Exponential bounds for the minimum contrast with some applications

Golubev, Yuri,
CMI
39, rue F. Joliot-Curie
13453 Marseille FRANCE,
golubev@gyptis.univ-mrs.fr

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

Abstract

The paper studies parametric minimum contrast estimates under rather general conditions. The quality if estimation is measured by the rate function related to the contrast which allows for stating the results without specifying the particular parametric structure of the model. This approach permits also to go far beyond the classical i.i.d. case and to obtain nonasymptotic upper bounds for the risk. These bounds apply even for small or moderate samples. They also cover the case of misspecified parametric models. Another important feature of the approach is that it works well in the case when the parametric set can be unbounded and non-compact. In the case of a smooth contrast, the obtained exponential bounds do not rely on the covering numbers and can be easily computed. We also illustrate how these bound can be used for statistical inference: bounding the estimation risk, constructing the confidence sets for the underlying parameters, establishing the concentration properties of the minimum contrast estimate.

The general results are specified to the case of a Gaussian contrast and of an i.i.d. sample. We also illustrate the approach by several popular examples including least squares and least absolute deviation contrasts and the problem of estimating the location of the change point. What we obtain in these examples slightly differs from usual asymptotic results known in the classical literature. This difference is due to the unboundness of the parameter set and a possible model misspecification.

Keywords: risk bound, quasi maximum likelihood, smooth contrast

AMS 2000 Subject Classification: 62F10 Secondary: 62F12, 62F25

1 Introduction

One of the most fundamental ideas in statistics is to describe the observation data using a simple parametric family $(I\!P_{\boldsymbol{\theta}}, \boldsymbol{\theta} \in \Theta)$, where Θ is a subset in a finite dimensional space, typically, in $I\!R^p$. In this situation, an unknown distribution $I\!P$ of the observations $Y \in I\!R^n$ is characterized by the value of the parameter $\boldsymbol{\theta} \in \Theta$ and the problem of statistical inferences about $I\!P$ is reduced to recovering $\boldsymbol{\theta}$. The classical parametric theory mostly focuses on the asymptotic properties of estimates as the sample size n tends to infinity. Typical results claim that the maximum likelihood and Bayes estimates are asymptotically optimal under certain regularity conditions (see e.g. [6]). The maximum likelihood and Bayes estimates can be generalized in several ways resulting in the so-called *minimum contrast* and *M-estimators* proposed in [8] (see also [9] for more detail). The idea behind this generalization is to estimate the underlying parameter $\boldsymbol{\theta}$ by minimizing over Θ of a *contrast function* which we denote by $-L(\boldsymbol{\theta})$:

$$\widetilde{\boldsymbol{\theta}} = \operatorname*{argmin}_{\boldsymbol{\theta}\in\Theta}\{-L(\boldsymbol{\theta})\} = \operatorname*{argmax}_{\boldsymbol{\theta}\in\Theta} L(\boldsymbol{\theta}).$$

The negative sign in this notation comes from the main example we have in mind when $L(\boldsymbol{\theta})$ is the log-likelihood or quasi log-likelihood. The contrast function is a random field on Θ which is somehow related to the underlying parametric family $(I\!P_{\boldsymbol{\theta}}, \boldsymbol{\theta} \in \Theta)$. A natural condition on this field is that its expectation under the "true" measure $I\!P_{\boldsymbol{\theta}_0}$ is minimized at the "true" parameter $\boldsymbol{\theta}_0$, i.e.

$$I\!E_{\boldsymbol{\theta}_0} L(\boldsymbol{\theta}_0) = \max_{\boldsymbol{\theta}\in\Theta} I\!E_{\boldsymbol{\theta}_0} L(\boldsymbol{\theta}). \tag{1.1}$$

When the real distribution $I\!P$ doesn't belong to the parametric family $(I\!P_{\boldsymbol{\theta}}, \boldsymbol{\theta} \in \Theta)$, then the "target" of estimation can be naturally defined as the point of minimum of $-I\!E\, L(\boldsymbol{\theta})$. We will see that this point $\boldsymbol{\theta}_0$ indeed minimizes a special distance between the underlying measure $I\!P$ and the measures $I\!P_{\boldsymbol{\theta}}$ from the given parametric family.

The classical parametric statistics focus mostly on statistical properties of the distance between $\widetilde{\boldsymbol{\theta}}$ and the true value $\boldsymbol{\theta}_0$. There is a vast literature on this issue. We only mention the book [6], which provides a comprehensive study of asymptotic properties of maximum likelihood and Bayessian estimators. Large deviation results about minimum contrast estimators can be found in [7] and [10], while subtle small sample size properties of these estimates are presented in [4] and [5]. There exists a number of important studies properties of the minimum contrast estimates in a quite general situation, when

the parameter set Θ is a subset of some functional space. We mention the papers [11], [2], [3], [1] and the other related references therein. The principal facts established in these papers claim some concentration properties for a rather general minimum contrast estimate in term of metric entropy properties of Θ.

The approach of this paper is a bit different in the sense that we study exponential moments, or more precisely the rate function, of $L(\widetilde{\boldsymbol{\theta}}) - L(\boldsymbol{\theta}_0)$ under rather general assumptions. To explain the main idea, denote $L(\boldsymbol{\theta}, \boldsymbol{\theta}') = L(\boldsymbol{\theta}) - L(\boldsymbol{\theta}')$ and compute the moment generating function $\log I\!\!E \exp\{\mu L(\boldsymbol{\theta}, \boldsymbol{\theta}_0)\}$ for this random variable. Let

$$\mu^*(\boldsymbol{\theta}, \boldsymbol{\theta}_0) \stackrel{\mathrm{def}}{=} \operatorname*{argmax}_{\mu} \{ - \log I\!\!E \exp[\mu L(\boldsymbol{\theta}, \boldsymbol{\theta}_0)] \}$$

then the rate function is defined as follows

$$\mathfrak{M}(\boldsymbol{\theta}, \boldsymbol{\theta}_0) \stackrel{\mathrm{def}}{=} - \log I\!\!E \exp[\mu^*(\boldsymbol{\theta}, \boldsymbol{\theta}_0) L(\boldsymbol{\theta}, \boldsymbol{\theta}_0)].$$

By the above definitions we obviously get the following identity

$$I\!\!E \exp\Big\{ \mu^*(\boldsymbol{\theta}, \boldsymbol{\theta}_0) L(\boldsymbol{\theta}, \boldsymbol{\theta}_0) + \mathfrak{M}(\boldsymbol{\theta}, \boldsymbol{\theta}_0) \Big\} = 1$$

which holds true for any $\boldsymbol{\theta} \in \Theta$. We aim to extend this pointwise result to the supremum over $\boldsymbol{\theta} \in \Theta$, that is, by replacing $\boldsymbol{\theta}$ with the estimate $\widetilde{\boldsymbol{\theta}}$. Unfortunately, in the general situation $I\!\!E \exp\{ \mu^*(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}_0) L(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}_0) + \mathfrak{M}(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}_0) \}$ explodes. However, it turns out that under some assumptions, for any $\rho, s \in [0, 1)$,

$$I\!\!E \exp\Big\{ \rho \big[\mu^*(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}_0) L(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}_0) + s \mathfrak{M}(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}_0) \big] \Big\} \le \mathfrak{Q}(\rho, s), \tag{1.2}$$

where the constant $\mathfrak{Q}(\rho, s)$ can be easily controlled in typical examples. Section 2.5 presents some useful corollaries of this inequality including the bound for the estimation risk of $\widetilde{\boldsymbol{\theta}}$ with a polynomial loss function, concentration properties of $\widetilde{\boldsymbol{\theta}}$, confidence sets for the target $\boldsymbol{\theta}_0$ based on the maximum contrast $L(\widetilde{\boldsymbol{\theta}})$. In the i.i.d. case considered in Section 5 the results also yield root-n consistency of $\widetilde{\boldsymbol{\theta}}$.

The basic inequality (1.2) allows to address the following questions important for many practical applications:

- how one can extend the classical asymptotic results to the "small sample size" situation. The upper bound (1.2) is nonasymptotic and it holds true for any sample size. However, in some applications it is required that the sample size should be larger than a fixed prescribed value.

- what happens if the parametric model is misspecified. In the i.i.d. case studied in Section 5, we show that the target of estimation is the point of the parametric family which minimizes the Kullback-Leibler divergence between the underlying measure and the given parametric family.

- what is the accuracy of estimation when the parameter set Θ is not compact. We present some examples in Section 6 illustrating that the quality of the minimum contrast estimates may heavily depend on Θ and on the behavior of the exponential moments of the contrast function for large $\boldsymbol{\theta}$. In general, the rate of convergence may differ from the classical root-n behavior.

Section 3 specifies the approach to the important case of a smooth contrast. In this situations the main conditions ensuring (1.2) are simplified and the final bound can be written as an integral over the parameter set. This allows, in particular, to avoid computing the entropy numbers. Section 5 illustrates how our approach applies to the classical i.i.d. case. Notice also that the main inequality (1.2) can be simplified in the case of a Gaussian contrast, see Section 4 for more detail.

Section 6 illustrates the applications of the general bound for the problem of mean, scale and median estimation. The last example in this section concerns the prominent change point problem. We particularly show that in the case when the size of the change is completely unknown, the rate of estimation of the location of the change differs from the well known parametric rate $1/n$ and it depends on the distance to the edge and involves some extra log-log factor.

2 Risk bound for the minimum contrast

This section presents a general exponential bound for the value of the minimal contrast in the rather general set-up which includes the parametric situation as an important special case. Let $\bigl(-L(\boldsymbol{\theta}), \boldsymbol{\theta} \in \Theta\bigr)$ be a *contrast function* of a finite dimensional parameter $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^p$. From now on we denote for brevity $L(\boldsymbol{\theta}) = L(\boldsymbol{\theta})$ and suppose that $L(\boldsymbol{\theta})$ is a separable integrable random field on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$. The *minimum contrast estimate* is defined as the minimizer of $-L(\boldsymbol{\theta})$ or, equivalently, by

$$\widetilde{\boldsymbol{\theta}} = \operatorname*{argmax}_{\boldsymbol{\theta} \in \Theta} L(\boldsymbol{\theta}).$$

The "target" of estimation is the value $\boldsymbol{\theta}_0$ which minimizes the expectation of the contrast $-L(\boldsymbol{\theta})$:

$$\boldsymbol{\theta}_0 = \operatorname*{argmax}_{\boldsymbol{\theta}\in\Theta} I\!\!E L(\boldsymbol{\theta}).$$

We also denote $L(\boldsymbol{\theta},\boldsymbol{\theta}') = L(\boldsymbol{\theta}) - L(\boldsymbol{\theta}')$. It is clear that $\widetilde{\boldsymbol{\theta}} = \operatorname{argmax}_{\boldsymbol{\theta}\in\Theta} L(\boldsymbol{\theta},\boldsymbol{\theta}')$ and $\boldsymbol{\theta}_0 = \operatorname{argmax}_{\boldsymbol{\theta}\in\Theta} I\!\!E L(\boldsymbol{\theta},\boldsymbol{\theta}')$ whatever $\boldsymbol{\theta}' \in \Theta$ is.

The main object of our study is the value of the *minimum contrast*, or, equivalently, the maximum of the random field $L(\boldsymbol{\theta},\boldsymbol{\theta}_0)$:

$$L(\widetilde{\boldsymbol{\theta}},\boldsymbol{\theta}_0) = \sup_{\boldsymbol{\theta}\in\Theta} L(\boldsymbol{\theta},\boldsymbol{\theta}_0).$$

By definition, $L(\widetilde{\boldsymbol{\theta}},\boldsymbol{\theta})$ is a non-negative random variable and we aim to show that it has bounded exponential moments under some natural conditions.

2.1 Examples

Although the particular structure of the contrast function is not important for our results, but it is useful to have some specific examples in mind.

2.1.1 Maximum Likelihood Estimate

One of the most popular examples of the contrast function is the *log-likelihood*. Let $(I\!P_{\boldsymbol{\theta}}, \boldsymbol{\theta}\in\Theta\subset I\!R^p)$ be a parametric family dominated by a measure $I\!P_0$. Let also $L(\boldsymbol{\theta})$ be the corresponding log-likelihood, i.e. $L(\boldsymbol{\theta}) = \log dI\!P_{\boldsymbol{\theta}}/dI\!P_0$, and $L(\boldsymbol{\theta}',\boldsymbol{\theta}) = L(\boldsymbol{\theta}') - L(\boldsymbol{\theta})$, the log-likelihood ratio. The parametric maximum likelihood estimate (MLE) $\widetilde{\boldsymbol{\theta}}$ is the maximizer of $L(\boldsymbol{\theta})$, $\widetilde{\boldsymbol{\theta}} = \operatorname{argmax}_{\boldsymbol{\theta}\in\Theta} L(\boldsymbol{\theta})$. The main object of the study in this case is the *fitted log-likelihood* $L(\widetilde{\boldsymbol{\theta}},\boldsymbol{\theta}_0) = \sup_{\boldsymbol{\theta}\in\Theta} L(\boldsymbol{\theta},\boldsymbol{\theta}_0)$ where $\boldsymbol{\theta}_0$ is the "true" parameter value.

It is worth mentioning that if the underlying measure $I\!P$ coincides with some $I\!P_{\boldsymbol{\theta}}$ from the given family $\mathcal{P}$ then $-I\!\!E\, L(\boldsymbol{\theta},\boldsymbol{\theta}_0) = -I\!\!E_{\boldsymbol{\theta}_0} L(\boldsymbol{\theta},\boldsymbol{\theta}_0)$ is nothing but the Kullback-Leibler divergence $\mathcal{K}(I\!P_{\boldsymbol{\theta}_0}, I\!P_{\boldsymbol{\theta}})$ between $I\!P_{\boldsymbol{\theta}_0}$ and $I\!P_{\boldsymbol{\theta}}$. It is well known that $\mathcal{K}(I\!P_{\boldsymbol{\theta}_0}, I\!P_{\boldsymbol{\theta}})$ is always non-negative and $\mathcal{K}(I\!P_{\boldsymbol{\theta}_0}, I\!P_{\boldsymbol{\theta}}) = 0$ if and only if the measure $I\!P_{\boldsymbol{\theta}_0}$ and $I\!P_{\boldsymbol{\theta}}$ coincide. This particularly implies the condition (1.1).

2.1.2 Quasi Maximum Likelihood Estimate

Extensions of maximum likelihood principle are often called quasi likelihood estimators. The quasi MLE $\widetilde{\boldsymbol{\theta}}$ minimizes a "quasi" log-likelihood $L(\boldsymbol{\theta})$ which corresponds to some parametric assumption while the real model does not necessarily follow this assumption. This means that the "true" measure $I\!P$ may not belong to the family $(I\!P_{\boldsymbol{\theta}}, \boldsymbol{\theta} \in \Theta)$. Define the target parameter $\boldsymbol{\theta}_0 \in \Theta$ by

$$\boldsymbol{\theta}_0 = \operatorname*{argmax}_{\boldsymbol{\theta} \in \Theta} I\!E L(\boldsymbol{\theta}).$$

The measure $I\!P_{\boldsymbol{\theta}_0}$ can be viewed as the closest to $I\!P$ from the given parametric family and $\widetilde{\boldsymbol{\theta}}$ is the empirical counterpart of $\boldsymbol{\theta}_0$. We will see that $\widetilde{\boldsymbol{\theta}}$ indeed estimates $\boldsymbol{\theta}_0$.

Typical example is given by the regression model

$$Y_i = f(X_i, \boldsymbol{\theta}) + \varepsilon_i.$$

If the errors ε_i are i.i.d. zero mean Gaussian random variables, then the MLE $\widetilde{\boldsymbol{\theta}}$ coincides with the *least squares estimate*, and it minimizes the *least squares contrast* $L(\boldsymbol{\theta}) = \sum_i [Y_i - f(X_i, \boldsymbol{\theta})]^2$. If we only assume that the errors are with mean zero and finite variance, then the least squares estimate can be treated within the quasi maximum likelihood approach. Similarly, the *least absolute deviation* contrast $L(\boldsymbol{\theta}) = \sum_i |Y_i - f(X_i, \boldsymbol{\theta})|$ can be interpreted as the quasi maximum likelihood for the regression with the Laplace distribution with the density $p(y) = e^{-|y|}/2$.

2.1.3 Partial Maximum Likelihood Estimate

Finally, we mention the partial likelihood approach when the parameter of interest $\boldsymbol{\theta}$ is estimated under the presence of a nuisance parameter η. Since the likelihood function depends on the both parameters, we write it in the form $L(\boldsymbol{\theta}; \eta)$. The full likelihood approach means that the likelihood is optimized w.r.t. the couple $(\boldsymbol{\theta}, \eta)$. Equivalently, one first optimize $L(\boldsymbol{\theta}; \eta)$ w.r.t. η leading to $L^*(\boldsymbol{\theta}) = \sup_\eta L(\boldsymbol{\theta}; \eta)$. Then $\widetilde{\boldsymbol{\theta}}$ is the point of maximum of the new contrast $L^*(\boldsymbol{\theta})$.

However, partial optimizing of $L(\boldsymbol{\theta}; \eta)$ w.r.t. η may be a difficult problem, especially if the nuisance parameter is high dimensional. A typical approach to overcome this difficulty is to make use of a simple pilot estimate $\widetilde{\eta}$ of η and then to optimize the *partial* likelihood $L(\boldsymbol{\theta}; \widetilde{\eta})$ w.r.t. the target parameter $\boldsymbol{\theta}$. Note also that the partial like-

lihood estimation when $L(\boldsymbol{\theta};\eta)$ is only optimized w.r.t. $\boldsymbol{\theta}$ with a probably misspecified parameter η is an important special case of the quasi likelihood approach.

2.2 Auxiliary notations and definitions

To proceed with the statistical analysis of the minimum contrast estimates, we need some additional notations and definitions. For any $\boldsymbol{\theta} \in \Theta$ denote

$$M(\boldsymbol{\theta}, \boldsymbol{\theta}_0) \stackrel{\mathrm{def}}{=} -I\!\!E\, L(\boldsymbol{\theta}, \boldsymbol{\theta}_0),$$

where $\boldsymbol{\theta}_0$ is the maximizer of $I\!\!E L(\boldsymbol{\theta})$. It is clear that $M(\boldsymbol{\theta}, \boldsymbol{\theta}_0) \geq 0$ and $M(\boldsymbol{\theta}_0, \boldsymbol{\theta}_0) = 0$. Define also for any $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \Theta$

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

The basic assumption we make is that the increments $L(\boldsymbol{\theta}, \boldsymbol{\theta}')$, or equivalently, their stochastic components $\zeta(\boldsymbol{\theta}, \boldsymbol{\theta}')$ have exponential moments. Define

$$\mathfrak{N}(\mu, \boldsymbol{\theta}, \boldsymbol{\theta}') \stackrel{\mathrm{def}}{=} \log I\!\!E \exp\bigl\{\mu\zeta(\boldsymbol{\theta}, \boldsymbol{\theta}')\bigr\}. \tag{2.1}$$

Notice that $\mathfrak{N}(\mu, \boldsymbol{\theta}, \boldsymbol{\theta}')$ may be equal to infinity for some $\mu > 0$, but it is assumed that for every pair $\boldsymbol{\theta}, \boldsymbol{\theta}'$ there exists $\mu > 0$ for which $\mathfrak{N}(\mu, \boldsymbol{\theta}, \boldsymbol{\theta}')$ is finite. Moreover we assume that the following condition holds

($\boldsymbol{EG}$) *For any $\boldsymbol{\theta} \in \Theta$ there exists a non-empty set $\Upsilon(\boldsymbol{\theta})$ in $(0, \infty)$, such that*

$$\mathfrak{N}(\mu, \boldsymbol{\theta}, \boldsymbol{\theta}_0) < \infty, \qquad \mu \in \Upsilon(\boldsymbol{\theta}).$$

For $\mu \in \Upsilon(\boldsymbol{\theta})$ denote

$$\mathfrak{M}(\mu, \boldsymbol{\theta}, \boldsymbol{\theta}_0) \stackrel{\mathrm{def}}{=} -\log I\!\!E \exp\bigl\{\mu L(\boldsymbol{\theta}, \boldsymbol{\theta}_0)\bigr\} = \mu M(\boldsymbol{\theta}, \boldsymbol{\theta}_0) - \mathfrak{N}(\mu, \boldsymbol{\theta}, \boldsymbol{\theta}_0).$$

Under the condition (EG) we can define optimal $\mu^*(\boldsymbol{\theta}, \boldsymbol{\theta}_0)$ by

$$\mu^*(\boldsymbol{\theta}, \boldsymbol{\theta}_0) \stackrel{\mathrm{def}}{=} \operatorname*{argmax}_{\mu} \mathfrak{M}(\mu, \boldsymbol{\theta}, \boldsymbol{\theta}_0) = \operatorname*{argmax}_{\mu} \{\mu M(\boldsymbol{\theta}, \boldsymbol{\theta}_0) - \mathfrak{N}(\mu, \boldsymbol{\theta}, \boldsymbol{\theta}_0)\}. \tag{2.2}$$

Denote for brevity $\mathfrak{M}^*(\boldsymbol{\theta}, \boldsymbol{\theta}_0) = \mathfrak{M}(\mu^*(\boldsymbol{\theta}, \boldsymbol{\theta}_0), \boldsymbol{\theta}, \boldsymbol{\theta}_0)$. Usually the functions $\mu^*(\boldsymbol{\theta}, \boldsymbol{\theta}_0)$ and $\mathfrak{M}^*(\boldsymbol{\theta}, \boldsymbol{\theta}_0)$ can be easily evaluated in a small neighborhood of the target parameter $\boldsymbol{\theta}_0$. However, it might be difficult to compute them for all $\boldsymbol{\theta} \in \Theta$. Therefore sometimes it is more convenient to deal with another function $\mu(\boldsymbol{\theta}, \boldsymbol{\theta}_0)$, which can be viewed as a

rough approximation of $\mu^*(\boldsymbol{\theta},\boldsymbol{\theta}_0)$. Section 6 provides some examples. Let $\mu(\boldsymbol{\theta},\boldsymbol{\theta}_0)$ be a given function such that $\mu(\boldsymbol{\theta},\boldsymbol{\theta}_0) \in \Upsilon(\boldsymbol{\theta})$. Then we can define

$$\begin{aligned}\mathfrak{N}(\boldsymbol{\theta},\boldsymbol{\theta}_0) &\stackrel{\text{def}}{=} \mathfrak{N}(\mu(\boldsymbol{\theta},\boldsymbol{\theta}_0),\boldsymbol{\theta},\boldsymbol{\theta}_0),\\ \mathfrak{M}(\boldsymbol{\theta},\boldsymbol{\theta}_0) &\stackrel{\text{def}}{=} -\log I\!\!E \exp\bigl\{\mu(\boldsymbol{\theta},\boldsymbol{\theta}_0)L(\boldsymbol{\theta},\boldsymbol{\theta}_0)\bigr\} = \mu(\boldsymbol{\theta},\boldsymbol{\theta}_0)M(\boldsymbol{\theta},\boldsymbol{\theta}_0) - \mathfrak{N}(\boldsymbol{\theta},\boldsymbol{\theta}_0).\end{aligned}$$

The most important requirement on $\mu(\boldsymbol{\theta},\boldsymbol{\theta}_0)$ is that $\mathfrak{M}(\boldsymbol{\theta},\boldsymbol{\theta}_0)$ is positive and grows to infinity as $\boldsymbol{\theta}$ moves away from $\boldsymbol{\theta}_0$. With these notations, the following identity holds for any given $\boldsymbol{\theta} \in \Theta$:

$$I\!\!E \exp\Bigl\{\mu(\boldsymbol{\theta},\boldsymbol{\theta}_0)L(\boldsymbol{\theta},\boldsymbol{\theta}_0) + \mathfrak{M}(\boldsymbol{\theta},\boldsymbol{\theta}_0)\Bigr\} = 1. \tag{2.3}$$

This means that the random variable $\mu(\boldsymbol{\theta},\boldsymbol{\theta}_0)L(\boldsymbol{\theta},\boldsymbol{\theta}_0)+\mathfrak{M}(\boldsymbol{\theta},\boldsymbol{\theta}_0)$ has bounded exponential moments for every $\boldsymbol{\theta}$. Our main goal in this paper is to establish a similar fact for the supremum of this function in $\boldsymbol{\theta} \in \Theta$. More precisely, we are interested in bounding the following function

$$\mathfrak{Q}(\rho,s) \stackrel{\text{def}}{=} I\!\!E \sup_{\boldsymbol{\theta}\in\Theta} \exp\Bigl\{\rho\bigl[\mu(\boldsymbol{\theta},\boldsymbol{\theta}_0)L(\boldsymbol{\theta},\boldsymbol{\theta}_0) + s\mathfrak{M}(\boldsymbol{\theta},\boldsymbol{\theta}_0)\bigr]\Bigr\}, \tag{2.4}$$

where $\rho, s \in [0,1)$.

2.3 The discrete case. A rough bound

We begin with an upper bound, which can be viewed as a simple corollary of (2.3).

Theorem 2.1. *Assume* (EG) *and let* Θ *be a discrete set, then*

$$\mathfrak{Q}(\rho,s) \le \sum_{\boldsymbol{\theta}\in\Theta} \exp\bigl\{-\rho(1-s)\mathfrak{M}(\boldsymbol{\theta},\boldsymbol{\theta}_0)\bigr\}. \tag{2.5}$$

Proof. Obviously

$$\sup_{\boldsymbol{\theta}\in\Theta} \exp\Bigl\{\rho\bigl[\mu(\boldsymbol{\theta},\boldsymbol{\theta}_0)L(\boldsymbol{\theta},\boldsymbol{\theta}_0) + s\mathfrak{M}(\boldsymbol{\theta},\boldsymbol{\theta}_0)\bigr]\Bigr\} \le \sum_{\boldsymbol{\theta}\in\Theta} \exp\Bigl\{\rho\bigl[\mu(\boldsymbol{\theta},\boldsymbol{\theta}_0)L(\boldsymbol{\theta},\boldsymbol{\theta}_0) + s\mathfrak{M}(\boldsymbol{\theta},\boldsymbol{\theta}_0)\bigr]\Bigr\}$$

and therefore

$$\mathfrak{Q}(\rho,s) \le \sum_{\boldsymbol{\theta}\in\Theta} I\!\!E \exp\Bigl\{\rho\bigl[\mu(\boldsymbol{\theta},\boldsymbol{\theta}_0)L(\boldsymbol{\theta},\boldsymbol{\theta}_0) + s\mathfrak{M}(\boldsymbol{\theta},\boldsymbol{\theta}_0)\bigr]\Bigr\}. \tag{2.6}$$

Now combining (2.3) with the Jensen inequality yields

$$I\!\!E \exp\Bigl\{\rho\bigl[\mu(\boldsymbol{\theta},\boldsymbol{\theta}_0)L(\boldsymbol{\theta},\boldsymbol{\theta}_0) + s\mathfrak{M}(\boldsymbol{\theta},\boldsymbol{\theta}_0)\bigr]\Bigr\} \le \exp\Bigl\{-\rho(1-s)\mathfrak{M}(\boldsymbol{\theta},\boldsymbol{\theta}_0)\Bigr\}$$

and substituting this in (2.6) completes the proof. □

Although Theorem 2.1 is a rather simple corollary of (2.3), the bound (2.5) yields a number of useful statistical corollaries. Some of them are presented in Section 2.5. Note however, that even in the discrete case, this bound may be too rough (see the example in Section 6.4). It is also clear that (2.5) has no sense in the continuous case. The next section demonstrates how the bound (2.5) can be extended to the case of an arbitrary parameter set.

2.4 The general exponential bound

The main idea of extending (2.5) is to evaluate the supremum of the contrast over the whole parameter set Θ by a supremum over a discrete ϵ-net $\mathcal{D}$ plus an extra term which controls the local fluctuations of the process $L(\boldsymbol{\theta})$.

Usually the local properties of the centered contrast $\zeta(\boldsymbol{\theta}) = L(\boldsymbol{\theta}) - I\!\!E L(\boldsymbol{\theta})$ are controlled by the variance $\mathfrak{V}^2(\boldsymbol{\theta},\boldsymbol{\theta}') = \operatorname{Var} L(\boldsymbol{\theta},\boldsymbol{\theta}')$, which defines a semi-metric on Θ see, e.g. [12]. However, in some cases, it is more convenient to deal with a slightly different metric which we denote $\mathfrak{S}(\boldsymbol{\theta},\boldsymbol{\theta}')$. This metric usually bounds the variance $\operatorname{Var} L(\boldsymbol{\theta},\boldsymbol{\theta}')$ from above and helps to control the local behavior of the function $\mathfrak{N}(\mu,\boldsymbol{\theta},\boldsymbol{\theta}')$. Sections 3 and 5 present some typical examples of choosing this metric. Below we assume that the metric $\mathfrak{S}(\cdot,\cdot)$ is fixed and defines for every $\boldsymbol{\theta}^\circ \in \Theta$ and every $\epsilon > 0$ a ball

$$B(\epsilon,\boldsymbol{\theta}^\circ) = \bigl\{\boldsymbol{\theta} : \mathfrak{S}(\boldsymbol{\theta},\boldsymbol{\theta}^\circ) \le \epsilon\bigr\}$$

with the center $\boldsymbol{\theta}^\circ \in \Theta$ and the radius ϵ. We say that a discrete set $\mathcal{D}$ is an ϵ-net in Θ, if

$$\Theta \subset \bigcup_{\boldsymbol{\theta}^\circ \in \mathcal{D}} B(\epsilon,\boldsymbol{\theta}^\circ). \tag{2.7}$$

To control of local fluctuations of the contrast process $L(\boldsymbol{\theta})$ within the balls $B(\epsilon,\boldsymbol{\theta}^\circ)$, we impose the following condition:

$(\boldsymbol{EL})$ *There exist $\epsilon > 0$ and a function $\varkappa(\lambda)$ such that for any $\boldsymbol{\theta}^\circ \in \Theta$*

$$\sup_{\boldsymbol{\theta},\boldsymbol{\theta}' \in B(\epsilon,\boldsymbol{\theta}^\circ)} \mathfrak{N}\Bigl(\frac{2\lambda}{\mathfrak{S}(\boldsymbol{\theta},\boldsymbol{\theta}')},\boldsymbol{\theta},\boldsymbol{\theta}'\Bigr) = \sup_{\boldsymbol{\theta},\boldsymbol{\theta}' \in B(\epsilon,\boldsymbol{\theta}^\circ)} \log I\!\!E \exp\bigl\{2\lambda\xi(\boldsymbol{\theta},\boldsymbol{\theta}')\bigr\} \le \varkappa(\lambda), \tag{2.8}$$

where

$$\xi(\boldsymbol{\theta},\boldsymbol{\theta}') \stackrel{\mathrm{def}}{=} \frac{L(\boldsymbol{\theta},\boldsymbol{\theta}') - I\!\!E L(\boldsymbol{\theta},\boldsymbol{\theta}')}{\mathfrak{S}(\boldsymbol{\theta},\boldsymbol{\theta}')} = \frac{\zeta(\boldsymbol{\theta},\boldsymbol{\theta}')}{\mathfrak{S}(\boldsymbol{\theta},\boldsymbol{\theta}')}\,.$$

By $\mathbb{N}_{\boldsymbol{\theta}^\circ}(\epsilon')$ for $\epsilon' \le \epsilon$ we denote the "local" covering number defined as the minimal number of balls $B(\epsilon', \cdot)$ required to cover the ball $B(\epsilon, \boldsymbol{\theta}^\circ)$. With this covering number we associate the local entropy

$$\mathfrak{e}(\boldsymbol{\theta}^\circ) \stackrel{\mathrm{def}}{=} \sum_{k=1}^{\infty} 2^{-k} \log \mathbb{N}_{\boldsymbol{\theta}^\circ}(2^{-k}\epsilon).$$

The next theorem generalizes the upper bound from Theorem 2.1.

Theorem 2.2. *Assume the condition* (EL), (EG), *and (2.7). Then*

$$\mathfrak{Q}(\rho, s) \le \sum_{\boldsymbol{\theta}^\circ \in \mathcal{D}} \exp\biggl\{ -\rho(1-s)\breve{\mathfrak{M}}(\boldsymbol{\theta}^\circ, \boldsymbol{\theta}_0) + (1-\rho)\biggl[\varkappa\biggl(\frac{\rho \overline{\mu}(\boldsymbol{\theta}^\circ)\epsilon}{1-\rho} \biggr) + \mathfrak{e}(\boldsymbol{\theta}^\circ) \biggr] \biggr\},$$

with

$$\overline{\mu}(\boldsymbol{\theta}^\circ) = \sup_{\boldsymbol{\theta} \in B(\epsilon, \boldsymbol{\theta}^\circ)} \mu(\boldsymbol{\theta}, \boldsymbol{\theta}_0), \qquad \breve{\mathfrak{M}}(\boldsymbol{\theta}^\circ, \boldsymbol{\theta}_0) = \inf_{\boldsymbol{\theta} \in B(\epsilon, \boldsymbol{\theta}^\circ)} \mathfrak{M}(\boldsymbol{\theta}, \boldsymbol{\theta}_0). \tag{2.9}$$

Compared with Theorem 2.1, the above exponential bound contains two additional terms: $\mathfrak{e}(\boldsymbol{\theta}^\circ)$ controls the local entropy of $B(\epsilon, \boldsymbol{\theta}^\circ)$, and $\varkappa(\cdot)$ bounds the local fluctuations of the contrast process over the ball $B(\epsilon, \boldsymbol{\theta}^\circ)$.

We also present another version of Theorem 2.2 which involves the standard chaining construction. Consider the sets

$$\mathcal{A}(r, \boldsymbol{\theta}_0) = \{ \boldsymbol{\theta} : \mathfrak{M}(\boldsymbol{\theta}, \boldsymbol{\theta}_0) \le r \}, \qquad r \ge 0. \tag{2.10}$$

In the next theorem we use a growing sequence of radii r_k, $k = 1, 2, \ldots$ such that $\lim_{n \to \infty} r_k = \infty$. A standard example is given by $r_k = 2^{k-1} r_1$. We also set for convenience $r_0 = 0$. Then the large deviations of the contrast $L(\boldsymbol{\theta})$ may be controlled separately for every set

$$\mathcal{C}_k = \mathcal{A}(r_{k+1}, \boldsymbol{\theta}_0) \setminus \mathcal{A}(r_k, \boldsymbol{\theta}_0). \tag{2.11}$$

Let $\mathbb{N}(\mathcal{C}_k)$ be the minimal numbers of balls $B(\epsilon, \boldsymbol{\theta}^\circ)$ from (2.7) required to cover $\mathcal{C}_k$. This means that there exists a discrete set $\mathcal{D}_k$ of cardinality at most $\mathbb{N}(\mathcal{C}_k)$ such that $\mathcal{C}_k \subseteq \bigcup_{\boldsymbol{\theta}^\circ \in \mathcal{D}_k} B(\epsilon, \boldsymbol{\theta}^\circ)$.

Theorem 2.3. *Under the conditions of Theorem 2.2 it holds*

$$\mathfrak{Q}(\rho, s) \le \sum_{k=0}^{\infty} \mathbb{N}(\mathcal{C}_k) \exp\biggl\{ -\rho(1-s) r_k + (1-\rho)\biggl[\varkappa\biggl(\frac{\rho \overline{\mu}_k \epsilon}{1-\rho} \biggr) + \mathfrak{e}_k \biggr] \biggr\},$$

with

$$\overline{\mu}_k = \sup_{\boldsymbol{\theta} \in \mathcal{C}_k} \mu(\boldsymbol{\theta}, \boldsymbol{\theta}_0), \quad \mathfrak{e}_k = \sup_{\boldsymbol{\theta} \in \mathcal{C}_k} \mathfrak{e}(\boldsymbol{\theta}).$$

This theorem can be proved by bounding $\mathfrak{Q}(\rho, s)$ from Theorem 2.2 on every coronae set $\mathcal{C}_k$ and using that $\mathfrak{M}(\boldsymbol{\theta}, \boldsymbol{\theta}_0) \geq r_k$ on $\mathcal{C}_k$.

2.5 Some corollaries

In this section, we demonstrate how the exponential bounds from Theorems 2.1–2.3 can be used in analyzing statistical performance of the minimum contrast estimator $\widetilde{\boldsymbol{\theta}} = \operatorname{argmax}_{\boldsymbol{\theta} \in \Theta} L(\boldsymbol{\theta})$. It obviously follows from the definition (2.4) of $\mathfrak{Q}(\rho, s)$ that

$$\mathbb{E} \exp\Big\{ \rho \big[\mu(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}_0) L(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}_0) + s \mathfrak{M}(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}_0) \big] \Big\} \leq \mathfrak{Q}(\rho, s), \quad 0 \leq \rho, s < 1. \tag{2.12}$$

Corollary 2.4. *For any* $\rho, s < 1$

$$\mathbb{E} \exp\Big\{ \rho \mu(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}_0) L(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}_0) \Big\} \leq \mathfrak{Q}(\rho, 0), \tag{2.13}$$

$$\mathbb{E} \exp\Big\{ \rho s \, \mathfrak{M}(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}_0) \Big\} \leq \mathfrak{Q}(\rho, s). \tag{2.14}$$

Proof. Substituting $s = 0$ in (2.12) yields the first bound. To prove the second inequality, notice that $L(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}_0) \geq 0$. Therefore the elementary inequality $\mathbf{1}\{x \geq 0\} \leq \exp(\mu x)$ for any $\mu > 0$ yields (see also (2.12))

$$\begin{aligned} \mathbb{E} \exp\big\{ \rho s \, \mathfrak{M}(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}_0) \big\} &= \mathbb{E} \exp\big\{ \rho s \, \mathfrak{M}(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}_0) \big\} \mathbf{1}\big\{ L(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}_0) \geq 0 \big\} \\ &\leq \mathbb{E} \exp\big\{ \rho s \, \mathfrak{M}(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}_0) + \rho \mu(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}_0) L(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}_0) \big\} \leq \mathfrak{Q}(\rho, s). \end{aligned}$$

□

The second assertion of Corollary 2.4 presents an exponential risk bound for the estimate $\widetilde{\boldsymbol{\theta}}$ for the "natural" loss function $\mathfrak{M}(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}_0)$. Clearly, the exponential bound (2.14) implies any polynomial moments of the loss $\big| \mathfrak{M}(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}_0) \big|^r$, see Lemma 7.5 for a precise formulation.

The assertion (2.13) can be used for establishing the concentration property of the estimate $\widetilde{\boldsymbol{\theta}}$. Consider the sets $\mathcal{A}(r, \boldsymbol{\theta}_0)$ from (2.10). The next result shows that the estimate $\widetilde{\boldsymbol{\theta}}$ deviates out of the set $\mathcal{A}(r, \boldsymbol{\theta}_0)$ for some $r > 0$ with the exponentially small probability of order $\exp\{-\rho s \, r\}$.

Corollary 2.5. *For any* $\rho, s < 1$, *it holds*

$$\mathbb{P}\Big(\widetilde{\boldsymbol{\theta}} \notin \mathcal{A}(r, \boldsymbol{\theta}_0) \Big) \leq \mathfrak{Q}(\rho, s) \exp\{-\rho s \, r\}.$$

Proof. Since $L(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}_0) \geq 0$ and $\mathfrak{M}(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}_0) \geq r$ for $\widetilde{\boldsymbol{\theta}} \notin \mathcal{A}(r, \boldsymbol{\theta}_0)$, we obtain

$$\mathrm{e}^{\rho s\, r} \mathbf{1}\Big(\widetilde{\boldsymbol{\theta}} \notin \mathcal{A}(r, \boldsymbol{\theta}_0)\Big) \leq \exp\Big\{ \rho \big[\mu(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}_0) L\big(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}_0\big) + s \mathfrak{M}(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}_0) \big] \Big\} \leq \mathfrak{Q}(\rho, s).$$

□

Finally we discuss confidence sets for the target $\boldsymbol{\theta}_0$. Since the inequality (2.13) claims that $L(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}_0)$ is bounded with exponential moments, we can use the following confidence set:

$$\mathcal{E}(\mathfrak{z}) = \big\{ \boldsymbol{\theta} \in \Theta : L(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}) \leq \mathfrak{z} \big\}.$$

To evaluate the covering probability, consider first the case when $\mu(\boldsymbol{\theta}, \boldsymbol{\theta}_0) \geq \mu_* > 0$ uniformly in $\boldsymbol{\theta} \in \Theta$.

Corollary 2.6. *Assume that* $\mu(\boldsymbol{\theta}, \boldsymbol{\theta}_0) \geq \mu_* > 0$. *Then for any* $\mathfrak{z} > 0$ *and any* $\rho < 1$

$$I\!P\big(\boldsymbol{\theta}_0 \notin \mathcal{E}(\mathfrak{z})\big) \leq \mathfrak{Q}(\rho, 0) \exp\big\{ -\rho \mu_* \mathfrak{z} \big\}.$$

Proof. Now the bound (2.13) and the exponential Chernov inequality imply

$$\begin{aligned} I\!P\big(\boldsymbol{\theta}_0 \notin \mathcal{E}(\mathfrak{z})\big) &= I\!P\big(L(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}) > \mathfrak{z}\big) \\ &\leq I\!E \exp\big\{ -\rho \mu(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}_0) \mathfrak{z} \big\} \exp\big\{ \rho \mu(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}_0) L(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}) \big\} \\ &\leq \exp\big\{ -\rho \mu_* \mathfrak{z} \big\} I\!E \exp\big\{ \rho \mu(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}_0) L(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}) \big\} \\ &\leq \mathfrak{Q}(\rho, 0) \exp\big\{ -\rho \mu_* \mathfrak{z} \big\} \end{aligned}$$

as required. □

In the case when the function $\mu(\boldsymbol{\theta}, \boldsymbol{\theta}_0)$ cannot be uniformly bounded from below by a positive constant, we assume that such a bound exists for every set $\mathcal{A}(r, \boldsymbol{\theta}_0)$. Denote

$$\mu_*(r) \stackrel{\mathrm{def}}{=} \inf_{\boldsymbol{\theta} \in \mathcal{A}(r, \boldsymbol{\theta}_0)} \mu(\boldsymbol{\theta}, \boldsymbol{\theta}_0).$$

Then

$$I\!P\big(\boldsymbol{\theta}_0 \notin \mathcal{E}(\mathfrak{z})\big) \leq I\!P\big(\boldsymbol{\theta}_0 \notin \mathcal{E}(\mathfrak{z}), \widetilde{\boldsymbol{\theta}} \in \mathcal{A}(r, \boldsymbol{\theta}_0)\big) + I\!P\big(\widetilde{\boldsymbol{\theta}} \notin \mathcal{A}(r, \boldsymbol{\theta}_0)\big)$$

and combining Corollaries 2.5–2.6 results in

Corollary 2.7. *For any* $\mathfrak{z} > 0$ *and any* $\rho, s < 1$

$$I\!P\big(\boldsymbol{\theta}_0 \notin \mathcal{E}(\mathfrak{z})\big) \leq \mathfrak{Q}(\rho, 0) \exp\big\{ -\rho \mu_*(r) \mathfrak{z} \big\} + \mathfrak{Q}(\rho, s) \exp\big\{ -\rho s\, r \big\}.$$

3 Exponential bounds for smooth contrasts

This section deals with the case when the contrast $L(\boldsymbol{\theta})$ is a smooth function of $\boldsymbol{\theta}$. In this situation, we show that the local condition (EL) is easy to verify. Moreover, the local balls $B(\epsilon, \boldsymbol{\theta})$ nearly coincide with the usual Euclidean ellipsoids and the local entropy can be easily bounded by an absolute constant only depending on the dimensionality p of Θ. In addition, one can avoid computing the covering numbers by replacing the sum over an ϵ-net by an integral.

Suppose that Θ is a convex set in $\mathbb{R}^p$ and the function $L(\boldsymbol{\theta})$ is differentiable w.r.t. $\boldsymbol{\theta}$. Below, $\nabla L(\boldsymbol{\theta})$ stands for the gradient of $L(\boldsymbol{\theta})$. Define

$$V(\boldsymbol{\theta}) \stackrel{\mathrm{def}}{=} \mathbb{E} \nabla \zeta(\boldsymbol{\theta}) \bigl[\nabla \zeta(\boldsymbol{\theta})\bigr]^{\top} \tag{3.1}$$

and

$$H(\lambda, \gamma, \boldsymbol{\theta}) \stackrel{\mathrm{def}}{=} \log \mathbb{E} \exp\biggl\{ 2\lambda \frac{\gamma^{\top} \nabla \zeta(\boldsymbol{\theta})}{\sqrt{\gamma^{\top} V(\boldsymbol{\theta}) \gamma}} \biggr\}. \tag{3.2}$$

It is easy to see that $H(0, \gamma, \boldsymbol{\theta}) = 0$, $\partial H(0, \gamma, \boldsymbol{\theta}) / \partial \lambda = 0$, and

$$\frac{\partial^2 H(\lambda, \gamma, \boldsymbol{\theta})}{\partial^2 \lambda}\biggr|_{\lambda=0} = \frac{4 \gamma^{\top} \operatorname{Var}\bigl\{\nabla L(\boldsymbol{\theta})\bigr\} \gamma}{\gamma^{\top} V(\boldsymbol{\theta}) \gamma} = 4.$$

So, $H(\lambda, \gamma, \boldsymbol{\theta}) \approx 2\lambda^2$ for small λ. This property justifies the following assumption:

$(\boldsymbol{ED})$ *There exists $\overline{\lambda} > 0$ such that for some $\nu_0 \geq 1$ uniformly in $\boldsymbol{\theta} \in \Theta$*

$$\sup_{|\lambda| \leq \overline{\lambda}} \sup_{\gamma \in \mathcal{S}^p} \lambda^{-2} H(\lambda, \gamma, \boldsymbol{\theta}) \leq 2\nu_0^2. \tag{3.3}$$

In typical situations the matrix $V(\boldsymbol{\theta})$ is rather large (proportional to the sample size, see Section 5). However, in some cases there are singular points $\boldsymbol{\theta}^{\circ} \in \Theta$ such that $V(\boldsymbol{\theta})$ becomes nearly degenerate when $\boldsymbol{\theta}$ approaches $\boldsymbol{\theta}^{\circ}$. In such cases the condition (3.3) is difficult to check and it is reasonable to replace the matrix $V(\boldsymbol{\theta})$ in this condition by its "regularization", see Section 6.3 for an example. Without loss of generality we assume that this "regularized" matrix $V(\cdot)$ satisfies $V(\boldsymbol{\theta}) \geq s_0 I$ for some $s_0 > 0$ which means that $\gamma^{\top} V(\boldsymbol{\theta}) \gamma \geq s_0$ for all unit vectors $\gamma \in \mathbb{R}^d$, or equivalently, the smallest eigenvalue $\lambda_{\min}[V(\boldsymbol{\theta})]$ is not smaller than s_0 for all $\boldsymbol{\theta} \in \Theta$.

Now we define the metric $\mathfrak{S}(\boldsymbol{\theta}, \boldsymbol{\theta}')$ by

$$\mathfrak{S}^2(\boldsymbol{\theta}, \boldsymbol{\theta}') \stackrel{\mathrm{def}}{=} \sup_{t \in [0,1]} (\boldsymbol{\theta} - \boldsymbol{\theta}')^{\top} V\bigl[(1-t)\boldsymbol{\theta}' + t\boldsymbol{\theta}\bigr] (\boldsymbol{\theta} - \boldsymbol{\theta}'). \tag{3.4}$$

Define also for every $\boldsymbol{\theta}^\circ \in \Theta$ the ellipsoid $B'(\epsilon, \boldsymbol{\theta}^\circ)$ and the Euclidean ball $B^*(\epsilon, \boldsymbol{\theta}^\circ)$ as

$$\begin{aligned} B'(\epsilon, \boldsymbol{\theta}^\circ) &= \Big\{ \boldsymbol{\theta} : (\boldsymbol{\theta} - \boldsymbol{\theta}^\circ)^\top V(\boldsymbol{\theta}^\circ) (\boldsymbol{\theta} - \boldsymbol{\theta}^\circ) \le \epsilon^2 \Big\}, \\ B^*(\epsilon, \boldsymbol{\theta}^\circ) &= \Big\{ \boldsymbol{\theta} : \| \boldsymbol{\theta} - \boldsymbol{\theta}^\circ \|^2 \le \epsilon^2 / \lambda_{\min}[V(\boldsymbol{\theta}^\circ)] \Big\}. \end{aligned}$$

Obviously $B(\epsilon, \boldsymbol{\theta}^\circ) \subseteq B'(\epsilon, \boldsymbol{\theta}^\circ)$ and $B^*(\epsilon, \boldsymbol{\theta}^\circ)$ is the smallest Euclidean ball containing $B'(\epsilon, \boldsymbol{\theta}^\circ)$.

In what follows, we assume that the radius ϵ can be chosen in such a way that the functions $V(\boldsymbol{\theta})$ and $\mathfrak{M}(\boldsymbol{\theta}, \boldsymbol{\theta}_0)$ have bounded fluctuations within the ball $B^*(\epsilon, \boldsymbol{\theta}^\circ)$ for every $\boldsymbol{\theta}^\circ \in \Theta$. More precisely, for a given function $f(\cdot)$ define its magnitude over $B^*(\epsilon, \boldsymbol{\theta}^\circ)$ by

$$\mathfrak{A}_\epsilon f(\boldsymbol{\theta}^\circ) \stackrel{\text{def}}{=} \sup_{\boldsymbol{\theta}, \boldsymbol{\theta}' \in B^*(\epsilon, \boldsymbol{\theta}^\circ)} \frac{f(\boldsymbol{\theta})}{f(\boldsymbol{\theta}')}.$$

Similarly, the magnitude of the matrix $V(\boldsymbol{\theta})$ over $B^*(\epsilon, \boldsymbol{\theta}^\circ)$ is computed as follows

$$\mathfrak{A}_\epsilon V(\boldsymbol{\theta}^\circ) \stackrel{\text{def}}{=} \sup_{\boldsymbol{\theta}, \boldsymbol{\theta}' \in B^*(\epsilon, \boldsymbol{\theta}^\circ)} \sup_{\gamma \in S^d} \frac{\gamma^\top V(\boldsymbol{\theta}) \gamma}{\gamma^\top V(\boldsymbol{\theta}') \gamma}.$$

Notice that under the condition $\mathfrak{A}_\epsilon V(\cdot) \le \nu_1$, the topology induced by the metric $\mathfrak{S}(\cdot, \cdot)$ is (locally) equivalent to the Euclidean topology and the set $B(\epsilon, \boldsymbol{\theta}^\circ)$ can be well approximated by the ellipsoid $B'(\epsilon, \boldsymbol{\theta}^\circ)$. This yields that computing the local entropy $\mathfrak{e}(\cdot)$ can be reduced to the Euclidean case, see Lemma 7.3 for more detail.

Now we are ready to state a result about deviations of the contrast process in the smooth case. For simplicity we assume that $\mu(\boldsymbol{\theta}, \boldsymbol{\theta}_0) \le \overline{\mu}$ for all $\boldsymbol{\theta}$ and some given $\overline{\mu}$.

Theorem 3.1. *Assume (EG) and (ED) with for some ν_0 and $\overline{\lambda} > 0$. Let also $\mu(\boldsymbol{\theta}, \boldsymbol{\theta}_0) \le \overline{\mu}$. Suppose that there is a constant $\epsilon^* \le \min\{\overline{\lambda}, \sqrt{2/\nu_0}\}/\overline{\mu}$ such that for a fixed $\nu_1 \ge 1$ and each $\boldsymbol{\theta} \in \Theta$ holds*

$$\mathfrak{A}_{\epsilon^*} V(\boldsymbol{\theta}) \le \nu_1, \quad \mathfrak{A}_{\epsilon^*} \big\{ 1 + \mathfrak{M}(\boldsymbol{\theta}, \boldsymbol{\theta}_0) \big\} \le \nu_1 . \tag{3.5}$$

Then, there exists a constant C_p which depends on p only such that for any $\rho, s < 1$

$$\mathfrak{Q}(\rho, s) \le \frac{C_p \nu_1^{p/2}}{\big| \epsilon^* (1 - \rho) \big|^p} \int_\Theta \exp\big\{ -\nu_1^{-1} \rho (1 - s) \mathfrak{M}(\boldsymbol{\theta}, \boldsymbol{\theta}_0) \big\} \sqrt{\det\big\{ V(\boldsymbol{\theta}) \big\}}\, d\boldsymbol{\theta}. \tag{3.6}$$

The bound (3.6) is only meaningful if the integral in the right hand-side of (3.6) is finite. Fortunately it can be easily bounded in typical situations. For instance, in the region Θ_1 close to $\boldsymbol{\theta}_0$, the function $\mathfrak{M}(\boldsymbol{\theta}, \boldsymbol{\theta}_0)$ is nearly quadratic because it is smooth and

satisfies $\mathfrak{M}(\boldsymbol{\theta}_0, \boldsymbol{\theta}_0) = 0$, $\nabla\mathfrak{M}(\boldsymbol{\theta}_0, \boldsymbol{\theta}_0) = 0$, and the integral in (3.6) is easy to evaluate. When $\boldsymbol{\theta}$ is far away from $\boldsymbol{\theta}_0$, the logarithmic growth rate of the function $\mathfrak{M}(\boldsymbol{\theta}, \boldsymbol{\theta}_0)$ is usually sufficient for bounding the integral. We postpone a precise formulation until Section 5 where the results are specified to the case of the i.i.d. contrast.

4 Exponential bounds for the Gaussian contrasts

Here we provide some general results in the case where the contrast difference $L(\boldsymbol{\theta}, \boldsymbol{\theta}')$ is a Gaussian random variable for any $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \Theta$. Then $L(\boldsymbol{\theta}, \boldsymbol{\theta}') \sim \mathcal{N}\bigl(-M(\boldsymbol{\theta}, \boldsymbol{\theta}'), \mathfrak{V}^2(\boldsymbol{\theta}, \boldsymbol{\theta}')\bigr)$ and $\xi(\boldsymbol{\theta}, \boldsymbol{\theta}') = \bigl\{L(\boldsymbol{\theta}, \boldsymbol{\theta}') - I\!\!E L(\boldsymbol{\theta}, \boldsymbol{\theta}')\bigr\}/\mathfrak{V}(\boldsymbol{\theta}, \boldsymbol{\theta}')$ is $\mathcal{N}(0, 1)$ so, that for any $\boldsymbol{\theta} \in \Theta$

$$\mathfrak{N}(\mu; \boldsymbol{\theta}, \boldsymbol{\theta}_0) = \log I\!\!E \exp\bigl\{\mu\zeta(\boldsymbol{\theta}, \boldsymbol{\theta}_0)\bigr\} = \mu^2\mathfrak{V}^2(\boldsymbol{\theta}, \boldsymbol{\theta}_0)/2$$

and the optimized values $\mu^*(\boldsymbol{\theta}, \boldsymbol{\theta}_0)$, $\mathfrak{M}^*(\boldsymbol{\theta}, \boldsymbol{\theta}_0)$ can be easily computed:

$$\begin{aligned} \mu^*(\boldsymbol{\theta}, \boldsymbol{\theta}_0) &= \operatorname*{argmax}_{\mu \geq 0}\left\{\mu M(\boldsymbol{\theta}, \boldsymbol{\theta}_0) - \mu^2\frac{\mathfrak{V}^2(\boldsymbol{\theta}, \boldsymbol{\theta}_0}{2}\right\} = \frac{M(\boldsymbol{\theta}, \boldsymbol{\theta}_0)}{\mathfrak{V}^2(\boldsymbol{\theta}, \boldsymbol{\theta}_0)}, \\ \mathfrak{M}^*(\boldsymbol{\theta}, \boldsymbol{\theta}_0) &= \frac{M^2(\boldsymbol{\theta}, \boldsymbol{\theta}_0)}{2\mathfrak{V}^2(\boldsymbol{\theta}, \boldsymbol{\theta}_0)}. \end{aligned} \tag{4.1}$$

Moreover, the local condition (EL) is fulfilled globally with $\mathfrak{S}(\boldsymbol{\theta}, \boldsymbol{\theta}') = \mathfrak{V}(\boldsymbol{\theta}, \boldsymbol{\theta}')$ and $\varkappa(\lambda) = 2\lambda^2$ and the condition (ED) is not required. Equivalently one can say that (ED) is fulfilled with $\overline{\lambda} = \infty$. Below we specify Theorem 2.2 to the two important cases: a smooth univariate contrast and the Gaussian one. We assume throughout that $\mu(\boldsymbol{\theta}, \boldsymbol{\theta}_0) = \mu^*(\boldsymbol{\theta}, \boldsymbol{\theta}_0)$ and $\mathfrak{M}(\boldsymbol{\theta}, \boldsymbol{\theta}_0) = \mathfrak{M}(\mu(\boldsymbol{\theta}, \boldsymbol{\theta}_0), \boldsymbol{\theta}, \boldsymbol{\theta}_0)$.

4.1 The univariate case

Let $\theta \in \Theta \subset I\!R^1$. Then for any $\epsilon > 0$ there exists a covering of the parameter set Θ by the non-overlapping local balls $B(\epsilon, \theta)$, that is, every point $\theta \in \Theta$ belongs to the only one ball $B(\epsilon, \theta)$. Let $\pi(d\theta)$ be a σ-finite measure on the parameter set Θ. Denote by $\Pi_\epsilon(\theta)$ the π-measure of the local ball $B(\epsilon, \theta) = \{\theta' : \mathfrak{V}(\theta, \theta') \leq \epsilon\}$.

For simplicity of formulation we assume in the next result that the local entropy $\mathfrak{e}(\theta)$ of every local ball $B(\epsilon, \theta)$ is uniformly bounded by some constant $\overline{\mathfrak{e}}$ and $\mu(\boldsymbol{\theta}, \boldsymbol{\theta}_0) \leq \overline{\mu}$.

Theorem 4.1. *Let $L(\theta)$ be a Gaussian contrast for $\theta \in \Theta \subset I\!R^1$. For given $\rho, s < 1$, $\nu_1 > 1$, choose $\epsilon \leq \overline{\mu}^{-1}\sqrt{(1-\rho)/2}$ such that for any $\theta \in \Theta$:*

$$\mathfrak{A}_\epsilon\bigl\{1 + \mathfrak{M}(\theta, \theta_0)\bigr\} \leq \nu_1, \qquad \mathfrak{A}_\epsilon\Pi_\epsilon(\theta) \leq \nu_1. \tag{4.2}$$

Let also the local entropy $\mathfrak{e}(\theta)$ *be bounded by* $\bar{\mathfrak{e}}$ *for all* θ. *Then*

$$\mathfrak{Q}(\rho, s) \leq \exp\{(1-\rho)\bar{\mathfrak{e}}\} \int_{\Theta} \exp\bigl\{-\nu_1^{-1}\rho(1-s)\mathfrak{M}(\theta, \theta_0)\bigr\} \frac{d\pi(\theta)}{\Pi_\epsilon(\theta)}$$

Proof. Consider a non-overlapping covering of the set Θ by the local balls $B(\theta^\circ)$, $\theta^\circ \in \mathcal{D}$. For any $\theta \in \mathcal{B}(\theta^\circ)$, the use of $\varkappa(\lambda) = 2\lambda^2$, $\overline{\mu}(\boldsymbol{\theta}^\circ) \leq \overline{\mu}$ yields

$$(1-\rho)\varkappa\biggl(\frac{\rho\overline{\mu}(\boldsymbol{\theta}^\circ)\epsilon}{1-\rho}\biggr) \leq \frac{2\rho^2\overline{\mu}^2\epsilon^2}{2(1-\rho)} \leq \rho^2 \leq 1.$$

Now the result follows from Theorem 2.2, see the proof of Theorem 3.1 for more detail. □

4.2 A smooth Gaussian contrast

When the contrast is smooth and Gaussian, Theorem 3.1 reads as follows:

Theorem 4.2. *Let* $L(\boldsymbol{\theta})$ *be a smooth Gaussian contrast,* $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^p$. *Let for some* $\nu_1 \geq 1$ *and* $\epsilon^* \leq 1$ *and any* $\boldsymbol{\theta} \in \Theta$ *hold:*

$$\mathfrak{A}_{\epsilon^*} V(\boldsymbol{\theta}) \leq \nu_1, \quad \mathfrak{A}_{\epsilon^*}\bigl\{1 + \mathfrak{M}(\boldsymbol{\theta}, \boldsymbol{\theta}_0)\bigr\} \leq \nu_1\,. \tag{4.3}$$

Let, in addition, the matrix $V(\boldsymbol{\theta})$ *be non-degenerated for every* $\boldsymbol{\theta} \in \Theta$. *Then there exists a constant* C_p *depending on* p *only and such that for any* $\rho, s < 1$

$$\mathfrak{Q}(\rho, s) \leq \frac{C_p \nu_1^{p/2}}{\epsilon^{*p}(1-\rho)^{p/2}} \int_{\Theta} \exp\bigl\{-\nu_1^{-1}\rho(1-s)\mathfrak{M}(\boldsymbol{\theta}, \boldsymbol{\theta}_0)\bigr\} \sqrt{\det\bigl\{V(\boldsymbol{\theta})\bigr\}}\, d\boldsymbol{\theta}.$$

This integral can be easily bounded by $C(1-\rho)^{-p/2}(1-s)^{-p/2}$ under additional conditions on the growth of the function $\mathfrak{M}(\boldsymbol{\theta}, \boldsymbol{\theta}_0)$, cf. Theorem 5.1.

4.3 MLE for a Gaussian model

If the Gaussian contrast $L(\boldsymbol{\theta})$ coincides, in addition, with the log-likelihood ratio, i.e. $L(\boldsymbol{\theta}) = \log d\mathbb{P}_{\boldsymbol{\theta}}/d\mathbb{P}_{\boldsymbol{\theta}_0}$, the equality $\mathbb{E}\exp\{L(\boldsymbol{\theta})\} = 1$ implies $M(\boldsymbol{\theta}, \boldsymbol{\theta}_0) = \mathfrak{V}^2(\boldsymbol{\theta}, \boldsymbol{\theta}_0)/2$ yielding $\mu^*(\boldsymbol{\theta}, \boldsymbol{\theta}_0) \equiv 1/2$, $\mathfrak{M}^*(\boldsymbol{\theta}, \boldsymbol{\theta}_0) = M(\boldsymbol{\theta}, \boldsymbol{\theta}_0)/4$. Then the exponential bound in the univariate case $\theta \in \Theta \subset \mathbb{R}^1$ has the following form

$$\begin{aligned} &\mathbb{E}\exp\Bigl\{\rho\bigl[L(\widetilde{\theta}, \theta_0)/2 + sM(\widetilde{\theta}, \theta_0)/4\bigr]\Bigr\} \\ &\quad \leq \exp\{(1-\rho)\bar{\mathfrak{e}}\} \int_{\Theta} \exp\Bigl\{-\frac{\rho(1-s)M(\theta, \theta_0)}{4\nu_1}\Bigr\} \frac{d\pi(\theta)}{\Pi_\epsilon(\theta)}. \end{aligned}$$

5 Quasi MLE for i.i.d. data

Let $Y^{(n)} = (Y_1, \ldots, Y_n)$ be an i.i.d. sample with a marginal distribution P. By $I\!P$ we denote the joint distribution of $Y^{(n)}$. Let also $\mathcal{P} = (P_{\boldsymbol{\theta}}, \boldsymbol{\theta} \in \Theta \subset I\!R^p)$ be a parametric family. The parametric hypothesis assumes that $P = P_{\boldsymbol{\theta}_0}$ for some $\boldsymbol{\theta}_0 \in \Theta$ and our goal is to recover this unknown parameter with the help of the data $Y^{(n)}$. The family $(P_{\boldsymbol{\theta}}, \boldsymbol{\theta} \in \Theta)$ and the underlying measure P are assumed dominated by some measure P_0. We denote by $p(y, \boldsymbol{\theta})$ and $p(y)$ the corresponding densities: $p(y, \boldsymbol{\theta}) = dP_{\boldsymbol{\theta}}/dP_0(y)$, $p(y) = dP/dP_0(y)$.

The parametric assumption leads to the estimate $\widetilde{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}_0$ defined by maximizing the corresponding log-likelihood i.e.,

$$\widetilde{\boldsymbol{\theta}} = \operatorname*{argmax}_{\boldsymbol{\theta} \in \Theta} L(\boldsymbol{\theta}) = \operatorname*{argmax}_{\boldsymbol{\theta} \in \Theta} \sum_{i=1}^{n} \ell(Y_i, \boldsymbol{\theta}),$$

where $\ell(Y, \boldsymbol{\theta}) = \log\bigl[p(Y, \boldsymbol{\theta})\bigr]$. In this section we discuss the quality of estimation in the case when the underlying measure P does not necessarily belongs to the parametric family $\mathcal{P}$. We will see that in this case the procedure estimates a point $\boldsymbol{\theta}_0$, which minimizes some special "distance" between P and $P_{\boldsymbol{\theta}}$ over $\boldsymbol{\theta} \in \Theta$.

We begin with some notations. Denote

$$\mathfrak{m}(\mu, \boldsymbol{\theta}, \boldsymbol{\theta}_0) = -\log E \exp\{\mu \ell(Y, \boldsymbol{\theta}, \boldsymbol{\theta}_0)\},$$

where $\ell(Y, \boldsymbol{\theta}, \boldsymbol{\theta}') = \ell(Y, \boldsymbol{\theta}) - \ell(Y, \boldsymbol{\theta}')$. The i.i.d. structure of the Y_i's implies

$$\mathfrak{M}(\mu, \boldsymbol{\theta}, \boldsymbol{\theta}_0) = n\, \mathfrak{m}(\mu, \boldsymbol{\theta}, \boldsymbol{\theta}_0).$$

This allows to define the function $\mu^*(\boldsymbol{\theta}, \boldsymbol{\theta}_0)$ or $\mu(\boldsymbol{\theta}, \boldsymbol{\theta}_0)$ in terms of the moment generating function $\mathfrak{m}(\cdot, \boldsymbol{\theta}, \boldsymbol{\theta}_0)$ of the marginal distribution P:

$$\mu^*(\boldsymbol{\theta}, \boldsymbol{\theta}_0) = \operatorname*{argmax}_{\mu} \mathfrak{m}(\mu, \boldsymbol{\theta}, \boldsymbol{\theta}_0)$$

and $\mu(\boldsymbol{\theta}, \boldsymbol{\theta}_0)$ can be viewed as a proxy for $\mu^*(\boldsymbol{\theta}, \boldsymbol{\theta}_0)$. Denote also

$$v(\boldsymbol{\theta}) = E\, \nabla\zeta_1(\boldsymbol{\theta})[\nabla\zeta_1(\boldsymbol{\theta})]^\top, \qquad h(\delta, \gamma; \boldsymbol{\theta}) = \log E \exp\biggl\{ 2\delta \frac{\gamma^\top \nabla\zeta_1(\boldsymbol{\theta})}{\sqrt{\gamma^\top v(\boldsymbol{\theta})\gamma}} \biggr\},$$

where $\zeta_1(\boldsymbol{\theta}) = \ell(Y_1, \boldsymbol{\theta}) - I\!E\ell(Y_1, \boldsymbol{\theta})$. Notice that if P coincides with $P_{\boldsymbol{\theta}}$, then $v(\boldsymbol{\theta})$ becomes the standard Fisher information matrix. One can easily check that $h(0, \gamma; \boldsymbol{\theta}) = 0$, $\partial h(0, \gamma; \boldsymbol{\theta})/\partial\delta = 0$ and $\partial^2 h(0, \gamma; \boldsymbol{\theta})/\partial^2\delta = 4$. It follows from Lemma 7.6 that for

every $\nu_0 > 1$ and $\boldsymbol{\theta} \in \Theta$ there exists $\overline{\delta}(\boldsymbol{\theta}) > 0$ such that $h(\delta, \gamma; \boldsymbol{\theta}) \leq 2\nu_0^2\delta^2$ for all unit vectors γ in $\mathbb{R}^d$. We assume a slightly stronger condition when $\overline{\delta}(\boldsymbol{\theta})$ can be taken the same for all $\boldsymbol{\theta}$, $\overline{\delta}(\cdot) \equiv \overline{\delta}$:

$$\sup_{\boldsymbol{\theta}\in\Theta} \sup_{\gamma\in S_p} h(\delta, \gamma; \boldsymbol{\theta}) \leq 2\nu_0^2\delta^2, \qquad \delta \leq \overline{\delta}. \tag{5.1}$$

In some cases, the matrix $v(\boldsymbol{\theta})$ has to be replaced by some its regularization $\overline{v}(\boldsymbol{\theta})$ to ensure this property, see Section 6.3 for an example.

Independence of the Y_i's implies $V(\boldsymbol{\theta}) = \mathrm{Cov}\bigl\{\nabla\zeta(\boldsymbol{\theta})\bigr\} = nv(\boldsymbol{\theta})$ and

$$H(\lambda, \gamma, \boldsymbol{\theta}) \stackrel{\mathrm{def}}{=} \log I\!\!E \exp\Bigl\{2\lambda \frac{\gamma^\top \nabla\zeta(\boldsymbol{\theta})}{\sqrt{\gamma^\top V(\boldsymbol{\theta})\gamma}}\Bigr\} = nh(n^{-1/2}\lambda, \gamma; \boldsymbol{\theta})$$

for any μ and any $\gamma \in S^p$. Therefore, if $n^{-1/2}\lambda \leq \overline{\delta}$ then by (5.1):

$$H(\lambda, \gamma, \boldsymbol{\theta}) \leq 2\nu_0^2\lambda^2$$

and the condition (ED) is fulfilled with $\overline{\lambda} \leq n^{1/2}\overline{\delta}$.

Therefore, one can easily rewrite the conditions of Theorem 3.1 in terms of the marginal distribution P.

Theorem 5.1. *Assume (5.1) for some $\overline{\delta}$ and $\nu_0 \geq 1$. Let $\mu(\boldsymbol{\theta}, \boldsymbol{\theta}_0) \leq \overline{\mu}$. Suppose that there are constants $\epsilon^* \leq \overline{\mu}^{-1} \min\bigl\{\overline{\lambda}, \sqrt{2/\nu_0}\bigr\}$ and $\nu_1 \geq 1$ such that for each $\boldsymbol{\theta} \in \Theta$*

$$\mathfrak{A}_{\epsilon^*} v(\boldsymbol{\theta}) \leq \nu_1, \quad \mathfrak{A}_{\epsilon^*}\bigl\{n^{-1} + \mathfrak{m}(\boldsymbol{\theta}, \boldsymbol{\theta}_0)\bigr\} \leq \nu_1\,. \tag{5.2}$$

Let also the matrix $v(\boldsymbol{\theta})$ be non-degenerated for every $\boldsymbol{\theta} \in \Theta$. Then for any $\rho, s < 1$

$$\mathfrak{Q}(\rho, s) \leq \frac{C_p \nu_1^{p/2} n^{p/2}}{\bigl|\epsilon^*(1-\rho)\bigr|^p} \int_\Theta \exp\bigl\{-\nu_1^{-1}\rho(1-s)n\,\mathfrak{m}(\boldsymbol{\theta}, \boldsymbol{\theta}_0)\bigr\} \sqrt{\det\bigl\{v(\boldsymbol{\theta})\bigr\}}\, d\boldsymbol{\theta}\,. \tag{5.3}$$

where a constant C_p only depends on p.

Moreover, the integral in (5.3) can be easily bounded in typical situations. We present one result of this sort in which we assume that the matrix $v(\boldsymbol{\theta})$ is uniformly bounded. An extension to an unbounded $v(\boldsymbol{\theta})$ is straightforward.

Introduce for any $r > 0$ the level set $\mathcal{A}_r \stackrel{\mathrm{def}}{=} \{\boldsymbol{\theta} : \mathfrak{m}(\boldsymbol{\theta}, \boldsymbol{\theta}_0) \leq r\}$.

Theorem 5.2. *Let the conditions of Theorem 5.1 be fulfilled. Suppose in addition that*

- *for some $r > 0$ there is a constant $\mathfrak{a}_r > 0$ such that*

$$\mathfrak{m}(\boldsymbol{\theta}, \boldsymbol{\theta}_0) \;\geq\; \mathfrak{a}_r^2 (\boldsymbol{\theta} - \boldsymbol{\theta}_0)^\top v_0 (\boldsymbol{\theta} - \boldsymbol{\theta}_0)/2, \quad \boldsymbol{\theta} \in \mathcal{A}_r, \tag{5.4}$$

  where v_0 is a strictly positive definite matrix such that $v(\boldsymbol{\theta}) \leq v_0$ for any $\boldsymbol{\theta} \in \Theta$;

- *there are constants* $\nu_2 > 0$ *and* $C(\nu_2)$ *such that*

$$\int_{\Theta} \exp\bigl\{-\nu_2 \mathfrak{m}(\boldsymbol{\theta}, \boldsymbol{\theta}_0)\bigr\} d\boldsymbol{\theta} \le C(\nu_2). \tag{5.5}$$

Then for any $r > 0$ *there is* $n(r)$ *such that with* $n \ge n(r)$ *it holds*

$$\mathfrak{Q}(\rho, s) \le \frac{C_p \nu_1^p}{\mathfrak{a}_r^p \rho^{p/2} \bigl|\epsilon^*(1-\rho)\bigr|^p (1-s)^{p/2}}. \tag{5.6}$$

Proof. The conditions (5.4) and (5.5) helps easily to bound from above the integral in Theorem 5.1. Indeed, changing the variable $\boldsymbol{\theta}$ by

$$\boldsymbol{u} = n^{1/2} \mathfrak{a}_r \rho^{1/2} (1-s)^{1/2} \nu_1^{-1/2} v_0^{1/2} (\boldsymbol{\theta} - \boldsymbol{\theta}_0)$$

yields for any $r > 0$

$$\begin{aligned} & n^{p/2} \int_{\mathcal{A}_r} \exp\bigl\{-\rho(1-s) n \mathfrak{m}(\boldsymbol{\theta}, \boldsymbol{\theta}_0)/\nu_1\bigr\} \sqrt{\det\bigl\{v(\boldsymbol{\theta})\bigr\}}\, d\boldsymbol{\theta} \\ & \quad \le \frac{\nu_1^{p/2}}{\mathfrak{a}_r^p \rho^{p/2} (1-s)^{p/2}} \int_{I\!R^p} \mathrm{e}^{-\|\boldsymbol{u}\|^2/2} d\boldsymbol{u} \le \frac{(2\pi\nu_1)^{p/2}}{\mathfrak{a}_r^p \rho^{p/2} (1-s)^{p/2}}. \end{aligned}$$

The integral over the complement $\Theta \setminus \mathcal{A}_r$ can be easily bounded using (5.5):

$$\begin{aligned} & n^{p/2} \int_{\Theta \setminus \mathcal{A}_r} \exp\bigl\{-\rho(1-s) n \mathfrak{m}(\boldsymbol{\theta}, \boldsymbol{\theta}_0)/\nu_1\bigr\} \sqrt{\det v(\boldsymbol{\theta})}\, d\boldsymbol{\theta} \\ & \quad \le C(\nu_2) \sqrt{\det(v_0)}\, n^{p/2} \exp\Bigl\{-\frac{n\rho(1-s)r}{\nu_1} + \nu_2\Bigr\}. \end{aligned}$$

If $n(r)$ fulfills $n(r)\rho(1-s)r/\nu_1 - (p/2)\log[n(r)] \ge 0$, then the latter bound decreases exponentially fast as n grows over $n(r)$. This yields (5.6) in view of (5.3). □

Remark 5.1. Usually the condition (5.5) can be easily verified if $\mathfrak{m}(\boldsymbol{\theta}, \boldsymbol{\theta}_0) \ge C \log(\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|)$ for some sufficiently large C and $\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|$.

6 Applications and examples

This section illustrates how the established exponential bounds can be applied to some particular situations. To simplify technical details, we do not try to cover the most general case. Rather we aim to show that the our basic conditions can be easily verified in typical situations.

6.1 The least squares contrast

Suppose we have at our disposal the following observations:

$$Y_i = \theta_0 + \varepsilon_i, \qquad i = 1, \ldots, n,$$

where ε_i are i.i.d. with a probability density $p(x)$ satisfying $I\!E\varepsilon_i = 0$ and $I\!E\varepsilon_i^2 = 1$. A particular example we have in mind is when ε_i follow the standard Laplace law with $p(y) = \mathrm{e}^{-|y|}/2$. We also assume that there is a positive $\delta > 0$ such that

$$h_1(\delta) \stackrel{\mathrm{def}}{=} \log I\!E \exp\bigl\{\delta|\varepsilon_1|\bigr\} < \infty.$$

Suppose that $\Theta = \mathbb{R}^1$ and θ_0 is estimated with the help of the standard least squares contrast

$$L(\theta) = -\frac{1}{2}\sum_{i=1}^{n}(Y_i - \theta)^2.$$

It is easy to see that

$$\begin{aligned} L(\theta, \theta_0) &= (\theta - \theta_0)\sum_i (Y_i - \theta_0) - \frac{n}{2}(\theta - \theta_0)^2, \\ \widetilde{\theta} &= \frac{1}{n}\sum_i Y_i, \quad L(\widetilde{\theta}, \theta) = \frac{n}{2}(\widetilde{\theta} - \theta)^2. \end{aligned}$$

Also $\zeta_1(\theta) \stackrel{\mathrm{def}}{=} \ell(Y_1, \theta) - I\!E\ell(Y_1, \theta) = \theta(Y_1 - \theta_0)$, and $\nabla\zeta_1(\theta) = \varepsilon_1$ yielding $v(\theta) \equiv 1$. Next, it is easy to check by simple algebra that

$$\begin{aligned} I\!E\ell(Y_1, \theta, \theta_0) &= -(\theta - \theta_0)^2/2, \\ \mathfrak{m}(\mu, \theta, \theta_0) &= \mu(\theta - \theta_0)^2/2 - h_1\bigl(\mu(\theta - \theta_0)\bigr). \end{aligned}$$

Define for $u \geq 0$ the functions $\mu_1(u)$ and $\mathfrak{m}_1(u)$ by

$$\mu_1(u) = \operatorname*{argmax}_{\mu}\bigl\{\mu u^2/2 - h_1(\mu u)\bigr\}, \qquad \mathfrak{m}_1(u) = \max_{\mu}\bigl\{\mu u^2/2 - h_1(\mu u)\bigr\}.$$

Then obviously for every $\theta \in \Theta$ and $u = \theta - \theta_0$

$$\mu^*(\theta, \theta_0) = \mu_1(u), \qquad \mathfrak{m}^*(\theta, \theta_0) = \mathfrak{m}_1(u).$$

To apply Theorem 5.1, we need a lower bound for $\mathfrak{m}_1(u)$. We first consider the case when ε_i follow the standard Laplace law. Then $h_1(u) = \log(1 - u^2)$ and

$$\mu_1(u) = \operatorname*{argmax}_{\mu}\bigl\{\mu u^2/2 - \log(1 - \mu^2 u^2)\bigr\} = 1\big/\bigl(2 + \sqrt{4 + u^2}\bigr).$$

This means that for small $|u|$, $\mu_1(u) \approx 1/4$. Therefore one can use sub-optimal $\mu_1(u) = 1/4$ for $|u| \le 3$ and $\mu_1(u) = 1/\{1+|u|\}$ for $|u| \ge 3$ leading to

$$\mathfrak{m}_1(u) \ge C \min\{u, u^2\} \tag{6.1}$$

for some fixed $C > 0$. These arguments apply in the general case as well. Indeed, Lemma 7.6 ensures that $h_1(\delta) \ge C\delta^2$ for all small $|\delta|$. It obviously implies $\mathfrak{m}_1(u) \ge Cu^2$ for small $|u|$. On the other hand, for large $|u|$

$$\mathfrak{m}_1(u) = \max_{\mu} \bigl\{ \mu u^2/2 - h_1(\mu u) \bigr\} \ge \delta|u|/2 - h_1(\delta) \ge \delta|u|/4.$$

Thus, (6.1) is verified and Theorem 5.1 yields for some constant C with $s = 0$ and $\widetilde{u} = \widetilde{\theta} - \theta_0$

$$I\!E \exp\Bigl\{ \rho \mu_1(\widetilde{u})\, n\widetilde{u}^2 \Bigr\} \le C(1-\rho)^{-1}.$$

It is worth mentioning that for the model with Laplace's errors ε_i, the exponential moment $I\!E \exp\{\mu n(\widetilde{\theta} - \theta_0)^2/2\}$ for the LSE $\widetilde{\theta}$ does not exist whatever $\mu > 0$ is. However, Theorem 5.2 ensures bounded exponential moments for $n(\widetilde{\theta} - \theta_0)^2 / \bigl[4 + (\widetilde{\theta} - \theta_0)^2\bigr]^{1/2}$. This particularly means that if Θ is a bounded set in $\mathbb{R}^1$, then projecting $\widetilde{\theta}$ on Θ, i.e. computing $\widetilde{\theta}_\Theta = \operatorname{argmin}_{\theta \in \Theta} |\widetilde{\theta} - \theta|$, ensures a bounded exponential moment of $\mu_\Theta L(\widetilde{\theta}_\Theta, \theta_0) = \mu_\Theta n(\widetilde{\theta}_\Theta - \theta_0)^2/2$ for some $\mu_\Theta > 0$.

6.2 Estimation in the exponential model

The exponential model assumes that the observations $Y_1, \ldots, Y_n$ are i.i.d. exponential random variables with $I\!P(Y_i > y) = \exp(-\theta_0 y)$, where $\theta_0 \in \mathbb{R}^+$ is an unknown parameter of interest. In this case, the maximum likelihood contrast is given by

$$L(\theta) = -\theta \sum_{i=1}^{n} Y_i + n \log(\theta)$$

yielding

$$\widetilde{\theta} = n \Big/ \sum_{i=1}^{n} Y_i, \qquad L(\widetilde{\theta}, \theta) = n \log(\widetilde{\theta}/\theta) + n(\theta/\widetilde{\theta} - 1) = n\mathcal{K}(\widetilde{\theta}, \theta),$$

where $\mathcal{K}(\theta, \theta')$ is the Kullback-Leibler divergence between the exponential laws P_θ and $P_{\theta'}$. In this example, we focus on statistical properties of $\widetilde{\theta}$ in the situation when Y_i are i.i.d. but not necessarily exponential. Instead, we assume that $I\!E \exp\{\overline{\delta} Y_1 / I\!E Y_1\} < \infty$

for some $\overline{\delta} > 0$. Denote $\theta_0 = 1/I\!\!E Y_1$ and notice that $I\!\!E L(\theta) = -n[\theta/\theta_0 + \log(\theta)]$ and θ_0 is the maximizer of $I\!\!E L(\theta)$ and thus, the target of estimation.

Let $p(y)$ be the density of $e_1 = \theta_0 Y_1$ on $I\!R^+$. Denoting $v = \operatorname{Var} Y_1$, one can easily compute

$$\zeta_1(\theta) \stackrel{\text{def}}{=} L(\theta) - I\!\!E L(\theta) = -\theta(Y_1 - 1/\theta_0), \quad \nabla\zeta_1(\theta) = -(Y_1 - 1/\theta_0) = -\theta_0^{-1}(e_1 - 1),$$

yielding $v(\theta) = I\!\!E\big[\nabla\zeta_1(\theta)\big]^2 \equiv v$. Let

$$h_1(\delta) \stackrel{\text{def}}{=} \log I\!\!E \exp\big\{-\delta(e_1 - 1)\big\} = \log I\!\!E \exp\big\{\delta\theta_0\nabla\zeta_1(\theta)\big\}.$$

It is easy to see that the condition (5.1) is satisfied with some $\nu_0^2 < \infty$. Define also

$$\begin{aligned} \mathfrak{m}_1(u) &= \max_{\mu}\big\{\mu[u - \log(1+u)] - h_1(\mu u)\big\}, \\ \mu_1(u) &= \operatorname*{argmax}_{\mu}\big\{\mu[u - \log(1+u)] - h_1(\mu u)\big\}. \end{aligned}$$

Then, with $u = \theta/\theta_0 - 1$, we have

$$\mathfrak{m}(\mu, \theta, \theta_0) = \mu[u - \log(1+u)] - h_1(\mu u).$$

and the optimal choice of $\mu(\theta, \theta_0)$ is given by $\mu^*(\theta, \theta_0) = \mu_1(u)$ leading to $\mathfrak{m}^*(\theta, \theta_0) = \mathfrak{m}_1(u)$ for $u = \theta/\theta_0 - 1$. Thus, to make use of Theorem 5.2, we bound from below $\mathfrak{m}_1(u)$. If the underlying density $p(y)$ is standard exponential, then

$$h_1(\delta) = \delta - \log(1+\delta),$$

and obviously

$$\mathfrak{m}(\mu, \theta, \theta_0) = \log(1 + \mu u) - \mu\log(1+u).$$

Simple algebra yields

$$\mu_1(u) = \operatorname*{argmax}_{\mu}\big\{\log(1+\mu u) - \mu\log(1+u)\big\} = \frac{u - \log(1+u)}{u\log(1+u)}.$$

A simplified choice is given by $\mu(u) \equiv 1/2$. This leads to

$$\mathfrak{m}(\theta, \theta_0) = \mathfrak{m}(u) \stackrel{\text{def}}{=} \log(1 + u/2) - 0.5\log(1+u) = \frac{1}{2}\log\left[1 + \frac{u^2}{4(1+u)}\right]$$

for $u = \theta/\theta_0 - 1 > -1$. It is easy to see that $\mathfrak{m}(u) \geq c_1 u^2$ for $|u| \leq 1$, and $\mathfrak{m}(u) \geq c_2\log(1+u)$ for $u \geq 1$ with some $c_1, c_2 > 0$. So, the conditions (5.2) and (5.4) of Theorem 5.1 is easy to verify and we get the following exponential inequality

$$I\!\!E \exp\big\{\rho L(\widetilde{\theta}, \theta_0)/2\big\} \equiv I\!\!E \exp\big\{\rho n\,\mathcal{K}(\widetilde{\theta}, \theta_0)/2\big\} \leq \frac{C}{1-\rho}. \tag{6.2}$$

An important feature of this result is that it applies uniformly over the whole (unbounded and non-compact) parameter set $[0, +\infty]$. Usually the accuracy bounds are stated only uniformly on compact and separated away from zero subsets of Θ.

Another corollary of the main result is that the true parameter θ_0 lies with a high probability in the confidence set $\mathcal{E}(\mathfrak{z})$ of the form

$$\mathcal{E}(\mathfrak{z}) = \{\theta \in \Theta : \theta/\widetilde{\theta} - 1 + \log(\widetilde{\theta}/\theta) \le \mathfrak{z}/n\}$$

with some sufficiently large $\mathfrak{z}$.

6.3 LAD contrast and median estimation

Suppose we are given a sample $Y^{(n)} = (Y_1, \ldots, Y_n)$. In the problem of median estimation these random variables are assumed i.i.d. and we are interested in estimating the median θ_0 which is a root of the equation

$$I\!P(Y_i \le \theta_0) = I\!P(Y_i \ge \theta_0).$$

Alternatively, the median minimizes the value $I\!E|Y_i - \theta|$. This remark leads to the natural estimate $\widetilde{\theta}$ of the median as the minimizer of the contrast $-L(\theta) = \sum_i |Y_i - \theta|$, i.e.

$$\widetilde{\theta} = \operatorname*{argmax}_{\theta} L(\theta) = \operatorname*{argmin}_{\theta} \sum_{i=1}^{n} |Y_i - \theta|.$$

If the Y_i's are i.i.d. with the Laplace density $\exp\bigl(-|y-\theta_0|\bigr)/2$, then $L(\theta)$ coincides (up to a constant factor) with the log-likelihood. In the general case, $L(\theta)$ can be treated as the quasi log-likelihood contrast.

Assume first that Y_i are i.i.d. with the density $p_\theta(y) = p(y-\theta)$ where $p(\cdot)$ is a centrally symmetric density. To simplify the notation, we also assume that $\theta_0 = 0$. The general case can be reduced to this one by a simple change of variables. The density $p(y)$ is supposed to be positive for all y and we denote $\lambda(y) = -(2y)^{-1}\log[2I\!P(Y > y)]$ for $y \ge 0$. Equivalently, we can write $I\!P(Y > y) = \mathrm{e}^{-2y\lambda(y)}/2$ for $y \ge 0$. The case with $\lambda(y) \ge \lambda_0 > 0$ corresponds to the light tails while $\lambda(y) \to 0$ as $|y| \to \infty$ means heavy tails of the distribution P. Below we focus on the most interesting case when $\lambda(y)$ is positive and monotonously decreases to zero in $y > 0$. Later we also briefly comment on the case when the Y_i's are not i.i.d.

Let

$$m(\theta) \stackrel{\text{def}}{=} I\!\!E|Y_1 - \theta|, \qquad q(\theta) \stackrel{\text{def}}{=} I\!\!P(Y_1 \le \theta) - I\!\!P(Y_1 > \theta).$$

Obviously $m'(\theta) \stackrel{\text{def}}{=} \partial m(\theta)/\partial\theta = q(\theta)$. It is also clear that $|q(\theta)| \le 1$. With $\zeta_i(\theta) \stackrel{\text{def}}{=} |Y_i - \theta| - I\!\!E|Y_i - \theta|$, holds

$$\begin{aligned} \nabla\zeta_i(\theta) &= \partial\zeta_i(\theta)/\partial\theta = \mathbf{1}(Y_i - \theta_0 \le \theta) - \mathbf{1}(Y_i - \theta_0 > \theta) - q(\theta) \\ v(\theta) &\stackrel{\text{def}}{=} I\!\!E\big|\nabla\zeta_i(\theta)\big|^2 = 1 - q^2(\theta). \end{aligned}$$

Note that $v(\theta)$ can be arbitrary small as θ goes to $-\infty$ or $+\infty$. Therefore, further we consider its upper bound $\overline{v}(\theta) \equiv 1$. Simple algebra yields for any $\delta > 0$

$$\begin{aligned} h(\delta,\theta) &\stackrel{\text{def}}{=} \log I\!\!E\exp\big\{2\delta\nabla\zeta_i(\theta)\big\} \\ &= \log\Big[1 - (1 - \mathrm{e}^{-4\delta})\big\{1 - q(\theta)\big\}/2\Big] + 2\delta\big\{1 - q(\theta)\big\} \end{aligned}$$

thus revealing that $h(\delta,\theta) \le 4\delta^2$ for $|\delta| \le \overline{\delta} = 1/\sqrt{2}$.

Next note that for $\theta \ge 0$ it holds

$$\ell'(y,\theta,\theta_0) \stackrel{\text{def}}{=} \frac{\partial}{\partial y}\ell(y,\theta,\theta_0) = \begin{cases} 0, & y \notin [0,\theta], \\ 2, & \text{otherwise}, \end{cases}$$

and $\ell(y,\theta,\theta_0) = -\theta$ for $y \notin [0,\theta]$. Therefore, integration by parts yields

$$\begin{aligned} I\!\!E\mathrm{e}^{\mu\ell(Y,\theta,\theta_0)} &= -\int \mathrm{e}^{\mu\ell(y,\theta,\theta_0)}\, dI\!\!P(Y > y) \\ &= e^{-\mu\theta} + \int \mu\ell'(y,\theta,\theta_0)\mathrm{e}^{\mu\ell(y,\theta,\theta_0)} I\!\!P(Y > y)\, dy \\ &= e^{-\mu\theta} + 2\mu\int_0^\theta \mathrm{e}^{\mu(2y-\theta)} I\!\!P(Y > y)\, dy \\ &= \mathrm{e}^{-\mu\theta} + \mu\mathrm{e}^{-\mu\theta}\int_0^\theta \mathrm{e}^{2y[\mu-\lambda(y)]}\, dy. \end{aligned}$$

and similarly for $\theta < \theta_0$. We now fix $\mu(\theta,\theta_0) = \lambda(\theta)$. Monotonicity of $\lambda(y)$ implies

$$I\!\!E\mathrm{e}^{\mu(\theta,\theta_0)\ell(Y,\theta,\theta_0)} \le \mathrm{e}^{-\theta\lambda(\theta)} + \lambda(\theta)\mathrm{e}^{-\theta\lambda(\theta)}\int_0^\theta \mathrm{e}^{y\lambda(\theta)-y\lambda(y)}dy \le \big\{1 + \theta\lambda(\theta)\big\}\mathrm{e}^{-\theta\lambda(\theta)}.$$

Therefore, for $\theta > 0$,

$$\mathfrak{m}(\theta,\theta_0) \ge \theta\lambda(\theta) - \log\big\{1 + \theta\lambda(\theta)\big\}.$$

The same low bound holds true for $\theta < 0$.

The continuity condition (5.2) from Theorem 5.1 is obviously fulfilled for $\mathfrak{m}(\theta,\theta_0)$ and $\overline{v}(\theta) \equiv 1$. Also the condition (5.5) is fulfilled as soon as $I\!\!E|Y_i|^\gamma < \infty$ for some $\gamma > 0$. Theorem 5.2 applied with $\rho = s$ leads to the bound for the loss $\widetilde{u} = |\widetilde{\theta} - \theta_0|$:

$$I\!\!E \exp\bigl\{\rho^2 n\bigl[\widetilde{u}\lambda(\widetilde{u}) - \log\{1 + \widetilde{u}\lambda(\widetilde{u})\}\bigr]\bigr\} \le \frac{C}{1-\rho}$$

provided that n is sufficiently large.

Finally notice that the case of independent but non i.i.d. observations can be again reduced to the considered case with the help of the averaged c.d.f. $P = n^{-1}\sum_i P_i$. So, the point θ_0 is now a root of the equation

$$\sum_{i=1}^n P_i(Y_i < \theta) = \sum_{i=1}^n P_i(Y_i > \theta).$$

6.4 Estimation of the location of a change point

Suppose we have at our disposal the noisy data

$$Y_k = A\,\mathbf{1}(k \le \theta) + \sigma\xi_k, \quad k = 1,\ldots,n, \tag{6.3}$$

where ξ_k is a standard white Gaussian noise. Our goal is to estimate the change point $\theta \in \Theta^n = \{1,\ldots,n-1\}$. We begin with the case when the amplitude A is known. To estimate θ, we use the maximum likelihood estimator

$$\widetilde{\theta}_A = \operatorname*{argmax}_{\theta\in\Theta^n} L_A(\theta),$$

where the maximum likelihood contrast is given by

$$L_A(\theta) = \frac{A}{\sigma^2}\sum_{k=1}^{\theta} Y_k - \frac{A^2}{2\sigma^2}\theta = \frac{A^2}{\sigma^2}\min(\theta,\theta_0) - \frac{A^2\theta}{2\sigma^2} + \frac{A}{\sigma}\sum_{k=1}^{\theta}\xi_k.$$

Note that this is a special case of a Gaussian likelihood contrast, see Subsection 4, with

$$M(\theta,\theta_0) = \frac{A^2}{2\sigma^2}|\theta-\theta_0|, \quad \mathfrak{V}^2(\theta,\theta') = \frac{A^2}{\sigma^2}|\theta-\theta'|.$$

Therefore, for $\rho < 1$, Theorem 2.1 implies

$$\begin{aligned} I\!\!E\exp\Bigl\{\rho^2\frac{A^2}{4\sigma^2}|\widetilde{\theta}-\theta_0|\Bigr\} &\le \sum_{\boldsymbol{\theta}\in\Theta}\exp\bigl\{-\frac{\rho(1-\rho)}{4}M(\theta,\theta_0)\bigr\} \\ &\le 2\sum_{k=0}^{\infty}\exp\Bigl\{-\frac{\rho(1-\rho)A^2}{8\sigma^2}k\Bigr\} = \frac{2}{1-C(\rho)} \end{aligned}$$

where $C(\rho) = \exp\{-\rho(1-\rho)A^2/(8\sigma^2)\}$. By Lemma 7.5

$$\mathbb{E}|\widetilde{\theta}_A - \theta_0|^r \le C_1(r)\big(\sigma^2/A^2\big)^r$$

with some constant $C_1(r)$.

Now we switch to the case when $A > 0$ is an unknown parameter. In this case, we cannot use the contrast $L_A(\theta)$ because it strongly depends on A. To get a reasonable contrast one can use the maximum likelihood principle. Considering A as a nuisance parameter and maximizing $L_A(\theta)$ w.r.t. $A \ge 0$, leads to the following estimate

$$\widetilde{\theta} = \operatorname*{argmax}_{\theta}\Big\{\max_{A\ge 0} L_A(\theta)\Big\} = \operatorname*{argmax}_{\theta} \frac{1}{2\sigma^2\theta}\Big[\sum_{k=1}^{\theta} Y_k\Big]_+^2,$$

where $[x]_+ = \max(x, 0)$. In what follows we deal with a slightly modified version of this estimator

$$\widetilde{\theta} = \operatorname*{argmax}_{\theta\in\Theta^n} L(\theta), \quad \text{with a new contrast} \quad L(\theta) = \frac{1}{\sigma\sqrt{\theta}}\sum_{k=1}^{\theta} Y_k,$$

which is again a Gaussian one. The model equation (6.3) allows to represent the contrast in the form

$$L(\theta) = \frac{1}{\sqrt{\theta}}\sum_{k=1}^{\theta}\xi_k + \frac{A\min(\theta,\theta_0)}{\sigma\sqrt{\theta}}.$$

It is easy to see that

$$M(\theta,\theta_0) = \mathfrak{a} d(\theta,\theta_0)$$

with $\mathfrak{a} = \sigma^{-1}A\sqrt{\theta_0}$, and

$$d(\theta,\theta') = 1 - \sqrt{\min\{\theta/\theta', \theta'/\theta\}} = \begin{cases} 1 - \sqrt{\theta/\theta'}, & \theta \le \theta', \\ 1 - \sqrt{\theta'/\theta}, & \theta \ge \theta'. \end{cases}$$

Similarly,

$$\mathfrak{V}^2(\theta,\theta') = \frac{2|\theta'-\theta|}{(\sqrt{\theta}+\sqrt{\theta'})\sqrt{\max(\theta,\theta')}} = 2d(\theta,\theta')$$

and obviously, $M(\theta,\theta_0) = \mathfrak{a}\mathfrak{V}^2(\theta,\theta_0)/2$. Also $\mathfrak{V}^2(\theta,\theta_0) \le 2$ for all θ. Next, by (4.1)

$$\mu^*(\theta,\theta_0) = \frac{M(\theta,\theta_0)}{\mathfrak{V}^2(\theta,\theta_0)} = \frac{\mathfrak{a}}{2}, \qquad \mathfrak{M}^*(\theta,\theta_0) = \frac{\mathfrak{a}^2}{8} d(\theta,\theta_0).$$

Note that for every $\theta \in \Theta$, the value $\mathfrak{M}^*(\theta,\theta_0)$ is bounded by $\mathfrak{a}^2/8$. So, this example is quite special in the sense that the Kullback-Leibler divergence between measures $I\!P_{\theta_0}$

and $I\!P_\theta$ does not grow to infinity with θ. We will see that this fact results in an extra loglog-factor in the bound for the minimum contrast.

For given $\epsilon > 0$ and $\theta^\circ \in \Theta$, the local ball $B(\epsilon, \theta^\circ) = \{\mathfrak{V}(\theta, \theta^\circ) \le \epsilon\}$ can be represented in the form

$$B(\epsilon, \theta^\circ) = \bigl\{\theta : \theta^\circ(1 - \epsilon^2/2)^2 \le \theta \le \theta^\circ(1 - \epsilon^2/2)^{-2}\bigr\}.$$

First of all, notice that this ball becomes the usual symmetric interval around $\log \theta^\circ$ for the parameter $\log \theta$:

$$B(\epsilon, \theta^\circ) = \{\theta : \bigl|\log \theta - \log \theta^\circ\bigr| \le -2 \log(1 - \epsilon^2/2)\}.$$

This immediately implies that the local entropy $\mathfrak{e}(\theta^\circ)$ is bounded by $\bar{\mathfrak{e}} = 1$ for all $\theta^\circ \in \Theta$.

Next, for the number $\Pi_\epsilon(\theta)$ of points θ in $B(\epsilon, \theta^\circ)$, it holds $\Pi_\epsilon(\theta) \approx K(\epsilon)\theta$ with $K(\epsilon) = (1 - \epsilon^2/2)^{-2} - (1 - \epsilon^2/2)^2 \ge \epsilon^2$ for $\epsilon \le 1$. Fix $\epsilon^2 = 1/2$. Then for every $\theta^\circ \notin B(2\epsilon, \theta_0)$, the magnitude of $1 + \mathfrak{M}(\theta, \theta_0)$ within the ball $B(\epsilon, \theta^\circ)$ is bounded by a fixed constant and the condition (4.2) is easily verified. This yields by Theorem 4.1

$$I\!E \exp\bigl\{\mathfrak{a}^2 d(\widetilde{\theta}, \theta_0)\bigr\} \le C_1 \sum_{\theta=1}^{n} \frac{1}{\Pi_\epsilon(\theta)} \le C_2 \sum_{\theta=1}^{n} \frac{1}{\theta} \le C_2 \log n, \tag{6.4}$$

thus resulting in

$$I\!E \exp\bigl\{\rho^2 \mathfrak{a}^2 d(\widetilde{\theta}, \theta_0)\bigr\} \le C_2 \log(n).$$

Combining this with Lemma 7.5 yields

$$I\!E\bigl\{\mathfrak{a}^2 d(\widetilde{\theta}, \theta_0)\bigr\}^r \le C |\log\log n|^r.$$

It is interesting to compare this result with the accuracy of the maximum likelihood method in the case, where the magnitude of jump A is known. One can see that there is a payment for the adaptation to the nuisance parameter A which is in form of an extra $\log\log$-factor. Another observation is that the accuracy of estimation strongly depends on the true location θ_0, more precisely, on the value $\mathfrak{a}^2 = A^2\theta_0/\sigma^2$. In the "classical" situation this value is of order n leading to the accuracy of order $n^{-1}\log\log(n)$. If this value becomes smaller in order than n, then the accuracy decreases with the same factor. In particular, if $A^2\theta_0/\sigma^2$ is of order one, then even consistency for the estimate $\widetilde{\theta}$ cannot be stated.

7 Proofs

This section collects proofs of the main theorems and some auxiliary facts.

The important technical result behind our bounds for contrast process is related to the behavior of the supremum of the stochastic component $\zeta(\boldsymbol{\theta}, \boldsymbol{\theta}^\diamond)$ of the contrast process $L(\boldsymbol{\theta}, \boldsymbol{\theta}^\diamond)$ over a local ball $B(\epsilon, \boldsymbol{\theta}^\circ) = \{\boldsymbol{\theta} : \mathfrak{S}(\boldsymbol{\theta}, \boldsymbol{\theta}^\circ) \le \epsilon\}$ with some ϵ.

Theorem 7.1. *Assume that $\zeta(\boldsymbol{\theta})$ is a separable process satisfying for some $\boldsymbol{\theta}^\circ$ the condition (EL). Then for any $\boldsymbol{\theta}^\diamond \in B(\epsilon, \boldsymbol{\theta}^\circ)$*

$$\log I\!\!E \exp\left\{ \frac{\lambda}{\epsilon} \sup_{\boldsymbol{\theta} \in B(\epsilon, \boldsymbol{\theta}^\circ)} \zeta(\boldsymbol{\theta}, \boldsymbol{\theta}^\diamond) \right\} \le \mathfrak{e}(\boldsymbol{\theta}^\circ) + \varkappa(\lambda).$$

Proof. The proof is based on the standard chaining argument (see e.g. [12]). Without loss of generality, we may assume that $\mathfrak{e}(\boldsymbol{\theta}^\circ) < \infty$. Then for any integer $k \ge 0$, there exists a $2^{-k}\epsilon$-net $\mathcal{D}_k(\epsilon)$ in the local ball $B(\epsilon, \boldsymbol{\theta}^\circ)$ having the cardinality $\mathbb{N}_{\boldsymbol{\theta}^\circ}(2^{-k}\epsilon)$. Using the nets $\mathcal{D}_k(\epsilon)$ with $k = 1, \ldots, K-1$, one can construct a chain connecting an arbitrary point $\boldsymbol{\theta}$ in $\mathcal{D}_K(\epsilon)$ and $\boldsymbol{\theta}'$. It means that one can find points $\boldsymbol{\theta}_k \in \mathcal{D}_k(\epsilon)$, $k = 1, \ldots, K-1$, such that $\mathfrak{S}(\boldsymbol{\theta}_k, \boldsymbol{\theta}_{k-1}) \le 2^{-k+1}\epsilon$ for $k = 1, \ldots, K$. Here we denoted for $\boldsymbol{\theta}_K = \boldsymbol{\theta}$, and $\boldsymbol{\theta}_0 = \boldsymbol{\theta}^\diamond$. Notice that $\boldsymbol{\theta}_k$ can be constructed recurrently $\boldsymbol{\theta}_{k-1} = \tau_{k-1}(\boldsymbol{\theta}_k)$, $k = K, \ldots, 1$, where

$$\tau_{k-1}(\boldsymbol{\theta}) = \operatorname*{argmin}_{\boldsymbol{\theta}' \in \mathcal{D}_{k-1}(\epsilon)} \mathfrak{S}(\boldsymbol{\theta}, \boldsymbol{\theta}').$$

It obviously holds

$$\zeta(\boldsymbol{\theta}, \boldsymbol{\theta}') = \sum_{k=1}^{K} \zeta(\boldsymbol{\theta}_k, \boldsymbol{\theta}_{k-1}).$$

In view of the definition of $\xi(\cdot, \cdot)$

$$\zeta(\boldsymbol{\theta}_k, \boldsymbol{\theta}_{k-1}) = \mathfrak{S}(\boldsymbol{\theta}_k, \boldsymbol{\theta}_{k-1}) \times \xi(\boldsymbol{\theta}_k, \boldsymbol{\theta}_{k-1}) = 2\epsilon\, c_k\, \xi(\boldsymbol{\theta}_k, \boldsymbol{\theta}_{k-1})$$

with $c_k = c_k(\boldsymbol{\theta}) = \mathfrak{S}(\boldsymbol{\theta}_k, \boldsymbol{\theta}_{k-1})/(2\epsilon) \le 2^{-k}$, thus resulting in

$$\begin{aligned} \sup_{\boldsymbol{\theta} \in \mathcal{D}_k(\epsilon)} \zeta(\boldsymbol{\theta}, \boldsymbol{\theta}^\circ) &\le \sum_{k=1}^{K} \sup_{\boldsymbol{\theta}' \in \mathcal{D}_k(\epsilon)} \zeta(\boldsymbol{\theta}', \tau_{k-1}(\boldsymbol{\theta}')) \\ &\le 2\epsilon \sum_{k=1}^{K} \sup_{\boldsymbol{\theta}' \in \mathcal{D}_k(\epsilon)} c_k \xi(\boldsymbol{\theta}', \tau_{k-1}(\boldsymbol{\theta}')). \end{aligned} \tag{7.1}$$

Because of $c_k \le 2^{-k}$, this yields by Lemma 7.4 and condition $(M\lambda)$:

$$\begin{aligned}
\log I\!\!E \exp\Big\{ \frac{\lambda}{\epsilon} \sup_{\boldsymbol{\theta} \in \mathcal{D}_K(\epsilon)} \zeta(\boldsymbol{\theta}, \boldsymbol{\theta}^\circ) \Big\} &\le \log I\!\!E \exp\Big\{ 2\lambda \sum_{k=1}^{K} \sup_{\boldsymbol{\theta}' \in \mathcal{D}_k(\epsilon)} c_k \xi(\boldsymbol{\theta}', \tau_{k-1}(\boldsymbol{\theta}')) \Big\} \\
&\le \sum_{k=1}^{K} 2^{-k} \log \Big[I\!\!E \exp\Big\{ \sup_{\boldsymbol{\theta}' \in \mathcal{D}_k(\epsilon)} 2^k c_k \times 2\lambda \xi(\boldsymbol{\theta}', \tau_{k-1}(\boldsymbol{\theta}')) \Big\} \Big] \\
&\le \sum_{k=1}^{K} 2^{-k} \log \Big[\sum_{\boldsymbol{\theta}' \in \mathcal{D}_k(\epsilon)} I\!\!E \exp\big\{ 2^k c_k \times 2\lambda \xi(\boldsymbol{\theta}', \tau_{k-1}(\boldsymbol{\theta}')) \big\} \Big] \\
&\le \sum_{k=1}^{K} 2^{-k} \big\{ \log \mathbb{N}_{\boldsymbol{\theta}^\circ}(2^{-k}\epsilon) + \varkappa(\lambda) \big\}.
\end{aligned}$$

These inequalities with the separability of $\zeta(\boldsymbol{\theta}, \boldsymbol{\theta}^\diamond)$ yield

$$\begin{aligned}
\log I\!\!E \exp\Big\{ \frac{\lambda}{\epsilon} \sup_{\boldsymbol{\theta} \in B(\epsilon, \boldsymbol{\theta}^\circ)} \zeta(\boldsymbol{\theta}, \boldsymbol{\theta}^\diamond) \Big\} &= \lim_{K \to \infty} \log I\!\!E \exp\Big\{ \frac{\lambda}{\epsilon} \sup_{\boldsymbol{\theta} \in \mathcal{D}_K(\epsilon)} \zeta(\boldsymbol{\theta}, \boldsymbol{\theta}^\diamond) \Big\} \\
&\le \sum_{k=1}^{\infty} 2^{-k} \big\{ \varkappa(\lambda) + \log \mathbb{N}_{\boldsymbol{\theta}^\circ}(2^{-k}\epsilon) \big\} \le \varkappa(\lambda) + \mathfrak{e}(\boldsymbol{\theta}^\circ)
\end{aligned}$$

which completes the proof of the theorem. □

7.1 Proof of Theorem 2.2

Fix a point $\boldsymbol{\theta}^\circ \in \mathcal{D}$. For given $\rho, s < 1$, denote

$$\boldsymbol{\theta}^\diamond = \operatorname*{argmax}_{\boldsymbol{\theta} \in B(\epsilon, \boldsymbol{\theta}^\circ)} \big\{ \mu(\boldsymbol{\theta}, \boldsymbol{\theta}_0) I\!\!E L(\boldsymbol{\theta}, \boldsymbol{\theta}_0) + s \mathfrak{M}(\boldsymbol{\theta}, \boldsymbol{\theta}_0) \big\}.$$

It is clear that with $\overline{\mu}(\boldsymbol{\theta}^\circ) = \sup_{\boldsymbol{\theta} \in B(\epsilon, \boldsymbol{\theta}^\circ)} \mu(\boldsymbol{\theta}, \boldsymbol{\theta}_0)$

$$\begin{aligned}
&\sup_{\boldsymbol{\theta} \in B(\epsilon, \boldsymbol{\theta}^\circ)} \Big\{ \rho \big[\mu(\boldsymbol{\theta}, \boldsymbol{\theta}_0) L\big(\boldsymbol{\theta}, \boldsymbol{\theta}_0\big) + s \mathfrak{M}(\boldsymbol{\theta}, \boldsymbol{\theta}_0) \big] \Big\} \\
&\quad \le \rho \big[\mu(\boldsymbol{\theta}^\diamond, \boldsymbol{\theta}_0) L(\boldsymbol{\theta}^\diamond, \boldsymbol{\theta}_0) + s \mathfrak{M}(\boldsymbol{\theta}^\diamond, \boldsymbol{\theta}_0) \big] + \rho \overline{\mu}(\boldsymbol{\theta}^\circ) \sup_{\boldsymbol{\theta} \in B(\epsilon, \boldsymbol{\theta}^\circ)} \zeta(\boldsymbol{\theta}, \boldsymbol{\theta}^\diamond).
\end{aligned}$$

Next, Theorem 7.1 implies for any μ and ϵ

$$\log I\!\!E \exp\Big\{ \mu \sup_{\boldsymbol{\theta} \in B(\epsilon, \boldsymbol{\theta}^\circ)} \zeta(\boldsymbol{\theta}, \boldsymbol{\theta}^\diamond) \Big\} \le \varkappa(\mu\epsilon) + \mathfrak{e}(\boldsymbol{\theta}^\circ).$$

By definition of $\breve{\mathfrak{M}}(\boldsymbol{\theta}^\circ, \boldsymbol{\theta}_0)$

$$\begin{aligned}
\log I\!\!E \exp\Big\{ \mu(\boldsymbol{\theta}^\diamond, \boldsymbol{\theta}_0) L(\boldsymbol{\theta}^\diamond, \boldsymbol{\theta}_0) \Big\} + s \mathfrak{M}(\boldsymbol{\theta}^\diamond, \boldsymbol{\theta}_0) &= -(1-s) \mathfrak{M}(\boldsymbol{\theta}^\diamond, \boldsymbol{\theta}_0) \\
&\le -(1-s) \breve{\mathfrak{M}}(\boldsymbol{\theta}^\circ, \boldsymbol{\theta}_0).
\end{aligned}$$

This yields by the Hölder inequality

$$\begin{aligned}
&\log I\!E \exp\Big\{ \sup_{\boldsymbol{\theta} \in B(\epsilon, \boldsymbol{\theta}^\circ)} \rho\big[\mu(\boldsymbol{\theta}, \boldsymbol{\theta}_0) L\big(\boldsymbol{\theta}, \boldsymbol{\theta}_0\big) + s \mathfrak{M}(\boldsymbol{\theta}, \boldsymbol{\theta}_0)\big] \Big\} \\
&\quad \le \log I\!E \exp\Big\{ \rho\big[\mu(\boldsymbol{\theta}^\diamond, \boldsymbol{\theta}_0) L(\boldsymbol{\theta}^\diamond, \boldsymbol{\theta}_0) + s \mathfrak{M}(\boldsymbol{\theta}^\diamond, \boldsymbol{\theta}_0)\big] + \rho \overline{\mu}(\boldsymbol{\theta}^\circ) \sup_{\boldsymbol{\theta} \in B(\epsilon, \boldsymbol{\theta}^\circ)} \zeta(\boldsymbol{\theta}, \boldsymbol{\theta}^\diamond) \Big\} \\
&\quad \le \rho \log I\!E \exp\Big\{ \mu(\boldsymbol{\theta}^\diamond, \boldsymbol{\theta}_0) L(\boldsymbol{\theta}^\diamond, \boldsymbol{\theta}_0) \Big\} + \rho s\, \mathfrak{M}(\boldsymbol{\theta}^\diamond, \boldsymbol{\theta}_0) \\
&\qquad + (1-\rho) \log I\!E \exp\Big\{ \frac{\rho \overline{\mu}(\boldsymbol{\theta}^\circ)}{1-\rho} \sup_{\boldsymbol{\theta} \in B(\epsilon, \boldsymbol{\theta}^\circ)} \zeta(\boldsymbol{\theta}, \boldsymbol{\theta}^\diamond) \Big\} \\
&\quad \le -\rho(1-s) \breve{\mathfrak{M}}(\boldsymbol{\theta}^\circ, \boldsymbol{\theta}_0) + (1-\rho) \mathfrak{e}(\boldsymbol{\theta}^\circ) + (1-\rho) \varkappa\Big(\frac{\rho \overline{\mu}(\boldsymbol{\theta}^\circ) \epsilon}{1-\rho} \Big).
\end{aligned}$$

Therefore, by (2.9)

$$\begin{aligned}
&I\!E \exp\Big\{ \sup_{\boldsymbol{\theta} \in \Theta} \rho\big[\mu(\boldsymbol{\theta}, \boldsymbol{\theta}_0) L(\boldsymbol{\theta}, \boldsymbol{\theta}_0) + s \mathfrak{M}(\boldsymbol{\theta}, \boldsymbol{\theta}_0)\big] \Big\} \qquad (7.2) \\
&\quad \le \sum_{\boldsymbol{\theta}^\circ \in \mathcal{D}} I\!E \exp\Big\{ \sup_{\boldsymbol{\theta} \in B(\epsilon, \boldsymbol{\theta}^\circ)} \rho\big[\mu(\boldsymbol{\theta}, \boldsymbol{\theta}_0) L(\boldsymbol{\theta}, \boldsymbol{\theta}_0) + s \mathfrak{M}(\boldsymbol{\theta}, \boldsymbol{\theta}_0)\big] \Big\} \\
&\quad \le \sum_{\boldsymbol{\theta}^\circ \in \mathcal{D}} \exp\Big\{ -\rho(1-s) \breve{\mathfrak{M}}(\boldsymbol{\theta}^\circ, \boldsymbol{\theta}_0) + (1-\rho) \varkappa\Big(\frac{\rho \overline{\mu}(\boldsymbol{\theta}^\circ) \epsilon}{1-\rho} \Big) + (1-\rho) \mathfrak{e}(\boldsymbol{\theta}^\circ) \Big\}
\end{aligned}$$

as required.

7.2 Proof of Theorems 3.1 and 4.2

In the proof by C_p we denote a generic constant (not necessarily the same) which only depends on the dimensionality p. First we show that the differentiability condition (ED) implies the local moment condition (EL).

Lemma 7.2. *Assume (ED) with some ν_0 and $\overline{\lambda}$. Then for any $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \Theta$ and any λ with $|\lambda| \le \overline{\lambda}$,*

$$\log I\!E \exp\Big\{ 2\lambda \frac{\zeta(\boldsymbol{\theta}, \boldsymbol{\theta}')}{\mathfrak{S}(\boldsymbol{\theta}, \boldsymbol{\theta}')} \Big\} \le 2\nu_0^2 \lambda^2. \qquad (7.3)$$

Proof. For $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \Theta$, denote for brevity $\boldsymbol{u} = \boldsymbol{\theta} - \boldsymbol{\theta}'$. With these notations, we obviously have

$$L(\boldsymbol{\theta}, \boldsymbol{\theta}') = \boldsymbol{u}^\top \int_0^1 \nabla L(\boldsymbol{\theta}' + t\boldsymbol{u}) dt.$$

Similar expressions hold for $I\!EL(\boldsymbol{\theta}, \boldsymbol{\theta}')$ and for $\zeta(\boldsymbol{\theta}, \boldsymbol{\theta}') = L(\boldsymbol{\theta}, \boldsymbol{\theta}') - I\!EL(\boldsymbol{\theta}, \boldsymbol{\theta}')$ i.e.,

$$\zeta(\boldsymbol{\theta}, \boldsymbol{\theta}') = \boldsymbol{u}^\top \int_0^1 \nabla \zeta(\boldsymbol{\theta}' + t\boldsymbol{u}) dt.$$

The definition of $\mathfrak{S}(\boldsymbol{\theta}, \boldsymbol{\theta}')$ implies for any $t \in [0,1]$

$$c(t) \stackrel{\text{def}}{=} \frac{\sqrt{\boldsymbol{u}^\top V(\boldsymbol{\theta} + t\boldsymbol{u})\boldsymbol{u}}}{\mathfrak{S}(\boldsymbol{\theta}, \boldsymbol{\theta}')} \leq 1,$$

and therefore Lemma 7.4 and (3.3) with $\gamma = \boldsymbol{u}/\|\boldsymbol{u}\|$ yield

$$\begin{aligned} \log I\!\!E \exp\left\{ 2\lambda \frac{\zeta(\boldsymbol{\theta}, \boldsymbol{\theta}')}{\mathfrak{S}(\boldsymbol{\theta}, \boldsymbol{\theta}')} \right\} &= \log I\!\!E \exp\left\{ 2\lambda \int_0^1 c(t) \frac{\gamma^\top \nabla \zeta(\boldsymbol{\theta} + t\boldsymbol{u})}{\sqrt{\gamma^\top V(\boldsymbol{\theta} + t\boldsymbol{u})\gamma}}\, dt \right\} \\ &\leq \int_0^1 c(t) \log I\!\!E \exp\left\{ 2\lambda \frac{\gamma^\top \nabla \zeta(\boldsymbol{\theta} + t\boldsymbol{u})}{\sqrt{\gamma^\top V(\boldsymbol{\theta} + t\boldsymbol{u})\gamma}} \right\} dt \leq 2\nu_0^2 \lambda^2. \end{aligned}$$

□

Due to the next lemma, the smoothness of the contrast implies that the topology induced by the metric $\mathfrak{S}(\cdot, \cdot)$ is (locally) equivalent to the Euclidean topology and computing the local entropy $\mathfrak{e}(\cdot)$ can be reduced to the Euclidean case.

Define also for every $\boldsymbol{\theta}^\circ \in \Theta$ the elliptic set $B'(\epsilon, \boldsymbol{\theta}^\circ)$ by

$$B'(\epsilon, \boldsymbol{\theta}^\circ) = \left\{ \boldsymbol{\theta} : (\boldsymbol{\theta} - \boldsymbol{\theta}^\circ)^\top V(\boldsymbol{\theta}^\circ)\, (\boldsymbol{\theta} - \boldsymbol{\theta}^\circ) \leq \epsilon^2 \right\}.$$

The definition of $B(\epsilon, \boldsymbol{\theta})$ implies that $B(\epsilon, \boldsymbol{\theta}^\circ) \subseteq B'(\epsilon, \boldsymbol{\theta}^\circ)$.

Lemma 7.3. *Assume* (ED) *with some* $\overline{\lambda}$, *and let, for some fixed* $\nu_1 \geq 1$, $\epsilon > 0$

$$\mathfrak{A}_\epsilon V(\boldsymbol{\theta}) \leq \nu_1, \qquad \boldsymbol{\theta} \in \Theta. \tag{7.4}$$

Then

- (EL) *is fulfilled with* $\varkappa(\lambda) \leq 2\nu_0^2\lambda^2$, *for* $\lambda \leq \overline{\lambda}$, *i.e. (7.3) holds for all* $\lambda \leq \overline{\lambda}$.
- $\sup_{\boldsymbol{\theta} \in \Theta} \mathfrak{e}(\boldsymbol{\theta}) \leq C(p, \nu_1)$, *where* $C(p, \nu_1)$ *is a constant depending on* p *and* ν_1.

Proof. The first claim is an immediate corollary of Lemma 7.2. Next, for each fixed $\boldsymbol{\theta}^\circ \in \Theta$, after the linear transformation, all the local balls $B'(2^{-k}\epsilon, \boldsymbol{\theta})$ with $\boldsymbol{\theta} \in B'(\epsilon, \boldsymbol{\theta}^\circ)$ and $k \geq 0$ become the usual Euclidean balls and the corresponding covering number is obviously bounded by a constant depending on the dimension p and ν_1 only. □

Now we are ready to proceed with the proof of Theorem 3.1. Define

$$\epsilon \stackrel{\text{def}}{=} \epsilon^*(1 - \rho)$$

and consider the ellipsoid $B'(\epsilon, \boldsymbol{\theta}^\circ) = \{\boldsymbol{\theta} : (\boldsymbol{\theta} - \boldsymbol{\theta}^\circ)^\top V(\boldsymbol{\theta}^\circ)\, (\boldsymbol{\theta} - \boldsymbol{\theta}^\circ) \leq \epsilon^2\}$. Its Lebesgue measure $\mathrm{mes}(B'(\epsilon, \boldsymbol{\theta}^\circ))$ is equal to $C_p \big/ \sqrt{\det\{\epsilon^{-2} V(\boldsymbol{\theta}^\circ)\}}$ where C_p is the volume of the

unit ball in $I\!R^p$. First we bound the value $1/\operatorname{mes}(B'(\epsilon,\boldsymbol{\theta}^\circ))$. The conditions (3.5) imply that for any $\boldsymbol{\theta} \in B'(\epsilon,\boldsymbol{\theta}^\circ)$

$$\det\bigl\{\epsilon^{-2}V(\boldsymbol{\theta}^\circ)\bigr\} \le \frac{\nu_1^p}{\bigl|\epsilon^*(1-\rho)\bigr|^{2p}}\det\bigl\{V(\boldsymbol{\theta})\bigr\}. \tag{7.5}$$

Since $\overline{\mu}(\boldsymbol{\theta}^\circ) \le \overline{\mu}$,

$$\frac{\rho\overline{\mu}(\boldsymbol{\theta}^\circ)\epsilon}{1-\rho} \le \epsilon^*\rho\overline{\mu} \le \overline{\lambda},$$

and therefore by (ED) and by Lemma 7.2 we obtain

$$(1-\rho)\,\varkappa\biggl(\frac{\rho\overline{\mu}(\boldsymbol{\theta}^\circ)\epsilon}{1-\rho}\biggr) \le (1-\rho)2\nu_0\bigl(\epsilon^*\rho\overline{\mu}\bigr)^2 \le 2\nu_0\rho\epsilon^{*2}\rho^2(1-\rho)\overline{\mu}^2 \le \nu_0\rho\epsilon^{*2}\overline{\mu}^2/2 \le \rho. \tag{7.6}$$

The condition (3.5) implies $1+\breve{\mathfrak{M}}(\boldsymbol{\theta}^\circ,\boldsymbol{\theta}_0) \ge \{1+\mathfrak{M}(\boldsymbol{\theta},\boldsymbol{\theta}_0)\}/\nu_1$, and hence $\breve{\mathfrak{M}}(\boldsymbol{\theta}^\circ,\boldsymbol{\theta}_0) \ge \nu_1^{-1} - 1 + \nu_1^{-1}\mathfrak{M}(\boldsymbol{\theta},\boldsymbol{\theta}_0)$. Combining this with (7.5) and (7.6) yields the following bound

$$\begin{aligned}
&\exp\biggl\{-\rho(1-s)\breve{\mathfrak{M}}(\boldsymbol{\theta}^\circ,\boldsymbol{\theta}_0) + (1-\rho)\varkappa\biggl(\frac{\rho\overline{\mu}(\boldsymbol{\theta}^\circ)\epsilon}{1-\rho}\biggr)\biggr\} \\
&\quad\le \frac{\mathrm{e}^{2\rho}}{\operatorname{mes}(B'(\epsilon,\boldsymbol{\theta}^\circ))}\int_{B'(\epsilon,\boldsymbol{\theta}^\circ)}\exp\bigl\{-\rho(1-s)\mathfrak{M}(\boldsymbol{\theta},\boldsymbol{\theta}_0)/\nu_1\bigr\}d\boldsymbol{\theta} \\
&\quad\le \frac{C_p\nu_1^{p/2}}{\bigl|\epsilon^*(1-\rho)\bigr|^p}\int_{B'(\epsilon,\boldsymbol{\theta}^\circ)}\exp\bigl\{-\rho(1-s)\mathfrak{M}(\boldsymbol{\theta},\boldsymbol{\theta}_0)/\nu_1\bigr\}\sqrt{\det\bigl\{V(\boldsymbol{\theta})\bigr\}}\,d\boldsymbol{\theta}\,.
\end{aligned} \tag{7.7}$$

Next, consider the set $B^*(\epsilon,\boldsymbol{\theta}^\circ) = \{\boldsymbol{\theta} : \|\boldsymbol{\theta}-\boldsymbol{\theta}^\circ\|^2 \le \epsilon^2/\lambda_{\min}[V(\boldsymbol{\theta}^\circ)]\}$. By condition (3.5), the matrix $V(\boldsymbol{\theta})$ is nearly constant within this set and hence the squared distance $\mathfrak{S}^2(\boldsymbol{\theta},\boldsymbol{\theta}')$ can be well approximated by $\bigl(\boldsymbol{\theta}-\boldsymbol{\theta}'\bigr)^\top V(\boldsymbol{\theta}^\circ)\bigl(\boldsymbol{\theta}-\boldsymbol{\theta}'\bigr)$. This enables to build easily an ϵ-net $\mathcal{D}(\epsilon,\boldsymbol{\theta}^\circ)$ in the ball $B^*(\epsilon,\boldsymbol{\theta}^\circ)$ such that every point $\boldsymbol{\theta}' \in B^*(\epsilon,\boldsymbol{\theta}^\circ)$ is covered by the balls $B'(\epsilon,\boldsymbol{\theta})$ for $\boldsymbol{\theta} \in \mathcal{D}(\epsilon,\boldsymbol{\theta}^\circ)$ only a finite number of times depending on dimensionality p. The use of (7.7) leads to

$$\begin{aligned}
&\sum_{\boldsymbol{\theta}\in\mathcal{D}(\epsilon,\boldsymbol{\theta}^\circ)}\exp\biggl\{-\rho(1-s)\breve{\mathfrak{M}}(\boldsymbol{\theta},\boldsymbol{\theta}_0) + (1-\rho)\varkappa\biggl(\frac{\rho\overline{\mu}(\boldsymbol{\theta})\epsilon}{1-\rho}\biggr)\biggr\} \\
&\quad\le \frac{C_p\nu_1^p}{\bigl|\epsilon^*(1-\rho)\bigr|^p}\int_{B^*(\epsilon,\boldsymbol{\theta}^\circ)}\exp\bigl\{-\rho(1-s)\mathfrak{M}(\boldsymbol{\theta},\boldsymbol{\theta}_0)/\nu_1\bigr\}\sqrt{\det\bigl\{V(\boldsymbol{\theta})\bigr\}}\,d\boldsymbol{\theta}\,.
\end{aligned} \tag{7.8}$$

To finish the proof, it remains to show that $\Theta \subseteq I\!R^p$ can be covered by the balls $B^*(\epsilon,\boldsymbol{\theta}^\circ)$. Define $r_k = \nu_1 k$ and consider the sets $\mathcal{A}_k = \{\boldsymbol{\theta} : \lambda_{\min}[V(\boldsymbol{\theta})] \in [r_{k-1}, r_k]\}$. Condition (3.5) ensures that for every $\boldsymbol{\theta}^\circ \in \mathcal{A}_k$ and every $\boldsymbol{\theta} \in B^*(\epsilon,\boldsymbol{\theta}^\circ)$, it holds $\nu_1^{-1}r_{k-1} \le \lambda_{\min}[V(\boldsymbol{\theta})] \le \nu_1 r_k$. This particularly implies that $B^*(\epsilon,\boldsymbol{\theta}^\circ)$ can only cross $\mathcal{A}_{k-1}$ and $\mathcal{A}_{k+1}$. Next, if $\boldsymbol{\theta}^\circ \in \mathcal{A}_k$ then $A(\epsilon/r_k,\boldsymbol{\theta}^\circ) \subseteq B^*(\epsilon,\boldsymbol{\theta}^\circ) \subseteq A(\epsilon/r_{k-1},\boldsymbol{\theta}^\circ)$ where

$A(a, \boldsymbol{\theta}^\circ) = \{\boldsymbol{\theta} : \|\boldsymbol{\theta}^\circ - \boldsymbol{\theta}\| \le a\}$. Clearly $\mathcal{A}_k$ can be covered by the balls $A(\epsilon/r_k, \boldsymbol{\theta}^\circ)$ and hence, by the balls $B^*(\epsilon, \boldsymbol{\theta}^\circ)$ so that every point in $\mathcal{A}_k$ is only covered by a finite number of times C_p which depends on p only. If $\mathcal{D}_k$ is the union of the sets $\mathcal{D}(\epsilon, \boldsymbol{\theta}^\circ)$ over these points $\boldsymbol{\theta}^\circ$, then the result (7.8) extends to the set $\mathcal{A}_k$. For the final extension to the whole parameter space it only remains to note that any ball $B^*(\epsilon, \boldsymbol{\theta}^\circ)$ can only intersect with at most two sets $\mathcal{A}_k$ for different k. Now the the assertion follows directly from Theorem 2.2 and Lemma 7.3.

For the proof of Theorem 4.2 only observe that $\overline{\lambda} = \infty$, $\nu_0 = 1$, and the choice of ϵ can be slightly refined: $\epsilon = \epsilon^*(1-\rho)^{1/2}\overline{\mu}$ it holds

$$\det\Big\{\epsilon^{-2} V(\boldsymbol{\theta}^\circ)\Big\} \le \left[\frac{\nu_1}{\epsilon^{*2}(1-\rho)}\right]^p \det\big\{V(\boldsymbol{\theta})\big\}$$

and

$$(1-\rho)\,\varkappa\left(\frac{\rho\overline{\mu}(\boldsymbol{\theta}^\circ)\epsilon}{1-\rho}\right) \le 2\big(\epsilon^*\rho\big)^2 \le 1.$$

The rest of the proof is the same as for Theorem 3.1.

7.3 Auxiliary facts

Lemma 7.4. *For any r.v.'s ξ_k and any nonnegative coefficients λ_k with $\Lambda = \sum_k \lambda_k \le 1$*

$$\log I\!\!E \exp\left(\sum_k \lambda_k \xi_k\right) \le \sum_k \lambda_k \log I\!\!E \mathrm{e}^{\xi_k} \tag{7.9}$$

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

$$\begin{aligned} I\!\!E \exp\left\{\frac{\Lambda}{\Lambda}\sum_k \lambda_k\big(\xi_k - \log I\!\!E \mathrm{e}^{\xi_k}\big)\right\} &\le I\!\!E^\Lambda \exp\left\{\frac{1}{\Lambda}\sum_k \lambda_k\big(\xi_k - \log I\!\!E \mathrm{e}^{\xi_k}\big)\right\} \\ &\le \left\{\frac{1}{\Lambda}\sum_k \lambda_k I\!\!E \exp\big(\xi_k - \log I\!\!E \mathrm{e}^{\xi_k}\big)\right\}^\Lambda = 1. \end{aligned}$$

□

Lemma 7.5. *Let ξ be a nonnegative random variable and*

$$\varphi(\lambda) = \log I\!\!E \exp\big(\lambda\xi\big)$$

for $\lambda \ge 0$. Then for any $r > 0$

$$\big(I\!\!E\xi^r\big)^{1/r} \le \inf_{\lambda:\,\varphi(\lambda)\ge r} \lambda^{-1}\varphi(\lambda). \tag{7.10}$$

In particular, if $\varphi(\lambda) \le a + \sigma^2\lambda^2$ for some $a, \sigma \ge 0$, then

$$\big(I\!\!E\xi^r\big)^{1/r} \le 2\sigma\sqrt{\max\{a, r/2\}}. \tag{7.11}$$

Proof. Consider the following function

$$f(x) = \begin{cases} \log^r(x) & \text{for } x \geq e^r, \\ xr^r/e^r & \text{for } x \leq e^r. \end{cases}$$

A simple algebra reveals that for $x > e^r$

$$\begin{aligned} f'(x) &= rx^{-1}\log^{r-1}(x), \\ f''(x) &= r(r-1)x^{-2}\log^{r-2}(x) - rx^{-2}\log^{r-1}(x) \\ &= rx^{-2}\big[r - 1 - \log(x)\big]\log^{r-2}(x) < 0. \end{aligned}$$

Since the function $f(x)$ is linear for $x \leq e^r$, it is concave for all $x \geq 0$. It is also easy to check that $[\log(x)]_+^r \leq f(x)$, because for $x \leq e^r$, the function $f(x)$ coincides with the tangent of $\log^r(x)$ at $x = e^r$. Therefore,

$$x^r = \lambda^{-r}\log^r\big(e^{\lambda x}\big) \leq \lambda^{-r} f(e^{\lambda x})$$

and the Jensen inequality implies for any $\lambda \geq 0$

$$I\!E\xi^r \leq \lambda^{-r} I\!E f(e^{\lambda\xi}) \leq \lambda^{-r} f\big(I\!E e^{\lambda\xi}\big) = \lambda^{-r} f\big(\mathrm{e}^{\varphi(\lambda)}\big). \tag{7.12}$$

If $\varphi(\lambda) \geq r$, then $f\big(\mathrm{e}^{\varphi(\lambda)}\big) = \log^r\big(e^{\varphi(\lambda)}\big) = \varphi^r(\lambda)$ and (7.10) follows from (7.12).

To prove (7.11), it remains to notice that the monotonicity of $f(\cdot)$ implies in view of (7.12)

$$\begin{aligned} (I\!E\xi^r)^{1/r} &\leq \inf_{\lambda:\, a+\sigma^2\lambda^2 \geq r} \left\{ \frac{a}{\lambda} + \sigma^2\lambda \right\} = \begin{cases} \sigma r (r-a)^{-1/2}, & a < r/2 \\ 2\sigma\sqrt{a}, & a \geq r/2 \end{cases} \\ &\leq \begin{cases} 2\sigma\sqrt{r/2}, & a < r/2 \\ 2\sigma\sqrt{a}, & a \geq r/2 \end{cases} \leq 2\sigma\sqrt{\max\{a, r/2\}}. \end{aligned}$$

□

Lemma 7.6. *Let a r.v.* ξ *fulfill* $I\!E\xi = 0$, $I\!E\xi^2 = 1$ *and* $I\!E\exp(\lambda_1|\xi|) = \varkappa < \infty$ *for some* $\lambda_1 > 0$. *Then for any* $\rho < 1$ *there is a constant* C_1 *depending on* $\varkappa$, λ_1 *and* ρ *only such that for* $\lambda < \rho\lambda_1$

$$\log I\!E\mathrm{e}^{\lambda\xi} \leq C_1\lambda^2/2.$$

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

$$\log I\!E\mathrm{e}^{\lambda\xi} \geq \rho\lambda^2/2.$$

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 $\log I\!\!E \exp(\lambda_1|\xi|) = \varkappa(\lambda_1) < \infty$. Then the function $h_0(\lambda) = I\!\!E \exp\bigl(\lambda\xi\bigr)$ fulfills $h_0(0) = 1$, $h_0'(0) = I\!\!E\xi = 0$, $h_0''(0) = 1$ and for $\lambda < \lambda_1$,

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

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

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

with $C_1 = \varkappa(\lambda_1)/\bigl\{2\lambda_1^2(1-\rho)^2\bigr\}$, and hence, $g(\lambda) = \log h_0(\lambda) \leq C_1\lambda^2/2$. □

References

[1] Birgé, L. (2006). Model selection via testing: an alternative to (penalized) maximum likelihood estimators. *Ann. Inst. Henri Poincaré, Probab. Stat.* **42**, No. 3, 273–325 (2006).

[2] Birgé, L., Massart, P. (1993). Rate of convergence for minimum contrast estimators. *Probab.Theory and Related Fields* **97**, 113–150.

[3] Birgé, L., Massart, P. (1998). Minimum contrast estimators on sieves: Exponential bounds and rates of convergence. *Bernoulli* **4**, No.3, 329–375 (1998).

[4] Field, C. A. (1982). Small sample asymptotic expansions for multivariate *M*-estimates. *Ann. Statist.* **10**, 672–689.

[5] Field, C. A. and Ronchetti, E. (1990). *Small sample asymptotics, IMS Lecture Notes-Monograph Ser.*, **13**.

[6] Ibragimov, I.A., Has'minskii, R.Z. (1981). *Statistical Estimation: Asymptotic Theory*. Springer-Verlag.

[7] Jensen, J. L. and Wood, A. (1998). Large deviation and other results for minimum contrast estimators. *Ann. Inst. Statist. Math.*, **50**, No. 4, 673–695.

[8] Huber, P. J. (1967). The behaviour of the maximum likelihood estimators under nonstandard conditions. *Proc. 5th Berkley Symp. Math. Stat. Prob.*, **1**, 221–233.

[9] Huber, P. J. (1981, 2004). *Robust Statistics.* Wiley.

[10] Sieders, A. and Dzhaparidze, K. (1987). A large deviation result for parameter estimators and its application to nonlinear regression analysis, *Ann. Statist.*, **15**, 1031-1049.

[11] Van de Geer, S. (1993). Hellinger-consistency of certain nonparametric maximum likelihood estimates. *Ann. Statist.* **21**, 14–44.

[12] Van der Vaart, A. and Wellner, J. (1996). *Weak convergence and empirical processes.* Springer-Verlag, NY.