\!R^p$ and $\lambda \le \mathtt{g}$, independence of the ε_i's implies that

$$\log I\!\!E \exp\Bigl\{ \frac{\lambda}{\|V_0 \boldsymbol{\gamma}\|} \boldsymbol{\gamma}^\top \sum \Psi_i \varepsilon_i \Bigr\} = \sum \log I\!\!E \exp\Bigl\{ \frac{\lambda \mathfrak{n}_i \Psi_i^\top \boldsymbol{\gamma}}{\|V_0 \boldsymbol{\gamma}\|} \varepsilon_i / \mathfrak{n}_i \Bigr\}. \tag{5.9}$$

By definition $\mathfrak{n}_i|\Psi_i^\top\gamma|/\|V_0\gamma\| \le N^{-1/2}$ and therefore, $\lambda\mathfrak{n}_i|\Psi_i^\top\gamma|/\|V_0\gamma\| \le \mathtt{g}_1$. Hence, (5.6) implies

$$\log I\!\!E \exp\Big\{ \frac{\lambda}{\|V_0\gamma\|}\gamma^\top \sum \Psi_i\varepsilon_i \Big\} \le \frac{\nu_0^2\lambda^2}{2\|V_0\gamma\|^2} \sum \mathfrak{n}_i^2|\Psi_i^\top\gamma|^2 = \frac{\nu_0^2\lambda^2}{2}\,, \tag{5.10}$$

and (ED_0) follows. □

It remains only 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 5.5. *It holds*

$$\begin{aligned} -I\!\!EL(\boldsymbol{\theta},\boldsymbol{\theta}^*) &= \sum \big\{ d(\Psi_i^\top\boldsymbol{\theta}) - d(\Psi_i^\top\boldsymbol{\theta}^*) - d'(\Psi_i^\top\boldsymbol{\theta}^*)\Psi_i^\top(\boldsymbol{\theta}-\boldsymbol{\theta}^*) \big\} \\ &= \mathcal{K}\big(I\!\!P_{\boldsymbol{\theta}^*}, I\!\!P_{\boldsymbol{\theta}}\big), \end{aligned} \tag{5.11}$$

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

$$-I\!\!EL(\boldsymbol{\theta},\boldsymbol{\theta}^*) = \frac{1}{2}\|D_{\boldsymbol{\theta}^\circ}(\boldsymbol{\theta}-\boldsymbol{\theta}^*)\|^2, \tag{5.12}$$

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

$$I\!\!EL(\boldsymbol{\theta},\boldsymbol{\theta}^*) = \sum \big\{ f_i\Psi_i^\top(\boldsymbol{\theta}-\boldsymbol{\theta}^*) - d(\Psi_i^\top\boldsymbol{\theta}) + d(\Psi_i^\top\boldsymbol{\theta}^*) \big\}.$$

As $\boldsymbol{\theta}^*$ is the extreme point of $I\!\!EL(\boldsymbol{\theta})$, it holds $\nabla I\!\!EL(\boldsymbol{\theta}^*) = \sum \big[f_i - d'(\Psi_i^\top\boldsymbol{\theta}^*)\big]\Psi_i = 0$ and (5.11) follows. The Taylor expansion of the second order around $\boldsymbol{\theta}^*$ yields the expansion (5.12). □

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 (5.7). 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 $\mathfrak{n}_i^2 = d''(\Psi_i^\top\boldsymbol{\theta}^*)$. 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 $-I\!\!EL(\boldsymbol{\theta},\boldsymbol{\theta}^*)$ 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 5.6. *Suppose that*

$$\|I_p - D_0^{-1} D^2(\boldsymbol{\theta}) D_0^{-1}\|_\infty \le \delta(\mathtt{r}), \qquad \boldsymbol{\theta} \in \Theta_0(\mathtt{r}). \tag{5.13}$$

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* $\boldsymbol{\xi}_{\boldsymbol{\epsilon}}$*:*

$$D^2_{\boldsymbol{\epsilon}} = (1-\delta) D_0^2, \qquad \boldsymbol{\xi}_{\boldsymbol{\epsilon}} = (1+\delta)^{1/2} \boldsymbol{\xi}$$
$$D^2_{\underline{\boldsymbol{\epsilon}}} = (1+\delta) D_0^2, \qquad \boldsymbol{\xi}_{\underline{\boldsymbol{\epsilon}}} = (1-\delta)^{1/2} \boldsymbol{\xi}$$

with

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

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

Theorem 5.7. *Let* (e_1) *hold. Then for* $\boldsymbol{\epsilon} = (\delta, 0)$ *with* $\delta \ge \delta(\mathtt{r})$ *and any* $\boldsymbol{\theta} \in \Theta_0(\mathtt{r})$

$$\mathbb{L}_{\underline{\boldsymbol{\epsilon}}}(\boldsymbol{\theta}, \boldsymbol{\theta}^*) \le L(\boldsymbol{\theta}, \boldsymbol{\theta}^*) \le \mathbb{L}_{\boldsymbol{\epsilon}}(\boldsymbol{\theta}, \boldsymbol{\theta}^*). \tag{5.14}$$

Moreover, for any $z > 0$ *and* $\mathfrak{z} > 0$*, it holds*

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

Linearity of the stochastic component $\zeta(\boldsymbol{\theta})$ in the considered GLM implies important fact that the quantities $\diamondsuit_{\boldsymbol{\epsilon}}(\mathtt{r}), \diamondsuit_{\underline{\boldsymbol{\epsilon}}}(\mathtt{r})$ in the majorization bound (5.14) vanish for any $\mathtt{r}$. However, the deterministic component is not quadratic in $\boldsymbol{\theta}$ unless the function $d(w)$ is quadratic. Therefore, the presented bounds are local and have to be accomplished with the large deviation bounds.

An interesting question, similarly to the i.i.d. case, is the minimal radius $\mathtt{r}$ of the local vicinity $\Theta_0(\mathtt{r})$ ensuring the desirable concentration property. We apply the sufficient condition (4.9) of Corollary 4.3 ensuring the concentration property. By Lemma 5.5, the function $\mathfrak{M}(\mu, \boldsymbol{\theta}, \boldsymbol{\theta}^*)$ can be decomposed as

$$\mathfrak{M}(\mu, \boldsymbol{\theta}, \boldsymbol{\theta}^*) = \mu \mathcal{K}\bigl(I\!\!P_{\boldsymbol{\theta}^*}, I\!\!P_{\boldsymbol{\theta}}\bigr) - \mathfrak{N}\bigl(\mu, \boldsymbol{\theta}, \boldsymbol{\theta}^*\bigr).$$

Let $\mu = \mu(\boldsymbol{\theta})$ be selected such that $\mu |\Psi_i^\top (\boldsymbol{\theta} - \boldsymbol{\theta}^*)| \mathfrak{n}_i \le \mathtt{g}_1$ for all i. Then the arguments from (5.9) and (5.10) yield

$$\mathfrak{N}\bigl(\mu, \boldsymbol{\theta}, \boldsymbol{\theta}^*\bigr) \le \sum \log I\!\!E \exp\bigl\{\mu \mathfrak{n}_i \Psi_i^\top (\boldsymbol{\theta} - \boldsymbol{\theta}^*) \varepsilon_i / \mathfrak{n}_i\bigr\}$$
$$\le \frac{\nu_0^2 \mu^2}{2} \sum \bigl|\mathfrak{n}_i \Psi_i^\top (\boldsymbol{\theta} - \boldsymbol{\theta}^*)\bigr|^2 = \frac{\nu_0^2 \mu^2}{2} \|V_0(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\|^2.$$

Therefore,

$$\mathfrak{M}(\mu,\boldsymbol{\theta},\boldsymbol{\theta}^*) \geq \mu\mathcal{K}\bigl(I\!P_{\boldsymbol{\theta}^*}, I\!P_{\boldsymbol{\theta}}\bigr) - \frac{\nu_0^2\mu^2}{2}\|V_0(\boldsymbol{\theta}-\boldsymbol{\theta}^*)\|^2$$

and a reasonable choice of $\mu = \mu(\boldsymbol{\theta})$ is given by $\mu(\boldsymbol{\theta}) = \nu_0^{-2}\mathcal{K}\bigl(I\!P_{\boldsymbol{\theta}^*}, I\!P_{\boldsymbol{\theta}}\bigr)/\|V_0(\boldsymbol{\theta}-\boldsymbol{\theta}^*)\|^2$ leading to

$$\mathfrak{M}^*(\boldsymbol{\theta},\boldsymbol{\theta}^*) \geq \frac{\nu_0^{-2}\mathcal{K}^2\bigl(I\!P_{\boldsymbol{\theta}^*}, I\!P_{\boldsymbol{\theta}}\bigr)}{2\|V_0(\boldsymbol{\theta}-\boldsymbol{\theta}^*)\|^2}.$$

So, the Kullback-Leibler divergence $\mathcal{K}\bigl(I\!P_{\boldsymbol{\theta}^*}, I\!P_{\boldsymbol{\theta}}\bigr)$ should of order at least

$$\|V_0(\boldsymbol{\theta}-\boldsymbol{\theta}^*)\|\sqrt{\|p\log\bigl(V_0(\boldsymbol{\theta}-\boldsymbol{\theta}^*)\|\bigr)}.$$

One can check that this condition is fulfilled if the convex function $d(\cdot)$ satisfies $d'(t) \geq \text{Const.}/t$ for some $\text{Const} > 0$ and if the effective sample size N from (5.8) is sufficiently large.

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 $\Psi_i \in I\!R^p$:

$$Y_i = \Psi_i^\top\boldsymbol{\theta} + \varepsilon_i \tag{5.15}$$

where ε_i denotes the ith individual error.

The study of the qMLE in the GLM (5.5) heavily relies on the assumption (e_1). If this assumption is not verified, then the proposed approach would not apply. An explicit structure of the qMLE, especially in the linear regression case, allows for direct study of the properties of $\widetilde{\boldsymbol{\theta}}$ under weaker moment assumptions than (e_1). However, we aim at establishing some exponential bounds and therefore, the condition of bounded exponential moments for each observation is really necessary within the least squares or generalized linear approach. In the case of heavily tailed data with only polynomial moments, one can obtain some convergence results for the LSE $\widetilde{\boldsymbol{\theta}}$, however an exponential bound is not available. In such cases, it is natural to use a robustified version of the contrast, e.g. the least absolute deviation (LAD) method. We consider the linear model (5.15) and suppose for a moment that the errors ε_i are i.i.d. and follow the double exponential (Laplace) distribution with the density $(1/2)e^{-|y|}$. Then the model (5.15) yields the log-likelihood

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

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

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

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 $P_i = \mathcal{L}(Y_i - \Psi_i^\top \boldsymbol{\theta}^*)$ has a positive density $p_i(y)$.

Note that the difference $L(\boldsymbol{\theta}) - L(\boldsymbol{\theta}^*)$ is bounded by $\frac{1}{2} \sum |\Psi_i^\top (\boldsymbol{\theta} - \boldsymbol{\theta}^*)|$ and condition (E) is fulfilled automatically. Next we check conditions (ED_0) and (ED_1). Denote $\xi_i(\boldsymbol{\theta}) = I\!\!1(Y_i - \Psi_i^\top \boldsymbol{\theta} \le 0) - q_i(\boldsymbol{\theta})$ 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

$$\nabla \zeta(\boldsymbol{\theta}) = \sum \xi_i(\boldsymbol{\theta}) \Psi_i. \tag{5.16}$$

This expression differs from the similar ones from the linear and generalized linear regression because the error terms ξ_i now depends on $\boldsymbol{\theta}$. First we check the global condition (ED). 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 $|\lambda| \le \mathtt{g}_1$

$$\begin{aligned} \log I\!\!E \exp(\lambda \xi) &= \log\big[q \exp\{\lambda(1-q)\} + (1-q)\exp(-\lambda q)\big] \\ &\le \nu_0^2 q(1-q)\lambda^2/2, \end{aligned} \tag{5.17}$$

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

$$\begin{aligned} \log I\!\!E \exp\{\rho \boldsymbol{\gamma}^\top \nabla \zeta(\boldsymbol{\theta})\} &\le \frac{\nu_0^2 \rho^2}{2} \sum q_i(\boldsymbol{\theta}) \big\{1 - q_i(\boldsymbol{\theta})\big\} |\Psi_i^\top \boldsymbol{\gamma}|^2 \\ &\le \nu_0^2 \rho^2 \|V(\boldsymbol{\theta}) \boldsymbol{\gamma}\|^2/2, \end{aligned} \tag{5.18}$$

where

$$V^2(\boldsymbol{\theta}) = \sum q_i(\boldsymbol{\theta}) \big\{1 - q_i(\boldsymbol{\theta})\big\} \Psi_i \Psi_i^\top. \tag{5.19}$$

Denote also

$$V^2 = \frac{1}{4} \sum \Psi_i \Psi_i^\top.$$

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

$$N^{-1/2} \stackrel{\text{def}}{=} \max_i \sup_{\boldsymbol{\gamma} \in I\!R^p} \frac{\Psi_i^\top \boldsymbol{\gamma}}{2\|V\boldsymbol{\gamma}\|}; \tag{5.20}$$

cf. (5.9).

5.3.1 A local central bound

Now we restrict ourselves to the elliptic vicinity $\Theta_0(\mathtt{r}) = \{\boldsymbol{\theta} : \|V_0(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\| \le \mathtt{r}\}$ of the central point $\boldsymbol{\theta}^*$ for $V_0 = V(\boldsymbol{\theta}^*)$ and some $\mathtt{r} > 0$. Define the matrix $V_0 = V(\boldsymbol{\theta}^*)$. Then condition (ED_0) with the matrix V_0 and $\mathtt{g} = N^{1/2}\mathtt{g}_1$ is fulfilled on $\Theta_0(\mathtt{r})$ due to (5.18). Next, for checking (ED_1) suppose the following regularity condition:

$$V \le \nu_1 V_0 \tag{5.21}$$

for some $\nu_1 \ge 1$. This condition implies the inequality $|\Psi_i^\top \boldsymbol{\gamma}| \le \nu_1 N^{-1/2} \|V_0 \boldsymbol{\gamma}\|$ for any vector $\boldsymbol{\gamma} \in I\!R^p$. By (5.16)

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

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 (5.17) yields similarly to (5.18)

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

with

$$\omega(\mathtt{r}) \stackrel{\text{def}}{=} 4\nu_1 \max_{i \le n} \sup_{\boldsymbol{\theta} \in \Theta_0(\mathtt{r})} q_i(\boldsymbol{\theta}, \boldsymbol{\theta}^*).$$

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

$$|q_i(\boldsymbol{\theta}) - q_i(\boldsymbol{\theta}^*)| \le C\bigl|\Psi_i^\top (\boldsymbol{\theta} - \boldsymbol{\theta}^*)\bigr| \le C\nu_1 N^{-1/2} \|V_0(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\| \le C\nu_1 N^{-1/2} \mathtt{r}.$$

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

Lemma 5.8. *It holds for any* $\boldsymbol{\theta}$

$$\frac{\partial^2}{\partial^2 \boldsymbol{\theta}} I\!\!E L(\boldsymbol{\theta}) = D^2(\boldsymbol{\theta}) \stackrel{\text{def}}{=} \sum p_i\bigl(\Psi_i^\top(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\bigr)\Psi_i\Psi_i^\top, \tag{5.22}$$

where $p_i(\cdot)$ *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*

$$\begin{aligned} -I\!\!E L(\boldsymbol{\theta}, \boldsymbol{\theta}^*) &= \frac{1}{2}\sum |\Psi_i^\top(\boldsymbol{\theta} - \boldsymbol{\theta}^*)|^2 p_i(\Psi_i^\top(\boldsymbol{\theta}^\circ - \boldsymbol{\theta}^*)) \\ &= (\boldsymbol{\theta} - \boldsymbol{\theta}^*)^\top D^2(\boldsymbol{\theta}^\circ)(\boldsymbol{\theta} - \boldsymbol{\theta}^*)/2. \end{aligned} \tag{5.23}$$

Proof. Obviously

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

The identity (5.22) is obtained by one more differentiation. By definition, $\boldsymbol{\theta}^*$ is the extreme point of $I\!\!E L(\boldsymbol{\theta})$. The equality $\nabla I\!\!E L(\boldsymbol{\theta}^*) = 0$ yields

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

Now (5.23) follows by the Taylor expansion of the second order at $\boldsymbol{\theta}^*$. □

Define

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

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 (5.13); see Lemma 5.6. Now all the local conditions are fulfilled yielding the general majorizing bound of Theorem 3.1 and all its corollaries.

Theorem 5.9. *Let* $V_0 = V(\boldsymbol{\theta}^*)$; *see* (5.19). *Assume* (5.21) *for* $V^2 = (1/4)\sum \Psi_i\Psi_i^\top$. *Fix any* $\delta \ge \delta(\mathtt{r})$ *and* $\varrho \ge 3\nu_0\omega(\mathtt{r})$. *Then Theorem 3.1 and its corollaries holds for the linear median estimation.*

This example is one more confirmation of the applicability of the the general approach: as soon as the local conditions have been checked the main local statements follow for free. It only remains to accomplish them by a large deviation bound, that is, to describe the local vicinity $\Theta_0(\mathtt{r})$ providing the prescribed concentration bound.

5.3.2 A large deviation bound

A sufficient condition for the concentration property is that the rate function $\mathfrak{M}^*(\boldsymbol{\theta}, \boldsymbol{\theta}^*)$ grows at least logarithmic in $\|\boldsymbol{\theta}^* - \boldsymbol{\theta}\|$. For $y > 0$, define for $A_y = [y, \infty]$

$$\lambda_i(y) = -(2y)^{-1}\log[P_i(A_y)] = -(2y)^{-1}\log\bigl[I\!\!P(Y_i - \Psi_i^\top\boldsymbol{\theta}^* > y)\bigr].$$

The case with $\lambda_i(y) \geq \lambda_0 > 0$ corresponds to light tails while $\lambda_i(y) \to 0$ as $|y| \to \infty$ means heavy tails of the distribution P_i. Below we focus on the most interesting case when $\lambda_i(y)$ is positive and monotonously decreases to zero in $y > 0$. For simplicity of presentation we also assume that $\lambda_i(y)$ is sufficiently regular and its first derivative $\lambda_i'(y)$ is uniformly continuous on $\mathbb{R}$. Define

$$\xi_i(\boldsymbol{\theta}, \boldsymbol{\theta}^*) \stackrel{\text{def}}{=} \bigl\{ |Y_i - \Psi_i^\top \boldsymbol{\theta}| - |Y_i - \Psi_i^\top \boldsymbol{\theta}^*| \bigr\}/2$$
$$f_i(\mu) \stackrel{\text{def}}{=} \log \mathbb{E} \exp\bigl\{ \mu \xi_i(\boldsymbol{\theta}, \boldsymbol{\theta}^*) \bigr\}.$$

As $|\xi_i(\boldsymbol{\theta}, \boldsymbol{\theta}^*)| \leq |\Psi_i^\top (\boldsymbol{\theta} - \boldsymbol{\theta}^*)|/2$ and $f_i(\mu)$ is analytic in μ, it holds for any $\mu \geq 0$ with $\mu \max_i |\Psi_i^\top (\boldsymbol{\theta} - \boldsymbol{\theta}^*)|/2 \leq 1$

$$f_i(\mu) \geq -\mu \mathbb{E} \xi_i(\boldsymbol{\theta}, \boldsymbol{\theta}^*) - \mu^2 \operatorname{Var} \xi_i(\boldsymbol{\theta}, \boldsymbol{\theta}^*).$$

This implies by independence of the ξ_i's

$$\begin{aligned} \mathfrak{M}(\mu, \boldsymbol{\theta}, \boldsymbol{\theta}^*) &\stackrel{\text{def}}{=} -\log \mathbb{E} \exp\bigl\{ \mu L(\boldsymbol{\theta}, \boldsymbol{\theta}^*) \bigr\} \\ &= \sum f_i(\mu) \geq -\mu \mathbb{E} L(\boldsymbol{\theta}, \boldsymbol{\theta}^*) - \mu^2 \operatorname{Var} L(\boldsymbol{\theta}, \boldsymbol{\theta}^*). \end{aligned}$$

The choice $\mu = \mu(\boldsymbol{\theta}) = -\mathbb{E} L(\boldsymbol{\theta}, \boldsymbol{\theta}^*) \big/ \bigl\{ 2 \operatorname{Var} L(\boldsymbol{\theta}, \boldsymbol{\theta}^*) \bigr\}$ yields

$$\mathfrak{M}^*(\boldsymbol{\theta}, \boldsymbol{\theta}^*) \geq \frac{\bigl| \mathbb{E} L(\boldsymbol{\theta}, \boldsymbol{\theta}^*) \bigr|^2}{4 \operatorname{Var} L(\boldsymbol{\theta}, \boldsymbol{\theta}^*)}.$$

Note that the inequality $|\xi_i(\boldsymbol{\theta}, \boldsymbol{\theta}^*)| \leq \bigl| \Psi_i^\top (\boldsymbol{\theta} - \boldsymbol{\theta}^*) \bigr|/2$ implies

$$\operatorname{Var} L(\boldsymbol{\theta}, \boldsymbol{\theta}^*) \leq \frac{1}{2} \max_i |\Psi_i^\top (\boldsymbol{\theta} - \boldsymbol{\theta}^*)| \sum \mathbb{E} |\xi_i(\boldsymbol{\theta}, \boldsymbol{\theta}^*)|.$$

This easily yields that the sufficient condition (4.9) of Corollary 4.3 is fulfilled in this situation if N from (5.20) is sufficiently large.

6 Deviation probability for quadratic forms

The approximation results of the previous sections rely on the probability of the form $\mathbb{P}\bigl(\|\boldsymbol{\xi}\| > \mathtt{y} \bigr)$ for a given random vector $\boldsymbol{\xi} \in \mathbb{R}^p$. The only condition imposed on this vector is that

$$\log \mathbb{E} \exp\bigl(\boldsymbol{\gamma}^\top \boldsymbol{\xi} \bigr) \leq \nu_0^2 \|\boldsymbol{\gamma}\|^2/2, \qquad \boldsymbol{\gamma} \in \mathbb{R}^p, \, \|\boldsymbol{\gamma}\| \leq \mathtt{g}.$$

To simplify the presentation we rewrite this condition as

$$\log \mathbb{E} \exp\bigl(\boldsymbol{\gamma}^\top \boldsymbol{\xi} \bigr) \leq \|\boldsymbol{\gamma}\|^2/2, \qquad \boldsymbol{\gamma} \in \mathbb{R}^p, \, \|\boldsymbol{\gamma}\| \leq \mathtt{g}. \tag{6.1}$$

The general case can be reduced to $\nu_0 = 1$ by rescaling $\boldsymbol{\xi}$ and $\mathtt{g}$:

$$\log I\!\!E \exp\bigl(\boldsymbol{\gamma}^\top \boldsymbol{\xi}/\nu_0\bigr) \le \|\boldsymbol{\gamma}\|^2/2, \qquad \boldsymbol{\gamma} \in I\!\!R^p,\, \|\boldsymbol{\gamma}\| \le \nu_0 \mathtt{g}$$

that is, $\nu_0^{-1}\boldsymbol{\xi}$ fulfills (6.1) with a slightly increased $\mathtt{g}$. In typical situations like in Section 5, the value $\mathtt{g}$ is large (of order root-n) while the value ν_0 is close to one.

6.1 Gaussian case

Our benchmark will be a deviation bound for $\|\boldsymbol{\xi}\|^2$ for a standard Gaussian vector $\boldsymbol{\xi}$. The ultimate goal is to show that under (6.1) the norm of the vector $\boldsymbol{\xi}$ exhibits behavior expected for a Gaussian vector, at least in the region of moderate deviations. For the reason of comparison, we begin by stating the result for a Gaussian vector $\boldsymbol{\xi}$.

Theorem 6.1. *Let $\boldsymbol{\xi}$ be a standard normal vector in $I\!R^p$. Then for any $u > 0$, it holds*

$$I\!\!P\bigl(\|\boldsymbol{\xi}\|^2 > p + u\bigr) \le \exp\bigl\{-(p/2)\phi(u/p)\bigr]\bigr\}$$

with

$$\phi(t) \stackrel{\text{def}}{=} t - \log(1+t).$$

Let $\phi^{-1}(\cdot)$ stand for the inverse of $\phi(\cdot)$. For any $\mathtt{x}$,

$$I\!\!P\bigl(\|\boldsymbol{\xi}\|^2 > p + \phi^{-1}(2\mathtt{x}/p)\bigr) \le \exp(-\mathtt{x}).$$

This particularly yields with $\varkappa = 6.6$

$$I\!\!P\bigl(\|\boldsymbol{\xi}\|^2 > p + \sqrt{\varkappa \mathtt{x} p} \vee (\varkappa \mathtt{x})\bigr) \le \exp(-\mathtt{x}).$$

Proof. The proof utilizes the following well known fact: for $\mu < 1$

$$\log I\!\!E \exp\bigl(\mu \|\boldsymbol{\xi}\|^2/2\bigr) = -0.5p \log(1-\mu).$$

It can be obtained by straightforward calculus. Now consider any $u > 0$. By the exponential Chebyshev inequality

$$\begin{aligned} I\!\!P\bigl(\|\boldsymbol{\xi}\|^2 > p + u\bigr) &\le \exp\bigl\{-\mu(p+u)/2\bigr\} I\!\!E \exp\bigl(\mu\|\boldsymbol{\xi}\|^2/2\bigr) \\ &= \exp\bigl\{-\mu(p+u)/2 - (p/2)\log(1-\mu)\bigr\}. \end{aligned} \tag{6.2}$$

It is easy to see that the value $\mu = u/(u+p)$ maximizes $\mu(p+u) + p\log(1-\mu)$ w.r.t. μ yielding

$$\mu(p+u) - p\log(1-\mu) = u - p\log(1+u/p).$$

Further we use that $x-\log(1+x) \geq a_0 x^2$ for $x \leq 1$ and $x-\log(1+x) \geq a_0 x$ for $x > 1$ with $a_0 = 1-\log(2) \geq 0.3$. This implies with $x = u/p$ for $u = \sqrt{\varkappa \mathtt{x} p}$ or $u = \varkappa \mathtt{x}$ and $\varkappa = 2/a_0 < 6.6$ that

$$I\!P\bigl(\|\boldsymbol{\xi}\|^2 \geq p + \sqrt{\varkappa \mathtt{x} p} \vee (\varkappa \mathtt{x})\bigr) \leq \exp(-\mathtt{x})$$

as required. □

The message of this result is that the squared norm of the Gaussian vector $\boldsymbol{\xi}$ concentrates around the value p and the deviation over the level $p + \sqrt{\mathtt{x} p}$ are exponentially small in $\mathtt{x}$.

A similar bound can be obtained for a norm of the vector $I\!B\boldsymbol{\xi}$ where $I\!B$ is some given matrix. For notational simplicity we assume that $I\!B$ is symmetric. Otherwise one should replace it with $(I\!B^\top A)^{1/2}$.

Theorem 6.2. *Let $\boldsymbol{\xi}$ be standard normal in $I\!R^p$. Then for every $\mathtt{x} > 0$ and any symmetric matrix $I\!B$, it holds with $\mathtt{p} = \operatorname{tr}(I\!B^2)$, $\mathtt{v}^2 = 2\operatorname{tr}(I\!B^4)$, and $a^* = \|I\!B^2\|_\infty$*

$$I\!P\bigl(\|I\!B\boldsymbol{\xi}\|^2 > \mathtt{p} + (2\mathtt{v}\mathtt{x}^{1/2}) \vee (6a^*\mathtt{x})\bigr) \leq \exp(-\mathtt{x}).$$

Proof. The matrix $I\!B^2$ can be represented as $U^\top \operatorname{diag}(a_1, \ldots, a_p) U$ for an orthogonal matrix U. The vector $\widetilde{\boldsymbol{\xi}} = U\boldsymbol{\xi}$ is also standard normal and $\|I\!B\boldsymbol{\xi}\|^2 = \widetilde{\boldsymbol{\xi}}^\top U I\!B^2 U^\top \widetilde{\boldsymbol{\xi}}$. This means that one can reduce the situation to the case of a diagonal matrix $I\!B^2 = \operatorname{diag}(a_1, \ldots, a_p)$. We can also assume without loss of generality that $a_1 \geq a_2 \geq \ldots \geq a_p$. The expressions for the quantities $\mathtt{p}$ and $\mathtt{v}^2$ simplifies to

$$\begin{aligned} \mathtt{p} &= \operatorname{tr}(I\!B^2) = a_1 + \ldots + a_p, \\ \mathtt{v}^2 &= 2\operatorname{tr}(I\!B^4) = 2(a_1^2 + \ldots + a_p^2). \end{aligned}$$

Moreover, rescaling the matrix $I\!B^2$ by a_1 reduces the situation to the case with $a_1 = 1$.

Lemma 6.3. *It holds*

$$I\!E\|I\!B\boldsymbol{\xi}\|^2 = \operatorname{tr}(I\!B^2), \qquad \operatorname{Var}\bigl(\|I\!B\boldsymbol{\xi}\|^2\bigr) = 2\operatorname{tr}(I\!B^4).$$

Moreover, for $\mu < 1$

$$I\!E \exp\bigl\{\mu\|I\!B\boldsymbol{\xi}\|^2/2\bigr\} = \det(1-\mu I\!B^2)^{-1/2} = \prod_{i=1}^{p}(1-\mu a_i)^{-1/2}. \tag{6.3}$$

Proof. If $I\!B^2$ is diagonal, then $\|I\!B\boldsymbol{\xi}\|^2 = \sum_i a_i \xi_i^2$ and the summands $a_i\xi_i^2$ are independent. It remains to note that $I\!E(a_i\xi_i^2) = a_i$, $\mathrm{Var}(a_i\xi_i^2) = 2a_i^2$, and for $\mu a_i < 1$,

$$I\!E \exp\bigl\{\mu a_i \xi_i^2/2\bigr\} = (1-\mu a_i)^{-1/2}$$

yielding (6.3). □

Given u, fix $\mu < 1$. The exponential Markov inequality yields

$$\begin{aligned} I\!P\bigl(\|I\!B\boldsymbol{\xi}\|^2 > \mathtt{p} + u\bigr) &\le \exp\Bigl\{-\frac{\mu(\mathtt{p}+u)}{2}\Bigr\} I\!E \exp\Bigl(\frac{\mu\|I\!B\boldsymbol{\xi}\|^2}{2}\Bigr) \\ &\le \exp\Bigl\{-\frac{\mu u}{2} - \frac{1}{2}\sum_{i=1}^{p}\bigl[\mu a_i + \log(1-\mu a_i)\bigr]\Bigr\}. \end{aligned}$$

We start with the case when $\mathtt{x}^{1/2} \le v/3$. Then $u = 2\mathtt{x}^{1/2}\mathtt{v}$ fulfills $u \le 2v^2/3$. Define $\mu = u/\mathtt{v}^2 \le 2/3$ and use that $t + \log(1-t) \ge -t^2$ for $t \le 2/3$. This implies

$$\begin{aligned} &I\!P\bigl(\|I\!B\boldsymbol{\xi}\|^2 > \mathtt{p} + u\bigr) \\ &\quad\le \exp\Bigl\{-\frac{\mu u}{2} + \frac{1}{2}\sum_{i=1}^{p}\mu^2 a_i^2\Bigr\} = \exp\bigl(-u^2/(4\mathtt{v}^2)\bigr) = e^{-\mathtt{x}}. \end{aligned} \tag{6.4}$$

Next, let $\mathtt{x}^{1/2} > v/3$. Set $\mu = 2/3$. It holds similarly to the above

$$\sum_{\mathcal{J}_1}\bigl[\mu a_i + \log(1-\mu a_i)\bigr] \ge -\sum_{\mathcal{J}_1}\mu^2 a_i^2 \ge -2\mathtt{v}^2/9 \ge -2\mathtt{x}.$$

Now, for $u = 6\mathtt{x}$ and $\mu u/2 = 2\mathtt{x}$, (6.4) implies

$$I\!P\bigl(\|I\!B\boldsymbol{\xi}\|^2 > \mathtt{p} + u\bigr) \le \exp\bigl\{-\bigl(2\mathtt{x} - \mathtt{x}\bigr)\bigr\} = \exp(-\mathtt{x})$$

as required. □

Below we establish similar bounds for a non-Gaussian vector $\boldsymbol{\xi}$ obeying (6.1).

6.2 A bound for the ℓ_2-norm

This section presents a general exponential bound for the probability $I\!P\bigl(\|\boldsymbol{\xi}\| > \mathtt{y}\bigr)$ under (6.1). Define the value

$$\mathtt{x}_c = 0.5\bigl[\mathtt{g}^2 - p\log\bigl(1 + \mathtt{g}^2/p\bigr)\bigr]. \tag{6.5}$$

Theorem 6.4. *Let $\boldsymbol{\xi} \in I\!R^p$ fulfill (6.1). Then it holds for each $\mathtt{x} \le \mathtt{x}_c$*

$$I\!P\bigl(\|\boldsymbol{\xi}\|^2 > p + \sqrt{\varkappa \mathtt{x} p} \vee (\varkappa\mathtt{x}), \|\boldsymbol{\xi}\|^2 \le p + \mathtt{g}^2\bigr) \le 2\exp(-\mathtt{x}),$$

where $\varkappa = 6.6$. *Moreover, for* $\mathtt{y}^2 \geq \mathtt{y}_c^2 \stackrel{\text{def}}{=} p + \mathtt{g}^2$, *it holds with* $\mathtt{g}_c = \mathtt{g}^2(p+\mathtt{g}^2)^{-1/2}$

$$\begin{aligned} I\!P\bigl(\|\boldsymbol{\xi}\|^2 > \mathtt{y}^2\bigr) &\leq 8.4 \exp\bigl\{-\mathtt{g}_c\mathtt{y}/2 - (p/2)\log(1-\mathtt{g}_c/\mathtt{y})\bigr\} \\ &\leq 8.4 \exp\bigl\{-\mathtt{x}_c - \mathtt{g}_c(\mathtt{y}-\mathtt{y}_c)/2\bigr\}. \end{aligned}$$

Proof. The main step of the proof is the following exponential bound.

Lemma 6.5. *Suppose* (6.1). *For any* $\mu < 1$ *with* $\mathtt{g}^2 > p\mu$, *it holds*

$$I\!E \exp\Bigl(\frac{\mu\|\boldsymbol{\xi}\|^2}{2}\Bigr)\, 1\!\!1\Bigl(\|\boldsymbol{\xi}\|^2 \leq \frac{\mathtt{g}^2}{\mu^2} - \frac{p}{\mu}\Bigr) \leq 2(1-\mu)^{-p/2}. \tag{6.6}$$

Proof. Let $\boldsymbol{\varepsilon}$ be a standard normal vector in $I\!R^p$ and $\boldsymbol{u} \in I\!R^p$. Given $\mathtt{r} > 0$, define $m(\boldsymbol{u}, r) = I\!P\bigl(\|\boldsymbol{\varepsilon} - \boldsymbol{u}\| \leq \mathtt{r}\bigr)$. This value can be rewritten as

$$\begin{aligned} m(\boldsymbol{u}, r) &= I\!P\bigl(\|\boldsymbol{\varepsilon} - \boldsymbol{u}\|^2 \leq \mathtt{r}^2\bigr) \\ &= I\!P\bigl(\|\boldsymbol{\varepsilon}\|^2 - I\!E\|\boldsymbol{\varepsilon}\|^2 - 2\boldsymbol{u}^\top\boldsymbol{\varepsilon} \leq \mathtt{r}^2 - \|\boldsymbol{u}\|^2 - I\!E\|\boldsymbol{\varepsilon}\|^2\bigr), \end{aligned}$$

and $m(\boldsymbol{u}, r) \geq 1/2$ for $\mathtt{r}^2 \geq \|\boldsymbol{u}\|^2 + I\!E\|\boldsymbol{\varepsilon}\|^2 = \|\boldsymbol{u}\|^2 + p$. Let us fix some $\boldsymbol{\xi}$ with $\mu\|\boldsymbol{\xi}\|^2 \leq \mathtt{g}^2/\mu - p$ and denote by $I\!P_{\boldsymbol{\xi}}$ the conditional probability given $\boldsymbol{\xi}$. It holds with $c_p = (2\pi)^{-p/2}$

$$\begin{aligned} & c_p \int \exp\Bigl(\boldsymbol{\gamma}^\top\boldsymbol{\xi} - \frac{\|\boldsymbol{\gamma}\|^2}{2\mu}\Bigr)\, 1\!\!1(\|\boldsymbol{\gamma}\| \leq \mathtt{g}) d\boldsymbol{\gamma} \\ &\quad = c_p \exp\bigl(\mu\|\boldsymbol{\xi}\|^2/2\bigr) \int \exp\Bigl(-\frac{1}{2}\bigl\|\mu^{-1/2}\boldsymbol{\gamma} - \mu^{1/2}\boldsymbol{\xi}\bigr\|^2\Bigr)\, 1\!\!1(\mu^{-1}\|\boldsymbol{\gamma}\|^2 \leq \mu^{-1}\mathtt{g}^2) d\boldsymbol{\gamma} \\ &\quad = \mu^{p/2} \exp\bigl(\mu\|\boldsymbol{\xi}\|^2/2\bigr) I\!P_{\boldsymbol{\xi}}\bigl(\|\boldsymbol{\varepsilon} + \mu^{1/2}\boldsymbol{\xi}\|^2 \leq \mu^{-1}\mathtt{g}^2\bigr) \\ &\quad \geq 0.5\mu^{p/2} \exp\bigl(\mu\|\boldsymbol{\xi}\|^2/2\bigr), \end{aligned}$$

because $\|\mu^{1/2}\boldsymbol{\xi}\|^2 + p \leq \mu^{-1}\mathtt{g}^2$. This implies in view of $p < \mathtt{g}^2/\mu$ that

$$\begin{aligned} & \exp\bigl(\mu\|\boldsymbol{\xi}\|^2/2\bigr)\, 1\!\!1\bigl(\|\boldsymbol{\xi}\|^2 \leq \mathtt{g}^2/\mu^2 - p/\mu\bigr) \\ &\quad \leq 2\mu^{-p/2} c_p \int \exp\Bigl(\boldsymbol{\gamma}^\top\boldsymbol{\xi} - \frac{\|\boldsymbol{\gamma}\|^2}{2\mu}\Bigr)\, 1\!\!1(\|\boldsymbol{\gamma}\| \leq \mathtt{g}) d\boldsymbol{\gamma}. \end{aligned}$$

Further, by (6.1)

$$\begin{aligned} & c_p I\!E \int \exp\Bigl(\boldsymbol{\gamma}^\top\boldsymbol{\xi} - \frac{1}{2\mu}\|\boldsymbol{\gamma}\|^2\Bigr)\, 1\!\!1(\|\boldsymbol{\gamma}\| \leq \mathtt{g}) d\boldsymbol{\gamma} \\ &\quad \leq c_p \int \exp\Bigl(-\frac{\mu^{-1}-1}{2}\|\boldsymbol{\gamma}\|^2\Bigr)\, 1\!\!1(\|\boldsymbol{\gamma}\| \leq \mathtt{g}) d\boldsymbol{\gamma} \\ &\quad \leq c_p \int \exp\Bigl(-\frac{\mu^{-1}-1}{2}\|\boldsymbol{\gamma}\|^2\Bigr) d\boldsymbol{\gamma} \\ &\quad \leq (\mu^{-1}-1)^{-p/2} \end{aligned}$$

and (6.6) follows. □

Due to this result, the scaled squared norm $\mu\|\boldsymbol{\xi}\|^2/2$ after a proper truncation possesses the same exponential moments as in the Gaussian case. A straightforward implication is the probability bound $I\!P\bigl(\|\boldsymbol{\xi}\|^2 > p+u\bigr)$ for moderate values u. Namely, given $u > 0$, define $\mu = u/(u+p)$. This value optimizes the inequality (6.2) in the Gaussian case. Now we can apply a similar bound under the constraints $\|\boldsymbol{\xi}\|^2 \le \mathtt{g}^2/\mu^2 - p/\mu$. Therefore, the bound is only meaningful if $p + u \le \mathtt{g}^2/\mu^2 - p/\mu$ with $\mu = u/(u+p)$. One can check that the largest value u for which this constraint is still valid, is given by $u = \mathtt{g}^2$. Hence, (6.6) yields for $u \le \mathtt{g}^2$

$$\begin{aligned} &I\!P\bigl(\|\boldsymbol{\xi}\|^2 > p+u, \|\boldsymbol{\xi}\|^2 \le p+\mathtt{g}^2\bigr) \\ &\quad\le \exp\Bigl\{-\frac{\mu(p+u)}{2}\Bigr\} I\!E \exp\Bigl(\frac{\mu\|\boldsymbol{\xi}\|^2}{2}\Bigr)\, 1\!\!1\Bigl(\|\boldsymbol{\xi}\|^2 \le \frac{\mathtt{g}^2}{\mu^2} - \frac{p}{\mu}\Bigr) \\ &\quad\le 2\exp\bigl\{-0.5\bigl[\mu(p+u) + p\log(1-\mu)\bigr]\bigr\} \\ &\quad= 2\exp\bigl\{-0.5\bigl[u - p\log(1+u/p)\bigr]\bigr\}. \end{aligned}$$

Similarly to the Gaussian case, this implies with $\varkappa = 6.6$ that

$$I\!P\bigl(\|\boldsymbol{\xi}\|^2 \ge p + \sqrt{\varkappa \mathtt{x} p} \vee (\varkappa\mathtt{x}), \|\boldsymbol{\xi}\|^2 \le p + \mathtt{g}^2\bigr) \le 2\exp(-\mathtt{x}).$$

The Gaussian case yields (6.1) with $\mathtt{g} = \infty$ and the result is done. In the non-Gaussian case with a finite $\mathtt{g}$, we have to accompany the moderate deviation bound with a large deviation bound $I\!P\bigl(\|\boldsymbol{\xi}\| > \mathtt{y}\bigr)$ for $\mathtt{y}^2 \ge p + \mathtt{g}^2$. This is done by combining the bound (6.6) with the standard slicing arguments.

Lemma 6.6. *Let* $\mu_0 \le \mathtt{g}^2/p$. *Define* $\mathtt{y}_0^2 = \mathtt{g}^2/\mu_0^2 - p/\mu_0$ *and* $\mathtt{g}_0^2 = \mathtt{g}^2 - \mu_0 p$. *It holds for* $\mathtt{y} \ge \mathtt{y}_0$

$$I\!P\bigl(\|\boldsymbol{\xi}\| > \mathtt{y}\bigr) \le 8.4(1 - \mathtt{g}_0/\mathtt{y})^{-p/2} \exp\bigl(-\mathtt{g}_0\mathtt{y}/2\bigr) \tag{6.7}$$

$$\le 8.4\exp\bigl\{-\mathtt{x}_0 - \mathtt{g}_0(\mathtt{y} - \mathtt{y}_0)/2\bigr\}. \tag{6.8}$$

with $\mathtt{x}_0$ *defined by*

$$2\mathtt{x}_0 = \mu_0\mathtt{y}_0^2 + p\log(1-\mu_0) = \mathtt{g}^2/\mu_0 - p + p\log(1-\mu_0).$$

Proof. Consider the growing sequence $\mathtt{y}_k$ with $\mathtt{y}_1 = \mathtt{y}$ and $\mathtt{g}_0\mathtt{y}_{k+1} = \mathtt{g}_0\mathtt{y} + k$. Define also $\mu_k = \mathtt{g}_0/\mathtt{y}_k$. In particular, $\mu_k \le \mu_1 = \mathtt{g}_0/\mathtt{y}$. Obviously

$$I\!P\bigl(\|\boldsymbol{\xi}\| > \mathtt{y}\bigr) = \sum_{k=1}^{\infty} I\!P\bigl(\|\boldsymbol{\xi}\| > \mathtt{y}_k, \|\boldsymbol{\xi}\| \le \mathtt{y}_{k+1}\bigr).$$

Now we try to evaluate every slicing probability in this expression. We use that

$$\mu_{k+1}\mathtt{y}_k^2 = \frac{(\mathtt{g}_0\mathtt{y}+k-1)^2}{\mathtt{g}_0\mathtt{y}+k} \geq \mathtt{g}_0\mathtt{y}+k-2,$$

and also $\mathtt{g}^2/\mu_k^2 - p/\mu_k \geq \mathtt{y}_k^2$ because $\mathtt{y} \geq \mathtt{y}_0$ and

$$\mathtt{g}^2/\mu_k^2 - p/\mu_k - \mathtt{y}_k^2 = \mu_k^{-2}(\mathtt{g}^2 - \mu_k p - \mathtt{g}_0^2) \geq \mu_k^{-2}(\mathtt{g}^2 - \mathtt{g}_0 p/\mathtt{y} - \mathtt{g}_0^2) \geq 0.$$

Hence by (6.6)

$$\begin{aligned}
I\!P\Big(\|\boldsymbol{\xi}\| > \mathtt{y}\Big) &\leq \sum_{k=1}^{\infty} I\!P\Big(\|\boldsymbol{\xi}\| > \mathtt{y}_k, \|\boldsymbol{\xi}\| \leq \mathtt{y}_{k+1}\Big) \\
&\leq \sum_{k=1}^{\infty} \exp\Big(-\frac{\mu_{k+1}\mathtt{y}_k^2}{2}\Big) I\!E \exp\Big(\frac{\mu_{k+1}\|\boldsymbol{\xi}\|^2}{2}\Big) 1\!\!1\big(\|\boldsymbol{\xi}\|^2 \leq \mathtt{y}_{k+1}^2\big) \\
&\leq \sum_{k=1}^{\infty} 2\big(1-\mu_{k+1}\big)^{-p/2} \exp\Big(-\frac{\mu_{k+1}\mathtt{y}_k^2}{2}\Big) \\
&\leq 2\big(1-\mu_1\big)^{-p/2} \sum_{k=1}^{\infty} \exp\Big(-\frac{\mathtt{g}_0\mathtt{y}+k-2}{2}\Big) \\
&= 2e^{1/2}(1-e^{-1/2})^{-1}(1-\mu_1)^{-p/2} \exp\big(-\mathtt{g}_0\mathtt{y}/2\big) \\
&\leq 8.4(1-\mu_1)^{-p/2} \exp\big(-\mathtt{g}_0\mathtt{y}/2\big)
\end{aligned}$$

and the first assertion follows. For $\mathtt{y} = \mathtt{y}_0$, it holds

$$\mathtt{g}_0\mathtt{y}_0 + p\log(1-\mu_0) = \mu_0\mathtt{y}_0^2 + p\log(1-\mu_0) = 2\mathtt{x}_0$$

and (6.7) implies $I\!P\big(\|\boldsymbol{\xi}\| > \mathtt{y}_0\big) \leq 8.4\exp(-\mathtt{x}_0)$. Now observe that the function $f(\mathtt{y}) = \mathtt{g}_0\mathtt{y}/2 + (p/2)\log\big(1 - \mathtt{g}_0/\mathtt{y}\big)$ fulfills $f(\mathtt{y}_0) = \mathtt{x}_0$ and $f'(\mathtt{y}) \geq \mathtt{g}_0/2$ yielding $f(\mathtt{y}) \geq \mathtt{x}_0 + \mathtt{g}_0(\mathtt{y}-\mathtt{y}_0)/2$. This implies (6.8). □

The statements of the theorem are obtained by applying the lemmas with $\mu_0 = \mu_c = \mathtt{g}^2/(p+\mathtt{g}^2)$. This also implies $\mathtt{y}_c^2 = p + \mathtt{g}^2$, $\mathtt{g}_c\mathtt{y}_c = \mathtt{g}^2$, $1 - \mu_c = p/(p+\mathtt{g}^2)$, and $\mathtt{x}_c$ from (6.5). □

The statements of Theorem 6.8 can be represented in the form:

Corollary 6.7. *Let $\boldsymbol{\xi}$ fulfill* (6.1). *Then it holds for $\mathtt{x} \leq \mathtt{x}_c = 0.5\big[\mathtt{g}^2 - p\log\big(1+\mathtt{g}^2/p\big)\big]$:*

$$I\!P\big(\|\boldsymbol{\xi}\|^2 \geq \mathfrak{z}(\mathtt{x},p)\big) \;\leq\; 2\mathrm{e}^{-\mathtt{x}} + 8.4\mathrm{e}^{-\mathtt{x}_c}, \tag{6.9}$$

$$\mathfrak{z}(\mathtt{x},p) \stackrel{\mathrm{def}}{=} \begin{cases} p + \sqrt{\varkappa\mathtt{x}p}, & \mathtt{x} \leq p/\varkappa, \\ p + \varkappa\mathtt{x} & p/\varkappa < \mathtt{x} \leq \mathtt{x}_c. \end{cases} \tag{6.10}$$

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

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

This result implicitly assumes that $p \leq \varkappa \mathtt{x}_c$ which is fulfilled if $u_0 = \mathtt{g}^2/p \geq 1$:

$$\varkappa \mathtt{x}_c = 0.5\varkappa\bigl[u_0 - \log(1+u_0)\bigr]p \geq 3.3\bigl[1 - \log(2)\bigr]p > p.$$

In the zone $\mathtt{x} \leq p/\varkappa$ we obtain sub-Gaussian behavior of the tail of $\|\boldsymbol{\xi}\|^2 - p$, in the zone $p/\varkappa < \mathtt{x} \leq \mathtt{x}_c$ it becomes sub-exponential. Note that the sub-exponential zone is empty if $\mathtt{g}^2 < p$.

For $\mathtt{x} \leq \mathtt{x}_c$, the function $\mathfrak{z}(\mathtt{x}, p)$ mimics the quantile behavior of the chi-squared distribution χ^2_p with p degrees of freedom. Moreover, increase the dimension p yields growth of the sub-Gaussian zone.

Finally, in the large deviation zone $\mathtt{x} > \mathtt{x}_c$ the deviation probability decays as $\mathrm{e}^{-c\mathtt{x}^{1/2}}$ for some fixed c. However, if the constant $\mathtt{g}$ in the condition (6.1) is sufficiently large relative to p, then $\mathtt{x}_c$ is large as well and the large deviation zone $\mathtt{x} > \mathtt{x}_c$ can be ignored at a small price of $8.4\mathrm{e}^{-\mathtt{x}_c}$ and one can focus on the deviation bound described by (6.9) and (6.10).

6.3 A bound for a quadratic form

Now we extend the result to more general bound for $\|I\!B\boldsymbol{\xi}\|^2 = \boldsymbol{\xi}^\top I\!B^2\boldsymbol{\xi}$ with a given matrix $I\!B$ and a vector $\boldsymbol{\xi}$ obeying the condition (6.1). Similarly to the Gaussian case we assume that $I\!B$ is symmetric. Define important characteristics of $I\!B$

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

For simplicity of formulation we suppose that $\lambda^* = 1$, otherwise one has to replace $\mathtt{p}$ and $\mathtt{v}^2$ with $\mathtt{p}/\lambda^*$ and $\mathtt{v}^2/\lambda^*$.

Let $\mathtt{g}$ be shown in (6.1). Define similarly to the ℓ_2-case $\mu_c = \mathtt{g}^2/(\mathtt{p} + \mathtt{g}^2)$. Further define the values $\mathtt{y}_c, \mathtt{g}_c$, and $\mathtt{x}_c$ by

$$\begin{aligned} \mathtt{y}_c^2 &\stackrel{\text{def}}{=} \mathtt{g}^2/\mu_c^2 - \mathtt{p}/\mu_c, \\ \mathtt{g}_c &\stackrel{\text{def}}{=} \mu_c \mathtt{y}_c = \sqrt{\mathtt{g}^2 - \mu_c \mathtt{p}}, \\ 2\mathtt{x}_c &\stackrel{\text{def}}{=} \mathtt{g}_c \mathtt{y}_c + \log\det\bigl(I_p - \mu_c I\!B^2\bigr). \end{aligned} \tag{6.11}$$

Theorem 6.8. *Let a random vector* $\boldsymbol{\xi}$ *in* $I\!R^p$ *fulfill* (6.1). *Then for each* $\mathtt{x} < \mathtt{x}_c$

$$I\!P\bigl(\|I\!B\boldsymbol{\xi}\|^2 > \mathtt{p} + (2\mathtt{v}\mathtt{x}^{1/2}) \vee (a_c\mathtt{x}), \|I\!B\boldsymbol{\xi}\|^2 \leq \mathtt{y}_c^2\bigr) \leq 2\exp(-\mathtt{x})$$

with $a_c \stackrel{\text{def}}{=} 6 \vee (4\mu_c^{-1})$. *Moreover, for* $\mathtt{y} \geq \mathtt{y}_c$, *it holds*

$$I\!P\bigl(\|I\!B\boldsymbol{\xi}\|^2 > \mathtt{y}^2\bigr) \leq 8.4 \exp\bigl(-\mathtt{x}_c - \mathtt{g}_c(\mathtt{y} - \mathtt{y}_c)/2\bigr).$$

Proof. The main steps of the proof are similar to the proof of Theorem 6.4.

Lemma 6.9. *Suppose* (6.1). *For any* $\mu < 1$ *with* $\mathtt{g}^2/\mu \geq \mathtt{p}$, *it holds*

$$I\!E \exp\bigl(\mu\|I\!B\boldsymbol{\xi}\|^2/2\bigr)\, 1\!\!1\bigl(\|I\!B^2\boldsymbol{\xi}\|^2 \leq \mathtt{g}^2/\mu^2 - \mathtt{p}/\mu\bigr) \leq 2\det(I_p - \mu I\!B^2)^{-1/2}. \tag{6.12}$$

Proof. With $c_p(I\!B) = \bigl(2\pi\bigr)^{-p/2} \det(I\!B^{-1})$

$$\begin{aligned}
& c_p(I\!B) \int \exp\Bigl(\boldsymbol{\gamma}^\top \boldsymbol{\xi} - \frac{1}{2\mu}\|I\!B^{-1}\boldsymbol{\gamma}\|^2\Bigr)\, 1\!\!1(\|\boldsymbol{\gamma}\| \leq \mathtt{g}) d\boldsymbol{\gamma} \\
&\quad = c_p(I\!B) \exp\Bigl(\frac{\mu\|I\!B\boldsymbol{\xi}\|^2}{2}\Bigr) \int \exp\Bigl(-\frac{1}{2}\bigl\|\mu^{1/2} I\!B\boldsymbol{\xi} - \mu^{-1/2} I\!B^{-1}\boldsymbol{\gamma}\bigr\|^2\Bigr)\, 1\!\!1(\|\boldsymbol{\gamma}\| \leq \mathtt{g}) d\boldsymbol{\gamma} \\
&\quad = \mu^{p/2} \exp\Bigl(\frac{\mu\|I\!B\boldsymbol{\xi}\|^2}{2}\Bigr) I\!P_{\boldsymbol{\xi}}\bigl(\|I\!B\boldsymbol{\varepsilon} - \mu^{1/2} I\!B^2\boldsymbol{\xi}\| \leq \mathtt{g}\mu^{-1/2}\bigr),
\end{aligned}$$

where $\boldsymbol{\varepsilon}$ denotes a standard normal vector in $I\!R^p$ and $I\!P_{\boldsymbol{\xi}}$ means the conditional expectation given $\boldsymbol{\xi}$. Moreover, for any $\boldsymbol{u} \in I\!R^p$ and $\mathtt{r} \geq \mathtt{p} + \|\boldsymbol{u}\|^2$, it holds in view of $I\!E\|I\!B\boldsymbol{\varepsilon}\|^2 = \mathtt{p}$

$$\begin{aligned}
I\!P\bigl(\|I\!B\boldsymbol{\varepsilon} - \boldsymbol{u}\| \leq \mathtt{r}\bigr) &= I\!P\bigl(\|I\!B\boldsymbol{\varepsilon}\|^2 - \mathtt{p} - 2\boldsymbol{u}^\top I\!B\boldsymbol{\varepsilon} \leq \mathtt{r}^2 - \|\boldsymbol{u}\|^2 - \mathtt{p}\bigr) \\
&\geq I\!P\bigl(\|I\!B\boldsymbol{\varepsilon}\|^2 - I\!E\|I\!B\boldsymbol{\varepsilon}\|^2 - 2\boldsymbol{u}^\top I\!B\boldsymbol{\varepsilon} \leq 0\bigr) \geq 1/2.
\end{aligned}$$

This implies

$$\begin{aligned}
& \exp\Bigl(\mu\|I\!B\boldsymbol{\xi}\|^2/2\Bigr)\, 1\!\!1\bigl(\|I\!B^2\boldsymbol{\xi}\|^2 \leq \mathtt{g}^2/\mu^2 - \mathtt{p}/\mu\bigr) \\
&\quad \leq 2\mu^{-p/2} c_p(I\!B) \int \exp\Bigl(\boldsymbol{\gamma}^\top \boldsymbol{\xi} - \frac{1}{2\mu}\|I\!B^{-1}\boldsymbol{\gamma}\|^2\Bigr)\, 1\!\!1(\|\boldsymbol{\gamma}\| \leq \mathtt{g}) d\boldsymbol{\gamma}.
\end{aligned}$$

Further, by (6.1)

$$\begin{aligned}
& c_p(I\!B) I\!E \int \exp\Bigl(\boldsymbol{\gamma}^\top \boldsymbol{\xi} - \frac{1}{2\mu}\|I\!B^{-1}\boldsymbol{\gamma}\|^2\Bigr)\, 1\!\!1(\|\boldsymbol{\gamma}\| \leq \mathtt{g}) d\boldsymbol{\gamma} \\
&\quad \leq c_p(I\!B) \int \exp\Bigl(\frac{\|\boldsymbol{\gamma}\|^2}{2} - \frac{1}{2\mu}\|I\!B^{-1}\boldsymbol{\gamma}\|^2\Bigr) d\boldsymbol{\gamma} \\
&\quad \leq \det(I\!B^{-1}) \det(\mu^{-1} I\!B^{-2} - I_p)^{-1/2} = \mu^{p/2} \det(I_p - \mu I\!B^2)^{-1/2}
\end{aligned}$$

and (6.12) follows. □

Now we evaluate the probability $I\!P\bigl(\|I\!B\boldsymbol{\xi}\| > \mathtt{y}\bigr)$ for moderate values of $\mathtt{y}$.

Lemma 6.10. *Let* $\mu_0 < 1 \wedge (\mathtt{g}^2/\mathtt{p})$. *With* $\mathtt{y}_0^2 = \mathtt{g}^2/\mu_0^2 - \mathtt{p}/\mu_0$, *it holds for any* $u > 0$

$$\begin{aligned} &I\!\!P\bigl(\|I\!\!B\boldsymbol{\xi}\|^2 > \mathtt{p} + u, \|I\!\!B^2\boldsymbol{\xi}\| \le \mathtt{y}_0\bigr) \\ &\quad \le 2\exp\bigl\{-0.5\mu_0(\mathtt{p}+u) - 0.5\log\det(I_p - \mu_0 I\!\!B^2)\bigr\}. \end{aligned} \tag{6.13}$$

In particular, if $I\!\!B^2$ *is diagonal, that is,* $I\!\!B^2 = \operatorname{diag}\bigl(a_1, \ldots, a_p\bigr)$, *then*

$$\begin{aligned} &I\!\!P\bigl(\|I\!\!B\boldsymbol{\xi}\|^2 > \mathtt{p} + u, \|I\!\!B^2\boldsymbol{\xi}\| \le \mathtt{y}_0\bigr) \\ &\quad \le 2\exp\Bigl\{-\frac{\mu_0 u}{2} - \frac{1}{2}\sum_{i=1}^{p}\bigl[\mu_0 a_i + \log\bigl(1-\mu_0 a_i\bigr)\bigr]\Bigr\}. \end{aligned} \tag{6.14}$$

Proof. The exponential Chebyshev inequality and (6.12) imply

$$\begin{aligned} &I\!\!P\bigl(\|I\!\!B\boldsymbol{\xi}\|^2 > \mathtt{p} + u, \|I\!\!B^2\boldsymbol{\xi}\| \le \mathtt{y}_0\bigr) \\ &\quad \le \exp\Bigl\{-\frac{\mu_0(\mathtt{p}+u)}{2}\Bigr\} I\!\!E\exp\Bigl(\frac{\mu_0\|I\!\!B\boldsymbol{\xi}\|^2}{2}\Bigr)\, 1\!\!1\Bigl(\|I\!\!B^2\boldsymbol{\xi}\|^2 \le \frac{\mathtt{g}^2}{\mu_0^2} - \frac{\mathtt{p}}{\mu_0}\Bigr) \\ &\quad \le 2\exp\bigl\{-0.5\mu_0(\mathtt{p}+u) - 0.5\log\det(I_p - \mu_0 I\!\!B^2)\bigr\}. \end{aligned}$$

Moreover, the standard change-of-basis arguments allow us to reduce the problem to the case of a diagonal matrix $I\!\!B^2 = \operatorname{diag}\bigl(a_1, \ldots, a_p\bigr)$ where $1 = a_1 \ge a_2 \ge \ldots \ge a_p > 0$. Note that $\mathtt{p} = a_1 + \ldots + a_p$. Then the claim (6.13) can be written in the form (6.14). □

Now we evaluate a large deviation probability that $\|I\!\!B\boldsymbol{\xi}\| > \mathtt{y}$ for a large $\mathtt{y}$. Note that the condition $\|I\!\!B^2\|_\infty \le 1$ implies $\|I\!\!B^2\boldsymbol{\xi}\| \le \|I\!\!B\boldsymbol{\xi}\|$. So, the bound (6.13) continues to hold when $\|I\!\!B^2\boldsymbol{\xi}\| \le \mathtt{y}_0$ is replaced by $\|I\!\!B\boldsymbol{\xi}\| \le \mathtt{y}_0$.

Lemma 6.11. *Let* $\mu_0 < 1$ *and* $\mu_0\mathtt{p} < \mathtt{g}^2$. *Define* $\mathtt{g}_0$ *by* $\mathtt{g}_0^2 = \mathtt{g}^2 - \mu_0\mathtt{p}$. *For any* $\mathtt{y} \ge \mathtt{y}_0 \stackrel{\text{def}}{=} \mathtt{g}_0/\mu_0$, *it holds*

$$\begin{aligned} I\!\!P\bigl(\|I\!\!B\boldsymbol{\xi}\| > \mathtt{y}\bigr) &\le 8.4\det\{I_p - (\mathtt{g}_0/\mathtt{y})I\!\!B^2\}^{-1/2}\exp\bigl(-\mathtt{g}_0\mathtt{y}/2\bigr). \\ &\le 8.4\exp\bigl(-\mathtt{x}_0 - \mathtt{g}_0(\mathtt{y}-\mathtt{y}_0)/2\bigr), \end{aligned} \tag{6.15}$$

where $\mathtt{x}_0$ *is defined by*

$$2\mathtt{x}_0 = \mathtt{g}_0\mathtt{y}_0 + \log\det\{I_p - (\mathtt{g}_0/\mathtt{y}_0)I\!\!B^2\}.$$

Proof. The slicing arguments of Lemma 6.6 apply here in the same manner. One has to replace $\|\boldsymbol{\xi}\|$ by $\|I\!\!B\boldsymbol{\xi}\|$ and $(1-\mu_1)^{-p/2}$ by $\det\{I_p - (\mathtt{g}_0/\mathtt{y})I\!\!B^2\}^{-1/2}$. We omit the details. In particular, with $\mathtt{y} = \mathtt{y}_0 = \mathtt{g}_0/\mu$, this yields

$$I\!\!P\bigl(\|I\!\!B\boldsymbol{\xi}\| > \mathtt{y}_0\bigr) \le 8.4\exp(-\mathtt{x}_0).$$

Moreover, for the function $f(\mathtt{y}) = \mathtt{g}_0\mathtt{y} + \log\det\{I_p - (\mathtt{g}_0/\mathtt{y})I\!B^2\}$, it holds $f'(\mathtt{y}) \geq \mathtt{g}_0$ and hence, $f(\mathtt{y}) \geq f(\mathtt{y}_0) + \mathtt{g}_0(\mathtt{y} - \mathtt{y}_0)$ for $\mathtt{y} > \mathtt{y}_0$. This implies (6.15). □

One important feature of the results of Lemma 6.10 and Lemma 6.11 is that the value $\mu_0 < 1 \wedge (\mathtt{g}^2/\mathtt{p})$ can be selected arbitrarily. In particular, for $\mathtt{y} \geq \mathtt{y}_c$, Lemma 6.11 with $\mu_0 = \mu_c$ yields the large deviation probability $I\!P\bigl(\|I\!B\boldsymbol{\xi}\| > \mathtt{y}\bigr)$. For bounding the probability $I\!P\bigl(\|I\!B\boldsymbol{\xi}\|^2 > \mathtt{p} + u,\ \|I\!B\boldsymbol{\xi}\| \leq \mathtt{y}_c\bigr)$, we use the inequality $\log(1-t) \geq -t - t^2$ for $t \leq 2/3$. It implies for $\mu \leq 2/3$ that

$$\begin{aligned} &-\log I\!P\bigl(\|I\!B\boldsymbol{\xi}\|^2 > \mathtt{p} + u, \|I\!B\boldsymbol{\xi}\| \leq \mathtt{y}_c\bigr) \\ &\quad\geq \mu(\mathtt{p} + u) + \sum_{i=1}^{p} \log\bigl(1 - \mu a_i\bigr) \\ &\quad\geq \mu(\mathtt{p} + u) - \sum_{i=1}^{p} (\mu a_i + \mu^2 a_i^2) \geq \mu u - \mu^2\mathtt{v}^2/2. \end{aligned} \tag{6.16}$$

Now we distinguish between $\mu_c \geq 2/3$ and $\mu_c < 2/3$ starting with $\mu_c \geq 2/3$. The bound (6.16) with $\mu_0 = (u/\mathtt{v}^2) \wedge (2/3) \leq \mu_c$ and with $u = (2\mathtt{v}\mathtt{x}^{1/2}) \vee (6\mathtt{x})$ yields

$$I\!P\bigl(\|I\!B\boldsymbol{\xi}\|^2 > \mathtt{p} + u,\ \|I\!B\boldsymbol{\xi}\| \leq \mathtt{y}_c\bigr) \leq 2\exp(-\mathtt{x});$$

see the proof of Theorem 6.2 for the Gaussian case.

Now consider $\mu_c < 2/3$. For $\mathtt{x}^{1/2} \leq \mu_c\mathtt{v}/2$, use $u = 2\mathtt{v}\mathtt{x}^{1/2}$ and $\mu_0 = u/\mathtt{v}^2$. It holds $\mu_0 = u/\mathtt{v}^2 \leq \mu_c$ and $u^2/(4\mathtt{v}^2) = \mathtt{x}$ yielding the desired bound by (6.16). For $\mathtt{x}^{1/2} > \mu_c\mathtt{v}/2$, we select again $\mu_0 = \mu_c$. It holds with $u = 4\mu_c^{-1}\mathtt{x}$ that $\mu_c u/2 - \mu_c^2\mathtt{v}^2/4 \geq 2\mathtt{x} - \mathtt{x} = \mathtt{x}$. This completes the proof. □

Now we describe the value $\mathfrak{z}(\mathtt{x}, I\!B)$ ensuring a small value for the large deviation probability $I\!P\bigl(\|I\!B\boldsymbol{\xi}\|^2 > \mathfrak{z}(\mathtt{x}, I\!B)\bigr)$. For ease of formulation, we suppose that $\mathtt{g}^2 \geq 2\mathtt{p}$ yielding $\mu_c^{-1} \leq 3/2$. The other case can be easily adjusted.

Corollary 6.12. *Let $\boldsymbol{\xi}$ fulfill* (6.1) *with $\mathtt{g}^2 \geq 2\mathtt{p}$. Then it holds for $\mathtt{x} \leq \mathtt{x}_c$ with $\mathtt{x}_c$ from* (6.11)*:*

$$\begin{aligned} I\!P\bigl(\|I\!B\boldsymbol{\xi}\|^2 \geq \mathfrak{z}(\mathtt{x}, I\!B)\bigr) \ &\leq\ 2\mathrm{e}^{-\mathtt{x}} + 8.4\mathrm{e}^{-\mathtt{x}_c}, \\ \mathfrak{z}(\mathtt{x}, I\!B) \ &\stackrel{\mathrm{def}}{=}\ \begin{cases} \mathtt{p} + \sqrt{2\mathtt{x}\mathtt{v}}, & \mathtt{x} \leq \mathtt{v}/18, \\ \mathtt{p} + 6\mathtt{x} & \mathtt{v}/18 < \mathtt{x} \leq \mathtt{x}_c. \end{cases} \end{aligned} \tag{6.17}$$

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

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

6.4 Rescaling and regularity condition

The result of Theorem 6.8 can be extended to a more general situation when the condition (6.1) is fulfilled for a vector $\boldsymbol{\zeta}$ rescaled by a matrix V_0. More precisely, let the random p-vector $\boldsymbol{\zeta}$ fulfills for some $p \times p$ matrix V_0 the condition

$$\sup_{\boldsymbol{\gamma} \in I\!R^p} \log I\!E \exp\left\{ \lambda \frac{\boldsymbol{\gamma}^\top \boldsymbol{\zeta}}{\|V_0 \boldsymbol{\gamma}\|} \right\} \leq \nu_0^2 \lambda^2 / 2, \qquad |\lambda| \leq \mathtt{g}, \tag{6.18}$$

with some constants $\mathtt{g} > 0$, $\nu_0 \geq 1$. Again, a simple change of variables reduces the case of an arbitrary $\nu_0 \geq 1$ to $\nu_0 = 1$. Our aim is to bound the squared norm $\|D_0^{-1} \boldsymbol{\zeta}\|^2$ of a vector $D_0^{-1} \boldsymbol{\zeta}$ for another $p \times p$ positive symmetric matrix D_0^2. Note that condition (6.18) implies (6.1) for the rescaled vector $\boldsymbol{\xi} = V_0^{-1} \boldsymbol{\zeta}$. This leads to bounding the quadratic form $\|D_0^{-1} V_0 \boldsymbol{\xi}\|^2 = \|I\!B \boldsymbol{\xi}\|^2$ with $I\!B^2 = D_0^{-1} V_0^2 D_0^{-1}$. It obviously holds

$$\mathtt{p} = \operatorname{tr}(I\!B^2) = \operatorname{tr}(D_0^{-2} V_0^2).$$

Now we can apply the result of Corollary 6.12.

Corollary 6.13. *Let $\boldsymbol{\zeta}$ fulfill* (6.18) *with some V_0 and $\mathtt{g}$. Given D_0, define $I\!B^2 = D_0^{-1} V_0^2 D_0^{-1}$, and let $\mathtt{g}^2 \geq 2\mathtt{p}$. Then it holds for $\mathtt{x} \leq \mathtt{x}_c$ with $\mathtt{x}_c$ from* (6.11)*:*

$$I\!P\bigl(\|D_0^{-1} \boldsymbol{\zeta}\|^2 \geq \mathfrak{z}(\mathtt{x}, I\!B) \bigr) \leq 2\mathrm{e}^{-\mathtt{x}} + 8.4 \mathrm{e}^{-\mathtt{x}_c},$$

with $\mathfrak{z}(\mathtt{x}, I\!B)$ from (6.17)*. For $\mathtt{x} > \mathtt{x}_c$*

$$I\!P\bigl(\|D_0^{-1} \boldsymbol{\zeta}\|^2 \geq \mathfrak{z}_c(\mathtt{x}, I\!B) \bigr) \leq 8.4 \mathrm{e}^{-\mathtt{x}}, \qquad \mathfrak{z}_c(\mathtt{x}, I\!B) \stackrel{\text{def}}{=} \bigl| \mathtt{y}_c + 2(\mathtt{x} - \mathtt{x}_c)/\mathtt{g}_c \bigr|^2.$$

Finally we briefly discuss the *regular* case with $D_0 \geq \mathfrak{a} V_0$ for some $\mathfrak{a} > 0$. This implies $\|I\!B\|_\infty \leq \mathfrak{a}^{-1}$ and

$$\mathtt{v}^2 = 2 \operatorname{tr}(I\!B^4) \leq 2 \mathfrak{a}^{-2} \mathtt{p}.$$

This together with $\mathtt{g}^2 \geq 2\mathtt{p}$ yields

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

7 Some results for empirical processes

This chapter presents some general results of the theory of empirical processes. Under the global conditions from Section 2 one can apply the well developed chaining arguments in Orlic spaces; see e.g. van der Vaart and Wellner (1996), Chapter ???. We follow the more recent approach inspired by the notions of generic chaining and majorizing measures due to M. Talagrand. The chaining arguments are replaced by the *pilling* device; see e.g. Talagrand (1996, 2001, 2005). The results are close to that of Bednorz (2006). We state the results in a slightly different form and present an independent and self-containing proof.

The first result states a bound for local fluctuations of the process $\mathcal{U}(\upsilon)$ given on a metric space Υ. Then this result will be used for bounding the maximum of the negatively drifted process $\mathcal{U}(\upsilon) - \mathcal{U}(\upsilon_0) - \rho d^2(\upsilon, \upsilon_0)$ over a vicinity $\Upsilon_\circ(\mathtt{r})$ of the central point υ_0. The behavior of $\mathcal{U}(\upsilon)$ outside of the local central set $\Upsilon_\circ(\mathtt{r})$ is described using the *upper function* method. Namely, we construct a multiscale deterministic function $\mathfrak{u}(\mu, \upsilon)$ ensuring that with probability at least $1 - \mathrm{e}^{-\mathtt{x}}$ it holds $\mu\mathcal{U}(\upsilon) + \mathfrak{u}(\mu, \upsilon) \le \mathfrak{z}(\mathtt{x})$ for all $\upsilon \notin \Upsilon_\circ(\mathtt{r})$ and $\mu \in \mathbb{M}$, where $\mathfrak{z}(\mathtt{x})$ grows linearly in $\mathtt{x}$.

7.1 A bound for local fluctuations

An important step in the whole construction is an exponential bound on the maximum of a random process $\mathcal{U}(\upsilon)$ under the exponential moment conditions on its increments. Let $d(\upsilon, \upsilon')$ be a semi-distance on Υ. We suppose the following condition to hold:

$(\mathcal{E}d)$ *There exist* $\mathtt{g} > 0$, $\mathtt{r}_1 > 0$, $\nu_0 \ge 1$, *such that for any* $\lambda \le \mathtt{g}$ *and* $\upsilon, \upsilon' \in \Upsilon$ *with* $d(\upsilon, \upsilon') \le \mathtt{r}_1$

$$\log I\!E \exp\left\{ \lambda \frac{\mathcal{U}(\upsilon) - \mathcal{U}(\upsilon')}{d(\upsilon, \upsilon')} \right\} \le \nu_0^2 \lambda^2 / 2.$$

Formulation of the result involves a sigma-finite measure π on the space Υ which is often called the *majorizing measure* and used in the *generic chaining* device; see Talagrand (2005). A typical example of choosing π is the Lebesgue measure on $I\!R^p$. Let Υ° be a subset of Υ, a sequence $\mathtt{r}_k$ be fixed with $\mathtt{r}_0 = \operatorname{diam}(\Upsilon^\circ)$ and $\mathtt{r}_k = \mathtt{r}_0 2^{-k}$. Let also $\mathcal{B}_k(\upsilon) \stackrel{\mathrm{def}}{=} \{\upsilon' \in \Upsilon^\circ : d(\upsilon, \upsilon') \le \mathtt{r}_k\}$ be the d-ball centered at υ of radius $\mathtt{r}_k$ and $\pi_k(\upsilon)$ denote its π-measure:

$$\pi_k(\upsilon) \stackrel{\mathrm{def}}{=} \int_{\mathcal{B}_k(\upsilon)} \pi(d\upsilon') = \int_{\Upsilon^\circ} I\!\!1\big(d(\upsilon, \upsilon') \le \mathtt{r}_k\big) \pi(d\upsilon').$$

Denote also

$$M_k \stackrel{\text{def}}{=} \max_{\boldsymbol{\upsilon} \in \Upsilon^\circ} \frac{\pi(\Upsilon^\circ)}{\pi_k(\boldsymbol{\upsilon})} \qquad k \geq 1. \tag{7.1}$$

Finally set $c_1 = 1/3$, $c_k = 2^{-k+2}/3$ for $k \geq 2$, and define the value $\mathbb{Q}(\Upsilon^\circ)$ by

$$\mathbb{Q}(\Upsilon^\circ) \stackrel{\text{def}}{=} \sum_{k=1}^{\infty} c_k \log(2M_k) = \frac{1}{3}\log(2M_1) + \frac{4}{3}\sum_{k=2}^{\infty} 2^{-k}\log(2M_k).$$

Theorem 7.1. *Suppose* $(\mathcal{E}d)$. *If* Υ° *is a central set with the center* $\boldsymbol{\upsilon}^\circ$ *and the radius* $\mathtt{r}_1$, *i.e.* $d(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^\circ) \leq \mathtt{r}_1$ *for all* $\boldsymbol{\upsilon} \in \Upsilon^\circ$, *then for* $\lambda \leq \mathtt{g}_0 \stackrel{\text{def}}{=} \nu_0 \mathtt{g}$

$$\log I\!\!E \exp\Big\{ \frac{\lambda}{3\nu_0 \mathtt{r}_1} \sup_{\boldsymbol{\upsilon} \in \Upsilon^\circ} \big| \mathcal{U}(\boldsymbol{\upsilon}) - \mathcal{U}(\boldsymbol{\upsilon}^\circ) \big| \Big\} \leq \lambda^2/2 + \mathbb{Q}(\Upsilon^\circ). \tag{7.2}$$

Proof. A simple change $\mathcal{U}(\cdot)$ with $\nu_0^{-1}\mathcal{U}(\cdot)$ and $\mathtt{g}$ with $\mathtt{g}_0 = \nu_0 \mathtt{g}$ allows for reducing the result to the case with $\nu_0 = 1$ which we assume below. Consider for $k \geq 1$ the smoothing operator $\mathbb{S}_k$ defined as

$$\mathbb{S}_k f(\boldsymbol{\upsilon}^\circ) = \frac{1}{\pi_k(\boldsymbol{\upsilon}^\circ)} \int_{\mathcal{B}_k(\boldsymbol{\upsilon}^\circ)} f(\boldsymbol{\upsilon}) \pi(d\boldsymbol{\upsilon}).$$

Further, define

$$\mathbb{S}_0 \mathcal{U}(\boldsymbol{\upsilon}) \equiv \mathcal{U}(\boldsymbol{\upsilon}^\circ)$$

so that $\mathbb{S}_0\mathcal{U}$ is a constant function and the same holds for $\mathbb{S}_k\mathbb{S}_{k-1}\ldots\mathbb{S}_0\mathcal{U}$ with any $k \geq 1$. If $f(\cdot) \leq g(\cdot)$ for two non-negative functions f and g, then $\mathbb{S}_k f(\cdot) \leq \mathbb{S}_k g(\cdot)$. Separability of the process $\mathcal{U}$ implies that $\lim_k \mathbb{S}_k \mathcal{U}(\boldsymbol{\upsilon}) = \mathcal{U}(\boldsymbol{\upsilon})$. We conclude that for each $\boldsymbol{\upsilon} \in \Upsilon^\circ$

$$\begin{aligned} \big| \mathcal{U}(\boldsymbol{\upsilon}) - \mathcal{U}(\boldsymbol{\upsilon}^\circ) \big| &= \lim_{k \to \infty} \big| \mathbb{S}_k \mathcal{U}(\boldsymbol{\upsilon}) - \mathbb{S}_k \ldots \mathbb{S}_0 \mathcal{U}(\boldsymbol{\upsilon}) \big| \\ &\leq \lim_{k \to \infty} \sum_{i=1}^{k} \big| \mathbb{S}_k \ldots \mathbb{S}_i (I - \mathbb{S}_{i-1}) \mathcal{U}(\boldsymbol{\upsilon}) \big| \leq \sum_{i=1}^{\infty} \xi_k^* . \end{aligned}$$

Here $\xi_k^* \stackrel{\text{def}}{=} \sup_{\boldsymbol{\upsilon} \in \Upsilon^\circ} \xi_k(\boldsymbol{\upsilon})$ for $k \geq 1$ with

$$\xi_1(\boldsymbol{\upsilon}) \equiv |\mathbb{S}_1 \mathcal{U}(\boldsymbol{\upsilon}) - \mathcal{U}(\boldsymbol{\upsilon}^\circ)|, \qquad \xi_k^* \stackrel{\text{def}}{=} |\mathbb{S}_k (I - \mathbb{S}_{k-1}) \mathcal{U}(\boldsymbol{\upsilon})|, \qquad k \geq 2$$

For a fixed point $\boldsymbol{\upsilon}^\sharp$, it holds

$$\xi_k(\boldsymbol{\upsilon}^\sharp) \leq \frac{1}{\pi_k(\boldsymbol{\upsilon}^\sharp)} \int_{\mathcal{B}_k(\boldsymbol{\upsilon}^\sharp)} \frac{1}{\pi_{k-1}(\boldsymbol{\upsilon})} \int_{\mathcal{B}_{k-1}(\boldsymbol{\upsilon})} \big| \mathcal{U}(\boldsymbol{\upsilon}) - \mathcal{U}(\boldsymbol{\upsilon}') \big| \pi(d\boldsymbol{\upsilon}') \pi(d\boldsymbol{\upsilon}).$$

For each $\upsilon' \in \mathcal{B}_{k-1}(\upsilon)$, it holds $d(\upsilon,\upsilon') \le \mathtt{r}_{k-1} = 2\mathtt{r}_k$ and

$$\bigl|\mathcal{U}(\upsilon) - \mathcal{U}(\upsilon')\bigr| \le \mathtt{r}_{k-1} \frac{\bigl|\mathcal{U}(\upsilon) - \mathcal{U}(\upsilon')\bigr|}{d(\upsilon,\upsilon')}.$$

This implies for each $\upsilon^\sharp \in \Upsilon^\circ$ and $k \ge 2$ by the Jensen inequality and (7.1)

$$\begin{aligned}
\exp\Bigl\{\frac{\lambda}{\mathtt{r}_{k-1}}\xi_k(\upsilon^\sharp)\Bigr\} &\le \int_{\mathcal{B}_k(\upsilon^\sharp)} \biggl(\int_{\mathcal{B}_{k-1}(\upsilon)} \exp\frac{\lambda\bigl|\mathcal{U}(\upsilon) - \mathcal{U}(\upsilon')\bigr|}{d(\upsilon,\upsilon')} \frac{\pi(d\upsilon')}{\pi_{k-1}(\upsilon)}\biggr)\frac{\pi(d\upsilon)}{\pi_k(\upsilon^\sharp)} \\
&\le M_k \int_{\Upsilon^\circ} \biggl(\int_{\mathcal{B}_{k-1}(\upsilon)} \exp\frac{\lambda\bigl|\mathcal{U}(\upsilon) - \mathcal{U}(\upsilon')\bigr|}{d(\upsilon,\upsilon')} \frac{\pi(d\upsilon')}{\pi_{k-1}(\upsilon)}\biggr)\frac{\pi(d\upsilon)}{\pi(\Upsilon^\circ)}.
\end{aligned}$$

As the right hand-side does not depend on $\upsilon^\sharp$, this yields for $\xi_k^* \stackrel{\mathrm{def}}{=} \sup_{\upsilon\in\Upsilon^\circ}\xi_k(\upsilon)$ by condition $(\mathcal{E}d)$ in view of $e^{|x|} \le e^x + e^{-x}$

$$\begin{aligned}
\mathbb{E}\exp\Bigl(\frac{\lambda}{\mathtt{r}_{k-1}}\xi_k^*\Bigr) &\le M_k \int_{\Upsilon^\circ} \biggl(\int_{\mathcal{B}_{k-1}(\upsilon)} \mathbb{E}\exp\frac{\lambda\bigl|\mathcal{U}(\upsilon) - \mathcal{U}(\upsilon')\bigr|}{d(\upsilon,\upsilon')} \frac{\pi(d\upsilon')}{\pi_{k-1}(\upsilon)}\biggr)\frac{\pi(d\upsilon)}{\pi(\Upsilon^\circ)} \\
&\le 2M_k \exp(\lambda^2/2) \int_{\Upsilon^\circ} \biggl(\int_{\mathcal{B}_{k-1}(\upsilon)} \frac{\pi(d\upsilon')}{\pi_{k-1}(\upsilon)}\biggr)\frac{\pi(d\upsilon)}{\pi(\Upsilon^\circ)} \\
&= 2M_k \exp(\lambda^2/2).
\end{aligned}$$

Further, the use of $d(\upsilon,\upsilon^\circ) \le \mathtt{r}_1$ for all $\upsilon \in \Upsilon^\circ$ yields by $(\mathcal{E}d)$

$$\mathbb{E}\exp\Bigl\{\frac{\lambda}{\mathtt{r}_1}|\mathcal{U}(\upsilon) - \mathcal{U}(\upsilon^\circ)|\Bigr\} \le 2\exp\bigl(\lambda^2/2\bigr) \tag{7.3}$$

and thus

$$\begin{aligned}
\mathbb{E}\exp\Bigl\{\frac{\lambda}{\mathtt{r}_1}|\mathbb{S}_1\mathcal{U}(\upsilon) - \mathcal{U}(\upsilon^\circ)|\Bigr\} &\le \frac{1}{\pi_1(\upsilon)} \int_{\mathcal{B}_1(\upsilon)} \mathbb{E}\exp\Bigl\{\frac{\lambda}{\mathtt{r}_1}|\mathcal{U}(\upsilon') - \mathcal{U}(\upsilon^\circ)|\Bigr\}\pi(d\upsilon') \\
&\le \frac{M_1}{\pi(\Upsilon^\circ)} \int_{\Upsilon^\circ} \mathbb{E}\exp\Bigl\{\frac{\lambda}{\mathtt{r}_1}|\mathcal{U}(\upsilon') - \mathcal{U}(\upsilon^\circ)|\Bigr\}\pi(d\upsilon').
\end{aligned}$$

This implies by (7.3) for $\xi_1^* \equiv \sup_{\upsilon\in\Upsilon^\circ}|\mathbb{S}_1\mathcal{U}(\upsilon) - \mathcal{U}(\upsilon^\circ)|$

$$\mathbb{E}\exp\Bigl(\frac{\lambda}{\mathtt{r}_1}\xi_1^*\Bigr) \le 2M_1\exp\bigl(\lambda^2/2\bigr).$$

Denote $c_1 = 1/3$ and $c_k = \mathtt{r}_{k-1}/(3\mathtt{r}_1) = 2^{-k+2}/3$ for $k \ge 2$. Then $\sum_{k=1}^\infty c_k = 1$ and it holds by the Hölder inequality; see Lemma 7.11 below:

$$\begin{aligned}
\log\mathbb{E}\exp\biggl(\frac{\lambda}{3\mathtt{r}_1}\sum_{k=1}^\infty \xi_k^*\biggr) &\le c_1 \log\mathbb{E}\exp\biggl(\frac{\lambda}{\mathtt{r}_1}\xi_k^*\biggr) + \sum_{k=2}^\infty c_k \log\mathbb{E}\exp\biggl(\frac{\lambda}{\mathtt{r}_{k-1}}\xi_k^*\biggr) \\
&\le \lambda^2/2 + c_1\log(2M_1) + \sum_{k=2}^\infty c_k\log(2M_k) \\
&< \lambda^2/2 + \mathbb{Q}(\Upsilon^\circ).
\end{aligned}$$

This implies the result. □

7.2 A local central bound

Due to the result of Theorem 7.1, the bound for the maximum of $\mathcal{U}(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}_0)$ over $\boldsymbol{\upsilon} \in \mathcal{B}_{\mathtt{r}}(\boldsymbol{\upsilon}_0)$ grows quadratically in $\mathtt{r}$. So, its applications to situations with $\mathtt{r}^2 \gg \mathbb{Q}(\Upsilon^\circ)$ are limited. The next result shows that introducing a negative quadratic drift helps to state a uniform in $\mathtt{r}$ local probability bound. Namely, the bound for the process $\mathcal{U}(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}_0) - \rho d^2(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}_0)/2$ with some positive ρ over a ball $\mathcal{B}_{\mathtt{r}}(\boldsymbol{\upsilon}_0)$ around the point $\boldsymbol{\upsilon}_0$ only depends on the drift coefficient ρ but not on $\mathtt{r}$. Here the generic chaining arguments are accomplished with the *slicing* technique. The idea is for a given $\mathtt{r}^* > 1$ to split the ball $\mathcal{B}_{\mathtt{r}^*}(\boldsymbol{\upsilon}_0)$ into the slices $\mathcal{B}_{\mathtt{r}+1}(\boldsymbol{\upsilon}_0) \setminus \mathcal{B}_{\mathtt{r}}(\boldsymbol{\upsilon}_0)$ and to apply Theorem 7.1 to each slice separately with a proper choice of the parameter λ.

Theorem 7.2. *Let $\mathtt{r}^*$ be such that $(\mathcal{E}d)$ holds on $\mathcal{B}_{\mathtt{r}^*}(\boldsymbol{\upsilon}_0)$. Let also $\mathbb{Q}(\Upsilon^\circ) \leq \mathbb{Q}$ for $\Upsilon^\circ = \mathcal{B}_{\mathtt{r}}(\boldsymbol{\upsilon}_0)$ with $\mathtt{r} \leq \mathtt{r}^*$. If $\rho > 0$ and $\mathfrak{z}$ are fixed to ensure $\sqrt{\rho \mathfrak{z}} \leq \mathtt{g}_0 = \nu_0 \mathtt{g}$ and $\rho(\mathfrak{z} - 1) \geq 2$, then it holds*

$$\begin{aligned} \log I\!P\biggl(\sup_{\boldsymbol{\upsilon} \in \mathcal{B}_{\mathtt{r}^*}(\boldsymbol{\upsilon}_0)} \biggl\{ \frac{1}{3\nu_0} \mathcal{U}(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}_0) - \frac{\rho}{2} d^2(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}_0) \biggr\} > \mathfrak{z} \biggr) \\ \leq -\rho(\mathfrak{z} - 1) + \log(4\mathfrak{z}) + \mathbb{Q}. \end{aligned} \tag{7.4}$$

Moreover, if $\sqrt{\rho \mathfrak{z}} > \mathtt{g}_0$, then

$$\begin{aligned} \log I\!P\biggl(\sup_{\boldsymbol{\upsilon} \in \mathcal{B}_{\mathtt{r}^*}(\boldsymbol{\upsilon}_0)} \biggl\{ \frac{1}{3\nu_0} \mathcal{U}(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}_0) - \frac{\rho}{2} d^2(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}_0) \biggr\} > \mathfrak{z} \biggr) \\ \leq -\mathtt{g}_0 \sqrt{\rho(\mathfrak{z} - 1)} + \mathtt{g}_0^2/2 + \log(4\mathfrak{z}) + \mathbb{Q}. \end{aligned} \tag{7.5}$$

Remark 7.1. Formally the bound applies even with $\mathtt{r}^* = \infty$ provided that $(\mathcal{E}d)$ is fulfilled on the whole set Υ°.

Proof. Denote

$$\mathtt{u}(\mathtt{r}) \stackrel{\text{def}}{=} \frac{1}{3\nu_0 \mathtt{r}} \sup_{\boldsymbol{\upsilon} \in \mathcal{B}_{\mathtt{r}}(\boldsymbol{\upsilon}_0)} \bigl\{ \mathcal{U}(\boldsymbol{\upsilon}) - \mathcal{U}(\boldsymbol{\upsilon}_0) \bigr\}.$$

Then we have to bound the probability

$$I\!P\bigl(\sup_{\mathtt{r} \leq \mathtt{r}^*} \bigl\{ \mathtt{r}\, \mathtt{u}(\mathtt{r}) - \rho \mathtt{r}^2/2 \bigr\} > \mathfrak{z} \bigr).$$

For each $\mathtt{r} \leq \mathtt{r}^*$ and $\lambda \leq \mathtt{g}_0$, it follows from (7.2) that

$$\log I\!E \exp\bigl\{ \lambda \mathtt{u}(\mathtt{r}) \bigr\} \leq \lambda^2/2 + \mathbb{Q}.$$

The choice $\lambda = \sqrt{\rho \mathfrak{z}}$ is admissible in view of $\sqrt{\rho \mathfrak{z}} \le \mathtt{g}_0$. This implies by the exponential Chebyshev inequality

$$\begin{aligned} \log I\!P\bigl(\mathtt{r}\,\mathtt{u}(\mathtt{r}) - \rho \mathtt{r}^2/2 \ge \mathfrak{z}\bigr) &\le -\lambda(\mathfrak{z}/\mathtt{r} + \rho\mathtt{r}/2) + \lambda^2/2 + \mathbb{Q} \\ &= -\rho\mathfrak{z}(x + x^{-1} - 1) + \mathbb{Q}, \end{aligned} \tag{7.6}$$

where $u = \sqrt{\rho/(2\mathfrak{z})}\,\mathtt{r}$. We now apply the slicing arguments w.r.t. $t = \rho\mathtt{r}^2/2 = \mathfrak{z}x^2$. By definition, $\mathtt{r}\mathtt{u}(\mathtt{r})$ increases in $\mathtt{r}$. We use that for any growing function $f(\cdot)$ and any $t \ge 0$, it holds

$$f(t) - t \le \int_t^{t+1} \bigl\{ f(s) - s + 1 \bigr\} ds$$

Therefore, for any $t > 0$, it holds by (7.6) in view of $dt = 2\mathfrak{z}\,x\,dx$

$$\begin{aligned} I\!P\biggl(\sup_{\mathtt{r} \le \mathtt{r}^*} \bigl\{ \mathtt{r}\,\mathtt{u}(\mathtt{r}) - \rho\mathtt{r}^2/2 \bigr\} > \mathfrak{z} \biggr) &\le \int_0^{t^*+1} I\!P\bigl\{ \mathtt{r}\,\mathtt{u}(\mathtt{r}) - t \ge \mathfrak{z} - 1 \bigr\} dt \\ &\le 2\mathfrak{z} \int_0^{t^*+1} \exp\bigl\{ -\rho(\mathfrak{z}-1)(x + x^{-1} - 1) + \mathbb{Q} \bigr\} x\,dx \\ &\le 2\mathfrak{z} e^{-b+\mathbb{Q}} \int_0^{\infty} \exp\bigl\{ -b(x + x^{-1} - 2) \bigr\}\, x\,dx \end{aligned}$$

with $b = \rho(\mathfrak{z} - 1)$ and $t^* = \rho \mathtt{r}^{*2}/2$. This implies for $b \ge 2$

$$\begin{aligned} I\!P\biggl(\sup_{\mathtt{r} \le \mathtt{r}^*} \bigl\{ \mathtt{r}\,\mathtt{u}(\mathtt{r}) - \rho\mathtt{r}^2/2 \bigr\} > \mathfrak{z} \biggr) &\le 2\mathfrak{z} e^{-b+\mathbb{Q}} \int_0^{\infty} \exp\bigl\{ -2(x + x^{-1} - 2) \bigr\} x\,dx \\ &\le 4\mathfrak{z} \exp\{ -\rho(\mathfrak{z} - 1) + \mathbb{Q} \} \end{aligned}$$

and (7.4) follows.

If $\sqrt{\rho\mathfrak{z}} > \mathtt{g}_0$, then select $\lambda = \mathtt{g}_0$. For $\mathtt{r} \le \mathtt{r}^*$

$$\begin{aligned} \log I\!P\bigl\{ \mathtt{r}\,\mathtt{u}(\mathtt{r}) - \rho\mathtt{r}^2/2 \ge \mathfrak{z} \bigr\} &= \log I\!P\bigl\{ \mathtt{u}(\mathtt{r}) > \mathfrak{z}/\mathtt{r} + \rho\mathtt{r}/2 \bigr\} \\ &\le -\lambda(\mathfrak{z}/\mathtt{r} + \rho\mathtt{r}/2) + \lambda^2/2 + \mathbb{Q} \\ &\le -\lambda\sqrt{\rho\mathfrak{z}}(x + x^{-1} - 2)/2 - \lambda\sqrt{\rho\mathfrak{z}} + \lambda^2/2 + \mathbb{Q}, \end{aligned}$$

where $u = \sqrt{\rho/\mathfrak{z}}\,\mathtt{r}$. This allows to bound in the same way as above

$$I\!P\biggl(\sup_{\mathtt{r} \le \mathtt{r}^*} \bigl\{ \mathtt{r}\,\mathtt{u}(\mathtt{r}) - \rho\mathtt{r}^2/2 \bigr\} > \mathfrak{z} \biggr) \le 4\mathfrak{z} \exp\bigl(-\lambda\sqrt{\rho(\mathfrak{z}-1)} + \lambda^2/2 + \mathbb{Q} \bigr)$$

yielding (7.5). □

This result can be used for describing the concentration bound for the maximum of $(3\nu_0)^{-1}\mathcal{U}(\boldsymbol{\upsilon},\boldsymbol{\upsilon}_0) - \rho d^2(\boldsymbol{\upsilon},\boldsymbol{\upsilon}_0)/2$. Namely, it suffices to find $\mathfrak{z}$ ensuring the prescribed deviation probability. We state the result for a special case with $\rho = 1$ and $\mathtt{g}_0 \geq 3$ which simplifies the notation.

Corollary 7.3. *Under the conditions of Theorem 7.2, it holds for* $\mathtt{x} \geq 0$ *with* $\mathtt{x}+\mathbb{Q} \geq 4$*:*

$$\mathbb{P}\Big(\sup_{\boldsymbol{\upsilon} \in \mathcal{B}_{\mathtt{r}^*}(\boldsymbol{\upsilon}_0)} \Big\{ \frac{1}{3\nu_0}\mathcal{U}(\boldsymbol{\upsilon},\boldsymbol{\upsilon}_0) - \frac{1}{2}d^2(\boldsymbol{\upsilon},\boldsymbol{\upsilon}_0) \Big\} > \mathfrak{z}_0(\mathtt{x},\mathbb{Q}) \Big) \leq \exp\bigl(-\mathtt{x}\bigr),$$

where with $\mathtt{g}_0 = \nu_0 \mathtt{g} \geq 2$

$$\mathfrak{z}_0(\mathtt{x},\mathbb{Q}) \stackrel{\text{def}}{=} \begin{cases} \bigl(1+\sqrt{\mathtt{x}+\mathbb{Q}}\bigr)^2 & \text{if } 1+\sqrt{\mathtt{x}+\mathbb{Q}} \leq \mathtt{g}_0, \\ 1 + \bigl(1+2\mathtt{g}_0^{-1}\bigr)^2 \Bigl(\mathtt{g}_0^{-1}(\mathtt{x}+\mathbb{Q}) + \mathtt{g}_0/2\Bigr)^2 & \text{otherwise.} \end{cases} \tag{7.7}$$

Proof. In view of (7.4), it suffices to check that $\mathfrak{z} = \bigl(1+\sqrt{\mathtt{x}+\mathbb{Q}}\bigr)^2$ ensures

$$\mathfrak{z} - 1 - \log(4\mathfrak{z}) - \mathbb{Q} \geq \mathtt{x}.$$

This follows from the inequality

$$(1+\mathtt{y})^2 - 1 - 2\log(2+2\mathtt{y}) \geq \mathtt{y}^2$$

with $\mathtt{y} = \sqrt{\mathtt{x}+\mathbb{Q}} \geq 2$. Similarly $\mathfrak{z} = 1 + \bigl(1+2\mathtt{g}_0^{-1}\bigr)^2\mathtt{y}^2$ for $\mathtt{y} = \mathtt{g}_0^{-1}(\mathtt{x}+\mathbb{Q}) + \mathtt{g}_0/2$ ensures

$$\mathtt{g}_0\sqrt{\mathfrak{z}-1} - \mathtt{g}_0^2/2 - \log(4\mathfrak{z}) - \mathbb{Q} \geq \mathtt{x}$$

in view of $\sqrt{\mathfrak{z}-1} \geq \mathtt{y}(1+2\mathtt{g}_0^{-1})$, $4\mathfrak{z} \leq (1+2\mathtt{y}+4\mathtt{g}_0^{-1}\mathtt{y})^2$, and

$$\mathtt{y} - \log(1+2\mathtt{y}+4\mathtt{g}_0^{-1}\mathtt{y}) \geq 0$$

for $\mathtt{y} \geq 2$ and $\mathtt{g}_0 \geq 3$. □

If $\mathtt{g} \gg \sqrt{\mathbb{Q}}$ and $\mathtt{x}$ is not too big then $\mathfrak{z}_0(\mathtt{x},\mathbb{Q})$ is of order $\mathtt{x}+\mathbb{Q}$. So, the main message of this result is that with a high probability the maximum of $(3\nu_0)^{-1}\mathcal{U}(\boldsymbol{\upsilon},\boldsymbol{\upsilon}_0) - d^2(\boldsymbol{\upsilon},\boldsymbol{\upsilon}_0)/2$ does not significantly exceed the level $\mathbb{Q}$.

7.3 A global upper function and concentration sets

The result of the previous section can be explained as a local upper function for the process $\mathcal{U}(\cdot)$. Indeed, in a vicinity $\mathcal{B}_{\mathtt{r}^*}(\boldsymbol{\upsilon}_0)$ of the central point $\boldsymbol{\upsilon}_0$, it holds $(3\nu_0)^{-1}\mathcal{U}(\boldsymbol{\upsilon}) \leq d^2(\boldsymbol{\upsilon},\boldsymbol{\upsilon}_0)/2 + \mathfrak{z}$ with a probability exponentially small in $\mathfrak{z}$. However, an extension of

this result on the whole set Υ is only possible under some quite restrictive conditions. This section presents one possible construction of an upper function for the process $\mathcal{U}(\cdot)$ on the complement of the local set $\Upsilon_\circ(\mathtt{r}^*)$. For simplifying the notations assume that $\mathcal{U}(\upsilon_0) \equiv 0$. Then $\mathcal{U}(\upsilon, \upsilon_0) = \mathcal{U}(\upsilon)$. We say that $\mathfrak{u}(\mu, \upsilon)$ is a *multiscale upper function* for $\mu\mathcal{U}(\cdot)$ on a subset Υ° of Υ if

$$\mathbb{P}\Big(\sup_{\mu\in\mathbb{M}} \sup_{\upsilon\in\Upsilon^\circ} \big\{\mu\mathcal{U}(\upsilon) - \mathfrak{u}(\mu,\upsilon)\big\} \geq \mathfrak{z}(\mathtt{x})\Big) \leq \mathrm{e}^{-\mathtt{x}},$$

for some fixed function $\mathfrak{z}(\mathtt{x})$. An upper function can be used for describing the concentration sets of the point of maximum $\widetilde{\upsilon} = \operatorname{argmax}_{\upsilon\in\Upsilon^\circ} \mathcal{U}(\upsilon)$; see Section 7.3.1 below.

For constructing such an upper function, the following condition is used which extends condition $(\mathcal{E}d)$:

$(\mathcal{E})$ *For any $\upsilon \in \Upsilon$ there exists $\mu \in \mathbb{M}$ such that*

$$\mathfrak{N}(\mu, \upsilon) \stackrel{\mathrm{def}}{=} \log \mathbb{E} \exp\big\{\mu\mathcal{U}(\upsilon)\big\} < \infty. \tag{7.8}$$

This condition can be used for building a simple pointwise upper function for $\mu\mathcal{U}(\upsilon)$. Indeed, (7.8) implies

$$\mathbb{E} \exp\big\{\mu\mathcal{U}(\upsilon) - \mathfrak{N}(\mu,\upsilon)\big\} = 1. \tag{7.9}$$

The next step is in extending this pointwise result to a uniform one. The standard approach is based on the notion of a ϵ-net which is a discrete set Υ°_ϵ providing that for any point $\upsilon \in \Upsilon$, there exists a point $\upsilon^\circ \in \Upsilon^\circ_\epsilon$ with $d(\upsilon, \upsilon^\circ) \leq \epsilon$. The upper function is first constructed on this discrete set Υ°_ϵ using (7.9) by increasing the pointwise bound with $\log N_\epsilon$, where the covering number N_ϵ is the cardinality of Υ°_ϵ. Then it is extended on the whole set Υ° using stochastic continuity of the process $\mathcal{U}(\cdot)$.

We apply a slightly different construction usually called *pilling*. Let the value $\mathtt{r}_1$ be fixed. Let also a measure π on Υ be fixed. By $\mathcal{B}_\mu(\upsilon)$ we denote the ball of radius $\mathtt{r}_1/\mu$ at $\upsilon \in \Upsilon$, while $\pi_\mu(\upsilon)$ denotes its π-measure. In our results the value $1/\pi_\mu(\upsilon)$ replaces the covering number N_ϵ with $\epsilon = \mathtt{r}_1/\mu$. Also define the constant ν_1 describing the local variability of $\pi_1(\cdot)$:

$$\nu_1 \stackrel{\mathrm{def}}{=} \sup_{\mu\in\mathbb{M}} \sup_{\upsilon^\circ\in\Upsilon^\circ} \sup_{\upsilon\in\mathcal{B}_\mu(\upsilon^\circ)} \frac{\pi_\mu(\upsilon)}{\pi_\mu(\upsilon^\circ)}. \tag{7.10}$$

For any fixed point υ°, the local maximum of the process $\mu\mathcal{U}$ over the ball $\mathcal{B}_\mu(\upsilon^\circ)$ can be bounded by combining the pointwise result (7.9) and the result of Theorem 7.1 for local fluctuations of the process $\mu\mathcal{U}$ within $\mathcal{B}_\mu(\upsilon^\circ)$. To get a global bound over Υ°, we introduce the so called *penalty* function $\mathfrak{t}_\mu(\upsilon)$ which accounts for the size of the set

Υ°. In the next result this function is allowed to be μ-dependent. However, in typical situations, one can apply a universal function $\mathfrak{t}(\boldsymbol{\upsilon})$; cf. the smooth case in Section 7.4.

Remind the definition of the smoothing operator $\mathbb{S}_\mu$:

$$\mathbb{S}_\mu f(\boldsymbol{\upsilon}^\circ) = \frac{1}{\pi_\mu(\boldsymbol{\upsilon}^\circ)} \int_{\mathcal{B}_\mu(\boldsymbol{\upsilon}^\circ)} f(\boldsymbol{\upsilon}) \pi(d\boldsymbol{\upsilon}).$$

The next result suggests a construction of the upper function $\mathfrak{u}(\mu, \boldsymbol{\upsilon})$. The construction involves a constant s which can be selected as the smallest value ensuring the bound $3\nu_0 \mathtt{r}_1/s \le \mathtt{g} \wedge \sqrt{2\mathbb{Q}}$.

Theorem 7.4. *Let the process $\mathcal{U}(\cdot)$ fulfill $(\mathcal{E})$ and $(\mathcal{E}d)$ for all $\boldsymbol{\upsilon}, \boldsymbol{\upsilon}' \in \Upsilon^\circ$ with $d(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}') \le \mathtt{r}_1$. If s is such that $3\nu_0 \mathtt{r}_1/s \le \mathtt{g} \wedge \sqrt{2\mathbb{Q}}$, then it holds for any $\mathtt{x} > 0$*

$$I\!P\Big(\sup_{\mu \in \mathbb{M}} \sup_{\boldsymbol{\upsilon} \in \Upsilon^\circ} \big\{ \mu \mathcal{U}(\boldsymbol{\upsilon}) - \mathbb{S}_\mu \mathfrak{N}(\mu, \boldsymbol{\upsilon}) - (1+s) \mathbb{S}_\mu \mathfrak{t}_\mu(\boldsymbol{\upsilon}) \big\} \ge \mathfrak{z}_1(\mathtt{x}) \Big) \le 2\mathrm{e}^{-\mathtt{x}}, \quad (7.11)$$

where $\mathfrak{z}_1(\mathtt{x})$ is a linear function in $\mathtt{x}$:

$$\mathfrak{z}_1(\mathtt{x}) \stackrel{\text{def}}{=} (1+s)\big\{ \mathtt{x} + \log\big(\nu_1 \mathcal{T}\big) \big\} + 2s\mathbb{Q},$$

and

$$\mathcal{T} \stackrel{\text{def}}{=} \sum_{\mu \in \mathbb{M}} \int_{\Upsilon^\circ} \exp\big\{ -\mathfrak{t}_\mu(\boldsymbol{\upsilon}) \big\} \frac{\pi(d\boldsymbol{\upsilon})}{\pi_\mu(\boldsymbol{\upsilon})}. \quad (7.12)$$

Proof. We bound $\mu \mathcal{U}(\boldsymbol{\upsilon})$ in two steps: first we evaluate $\mathbb{S}_\mu \big[\mu \mathcal{U}(\boldsymbol{\upsilon}^\circ) - \mathfrak{N}(\mu, \boldsymbol{\upsilon}^\circ) \big]$ and then $\mu \big[\mathcal{U}(\boldsymbol{\upsilon}^\circ) - \mathbb{S}_\mu \mathcal{U}(\boldsymbol{\upsilon}^\circ) \big]$. Convexity of the exp-function implies by the Jensen inequality

$$\begin{aligned}
&\exp\big\{ \mathbb{S}_\mu \big[\mu \mathcal{U}(\boldsymbol{\upsilon}^\circ) - \mathfrak{N}(\mu, \boldsymbol{\upsilon}^\circ) - \mathfrak{t}_\mu(\boldsymbol{\upsilon}^\circ) \big] \big\} \\
&\quad \le \mathbb{S}_\mu \exp\big\{ \mu \mathcal{U}(\boldsymbol{\upsilon}^\circ) - \mathfrak{N}(\mu, \boldsymbol{\upsilon}^\circ) - \mathfrak{t}_\mu(\boldsymbol{\upsilon}^\circ) \big\} \\
&\quad \le \int_{\mathcal{B}_\mu(\boldsymbol{\upsilon}^\circ)} \exp\big\{ \mu \mathcal{U}(\boldsymbol{\upsilon}) - \mathfrak{N}(\mu, \boldsymbol{\upsilon}) - \mathfrak{t}_\mu(\boldsymbol{\upsilon}) \big\} \frac{\pi(d\boldsymbol{\upsilon})}{\pi_\mu(\boldsymbol{\upsilon}^\circ)} \\
&\quad \le \nu_1 \int_{\mathcal{B}_\mu(\boldsymbol{\upsilon}^\circ)} \exp\big\{ \mu \mathcal{U}(\boldsymbol{\upsilon}) - \mathfrak{N}(\mu, \boldsymbol{\upsilon}) - \mathfrak{t}_\mu(\boldsymbol{\upsilon}) \big\} \frac{\pi(d\boldsymbol{\upsilon})}{\pi_\mu(\boldsymbol{\upsilon})} \\
&\quad \le \nu_1 \int_{\Upsilon^\circ} \exp\big\{ \mu \mathcal{U}(\boldsymbol{\upsilon}) - \mathfrak{N}(\mu, \boldsymbol{\upsilon}) - \mathfrak{t}_\mu(\boldsymbol{\upsilon}) \big\} \frac{\pi(d\boldsymbol{\upsilon})}{\pi_\mu(\boldsymbol{\upsilon})}.
\end{aligned}$$

As the right hand-side does not depend on $\boldsymbol{\upsilon}^\circ$, the bound applies to the maximum of this expression over $\boldsymbol{\upsilon}^\circ$. This implies in view of $I\!E \exp\big\{ \mu \mathcal{U}(\boldsymbol{\upsilon}) - \mathfrak{N}(\mu, \boldsymbol{\upsilon}) \big\} = 1$

$$I\!E \exp \sup_{\boldsymbol{\upsilon} \in \Upsilon^\circ} \big\{ \mathbb{S}_\mu \big[\mu \mathcal{U}(\boldsymbol{\upsilon}) - \mathfrak{N}(\mu, \boldsymbol{\upsilon}) - \mathfrak{t}_\mu(\boldsymbol{\upsilon}) \big] \big\} \le \nu_1 \int_{\Upsilon^\circ} \exp\big\{ -\mathfrak{t}_\mu(\boldsymbol{\upsilon}) \big\} \frac{\pi(d\boldsymbol{\upsilon})}{\pi_\mu(\boldsymbol{\upsilon})}.$$

As the sup over $\mu \in \mathbb{M}$ is not larger than the sum of the exponential terms, it holds

$$\mathbb{E} \exp \sup_{\boldsymbol{\upsilon} \in \Upsilon^\circ} \sup_{\mu \in \mathbb{M}} \bigl\{ \mathbb{S}_\mu \bigl[\mu \mathcal{U}(\boldsymbol{\upsilon}) - \mathfrak{N}(\mu, \boldsymbol{\upsilon}) - \mathfrak{t}_\mu(\boldsymbol{\upsilon}) \bigr] \bigr\} \leq \nu_1 \mathcal{T}.$$

This bound implies for each $\mathtt{x} > 0$ with probability at least $1 - e^{-\mathtt{x}}$

$$\mathbb{S}_\mu \bigl[\mu \mathcal{U}(\boldsymbol{\upsilon}) - \mathfrak{N}(\mu, \boldsymbol{\upsilon}) - \mathfrak{t}_\mu(\boldsymbol{\upsilon}) \bigr] \leq \mathtt{x} + \log\bigl(\nu_1 \mathcal{T}\bigr), \qquad \boldsymbol{\upsilon} \in \Upsilon^\circ. \tag{7.13}$$

Now define

$$w_\mu(\boldsymbol{\upsilon}^\circ) \stackrel{\text{def}}{=} \sup_{\boldsymbol{\upsilon} \in \mathcal{B}_\mu(\boldsymbol{\upsilon}^\circ)} \mu \bigl| \mathcal{U}(\boldsymbol{\upsilon}) - \mathcal{U}(\boldsymbol{\upsilon}^\circ) \bigr|.$$

With $\lambda \stackrel{\text{def}}{=} (3\nu_0 \mathtt{r}_1 / s) \wedge \sqrt{2\mathbb{Q}}$, it holds by Theorem 7.1 in view of $\lambda^2/2 \leq \mathbb{Q}$

$$\log \mathbb{E} \exp\bigl\{ s^{-1} w_\mu(\boldsymbol{\upsilon}) \bigr\} \leq 2\mathbb{Q}.$$

Then

$$\begin{aligned}
&\exp\bigl\{ s^{-1} \mu \bigl[\mathcal{U}(\boldsymbol{\upsilon}^\circ) - \mathbb{S}_\mu \mathcal{U}(\boldsymbol{\upsilon}^\circ) \bigr] - \mathbb{S}_\mu \mathfrak{t}_\mu(\boldsymbol{\upsilon}) \bigr\} \\
&\quad = \exp \int \bigl\{ s^{-1} \mu \bigl[\mathcal{U}(\boldsymbol{\upsilon}^\circ) - \mathcal{U}(\boldsymbol{\upsilon}) \bigr] - \mathfrak{t}_\mu(\boldsymbol{\upsilon}) \bigr\} \frac{\pi(d\boldsymbol{\upsilon})}{\pi_\mu(\boldsymbol{\upsilon}^\circ)} \\
&\quad \leq \int_{\mathcal{B}_\mu(\boldsymbol{\upsilon}^\circ)} \exp\bigl\{ s^{-1} w_\mu(\boldsymbol{\upsilon}) - \mathfrak{t}_\mu(\boldsymbol{\upsilon}) \bigr\} \frac{\pi(d\boldsymbol{\upsilon})}{\pi_\mu(\boldsymbol{\upsilon}^\circ)} \\
&\quad \leq \nu_1 \int_{\mathcal{B}_\mu(\boldsymbol{\upsilon}^\circ)} \exp\bigl\{ s^{-1} w_\mu(\boldsymbol{\upsilon}) - \mathfrak{t}_\mu(\boldsymbol{\upsilon}) \bigr\} \frac{\pi(d\boldsymbol{\upsilon})}{\pi_\mu(\boldsymbol{\upsilon})} \\
&\quad \leq \nu_1 \int_{\Upsilon^\circ} \exp\bigl\{ s^{-1} w_\mu(\boldsymbol{\upsilon}) - \mathfrak{t}_\mu(\boldsymbol{\upsilon}) \bigr\} \frac{\pi(d\boldsymbol{\upsilon})}{\pi_\mu(\boldsymbol{\upsilon})} \, .
\end{aligned}$$

As the right hand-side does not depend on $\boldsymbol{\upsilon}^\circ$, the bound applies to the maximum of this expression over $\boldsymbol{\upsilon}^\circ$. This implies

$$\mathbb{E} \exp \sup_{\boldsymbol{\upsilon} \in \Upsilon^\circ} \sup_{\mu \in \mathbb{M}} \bigl\{ s^{-1} \mu \bigl[\mathcal{U}(\boldsymbol{\upsilon}^\circ) - \mathbb{S}_\mu \mathcal{U}(\boldsymbol{\upsilon}^\circ) \bigr] - \mathbb{S}_\mu \mathfrak{t}_\mu(\boldsymbol{\upsilon}) \bigr\} \leq \nu_1 \mathcal{T} \exp(2\mathbb{Q}).$$

This implies similarly to (7.13) with probability at least $1 - e^{-\mathtt{x}}$:

$$s^{-1} \mu \bigl[\mathcal{U}(\boldsymbol{\upsilon}) - \mathbb{S}_\mu \mathcal{U}(\boldsymbol{\upsilon}) \bigr] - \mathbb{S}_\mu \mathfrak{t}_\mu(\boldsymbol{\upsilon}) \leq \mathtt{x} + 2\mathbb{Q} + \log\bigl(\nu_1 \mathcal{T}\bigr), \qquad \boldsymbol{\upsilon} \in \Upsilon^\circ.$$

Combining these two bounds yields (7.11) with probability at least $1 - 2e^{-\mathtt{x}}$. □

Remark 7.2. It is interesting to compare the uniform bound (7.11) of Theorem 7.4 and the pointwise bound (7.9): at which prise the pointwise result can be extended to a global one. The proposed construction involves two additional terms. One of them is

proportional to the local entropy $\mathbb{Q}$ and it comes from the local bound of Theorem 7.1 as the price for taking the local supremum. The second term is proportional to $\mathfrak{t}_\mu(\boldsymbol{\upsilon}) + \log(\mathcal{T})$ with $\mathcal{T}$ from (7.12) and it is responsible for extending the local maximum into the global one over Υ°.

7.3.1 Hitting probability

Let $M(\boldsymbol{\upsilon})$ be a deterministic *boundary* function. We aim at bounding the probability that a process $\mathcal{U}(\boldsymbol{\upsilon})$ on Υ° hits this boundary on the set Υ°. This precisely means the probability that $\sup_{\boldsymbol{\upsilon}\in\Upsilon^\circ}\{\mathcal{U}(\boldsymbol{\upsilon}) - M(\boldsymbol{\upsilon})\} \geq 0$. A particularly interesting problem is to describe for each $\mathtt{x} > 0$ the value $\mathfrak{z}_1(\mathtt{x})$ ensuring that

$$\mathbb{P}\Big(\sup_{\boldsymbol{\upsilon}\in\Upsilon^\circ} \{\mathcal{U}(\boldsymbol{\upsilon}) - M(\boldsymbol{\upsilon})\} \geq \mathfrak{z}_1(\mathtt{x}) \Big) \leq \mathrm{e}^{-\mathtt{x}}.$$

Let $\mathfrak{u}(\mu, \boldsymbol{\upsilon})$ be the multiscale upper function for $\mathcal{U}(\boldsymbol{\upsilon})$:

$$\mathbb{P}\Big\{ \sup_{\mu\in\mathbb{M}} \sup_{\boldsymbol{\upsilon}\in\Upsilon^\circ} \big[\mu\,\mathcal{U}(\boldsymbol{\upsilon}) - \mathfrak{u}(\mu, \boldsymbol{\upsilon})\big] \geq \mathfrak{z}_1(\mathtt{x}) \Big\} \leq 2\mathrm{e}^{-\mathtt{x}}; \tag{7.14}$$

cf. (7.11). Define

$$\begin{aligned} \mathcal{C}(\mu, \boldsymbol{\upsilon}) &\stackrel{\text{def}}{=} -\mathfrak{u}(\mu, \boldsymbol{\upsilon}) + \mu M(\boldsymbol{\upsilon}), \\ \mu^*(\boldsymbol{\upsilon}) &\stackrel{\text{def}}{=} \operatorname*{argmax}_{\mu\in\mathbb{M}} \mathcal{C}(\mu, \boldsymbol{\upsilon}), \\ \mathcal{C}^*(\boldsymbol{\upsilon}) &\stackrel{\text{def}}{=} \max_{\mu\in\mathbb{M}} \mathcal{C}(\mu, \boldsymbol{\upsilon}) = \mathcal{C}(\mu^*(\boldsymbol{\upsilon}), \boldsymbol{\upsilon}). \end{aligned}$$

The studied hitting probability can be described via the value

$$\mathfrak{g}(\Upsilon^\circ) \stackrel{\text{def}}{=} \inf_{\boldsymbol{\upsilon}\in\Upsilon^\circ} \mathcal{C}^*(\boldsymbol{\upsilon}) = \inf_{\boldsymbol{\upsilon}\in\Upsilon^\circ} \max_{\mu\in\mathbb{M}} \{-\mathfrak{u}(\mu, \boldsymbol{\upsilon}) + \mu M(\boldsymbol{\upsilon})\}.$$

The larger this value is, the smaller is the bound for the hitting probability. More precisely, let a fixed $\mathtt{x}$ and the corresponding $\mathfrak{z}_1(\mathtt{x})$ in (7.14) be fixed. If for each $\boldsymbol{\upsilon} \in \Upsilon^\circ$, the inequality $\mu M(\boldsymbol{\upsilon}) \geq \mathfrak{u}(\mu, \boldsymbol{\upsilon})$ holds with a properly selected $\mu = \mu(\boldsymbol{\upsilon})$, then the hitting probability is bounded by $2\mathrm{e}^{-\mathtt{x}}$.

Theorem 7.5. *Suppose* (7.14). *If* $\mathfrak{g}(\Upsilon^\circ) \geq \mathfrak{z}_1(\mathtt{x})$, *then*

$$\mathbb{P}\Big(\sup_{\boldsymbol{\upsilon}\in\Upsilon^\circ} \{\mathcal{U}(\boldsymbol{\upsilon}) - M(\boldsymbol{\upsilon})\} \geq 0 \Big) \leq 2\mathrm{e}^{-\mathtt{x}}.$$

Proof. For each $\upsilon \in \Upsilon^\circ$, in view of $\mathfrak{C}^*(\upsilon) \geq \mathfrak{g}(\Upsilon^\circ)$, it holds

$$\begin{aligned}
\{\mathcal{U}(\upsilon) - M(\upsilon) \geq 0\} &= \{\mu^*(\upsilon)[\mathcal{U}(\upsilon) - M(\upsilon)] \geq 0\} \\
&\subseteq \{\mu^*(\upsilon)[\mathcal{U}(\upsilon) - M(\upsilon)] + \mathfrak{C}^*(\upsilon) \geq \mathfrak{g}(\Upsilon^\circ)\} \\
&= \{\mu^*(\upsilon)\mathcal{U}(\upsilon) - \mathfrak{u}(\mu^*(\upsilon), \upsilon) \geq \mathfrak{z}_1(\mathtt{x})\} \\
&\subseteq \{\max_{\mu}[\mu\,\mathcal{U}(\upsilon) - \mathfrak{u}(\mu, \upsilon)] \geq \mathfrak{z}_1(\mathtt{x})\},
\end{aligned}$$

and the result follows from (7.14). □

7.4 Finite-dimensional smooth case

Here we discuss the special case when Υ is an open subset in $\mathbb{R}^p$, the stochastic process $\mathcal{U}(\upsilon)$ is absolutely continuous and its gradient $\nabla\mathcal{U}(\upsilon) \stackrel{\text{def}}{=} d\mathcal{U}(\upsilon)/d\upsilon$ has bounded exponential moments.

($\mathcal{E}D$) *There exist* $\mathtt{g} > 0$, $\nu_0 \geq 1$, *and for each* $\upsilon \in \Upsilon$, *a symmetric non-negative matrix* $H(\upsilon)$ *such that for any* $\lambda \leq \mathtt{g}$ *and any unit vector* $\gamma \in \mathbb{R}^p$, *it holds*

$$\log \mathbb{E} \exp\Big\{\lambda \frac{\gamma^\top \nabla\mathcal{U}(\upsilon)}{\|H(\upsilon)\gamma\|}\Big\} \leq \nu_0^2\lambda^2/2.$$

A natural candidate for $H^2(\upsilon)$ is the covariance matrix $\mathrm{Var}\big(\nabla\mathcal{U}(\upsilon)\big)$ provided that this matrix is well posed. Then the constant ν_0 can be taken close to one by reducing the value $\mathtt{g}$; see Lemma 7.12 below.

In what follows we fix a subset Υ° of Υ and establish a bound for the maximum of the process $\mathcal{U}(\upsilon, \upsilon^\circ) = \mathcal{U}(\upsilon) - \mathcal{U}(\upsilon^\circ)$ on Υ° for a fixed point υ°. We will assume existence of a dominating matrix $H^* = H^*(\Upsilon^\circ)$ such that $H(\upsilon) \preceq H^*$ for all $\upsilon \in \Upsilon^\circ$. We also assume that π is the Lebesgue measure on Υ. First we show that the differentiability condition $(\mathcal{E}D)$ implies $(\mathcal{E}d)$.

Lemma 7.6. *Assume that* $(\mathcal{E}D)$ *holds with some* $\mathtt{g}$ *and* $H(\upsilon) \preceq H^*$ *for* $\upsilon \in \Upsilon^\circ$. *Consider any* $\upsilon, \upsilon^\circ \in \Upsilon^\circ$. *Then it holds for* $|\lambda| \leq \mathtt{g}$

$$\log \mathbb{E} \exp\left\{\lambda \frac{\mathcal{U}(\upsilon, \upsilon^\circ)}{\|H^*(\upsilon - \upsilon^\circ)\|}\right\} \leq \frac{\nu_0^2\lambda^2}{2}.$$

Proof. Denote $\delta = \|\upsilon - \upsilon^\circ\|$, $\gamma = (\upsilon - \upsilon^\circ)/\delta$. Then

$$\mathcal{U}(\upsilon, \upsilon^\circ) = \delta\gamma^\top \int_0^1 \nabla\mathcal{U}(\upsilon^\circ + t\delta\gamma)dt$$

and $\|H^*(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^\circ)\| = \delta\|H^*\boldsymbol{\gamma}\|$. Now the Hölder inequality and $(\mathcal{E}D)$ yield

$$\begin{aligned}
&\mathbb{E}\exp\Big\{\lambda\frac{\mathcal{U}(\boldsymbol{\upsilon},\boldsymbol{\upsilon}^\circ)}{\|H^*(\boldsymbol{\upsilon}-\boldsymbol{\upsilon}^\circ)\|} - \frac{\nu_0^2\lambda^2}{2}\Big\} \\
&\quad = \mathbb{E}\exp\Big\{\int_0^1\Big[\lambda\frac{\boldsymbol{\gamma}^\top\nabla\mathcal{U}(\boldsymbol{\upsilon}^\circ + t\delta\boldsymbol{\gamma})}{\|H^*\boldsymbol{\gamma}\|} - \frac{\nu_0^2\lambda^2}{2}\Big]dt\Big\} \\
&\quad \le \int_0^1 \mathbb{E}\exp\Big\{\lambda\frac{\boldsymbol{\gamma}^\top\nabla\mathcal{U}(\boldsymbol{\upsilon}^\circ + t\delta\boldsymbol{\gamma})}{\|H^*\boldsymbol{\gamma}\|} - \frac{\nu_0^2\lambda^2}{2}\Big\}dt \le 1
\end{aligned}$$

as required. □

The result of Lemma 7.6 enables us to define $d(\boldsymbol{\upsilon},\boldsymbol{\upsilon}') = \|H^*(\boldsymbol{\upsilon}-\boldsymbol{\upsilon}^\circ)\|$ so that the corresponding ball coincides with the ellipsoid $B(\mathtt{r},\boldsymbol{\upsilon}^\circ)$. Now we bound the value $\mathbb{Q}(\Upsilon^\circ)$ for $\Upsilon^\circ = B(\mathtt{r}_1,\boldsymbol{\upsilon}^\circ)$.

Lemma 7.7. *Let $\Upsilon^\circ = B(\mathtt{r}_1,\boldsymbol{\upsilon}^\circ)$. Under the conditions of Lemma 7.6, it holds $\mathbb{Q}(\Upsilon^\circ) \le \mathfrak{c}_1 p$, where $\mathfrak{c}_1 = 2$ for $p \ge 2$, and $\mathfrak{c}_1 = 2.4$ for $p = 1$.*

Proof. The set Υ° coincides with the ellipsoid $B(\mathtt{r}_1,\boldsymbol{\upsilon}^\circ)$ while the d-ball $\mathcal{B}_k(\boldsymbol{\upsilon})$ coincides with the ellipsoid $B(\mathtt{r}_k,\boldsymbol{\upsilon}^\circ)$ for each $k \ge 2$. By change of variables, the study can be reduced to the case with $\boldsymbol{\upsilon}^\circ = 0$, $H^* \equiv \boldsymbol{I}_p$, $\mathtt{r}_1 = 1$, so that $B(\mathtt{r},\boldsymbol{\upsilon})$ is the usual Euclidean ball in $\mathbb{R}^p$ of radius $\mathtt{r}$. It is obvious that the measure of the overlap of two balls $B(1,0)$ and $B(2^{-k+1},\boldsymbol{\upsilon})$ for $\|\boldsymbol{\upsilon}\| \le 1$ is minimized when $\|\boldsymbol{\upsilon}\| = 1$, and this value is the same for all such $\boldsymbol{\upsilon}$. Define the number $a_{k,p}$ by

$$a_{k,p}^p \stackrel{\text{def}}{=} \frac{\pi\big(B(1,0)\big)}{\pi\big(B(1,0)\cap B(2^{-k+1},\boldsymbol{\upsilon})\big)}$$

for any $\boldsymbol{\upsilon}$ with $\|\boldsymbol{\upsilon}\| = 1$. It is easy to see that $a_{k,1} = 2^k$ and $a_{k,p}$ decreases with p. So, $M_k \le 2^{pk}$ and

$$\begin{aligned}
\mathbb{Q}(\Upsilon^\circ) &\le \frac{1}{3}\log(2^{1+p}) + \frac{4}{3}\sum_{k=2}^{\infty} 2^{-k}\log(2^{1+kp}) \\
&= \frac{\log 2}{3}\Big[3 + p + 2p\sum_{k=1}^{\infty}(k+1)2^{-k}\Big] = (3+7p)\frac{\log 2}{3} \le \mathfrak{c}_1 p,
\end{aligned}$$

where $\mathfrak{c}_1 = 2$ for $p \ge 2$, and $\mathfrak{c}_1 = 2.4$ for $p = 1$, and the result follows. □

7.4.1 Local central bound

Here we specify the local bounds of Theorem 7.1 and the central result of Corollary 7.3 to the smooth case.

Theorem 7.8. *Suppose* $(\mathcal{E}d)$. *For any* $\lambda \le \nu_0 \mathtt{g}$, $\mathtt{r}_1 > 0$, *and* $\boldsymbol{\upsilon}^\circ \in \Upsilon$

$$\log I\!\!E \exp\Big\{ \frac{\lambda}{3\nu_0 \mathtt{r}_1} \sup_{\boldsymbol{\upsilon} \in B(\mathtt{r}_1, \boldsymbol{\upsilon}^\circ)} \big| \mathcal{U}(\boldsymbol{\upsilon}) - \mathcal{U}(\boldsymbol{\upsilon}^\circ) \big| \Big\} \le \lambda^2/2 + \mathbb{Q},$$

where $\mathbb{Q} = \mathfrak{c}_1 p$.

We consider the local sets of the elliptic form $\Upsilon_\circ(\mathtt{r}) \stackrel{\text{def}}{=} \{\boldsymbol{\upsilon} : \|H_0(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}_0)\| \le \mathtt{r}\}$, where H_0 dominates $H(\boldsymbol{\upsilon})$ on this set: $H(\boldsymbol{\upsilon}) \preceq H_0$.

Theorem 7.9. *Let* $(\mathcal{E}D)$ *hold with some* $\mathtt{g}$ *and a matrix* $H(\boldsymbol{\upsilon})$. *Suppose that* $H(\boldsymbol{\upsilon}) \preceq H_0$ *for all* $\boldsymbol{\upsilon} \in \Upsilon_\circ(\mathtt{r})$. *Then*

$$I\!\!P\Big(\sup_{\boldsymbol{\upsilon} \in \Upsilon_\circ(\mathtt{r})} \Big\{ \frac{1}{3\nu_0} \mathcal{U}(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}_0) - \frac{1}{2} \|H_0(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}_0)\|^2 \Big\} \ge \mathfrak{z}_0(\mathtt{x}, p) \Big) \le \exp(-\mathtt{x}), \qquad (7.15)$$

where $\mathfrak{z}_0(\mathtt{x}, p)$ *coincides with* $\mathfrak{z}_0(\mathtt{x}, \mathbb{Q})$ *from* (7.7) *with* $\mathbb{Q} = \mathfrak{c}_1 p$.

Remark 7.3. An important feature of the established result is that the bound in the right hand-side of (7.15) does not depend on the value $\mathtt{r}$ describing the radius of the local vicinity around the central point $\boldsymbol{\upsilon}_0$. In the ideal case one would apply this result with $\mathtt{r} = \infty$ provided that the conditions $H(\boldsymbol{\upsilon}) \le H_0$ is fulfilled uniformly over Υ.

Proof. Lemma 7.7 implies $(\mathcal{E}d)$ with $d(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}_0) = \|H_0(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}_0)\|^2/2$. Now the result follows from Corollary 7.3. □

7.4.2 A global upper function

Now we specify the general result of Theorem 7.4 to the smooth case. To make the formulation more transparent, the matrix $H(\boldsymbol{\upsilon})$ from condition $(\mathcal{E}D)$ is assumed to be uniformly bounded by a fixed matrix H^*. Let $\mathtt{r}$ be fixed with $\mathtt{r}^2 \ge p/2$. We aim to build un upper function for the process $\mathcal{U}(\cdot)$ on the complement of the central set $\Upsilon_\circ(\mathtt{r}) \stackrel{\text{def}}{=} \{\boldsymbol{\upsilon} : \|H_0 \boldsymbol{\upsilon}\| \le \mathtt{r}\}$. The penalty function $\mathfrak{t}(\boldsymbol{\upsilon})$ is taken independent of μ as a logarithmic function of $\|H_0 \boldsymbol{\upsilon}\|$.

Theorem 7.10. *Assume* $(\mathcal{E})$ *and* $(\mathcal{E}D)$ *with* $H(\boldsymbol{\upsilon}) \preceq H^*$ *for all* $\boldsymbol{\upsilon} \in \Upsilon$. *Let* $\mathtt{r}_1 \ge 1$ *be such that* $\det(\mathtt{r}_1^{-1} H^*) \le \det(H_0)$, *and* s *be such that* $3\nu_0 \mathtt{r}_1 / s \le \mathtt{g} \wedge \sqrt{2\mathfrak{c}_1 p}$. *Given* $\mathtt{r} \ge \sqrt{p/2}$, *define*

$$\mathfrak{t}(\boldsymbol{\upsilon}) \stackrel{\text{def}}{=} (p+2) \log\big(\|H_0 \boldsymbol{\upsilon}\|\big), \qquad \boldsymbol{\upsilon} \in \Upsilon.$$

Then for any $\mathtt{x} > 0$, *it holds with probability at least* $1 - 2\mathrm{M}\mathrm{e}^{-\mathtt{x}}$ *for* $\mathrm{M} = \sum_{\mu \in \mathbb{M}} \mu^p$

$$\mu \mathcal{U}(\boldsymbol{\upsilon}) - \mathbb{S}_\mu \mathfrak{N}(\mu, \boldsymbol{\upsilon}) - (1+s)\mathfrak{t}(\boldsymbol{\upsilon}) \le (1+s)\mathtt{x} + 2s\mathfrak{c}_1 p, \qquad \boldsymbol{\upsilon} \in \Upsilon \setminus \Upsilon_\circ(\mathtt{r}). \quad (7.16)$$

Proof. For the measure Lebesgue measure π on $\mathbb{R}^p$, it holds for $\mathcal{B}_\mu(\upsilon^\circ) = B(\mathtt{r}_1/\mu, \upsilon^\circ)$:

$$\frac{1}{\pi\bigl(\mathcal{B}_\mu(\upsilon^\circ)\bigr)} = \frac{\det(\mathtt{r}_1^{-1}\mu H^*)}{\omega_p} \le \frac{\mu^p \det(H_0)}{\omega_p},$$

where ω_p is the measure of the unit ball in $\mathbb{R}^p$. In particular, this measure does not depend on the location υ° and thus, $\nu_1 = 1$; see (7.10). The change of variables yields

$$\begin{aligned}
\mathcal{T} &= \sum_{\mu\in\mathbb{M}} \int_{\Upsilon^\circ} \frac{1}{\pi(\mathcal{B}_\mu(\upsilon))} \exp\bigl\{-\mathfrak{t}(\upsilon)\bigr\} d\pi(\upsilon) \\
&\le \sum_{\mu\in\mathbb{M}} \mu^p \int_{\mathbb{R}^p} \frac{\det(H_0)}{\omega_p} \|H_0\upsilon\|^{-p-2} \mathbb{1}\bigl(\|H_0\upsilon\| \ge \mathtt{r}\bigr) d\pi(\upsilon) \\
&= \frac{\mathrm{M}}{\omega_p} \int_{\|\boldsymbol{u}\|\ge\mathtt{r}} \|\boldsymbol{u}\|^{-p-2} d\boldsymbol{u} \le \mathrm{M}p/(2\mathtt{r}^2) \le \mathrm{M}.
\end{aligned}$$

Symmetricity arguments imply $\mathbb{S}_\mu\|H_0\upsilon\| = \|H_0\upsilon\|$ and concavity of the log-function yields $\mathbb{S}_\mu\mathfrak{t}(\upsilon) \le \mathfrak{t}(\upsilon)$. Now the result (7.16) follows from Theorem 7.4 in view of $\mathbb{Q} \le \mathfrak{c}_1 p$. □

7.5 Auxiliary facts

Lemma 7.11. *For any r.v.'s ξ_k and $\lambda_k \ge 0$ such that $\Lambda = \sum_k \lambda_k \le 1$*

$$\log \mathbb{E}\exp\biggl(\sum_k \lambda_k \xi_k\biggr) \le \sum_k \lambda_k \log \mathbb{E}\mathrm{e}^{\xi_k}.$$

Proof. Convexity of e^x and concavity of x^Λ imply

$$\begin{aligned}
\mathbb{E}\exp\biggl\{\frac{\Lambda}{\Lambda}\sum_k \lambda_k\bigl(\xi_k - \log\mathbb{E}\mathrm{e}^{\xi_k}\bigr)\biggr\} &\le \mathbb{E}^\Lambda \exp\biggl\{\frac{1}{\Lambda}\sum_k \lambda_k\bigl(\xi_k - \log\mathbb{E}\mathrm{e}^{\xi_k}\bigr)\biggr\} \\
&\le \biggl\{\frac{1}{\Lambda}\sum_k \lambda_k \mathbb{E}\exp\bigl(\xi_k - \log\mathbb{E}\mathrm{e}^{\xi_k}\bigr)\biggr\}^\Lambda = 1.
\end{aligned}$$

□

Lemma 7.12. *Let a r.v. ξ fulfill $\mathbb{E}\xi = 0$, $\mathbb{E}\xi^2 = 1$ and $\mathbb{E}\exp(\lambda_1|\xi|) = \varkappa < \infty$ for some $\lambda_1 > 0$. Then for any $\varrho < 1$ there is a constant C_1 depending on $\varkappa$, λ_1 and ϱ only such that for $\lambda < \varrho\lambda_1$*

$$\log \mathbb{E}\mathrm{e}^{\lambda\xi} \le C_1\lambda^2/2.$$

Moreover, there is a constant $\lambda_2 > 0$ such that for all $\lambda \le \lambda_2$

$$\log \mathbb{E}\mathrm{e}^{\lambda\xi} \ge \varrho\lambda^2/2.$$

Proof. Define $h(x) = (\lambda - \lambda_1)x + m\log(x)$ for $m \geq 0$ and $\lambda < \lambda_1$. It is easy to see by a simple algebra that

$$\max_{x \geq 0} h(x) = -m + m\log\frac{m}{\lambda_1 - \lambda}.$$

Therefore for any $x \geq 0$

$$\lambda x + m\log(x) \leq \lambda_1 x + \log\left(\frac{m}{\mathrm{e}(\lambda_1 - \lambda)}\right)^m.$$

This implies for all $\lambda < \lambda_1$

$$I\!\!E|\xi|^m \exp(\lambda|\xi|) \leq \left(\frac{m}{\mathrm{e}(\lambda_1 - \lambda)}\right)^m I\!\!E \exp(\lambda_1|\xi|).$$

Suppose now that for some $\lambda_1 > 0$, it holds $I\!\!E \exp(\lambda_1|\xi|) = \varkappa(\lambda_1) < \infty$. Then the function $h_0(\lambda) = I\!\!E \exp(\lambda\xi)$ fulfills $h_0(0) = 1$, $h_0'(0) = I\!\!E\xi = 0$, $h_0''(0) = 1$ and for $\lambda < \lambda_1$,

$$h_0''(\lambda) = I\!\!E\xi^2 \mathrm{e}^{\lambda\xi} \leq I\!\!E\xi^2 \mathrm{e}^{\lambda|\xi|} \leq \frac{1}{(\lambda_1 - \lambda)^2} I\!\!E \exp(\lambda_1|\xi|).$$

This implies by the Taylor expansion for $\lambda < \varrho\lambda_1$ that

$$h_0(\lambda) \leq 1 + C_1\lambda^2/2$$

with $C_1 = \varkappa(\lambda_1)/\left\{\lambda_1^2(1-\varrho)^2\right\}$, and hence, $\log h_0(\lambda) \leq C_1\lambda^2/2$. □

References

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. .

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.

SFB 649 Discussion Paper Series 2011

For a complete list of Discussion Papers published by the SFB 649, please visit http://sfb649.wiwi.hu-berlin.de.

001 "Localising temperature risk" by Wolfgang Karl Härdle, Brenda López Cabrera, Ostap Okhrin and Weining Wang, January 2011.
002 "A Confidence Corridor for Sparse Longitudinal Data Curves" by Shuzhuan Zheng, Lijian Yang and Wolfgang Karl Härdle, January 2011.
003 "Mean Volatility Regressions" by Lu Lin, Feng Li, Lixing Zhu and Wolfgang Karl Härdle, January 2011.
004 "A Confidence Corridor for Expectile Functions" by Esra Akdeniz Duran, Mengmeng Guo and Wolfgang Karl Härdle, January 2011.
005 "Local Quantile Regression" by Wolfgang Karl Härdle, Vladimir Spokoiny and Weining Wang, January 2011.
006 "Sticky Information and Determinacy" by Alexander Meyer-Gohde, January 2011.
007 "Mean-Variance Cointegration and the Expectations Hypothesis" by Till Strohsal and Enzo Weber, February 2011.
008 "Monetary Policy, Trend Inflation and Inflation Persistence" by Fang Yao, February 2011.
009 "Exclusion in the All-Pay Auction: An Experimental Investigation" by Dietmar Fehr and Julia Schmid, February 2011.
010 "Unwillingness to Pay for Privacy: A Field Experiment" by Alastair R. Beresford, Dorothea Kübler and Sören Preibusch, February 2011.
011 "Human Capital Formation on Skill-Specific Labor Markets" by Runli Xie, February 2011.
012 "A strategic mediator who is biased into the same direction as the expert can improve information transmission" by Lydia Mechtenberg and Johannes Münster, March 2011.
013 "Spatial Risk Premium on Weather Derivatives and Hedging Weather Exposure in Electricity" by Wolfgang Karl Härdle and Maria Osipenko, March 2011.
014 "Difference based Ridge and Liu type Estimators in Semiparametric Regression Models" by Esra Akdeniz Duran, Wolfgang Karl Härdle and Maria Osipenko, March 2011.
015 "Short-Term Herding of Institutional Traders: New Evidence from the German Stock Market" by Stephanie Kremer and Dieter Nautz, March 2011.
016 "Oracally Efficient Two-Step Estimation of Generalized Additive Model" by Rong Liu, Lijian Yang and Wolfgang Karl Härdle, March 2011.
017 "The Law of Attraction: Bilateral Search and Horizontal Heterogeneity" by Dirk Hofmann and Salmai Qari, March 2011.
018 "Can crop yield risk be globally diversified?" by Xiaoliang Liu, Wei Xu and Martin Odening, March 2011.
019 "What Drives the Relationship Between Inflation and Price Dispersion? Market Power vs. Price Rigidity" by Sascha Becker, March 2011.
020 "How Computational Statistics Became the Backbone of Modern Data Science" by James E. Gentle, Wolfgang Härdle and Yuichi Mori, May 2011.
021 "Customer Reactions in Out-of-Stock Situations – Do promotion-induced phantom positions alleviate the similarity substitution hypothesis?" by Jana Luisa Diels and Nicole Wiebach, May 2011.

SFB 649, Spandauer Str. 1, D-10178 Berlin
http://sfb649.wiwi.hu-berlin.de

This research was supported by the Deutsche Forschungsgemeinschaft through the SFB 649 "Economic Risk".

SFB 649 Discussion Paper Series 2011

For a complete list of Discussion Papers published by the SFB 649, please visit http://sfb649.wiwi.hu-berlin.de.

022 "Extreme value models in a conditional duration intensity framework" by Rodrigo Herrera and Bernhard Schipp, May 2011.
023 "Forecasting Corporate Distress in the Asian and Pacific Region" by Russ Moro, Wolfgang Härdle, Saeideh Aliakbari and Linda Hoffmann, May 2011.
024 "Identifying the Effect of Temporal Work Flexibility on Parental Time with Children" by Juliane Scheffel, May 2011.
025 "How do Unusual Working Schedules Affect Social Life?" by Juliane Scheffel, May 2011.
026 "Compensation of Unusual Working Schedules" by Juliane Scheffel, May 2011.
027 "Estimation of the characteristics of a Lévy process observed at arbitrary frequency" by Johanna Kappus and Markus Reiß, May 2011.
028 "Asymptotic equivalence and sufficiency for volatility estimation under microstructure noise" by Markus Reiß, May 2011.
029 "Pointwise adaptive estimation for quantile regression" by Markus Reiß, Yves Rozenholc and Charles A. Cuenod, May 2011.
030 "Developing web-based tools for the teaching of statistics: Our Wikis and the German Wikipedia" by Sigbert Klinke, May 2011.
031 "What Explains the German Labor Market Miracle in the Great Recession?" by Michael C. Burda and Jennifer Hunt, June 2011.
032 "The information content of central bank interest rate projections: Evidence from New Zealand" by Gunda-Alexandra Detmers and Dieter Nautz, June 2011.
033 "Asymptotics of Asynchronicity" by Markus Bibinger, June 2011.
034 "An estimator for the quadratic covariation of asynchronously observed Itô processes with noise: Asymptotic distribution theory" by Markus Bibinger, June 2011.
035 "The economics of TARGET2 balances" by Ulrich Bindseil and Philipp Johann König, June 2011.
036 "An Indicator for National Systems of Innovation - Methodology and Application to 17 Industrialized Countries" by Heike Belitz, Marius Clemens, Christian von Hirschhausen, Jens Schmidt-Ehmcke, Axel Werwatz and Petra Zloczysti, June 2011.
037 "Neurobiology of value integration: When value impacts valuation" by Soyoung Q. Park, Thorsten Kahnt, Jörg Rieskamp and Hauke R. Heekeren, June 2011.
038 "The Neural Basis of Following Advice" by Guido Biele, Jörg Rieskamp, Lea K. Krugel and Hauke R. Heekeren, June 2011.
039 "The Persistence of "Bad" Precedents and the Need for Communication: A Coordination Experiment" by Dietmar Fehr, June 2011.
040 "News-driven Business Cycles in SVARs" by Patrick Bunk, July 2011.
041 "The Basel III framework for liquidity standards and monetary policy implementation" by Ulrich Bindseil and Jeroen Lamoot, July 2011.
042 "Pollution permits, Strategic Trading and Dynamic Technology Adoption" by Santiago Moreno-Bromberg and Luca Taschini, July 2011.
043 "CRRA Utility Maximization under Risk Constraints" by Santiago Moreno-Bromberg, Traian A. Pirvu and Anthony Réveillac, July 2011.

SFB 649, Spandauer Str. 1, D-10178 Berlin
http://sfb649.wiwi.hu-berlin.de

This research was supported by the Deutsche Forschungsgemeinschaft through the SFB 649 "Economic Risk".

SFB 649 Discussion Paper Series 2011

For a complete list of Discussion Papers published by the SFB 649, please visit http://sfb649.wiwi.hu-berlin.de.

044 "Predicting Bid-Ask Spreads Using Long Memory Autoregressive Conditional Poisson Models" by Axel Groß-Klußmann and Nikolaus Hautsch, July 2011.
045 "Bayesian Networks and Sex-related Homicides" by Stephan Stahlschmidt, Helmut Tausendteufel and Wolfgang K. Härdle, July 2011.
046 "The Regulation of Interdependent Markets", by Raffaele Fiocco and Carlo Scarpa, July 2011.
047 "Bargaining and Collusion in a Regulatory Model", by Raffaele Fiocco and Mario Gilli, July 2011.
048 "Large Vector Auto Regressions", by Song Song and Peter J. Bickel, August 2011.
049 "Monetary Policy, Determinacy, and the Natural Rate Hypothesis", by Alexander Meyer-Gohde, August 2011.
050 "The impact of context and promotion on consumer responses and preferences in out-of-stock situations", by Nicole Wiebach and Jana L. Diels, August 2011.
051 "A Network Model of Financial System Resilience", by Kartik Anand, Prasanna Gai, Sujit Kapadia, Simon Brennan and Matthew Willison, August 2011.
052 "Rollover risk, network structure and systemic financial crises", by Kartik Anand, Prasanna Gai and Matteo Marsili, August 2011.
053 "When to Cross the Spread: Curve Following with Singular Control" by Felix Naujokat and Ulrich Horst, August 2011.
054 "TVICA - Time Varying Independent Component Analysis and Its Application to Financial Data" by Ray-Bing Chen, Ying Chen and Wolfgang K. Härdle, August 2011.
055 "Pricing Chinese rain: a multi-site multi-period equilibrium pricing model for rainfall derivatives" by Wolfgang K. Härdle and Maria Osipenko, August 2011.
056 "Limit Order Flow, Market Impact and Optimal Order Sizes: Evidence from NASDAQ TotalView-ITCH Data" by Nikolaus Hautsch and Ruihong Huang, August 2011.
057 "Optimal Display of Iceberg Orders" by Gökhan Cebiroğlu and Ulrich Horst, August 2011.
058 "Optimal liquidation in dark pools" by Peter Kratz and Torsten Schöneborn, September 2011.
059 "The Merit of High-Frequency Data in Portfolio Allocation" by Nikolaus Hautsch, Lada M. Kyj and Peter Malec, September 2011.
060 "On the Continuation of the Great Moderation: New evidence from G7 Countries" by Wenjuan Chen, September 2011.
061 "Forward-backward systems for expected utility maximization" by Ulrich Horst, Ying Hu, Peter Imkeller, Anthony Réveillac and Jianing Zhang.
062 "On heterogeneous latent class models with applications to the analysis of rating scores" by Aurélie Bertrand and Christian M. Hafner, October 2011.
063 "Multivariate Volatility Modeling of Electricity Futures" by Luc Bauwens, Christian Hafner and Diane Pierret, October 2011.

SFB 649, Spandauer Str. 1, D-10178 Berlin
http://sfb649.wiwi.hu-berlin.de

This research was supported by the Deutsche Forschungsgemeinschaft through the SFB 649 "Economic Risk".

SFB 649 Discussion Paper Series 2011

For a complete list of Discussion Papers published by the SFB 649, please visit http://sfb649.wiwi.hu-berlin.de.

064 "Semiparametric Estimation with Generated Covariates" by Enno Mammen, Christoph Rothe and Melanie Schienle, October 2011.
065 "Linking corporate reputation and shareholder value using the publication of reputation rankings" by Sven Tischer and Lutz Hildebrandt, October 2011.
066 "Monitoring, Information Technology and the Labor Share" by Dorothee Schneider, October 2011.
067 "Minimal Supersolutions of BSDEs with Lower Semicontinuous Generators" by Gregor Heyne, Michael Kupper and Christoph Mainberger, October 2011.
068 "Bargaining, Openness, and the Labor Share" by Dorothee Schneider, October 2011.
069 "The Labor Share: A Review of Theory and Evidence" by Dorothee Schneider, October 2011.
070 "The Power of Sunspots: An Experimental Analysis" by Dietmar Fehr, Frank Heinemann and Aniol Llorente-Saguer, October 2011.
071 "Econometric analysis of volatile art markets" by Fabian Y. R. P. Bocart and Christian M. Hafner, October 2011.
072 "Financial Network Systemic Risk Contributions" by Nikolaus Hautsch, Julia Schaumburg and Melanie Schienle, October 2011.
073 "Calibration of self-decomposable Lévy models" by Mathias Trabs, November 2011.
074 "Time-Varying Occupational Contents: An Additional Link between Occupational Task Profiles and Individual Wages" by Alexandra Fedorets, November 2011.
075 "Changes in Occupational Demand Structure and their Impact on Individual Wages" by Alexandra Fedorets, November 2011.
076 "Nonparametric Nonstationary Regression with Many Covariates" by Melanie Schienle, November 2011.
077 "Increasing Weather Risk: Fact or Fiction?" by Weining Wang, Ihtiyor Bobojonov, Wolfgang Karl Härdle and Martin Odening, November 2011.
078 "Spatially Adaptive Density Estimation by Localised Haar Projections" by Florian Gach, Richard Nickl and Vladimir Spokoiny, November 2011.
079 "Martingale approach in pricing and hedging European options under regime-switching" by Grigori N. Milstein and Vladimir Spokoiny, November 2011.
080 "Sparse Non Gaussian Component Analysis by Semidefinite Programming" by Elmar Diederichs, Anatoli Juditsky, Arkadi Nemirovski and Vladimir Spokoiny, November 2011.
081 "Parametric estimation. Finite sample theory" by Vladimir Spokoiny, November 2011.

SFB 649, Spandauer Str. 1, D-10178 Berlin
http://sfb649.wiwi.hu-berlin.de

This research was supported by the Deutsche Forschungsgemeinschaft through the SFB 649 "Economic Risk".