Bayesian Analysis (2015) **10**, Number 3, pp. 665–710

Finite Sample Bernstein – von Mises Theorem for Semiparametric Problems

Maxim Panov* and Vladimir Spokoiny†

Abstract. The classical parametric and semiparametric Bernstein – von Mises (BvM) results are reconsidered in a non-classical setup allowing finite samples and model misspecification. In the case of a finite dimensional nuisance parameter we obtain an upper bound on the error of Gaussian approximation of the posterior distribution for the target parameter which is explicit in the dimension of the nuisance and target parameters. This helps to identify the so called *critical dimension* p_n of the full parameter for which the BvM result is applicable. In the important i.i.d. case, we show that the condition "p_n^3/n is small" is sufficient for the BvM result to be valid under general assumptions on the model. We also provide an example of a model with the phase transition effect: the statement of the BvM theorem fails when the dimension p_n approaches $n^{1/3}$. The results are extended to the case of infinite dimensional parameters with the nuisance parameter from a Sobolev class.

Keywords: prior, posterior, Bayesian inference, semiparametric, critical dimension.

1 Introduction

The prominent Bernstein – von Mises (BvM) theorem claims that the posterior measure is asymptotically normal with the mean close to the maximum likelihood estimator (MLE) and the posterior variance is nearly the inverse of the total Fisher information matrix. The BvM result provides a theoretical background for Bayesian computations of the MLE and its variance. Also it justifies usage of elliptic credible sets based on the first two moments of the posterior. The classical version of the BvM Theorem is stated for the standard parametric setup with a fixed parametric model and large samples; see Le Cam and Yang (1990); van der Vaart (2000) for a detailed overview. However, in modern statistics applications one often faces very complicated models involving a lot of parameters and with a limited sample size. This requires an extension of the classical results to this non-classical situation. We mention Cox (1993); Freedman (1999); Ghosal (1999); Johnstone (2010) and references therein for some special phenomena arising in the Bayesian analysis when the parameter dimension increases. Already consistency of the posterior distribution in nonparametric and semiparametric models is a nontrivial problem; cf. Schwartz (1965) and Barron et al. (1996). Asymptotic normality of the posterior measure for these classes of models is even more challenging; see e.g. Shen

*Moscow Institute of Physics and Technology, Institute for Information Transmission Problems of RAS, Datadvance Company, Pokrovsky blvd. 3 building 1B, 109028 Moscow, Russia, panov.maxim@gmail.com

†Weierstrass Institute and Humboldt University Berlin, Moscow Institute of Physics and Technology, Institute for Information Transmission Problems of RAS, Mohrenstr. 39, 10117 Berlin, Germany, spokoiny@wias-berlin.de

 DOI: 10.1214/14-BA926

(2002). Some results for particular semi and nonparametric problems are available from Kim and Lee (2004); Kim (2006); Leahu (2011); Castillo and Nickl (2013). Cheng and Kosorok (2008) obtained a version of the BvM statement based on a high order expansion of the profile sampler. The recent paper by Bickel and Kleijn (2012) extends the BvM statement from the classical parametric case to a rather general i.i.d. framework. Castillo (2012) studies the semiparametric BvM result for Gaussian process functional priors. In Rivoirard and Rousseau (2012) a semiparametric BvM theorem is derived for linear functionals of density and in forthcoming work (Castillo and Rousseau, 2013) the result is generalized to a broad class of models and functionals. However, all these results are limited to the asymptotic setup and to some special classes of models like i.i.d. or Gaussian.

In this paper we reconsider the BvM result for the parametric component of a general semiparametric model. An important feature of the study is that the sample size is fixed, we proceed with just one sample. A finite sample theory is especially challenging because most notions, methods and tools in the classical theory are formulated in the asymptotic setup with growing sample size. Only a few finite sample general results are available; see e.g. the recent paper by Boucheron and Massart (2011). This paper focuses on the semiparametric problem when the full parameter is large or infinite dimensional but the target is low dimensional. In the Bayesian framework, the aim is the marginal of the posterior corresponding to the target parameter; cf. Castillo (2012). Typical examples are provided by functional estimation, estimation of a function at a point, or simply by estimating a given subvector of the parameter vector. An interesting feature of the semiparametric BvM result is that the nuisance parameter appears only via the effective score and the related efficient Fisher information; cf. Bickel and Kleijn (2012). The methods of study heavily rely on the notion of the hardest parametric submodel. In addition, one assumes that an estimate of the nuisance parameter is available which ensures a certain accuracy of estimation; see Cheng and Kosorok (2008) or Bickel and Kleijn (2012). This essentially simplifies the study but does not allow to derive a qualitative relation between the full dimension of the parameter space and the total available information in the data.

Some recent results study the impact of a growing parameter dimension p_n on the quality of Gaussian approximation of the posterior. We mention Ghosal (1999, 2000), Boucheron and Gassiat (2009), Johnstone (2010) and Bontemps (2011) for specific examples. See the discussion after Theorem 4 below for more details.

In this paper we show that the *bracketing* approach of Spokoiny (2012) can be used for obtaining a finite sample semiparametric version of the Bernstein – von Mises theorem even if the full parameter dimension grows with the sample size. The ultimate goal of this paper is to quantify the so called critical parameter dimension for which the BvM result can be applied. Our approach neither relies on a pilot estimate of the nuisance and target parameter nor involves the notion of the hardest parametric submodel. In the case of finite dimensional nuisance the obtained results only require some smoothness of the log-likelihood function, its finite exponential moments, and some identifiability conditions. Further we specify this result to the i.i.d. setup and show that the imposed conditions are satisfied if p_n^3/n is small. We present an example showing

that the dimension $p_n = O(n^{1/3})$ is indeed critical and the BvM result starts to fail if p_n grows over $n^{1/3}$. If the nuisance is infinite dimensional then additionally some smoothness of the nonparametric part is required. We state the general BvM results and show the validity of the BvM result for linear and generalized linear models with the nuisance parameter from the Sobolev class and a uniform sieve prior distribution on the parameter space.

Now we describe our setup. Let $\boldsymbol{Y}$ denote the observed random data, and $I\!P$ denote the data distribution. The parametric statistical model assumes that the unknown data distribution $I\!P$ belongs to a given parametric family $(I\!P_{\boldsymbol{\upsilon}})$:

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

where Υ is some parameter space and $\boldsymbol{\upsilon}^* \in \Upsilon$ is the true value of parameter. In the semiparametric framework, one attempts to recover only a low dimensional component $\boldsymbol{\theta}$ of the whole parameter $\boldsymbol{\upsilon}$. This means that the target of estimation is $\boldsymbol{\theta}^* \stackrel{\text{def}}{=} \Pi_0 \boldsymbol{\upsilon}^*$ for some mapping $\Pi_0 : \Upsilon \to I\!R^q$, and $q \in \mathbb{N}$ stands for the dimension of the target. Usually in the classical semiparametric setup, the vector $\boldsymbol{\upsilon}$ is represented as $\boldsymbol{\upsilon} = (\boldsymbol{\theta}, \boldsymbol{\eta})$, where $\boldsymbol{\theta}$ is the target of analysis while $\boldsymbol{\eta}$ is the *nuisance parameter*. We refer to this situation as the $(\boldsymbol{\theta}, \boldsymbol{\eta})$-setup and our presentation follows this setting. An extension to the $\boldsymbol{\upsilon}$-setup with $\boldsymbol{\theta} = \Pi_0 \boldsymbol{\upsilon}$ is straightforward. Also for simplicity we first develop our results for the case when the total parameter space Υ is a subset of the Euclidean space of dimensionality p.

Another issue addressed in this paper is the model misspecification. In the most of practical problems, it is unrealistic to expect that the model assumptions are exactly fulfilled, even if some rich nonparametric models are used. This means that the true data distribution $I\!P$ does not belong to the considered family $(I\!P_{\boldsymbol{\upsilon}}, \boldsymbol{\upsilon} \in \Upsilon)$. The "true" value $\boldsymbol{\upsilon}^*$ of the parameter $\boldsymbol{\upsilon}$ can defined by

$$\boldsymbol{\upsilon}^* \stackrel{\text{def}}{=} \operatorname*{argmax}_{\boldsymbol{\upsilon} \in \Upsilon} I\!E \mathcal{L}(\boldsymbol{\upsilon}), \tag{1}$$

where $\mathcal{L}(\boldsymbol{\upsilon}) = \log \frac{dI\!P_{\boldsymbol{\upsilon}}}{d\boldsymbol{\mu}_0}(\boldsymbol{Y})$ is the log-likelihood function of the family $(I\!P_{\boldsymbol{\upsilon}})$ for some dominating measure $\boldsymbol{\mu}_0$. Under model misspecification, $\boldsymbol{\upsilon}^*$ defines the best parametric fit to $I\!P$ by the considered family; cf. Chernozhukov and Hong (2003), Kleijn and van der Vaart (2006, 2012) and references therein. The target $\boldsymbol{\theta}^*$ is defined by the mapping Π_0:

$$\boldsymbol{\theta}^* \stackrel{\text{def}}{=} \Pi_0 \boldsymbol{\upsilon}^*.$$

Now we switch to the Bayesian set-up. Let π be a prior measure on the parameter set Υ. Below we study the properties of the posterior measure which is the random measure on Υ describing the conditional distribution of $\boldsymbol{\upsilon}$ given $\boldsymbol{Y}$ and obtained by normalization of the product $\exp\{\mathcal{L}(\boldsymbol{\upsilon})\}\pi(d\boldsymbol{\upsilon})$. This relation is usually written as

$$\boldsymbol{\upsilon} \mid \boldsymbol{Y} \propto \exp\{\mathcal{L}(\boldsymbol{\upsilon})\}\, \pi(d\boldsymbol{\upsilon}). \tag{2}$$

An important feature of our analysis is that $\mathcal{L}(\boldsymbol{\upsilon})$ is not assumed to be the true log-likelihood. This means that a model misspecification is possible and the underlying data distribution can be beyond the considered parametric family. In this sense, the Bayes formula (2) describes a *quasi posterior*; Chernozhukov and Hong (2003). Below we show that smoothness of the log-likelihood function $\mathcal{L}(\boldsymbol{\upsilon})$ ensures a kind of a Gaussian approximation of the posterior measure. Our focus is to describe the accuracy of such approximation as a function of the parameter dimension p and the other important characteristics of the model.

We suppose that the prior measure π has a positive density $\pi(\boldsymbol{\upsilon})$ w.r.t. to the Lebesgue measure on Υ: $\pi(d\boldsymbol{\upsilon}) = \pi(\boldsymbol{\upsilon})d\boldsymbol{\upsilon}$. Then (2) can be written as

$$\boldsymbol{\upsilon} \mid \boldsymbol{Y} \propto \exp\bigl\{\mathcal{L}(\boldsymbol{\upsilon})\bigr\}\,\pi(\boldsymbol{\upsilon}). \tag{3}$$

The famous Bernstein – von Mises (BvM) theorem claims that the posterior centered by any efficient estimator $\widetilde{\boldsymbol{\upsilon}}$ of the parameter $\boldsymbol{\upsilon}^*$ (for example the MLE) and scaled by the total Fisher information matrix is nearly standard normal:

$$\mathcal{D}_0(\boldsymbol{\upsilon} - \widetilde{\boldsymbol{\upsilon}}) \mid \boldsymbol{Y} \xrightarrow{w} \mathcal{N}(0, \boldsymbol{I}_p)\,,$$

where $\boldsymbol{I}_p$ is an identity matrix of dimension p.

An important feature of the posterior distribution is that it is entirely known and can be numerically assessed. If we know in addition that the posterior is nearly normal, it suffices to compute its mean and variance for building the concentration and credible sets. The BvM result does not require the prior distribution to be proper and the phenomenon can be observed in the case of improper priors as well (for examples, see Bochkina and Green (2014)).

In this work we investigate the properties of the posterior distribution for the target parameter $\boldsymbol{\vartheta} \bigl| \boldsymbol{Y} = \Pi_0 \boldsymbol{\upsilon} \bigr| \boldsymbol{Y}$. In this case (3) can be written as

$$\boldsymbol{\vartheta} \mid \boldsymbol{Y} \propto \int \exp\bigl\{\mathcal{L}(\boldsymbol{\upsilon})\bigr\}\,\pi(\boldsymbol{\upsilon})d\boldsymbol{\eta}. \tag{4}$$

The BvM result in this case transforms into

$$\breve{D}_0(\boldsymbol{\vartheta} - \widetilde{\boldsymbol{\theta}}) \mid \boldsymbol{Y} \xrightarrow{w} \mathcal{N}(0, \boldsymbol{I}_q)\,,$$

where $\boldsymbol{I}_q$ is an identity matrix of dimension q, $\widetilde{\boldsymbol{\theta}} = \Pi_0 \widetilde{\boldsymbol{\upsilon}}$, and $\breve{D}_0^2$ is given in (6).

We consider two important classes of priors, namely non-informative and flat Gaussian priors. Our goal is to show under mild conditions that the posterior distribution of the target parameter (4) is close to a prescribed Gaussian law even for finite samples. The other important issue is to specify the conditions on the sample size and the dimension of the parameter space for which the BvM result is still applicable.

2 BvM Theorem with a finite dimensional nuisance

This section presents our main results for the case of a finite dimensional parameter υ, i.e. $\dim(\Upsilon) = p < \infty$. One of the main elements of our construction is a $p \times p$ matrix $\mathcal{D}_0^2$ which is defined similarly to the Fisher information matrix:

$$\mathcal{D}_0^2 \stackrel{\text{def}}{=} -\nabla^2 I\!\!E\mathcal{L}(\upsilon^*). \tag{5}$$

Here and in what follows we work under conditions which are close to the classical conditions of a *regular parametric family* (see Ibragimov and Khas'minskij (1981)) and assume that the log-likelihood function $\mathcal{L}(\upsilon)$ is sufficiently smooth in υ, $\nabla\mathcal{L}(\upsilon)$ stands for its gradient and $\nabla^2 I\!\!E\mathcal{L}(\upsilon)$ for the Hessian of the expectation $I\!\!E\mathcal{L}(\upsilon)$, and the true value υ^* is due to (1). Also define the score

$$\boldsymbol{\xi} \stackrel{\text{def}}{=} \mathcal{D}_0^{-1}\nabla\mathcal{L}(\upsilon^*).$$

Under our conditions we can permute expectation and differentiation of the likelihood and thus the definition of υ^* implies $\nabla I\!\!E\mathcal{L}(\upsilon^*) = 0$ and hence, $I\!\!E\boldsymbol{\xi} = 0$.

For the $(\boldsymbol{\theta}, \boldsymbol{\eta})$-setup, we consider the block representation of the vector $\nabla\mathcal{L}(\upsilon^*)$ and of the matrix and $\mathcal{D}_0^2$ from (5):

$$\nabla\mathcal{L}(\upsilon^*) = \begin{pmatrix} \nabla_{\boldsymbol{\theta}} \\ \nabla_{\boldsymbol{\eta}} \end{pmatrix}, \quad \mathcal{D}_0^2 = \begin{pmatrix} D_0^2 & A_0 \\ A_0^\top & H_0^2 \end{pmatrix}.$$

Define also the $q \times q$ matrix $\breve{D}_0^2$ and random vectors $\breve{\nabla}_{\boldsymbol{\theta}}, \breve{\boldsymbol{\xi}} \in I\!R^q$ as

$$\breve{D}_0^2 \stackrel{\text{def}}{=} D_0^2 - A_0 H_0^{-2} A_0^\top, \tag{6}$$

$$\breve{\nabla}_{\boldsymbol{\theta}} \stackrel{\text{def}}{=} \nabla_{\boldsymbol{\theta}} - A_0 H_0^{-2} \nabla_{\boldsymbol{\eta}},$$

$$\breve{\boldsymbol{\xi}} \stackrel{\text{def}}{=} \breve{D}_0^{-1} \breve{\nabla}_{\boldsymbol{\theta}}. \tag{7}$$

The $q \times q$ matrix $\breve{D}_0^2$ is usually called the efficient Fisher information matrix, while the random vector $\breve{\boldsymbol{\xi}} \in I\!R^q$ is the efficient score. Everywhere in the text for a vector $\boldsymbol{a}$ we denote by $\|\boldsymbol{a}\|$ its Euclidean norm and for a matrix A we denote by $\|A\|$ its operator norm.

2.1 Conditions

Our results assume a number of conditions to be satisfied. The list is essentially as in Spokoiny (2012), one can find there some discussion and examples showing that the conditions are not restrictive and are fulfilled in most classical models used in statistical studies like i.i.d., regression or generalized linear models. The conditions are split into

local and global. The local conditions only describe the properties of the process $\mathcal{L}(\boldsymbol{\upsilon})$ for $\boldsymbol{\upsilon} \in \Upsilon_0(\mathtt{r}_0)$ with some fixed value $\mathtt{r}_0$:

$$\Upsilon_0(\mathtt{r}_0) \stackrel{\text{def}}{=} \bigl\{\boldsymbol{\upsilon} \in \Upsilon\colon \|\mathcal{D}_0(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^*)\| \le \mathtt{r}_0\bigr\}.$$

The global conditions have to be fulfilled on the whole Υ. Define the stochastic component $\zeta(\boldsymbol{\upsilon})$ of $\mathcal{L}(\boldsymbol{\upsilon})$:

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

and introduce the notation

$$L(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^*) \stackrel{\text{def}}{=} \mathcal{L}(\boldsymbol{\upsilon}) - \mathcal{L}(\boldsymbol{\upsilon}^*)$$

for the (quasi) log-likelihood ratio. We start with some exponential moments conditions.

($\boldsymbol{ED_0}$) There exists a constant $\nu_0 > 0$, a positive symmetric $p \times p$ matrix $\mathcal{V}_0^2$ satisfying $\operatorname{Var}\{\nabla\zeta(\boldsymbol{\upsilon}^*)\} \le \mathcal{V}_0^2$, and a constant $\mathtt{g} > 0$ such that

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

($\boldsymbol{ED_2}$) There exists a constant $\nu_0 > 0$, a constant $\omega > 0$ and for each $\mathtt{r} > 0$ a constant $\mathtt{g}(\mathtt{r}) > 0$ such that for all $\boldsymbol{\upsilon} \in \Upsilon_0(\mathtt{r})$:

$$\sup_{\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2 \in I\!R^p} \log I\!\!E \exp\left\{\frac{\mu}{\omega} \frac{\boldsymbol{\gamma}_1^\top \nabla^2\zeta(\boldsymbol{\upsilon})\boldsymbol{\gamma}_2}{\|\mathcal{D}_0\boldsymbol{\gamma}_1\| \cdot \|\mathcal{D}_0\boldsymbol{\gamma}_2\|}\right\} \le \frac{\nu_0^2\mu^2}{2}, \ |\mu| \le \mathtt{g}(\mathtt{r}).$$

The next condition is needed to ensure some smoothness properties of expected log-likelihood $I\!\!E\mathcal{L}(\boldsymbol{\upsilon})$ in the local zone $\boldsymbol{\upsilon} \in \Upsilon_0(\mathtt{r}_0)$. Define

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

Then $\mathcal{D}_0^2 = \mathcal{D}_0^2(\boldsymbol{\upsilon}^*)$.

($\boldsymbol{\mathcal{L}_0}$) There exists a constant $\delta(\mathtt{r})$ such that it holds on the set $\Upsilon_0(\mathtt{r})$ for all $\mathtt{r} \le \mathtt{r}_0$

$$\bigl\|\mathcal{D}_0^{-1}\mathcal{D}_0^2(\boldsymbol{\upsilon})\mathcal{D}_0^{-1} - \boldsymbol{I}_p\bigr\| \le \delta(\mathtt{r}).$$

The global identification condition is:

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

$$-I\!\!EL(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^*) \ge \mathtt{r}^2\mathtt{b}(\mathtt{r}) \quad \text{for all } \boldsymbol{\upsilon} \text{ with } \mathtt{r} = \|\mathcal{D}_0(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^*)\|.$$

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

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

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

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

$$\mathfrak{a}^2 D_0^2 \geq V_0^2, \qquad \mathfrak{a}^2 H_0^2 \geq Q_0^2, \qquad \mathfrak{a}^2 \mathcal{D}_0^2 \geq \mathcal{V}_0^2. \tag{8}$$

and

$$\| D_0^{-1} A_0 H_0^{-2} A_0^\top D_0^{-1} \| \leq \nu. \tag{9}$$

The quantity ν bounds the angle between the target and nuisance subspaces in the tangent space. The regularity condition $(\mathcal{I})$ ensures that this angle is not too small and hence, the target and nuisance parameters are identifiable. In particular, the matrix $\breve{D}_0^2$ from (6) is well posed under $(\mathcal{I})$. The bounds in (8) are given with the same constant $\mathfrak{a}$ only for simplifying the notation. One can show that the last bound on $\mathcal{D}_0^2$ follows from the first two and (9) with another constant $\mathfrak{a}'$ depending on $\mathfrak{a}$ and ν only.

2.2 The main results

First we state the BvM result about the properties of the $\boldsymbol{\vartheta}$-posterior given by (4) in the case of a uniform prior that is, $\pi(\boldsymbol{\upsilon}) \equiv 1$ on Υ. Define

$$\overline{\boldsymbol{\vartheta}} \stackrel{\text{def}}{=} I\!\!E(\boldsymbol{\vartheta} \,|\, \boldsymbol{Y}), \qquad \mathfrak{S}^2 \stackrel{\text{def}}{=} \operatorname{Cov}(\boldsymbol{\vartheta} \,|\, \boldsymbol{Y}) \stackrel{\text{def}}{=} I\!\!E\{(\boldsymbol{\vartheta} - \overline{\boldsymbol{\vartheta}})(\boldsymbol{\vartheta} - \overline{\boldsymbol{\vartheta}})^\top \,|\, \boldsymbol{Y}\}. \tag{10}$$

Also define

$$\boldsymbol{\theta}^\circ \stackrel{\text{def}}{=} \boldsymbol{\theta}^* + \breve{D}_0^{-1} \breve{\boldsymbol{\xi}},$$

where $\breve{D}_0$ and $\breve{\boldsymbol{\xi}}$ are defined by (6) and (7) respectively. The random point $\boldsymbol{\theta}^\circ$ can be viewed as a first order approximation of the profile MLE $\widetilde{\boldsymbol{\theta}}$. Below we present a version of the BvM result in the considered nonasymptotic setup which claims that $\overline{\boldsymbol{\vartheta}}$ is close to $\boldsymbol{\theta}^\circ$, $\mathfrak{S}^2$ is nearly equal to $\breve{D}_0^{-2}$, and $\breve{D}_0(\boldsymbol{\vartheta} - \boldsymbol{\theta}^\circ)$ is nearly standard normal conditionally on $\boldsymbol{Y}$.

We suppose that a large constant $\mathtt{x}$ is fixed which specifies random events $\Omega(\mathtt{x})$ of *dominating probability*. We say that a generic random set $\Omega(\mathtt{x})$ is of dominating probability if

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

The notation $\mathtt{C}$ is for a generic absolute constant and $\mathtt{x}$ is for a positive value ensuring that $\mathrm{e}^{-\mathtt{x}}$ is negligible. The exact values of $\mathtt{C}$ will be specified in each particular case. The formulation of the results also involve the radius $\mathtt{r}_0$ and the spread $\Delta(\mathtt{r}_0, \mathtt{x})$. The radius $\mathtt{r}_0$ separates the local zone $\Upsilon_\circ(\mathtt{r}_0)$ which is a vicinity of the central point $\boldsymbol{\upsilon}^*$, and its complement $\Upsilon \setminus \Upsilon_\circ(\mathtt{r}_0)$ for which we establish a large deviation result. The *spread* value $\Delta(\mathtt{r}_0, \mathtt{x})$ measures the quality of local approximation of the log-likelihood $L(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^*)$ by a quadratic process $\mathbb{L}(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^*)$:

$$\Delta(\mathtt{r}_0, \mathtt{x}) \stackrel{\mathrm{def}}{=} \bigl\{ \delta(\mathtt{r}_0) + 6\nu_0 \, z_{\mathbb{H}}(\mathtt{x}) \, \omega \bigr\} \, \mathtt{r}_0^2 .$$

Here the term $\delta(\mathtt{r}_0)\mathtt{r}_0^2$ measures the error of a quadratic approximation of the expected log-likelihood $L(\boldsymbol{\upsilon})$ due to $(\mathcal{L}_0)$, while the second term $6\nu_0 \, z_{\mathbb{H}}(\mathtt{x}) \, \omega \, \mathtt{r}_0^2$ controls the stochastic term and involves the entropy of the parameter space which is involved in the definition of $z_{\mathbb{H}}(\mathtt{x})$. A precise formulation is given in Theorem 9 below.

Theorem 1. *Suppose the conditions of Section 2.1. Let the prior be uniform on Υ. Then there exists a random event $\Omega(\mathtt{x})$ of probability at least $1 - 4\mathrm{e}^{-\mathtt{x}}$ such that it holds on $\Omega(\mathtt{x})$*

$$\begin{aligned} \| \breve{D}_0 (\overline{\boldsymbol{\vartheta}} - \boldsymbol{\theta}^\circ) \|^2 &\leq 4\Delta(\mathtt{r}_0, \mathtt{x}) + 16\mathrm{e}^{-\mathtt{x}}, \\ \bigl\| I_q - \breve{D}_0 \mathfrak{S}^2 \breve{D}_0 \bigr\| &\leq 4\Delta(\mathtt{r}_0, \mathtt{x}) + 16\mathrm{e}^{-\mathtt{x}}, \end{aligned}$$

where $\overline{\boldsymbol{\vartheta}}$ and $\mathfrak{S}^2$ are from (10).

Moreover, on $\Omega(\mathtt{x})$ for any measurable set $A \subseteq \mathbb{R}^q$

$$\begin{aligned} &\exp\bigl(-2\Delta(\mathtt{r}_0, \mathtt{x}) - 8\mathrm{e}^{-\mathtt{x}}\bigr) I\!P\bigl(\boldsymbol{\gamma} \in A\bigr) - \mathrm{e}^{-\mathtt{x}} \\ &\quad \leq I\!P\bigl(\breve{D}_0 (\boldsymbol{\vartheta} - \boldsymbol{\theta}^\circ) \in A \,\big|\, \boldsymbol{Y}\bigr) \\ &\qquad \leq \exp\bigl(2\Delta(\mathtt{r}_0, \mathtt{x}) + 5\mathrm{e}^{-\mathtt{x}}\bigr) I\!P\bigl(\boldsymbol{\gamma} \in A\bigr), \end{aligned}$$

where $\boldsymbol{\gamma}$ is a standard Gaussian vector in $\mathbb{R}^q$.

The condition "$\Delta(\mathtt{r}_0, \mathtt{x})$ is small" yields the desirable BvM result, that is, the posterior measure after centering and standardization is close in total variation to the standard normal law. The classical asymptotic results immediately follow for many classical models (see discussion in Section 4). The next corollary extends the previous result by using empirically computable objects.

Corollary 1. *Under the conditions of Theorem 1 for any measurable set $A \subseteq \mathbb{R}^q$ a random event $\Omega(\mathtt{x})$ of a dominating probability at least $1 - 4\mathrm{e}^{-\mathtt{x}}$*

$$\begin{aligned} &\exp\bigl(-2\Delta(\mathtt{r}_0, \mathtt{x}) - 8\mathrm{e}^{-\mathtt{x}}\bigr) \bigl\{ I\!P\bigl(\boldsymbol{\gamma} \in A\bigr) - \tau \bigr\} - \mathrm{e}^{-\mathtt{x}} \\ &\quad \leq I\!P\bigl(\mathfrak{S}^{-1} (\boldsymbol{\vartheta} - \overline{\boldsymbol{\vartheta}}) \in A \,\big|\, \boldsymbol{Y}\bigr) \\ &\qquad \leq \exp\bigl(2\Delta(\mathtt{r}_0, \mathtt{x}) + 5\mathrm{e}^{-\mathtt{x}}\bigr) \bigl\{ I\!P\bigl(\boldsymbol{\gamma} \in A\bigr) + \tau \bigr\}, \end{aligned}$$

where $\boldsymbol{\gamma}$ *is a standard Gaussian vector in* $\mathbb{R}^q$ *and*

$$\tau \stackrel{\text{def}}{=} \frac{1}{2}\Big(q\Delta^2(\mathtt{r}_0,\mathtt{x}) + \big\{1+\Delta(\mathtt{r}_0,\mathtt{x})\big\}^2\Delta^2(\mathtt{r}_0,\mathtt{x})\Big).$$

This corollary is important as in practical applications we do not know matrix $\breve{D}_0$ and vector $\boldsymbol{\theta}^\circ$, but matrix $\mathfrak{S}^{-1}$ and vector $\overline{\boldsymbol{\vartheta}}$ can be computed by numerical computations. If dimension q is fixed the result becomes informative under the condition "$\Delta(\mathtt{r}_0,\mathtt{x})$ is small". Moreover, the statement can be extended to situations when the target dimension q grows but $\Delta(\mathtt{r}_0,\mathtt{x})\,q^{1/2}$ is still small.

2.3 Extension to a flat Gaussian prior

The previous results for a non-informative prior can be extended to the case of a flat prior $\Pi(d\boldsymbol{\upsilon})$. To be more specific we restrict ourselves to the case of a Gaussian prior. This is a prototypic situation because any smooth prior can be locally approximated by a Gaussian one. Without loss of generality the prior mean will be set to zero:

$$\Pi = \mathcal{N}(0, G^{-2})$$

with the density

$$\pi(\boldsymbol{\upsilon}) \propto \exp\big\{-\|G\boldsymbol{\upsilon}\|^2/2\big\}$$

for some positive symmetric matrix G^2.

The non-informative prior can be viewed as a limiting case of a Gaussian prior as $G \to 0$. We are interested in quantifying this relation. How small should G be to ensure the BvM result? To explain the result, we first consider the Gaussian case when $I\!P_{\boldsymbol{\upsilon}} = \mathcal{N}(\boldsymbol{\upsilon}, \mathcal{D}_0^{-2})$ and $\boldsymbol{\upsilon}^*$ is the true point. It is well known that in this situation the non-informative prior leads to the Gaussian posterior $\mathcal{N}(\boldsymbol{\upsilon}^*, \mathcal{D}_0^{-2})$, while the Gaussian prior Π yields again the Gaussian posterior with the covariance $\mathcal{D}_G^{-2}$ for $\mathcal{D}_G^2 = \mathcal{D}_0^2 + G^2$ and mean $\boldsymbol{\upsilon}_G^* = \mathcal{D}_G^{-2}\mathcal{D}_0^2\boldsymbol{\upsilon}^*$. Therefore, the prior does not significantly affect the posterior if two Gaussian measures $\mathcal{N}(\boldsymbol{\upsilon}^*, \mathcal{D}_0^{-2})$ and $\mathcal{N}(\boldsymbol{\upsilon}_G^*, \mathcal{D}_G^{-2})$ are nearly equivalent. The corresponding condition is represented in Lemma 8. It requires the values $\|\mathcal{D}_0^{-1}\mathcal{D}_G^2\mathcal{D}_0^{-1} - \boldsymbol{I}_q\| = \|\mathcal{D}_0^{-1}G^2\mathcal{D}_0^{-1}\|$, $\operatorname{tr}(\mathcal{D}_0^{-1}\mathcal{D}_G^2\mathcal{D}_0^{-1} - \boldsymbol{I}_q)^2 = \operatorname{tr}(\mathcal{D}_0^{-1}G^2\mathcal{D}_0^{-1})^2$ and $\|\mathcal{D}_G(\boldsymbol{\upsilon}^* - \boldsymbol{\upsilon}_G^*)\| \asymp \|\mathcal{D}_G^{-1}G^2\boldsymbol{\upsilon}^*\|$ to be small.

Theorem 2. *Suppose the conditions of Theorem 1. Let also* $\Pi = \mathcal{N}(0, G^{-2})$ *be a Gaussian prior measure on* $I\!R^p$ *such that*

$$\|\mathcal{D}_0^{-1}G^2\mathcal{D}_0^{-1}\| \le \boldsymbol{\epsilon} \le 1/2, \quad \operatorname{tr}\big(\mathcal{D}_0^{-1}G^2\mathcal{D}_0^{-1}\big)^2 \le \delta^2, \quad \|\mathcal{D}_G^{-1}G^2\boldsymbol{\upsilon}^*\| \le \beta,$$

where δ *and* β *are given constants. Then it holds on a set* $\Omega(\mathtt{x})$ *of probability* $1 - 5\mathrm{e}^{-\mathtt{x}}$

$$I\!P\big(\breve{D}_0(\boldsymbol{\vartheta} - \boldsymbol{\theta}^\circ) \in A \,\big|\, \boldsymbol{Y}\big) \ge \exp\big(-2\Delta(\mathtt{r}_0,\mathtt{x}) - 8\mathrm{e}^{-\mathtt{x}}\big)\big\{I\!P\big(\boldsymbol{\gamma} \in A\big) - \tau\big\} - \mathrm{e}^{-\mathtt{x}},$$

$$I\!P\big(\breve{D}_0(\boldsymbol{\vartheta} - \boldsymbol{\theta}^\circ) \in A \,\big|\, \boldsymbol{Y}\big) \le \exp\big(2\Delta(\mathtt{r}_0,\mathtt{x}) + 5\mathrm{e}^{-\mathtt{x}}\big)\big\{I\!P\big(\boldsymbol{\gamma} \in A\big) + \tau\big\} + \mathrm{e}^{-\mathtt{x}},$$

where

$$\tau \stackrel{\text{def}}{=} \frac{1}{2}\sqrt{(1+\epsilon)\bigl(3\beta+\epsilon z_B(\mathtt{x})\bigr)^2+\delta^2}$$

and the quantile function $z_B(\mathtt{x})$ *is defined in* (33).

Similar conditions and results can be found in the literature for more specific models. In particular, Bontemps (2011) or Johnstone (2010) explored the Gaussian case; see Section 5.1 below for a more detailed comparison.

3 Infinite dimensional nuisance

This section describes how previous results can be extended to the case where the nuisance is infinite dimensional. More specifically, we consider the $(\boldsymbol{\theta},\boldsymbol{f})$-setup, where $\boldsymbol{\theta}\in\Theta\subseteq\mathbb{R}^q$ and $\boldsymbol{f}\in\mathcal{H}$ for some Hilbert space $\mathcal{H}$. Suppose that in $\mathcal{H}$ exists a countable basis $\boldsymbol{e}_1,\boldsymbol{e}_2,\ldots$. Then

$$\boldsymbol{f}=\boldsymbol{f}(\boldsymbol{\phi})=\sum_{j=1}^{\infty}\phi_j\boldsymbol{e}_j\in\mathcal{H},$$

where a vector $\boldsymbol{\phi}=\{\phi_j\}_{j=1}^{\infty}\in\ell_2$ and $\phi_j=\langle\boldsymbol{f},\boldsymbol{e}_j\rangle$.

Let the likelihood for the full model be $\mathcal{L}(\boldsymbol{\theta},\boldsymbol{f})$. Denote with $\boldsymbol{\upsilon}=(\boldsymbol{\theta},\boldsymbol{\phi})$

$$\mathcal{L}(\boldsymbol{\upsilon})=\mathcal{L}(\boldsymbol{\theta},\boldsymbol{\phi})=\mathcal{L}(\boldsymbol{\theta},\boldsymbol{f}(\boldsymbol{\phi})).$$

The underlying full and target parameters can be defined by maximizing the expected log-likelihood:

$$\boldsymbol{\upsilon}^*\stackrel{\text{def}}{=}\operatorname*{argmax}_{\boldsymbol{\upsilon}=(\boldsymbol{\theta},\boldsymbol{\phi})}I\!\!E\mathcal{L}(\boldsymbol{\upsilon}),\qquad \boldsymbol{\theta}^*\stackrel{\text{def}}{=}\Pi_0\boldsymbol{\upsilon}^*. \tag{11}$$

Also define the information matrix $\mathcal{D}_0^2$ for the full parameter $\boldsymbol{\upsilon}$ and the efficient information matrix $\breve{D}^2$ for the target $\boldsymbol{\theta}$:

$$\mathcal{D}_0^2\stackrel{\text{def}}{=}\nabla^2 I\!\!E[\mathcal{L}(\boldsymbol{\upsilon}^*)]\in Lin(\ell_2,\ell_2),$$
$$\breve{D}_0^2\stackrel{\text{def}}{=}\bigl(\Pi_0\mathcal{D}_0^{-2}\Pi_0^\top\bigr)^{-1}\in I\!R^{q\times q},$$

where $Lin(\ell_2,\ell_2)$ is the space of linear operators from ℓ_2 to ℓ_2.

We apply the sieve approach and use an uninformative finite dimensional prior for the parameters $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$. The question under consideration is how the sieve truncation affects the posterior properties.

Let $\boldsymbol{\eta} = \{\eta_j\}_{j=1}^m$ be the sieve projection of the full parameter $\boldsymbol{\phi}$. For notational convenience represent $\boldsymbol{\phi} = (\boldsymbol{\eta}, \varkappa)$ and the sieve approximation corresponds to $\varkappa \equiv 0$. Write the "true" point $\boldsymbol{\upsilon}^*$ in the form $\boldsymbol{\upsilon}^* = (\boldsymbol{\theta}^*, \boldsymbol{\eta}^*, \varkappa^*)$. Approximation of the functional nuisance parameter $\boldsymbol{\phi}$ by the m-dimensional parameter $\boldsymbol{\eta}$ leads to two sources of bias due to projection of the functional parameter onto the finite dimensional space spanned by the first m basis functions. The first one is caused by ignoring the trimmed component $\varkappa$. The "sieved" target $\boldsymbol{\theta}^*_{\mathtt{s}}$ defined as

$$\boldsymbol{\theta}^*_{\mathtt{s}} \stackrel{\text{def}}{=} \operatorname*{argmax}_{\boldsymbol{\theta}} \max_{\boldsymbol{\eta}} I\!\!E L(\boldsymbol{\theta}, \boldsymbol{\eta}, 0),$$

can be different from the truth $\boldsymbol{\theta}^*$. The other one is caused by change of the efficient Fisher information $\breve{D}^2$ by its sieve counterpart $\breve{D}^2_{\mathtt{s}}$. The bias terms can be bounded under smoothness assumptions on the model and on the functional nuisance parameter $\boldsymbol{f}$ using the standard methods of approximation theory. To avoid tedious calculus we simply assume a kind of consistency of the sieve approximation. The notation simplifies if we also assume that the basis $\boldsymbol{e}_m$ in the space $\mathcal{H}$ is selected to provide orthogonality of the corresponding block H^2 of the Fisher information matrix, that is, $H^2 = \boldsymbol{I}_{\boldsymbol{f}}$. Of course, this implies the same structure of its blocks $H^2_{\boldsymbol{\eta}}$ and $H^2_{\varkappa}$. Note that the general situation can be reduced to this orthogonal case by a simple linear transformation of the nuisance parameter $\boldsymbol{\phi}$. This allows to represent the total Fisher matrix in the form

$$\mathfrak{D}^2_0 = \begin{pmatrix} D^2_{\boldsymbol{\theta}} & A_{\boldsymbol{\theta}\boldsymbol{\eta}} & A_{\boldsymbol{\theta}\varkappa} \\ A^\top_{\boldsymbol{\theta}\boldsymbol{\eta}} & \boldsymbol{I}_{\boldsymbol{\eta}} & 0 \\ A^\top_{\boldsymbol{\theta}\varkappa} & 0 & \boldsymbol{I}_{\varkappa} \end{pmatrix}. \tag{12}$$

First we specify the required conditions. The first one is a kind of semi parametric identifiability and it allows to separate the target and the nuisance parameters. Formally it requires that the angle between two tangent subspaces for these parameters is separated away from zero:

($\boldsymbol{\mathcal{I}}^{\mathtt{s}}$) For $\nu < 1$, it holds

$$\|D^{-1}_{\boldsymbol{\theta}} A_{\boldsymbol{\theta}\boldsymbol{\eta}} A^\top_{\boldsymbol{\theta}\boldsymbol{\eta}} D^{-1}_{\boldsymbol{\theta}}\| \le \nu.$$

The smoothness conditions on $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ are expressed via the component $\varkappa^*$ of $\boldsymbol{\upsilon}^*$ and the block $A_{\boldsymbol{\theta}\varkappa}$ of $\mathfrak{D}^2_0$.

($\boldsymbol{B}$) It holds with $\rho_{\mathtt{s}}$, $b_{\mathtt{s}} \le 1/2$

$$\|D^{-1}_{\boldsymbol{\theta}} A_{\boldsymbol{\theta}\varkappa} \varkappa^*\| \le \rho_{\mathtt{s}},$$
$$\|D^{-1}_{\boldsymbol{\theta}} A_{\boldsymbol{\theta}\varkappa} A^\top_{\boldsymbol{\theta}\varkappa} D^{-1}_{\boldsymbol{\theta}}\| \le b_{\mathtt{s}} \le 1/2.$$

For validity of our results we will need that the value of m is fixed in a proper way ensuring that the values of $\rho_{\mathtt{s}}$ and $b_{\mathtt{s}}$ are sufficiently small. These values can be upper bounded under the usual smoothness conditions on $\boldsymbol{f}$, e.g. if $\boldsymbol{f}$ belongs to a Sobolev ball with a certain regularity; cf. Bontemps (2011); Bickel and Kleijn (2012); Castillo (2012). See also an example of computing the quantities $\rho_{\mathtt{s}}$ and $b_{\mathtt{s}}$ in Section 5.3 below.

Consider a non-informative sieve prior which is uniform on $(\boldsymbol{\theta}, \boldsymbol{\eta})$ sieve component of the full parameter and puts singular mass at point 0 for the remaining components of the nuisance parameter $\{\eta_j\}_{j=m+1}^{\infty}$. We focus on the posterior distribution of the target parameter. It is assumed that the conditions of Theorem 1 and Corollary 1 are fulfilled for the sieve prior. For the efficient Fisher information matrix $\breve{D}_{\mathtt{s}}^2$ and the vectors $\boldsymbol{\theta}_{\mathtt{s}}^*$ and $\boldsymbol{\theta}_{\mathtt{s}}^\circ$ defined as

$$\breve{D}_{\mathtt{s}}^2 \stackrel{\text{def}}{=} D_{\boldsymbol{\theta}}^2 - A_{\boldsymbol{\theta}\boldsymbol{\eta}} A_{\boldsymbol{\theta}\boldsymbol{\eta}}^\top,$$

$$\boldsymbol{\theta}_{\mathtt{s}}^\circ \stackrel{\text{def}}{=} \boldsymbol{\theta}_{\mathtt{s}}^* + \breve{D}_{\mathtt{s}}^{-1} \breve{\boldsymbol{\xi}}_{\mathtt{s}} = \boldsymbol{\theta}_{\mathtt{s}}^* + \breve{D}_{\mathtt{s}}^{-1} \bigl(\nabla_{\boldsymbol{\theta}} - A_{\boldsymbol{\theta}\boldsymbol{\eta}} \nabla_{\boldsymbol{\eta}} \bigr),$$

Theorem 1 ensures the BvM result for the sieve non-informative prior on $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$: the $\boldsymbol{\theta}$-posterior is approximated by the Gaussian measure $\mathcal{N}(\boldsymbol{\theta}_{\mathtt{s}}^\circ, \breve{D}_{\mathtt{s}}^{-2})$. The main question is whether the side truncation introduces a significant bias in the posterior distribution. For the full semiparametric model, define

$$\breve{D}^2 \stackrel{\text{def}}{=} D_{\boldsymbol{\theta}}^2 - A_{\boldsymbol{\theta}\boldsymbol{\phi}} A_{\boldsymbol{\theta}\boldsymbol{\phi}}^\top,$$

$$\boldsymbol{\theta}^\circ \stackrel{\text{def}}{=} \boldsymbol{\theta}^* + \breve{D}^{-1} \breve{\boldsymbol{\xi}} = \boldsymbol{\theta}^* + \breve{D}^{-1} (\nabla_{\boldsymbol{\theta}} - A_{\boldsymbol{\theta}\boldsymbol{\phi}} \nabla_{\boldsymbol{\phi}}).$$

The vector $\boldsymbol{\theta}^\circ$ is the efficient score and $\breve{D}^2$ is the efficient Fisher matrix and they naturally appear in the infinite dimensional Gaussian case as the posterior mean and influence matrix for the improper non-informative prior. The next result accomplishes Theorem 1. Under our identifiability condition $(\mathcal{I}^{\mathtt{s}})$ and the smoothness condition (B), it allows to measure the distance between the Gaussian measure $\mathcal{N}(\boldsymbol{\theta}_{\mathtt{s}}^\circ, \breve{D}_{\mathtt{s}}^{-2})$ which approximates the sieve posterior, with the Gaussian measure $\mathcal{N}(\boldsymbol{\theta}^\circ, \breve{D}^{-2})$ corresponding to the full-dimensional prior. Due to Lemma 8 these two measures are close to each other if the ratio of two matrices $\breve{D}_{\mathtt{s}}^{-1} \breve{D}^2 \breve{D}_{\mathtt{s}}^{-1}$ is close to identity and the normalized mean difference $\breve{D}(\boldsymbol{\theta}^\circ - \boldsymbol{\theta}_{\mathtt{s}}^\circ)$ is small.

Theorem 3. *Consider a semiparametric model with a quasi log-likelihood $L(\boldsymbol{\theta}, \boldsymbol{\phi})$. The true value $(\boldsymbol{\theta}^*, \boldsymbol{\phi}^*)$ is given by (11). Let $\mathcal{D}_0^2 \stackrel{\text{def}}{=} -\nabla^2 I\!\!E L(\boldsymbol{\theta}^*, \boldsymbol{\phi}^*)$ be the corresponding Fisher operator. Suppose that the nuisance parameter $\boldsymbol{\phi}$ is rescaled to ensure that the corresponding $\boldsymbol{\phi}$-block of $\mathcal{D}_0^2$ is identity. Let $(\boldsymbol{\eta}, 0)$ be a sieve approximation of the functional nuisance parameter $\boldsymbol{\phi} = (\boldsymbol{\eta}, \varkappa)$, and (12) be the related block representation of $\mathcal{D}_0^2$. Suppose the identifiability condition $(\mathcal{I}^{\mathtt{s}})$ and the smoothness condition (B). Then the efficient Fisher information matrices $\breve{D}^2 = D_{\boldsymbol{\theta}}^2 - A_{\boldsymbol{\theta}\boldsymbol{\phi}} A_{\boldsymbol{\theta}\boldsymbol{\phi}}^\top$ and $\breve{D}_{\mathtt{s}}^2 = D_{\boldsymbol{\theta}}^2 - A_{\boldsymbol{\theta}\boldsymbol{\eta}} A_{\boldsymbol{\theta}\boldsymbol{\eta}}^\top$ in the full and sieve models are related by*

$$\bigl\| \breve{D}_{\mathtt{s}}^{-1} \breve{D}^2 \breve{D}_{\mathtt{s}}^{-1} - \boldsymbol{I}_q \bigr\| \leq (1-\nu)^{-1} \rho_{\mathtt{s}},$$

$$\mathrm{tr}\bigl\{ (\breve{D}_{\mathtt{s}}^{-1} \breve{D}^2 \breve{D}_{\mathtt{s}}^{-1} - \boldsymbol{I}_q)^2 \bigr\} \leq (1-\nu)^{-2} q\, \rho_{\mathtt{s}}^2.$$

The target parameter $\boldsymbol{\theta}^*$ *and its sieve counterpart* $\boldsymbol{\theta}^*_{\mathtt{s}}$ *are related by*

$$\bigl\|\breve{D}(\boldsymbol{\theta}^* - \boldsymbol{\theta}^*_{\mathtt{s}})\bigr\| \leq (1-\nu)^{-1} b_{\mathtt{s}} + \delta(\mathtt{r}_{\mathtt{s}})\mathtt{r}_{\mathtt{s}}\,, \tag{13}$$

where $\mathtt{r}_{\mathtt{s}} = \|\varkappa^*\|$. *Moreover, if* $\boldsymbol{\theta}^\circ = \boldsymbol{\theta}^* + \breve{D}^{-1}\breve{\boldsymbol{\xi}}$ *and* $\boldsymbol{\theta}^\circ_{\mathtt{s}} = \boldsymbol{\theta}^*_{\mathtt{s}} + \breve{D}^{-1}_{\mathtt{s}}\breve{\boldsymbol{\xi}}_{\mathtt{s}}$, *then it holds with* $\boldsymbol{\xi}_\varkappa \stackrel{\mathrm{def}}{=} \rho_{\mathtt{s}}^{-1} D_{\boldsymbol{\theta}}^{-1} A_{\boldsymbol{\theta}\varkappa} \nabla_\varkappa$

$$\bigl\|\breve{D}(\boldsymbol{\theta}^\circ - \boldsymbol{\theta}^\circ_{\mathtt{s}})\bigr\| \leq (1-\nu)^{-1}\rho_{\mathtt{s}}\bigl(\|\breve{\boldsymbol{\xi}}_{\mathtt{s}}\| + \|\boldsymbol{\xi}_\varkappa\|\bigr) + (1-\nu)^{-1} b_{\mathtt{s}} + \delta(\mathtt{r}_{\mathtt{s}})\mathtt{r}_{\mathtt{s}}\,. \tag{14}$$

Finally, under (ED_0) *and* $(\mathcal{I})$, *it holds on a set* $\Omega(\mathtt{x})$ *with* $\mathbb{P}\bigl(\Omega(\mathtt{x})\bigr) \geq 1 - 4\mathrm{e}^{\mathtt{x}}$

$$\bigl\|\breve{D}(\boldsymbol{\theta}^\circ - \boldsymbol{\theta}^\circ_{\mathtt{s}})\bigr\| \leq 2\mathfrak{a}(1-\nu)^{-1}\rho_{\mathtt{s}}\bigl(q^{1/2} + 2\mathtt{x}\bigr) + (1-\nu)^{-1} b_{\mathtt{s}} + \delta(\mathtt{r}_{\mathtt{s}})\mathtt{r}_{\mathtt{s}}\,. \tag{15}$$

We conclude that the sieve prior does a good job if the quantities $q^{1/2}\rho_{\mathtt{s}}$ and $b_{\mathtt{s}}$ are small and $\|\varkappa^*\|$ is not large.

4 The i.i.d. case and critical dimension

This section comments on how the previously obtained general results can be linked to the classical asymptotic results in the statistical literature. The nice feature of the whole approach based on the local bracketing is that all the results are stated under the same list of conditions: once checked one can directly apply any of the mentioned results. Typical examples include i.i.d., generalized linear models (GLM), and median regression models. Here we briefly discuss how the BvM result can be applied to one typical case, namely, to an i.i.d. experiment.

Let $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^\top$ be an i.i.d. sample from a measure P. Here we suppose the conditions of Section 5.1 in Spokoiny (2012) on P and $(P_{\boldsymbol{\upsilon}})$ to be fulfilled. We admit that the parametric assumption $P \in (P_{\boldsymbol{\upsilon}}, \boldsymbol{\upsilon} \in \Upsilon)$ can be misspecified and consider the asymptotic setup with the full dimension $p = p_n$ which depends on n and grows to infinity as $p_n \to \infty$.

Theorem 4. *Suppose the conditions of Section 5.1 in Spokoiny (2012). Let also* $p_n \to \infty$ *and* $p_n^3/n \to 0$. *Then the result of Theorem 1 holds with* $\Delta(\mathtt{r}_0, \mathtt{x}) = \mathtt{C}\sqrt{p_n^3/n}$, $\mathcal{D}_0^2 = n\mathbb{F}_{\boldsymbol{\upsilon}^*}$, *where* $\mathbb{F}_{\boldsymbol{\upsilon}^*}$ *is the Fisher information of* $(P_{\boldsymbol{\upsilon}})$ *at* $\boldsymbol{\upsilon}^*$.

A similar result about asymptotic normality of the posterior in a linear regression model can be found in Ghosal (1999). However, the convergence is proved under the condition $p_n^4 \log(p_n)/n \to 0$ which appears to be too strong. Ghosal (2000) showed that the dimensionality constraint can be relaxed to $p_n^3/n \to 0$ for exponential models with a product structure. Boucheron and Gassiat (2009) proved the BvM result in a specific class of i.i.d. model with discrete probability distribution under the condition $p_n^3/n \to 0$. Further examples and the related conditions for Gaussian models are presented in Johnstone (2010).

4.1 Critical dimension

This section discusses the issue of a *critical dimension*. Namely we show that the condition $p_n = o(n^{1/3})$ in Theorem 4 for the validity of the BvM result cannot be dropped or relaxed in a general situation. Namely, we present an example for which $p_n^3/n \geq \beta^2 > 0$ and the posterior distribution does not concentrate around MLE.

Let n and p_n be such that $M_n = n/p_n$ is an integer. We consider a simple Poissonian model with $Y_i \sim \mathrm{Poisson}(\upsilon_j)$ for $i \in \mathcal{I}_j$, where $\mathcal{I}_j \stackrel{\mathrm{def}}{=} \{i : \lceil i/M_n \rceil = j\}$ for $j = 1, \ldots, p_n$ and $\lceil x \rceil$ is the nearest integer greater or equal to x. Let also $u_j = \log \upsilon_j$ be the canonical parameter. The log-likelihood $\mathcal{L}(\boldsymbol{u})$ with $\boldsymbol{u} = (u_1, \ldots, u_{p_n})$ reads as

$$\mathcal{L}(\boldsymbol{u}) = \sum_{j=1}^{p_n} \bigl(Z_j u_j - M_n e^{u_j} \bigr),$$

where

$$Z_j \stackrel{\mathrm{def}}{=} \sum_{i \in \mathcal{I}_j} Y_i \,.$$

We consider the problem of estimating the mean of the u_j's:

$$\theta = \frac{1}{p_n} \bigl(u_1 + \ldots + u_{p_n} \bigr).$$

Below we study this problem in the asymptotic setup with $p_n \to \infty$ as $n \to \infty$ when the underlying measure $I\!P$ corresponds to $u_1^* = \ldots = u_{p_n}^* = u^*$ for some u^* yielding $\theta^* = u^*$. The value u^* will be specified later. We consider an i.i.d. exponential prior on the parameters υ_j of Poisson distribution:

$$\upsilon_j \sim \mathrm{Exp}(\mu).$$

Below we allow that μ may depend on n. Our results are valid for $\mu \leq \mathtt{C} \sqrt{\frac{n}{\log n}}$. The posterior is Gamma distributed:

$$\upsilon_j \,\big|\, \boldsymbol{Y} \sim \mathrm{Gamma}(\alpha_j, \mu_j),$$

where $\alpha_j = 1 + \sum_{i \in \mathcal{I}_j} Y_i$, $\mu_j = \frac{\mu}{M_n \mu + 1}$.

First we describe the profile maximum likelihood estimator $\widetilde{\theta}_n$ of the target parameter θ. The MLE for the full parameter $\boldsymbol{\upsilon}$ reads as $\widetilde{\boldsymbol{\upsilon}} = (\widetilde{\upsilon}_1, \ldots, \widetilde{\upsilon}_{p_n})^\top$ with

$$\widetilde{\upsilon}_j = Z_j / M_n.$$

Thus, the profile MLE $\widetilde{\theta}_n$ reads as

$$\widetilde{\theta}_n = \frac{1}{p_n} \sum_{j=1}^{p_n} \log(\widetilde{\upsilon}_j).$$

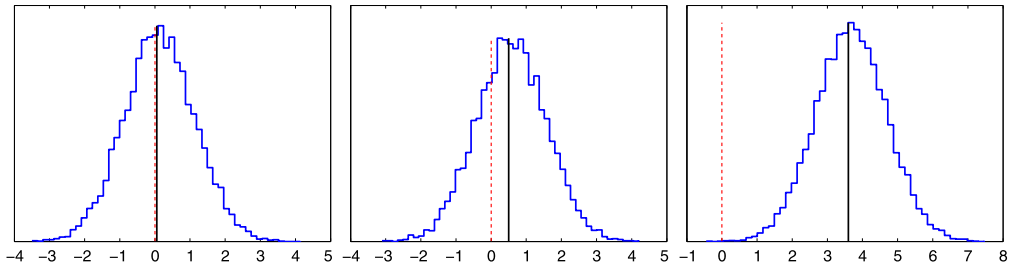

Figure 1: Posterior distribution of $\beta_n^{-1} p_n(\theta - \widetilde{\theta}_n)$ for $\beta_n = 1/\log(p_n)$, $\beta_n = 1$, and $\beta_n = \log(p_n)$. Solid line is for posterior mean and dashed line is for true mean.

Furthermore, the efficient Fisher information $\breve{D}_0^2$ is equal to $p_n^{-1} n$; see Lemma 11 below. As $\widetilde{\theta}_n$ is the profile MLE it is efficient with the asymptotic variance equal to $\breve{D}_0^{-2}$.

Theorem 5. *Let* $Y_i \sim \mathrm{Poisson}(\upsilon^*)$ *for all* $i = 1, \ldots, n$, $\upsilon^* = 1/p_n$. *Then*

1. *If* $p_n^3/n \to 0$ *as* $p_n \to \infty$, *then*

$$p_n^{1/2} n^{1/2} (\theta - \widetilde{\theta}_n) \,\big|\, \boldsymbol{Y} \xrightarrow{w} \mathcal{N}(0, 1).$$

2. *Let* $p_n^3/n \equiv \beta > 0$. *Then*

$$p_n^{1/2} n^{1/2} (\theta - \widetilde{\theta}_n) \,\big|\, \boldsymbol{Y} \xrightarrow{w} \mathcal{N}(\beta/2, 1).$$

3. *If* $p_n^3/n \to \infty$, *but* $p_n^4/n^{3/2} \to 0$, *then*

$$p_n^{1/2} n^{1/2} (\theta - \widetilde{\theta}_n) \,\big|\, \boldsymbol{Y} \xrightarrow{w} \infty.$$

We carried out a series of experiments to numerically demonstrate the results of Theorem 5. The dimension of parameter space was fixed $p_n = 10000$. Three cases were considered:

1. $p_n^{3/2}/n^{1/2} = \frac{1}{\log p_n}$, which corresponds to $p_n^3/n \to 0, n \to \infty$.
2. $p_n^{3/2}/n^{1/2} \equiv 1$.
3. $p_n^{3/2}/n^{1/2} = \log p_n$, which corresponds to $p_n^3/n \to \infty, n \to \infty$.

For each sample 10000 realizations of $\boldsymbol{Y}$ were generated from the exponential distribution $\mathrm{Exp}(\upsilon_*)$ and so were corresponding posterior values $\theta \,\big|\, \boldsymbol{Y}$. The resulting posterior distribution for three cases is demonstrated in Figure 1. It can be easily seen that the results of Theorem 5 are numerically confirmed.

5 Examples

This section presents a number of examples illustrating the general results of Section 2.

5.1 Linear Gaussian regression and a flat Gaussian prior

Let $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^\top$ be a random vector in $\mathbb{R}^n$ following the equation

$$\boldsymbol{Y} = \boldsymbol{f} + \boldsymbol{\varepsilon} = \Psi^\top \boldsymbol{\upsilon}^* + \boldsymbol{\varepsilon}, \tag{16}$$

where the errors $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^\top$ are independent zero mean, and Ψ is a given $p \times n$ design matrix. The mean vector $\boldsymbol{f} \in \mathbb{R}^n$ is unknown and the second equation in (16) means that it belongs to some given p-dimensional linear subspace in $\mathbb{R}^n$: $\boldsymbol{f} = \Psi^\top \boldsymbol{\upsilon}^*$ for an unknown target vector $\boldsymbol{\upsilon}^* \in \mathbb{R}^p$. We write the matrix Ψ in the form $\Psi = \{\Psi_1, \ldots, \Psi_n\}$, so that $f_i = \Psi_i^\top \boldsymbol{\upsilon}^*$. Below we suppose that $n > p$ and the rank of Ψ is p, or, equivalently, the rows of Ψ are linearly independent vectors in $\mathbb{R}^n$.

First, we consider the Gaussian case, i.e. $\varepsilon_i \in \mathcal{N}(0, \sigma_n^2 \boldsymbol{I}_n)$, $i = 1, \ldots, n$, with $\boldsymbol{I}_n$ being the $n \times n$ identity matrix. The variance of observations σ_n^2 is known but may depend on sample size n. For ease of comparison, we assume that the design matrix Ψ fulfills $\Psi^\top \Psi = \boldsymbol{I}_p$. In our notation, this implies $\mathcal{D}_0^2 = \sigma_n^{-2} \boldsymbol{I}_p$.

For a Gaussian prior, the posterior is exactly Gaussian and the Fisher and Wilks approximations are also exact. If a non-informative prior is used then the only necessary condition for validity of the BvM result is $p = p_n = o(n)$; see Bontemps (2011). Johnstone (2010) showed that the BvM result can be extended to the situation with a growing parameter dimension $p = p_n$ and a flat Gaussian prior with a covariance matrix $\tau_n^2 \boldsymbol{I}_n$ under the condition "$(\sigma_n/\tau_n)^4 p_n$ is small" and "$(\sigma_n/\tau_n^2)\|\boldsymbol{\upsilon}^*\|$ is small". Our results from Section 2 cover the case of a flat Gaussian prior and the only conditions to check for the BvM Theorem are that "$\operatorname{tr}(\mathcal{D}_0^{-2} G^2) = (\sigma_n/\tau_n)^4 p_n$ is small" and "$\|\mathcal{D}_G^{-1} G^2 \boldsymbol{\upsilon}^*\| = (\sigma_n/\tau_n^2)\|\boldsymbol{\upsilon}^*\|$ is small". One can see that our general results well apply to the case of growing dimension and are as sharp as the existing results obtained for a very special Gaussian case.

5.2 Linear non-Gaussian regression

Now consider a more general situation, when errors ε_i in (16) follow a general distribution with a given density $f(\cdot)$: $\varepsilon_i \sim P_f$. Denote $h(x) = \log f(x)$. The log-likelihood function for this problem reads as

$$\mathcal{L}(\boldsymbol{\upsilon}) = \sum_{i=1}^n \log f(Y_i - \Psi_i^\top \boldsymbol{\upsilon}) = \sum_{i=1}^n h(Y_i - \Psi_i^\top \boldsymbol{\upsilon}). \tag{17}$$

Suppose that $h(z)$ is twice continuously differentiable and let

$$\mathrm{h}^2 \stackrel{\mathrm{def}}{=} \int h''(z) f(z) dz < \infty.$$

If the model is correctly specified, then

$$\mathcal{D}_0^2 = -\nabla^2 I\!\!E \sum_{i=1}^n h(Y_i - \Psi_i^\top \upsilon^*) = \int h''(z) f(z) dz \cdot \sum_{i=1}^n \Psi_i \Psi_i^\top = \mathrm{h}^2 \sum_{i=1}^n \Psi_i \Psi_i^\top. \tag{18}$$

Similarly,

$$\mathcal{D}^2(\upsilon) = -\nabla^2 I\!\!E \sum_{i=1}^n h(Y_i - \Psi_i^\top \upsilon) = \sum_{i=1}^n \Psi_i \Psi_i^\top \int h''(z - \Psi_i^\top (\upsilon - \upsilon^*)) f(z) dz.$$

Conditions from Section 2.1 need to be checked in order to apply general results from Section 2.2. We assume some exponential moment conditions on distribution P_f:

(e_0) *There exist some constants ν_0 and $\mathtt{g}_1 > 0$ such that for $\varepsilon \sim P_f$ it holds*

$$\log I\!\!E \exp\bigl(\mu h'(\varepsilon)/\mathrm{h}\bigr) \le \nu_0^2 \mu^2/2, \qquad |\mu| \le \mathtt{g}_1.$$

Condition (e_0) effectively means that the distribution of error ε has exponentially decreasing tail. Under e_0 condition (ED_0) is satisfied due to the following lemma.

Lemma 1. *Assume (e_0) and let $\mathcal{V}_0^2 = \mathcal{D}_0^2 = \mathrm{h}^2 \sum_{i=1}^n \Psi_i \Psi_i^\top$. Then condition (ED_0) follows from (e_0) with this $\mathcal{V}_0^2$ and $\mathtt{g} = \mathtt{g}_1 N_1^{1/2}$, where*

$$N_1^{-1/2} \stackrel{\mathrm{def}}{=} \max_i \sup_{\gamma \in I\!R^p} \frac{\mathrm{h} |\Psi_i^\top \gamma|}{\|\mathcal{D}_0 \gamma\|}. \tag{19}$$

$(\mathcal{L}_0)$ and $(\mathcal{L}\mathtt{r})$ are also valid under the Lipschitz continuity assumption on $h''(\cdot)$.

Lemma 2. *Let*

$$|h''(z) - h''(z_0)| \le L |z - z_0|, \ z, z_0 \in \mathbb{R}.$$

Then

$$\delta(\mathtt{r}) \le \frac{L\,\mathtt{r}}{\mathrm{h}\, N_1^{1/2}},$$

where N_1 is defined by (19).

The next interesting question is checking the condition (ED_2). The stochastic part of the likelihood reads as follows:

$$\zeta(\upsilon) = \sum_{i=1}^n h(Y_i - \Psi_i^\top \upsilon) - I\!\!E h(Y_i - \Psi_i^\top \upsilon).$$

To check the condition (ED_2) we need to compute the Hessian of the $\zeta(\upsilon)$:

$$\nabla^2\zeta(\upsilon) = \sum_{i=1}^{n}\bigl\{h''(Y_i - \Psi_i^\top\upsilon) - I\!\!Eh''(Y_i - \Psi_i^\top\upsilon)\bigr\}\Psi_i\Psi_i^\top.$$

Finally, we need to impose the mild condition on marginal likelihood:

(e_2) *There exists some constant ν_0 and for every $\mathtt{r} > 0$ there exists $\mathtt{g}(\mathtt{r}) > 0$, such that for all numbers δ with $|\delta| \le N_2^{-1/2}\mathtt{r}$*

$$\log I\!\!E\exp\Bigl(\frac{\mu}{\mathfrak{s}_i}\bigl\{h''(Y_i+\delta) - I\!\!Eh''(Y_i+\delta)\bigr\}\Bigr) \le \frac{\nu_0^2\mu^2}{2}, \qquad |\mu| \le \mathtt{g}(\mathtt{r}),$$

where $\mathfrak{s}_i$ are some known values and

$$N_2^{-1/2} \stackrel{\mathrm{def}}{=} \max_i \sup_{\gamma\in I\!R^p}\frac{\mathfrak{s}_i|\Psi_i^\top\gamma|}{\|\mathcal{D}_0\gamma\|}.$$

Then

$$\begin{aligned}
&\sup_{\gamma_1,\gamma_2\in I\!R^p}\log I\!\!E\exp\left\{\frac{\mu}{\omega}\frac{\gamma_1^\top\nabla^2\zeta(\upsilon)\gamma_2}{\|\mathcal{D}_0\gamma_1\|\cdot\|\mathcal{D}_0\gamma_2\|}\right\}\\
&\quad= \sup_{\gamma_1,\gamma_2\in I\!R^p}\log I\!\!E\exp\left\{\frac{\mu}{\omega}\frac{\sum_{i=1}^n(h''(Y_i-\Psi_i^\top\upsilon)-I\!\!Eh''(Y_i-\Psi_i^\top\upsilon))\gamma_1^\top\Psi_i\Psi_i^\top\gamma_2}{\|\mathcal{D}_0\gamma_1\|\cdot\|\mathcal{D}_0\gamma_2\|}\right\}\\
&\quad\le \sup_{\gamma_1,\gamma_2\in I\!R^p}\frac{\nu_0^2\mu^2}{2\omega^2}\frac{\sum_{i=1}^n\mathfrak{s}_i^2|\Psi_i^\top\gamma_1|^2\cdot|\Psi_i^\top\gamma_2|^2}{\|\mathcal{D}_0\gamma_1\|^2\cdot\|\mathcal{D}_0\gamma_2\|^2}\\
&\quad\le \frac{\nu_0^2\mu^2}{2\omega^2}\frac{n}{N_2^2}.
\end{aligned}$$

It easily follows that $\omega = \frac{\sqrt{n}}{N_2}$ does the job, and (ED_2) follows. In regular situations N_2 is of order of sample size n and then $\omega \sim n^{-1/2}$.

Finally, turn to the semiparametric setup with $\upsilon = (\boldsymbol{\theta},\boldsymbol{\eta}) \in \mathbb{R}^p$. Assume the identifiability condition $(\mathcal{I})$ from Section 2.1. Then our general results yield the following semiparametric BvM Theorem for the linear model (16).

Theorem 6. *Let (e_0), (e_2), conditions of Lemma 2 and condition $(\mathcal{I})$ for matrix $\mathcal{D}_0^2$ from (18) hold. Then the results of Theorem 1 hold for linear model (16) with $\mathtt{r}_0^2 \ge \mathtt{C}(p+\mathtt{x})$ and*

$$\Delta(\mathtt{r}_0,\mathtt{x}) = \bigl\{\delta(\mathtt{r}_0) + 6\nu_0\, z_{I\!\!H}(\mathtt{x})\,\omega\bigr\}\,\mathtt{r}_0^2 \le \left\{\frac{L}{\mathtt{h}^2}\frac{\mathtt{r}_0}{N_1^{1/2}} + 6\nu_0\, z_{I\!\!H}(\mathtt{x})\,\frac{\sqrt{n}}{N_2}\right\}\mathtt{r}_0^2.$$

5.3 Semiparametric non-Gaussian linear regression

This section specifies the obtained results for the linear non-Gaussian model (16) with a semiparametric regression function

$$\boldsymbol{f}^* = \Psi^\top \boldsymbol{\theta}^* + \boldsymbol{g}^*, \tag{20}$$

where $\boldsymbol{\theta}^* \in \mathbb{R}^q$ is an unknown target vector and $\Psi = (\Psi_1, \ldots, \Psi_n)$ is a $q \times n$ matrix with $\Psi_i = (\psi_1(X_i), \ldots, \psi_q(X_i))^\top \in \mathbb{R}^q$ for given basis functions $\{\psi_j(\cdot), j = 1, \ldots, q\}$ and the design points X_i, $i = 1, \ldots, n$. Without loss of generality we can assume that these basis functions are design orthonormal:

$$\sum_i \psi_j(X_i)\psi_{j'}(X_i) = \delta_{j,j'}. \tag{21}$$

The general case can be reduced to this one by usual rotation and rescaling. Similarly, we suppose that the entries of the nuisance vector $\boldsymbol{g}^* = \{g^*(X_1), \ldots, g^*(X_n)\}^\top$ are the values at the design points X_i of a function g^* which is an element of a functional space. This means that $g^* = g(x) = \sum_{k=1}^\infty \eta_k \varphi_k(x)$ for a given functional basis $\{\varphi_k\}_{k=1}^\infty$ (e.g. Fourier, wavelet, etc.) and an infinite dimensional nuisance parameter vector $\boldsymbol{\eta} = \{\eta_1, \eta_2, \ldots, \eta_m, \ldots\}^\top$.

Furthermore, we assume that g^* is smooth, that is, it can be well approximated by finite sums $g_m^*(\cdot) = \sum_{k=1}^m \eta_k \varphi_k(\cdot)$ in the sense that

$$\|g^* - g_m^*\| \le \gamma_m. \tag{22}$$

For instance, if $\mathcal{F}_s$ is a Sobolev ball, that is,

$$g^* \in \mathcal{F}_s(\mathtt{C}) \stackrel{\text{def}}{=} \bigl\{ g(x) = \sum_{k=1}^\infty \eta_k \varphi_k(x) \colon \sum_{k=1}^\infty \eta_k^2 k^{2s} \le \mathtt{C} \bigr\},$$

then $\gamma_m \le (m+1)^{-s}$.

In addition we suppose the same smoothness condition for each basis function ψ_j for $j = 1, \ldots, q$, that is,

$$\|\psi_j - \psi_{j,m}\| \le \gamma_m, \tag{23}$$

where $\varphi_{j,m} = \Pi_m \varphi_j$ and Π_m is the projector on the spaced spanned by the first m basis functions $\varphi_1, \ldots, \varphi_m$.

To avoid the identifiability problem, we restrict the expansion of the function g to the first M coefficients for a large number M which may depend on the sample size n. For instance, one can take $M = n/\log(n)$. Alternatively, $M = n^a$ for some $a < 1$. Also, to simplify the presentation, we suppose that basis functions $\varphi_k(\cdot)$ are orthonormal in the sense that

$$\sum_i \varphi_{k'}(X_i)\varphi_k(X_i) = \delta_{k',k}. \tag{24}$$

Define also $\Phi_i = \{\varphi_1(X_i), \ldots, \varphi_q(X_i)\}^\top \in \mathbb{R}^q$ for $i = 1, 2, \ldots, n$. Then the full parameter of the model is $\upsilon = (\theta, \eta) \in I\!R^p$ for $p = q + M$, and the decomposition (20) can be rewritten as

$$f = \Xi \upsilon,$$

where $\Xi_i = (\Psi_i^\top, \Phi_i^\top)^\top$, $i = 1, \ldots, n$, and $\Xi = (\Xi_1, \ldots, \Xi_n)^\top$. The full Fisher information matrix reads as

$$\mathcal{D}_0^2 = \begin{pmatrix} D_0^2 & A_0 \\ A_0^\top & H_0^2 \end{pmatrix} = \mathrm{h}^2 \begin{pmatrix} \sum_{i=1}^n \Psi_i \Psi_i^\top & \sum_{i=1}^n \Psi_i \Phi_i^\top \\ \sum_{i=1}^n \Phi_i \Psi_i^\top & \sum_{i=1}^n \Phi_i \Phi_i^\top \end{pmatrix}.$$

Due to the orthogonality conditions (21) and (24), the blocks D_0^2 and H_0^2 are proportional to identity: $D_0^2 = \mathrm{h}^2 I_q$, $H_0^2 = \mathrm{h}^2 I_M$. In what follows, we assume $\mathrm{h} = 1$, the extension to the general case is trivial. The identifiability condition (9) can be written as

$$\|A_0 A_0^\top\| \leq \nu. \tag{25}$$

The condition (22) implies by orthogonality of the basis functions φ_m

$$\|A_{\theta\varkappa}\varkappa^*\|^2 = \|g^* - g_m^*\|^2 \leq \gamma_m^2.$$

Similarly, the smoothness condition (23) implies that each row of $A_{\theta\varkappa}$ is bounded in norm by γ_m. Therefore,

$$\left\|A_{\theta\varkappa} A_{\theta\varkappa}^\top\right\| \leq \mathrm{tr}(A_{\theta\varkappa} A_{\theta\varkappa}^\top) \leq q\gamma_m^2.$$

Theorem 7. *Consider the model* (16) *with the log-likelihood* (17) *and the semi parametric regression function from* (20). *Suppose the indentifiability condition* (25) *and the smoothness conditions* (22), (23). *Then the result of Theorem 3 holds with* $\rho_{\mathtt{s}} = b_{\mathtt{s}} = (1 - \nu)^{-1} q^{1/2} \gamma_m^2$.

Remark 1. In the case where all the functions $g^*, \psi_j, j = 1, \ldots, q$ are from the Sobolev ball with smoothness s, we can conclude that $\gamma_m \leq (m+1)^{-s}$ and $\rho_{\mathtt{s}} = b_{\mathtt{s}} \leq (1 - \nu)^{-1} q^{1/2} (m+1)^{-2s}$.

5.4 Generalized linear modeling

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

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

$$Y_i \sim P_{\Psi_i^\top \upsilon^*}. \tag{26}$$

The corresponding log-density of a GLM reads as

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

Under $\mathbb{P}_{\boldsymbol{\theta}^*}$ each observation Y_i follows (26), in particular, $\mathbb{E}Y_i = d'(\Psi_i^\top \upsilon^*)$. However, similarly to the previous sections, it is accepted that the parametric model (26) is misspecified. Response misspecification means that the vector $\boldsymbol{f} \stackrel{\text{def}}{=} \mathbb{E}\boldsymbol{Y}$ cannot be represented in the form $d'(\Psi^\top \upsilon)$ whatever υ is. The other sort of misspecification concerns the data distribution. The model (26) assumes that the Y_i's are independent and the marginal distribution belongs to the given parametric family P. In what follows, we only assume independent data having certain exponential moments. The target of estimation υ^* is defined by

$$\upsilon^* \stackrel{\text{def}}{=} \operatorname*{argmax}_{\upsilon} \mathbb{E}\mathcal{L}(\upsilon).$$

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

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

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

$$\upsilon^* = \operatorname*{argmax}_{\upsilon} \mathbb{E}\mathcal{L}(\upsilon) = \operatorname*{argmax}_{\upsilon} \sum \bigl\{ f_i \Psi_i^\top \upsilon - d(\Psi_i^\top \upsilon) \bigr\}$$

with $f_i = \mathbb{E}Y_i$. The function $\mathbb{E}\mathcal{L}(\upsilon)$ is concave as well and the vector υ^* is also well defined.

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

(e_0) *There exist some constants ν_0 and $\mathtt{g}_1 > 0$, and for every i a constant $\mathfrak{s}_i$ such that $\mathbb{E}\bigl(\varepsilon_i/\mathfrak{s}_i\bigr)^2 \le 1$ and*

$$\log \mathbb{E} \exp\bigl(\mu \varepsilon_i/\mathfrak{s}_i\bigr) \le \nu_0^2 \mu^2/2, \qquad |\mu| \le \mathtt{g}_1. \tag{27}$$

A natural candidate for $\mathfrak{s}_i$ is σ_i where $\sigma_i^2 = I\!\!E\varepsilon_i^2$ is the variance of ε_i. Under (27), introduce a $q \times q$ matrix $\mathcal{V}_0$ defined by

$$\mathcal{V}_0^2 \stackrel{\text{def}}{=} \sum \mathfrak{s}_i^2 \Psi_i \Psi_i^\top . \tag{28}$$

Condition (e_0) effectively means that each error term $\varepsilon_i = Y_i - I\!\!EY_i$ has some bounded exponential moments: for $|\lambda| \le \mathfrak{g}_1$, it holds $f(\lambda) \stackrel{\text{def}}{=} \log I\!\!E \exp\bigl(\lambda\varepsilon_i/\mathfrak{s}_i\bigr) < \infty$. In words, condition (e_0) requires a light (exponentially decreasing) tail for the marginal distribution of each ε_i.

Define also

$$N_1^{-1/2} \stackrel{\text{def}}{=} \max_i \sup_{\boldsymbol{\gamma} \in I\!R^p} \frac{\mathfrak{s}_i |\Psi_i^\top \boldsymbol{\gamma}|}{\|\mathcal{V}_0 \boldsymbol{\gamma}\|} . \tag{29}$$

Now conditions are satisfied due to following lemma, see Spokoiny (2012) for proof.

Lemma 3. *Assume* (e_0) *and let* $\mathcal{V}_0^2$ *be defined by* (28) *and* N_1 *by* (29). *Then condition* (ED_0) *follows from* (e_0) *with this* $\mathcal{V}_0^2$ *and* $\mathfrak{g} = \mathfrak{g}_1 N_1^{1/2}$. *Moreover, the stochastic component* $\zeta(\boldsymbol{\upsilon})$ *is linear in* $\boldsymbol{\upsilon}$ *and the condition* (ED_2) *is fulfilled with* $\omega(\mathbf{r}) \equiv 0$.

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

Lemma 4. *It holds*

$$-I\!\!EL(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^*) = \sum \bigl\{ d(\Psi_i^\top \boldsymbol{\upsilon}) - d(\Psi_i^\top \boldsymbol{\upsilon}^*) - d'(\Psi_i^\top \boldsymbol{\upsilon}^*) \Psi_i^\top (\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^*) \bigr\} = \mathcal{K}\bigl(I\!\!P_{\boldsymbol{\upsilon}^*}, I\!\!P_{\boldsymbol{\upsilon}}\bigr),$$

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

$$-I\!\!EL(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^*) = \|\mathcal{D}(\boldsymbol{\upsilon}^\circ)(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^*)\|^2/2,$$

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

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

The proof of this lemma can also be found in Spokoiny (2012). Define now the matrix $\mathcal{D}_0^2$ by

$$\mathcal{D}_0^2 \stackrel{\text{def}}{=} \mathcal{D}^2(\boldsymbol{\upsilon}^*) = \sum d''(\Psi_i^\top \boldsymbol{\upsilon}^*) \Psi_i \Psi_i^\top .$$

Let also $\mathcal{V}_0^2$ be defined by (28). Note that the matrices $\mathcal{D}_0^2$ and $\mathcal{V}_0^2$ coincide if the model $Y_i \sim P_{\Psi_i^\top \boldsymbol{\upsilon}^*}$ is correctly specified and $\mathfrak{s}_i^2 = d''(\Psi_i^\top \boldsymbol{\upsilon}^*)$. The matrix $\mathcal{D}_0^2$ describes a local

elliptic neighborhood of the central point υ^* in the form $\Upsilon_0(\mathtt{r}) = \{\upsilon : \|\mathcal{D}_0(\upsilon - \upsilon^*)\| \le \mathtt{r}\}$. If the matrix function $\mathcal{D}^2(\upsilon)$ is continuous in this vicinity $\Upsilon_0(\mathtt{r})$ then the value $\delta(\mathtt{r})$ measuring the approximation quality of $-I\!\!E L(\upsilon, \upsilon^*)$ by the quadratic function $\|\mathcal{D}_0(\upsilon - \upsilon^*)\|^2/2$ is small and the identifiability condition $(\mathcal{L}_0)$ is fulfilled on $\Upsilon_0(\mathtt{r})$. The following lemma gives bounds for $\delta(\mathtt{r})$.

Lemma 5. *Let $d''(z)$ be Lipschitz continuous:*

$$|d''(z) - d''(z_0)| \le L|z - z_0|, \quad z, z_0 \in \mathbb{R}$$

Then

$$\delta(\mathtt{r}) \le L\frac{\mathtt{r}}{N_2^{1/2}},$$

where

$$N_2^{-1/2} \stackrel{\mathrm{def}}{=} \max_i \sup_{\gamma \in I\!R^p} \frac{|\Psi_i^\top \gamma|}{d''(\Psi_i^\top \upsilon^*) \cdot \|\mathcal{D}_0 \gamma\|}.$$

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

Theorem 8. *Let (e_0) and conditions of Lemma 5 hold. Then the results of Theorem 1 hold for GLM with*

$$\Delta(\mathtt{r}_0, \mathtt{x}) \le L\frac{\mathtt{r}_0^3}{N_2^{1/2}}.$$

If the function $d(w)$ is quadratic then the approximation error δ vanishes as well and then quadratic approximation is valid globally, a localization step is not required. However, if $d(w)$ is not quadratic, the result applies only locally and it has to be accomplished with a large deviation bound. The GLM structure is helpful in the large deviation zone as well. Indeed, the identifiability condition $(\mathcal{L}\mathtt{r})$ easily follows from Lemma 4: it suffices to bound from below the matrix $\mathcal{D}(\upsilon)$ for $\upsilon \in \Upsilon_0(\mathtt{r})$:

$$\mathcal{D}(\upsilon) \ge \mathtt{b}(\mathtt{r})\mathcal{D}_0, \qquad \upsilon \in \Upsilon_0(\mathtt{r}).$$

An interesting question, similarly to the i.i.d. case, is the minimal radius $\mathtt{r}_0$ of the local vicinity $\Upsilon_0(\mathtt{r}_0)$ ensuring the desirable concentration property. The required value conditions are fulfilled for $\mathtt{r}^2 \ge \mathtt{r}_0^2 = \mathtt{C}(\mathtt{x} + p)$, where $\mathtt{C}$ only depends on $\nu_0, \mathtt{b}$, and $\mathtt{g}$. Thus, the results are valid if

$$\delta(\mathtt{r}_0)\mathtt{r}_0^2 = \mathtt{C}\frac{\mathtt{r}_0^3}{N_1^{1/2}} = \mathtt{C}\frac{(\mathtt{x} + p)^{3/2}}{N_1^{1/2}}$$

is small.

The GLM model also allows a semiparametric extension, i.e.

$$w_i = \Psi_i^\top \theta^* + g^*(X_i),$$

where $g^*(\cdot)$ is from a Sobolev class. This setup differs from Section 5.3 only by few technicalities and leads to similar theoretical results.

6 Supplementary

This section contains the imposed conditions and some supplementary statements which are of some interest by themselves.

6.1 Bracketing and upper function devices

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

$$\Upsilon_0(\mathtt{r}) \stackrel{\text{def}}{=} \bigl\{\upsilon : (\upsilon - \upsilon^*)^\top \mathcal{D}_0^2 (\upsilon - \upsilon^*) \le \mathtt{r}^2\bigr\}. \tag{30}$$

Define the quadratic processes $\mathbb{L}(\upsilon, \upsilon^*)$:

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

The next result states the local bracketing bound. The formulation assumes that some value $\mathtt{x}$ is fixed such that $\mathrm{e}^{-\mathtt{x}}$ is sufficiently small. If the dimension p is large, one can select $\mathtt{x} = \mathtt{C}\log(p)$. We assume that a value $\mathtt{r} = \mathtt{r}_0$ is fixed which separates the local and global zones.

Theorem 9. *Suppose the conditions* (ED_0), (ED_2), $(\mathcal{L}_0)$, *and* $(\mathcal{I})$ *from Section 2.1 hold for some* $\mathtt{r}_0 > 0$. *Then on a random set* $\Omega_{\mathtt{r}_0}(\mathtt{x})$ *of dominating probability at least* $1 - \mathrm{e}^{-\mathtt{x}}$

$$|L(\upsilon, \upsilon^*) - \mathbb{L}(\upsilon, \upsilon^*)| \le \Delta(\mathtt{r}_0, \mathtt{x}), \quad \upsilon \in \Upsilon_0(\mathtt{r}_0), \tag{31}$$

where

$$\begin{aligned} \Delta(\mathtt{r}_0, \mathtt{x}) &\stackrel{\text{def}}{=} \bigl\{\delta(\mathtt{r}_0) + 6\nu_0\, z_{\mathbb{H}}(\mathtt{x})\,\omega\bigr\}\,\mathtt{r}_0^2, \\ z_{\mathbb{H}}(\mathtt{x}) &\stackrel{\text{def}}{=} 2p^{1/2} + \sqrt{2\mathtt{x}} + \mathtt{g}^{-1}(\mathtt{g}^{-2}\mathtt{x} + 1)4p, \end{aligned} \tag{32}$$

and $\Upsilon_0(\mathtt{r}_0)$ *is defined in* (30). *Moreover, the random vector* $\boldsymbol{\xi} = \mathcal{D}_0^{-1}\nabla\mathcal{L}(\upsilon^*)$ *fulfills on a random set* $\Omega_B(\mathtt{x})$ *of dominating probability at least* $1 - 2\mathrm{e}^{-\mathtt{x}}$

$$\|\boldsymbol{\xi}\|^2 \le z_B^2(\mathtt{x}), \tag{33}$$

where $z_B^2(\mathtt{x}) \stackrel{\text{def}}{=} \mathtt{p}_B + 6\lambda_B \mathtt{x}$,

$$B \stackrel{\text{def}}{=} \mathcal{D}_0^{-1}\mathcal{V}_0^2\mathcal{D}_0^{-1}, \qquad \mathtt{p}_B \stackrel{\text{def}}{=} \operatorname{tr}(B), \qquad \lambda_B \stackrel{\text{def}}{=} \lambda_{\max}(B).$$

Furthermore, assume $(\mathcal{L}\mathtt{r})$ *with* $\mathtt{b}(\mathtt{r}) \equiv \mathtt{b}$ *yielding*

$$-I\!\!E L(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^*) \geq \mathtt{b}\,\|\mathcal{D}_0(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^*)\|^2$$

for each $\boldsymbol{\upsilon} \in \Upsilon \setminus \Upsilon_0(\mathtt{r}_0)$. *Let also*

$$\mathtt{r} \geq \frac{2}{\mathtt{b}} \Big\{ z_B(\mathtt{x}) + 6\nu_0\, z_{\mathbb{H}}\big(\mathtt{x} + \log(2\mathtt{r}/\mathtt{r}_0)\big)\, \omega \Big\}, \qquad \mathtt{r} \geq \mathtt{r}_0$$

with $z_{\mathbb{H}}(\mathtt{x})$ *from* (32). *Then,*

$$L(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^*) \leq -\mathtt{b}\,\|\mathcal{D}_0(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^*)\|^2/2, \qquad \boldsymbol{\upsilon} \in \Upsilon \setminus \Upsilon_0(\mathtt{r}_0). \tag{34}$$

holds on a random set $\Omega(\mathtt{x})$ *of probability at least* $1 - 4\mathrm{e}^{-\mathtt{x}}$.

The result (31) is an improved version of the approximation bound obtained in Spokoiny (2012), Theorem 3.1. The result (33) can be found in the supplement to Spokoiny (2012). The result (34) is very similar to Theorem 4.2 from Spokoiny (2012).

6.2 Tail posterior probability for full parameter space

The next step in our analysis is to check that $\boldsymbol{\upsilon}$ concentrates in a small vicinity $\Upsilon_0(\mathtt{r}_0)$ of the central point $\boldsymbol{\upsilon}^*$ with a properly selected $\mathtt{r}_0$. The concentration properties of the posterior will be described by using the random quantity

$$\rho^*(\mathtt{r}_0) = \frac{\int_{\Upsilon \setminus \Upsilon_0(\mathtt{r}_0)} \exp\{L(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^*)\} d\boldsymbol{\upsilon}}{\int_{\Upsilon_0(\mathtt{r}_0)} \exp\{L(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^*)\} d\boldsymbol{\upsilon}}.$$

Theorem 10. *Suppose the conditions of Theorem 9. Then it holds on* $\Omega_{\mathtt{r}_0}(\mathtt{x})$

$$\rho^*(\mathtt{r}_0) \leq \exp\{2\Delta(\mathtt{r}_0, \mathtt{x}) + \nu(\mathtt{r}_0)\}\; \mathtt{b}^{-p/2} I\!\!P\big(\|\boldsymbol{\gamma}\|^2 \geq \mathtt{b}\mathtt{r}_0^2\big), \tag{35}$$

with

$$\nu(\mathtt{r}_0) \stackrel{\text{def}}{=} -\log I\!\!P\big(\|\boldsymbol{\gamma} + \boldsymbol{\xi}\| \leq \mathtt{r}_0 \,\big|\, \boldsymbol{Y}\big).$$

If $\mathtt{r}_0 \geq z_B(\mathtt{x}) + z(p, \mathtt{x})$, *then on* $\Omega(\mathtt{x})$

$$\nu(\mathtt{r}_0) \leq 2\mathrm{e}^{-\mathtt{x}}. \tag{36}$$

This result together with Theorem 10 and Lemma 7 yields simple sufficient conditions on the value $\mathtt{r}_0$ which ensures the concentration of the posterior on $\Upsilon_0(\mathtt{r}_0)$.

Corollary 2. *Assume the conditions of Theorem 10. Then the additional inequality* $\mathtt{br}_0^2 \geq z^2(p, \mathtt{x} + \frac{p}{2}\log\frac{e}{b})$ *ensures on a random set* $\Omega(\mathtt{x})$ *of probability at least* $1 - 4\mathrm{e}^{-\mathtt{x}}$

$$\rho^*(\mathtt{r}_0) \leq \exp\{2\Delta(\mathtt{r}_0, \mathtt{x}) + 2\mathrm{e}^{-\mathtt{x}} - \mathtt{x}\}.$$

6.3 Tail posterior probability for target parameter

The next major step in our analysis is to check that $\boldsymbol{\theta}$ concentrates in a small vicinity $\Theta_0(\mathtt{r}_0) = \{\boldsymbol{\theta}\colon \|\breve{D}_0(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\| \leq \mathtt{r}_0\}$ of the central point $\boldsymbol{\theta}^* = \Pi_0 \boldsymbol{\upsilon}^*$ with a properly selected $\mathtt{r}_0$. The concentration properties of the posterior will be described by using the random quantity

$$\rho(\mathtt{r}_0) \stackrel{\mathrm{def}}{=} \frac{\int_\Upsilon \exp\{L(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^*)\}\pi(\boldsymbol{\upsilon})\, \mathbb{1}\{\boldsymbol{\theta} \notin \Theta_0(\mathtt{r}_0)\}d\boldsymbol{\upsilon}}{\int_\Upsilon \exp\{L(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^*)\}\pi(\boldsymbol{\upsilon})\, \mathbb{1}\{\boldsymbol{\theta} \in \Theta_0(\mathtt{r}_0)\}d\boldsymbol{\upsilon}}.$$

In what follows we suppose that prior is uniform, i.e. $\pi(\boldsymbol{\upsilon}) \equiv 1$, $\boldsymbol{\upsilon} \in \Upsilon$. This results in the following representation for $\rho(\mathtt{r}_0)$:

$$\rho(\mathtt{r}_0) = \frac{\int_\Upsilon \exp\{L(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^*)\}\, \mathbb{1}\{\boldsymbol{\theta} \notin \Theta_0(\mathtt{r}_0)\}d\boldsymbol{\upsilon}}{\int_\Upsilon \exp\{L(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^*)\}\, \mathbb{1}\{\boldsymbol{\theta} \in \Theta_0(\mathtt{r}_0)\}d\boldsymbol{\upsilon}}. \tag{37}$$

Obviously $I\!P(\boldsymbol{\vartheta} \notin \Theta_0(\mathtt{r}_0) \,|\, \boldsymbol{Y}) \leq \rho(\mathtt{r}_0)$. Therefore, small values of $\rho(\mathtt{r}_0)$ indicate a small posterior probability of the large deviation set $\{\boldsymbol{\vartheta} \notin \Theta_0(\mathtt{r}_0)\}$.

Theorem 11. *Suppose* (31). *Then for* $\mathtt{br}_0^2 \geq z^2(p, \mathtt{x} + \frac{p}{2}\log\frac{e}{b})$ *on* $\Omega(\mathtt{x})$ *of probability at least* $1 - 4\mathrm{e}^{-\mathtt{x}}$

$$\rho(\mathtt{r}_0) \leq \rho^*(\mathtt{r}_0) \leq \exp\{2\Delta(\mathtt{r}_0, \mathtt{x}) + 2\mathrm{e}^{-\mathtt{x}} - \mathtt{x}\}.$$

6.4 Local Gaussian approximation of the posterior: upper bound

It is convenient to introduce local conditional expectation: for a random variable η, define

$$I\!E^\circ \eta \stackrel{\mathrm{def}}{=} I\!E\Big[\eta\, \mathbb{1}\{\boldsymbol{\vartheta} \in \Theta_0(\mathtt{r}_0)\} \,\Big|\, \boldsymbol{Y}\Big].$$

The following theorem gives an exact statement about the upper bound of this posterior expectation. Let $\breve{\boldsymbol{\xi}}$ be from (7) and $\boldsymbol{\theta}^\circ = \boldsymbol{\theta}^* + \breve{D}_0^{-1}\breve{\boldsymbol{\xi}}$.

Theorem 12. *Suppose* (31). *Then for any* $f\colon \mathbb{R}^q \to \mathbb{R}_+$ *it holds on* $\Omega_{\mathtt{r}_0}(\mathtt{x})$

$$I\!E^\circ f\big(\breve{D}_0(\boldsymbol{\vartheta} - \boldsymbol{\theta}^\circ)\big) \leq \exp\{\Delta^+(\mathtt{r}_0, \mathtt{x})\}\, I\!E f(\boldsymbol{\gamma}), \tag{38}$$

where $\boldsymbol{\gamma} \sim \mathcal{N}(0, \boldsymbol{I}_q)$ *and*

$$\Delta^+(\mathtt{r}_0, \mathtt{x}) \stackrel{\text{def}}{=} 2\Delta(\mathtt{r}_0, \mathtt{x}) + \nu(\mathtt{r}_0) + \rho_f(\mathtt{r}_0),$$
$$\rho_f(\mathtt{r}_0) \stackrel{\text{def}}{=} \frac{\int_{\Upsilon \setminus \Upsilon_0(\mathtt{r}_0)} \exp\{L(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^*)\}\, f\bigl(\breve{D}_0(\boldsymbol{\theta} - \boldsymbol{\theta}^\circ)\bigr)\, d\boldsymbol{\upsilon}}{\int_{\Upsilon_0(\mathtt{r}_0)} \exp\{L(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^*)\}\, f\bigl(\breve{D}_0(\boldsymbol{\theta} - \boldsymbol{\theta}^\circ)\bigr)\, d\boldsymbol{\upsilon}}.$$

Define for random event $\eta \in A \subseteq \mathbb{R}^q$:

$$I\!\!P^\circ(\eta \in A) = I\!\!E^\circ\, \mathbb{1}\{\eta \in A\}.$$

The next result considers a special case with $f(\boldsymbol{u}) = \bigl|\boldsymbol{\lambda}^\top \boldsymbol{u}\bigr|^2$ and $f(\boldsymbol{u}) = \mathbb{1}(\boldsymbol{u} \in A)$ for any measurable set A.

Corollary 3. *For any* $\boldsymbol{\lambda} \in I\!R^q$, *it holds on* $\Omega_{\mathtt{r}_0}(\mathtt{x})$

$$I\!\!E^\circ \bigl|\boldsymbol{\lambda}^\top \breve{D}_0(\boldsymbol{\vartheta} - \boldsymbol{\theta}^\circ)\bigr|^2 \le \exp\{\Delta^+(\mathtt{r}_0, \mathtt{x})\} \|\boldsymbol{\lambda}\|^2.$$

For any measurable set $A \subseteq \mathbb{R}^q$, *it holds on* $\Omega_{\mathtt{r}_0}(\mathtt{x})$

$$I\!\!P^\circ\bigl(\breve{D}_0(\boldsymbol{\vartheta} - \boldsymbol{\theta}^\circ) \in A\bigr) \le \exp\{\Delta^+(\mathtt{r}_0, \mathtt{x})\} I\!\!P\bigl(\boldsymbol{\gamma} \in A\bigr). \tag{39}$$

On $\Omega(\mathtt{x})$ *one obtains*

$$\Delta^+(\mathtt{r}_0, \mathtt{x}) \le 2\Delta(\mathtt{r}_0, \mathtt{x}) + 2\mathrm{e}^{-\mathtt{x}} + 2\exp\{\Delta(\mathtt{r}_0, \mathtt{x}) + 4\mathrm{e}^{-\mathtt{x}} - \mathtt{x}\}.$$

The next corollary describes an upper bound for the posterior probability in case of changing of scaling.

Corollary 4. *Let* D_1 *be a symmetric* $q \times q$ *matrix such that* $\|I - D_1^{-1} \breve{D}_0^2 D_1^{-1}\| \le \alpha$. *Let also* $\widehat{\boldsymbol{\theta}} \in \mathbb{R}^q$ *be such that* $\|\breve{D}_0(\boldsymbol{\theta}^\circ - \widehat{\boldsymbol{\theta}})\| \le \beta$. *Then for any measurable set* $A \subseteq I\!R^q$, *it holds on* $\Omega(\mathtt{x})$ *with* $\boldsymbol{\delta}_0 \stackrel{\text{def}}{=} D_1(\boldsymbol{\theta}^\circ - \widehat{\boldsymbol{\theta}})$

$$\begin{aligned} I\!\!P^\circ\bigl(D_1(\boldsymbol{\vartheta} - \widehat{\boldsymbol{\theta}}) \in A\bigr) &\le \exp\{\Delta^+(\mathtt{r}_0, \mathtt{x})\} I\!\!P\bigl(D_1 \breve{D}_0^{-1} \boldsymbol{\gamma} + \boldsymbol{\delta}_0 \in A\bigr) \\ &\le \exp\{\Delta^+(\mathtt{r}_0, \mathtt{x})\} \Bigl(I\!\!P\bigl(\boldsymbol{\gamma} \in A\bigr) + \frac{1}{2}\sqrt{\alpha^2 q + (1+\alpha)^2 \beta^2}\Bigr). \end{aligned} \tag{40}$$

6.5 Local Gaussian approximation of the posterior: lower bound

Now we present a local lower bound for the posterior measure.

Theorem 13. *Suppose* (31). *Then for any* $f\colon \mathbb{R}^q \to \mathbb{R}_+$ *it holds on* $\Omega_{\mathtt{r}_0}(\mathtt{x})$

$$I\!\!E^\circ f\bigl(\breve{D}_0(\boldsymbol{\vartheta} - \boldsymbol{\theta}^\circ)\bigr) \ge \exp\{-\Delta^-(\mathtt{r}_0, \mathtt{x})\}\, I\!\!E\bigl\{f(\boldsymbol{\gamma})\, \mathbb{1}\bigl(\|\boldsymbol{\gamma} + \breve{\boldsymbol{\xi}}\| \le \mathtt{r}_0\bigr)\bigr\}, \tag{41}$$

where

$$\Delta^{-}(\mathtt{r}_0,\mathtt{x}) \stackrel{\text{def}}{=} 2\Delta(\mathtt{r}_0,\mathtt{x}) + \nu(\mathtt{r}_0) + \rho^*(\mathtt{r}_0) + 2\widetilde{\rho}_f(\mathtt{r}_0),$$
$$\widetilde{\rho}_f(\mathtt{r}_0) \stackrel{\text{def}}{=} \frac{\int_{\mathbb{R}^p \setminus \Upsilon_0(\mathtt{r}_0)} \exp\{\mathbb{L}(\boldsymbol{\upsilon},\boldsymbol{\upsilon}^*)\}\, f\bigl(\breve{D}_0(\boldsymbol{\theta}-\boldsymbol{\theta}^\circ)\bigr)\, d\boldsymbol{\upsilon}}{\int_{\Upsilon_0(\mathtt{r}_0)} \exp\{\mathbb{L}(\boldsymbol{\upsilon},\boldsymbol{\upsilon}^*)\}\, f\bigl(\breve{D}_0(\boldsymbol{\theta}-\boldsymbol{\theta}^\circ)\bigr)\, d\boldsymbol{\upsilon}}. \tag{42}$$

This result means that posterior measure can be bounded from below by the standard normal law up to (small) multiplicative and additive constants. As a corollary, we state the result for quadratic and indicator functions $f(\boldsymbol{u})$. The proof is similar to Corollary 4 and Corollary 3.

Corollary 5. *For any $\boldsymbol{\lambda} \in I\!R^q$, it holds on $\Omega_{\mathtt{r}_0}(\mathtt{x})$*

$$I\!E^\circ \bigl|\boldsymbol{\lambda}^\top \breve{D}_0(\boldsymbol{\vartheta}-\boldsymbol{\theta}^\circ)\bigr|^2 \geq \exp\{-\Delta^{-}(\mathtt{r}_0,\mathtt{x}) + \mathrm{e}^{-\mathtt{x}}\}\|\boldsymbol{\lambda}\|^2.$$

For any measurable set $A \subseteq \mathbb{R}^q$, it holds on $\Omega_{\mathtt{r}_0}(\mathtt{x})$

$$I\!P^\circ\bigl(\breve{D}_0(\boldsymbol{\vartheta}-\boldsymbol{\theta}^\circ) \in A\bigr) \geq \exp\{\Delta^{-}(\mathtt{r}_0,\mathtt{x})\} I\!P\bigl(\boldsymbol{\gamma} \in A\bigr) - \mathrm{e}^{-\mathtt{x}}. \tag{43}$$

Let D_1^2 be a symmetric $q \times q$ matrix such that $\|I - D_1^{-1}\breve{D}_0^2 D_1^{-1}\| \leq \alpha$ and let $\widehat{\boldsymbol{\theta}} \in \mathbb{R}^q$ be such that $\|\breve{D}_0(\boldsymbol{\theta}^\circ - \widehat{\boldsymbol{\theta}})\| \leq \beta$. Define $\boldsymbol{\delta}_0 \stackrel{\text{def}}{=} D_1(\boldsymbol{\theta}^\circ - \widehat{\boldsymbol{\theta}})$. Then for any measurable subset A in $I\!R^q$, it holds on $\Omega(\mathtt{x})$

$$\begin{aligned} I\!P^\circ\bigl(D_1(\boldsymbol{\vartheta}-\widehat{\boldsymbol{\theta}}) \in A\bigr) &\geq \exp\{\Delta^{-}(\mathtt{r}_0,\mathtt{x})\} I\!P\bigl(D_1\breve{D}_0^{-1}\boldsymbol{\gamma} + \boldsymbol{\delta}_0 \in A\bigr) - \mathrm{e}^{-\mathtt{x}} \\ &\geq \exp\{\Delta^{-}(\mathtt{r}_0,\mathtt{x})\}\Bigl\{I\!P\bigl(\boldsymbol{\gamma} \in A\bigr) - \frac{1}{2}\sqrt{\alpha^2 q + (1+\alpha)^2\beta^2}\Bigr\} - \mathrm{e}^{-\mathtt{x}}, \\ \Delta^{-}(\mathtt{r}_0,\mathtt{x}) &\leq 2\Delta(\mathtt{r}_0,\mathtt{x}) + 3\mathrm{e}^{-\mathtt{x}} + 4\exp\{\Delta(\mathtt{r}_0,\mathtt{x}) + 4\mathrm{e}^{-\mathtt{x}} - \mathtt{x}\}. \end{aligned}$$

7 Proofs

This appendix collects the proofs of the results.

7.1 Some inequalities for the normal law

This section collects some simple but useful facts about the properties of the multivariate standard normal distribution. Many similar results can be found in the literature, we present the proofs to keep the presentation self-contained. Everywhere in this section $\boldsymbol{\gamma}$ means a standard normal vector in $I\!R^q$.

Lemma 6. *For any $\boldsymbol{u} \in I\!R^q$, any unit vector $\boldsymbol{a} \in I\!R^q$, and any $z > 0$, it holds*

$$I\!P\bigl(\|\boldsymbol{\gamma}-\boldsymbol{u}\| \geq z\bigr) \leq \exp\{-z^2/4 + q/2 + \|\boldsymbol{u}\|^2/2\}, \tag{44}$$
$$I\!E\bigl\{|\boldsymbol{\gamma}^\top \boldsymbol{a}|^2\, \mathbb{1}\bigl(\|\boldsymbol{\gamma}-\boldsymbol{u}\| \geq z\bigr)\bigr\} \leq (2 + |\boldsymbol{u}^\top \boldsymbol{a}|^2)\exp\{-z^2/4 + q/2 + \|\boldsymbol{u}\|^2/2\}. \tag{45}$$

Proof. By the exponential Chebyshev inequality, for any $\lambda < 1$

$$\begin{aligned} I\!P\bigl(\|\boldsymbol{\gamma} - \boldsymbol{u}\| \geq z\bigr) &\leq \exp\bigl(-\lambda z^2/2\bigr) I\!E \exp\bigl(\lambda \|\boldsymbol{\gamma} - \boldsymbol{u}\|^2/2\bigr) \\ &= \exp\Bigl\{-\frac{\lambda z^2}{2} - \frac{q}{2}\log(1-\lambda) + \frac{\lambda}{2(1-\lambda)}\|\boldsymbol{u}\|^2\Bigr\}. \end{aligned}$$

In particular, with $\lambda = 1/2$, this implies (44). Further, for $\|\boldsymbol{a}\| = 1$

$$\begin{aligned} I\!E\bigl\{|\boldsymbol{\gamma}^\top \boldsymbol{a}|^2 \, 1\!\!1(\|\boldsymbol{\gamma} - \boldsymbol{u}\| \geq z)\bigr\} &\leq \exp\bigl(-z^2/4\bigr) I\!E\bigl\{|\boldsymbol{\gamma}^\top \boldsymbol{a}|^2 \exp\bigl(\|\boldsymbol{\gamma} - \boldsymbol{u}\|^2/4\bigr)\bigr\} \\ &\leq (2 + |\boldsymbol{u}^\top \boldsymbol{a}|^2) \exp\bigl(-z^2/4 + q/2 + \|\boldsymbol{u}\|^2/2\bigr) \end{aligned}$$

and (45) follows. □

The next result explains the concentration effect for the norm $\|\boldsymbol{\xi}\|^2$ of a Gaussian vector. We use a version from Spokoiny (2012).

Lemma 7. *For each* $\mathtt{x}$,

$$I\!P\bigl(\|\boldsymbol{\gamma}\| \geq z(q, \mathtt{x})\bigr) \leq \exp(-\mathtt{x}), \qquad I\!P\bigl(\|\boldsymbol{\gamma}\| \leq z_1(q, \mathtt{x})\bigr) \leq \exp(-\mathtt{x}),$$

where

$$z^2(q, \mathtt{x}) \stackrel{\text{def}}{=} q + \sqrt{6.6 q \mathtt{x}} \vee (6.6\mathtt{x}), \qquad z_1^2(q, \mathtt{x}) \stackrel{\text{def}}{=} q - 2\sqrt{q\mathtt{x}}.$$

The next lemma bounds from above the Kullback-Leibler divergence between two normal distributions.

Lemma 8. *Let* $I\!P = \mathcal{N}(\boldsymbol{b}, \Sigma)$ *and* $I\!P^\circ = \mathcal{N}(\boldsymbol{b}^\circ, \Sigma^\circ)$ *for some non-degenerate matrices* Σ *and* Σ°. *If*

$$\|\Sigma^{-1/2}\Sigma^\circ\Sigma^{-1/2} - I_q\| \leq \boldsymbol{\epsilon} \leq 1/2, \qquad \operatorname{tr}\bigl(\Sigma^{-1/2}\Sigma^\circ\Sigma^{-1/2} - I_q\bigr)^2 \leq \delta^2,$$

then

$$\begin{aligned} \mathcal{K}(I\!P, I\!P^\circ) = -I\!E_0 \log \frac{dI\!P^\circ}{dI\!P} &\leq \frac{\delta^2}{2} + \frac{1}{2}(\boldsymbol{b} - \boldsymbol{b}^\circ)^\top \Sigma^\circ (\boldsymbol{b} - \boldsymbol{b}^\circ) \\ &\leq \frac{\delta^2}{2} + \frac{1+\boldsymbol{\epsilon}}{2}(\boldsymbol{b} - \boldsymbol{b}^\circ)^\top \Sigma (\boldsymbol{b} - \boldsymbol{b}^\circ). \end{aligned}$$

For any measurable set $A \subset I\!R^q$, *it holds*

$$\bigl|I\!P(A) - I\!P^\circ(A)\bigr| \leq \sqrt{\mathcal{K}(I\!P, I\!P^\circ)/2} \leq \frac{1}{2}\sqrt{\delta^2 + (1+\boldsymbol{\epsilon})(\boldsymbol{b} - \boldsymbol{b}^\circ)^\top \Sigma (\boldsymbol{b} - \boldsymbol{b}^\circ)}.$$

Proof. The change of variables $\boldsymbol{u} = \Sigma^{-1/2}(\boldsymbol{x} - \boldsymbol{b})$ reduces the general case to the situation when $I\!P$ is standard normal in $I\!R^q$ while $P_1 = \mathcal{N}(\boldsymbol{\beta}, B)$ with $\boldsymbol{\beta} = \Sigma^{1/2}(\boldsymbol{b}^\circ - \boldsymbol{b})$ and $B \stackrel{\text{def}}{=} \Sigma^{-1/2}\Sigma^\circ\Sigma^{-1/2}$

$$2\log\frac{dI\!P^\circ}{dI\!P}(\boldsymbol{\gamma}) = \log\det(B) - (\boldsymbol{\gamma}-\boldsymbol{\beta})^\top B(\boldsymbol{\gamma}-\boldsymbol{\beta}) + \|\boldsymbol{\gamma}\|^2$$

with $\boldsymbol{\gamma}$ standard normal and

$$2\mathcal{K}(I\!P, I\!P^\circ) = -2I\!E_0 \log\frac{dI\!P^\circ}{dI\!P} = -\log\det(B) + \operatorname{tr}(B - \boldsymbol{I}_q) + \boldsymbol{\beta}^\top B\boldsymbol{\beta}.$$

Let a_j be the jth eigenvalue of $B - \boldsymbol{I}_q$. From $\|B - \boldsymbol{I}_q\| \le 1/2$ it follows $|a_j| \le 1/2$ and

$$\begin{aligned} 2\mathcal{K}(I\!P, I\!P^\circ) &= \boldsymbol{\beta}^\top B\boldsymbol{\beta} + \sum_{j=1}^q \bigl\{a_j - \log(1+a_j)\bigr\} \le \boldsymbol{\beta}^\top B\boldsymbol{\beta} + \sum_{j=1}^q a_j^2 \\ &\le \boldsymbol{\beta}^\top B\boldsymbol{\beta} + \operatorname{tr}(B - \boldsymbol{I}_q)^2 \le \boldsymbol{\beta}^\top B\boldsymbol{\beta} + \delta^2. \end{aligned}$$

This implies by Pinsker's inequality

$$\sup_A |I\!P(A) - I\!P^\circ(A)| \le \sqrt{\frac{1}{2}\mathcal{K}(I\!P, I\!P^\circ)} \le \frac{1}{2}\sqrt{\delta^2 + \boldsymbol{\beta}^\top B\boldsymbol{\beta}}$$

as required. □

7.2 Proof of Theorem 10

Define $\mathfrak{u}(\boldsymbol{\upsilon}) = \mathfrak{b}\,\|\mathcal{D}_0(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^*)\|^2/2$. Now, by a change of variables, one obtains

$$\begin{aligned} &\frac{\mathfrak{b}^{p/2}\det(\mathcal{D}_0)}{(2\pi)^{p/2}} \int_{\Upsilon\setminus\Upsilon_0(\mathfrak{r}_0)} \exp\bigl\{-\mathfrak{u}(\boldsymbol{\upsilon})\bigr\} d\boldsymbol{\upsilon} \\ &\quad\le \frac{\mathfrak{b}^{p/2}\det(\mathcal{D}_0)}{(2\pi)^{p/2}} \int_{\Upsilon\setminus\Upsilon_0(\mathfrak{r}_0)} \exp\bigl\{-\mathfrak{b}\,\|\mathcal{D}_0(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^*)\|^2/2\bigr\} d\boldsymbol{\upsilon} = I\!P\bigl(\|\boldsymbol{\gamma}\|^2 \ge \mathfrak{b}\mathfrak{r}_0^2\bigr). \end{aligned}$$

For the integral in the numerator of (37), it holds on $\Omega(\mathtt{x})$ by (34)

$$\int_{\Upsilon\setminus\Upsilon_0(\mathfrak{r}_0)} \exp\bigl\{L(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^*)\bigr\}\, d\boldsymbol{\upsilon} \le \int_{\Upsilon\setminus\Upsilon_0(\mathfrak{r}_0)} \exp\bigl\{-\mathfrak{u}(\boldsymbol{\upsilon})\bigr\}\, d\boldsymbol{\upsilon}.$$

For the integral in the denominator it holds

$$\begin{aligned} &\int_{\Upsilon_0(\mathfrak{r}_0)} \exp\bigl\{L(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^*)\bigr\}\, d\boldsymbol{\upsilon} \\ &\quad\ge \exp\{-\Delta(\mathfrak{r}_0, \mathtt{x}) - m(\boldsymbol{\xi})\} \int_{\Upsilon_0(\mathfrak{r}_0)} \exp\bigl\{\mathbb{L}(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^*) + m(\boldsymbol{\xi})\bigr\}\, d\boldsymbol{\upsilon}. \end{aligned} \tag{46}$$

Inequality (46) implies by definition of $\nu(\mathtt{r}_0)$:

$$\int_{\Upsilon_0(\mathtt{r}_0)} \exp\{L(\upsilon, \upsilon^*)\}\, d\upsilon \geq \exp\bigl\{-\Delta(\mathtt{r}_0, \mathtt{x}) - m(\boldsymbol{\xi}) - \nu(\mathtt{r}_0)\bigr\}. \tag{47}$$

The bound (47) for the local integral $\int_{\Upsilon_0(\mathtt{r}_0)} \exp\bigl\{L(\upsilon, \upsilon^*)\bigr\} d\upsilon$ implies that

$$\rho^*(\mathtt{r}_0) \leq \exp\bigl\{\Delta(\mathtt{r}_0, \mathtt{x}) + \nu(\mathtt{r}_0) + m(\boldsymbol{\xi})\bigr\} \int_{\Upsilon \setminus \Upsilon_0(\mathtt{r}_0)} \exp\bigl\{-\mathtt{u}(\upsilon)\bigr\} d\upsilon.$$

Finally

$$\exp\bigl\{m(\boldsymbol{\xi})\bigr\} = \exp\bigl\{-\|\boldsymbol{\xi}\|^2/2\bigr\}\, (2\pi)^{-p/2}\, \det(\mathcal{D}_0) \leq (2\pi)^{-p/2}\, \det(\mathcal{D}_0)$$

and the assertion (35) follows. The bound (36) is also straightforward:

$$\begin{aligned} \nu(\mathtt{r}_0) &= -\log I\!P\bigl(\|\boldsymbol{\gamma} + \boldsymbol{\xi}\| \leq \mathtt{r}_0 \,\big|\, \boldsymbol{Y}\bigr) \leq -\log I\!P\bigl(\|\boldsymbol{\gamma}\| + \|\boldsymbol{\xi}\| \leq \mathtt{r}_0 \,\big|\, \boldsymbol{Y}\bigr) \\ &\leq -\log I\!P\bigl(\|\boldsymbol{\gamma}\| \leq z(p, \mathtt{x}) \,\big|\, \boldsymbol{Y}\bigr) \leq 2\mathrm{e}^{-\mathtt{x}}. \end{aligned}$$

7.3 Proof of Theorem 11

Obviously $\bigl\{\boldsymbol{\theta} \notin \Theta_0(\mathtt{r}_0),\, \upsilon \in \Upsilon\bigr\} \subset \bigl\{\Upsilon \setminus \Upsilon_0(\mathtt{r}_0)\bigr\}$. Therefore, it holds for the integral in the numerator of (37) in view of (34)

$$\int_{\Upsilon} \exp\bigl\{L(\upsilon, \upsilon^*)\bigr\}\, \mathbb{1}\bigl\{\boldsymbol{\theta} \notin \Theta_0(\mathtt{r}_0)\bigr\}\, d\upsilon \leq \int_{\Upsilon \setminus \Upsilon_0(\mathtt{r}_0)} \exp\bigl\{L(\upsilon, \upsilon^*)\bigr\}\, d\upsilon.$$

For the denominator, the inclusion $\Upsilon_0(\mathtt{r}_0) \subset \bigl\{\boldsymbol{\theta} \in \Theta_0(\mathtt{r}_0), \upsilon \in \Upsilon\bigr\}$ and (34) imply

$$\int_{\Upsilon} \exp\bigl\{L(\upsilon, \upsilon^*)\bigr\}\, \mathbb{1}\bigl\{\boldsymbol{\theta} \in \Theta_0(\mathtt{r}_0)\bigr\}\, d\upsilon \geq \int_{\Upsilon_0(\mathtt{r}_0)} \exp\bigl\{L(\upsilon, \upsilon^*)\bigr\}\, d\upsilon.$$

Finally

$$\rho(\mathtt{r}_0) = \frac{\int_{\Upsilon} \exp\bigl\{L(\upsilon, \upsilon^*)\bigr\}\, \mathbb{1}\bigl\{\boldsymbol{\theta} \notin \Theta_0(\mathtt{r}_0)\bigr\} d\upsilon}{\int_{\Upsilon} \exp\bigl\{L(\upsilon, \upsilon^*)\bigr\}\, \mathbb{1}\bigl\{\boldsymbol{\theta} \in \Theta_0(\mathtt{r}_0)\bigr\} d\upsilon} \leq \frac{\int_{\Upsilon \setminus \Upsilon_0(\mathtt{r}_0)} \exp\bigl\{L(\upsilon, \upsilon^*)\bigr\} d\upsilon}{\int_{\Upsilon_0(\mathtt{r}_0)} \exp\bigl\{L(\upsilon, \upsilon^*)\bigr\} d\upsilon} = \rho^*(\mathtt{r}_0),$$

and the assertion follows from Theorem 10.

7.4 Proof of Theorem 12

We use that $\mathbb{L}(\upsilon, \upsilon^*) = \boldsymbol{\xi}^\top \mathcal{D}_0(\upsilon - \upsilon^*) - \|\mathcal{D}_0(\upsilon - \upsilon^*)\|^2/2$ is proportional to the density of a Gaussian distribution. More precisely, define

$$m(\boldsymbol{\xi}) \stackrel{\text{def}}{=} -\|\boldsymbol{\xi}\|^2/2 + \log(\det \mathcal{D}_0) - p \log(\sqrt{2\pi}).$$

Then

$$m(\boldsymbol{\xi}) + \mathbb{L}(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^*) \, = \, -\|\mathcal{D}_0(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^*) - \boldsymbol{\xi}\|^2/2 + \log(\det \mathcal{D}_0) - p \log(\sqrt{2\pi})$$

is (conditionally on $\boldsymbol{Y}$) the log-density of the normal law with the mean $\boldsymbol{\upsilon}_0 = \boldsymbol{\upsilon}^* + \mathcal{D}_0^{-1}\boldsymbol{\xi}$ and the covariance matrix $\mathcal{D}_0^{-2}$. If we perform integration and leave only $\boldsymbol{\theta}$ part of $\boldsymbol{\upsilon}$ then $m(\boldsymbol{\xi}) + \mathbb{L}(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^*)$ is (conditionally on $\boldsymbol{Y}$) the log-density of the normal law with the mean $\boldsymbol{\theta}^\circ = \breve{D}_0^{-1}\breve{\boldsymbol{\xi}} + \boldsymbol{\theta}^*$ and the covariance matrix $\breve{D}_0^{-2}$. So, for any nonnegative function $f : \mathbb{R}^q \to \mathbb{R}_+$ we get

$$\begin{aligned}
&\int_{\Upsilon} \exp\bigl\{L(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^*) + m(\boldsymbol{\xi})\bigr\} \, f\bigl(\breve{D}_0(\boldsymbol{\theta} - \boldsymbol{\theta}^\circ)\bigr) \, d\boldsymbol{\upsilon} \\
&\quad = \int_{\Upsilon_0(\mathtt{r}_0)} \exp\bigl\{L(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^*) + m(\boldsymbol{\xi})\bigr\} \, f\bigl(\breve{D}_0(\boldsymbol{\theta} - \boldsymbol{\theta}^\circ)\bigr) \, d\boldsymbol{\upsilon} \\
&\qquad + \int_{\Upsilon \setminus \Upsilon_0(\mathtt{r}_0)} \exp\bigl\{L(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^*) + m(\boldsymbol{\xi})\bigr\} \, f\bigl(\breve{D}_0(\boldsymbol{\theta} - \boldsymbol{\theta}^\circ)\bigr) \, d\boldsymbol{\upsilon} \\
&\quad = \bigl(1 + \rho_f(\mathtt{r}_0)\bigr) \int_{\Upsilon_0(\mathtt{r}_0)} \exp\bigl\{L(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^*) + m(\boldsymbol{\xi})\bigr\} \, f\bigl(\breve{D}_0(\boldsymbol{\theta} - \boldsymbol{\theta}^\circ)\bigr) \, d\boldsymbol{\upsilon} \\
&\quad \le \mathrm{e}^{\Delta(\mathtt{r}_0, \mathtt{x}) + \rho_f(\mathtt{r}_0)} \int_{\Upsilon_0(\mathtt{r}_0)} \exp\bigl\{\mathbb{L}(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^*) + m(\boldsymbol{\xi})\bigr\} \, f\bigl(\breve{D}_0(\boldsymbol{\theta} - \boldsymbol{\theta}^\circ)\bigr) \, d\boldsymbol{\upsilon} \\
&\quad \le \mathrm{e}^{\Delta(\mathtt{r}_0, \mathtt{x}) + \rho_f(\mathtt{r}_0)} \int_{\mathbb{R}^p} \exp\bigl\{\mathbb{L}(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^*) + m(\boldsymbol{\xi})\bigr\} \, f\bigl(\breve{D}_0(\boldsymbol{\theta} - \boldsymbol{\theta}^\circ)\bigr) \, d\boldsymbol{\upsilon} \\
&\quad = \mathrm{e}^{\Delta(\mathtt{r}_0, \mathtt{x}) + \rho_f(\mathtt{r}_0)} \, I\!\!E f(\boldsymbol{\gamma}).
\end{aligned}$$

Thus,

$$\int_{\Upsilon} \exp\bigl\{L(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^*)\bigr\} \, f\bigl(\breve{D}_0(\boldsymbol{\theta} - \boldsymbol{\theta}^\circ)\bigr) \, d\boldsymbol{\upsilon} \, \le \, \exp\{\Delta(\mathtt{r}_0, \mathtt{x}) - m(\boldsymbol{\xi}) + \rho_f(\mathtt{r}_0)\} \, I\!\!E f(\boldsymbol{\gamma}). \tag{48}$$

Now (48) and (47) imply

$$\frac{\int_{\Upsilon} \exp\bigl\{L(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^*)\bigr\} f\bigl(\breve{D}_{\boldsymbol{\epsilon}}(\boldsymbol{\theta} - \boldsymbol{\theta}_{\boldsymbol{\epsilon}})\bigr) \, d\boldsymbol{\upsilon}}{\int_{\Upsilon} \exp\bigl\{L(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^*)\bigr\} \, d\boldsymbol{\upsilon}} \, \le \, \exp\bigl\{2\Delta(\mathtt{r}_0, \mathtt{x}) + \nu(\mathtt{r}_0) + \rho_f(\mathtt{r}_0)\bigr\} \, I\!\!E f(\boldsymbol{\gamma})$$

and (38) follows by definition of $\Delta^+(\mathtt{r}_0, \mathtt{x})$.

7.5 Proof of Corollary 3

As a direct implication of (38) one easily gets

$$I\!\!E^\circ \bigl|\boldsymbol{\lambda}^\top \breve{D}_0(\boldsymbol{\vartheta} - \boldsymbol{\theta}^\circ)\bigr|^2 \, \le \, \exp(\Delta^+(\mathtt{r}_0, \mathtt{x}))\|\boldsymbol{\lambda}\|^2.$$

The only important step is to show that $\rho_{x^2}(\mathtt{r}_0)$ is small. Denote

$$\boldsymbol{\lambda}_0 = \mathcal{D}_0^{-1} \begin{pmatrix} \breve{D}_0 \boldsymbol{\lambda} \\ \mathbf{0} \end{pmatrix}$$

and $\mathbf{0}$ is a zero vector of dimension $(p-q)$. We proceed separately with the numerator and denominator. For the numerator by (44) and (45) on $\Omega(\mathtt{x})$

$$\begin{aligned}
&\int_{\Upsilon \setminus \Upsilon_0(\mathtt{r}_0)} \bigl| \boldsymbol{\lambda}^\top \breve{D}_0 (\boldsymbol{\theta} - \boldsymbol{\theta}^\circ) \bigr|^2 \, \exp\bigl\{ L(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^*) \bigr\} \, d\boldsymbol{\upsilon} \\
&\leq \int_{\Upsilon \setminus \Upsilon_0(\mathtt{r}_0)} \bigl| \boldsymbol{\lambda}^\top \breve{D}_0 (\boldsymbol{\theta} - \boldsymbol{\theta}^\circ) \bigr|^2 \, \exp\bigl\{ -\mathtt{b} \| \mathcal{D}_0 (\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^*) \|^2 / 2 \bigr\} \, d\boldsymbol{\upsilon} \\
&= \int_{\Upsilon \setminus \Upsilon_0(\mathtt{r}_0)} \bigl| \boldsymbol{\lambda}_0^\top \mathcal{D}_0 (\boldsymbol{\upsilon} - \boldsymbol{\upsilon}_0) \bigr|^2 \, \exp\bigl\{ -\mathtt{b} \| \mathcal{D}_0 (\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^*) \|^2 / 2 \bigr\} d\boldsymbol{\upsilon} \\
&= \exp\bigl\{ (-(p/2+2) \log \mathtt{b} - \log(\det \mathcal{D}_0) + p \log(\sqrt{2\pi}) \bigr\} I\!\!E \bigl| \boldsymbol{\lambda}_0^\top (\boldsymbol{\gamma} + \boldsymbol{\xi}) \bigr|^2 \, \mathbb{1}(\|\boldsymbol{\gamma}\|^2 \geq \mathtt{b} \mathtt{r}_0^2) \\
&\leq (4 + 2\|\boldsymbol{\xi}\|^2) \exp\Bigl\{ -(p/2+2) \log \mathtt{b} - \log(\det \mathcal{D}_0) + p \log(\sqrt{2\pi}) - \mathtt{b} \mathtt{r}_0^2 / 4 + p/2 \Bigr\} \bigl\| \boldsymbol{\lambda}_0 \bigr\|^2 \\
&\leq \exp\Bigl\{ \|\boldsymbol{\xi}\|^2/2 + (p/2+2) \log(\mathrm{e}/\mathtt{b}) - \log(\det \mathcal{D}_0) + p \log(\sqrt{2\pi}) - \mathtt{b} \mathtt{r}_0^2 / 4 \Bigr\} \bigl\| \boldsymbol{\lambda}_0 \bigr\|^2 \\
&\leq \exp\bigl\{ -\log(\det \mathcal{D}_0) + p \log(\sqrt{2\pi}) - \mathtt{x} \bigr\} \bigl\| \boldsymbol{\lambda}_0 \bigr\|^2
\end{aligned}$$

for $\mathtt{b} \mathtt{r}_0^2 \geq (2p+4) \log(\mathrm{e}/\mathtt{b}) + 2 z_B(\mathtt{x}) + 4\mathtt{x}$. For the denominator, it holds on $\Omega(\mathtt{x})$

$$\begin{aligned}
&\int_{\Upsilon_0(\mathtt{r}_0)} \bigl| \boldsymbol{\lambda}^\top \breve{D}_0 (\boldsymbol{\theta} - \boldsymbol{\theta}^\circ) \bigr|^2 \, \exp\bigl\{ L(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^*) \bigr\} d\boldsymbol{\upsilon} \\
&\quad \geq \mathrm{e}^{-\Delta(\mathtt{r}_0, \mathtt{x})} \int_{\Upsilon_0(\mathtt{r}_0)} \bigl| \boldsymbol{\lambda}^\top \breve{D}_0 (\boldsymbol{\theta} - \boldsymbol{\theta}^\circ) \bigr|^2 \, \exp\bigl\{ \mathbb{L}(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^*) \bigr\} d\boldsymbol{\upsilon} \\
&\quad = \mathrm{e}^{-\Delta(\mathtt{r}_0, \mathtt{x})} \int_{\Upsilon_0(\mathtt{r}_0)} \bigl| \boldsymbol{\lambda}_0^\top \mathcal{D}_0 (\boldsymbol{\upsilon} - \boldsymbol{\upsilon}_0) \bigr|^2 \, \exp\bigl\{ \mathbb{L}(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^*) \bigr\} d\boldsymbol{\upsilon} \\
&\quad = \mathrm{e}^{-\Delta(\mathtt{r}_0, \mathtt{x}) - m(\boldsymbol{\xi})} I\!\!E \bigl| \boldsymbol{\lambda}_0^\top (\boldsymbol{\gamma} + \boldsymbol{\xi}) \bigr|^2 \, \mathbb{1}(\|\boldsymbol{\gamma} + \boldsymbol{\xi}\|^2 \leq \mathtt{r}_0^2) \\
&\quad = \mathrm{e}^{-\Delta(\mathtt{r}_0, \mathtt{x}) - m(\boldsymbol{\xi})} \Bigl\{ I\!\!E \bigl| \boldsymbol{\lambda}_0^\top (\boldsymbol{\gamma} + \boldsymbol{\xi}) \bigr|^2 - I\!\!E \bigl| \boldsymbol{\lambda}_0^\top (\boldsymbol{\gamma} + \boldsymbol{\xi}) \bigr|^2 \, \mathbb{1}(\|\boldsymbol{\gamma} + \boldsymbol{\xi}\|^2 \geq \mathtt{r}_0^2) \Bigr\} \\
&\quad \geq \mathrm{e}^{-\Delta(\mathtt{r}_0, \mathtt{x}) - m(\boldsymbol{\xi})} \Bigl\{ \bigl\| \boldsymbol{\lambda}_0 \bigr\|^2 + \bigl| \boldsymbol{\lambda}_0^\top \boldsymbol{\xi} \bigr|^2 - 2 \bigl\| \boldsymbol{\lambda}_0 \bigr\|^2 \exp\bigl\{ -\mathtt{r}_0^2/4 + p/2 + \|\boldsymbol{\xi}\|^2/2 \bigr\} \Bigr\} \\
&\quad \geq \mathrm{e}^{-\Delta(\mathtt{r}_0, \mathtt{x}) - m(\boldsymbol{\xi})} \bigl\| \boldsymbol{\lambda}_0 \bigr\|^2 \{ 1 - 2\mathrm{e}^{-\mathtt{x}} \} \geq \exp\bigl\{ -\Delta(\mathtt{r}_0, \mathtt{x}) - m(\boldsymbol{\xi}) - 4\mathrm{e}^{-\mathtt{x}} \bigr\} \bigl\| \boldsymbol{\lambda}_0 \bigr\|^2
\end{aligned}$$

for $\mathtt{r}_0^2 \geq 2p + 2z_B(\mathtt{x}) + 4\mathtt{x}$ on $\Omega(\mathtt{x})$. This yields on $\Omega(\mathtt{x})$

$$\begin{aligned}
\rho_{x^2}(\mathtt{r}_0) &= \frac{\int_{\Upsilon \setminus \Upsilon_0(\mathtt{r}_0)} \bigl|\boldsymbol{\lambda}^\top \breve{D}_0(\boldsymbol{\theta} - \boldsymbol{\theta}^\circ)\bigr|^2 \exp\bigl\{L(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^*)\bigr\}\, d\boldsymbol{\upsilon}}{\int_{\Upsilon_0(\mathtt{r}_0)} \bigl|\boldsymbol{\lambda}^\top \breve{D}_0(\boldsymbol{\theta} - \boldsymbol{\theta}^\circ)\bigr|^2 \exp\bigl\{L(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^*)\bigr\} d\boldsymbol{\upsilon}} \\
&\leq \frac{2\exp\bigl\{-\log(\det \mathcal{D}_0) + p\log(\sqrt{2\pi})\bigr\} - \mathtt{x}\bigr\} \|\boldsymbol{\lambda}_0\|^2}{\exp\bigl\{-\Delta(\mathtt{r}_0, \mathtt{x}) + m(\boldsymbol{\xi}) - 4\mathrm{e}^{-\mathtt{x}}\bigr\} \|\boldsymbol{\lambda}_0\|^2} \\
&= 2\exp\bigl\{\Delta(\mathtt{r}_0, \mathtt{x}) - \|\boldsymbol{\xi}\|^2/2 \ + 4\mathrm{e}^{-\mathtt{x}} - \mathtt{x}\bigr\} \leq 2\exp\bigl\{\Delta(\mathtt{r}_0, \mathtt{x}) + 4\mathrm{e}^{-\mathtt{x}} - \mathtt{x}\bigr\}.
\end{aligned}$$

7.6 Proof of Corollary 4

The first statement follows from Theorem 12 with $f(\boldsymbol{u}) = \mathbb{1}\bigl(D_1 \breve{D}_0^{-1} \boldsymbol{u} + \boldsymbol{\delta}_0 \in A\bigr)$. Further, it holds on $\Omega(\mathtt{x})$ for $\boldsymbol{\delta}_0 \stackrel{\text{def}}{=} D_1(\boldsymbol{\theta}^\circ - \widehat{\boldsymbol{\theta}})$

$$\|\boldsymbol{\delta}_0\|^2 = \|D_1(\boldsymbol{\theta}^\circ - \widehat{\boldsymbol{\theta}})\|^2 \leq (1+\alpha)\|\breve{D}_0(\boldsymbol{\theta}^\circ - \widehat{\boldsymbol{\theta}})\|^2 \leq (1+\alpha)\beta^2.$$

For proving (40), we apply Pinsker's inequality to two normal distributions. Let $\boldsymbol{\gamma}$ be standard normal in $\mathbb{R}^q$. The random variable $D_1 \breve{D}_0^{-1}\boldsymbol{\gamma} + \boldsymbol{\delta}_0$ is normal with mean $\mathcal{N}(\boldsymbol{\delta}_0, B_1^{-1})$ with $B_1^{-1} \stackrel{\text{def}}{=} D_1 \breve{D}_0^{-2} D_1$. Obviously $\|I_q - B_1\| = \|I_q - D_1^{-1}\breve{D}_0^2 D_1^{-1}\| \leq \alpha$. Thus, by Lemma 8 for any measurable set A, it holds

$$\mathbb{P}\bigl(D_1 \breve{D}_0^{-1}\boldsymbol{\gamma} + \boldsymbol{\delta}_0 \in A \,\big|\, \boldsymbol{Y}\bigr) \leq \mathbb{P}\bigl(\boldsymbol{\gamma} \in A\bigr) + \frac{1}{2}\sqrt{\alpha^2 q + (1+\alpha)^2\beta^2}.$$

7.7 Proof of Theorem 13

As in the proof of Theorem 12, for any nonnegative function $f : \mathbb{R}^q \to \mathbb{R}_+$, it holds

$$\begin{aligned}
&\int_{\Upsilon} \exp\bigl\{L(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^*)\bigr\}\, f\bigl(\breve{D}_0(\boldsymbol{\theta} - \boldsymbol{\theta}^\circ)\bigr)\, \mathbb{1}\bigl\{\boldsymbol{\theta} \in \Theta_0(\mathtt{r}_0)\bigr\}\, d\boldsymbol{\upsilon} \\
&\quad \geq \int_{\Upsilon_0(\mathtt{r}_0)} \exp\bigl\{L(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^*)\bigr\}\, f\bigl(\breve{D}_0(\boldsymbol{\theta} - \boldsymbol{\theta}^\circ)\bigr)\, d\boldsymbol{\upsilon} \\
&\quad \geq \exp\{-\Delta(\mathtt{r}_0, \mathtt{x}) - m(\boldsymbol{\xi})\} \int_{\Upsilon_0(\mathtt{r}_0)} \exp\bigl\{\mathbb{L}(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^*)\bigr\}\, f\bigl(\breve{D}_0(\boldsymbol{\theta} - \boldsymbol{\theta}^\circ)\bigr)\, d\boldsymbol{\upsilon} \\
&\quad \geq \exp\{-\Delta(\mathtt{r}_0, \mathtt{x}) - m(\boldsymbol{\xi})\} \int_{\mathbb{R}^p} \exp\bigl\{\mathbb{L}(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^*)\bigr\}\, f\bigl(\breve{D}_0(\boldsymbol{\theta} - \boldsymbol{\theta}^\circ)\bigr)\, d\boldsymbol{\upsilon} \\
&\qquad - \exp\{-\Delta(\mathtt{r}_0, \mathtt{x}) - m(\boldsymbol{\xi})\} \int_{\mathbb{R}^p \setminus \Upsilon_0(\mathtt{r}_0)} \exp\bigl\{\mathbb{L}(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^*)\bigr\}\, f\bigl(\breve{D}_0(\boldsymbol{\theta} - \boldsymbol{\theta}^\circ)\bigr)\, d\boldsymbol{\upsilon} \\
&\quad = \exp\{-\Delta(\mathtt{r}_0, \mathtt{x}) - m(\boldsymbol{\xi})\}(1 - \widetilde{\rho}_f(\mathtt{r}_0)) \int_{\mathbb{R}^p} \exp\bigl\{\mathbb{L}(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^*)\bigr\}\, f\bigl(\breve{D}_0(\boldsymbol{\theta} - \boldsymbol{\theta}^\circ)\bigr)\, d\boldsymbol{\upsilon}.
\end{aligned}$$

The definition (42) implies

$$\begin{aligned}
&\int_{\Upsilon} \exp\{L(\upsilon,\upsilon^*)\}\, f\bigl(\breve{D}_0(\theta-\theta^\circ)\bigr)\, \mathbb{1}\{\theta\in\Theta_0(\mathtt{r}_0)\}\, d\upsilon \\
&\quad \geq \exp\{-\Delta(\mathtt{r}_0,\mathtt{x})-m(\xi)\}(1-\widetilde{\rho}_f(\mathtt{r}_0))\int_{\Theta_0(\mathtt{r}_0)\times\mathbb{R}^{(p-q)}} \exp\{\mathbb{L}(\upsilon,\upsilon^*)\} f\bigl(\breve{D}_0(\theta-\theta^\circ)\bigr) d\upsilon \\
&\quad \geq \exp\{-\Delta(\mathtt{r}_0,\mathtt{x})-m(\xi)-2\widetilde{\rho}_f(\mathtt{r}_0)\}\, I\!\!E f(\gamma)\, \mathbb{1}\{\|\gamma+\breve{\xi}\|\leq \mathtt{r}_0\}. \qquad (49)
\end{aligned}$$

Here we used that $1-\alpha\geq \mathrm{e}^{-2\alpha}$ for $0\leq\alpha\leq\frac{1}{2}$. Similarly,

$$\begin{aligned}
\int_{\Upsilon} \exp\{L(\upsilon,\upsilon^*)\}\, d\upsilon &= \int_{\Upsilon_0(\mathtt{r}_0)} \exp\{L(\upsilon,\upsilon^*)\}\, d\upsilon + \int_{\Upsilon\setminus\Upsilon_0(\mathtt{r}_0)} \exp\{L(\upsilon,\upsilon^*)\}\, d\upsilon \\
&= \{1+\rho^*(\mathtt{r}_0)\}\int_{\Upsilon_0(\mathtt{r}_0)} \exp\{L(\upsilon,\upsilon^*)\}\, d\upsilon \\
&\leq \{1+\rho^*(\mathtt{r}_0)\}\exp\{\Delta(\mathtt{r}_0,\mathtt{x})-m(\xi)\}\, I\!\!P\bigl(\|\gamma+\xi\|\leq \mathtt{r}_0 \,\big|\, Y\bigr),
\end{aligned}$$

and finally

$$\int_{\Upsilon} \exp\{L(\upsilon,\upsilon^*)\}\, d\upsilon \leq \exp\{\Delta(\mathtt{r}_0,\mathtt{x})-m(\xi)+\nu(\mathtt{r}_0)+\rho^*(\mathtt{r}_0)\}. \qquad (50)$$

The bounds (49) and (50) imply

$$\begin{aligned}
&\frac{\int_{\Upsilon}\exp\{L(\upsilon,\upsilon^*)\} f\bigl(\breve{D}_0(\theta-\theta^\circ)\bigr)\, d\upsilon}{\int_{\Upsilon}\exp\{L(\upsilon,\upsilon^*)\}\, d\upsilon} \\
&\quad \geq \frac{\exp\{-\Delta(\mathtt{r}_0,\mathtt{x})-m(\xi)-2\widetilde{\rho}_f(\mathtt{r}_0)\}\, I\!\!E f(\gamma)\, \mathbb{1}\{\|\gamma+\breve{\xi}\|\leq \mathtt{r}_0\}}{\exp\{\Delta(\mathtt{r}_0,\mathtt{x})-m(\xi)+\nu(\mathtt{r}_0)+\rho^*(\mathtt{r}_0)\}} \\
&\quad \geq \exp\{-2\Delta(\mathtt{r}_0,\mathtt{x})-2\widetilde{\rho}_f(\mathtt{r}_0)-\nu(\mathtt{r}_0)-\rho^*(\mathtt{r}_0)\} I\!\!E f(\gamma)\, \mathbb{1}\{\|\gamma+\breve{\xi}\|\leq \mathtt{r}_0\}.
\end{aligned}$$

This yields (41).

7.8 Proof of Theorem 1

Due to our previous results, it is convenient to decompose the r.v. ϑ in the form

$$\vartheta = \vartheta\, \mathbb{1}\{\vartheta\in\Theta_0(\mathtt{r}_0)\} + \vartheta\, \mathbb{1}\{\vartheta\notin\Theta_0(\mathtt{r}_0)\} = \vartheta^\circ + \vartheta^c.$$

The large deviation result yields that the posterior distribution of the part ϑ^c is negligible provided a proper choice of $\mathtt{r}_0$. Below we show that ϑ° is nearly normal which yields the BvM result. Define

$$\vartheta^\circ \stackrel{\text{def}}{=} I\!\!E^\circ \vartheta, \qquad \mathfrak{S}_\circ^2 \stackrel{\text{def}}{=} \mathrm{Cov}(\vartheta^\circ) \stackrel{\text{def}}{=} I\!\!E^\circ\bigl\{(\vartheta-\vartheta^\circ)(\vartheta-\vartheta^\circ)^\top\bigr\}.$$

It suffices to show that holds on $\Omega(\mathtt{x})$

$$\|\breve{D}_0(\boldsymbol{\vartheta}^\circ - \boldsymbol{\theta}^\circ)\|^2 \leq 2\Delta^*$$
$$\bigl\|I_q - \breve{D}_0 \mathfrak{S}_\circ^2 \breve{D}_0\bigr\| \leq 2\Delta^*,$$

where $\Delta^* = \max\bigl\{\Delta^+, \Delta^-\bigr\}$.

Consider $\boldsymbol{\eta} \stackrel{\text{def}}{=} \breve{D}_0(\boldsymbol{\vartheta} - \boldsymbol{\theta}^\circ)$. Corollaries 3 and 5 yield for any $\boldsymbol{\lambda} \in I\!R^q$ that

$$\|\boldsymbol{\lambda}\|^2 \exp(-\Delta^-) \leq I\!E^\circ \bigl|\boldsymbol{\lambda}^\top \boldsymbol{\eta}\bigr|^2 \leq \|\boldsymbol{\lambda}\|^2 \exp(\Delta^+). \tag{51}$$

Define the first two moments of $\boldsymbol{\eta}$:

$$\overline{\boldsymbol{\eta}} \stackrel{\text{def}}{=} I\!E^\circ \boldsymbol{\eta}, \qquad S_\circ^2 \stackrel{\text{def}}{=} I\!E^\circ \bigl\{(\boldsymbol{\eta} - \overline{\boldsymbol{\eta}})(\boldsymbol{\eta} - \overline{\boldsymbol{\eta}})^\top\bigr\} = \breve{D}_0 \mathfrak{S}_\circ^2 \breve{D}_0.$$

Use the following technical statement.

Lemma 9. *Assume* (51). *Then with* $\Delta^* = \max\bigl\{\Delta^+, \Delta^-\bigr\} \leq 1/2$

$$\|\overline{\boldsymbol{\eta}}\|^2 \leq 2\Delta^*, \qquad \|S_\circ^2 - I_q\| \leq 2\Delta^*. \tag{52}$$

Proof. Let $\boldsymbol{u}$ be any unit vector in $I\!R^q$. We obtain from (51)

$$\exp(-\Delta^-) \leq I\!E^\circ \bigl|\boldsymbol{u}^\top \boldsymbol{\eta}\bigr|^2 \leq \exp(\Delta^+).$$

Note now that

$$I\!E^\circ \bigl|\boldsymbol{u}^\top \boldsymbol{\eta}\bigr|^2 = \boldsymbol{u}^\top S_\circ^2 \boldsymbol{u} + |\boldsymbol{u}^\top \overline{\boldsymbol{\eta}}|^2.$$

Hence

$$\exp(-\Delta^-) \leq \boldsymbol{u}^\top S_\circ^2 \boldsymbol{u} + |\boldsymbol{u}^\top \overline{\boldsymbol{\eta}}|^2 \leq \exp(\Delta^+). \tag{53}$$

In a similar way with $\boldsymbol{u} = \overline{\boldsymbol{\eta}}/\|\overline{\boldsymbol{\eta}}\|$ and $\boldsymbol{\gamma} \sim \mathcal{N}(0, I_q)$

$$I\!E^\circ \bigl|\boldsymbol{u}^\top (\boldsymbol{\eta} - \overline{\boldsymbol{\eta}})\bigr|^2 \geq \mathrm{e}^{-\Delta^-} I\!E \bigl|\boldsymbol{u}^\top (\boldsymbol{\gamma} - \overline{\boldsymbol{\eta}})\bigr|^2 = \mathrm{e}^{-\Delta^-} \bigl(1 + \|\overline{\boldsymbol{\eta}}\|^2\bigr)$$

yielding

$$\boldsymbol{u}^\top S_\circ^2 \boldsymbol{u} \geq \bigl(1 + \|\overline{\boldsymbol{\eta}}\|^2\bigr) \exp(-\Delta^-).$$

This inequality contradicts (53) if $\|\overline{\boldsymbol{\eta}}\|^2 > 2\Delta^* > 1$, and (52) follows. □

The bound for the first moment implies with $\boldsymbol{\vartheta}^\circ = I\!\!E^\circ \boldsymbol{\vartheta}$

$$\bigl\| \breve{D}_0 (\boldsymbol{\vartheta}^\circ - \boldsymbol{\theta}^\circ) \bigr\|^2 \le 2\Delta^*$$

while the second bound yields with

$$\bigl\| \breve{D}_0 \mathfrak{S}_\circ^2 \breve{D}_0 - I_q \bigr\| \le 2\Delta^*.$$

The last result follows from (39) and (43) with an additional assumption that $\mathtt{x}$ is large enough to ensure $\Delta^+(\mathtt{r}_0, \mathtt{x}) \le 2\Delta(\mathtt{r}_0, \mathtt{x}) + 5\mathrm{e}^{-\mathtt{x}}$ and $\Delta^-(\mathtt{r}_0, \mathtt{x}) \ge 2\Delta(\mathtt{r}_0, \mathtt{x}) - 8\mathrm{e}^{-\mathtt{x}}$.

7.9 Proof of Theorem 2

Define $L_G(\boldsymbol{\upsilon}) = L(\boldsymbol{\upsilon}) - \|G\boldsymbol{\upsilon}\|^2/2$. The stochastic component of $L_G(\boldsymbol{\upsilon})$ coincides with one of $L(\boldsymbol{\upsilon})$. Also the quadratic term $\|G\boldsymbol{\upsilon}\|^2/2$ does not deteriorate the smoothness properties of the expected process $I\!\!E L_G(\boldsymbol{\upsilon})$. In particular, one can locally approximate $I\!\!E L_G(\boldsymbol{\upsilon}_G^\circ, \boldsymbol{\upsilon})$ by a quadratic function $\bigl\| \mathcal{D}_G(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}_G^\circ) \bigr\|^2/2$ with

$$\boldsymbol{\upsilon}_G^\circ \stackrel{\text{def}}{=} \boldsymbol{\upsilon}_G^* + \mathcal{D}_G^{-1} \boldsymbol{\xi}_G, \quad \boldsymbol{\xi}_G \stackrel{\text{def}}{=} \mathcal{D}_G^{-1} \nabla I\!\!E L_G(\boldsymbol{\upsilon}^*) = \mathcal{D}_G^{-1} \nabla I\!\!E L(\boldsymbol{\upsilon}^*) + \mathcal{D}_G^{-1} G^2 \boldsymbol{\upsilon}^*,$$

$$\boldsymbol{\upsilon}_G^* \stackrel{\text{def}}{=} \operatorname*{argmax}_{\boldsymbol{\upsilon}} I\!\!E L_G(\boldsymbol{\upsilon}), \quad \mathcal{D}_G^2 \stackrel{\text{def}}{=} -\nabla^2 I\!\!E L(\boldsymbol{\upsilon}^*) + G^2 = \mathcal{D}_0^2 + G^2.$$

Now one can easily see that all the conditions of Theorem 1 are fulfilled for the process $L_G(\boldsymbol{\upsilon})$ when $\boldsymbol{\upsilon}^\circ$ is replaced by $\boldsymbol{\upsilon}_G^\circ$ and $\mathcal{D}_0$ by $\mathcal{D}_G$. The result approximates the posterior $\boldsymbol{\upsilon} \,\big|\, \boldsymbol{Y}$ for the Gaussian prior Π by the normal law $\mathcal{N}(\boldsymbol{\upsilon}_G^\circ, \mathcal{D}_G^{-2})$. Now the final result follows by Lemma 8 if we can bound $\|\mathcal{D}_0^{-1} \mathcal{D}_G^2 \mathcal{D}_0^{-1} - I_p\|$ and $\|\mathcal{D}_G(\boldsymbol{\upsilon}^\circ - \boldsymbol{\upsilon}_G^\circ)\|$. By definition

$$\mathcal{D}_0^{-1} \mathcal{D}_G^2 \mathcal{D}_0^{-1} - I_p = \mathcal{D}_0^{-1} G^2 \mathcal{D}_0^{-1}.$$

Further, the definition ensures $\nabla I\!\!E L_G(\boldsymbol{\upsilon}) = \nabla I\!\!E L(\boldsymbol{\upsilon}) - G^2 \boldsymbol{\upsilon}$ for any $\boldsymbol{\upsilon}$ and also $\nabla I\!\!E L(\boldsymbol{\upsilon}^*) = \nabla I\!\!E L_G(\boldsymbol{\upsilon}_G^*) = 0$. This implies

$$\nabla I\!\!E L_G(\boldsymbol{\upsilon}_G^*) - \nabla I\!\!E L_G(\boldsymbol{\upsilon}^*) = \mathcal{D}_G^2(\breve{\boldsymbol{\upsilon}})(\boldsymbol{\upsilon}_G^* - \boldsymbol{\upsilon}^*),$$

where $\breve{\boldsymbol{\upsilon}}$ is a point from the interval connecting $\boldsymbol{\upsilon}^*$ and $\boldsymbol{\upsilon}_G^*$ and $\mathcal{D}_G^2(\breve{\boldsymbol{\upsilon}}) = -\nabla^2 I\!\!E L(\breve{\boldsymbol{\upsilon}}) + G^2$. Therefore,

$$\mathcal{D}_G^2(\breve{\boldsymbol{\upsilon}})(\boldsymbol{\upsilon}_G^* - \boldsymbol{\upsilon}^*) = G^2 \boldsymbol{\upsilon}^*. \tag{54}$$

Let $\mathtt{r}_G = \|\mathcal{D}_G(\boldsymbol{\upsilon}^\circ - \boldsymbol{\upsilon}_G^\circ)\|$. It holds by $(\mathcal{L}_0)$ and (54) with probability $\ge 1 - 2\mathrm{e}^{-\mathtt{x}}$

$$\begin{aligned}
\mathtt{r}_G &= \|\mathcal{D}_G(\boldsymbol{\upsilon}^\circ - \boldsymbol{\upsilon}_G^\circ)\| \le \|\mathcal{D}_G(\boldsymbol{\upsilon}^* - \boldsymbol{\upsilon}_G^*)\| + \|\boldsymbol{\xi}_G - \mathcal{D}_G \mathcal{D}_0^{-1} \boldsymbol{\xi}\| \\
&\le \|\mathcal{D}_G \mathcal{D}_G^{-2}(\breve{\boldsymbol{\upsilon}}) \mathcal{D}_G \, \mathcal{D}_G^{-1} G^2 \boldsymbol{\upsilon}^*\| + \|\mathcal{D}_G^{-1} \mathcal{D}_0 \boldsymbol{\xi} + \mathcal{D}_G^{-1} G^2 \boldsymbol{\upsilon}^* - \mathcal{D}_G \mathcal{D}_0^{-1} \boldsymbol{\xi}\| \\
&\le \bigl\{1 - \delta(\mathtt{r}_G)\bigr\}^{-1} \beta + \beta + \|\mathcal{D}_G^{-1} \bigl(\mathcal{D}_0^2 - \mathcal{D}_G^2\bigr) \mathcal{D}_0^{-1}\| \cdot \|\boldsymbol{\xi}\| \\
&\le \bigl\{1 - \delta(\mathtt{r}_G)\bigr\}^{-1} \beta + \beta + \boldsymbol{\epsilon} z_B(\mathtt{x}),
\end{aligned}$$

where $z_B(\mathtt{x})$ is defined by (33). In particular, $\delta(\mathtt{r}_G) \le 1/2$ and $\boldsymbol{\epsilon} \le 1/2$ imply $\mathtt{r}_G \le 3\beta + \boldsymbol{\epsilon} z_B(\mathtt{x})$. This yields by Lemma 8 for any $A^* \subset \mathbb{R}^p$:

$$\bigl|I\!\!P_0(A^*) - I\!\!P_G(A^*)\bigr| \le \frac{1}{2}\sqrt{\delta^2 + (1+\boldsymbol{\epsilon})(3\beta + \boldsymbol{\epsilon} z_B(\mathtt{x}))}, \tag{55}$$

where measure $I\!\!P_0$ stands for Gaussian measure $\mathcal{N}(\boldsymbol{\upsilon}^\circ, \mathcal{D}_0^{-2})$ and measure $I\!\!P_G$ stands for Gaussian measure $\mathcal{N}(\boldsymbol{\upsilon}_G^\circ, \mathcal{D}_G^{-2})$. Now we can take $A^* = A \times \mathbb{R}^{(p-q)}$. This yields by (55):

$$\bigl|I\!\!P_0(A^*) - I\!\!P_G(A^*)\bigr| = \bigl|\breve{I\!\!P}_0(A) - \breve{I\!\!P}_G(A)\bigr| \le \frac{1}{2}\sqrt{\delta^2 + (1+\boldsymbol{\epsilon})(3\beta + \boldsymbol{\epsilon} z_B(\mathtt{x}))},$$

where measure $\breve{I\!\!P}_0$ stands for Gaussian measure $\mathcal{N}(\boldsymbol{\theta}^\circ, \breve{D}_0^{-2})$ and measure $\breve{I\!\!P}_G$ stands for Gaussian measure $\mathcal{N}(\boldsymbol{\theta}_G^\circ, \breve{D}_G^{-2})$. This yields the result of the theorem.

7.10 Proof of Theorem 3

First we compare the deterministic quantities $\breve{D}^2$ and $\boldsymbol{\theta}^*$ with their sieve counterparts. The identifiability condition $(\mathcal{I}^{\mathtt{s}})$ guarantees for $\breve{D}^2_{\mathtt{s}} = D^2_{\boldsymbol{\theta}} - A_{\boldsymbol{\theta}\boldsymbol{\eta}} A^\top_{\boldsymbol{\theta}\boldsymbol{\eta}}$ that

$$\|D_{\boldsymbol{\theta}}^{-1} \breve{D}^2_{\mathtt{s}} D_{\boldsymbol{\theta}}^{-1}\| \ge 1 - \nu. \tag{56}$$

Further $\breve{D}^2 = D^2_{\boldsymbol{\theta}} - A_{\boldsymbol{\theta}\boldsymbol{\phi}} A^\top_{\boldsymbol{\theta}\boldsymbol{\phi}}$ and $\breve{D}^2_{\mathtt{s}} = D^2_{\boldsymbol{\theta}} - A_{\boldsymbol{\theta}\boldsymbol{\eta}} A^\top_{\boldsymbol{\theta}\boldsymbol{\eta}}$

$$\breve{D}^2_{\mathtt{s}} - \breve{D}^2 = D^2_{\boldsymbol{\theta}} - A_{\boldsymbol{\theta}\boldsymbol{\eta}} A^\top_{\boldsymbol{\theta}\boldsymbol{\eta}} - (D^2_{\boldsymbol{\theta}} - A_{\boldsymbol{\theta}\boldsymbol{\phi}} A^\top_{\boldsymbol{\theta}\boldsymbol{\phi}}) = A_{\boldsymbol{\theta}\boldsymbol{\varkappa}} A^\top_{\boldsymbol{\theta}\boldsymbol{\varkappa}}.$$

The use of the smoothness condition (B) implies

$$\begin{aligned} \bigl\|D_{\boldsymbol{\theta}}^{-1}(\breve{D}^2_{\mathtt{s}} - \breve{D}^2) D_{\boldsymbol{\theta}}^{-1}\bigr\| &= \bigl\|D_{\boldsymbol{\theta}}^{-1} A_{\boldsymbol{\theta}\boldsymbol{\varkappa}} A^\top_{\boldsymbol{\theta}\boldsymbol{\varkappa}} D_{\boldsymbol{\theta}}^{-1}\bigr\| \le \rho_{\mathtt{s}}, \\ \bigl\|\breve{D}^{-1}\bigl(\breve{D}^2_{\mathtt{s}} - \breve{D}^2\bigr)\breve{D}^{-1}\bigr\| &\le (1-\nu)^{-1}\bigl\|D_{\boldsymbol{\theta}}^{-1} A_{\boldsymbol{\theta}\boldsymbol{\varkappa}} A^\top_{\boldsymbol{\theta}\boldsymbol{\varkappa}} D_{\boldsymbol{\theta}}^{-1}\bigr\| \le (1-\nu)^{-1}\rho_{\mathtt{s}}. \end{aligned} \tag{57}$$

Next we bound the bias $\boldsymbol{\theta}^* - \boldsymbol{\theta}^*_{\mathtt{s}}$ introduced in $\boldsymbol{\theta}^*$ by truncation. Consider first the Gaussian likelihood with

$$I\!\!E L(\boldsymbol{\upsilon}^*, \boldsymbol{\varkappa}^*) - I\!\!E L(\boldsymbol{\upsilon}, \boldsymbol{\varkappa}) = \|\mathcal{D}_0\bigl\{(\boldsymbol{\upsilon}, \boldsymbol{\varkappa}) - (\boldsymbol{\upsilon}^*, \boldsymbol{\varkappa}^*)\bigr\}\|^2/2.$$

Define $\boldsymbol{\upsilon}^*_{\mathtt{s}} = (\boldsymbol{\theta}^*_{\mathtt{s}}, \boldsymbol{\eta}^*_{\mathtt{s}})$ as the minimizer of the quadratic function $\|\mathcal{D}_0\bigl((\boldsymbol{\upsilon}, 0) - (\boldsymbol{\upsilon}^*, \boldsymbol{\varkappa}^*)\bigr)\|^2$ over the set of parameters $(\boldsymbol{\upsilon}, 0) = (\boldsymbol{\theta}, \boldsymbol{\eta}, 0)$:

$$\begin{aligned} \boldsymbol{\upsilon}^*_{\mathtt{s}} &= \operatorname*{argmin}_{\boldsymbol{\upsilon} = (\boldsymbol{\theta}, \boldsymbol{\eta})} \|\mathcal{D}_0\bigl((\boldsymbol{\upsilon}, 0) - (\boldsymbol{\upsilon}^*, \boldsymbol{\varkappa}^*)\bigr)\|^2 \\ &= \operatorname*{argmin}_{(\boldsymbol{\theta}, \boldsymbol{\eta})} \bigl\{(\boldsymbol{\theta} - \boldsymbol{\theta}^*)^\top D^2_{\boldsymbol{\theta}} (\boldsymbol{\theta} - \boldsymbol{\theta}^*) + \|\boldsymbol{\eta} - \boldsymbol{\eta}^*\|^2 \\ &\qquad + 2(\boldsymbol{\theta} - \boldsymbol{\theta}^*)^\top A_{\boldsymbol{\theta}\boldsymbol{\eta}} (\boldsymbol{\eta} - \boldsymbol{\eta}^*) - 2(\boldsymbol{\theta} - \boldsymbol{\theta}^*)^\top A_{\boldsymbol{\theta}\boldsymbol{\varkappa}} \boldsymbol{\varkappa}^* \bigr\}. \end{aligned}$$

This implies by direct calculus that $\upsilon_{\mathtt{s}}^* = (\boldsymbol{\theta}_{\mathtt{s}}^*, \boldsymbol{\eta}_{\mathtt{s}}^*)$ fulfills

$$\begin{aligned} \boldsymbol{\eta}_{\mathtt{s}}^* - \boldsymbol{\eta}^* &= -A_{\boldsymbol{\theta}\boldsymbol{\eta}}^\top (\boldsymbol{\theta}_{\mathtt{s}}^* - \boldsymbol{\theta}^*), \\ \boldsymbol{\theta}_{\mathtt{s}}^* - \boldsymbol{\theta}^* &= \bigl(D_{\boldsymbol{\theta}}^2 - A_{\boldsymbol{\theta}\boldsymbol{\eta}} A_{\boldsymbol{\theta}\boldsymbol{\eta}}^\top\bigr)^{-1} A_{\boldsymbol{\theta}\varkappa} \varkappa^* = \breve{D}_{\mathtt{s}}^{-2} A_{\boldsymbol{\theta}\varkappa} \varkappa^* . \end{aligned} \tag{58}$$

The identifiability condition (56) and the smoothness conditions (22), (23) imply

$$\begin{aligned} \|\breve{D}_{\mathtt{s}} (\boldsymbol{\theta}_{\mathtt{s}}^* - \boldsymbol{\theta}^*)\| &= \|\breve{D}_{\mathtt{s}}^{-1} A_{\boldsymbol{\theta}\varkappa} \varkappa^*\| \\ &\le (1-\nu)^{-1} \bigl\| D_{\boldsymbol{\theta}}^{-1} A_{\boldsymbol{\theta}\varkappa} \varkappa^* \bigr\| \le (1-\nu)^{-1} b_{\mathtt{s}}. \end{aligned} \tag{59}$$

In the general non-Gaussian case, we use that

$$\bigl\| \mathcal{D}_0^{-1} \bigl\{ \nabla I\!\!E L(\upsilon^*, \varkappa^*) - \nabla I\!\!E L(\upsilon_{\mathtt{s}}^*, 0) \bigr\} - \mathcal{D}_0 \bigl\{ (\upsilon^*, \varkappa^*) - (\upsilon_{\mathtt{s}}^*, 0) \bigr\} \bigr\| \le \delta(\mathtt{r}_{\mathtt{s}}) \mathtt{r}_{\mathtt{s}}$$

with

$$\mathtt{r}_{\mathtt{s}} = \bigl\| \mathcal{D}_0 \bigl\{ (\upsilon^*, \varkappa^*) - (\upsilon_{\mathtt{s}}^*, 0) \bigr\} \bigr\| \le \bigl\| \mathcal{D}_0 \bigl\{ (\upsilon^*, \varkappa^*) - (\upsilon^*, 0) \bigr\} \bigr\| = \|\varkappa^*\| .$$

The definition of $(\boldsymbol{\theta}^*, \boldsymbol{\eta}^*, \varkappa^*)$ and $\upsilon_{\mathtt{s}}^*$ imply that $\nabla_{\upsilon} I\!\!E L(\boldsymbol{\theta}^*, \boldsymbol{\eta}^*, \varkappa^*) = 0$ and $\nabla_{\upsilon} I\!\!E L(\upsilon_{\mathtt{s}}^*, 0) = 0$. Now, projecting on the $\boldsymbol{\theta}$-subspace implies similarly to (58)

$$\bigl\| \breve{D}_{\mathtt{s}} (\boldsymbol{\theta}^* - \boldsymbol{\theta}_{\mathtt{s}}^*) - \breve{D}_{\mathtt{s}}^{-1} A_{\boldsymbol{\theta}\varkappa} \varkappa^* \bigr\| \le \delta(\mathtt{r}_{\mathtt{s}}) \mathtt{r}_{\mathtt{s}} .$$

This implies by (59)

$$\bigl\| \breve{D}_{\mathtt{s}} (\boldsymbol{\theta}^* - \boldsymbol{\theta}_{\mathtt{s}}^*) \bigr\| \le (1-\nu)^{-1} b_{\mathtt{s}} + \delta(\mathtt{r}_{\mathtt{s}}) \mathtt{r}_{\mathtt{s}} ,$$

which completes the proof of (13).

Now we consider the stochastic part $\breve{D}_{\mathtt{s}}^{-1} \breve{\boldsymbol{\xi}}_{\mathtt{s}} - \breve{D}^{-1} \breve{\boldsymbol{\xi}}$ of $\boldsymbol{\theta}^\circ - \boldsymbol{\theta}_{\mathtt{s}}^\circ$. The bounds (57) and (56) imply for $\breve{\boldsymbol{\xi}} = \nabla_{\boldsymbol{\theta}} - A_{\boldsymbol{\theta}\boldsymbol{\eta}} \nabla_{\boldsymbol{\eta}} - A_{\boldsymbol{\theta}\varkappa} \nabla_{\varkappa}$ and $\breve{\boldsymbol{\xi}}_{\mathtt{s}} = \nabla_{\boldsymbol{\theta}} - A_{\boldsymbol{\theta}\boldsymbol{\eta}} \nabla_{\boldsymbol{\eta}}$

$$\begin{aligned} \Bigl\| \breve{D} \bigl\{ \breve{D}_{\mathtt{s}}^{-1} \breve{\boldsymbol{\xi}}_{\mathtt{s}} - \breve{D}^{-1} \breve{\boldsymbol{\xi}} \bigr\} \Bigr\| &= \Bigl\| \breve{D} \breve{D}_{\mathtt{s}}^{-2} (\nabla_{\boldsymbol{\theta}} - A_{\boldsymbol{\theta}\boldsymbol{\eta}} \nabla_{\boldsymbol{\eta}}) - \breve{D}^{-1} (\nabla_{\boldsymbol{\theta}} - A_{\boldsymbol{\theta}\boldsymbol{\eta}} \nabla_{\boldsymbol{\eta}} - A_{\boldsymbol{\theta}\varkappa} \nabla_{\varkappa}) \Bigr\| \\ &= \Bigl\| \bigl\{ \breve{D} \breve{D}_{\mathtt{s}}^{-2} - \breve{D}^{-1} \bigr\} (\nabla_{\boldsymbol{\theta}} - A_{\boldsymbol{\theta}\boldsymbol{\eta}} \nabla_{\boldsymbol{\eta}}) + \breve{D}^{-1} A_{\boldsymbol{\theta}\varkappa} \nabla_{\varkappa} \Bigr\| \\ &\le \Bigl\| \bigl\{ I_q - \breve{D}^{-1} \breve{D}_{\mathtt{s}}^2 \breve{D}^{-1} \bigr\} \breve{D} \breve{D}_{\mathtt{s}}^{-2} (\nabla_{\boldsymbol{\theta}} - A_{\boldsymbol{\theta}\boldsymbol{\eta}} \nabla_{\boldsymbol{\eta}}) \Bigr\| + \bigl\| \breve{D}^{-1} A_{\boldsymbol{\theta}\varkappa} \nabla_{\varkappa} \bigr\| \\ &\le (1-\nu)^{-1} \rho_{\mathtt{s}} \|\breve{\boldsymbol{\xi}}_{\mathtt{s}}\| + (1-\nu)^{-1} \bigl\| D_{\boldsymbol{\theta}}^{-1} A_{\boldsymbol{\theta}\varkappa} \nabla_{\varkappa} \bigr\| \\ &= (1-\nu)^{-1} \rho_{\mathtt{s}} \bigl(\|\breve{\boldsymbol{\xi}}_{\mathtt{s}}\| + \|\boldsymbol{\xi}_{\varkappa}\| \bigr). \end{aligned}$$

Here $\boldsymbol{\xi}_{\varkappa} = \rho_{\mathtt{s}}^{-1} D_{\boldsymbol{\theta}}^{-1} A_{\boldsymbol{\theta}\varkappa} \nabla_{\varkappa}$ and we have also used that $\breve{D}_{\mathtt{s}}^2 \ge \breve{D}^2$. This proves (14). It remains to check (15). If the full model is true then $\mathrm{Var}(\breve{\boldsymbol{\xi}}_{\mathtt{s}}) = I_q$ and under (ED_0), it holds on a set of probability at least $1 - 2\mathrm{e}^{-\mathtt{x}}$

$$\|\breve{\boldsymbol{\xi}}_{\mathtt{s}}\| \le \sqrt{q} + 2\mathtt{x}; \tag{60}$$

see Spokoiny (2012). Similarly, under the correct model specification, it holds $\mathrm{Var}(\nabla_{\varkappa}) = \boldsymbol{I}_{\varkappa}$. For $\boldsymbol{\xi}_{\varkappa} = \rho_{\mathtt{s}}^{-1} D_{\boldsymbol{\theta}}^{-1} A_{\boldsymbol{\theta}\varkappa} \nabla_{\varkappa}$, it holds by (B)

$$\mathrm{Var}(\boldsymbol{\xi}_{\varkappa}) \leq \rho_{\mathtt{s}}^{-2} \bigl\| D_{\boldsymbol{\theta}}^{-1} A_{\boldsymbol{\theta}\varkappa} A_{\boldsymbol{\theta}\varkappa}^{\top} D_{\boldsymbol{\theta}}^{-1} \bigr\| \leq 1,$$

and hence, the use of (ED_0) implies with a dominating probability $1 - 2\mathrm{e}^{-\mathtt{x}}$

$$\bigl\| \boldsymbol{\xi}_{\varkappa} \bigr\| \leq \sqrt{q} + 2\mathtt{x}$$

similarly to (60). In the general situation, if the parametric assumption is not exactly true, we still can use (ED_0). It ensures $\mathrm{Var}(\check{\boldsymbol{\xi}}_{\mathtt{s}}) \leq \mathfrak{a}^2 \boldsymbol{I}_q$, $\mathrm{Var}(\boldsymbol{\xi}_{\varkappa}) \leq \mathfrak{a}^2 \boldsymbol{I}_q$, and

$$\| \check{\boldsymbol{\xi}}_{\mathtt{s}} \| \leq \mathfrak{a}\bigl(\sqrt{q} + 2\mathtt{x}\bigr), \qquad \| \boldsymbol{\xi}_{\varkappa} \| \leq \mathfrak{a}\bigl(\sqrt{q} + 2\mathtt{x}\bigr).$$

This yields the last claim of the theorem.

7.11 Proof of Theorem 5

First we check that the required conditions of Section 2.1 are fulfilled in the considered example. This can be easily done if we slightly change the definition of the local set $\Upsilon_0(\mathtt{r}_0)$. Namely, for $\boldsymbol{u}^* = (u_1^*, \ldots, u_{p_n}^*)^{\top}$, define $\Upsilon_0(\sqrt{\mathfrak{z}})$ as a rectangle

$$\Upsilon_0(\sqrt{\mathfrak{z}}) \stackrel{\mathrm{def}}{=} \bigl\{ \boldsymbol{u} : \; M_n \mathcal{K}(u_j, u_j^*) \leq \mathfrak{z}, \; j = 1, \ldots, p_n \bigr\}.$$

Here $\mathcal{K}(u, u^*)$ is the Kullback-Leibler divergence for the Poisson family:

$$\mathcal{K}(u, u^*) = \mathrm{e}^{u}(u - u^*) - \mathrm{e}^{u} + \mathrm{e}^{u^*}.$$

Lemma 10. *Let $\mathfrak{z}_n$ be such that $2p_n \mathrm{e}^{-\mathfrak{z}_n} \leq 1/2$. Then it holds*

$$I\!P\bigl(\widetilde{\boldsymbol{u}} \in \Upsilon_0(\sqrt{\mathfrak{z}_n})\bigr) \geq 1 - 4p_n \mathrm{e}^{-\mathfrak{z}_n}. \tag{61}$$

In particular, the choice $\mathfrak{z}_n = \mathtt{x}_n + \log(p_n)$ with $\mathtt{x}_n = \mathtt{C} \log n$ provides

$$I\!P\bigl(\widetilde{\boldsymbol{u}} \in \Upsilon_0(\sqrt{\mathfrak{z}_n})\bigr) \geq 1 - 4\mathrm{e}^{-\mathtt{x}_n}. \tag{62}$$

Proof. We use the bound from Polzehl and Spokoiny (2006)

$$I\!P\bigl(M_n \mathcal{K}(\widetilde{u}_j, u_j^*) > \mathfrak{z}_n\bigr) \leq 2\mathrm{e}^{-\mathfrak{z}_n}.$$

This yields

$$I\!P\bigl(\widetilde{\boldsymbol{u}} \in \Upsilon_0(\sqrt{\mathfrak{z}_n})\bigr) \geq \bigl(1 - 2\mathrm{e}^{-\mathfrak{z}_n}\bigr)^{p_n}.$$

Now the elementary inequalities $\log(1-\alpha) \geq -2\alpha$ for $0 \leq \alpha \leq 1/2$ and $\mathrm{e}^{-\delta} \geq 1-\delta$ for $\delta \geq 0$ applied with $\alpha_n = 2\mathrm{e}^{-\mathfrak{z}_n}$ and $\delta_n = 2\alpha_n p_n$ imply

$$(1-\alpha_n)^{p_n} = \mathrm{e}^{\log(1-\alpha_n)p_n} \geq \mathrm{e}^{-2\alpha_n p_n} \geq 1 - 2\alpha_n p_n$$

and (61) follows. □

In the special case $u_1^* = \ldots = u_{p_n}^* = u^*$, the set $\Upsilon_0(\sqrt{\mathfrak{z}})$ is a cube which can be also viewed as a ball in the sup-norm. Moreover, if $\mathfrak{z}_n/(M_n \mathrm{e}^{u^*}) \leq 1/2$, this cube is contained in the cube $\{\boldsymbol{u} : \|\boldsymbol{u} - \boldsymbol{u}^*\| \leq \sqrt{\mathfrak{z}_n/(M_n \mathrm{e}^{u^*})}\}$ in view of $\mathrm{e}^x - 1 - x \leq a^2 \leq 1/2$ for $|x| \leq a \leq 1$. The concentration bound (62) enables us to check the local conditions only on the cube $\Upsilon_0(\sqrt{\mathfrak{z}_n})$. Especially the condition $(\mathcal{ED}_1)$ is trivially fulfilled because $\boldsymbol{\zeta}(\boldsymbol{u}) = \mathcal{L}(\boldsymbol{u}) - I\!\!E\mathcal{L}(\boldsymbol{u})$ is linear in $\boldsymbol{u}$ and θ is a linear functional of $\boldsymbol{u}$. Condition $(\mathcal{L}_0)$ can be checked on $\Upsilon_0(\sqrt{\mathfrak{z}_n})$ with $\delta(\mathfrak{z}_n) = \sqrt{\mathfrak{z}_n/(M_n \mathrm{e}^{u^*})}$.

It remains to compute the value $\breve{D}_0^2$. Define $\beta_n = p_n/M_n^{1/2} = p_n^{3/2}/n^{1/2}$. If $n = p_n^3$, then $\beta_n = 1$.

Lemma 11. *Let $v^* = 1/p_n$. Then it holds*

$$\breve{D}_0^2 = p_n^2 \beta_n^{-2}.$$

Now we are ready to finalize the proof Theorem 5.

Proof. Let β_n be bounded. The definition implies

$$p_n(\theta - \widetilde{\theta}_n) = \sum_{j=1}^{p_n} \log\Big(\frac{\upsilon_j}{Z_j/M_n}\Big).$$

The posterior distribution $\upsilon_j \,|\, \boldsymbol{Y}$ is $\mathrm{Gamma}(\alpha_j, \mu_j)$ with $\alpha_j = 1 + Z_j$ and $\mu_j = \frac{\mu}{M_n\mu+1}$. We use following decomposition

$$\frac{\upsilon_j}{Z_j/M_n} = \frac{M_n \mu_j \alpha_j}{\alpha_j - 1}\big(1 + \alpha_j^{-1/2}\gamma_j\big),$$

where

$$\gamma_j \stackrel{\mathrm{def}}{=} (\alpha_j \mu_j^2)^{-1/2}\big(\upsilon_j - \alpha_j \mu_j\big)$$

has zero mean and unit variance. We can use the Taylor expansion

$$p_n(\theta - \widetilde{\theta}_n) = \sum_{j=1}^{p_n} \log\Big(1 - \frac{1}{M_n\mu + 1}\Big) + \sum_{j=1}^{p_n} \log\Big(1 + \frac{1}{\alpha_j - 1}\Big) + \sum_{j=1}^{p_n} \log\Big(1 + \alpha_j^{-1/2}\gamma_j\Big).$$

Now take into account properties of the real data distribution.

$$\alpha_j = \frac{M_n}{p_n}\Big(1 + \sqrt{\frac{p_n}{M_n}}\delta_j\Big),$$

where δ_j is asymptotically standard normal.

Suppose now that $\beta_n^3/\sqrt{p_n} \to 0$ as $p_n \to \infty$. Then $M_n/p_n = \big(\sqrt{p_n}/\beta_n^3\big)^{2/3} p_n^{2/3} \to \infty$ as $p_n \to \infty$. Thus for p_n sufficient large, $\alpha_j \approx M_n/p_n$. Moreover, it holds for p_n sufficiently large that $\max_{j=1,\ldots,p_n} \alpha_j^{-1/2}|\gamma_j| \le 1/2$ with a high probability. Below we can restrict ourselves to the case when $\alpha_j^{-1/2}|\gamma_j| \le 1/2$. This allows to use the Taylor expansion

$$\begin{aligned} p_n\big(\theta - \widetilde{\theta}_n\big) &= \sum_{j=1}^{p_n} \log\Big(1 - \frac{1}{M_n\mu + 1}\Big) + \sum_{j=1}^{p_n} \log\Big(1 + \frac{1}{\alpha_j - 1}\Big) + \sum_{j=1}^{p_n} \log\big(1 + \frac{\gamma_j}{\sqrt{\alpha_j}}\big) \\ &= \sum_{j=1}^{p_n} \frac{1}{\alpha_j - 1} + \sum_{j=1}^{p_n} \frac{1}{\sqrt{\alpha_j}}\gamma_j - \sum_{j=1}^{p_n} \frac{1}{2\alpha_j}\gamma_j^2 + R. \end{aligned}$$

One can easily check that the remainder R is of order $\beta_n^3/\sqrt{p_n} \to 0$. Moreover, $p_n^{-1/2}\sum_{j=1}^{p_n}\gamma_j$ is asymptotically standard normal, while $p_n^{-1}\sum_{j=1}^{p_n}\gamma_j^2 \xrightarrow{I\!P} 1$. The central limit theorem here can be easily checked because of the Lyapunov condition being valid. Also $\sum_{j=1}^{p_n}(\alpha_j - 1)^{-1} = \frac{p_n^2}{M_n} + o_n(\beta_n^2)$. Now check what happens if $\beta_n \to 0$:

$$\beta_n^{-1} p_n\big(\theta - \widetilde{\theta}_n\big) = \beta_n + \frac{1}{\sqrt{p_n}}\sum_{j=1}^{p_n}\gamma_j - \frac{\beta_n}{2p_n}\sum_{j=1}^{p_n}\gamma_j^2 + o_n(1) \xrightarrow{w} \mathcal{N}(0,1).$$

Similarly, with $\beta_n \equiv \beta$,

$$\beta^{-1} p_n\big(\theta - \widetilde{\theta}_n\big) = \beta + \frac{1}{\sqrt{p_n}}\sum_{j=1}^{p_n}\gamma_j - \frac{\beta_n}{2p_n}\sum_{j=1}^{p_n}\gamma_j^2 + o_n(1) \xrightarrow{w} \mathcal{N}(\beta/2,1).$$

This proves the result for $\beta_n \equiv \beta$. Finally in the case when β_n grows to infinity, but $\beta_n^3/\sqrt{p_n} \to 0$, then $\beta_n^{-1}(\theta - \widetilde{\theta}_n) \xrightarrow{I\!P} \infty$. □

7.12 Proof of Lemma 11

Let $\overline{u}_j = u_j - u_j^*$. Then

$$L(\boldsymbol{u}, \boldsymbol{u}^*) = \mathcal{L}(\boldsymbol{u}) - \mathcal{L}(\boldsymbol{u}^*) = \sum_{j=1}^{p_n}\big\{Z_j\overline{u}_j - M_n p_n^{-1}(\mathrm{e}^{\overline{u}_j} - 1)\big\}.$$

The expected value of Z_j is M_n/p_n which leads to the following expectation of likelihood:

$$\mathbb{E}L(\boldsymbol{u},\boldsymbol{u}^*) = \frac{M_n}{p_n}\sum_{j=1}^{p_n}\bigl(\overline{u}_j - (e^{\overline{u}_j}-1)\bigr) = -\frac{M_n}{p_n}\sum_{j=1}^{p_n}\frac{\overline{u}_j^2}{2} + O\bigl(\|\overline{\boldsymbol{u}}\|^3\bigr).$$

Then we substitute $\overline{u}_1 = p_n\overline{\theta} - \sum_{j=2}^{p_n}\overline{u}_j$, where $\overline{\theta} = \theta - \theta^*$. Thus we get

$$\mathbb{E}L(\boldsymbol{u},\boldsymbol{u}^*) = -\frac{M_n}{p_n}\frac{1}{2}\bigl(p_n\overline{\theta} - \sum_{j=2}^{p_n}\overline{u}_j\bigr)^2 - \frac{M_n}{p_n}\sum_{j=2}^{p_n}\frac{\overline{u}_j^2}{2} + O\bigl(\|\overline{\boldsymbol{u}}\|^3\bigr).$$

This Taylor expansion allows us to compute components of the Fisher information matrix:

$$\mathcal{D}_0^2 = -\nabla^2\mathbb{E}\mathcal{L}(\boldsymbol{u}^*) = \frac{M_n}{p_n}\begin{pmatrix} p_n^2 & -p_n & \ldots & \ldots & -p_n \\ -p_n & 2 & 1 & \ldots & 1 \\ \vdots & 1 & \ddots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & 1 \\ -p_n & 1 & \ldots & 1 & 2 \end{pmatrix}.$$

The Fisher information for the target parameter θ can be computed as follows:

$$\breve{D}_0^2 = M_n p_n\bigl(1 - \boldsymbol{e}^\top\mathbb{H}^{-1}\boldsymbol{e}\bigr),$$

where $\boldsymbol{e} = (1,\ldots,1)^\top$ and $\mathbb{H} = I + E$ with $E = \boldsymbol{e}\boldsymbol{e}^\top$ being the matrix of ones of size $(p_n-1)\times(p_n-1)$. It follows

$$\boldsymbol{e}^\top\mathbb{H}^{-1}\boldsymbol{e} = \mathrm{tr}\bigl(\boldsymbol{e}^\top\mathbb{H}^{-1}\boldsymbol{e}\bigr) = \mathrm{tr}\bigl(\mathbb{H}^{-1}\boldsymbol{e}\boldsymbol{e}^\top\bigr) = \mathrm{tr}\bigl((E+\boldsymbol{I})^{-1}E\bigr).$$

Further, $(E+I)^{-1}E = I - (E+I)^{-1}$ yielding

$$\boldsymbol{e}^\top\mathbb{H}^{-1}\boldsymbol{e} = \mathrm{tr}\bigl\{\boldsymbol{I} - (E+\boldsymbol{I})^{-1}\bigr\} = (p_n-1) - \mathrm{tr}\bigl\{(E+\boldsymbol{I})^{-1}\bigr\} = (p_n-1) - \sum_{j=1}^{p_n}\lambda_j,$$

where λ_j are eigenvalues of matrix $(E+\boldsymbol{I})^{-1}$. It is easy to see that $\lambda_1 = p_n^{-1}$ while $\lambda_2 = \cdots = \lambda_{p_n-1} = 1$. Thus

$$\begin{aligned} \boldsymbol{e}^\top\mathbb{H}^{-1}\boldsymbol{e} &= (p_n-1) - \bigl\{p_n^{-1} + (p_n-2)\bigr\} = 1 - p_n^{-1}, \\ \breve{D}_0^2 &= M_n p_n\bigl(1 - \boldsymbol{e}^\top\mathbb{H}^{-1}\boldsymbol{e}\bigr) = M_n p_n\bigl\{1 - (1-p_n^{-1})\bigr\} = M_n = p_n^2\beta_n^{-2}, \end{aligned}$$

which completes the proof.

7.13 Proof of Lemma 1

It holds

$$\begin{aligned}
&\sup_{\boldsymbol{\gamma}\in I\!R^p} \log I\!E \exp\left\{ \mu \frac{\boldsymbol{\gamma}^\top \nabla \zeta(\boldsymbol{\upsilon}^*)}{\|\mathcal{D}_0 \boldsymbol{\gamma}\|} \right\} \\
&\quad = \sup_{\boldsymbol{\gamma}\in I\!R^p} \log I\!E \exp\left\{ \mu \frac{\sum_{i=1}^n (h'(Y_i - \Psi_i^\top \boldsymbol{\upsilon}^*) - I\!E h'(Y_i - \Psi_i^\top \boldsymbol{\upsilon}^*)) \boldsymbol{\gamma}^\top \Psi_i}{\|\mathcal{D}_0 \boldsymbol{\gamma}\|} \right\} \\
&\quad = \sup_{\boldsymbol{\gamma}\in I\!R^p} \log I\!E \exp\left\{ \frac{\mu \mathrm{h} \sum_{i=1}^n \boldsymbol{\gamma}^\top \Psi_i}{\|\mathcal{D}_0 \boldsymbol{\gamma}\|} h'(\varepsilon)/\mathrm{h} \right\}.
\end{aligned}$$

By definition $\mathrm{h}|\boldsymbol{\gamma}^\top \Psi_i|/\|\mathcal{D}_0 \boldsymbol{\gamma}\| \le N_1^{-1/2}$ and hence $\mu \mathrm{h}|\boldsymbol{\gamma}^\top \Psi_i|/\|\mathcal{D}_0\|\boldsymbol{\gamma} \le \mathtt{g}_1$. Thus,

$$\sup_{\boldsymbol{\gamma}\in I\!R^p} \log I\!E \exp\left\{ \mu \frac{\boldsymbol{\gamma}^\top \nabla \zeta(\boldsymbol{\upsilon}^*)}{\|\mathcal{D}_0 \boldsymbol{\gamma}\|} \right\} \le \sup_{\boldsymbol{\gamma}\in I\!R^p} \frac{\nu_0^2 \mu^2 \mathrm{h}^2}{2} \frac{\sum_{i=1}^n |\Psi_i^\top \boldsymbol{\gamma}|^2}{\|\mathcal{D}_0 \boldsymbol{\gamma}\|^2} = \frac{\nu_0^2 \mu^2}{2}.$$

7.14 Proof of Lemma 5

It holds

$$\begin{aligned}
\|\mathcal{D}_0^{-1}(\mathcal{D}^2(\boldsymbol{\upsilon}) - \mathcal{D}_0^2)\mathcal{D}_0^{-1}\| &= \sup_{\boldsymbol{\gamma}\in\mathbb{R}^p:\ \|\boldsymbol{\gamma}\|=1} |\boldsymbol{\gamma}^\top \mathcal{D}_0^{-1}(\mathcal{D}^2(\boldsymbol{\upsilon}) - \mathcal{D}_0^2)\mathcal{D}_0^{-1}\boldsymbol{\gamma}| \\
&\le \sup_{\boldsymbol{\gamma}\in\mathbb{R}^p:\ \|\boldsymbol{\gamma}\|=1} \Big| \sum_{i=1}^n (d''(\Psi_i^\top \boldsymbol{\upsilon}) - d''(\Psi_i^\top \boldsymbol{\upsilon}^*)) \boldsymbol{\gamma}^\top \mathcal{D}_0^{-1} \Psi_i \Psi_i^\top \mathcal{D}_0^{-1} \boldsymbol{\gamma} \Big| \\
&\le \sup_{\boldsymbol{\gamma}\in\mathbb{R}^p:\ \|\boldsymbol{\gamma}\|=1} \sum_{i=1}^n |d''(\Psi_i^\top \boldsymbol{\upsilon}) - d''(\Psi_i^\top \boldsymbol{\upsilon}^*)| \boldsymbol{\gamma}^\top \mathcal{D}_0^{-1} \Psi_i \Psi_i^\top \mathcal{D}_0^{-1} \boldsymbol{\gamma} \\
&\le \sup_{\boldsymbol{\gamma}\in\mathbb{R}^p:\ \|\boldsymbol{\gamma}\|=1} \sum_{i=1}^n L |\Psi_i^\top (\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^*)| \boldsymbol{\gamma}^\top \mathcal{D}_0^{-1} \Psi_i \Psi_i^\top \mathcal{D}_0^{-1} \boldsymbol{\gamma} \\
&\le L N^{-1/2} \|\mathcal{D}_0(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^*)\| \sup_{\boldsymbol{\gamma}\in\mathbb{R}^p:\ \|\boldsymbol{\gamma}\|=1} \boldsymbol{\gamma}^\top \mathcal{D}_0^{-1} \Big(\sum_{i=1}^n d''(\Psi_i^\top \boldsymbol{\upsilon}^*) \Psi_i \Psi_i^\top \Big) \mathcal{D}_0^{-1} \boldsymbol{\gamma} \\
&\le L \frac{\mathtt{r}}{N_2^{1/2}}.
\end{aligned}$$

References

Barron, A., Schervish, M. J., and Wasserman, L. (1996). "The Consistency of Posterior Distributions in Nonparametric Problems." *The Annals of Statistics*, 27: 536–561. 665

Bickel, P. J. and Kleijn, B. J. K. (2012). "The semiparametric Bernstein-von Mises theorem." *The Annals of Statistics*, 40(1): 206–237. 666, 676

Bochkina, N. and Green, P. J. (2014). "The Bernstein-von Mises theorem and non-regular models." *The Annals of Statistics*, 42(5): 1850–1878. 668

Bontemps, D. (2011). "Bernstein–von Mises theorem for Gaussian regression with increasing number of regressors." *The Annals of Statistics*, 39(5): 2557–2584. 666, 674, 676, 680

Boucheron, S. and Gassiat, E. (2009). "A Bernstein-von Mises theorem for discrete probability distributions." *Electronic Journal of Statistics*, 3: 114–148. 666, 677

Boucheron, S. and Massart, P. (2011). "A high-dimensional Wilks phenomenon." *Probability Theory and Related Fields*, 150: 405–433. 666

Castillo, I. (2012). "A semiparametric Bernstein–von Mises theorem for Gaussian process priors." *Probability Theory and Related Fields*, 152: 53–99. 666, 676

Castillo, I. and Nickl, R. (2013). "Nonparametric Bernstein–von Mises theorems in Gaussian white noise." *The Annals of Statistics*, 41(4): 1999–2028. 666

Castillo, I. and Rousseau, J. (2013). "A General Bernstein–von Mises Theorem in semiparametric models." Available at arXiv:1305.4482 [math.ST]. 666

Cheng, G. and Kosorok, M. R. (2008). "General frequentist properties of the posterior profile distribution." *The Annals of Statistics*, 36(4): 1819–1853. 666

Chernozhukov, V. and Hong, H. (2003). "An MCMC approach to classical estimation." *Journal of Econometrics*, 115(2): 293–346. 667, 668

Cox, D. D. (1993). "An analysis of Bayesian inference for nonparametric regression." *The Annals of Statistics*, 21(2): 903–923. 665

Freedman, D. (1999). "On the Bernstein-von Mises theorem with infinite-dimensional parameters." *The Annals of Statistics*, 27(4): 1119–1140. 665

Ghosal, S. (1999). "Asymptotic normality of posterior distributions in high-dimensional linear models." *Bernoulli*, 5(2): 315–331. 665, 666, 677

— (2000). "Asymptotic normality of posterior distributions for exponential families when the number of parameters tends to infinity." *Journal of Multivariate Analysis*, 74(1): 49–68. 666, 677

Ibragimov, I. and Khas'minskij, R. (1981). *Statistical estimation. Asymptotic theory. Translated from the Russian by Samuel Kotz*. New York–Heidelberg–Berlin: Springer-Verlag. 669

Johnstone, I. M. (2010). "High dimensional Bernstein–von Mises: simple examples." In *Borrowing strength: theory powering applications—a Festschrift for Lawrence D. Brown*, volume 6 of *Institute of Mathematical Statistics Collections*, 87–98. Beachwood, OH: Institute of Mathematical Statistics. 665, 666, 674, 677, 680

Kim, Y. (2006). "The Bernstein–von Mises theorem for the proportional hazard model." *The Annals of Statistics*, 34(4): 1678–1700. 666

Kim, Y. and Lee, J. (2004). "A Bernstein–von Mises theorem in the nonparametric right-censoring model." *The Annals of Statistics*, 32(4): 1492–1512. 666

Kleijn, B. J. K. and van der Vaart, A. W. (2006). "Misspecification in infinite-dimensional Bayesian statistics." *The Annals of Statistics*, 34(2): 837–877. 667

— (2012). "The Bernstein-von-Mises theorem under misspecification." *Electronic Journal of Statistics*, 6: 354–381. 667

Le Cam, L. and Yang, G. L. (1990). *Asymptotics in Statistics: Some Basic Concepts*. Springer in Statistics. 665

Leahu, H. (2011). "On the Bernstein-von Mises phenomenon in the Gaussian white noise model." *Electronic Journal of Statistics*, 5: 373–404. 666

McCullagh, P. and Nelder, J. (1989). *Generalized linear models. 2nd ed.*. Monographs on Statistics and Applied Probability. 37. London etc.: Chapman and Hall. 684

Polzehl, J. and Spokoiny, V. (2006). "Propagation-separation approach for local likelihood estimation." *Probability Theory and Related Fields*, 135(3): 335–362. 704

Rivoirard, V. and Rousseau, J. (2012). "Bernstein–von Mises theorem for linear functionals of the density." *The Annals of Statistics*, 40(3): 1489–1523. 666

Schwartz, L. (1965). "On Bayes Procedures." *Probability Theory and Related Fields*, 4(1): 10–26. 665

Shen, X. (2002). "Asymptotic normality of semiparametric and nonparametric posterior distributions." *Journal of the American Statistical Association*, 97(457): 222–235. 665

Spokoiny, V. (2012). "Parametric estimation. Finite sample theory." *The Annals of Statistics*, 40(6): 2877–2909. 666, 669, 671, 677, 686, 688, 689, 693, 704

van der Vaart, A. W. (2000). *Asymptotic Statistics (Cambridge Series in Statistical and Probabilistic Mathematics)*. Cambridge University Press. 665

Acknowledgments

The authors would like to thank the editors and the two anonymous reviewers for their constructive comments and suggestions on the original version of this paper. The research results of sections 3 and 5 are developed in IITP RAS with the support of the grant by the Russian Science Foundation (project 14-50-00150). The research results of sections 2 and 4 are partially supported by Laboratory for Structural Methods of Data Analysis in Predictive Modeling, MIPT, RF government grant, ag. 11.G34.31.0073, and German Research Foundation (DFG) through the Collaborative Research Center 649 "Economic Risk".