A Bernstein - von Mises Theorem for growing parameter
dimension

Vladimir Spokoiny∗
Weierstrass-Institute,
Humboldt University Berlin,
Moscow Institute of Physics and Technology
Mohrenstr. 39, 10117 Berlin, Germany,
spokoiny@wias-berlin.de
August 13, 2013

Abstract

The prominent Bernstein – von Mises (BvM) Theorem claims that the posterior distribution is asymptotically normal and its mean is nearly the maximum likelihood estimator (MLE), while its variance is nearly the inverse of the total Fisher information matrix, as for the MLE. This paper revisits the classical result from different viewpoints. Particular issues to address are: nonasymptotic framework with just one finite sample, possible model misspecification, and a large parameter dimension. In particular, in the case of an i.i.d. sample, the BvM result can be stated for any smooth parametric family provided that the dimension $p$ of the parameter space satisfies the condition “$p^3/n$ is small”.

AMS 2000 Subject Classification: Primary 62F15. Secondary 62F25

Keywords: posterior, concentration, bracketing, Gaussian approximation

∗The author is partially supported by Laboratory for Structural Methods of Data Analysis in Predictive Modeling, MIPT, RF government grant, ag. 11.G34.31.0073. Financial support by the German Research Foundation (DFG) through the Collaborative Research Center 649 “Economic Risk” is gratefully acknowledged. The author thanks the AE and two anonymous referees for very useful comments and suggestions.
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 variance close to the variance of the MLE. This explains why this result is often considered as the Bayes counterpart of the frequentist Fisher Theorem about asymptotic normality of the MLE. The BvM result provides a theoretical background for different Bayesian procedures. In particularly, one can use Bayesian computations for evaluation of the MLE and its variance. Also one can build elliptic credible sets using 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 Kleijn and van der Vaart (2012) for a detailed historical overview around the BvM result. Modern statistical problems require to consider situations with very complicated models involving a lot of parameters and with limited sample size. These applications motivated further research focused on properties of the posterior distribution in non-classical situations. There is a number of papers in this direction recently appeared. We mention Ghosal et al. (2000); Ghosal and van der Vaart (2007) for a general theory in the i.i.d. case; Ghosal (1999), Ghosal (2000) for high dimensional linear models; Boucheron and Gassiat (2009), Kim (2006) for non-Gaussian models; Shen (2002), Bickel and Kleijn (2012), and Castillo (2012) for the semiparametric version of the BvM result; Kleijn and van der Vaart (2006), Bunke and Milhaud (1998), for the misspecified parametric case, among many others. Probably the most general framework for the BvM result is given in Kleijn and van der Vaart (2012) in terms of the so called stochastic LAN conditions. This condition extends the classical LAN condition and basically means a kind of quadratic expansion of the log-likelihood in a root-n vicinity of the central point. The approach applies even if the parametric assumption is misspecified, however, the approach requires a fixed parametric model and large samples.

Extensions to nonparametric models leads to situations with infinite or growing parameter dimension $p$ as well as some modeling bias (model misspecification). There exist some BvM results for this situation, see e.g. Freedman (1999) and Ghosal (1999, 2000) for linear models or Bontemps (2011) for Gaussian regression. This paper attempts to investigate the problem of applicability of the BvM result in the general situation when the parameter dimension $p$ is large and may grow with the sample size. It appears that the standard LAN arguments do not apply in this situation because all the small terms in the LAN expansion may depend on the dimension $p$. Recently Spokoiny (2012) offered a version of the Le Cam LAN theory which applies to finite samples, large parameter dimension, and is robust against possible model misspecification. An important benefit
of the bracketing approach of Spokoiny (2012) is that all the error terms in the expansion are given explicitly. In particular, one can track the impact of the dimension $p$ as well as some other parameters like model identifiability and regularity. This paper follows the same path and our main result can be viewed as a finite sample version of the BvM theorem. All the error terms in the BvM expansion are given explicitly up to some absolute constants. In a special case of an i.i.d. sample, one can precisely control how the error terms depend on the sample size $n$ and the dimension $p$ and judge about the applicability of the approach when $p$ grows with $n$.

Another important issue is a possible model misspecification. The classical parametric theory is essentially parametric in the sense that it requires the parametric assumption to be exactly fulfilled. Only few results are available which admit some violation from the parametric assumption; see e.g. Kleijn and van der Vaart (2006), Kleijn and van der Vaart (2012). The latter paper extends the BvM result from the classical LAN framework to the so called stochastic LAN under a possible model misspecification. The bracketing approach of Spokoiny (2012) is similar to the stochastic LAN but is stated for finite samples and is not limited to root-n vicinity of the true point. Both issues are essential and useful when the situation with growing parameter dimension is considered. This study admits from the very beginning that the parametric specification is probably wrong. However, if the model is correct, the presented bounds imply the classical asymptotic BvM result. We briefly comment on the use of Bayesian credible sets as frequentist confidence sets. The classical asymptotic results like Fisher or Wilks expansion justify this use. However, for finite samples and possible model misspecification, one has to be careful and relation between posterior probability of elliptic credible sets and their coverage probability is not so obvious; see Section 2.4 for more discussion.

First we specify our set-up following to Spokoiny (2012). Let $Y$ denote the observed data and $P$ mean their distribution. A general parametric assumption (PA) means that $P$ belongs to $p$-dimensional family $(P_{\theta}, \theta \in \Theta \subseteq \mathbb{R}^p)$ dominated by a measure $\mu_0$. This family yields the log-likelihood function $L(\theta) = L(Y, \theta) \overset{\text{def}}{=} \log \frac{dP_{\theta}}{d\mu_0}(Y)$. The PA can be misspecified, so, in general, $L(\theta)$ is a quasi log-likelihood. The classical likelihood principle suggests to estimate $\theta$ by maximizing the function $L(\theta)$:

$$\tilde{\theta} \overset{\text{def}}{=} \arg\max_{\theta \in \Theta} L(\theta).$$

If $P \not\in (P_{\theta})$, then the (quasi) MLE estimate $\tilde{\theta}$ from (1.1) is still meaningful and it appears to be an estimate of the value $\theta^*$ defined by maximizing the expected value of $L(\theta)$:

$$\theta^* \overset{\text{def}}{=} \arg\max_{\theta \in \Theta} \mathbb{E} L(\theta).$$

(1.2)
\( \theta^* \) is the true value in the parametric situation and can be viewed as the parameter of the best parametric fit in the general case. Spokoiny (2012) states the following extension of the prominent Fisher expansion of the qMLE \( \tilde{\theta} \):

\[
D_0(\tilde{\theta} - \theta^*) \approx \xi \overset{\text{def}}{=} D_0^{-1}\nabla L(\theta^*),
\]

where \( \nabla L(\theta) = \frac{dL}{d\theta}(\theta) \) and \( D_0^2 \overset{\text{def}}{=} -\nabla^2 L(\theta^*) \) is the analog of the total Fisher information matrix. In the classical situation, the standardized score \( \xi \) is asymptotically standard normal yielding asymptotic root-n normality and efficiency of the MLE \( \tilde{\theta} \).

Now we switch to the Bayes setup with a random element \( \vartheta \) following a prior measure \( \Pi \) on the parameter set \( \Theta \). The posterior describes the conditional distribution of \( \vartheta \) given \( Y \) obtained by normalization of the product \( \exp\{L(\theta)\} \Pi(d\theta) \). This relation is usually written as

\[
\vartheta \mid Y \propto \exp\{L(\theta)\} \Pi(d\theta).
\]

An important feature of our analysis is that \( L(\theta) \) 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 (1.4) describes a quasi posterior; cf. Chernozhukov and Hong (2003). Section 2 studies some general properties of the posterior measure focusing on the case of a non-informative prior \( \Pi \). The main result claims that the distribution of \( D_0(\vartheta - \theta^*) - \xi \) given \( Y \) is nearly standard normal. Comparing with (1.3) indicates that the posterior is nearly centered at the qMLE \( \tilde{\theta} \) and its variance is close to \( D_0^{-2} \). So, our result extends the BvM theorem to the considered general setup. Section 2.5 comments how the result can be extended to the case of any regular prior. The study is nonasymptotic and all “small” terms are carefully described up to an absolute constant. This helps to understand how the parameter dimension is involved and particularly address the question of a critical dimension. It appears that in the special i.i.d. case with \( n \) observations the presented BvM result continues to hold as long as \( p^3/n \) is small.

The paper is organized as follows. Section 2.2 states our main result which can be viewed as a version of the BvM theorem for the case of a non-informative prior. Sections 2.3 and 2.4 discuss how this result can be used for building the elliptical credible sets and how these sets can also be used as frequentist confidence sets. Some extensions to a regular prior are given in Section 2.5. The results are illustrated on the cases of an i.i.d. sample in Section 2.6. In particular, we show that the BvM result holds if the sample size \( n \) is larger in order than the cube of the parameter dimension \( p \). In this case the classical asymptotic BvM statement follows immediately from our results. The imposed conditions are collected in Section A, Section B presents some auxiliary
results on the properties of the posterior, and the proofs are collected in Section C of the Appendix.

2 Main results

This section presents our main results which state the Bernstein–von Mises Theorem for a non-classical and non-asymptotic framework. First we discuss the special case of a non-informative prior given by the Lebesgue measure \( \pi(\theta) \equiv 1 \) on \( \mathbb{R}^p \). Then we extend the results to any prior with a continuous density on a vicinity of the central point \( \theta^* \).

The study uses some recent results from Spokoiny (2012). Section 2.1 provides a brief overview of the bracketing and upper function devices required for our study. The main result is stated in Section 2.2, then we discuss some corollaries and extensions.

2.1 Main tools. Local bracketing and upper function devices

Introduce the notation \( L(\theta, \theta^*) = L(\theta) - L(\theta^*) \) for the (quasi) log-likelihood ratio. The main step in the approach of Spokoiny (2012) is the following uniform local bracketing result:

\[
\mathbb{L}_\epsilon(\theta, \theta^*) - \diamond \leq L(\theta, \theta^*) \leq \mathbb{L}_\epsilon(\theta, \theta^*) + \diamond, \quad \theta \in \Theta_0. \tag{2.1}
\]

Here \( \mathbb{L}_\epsilon(\theta, \theta^*) \) and \( \mathbb{L}_\epsilon(\theta, \theta^*) \) are quadratic in \( \theta - \theta^* \) expressions, \( \diamond \) and \( \diamond \) are small errors only depending on \( \Theta_0 \) which is a local vicinity of the central point \( \theta^* \). This result can be viewed as an extension of the famous Le Cam local asymptotic normality (LAN) condition. The LAN condition postulates an approximation of the log-likelihood \( L(\theta) \) by one quadratic process; see e.g. Ibragimov and Khas’minskij (1981) or Kleijn and van der Vaart (2012) for an extension of this condition (stochastic LAN). The use of the bracketing device with two different quadratic expressions allows to control of the error terms \( \diamond, \diamond \) even for relatively large neighborhoods \( \Theta_0 \) of \( \theta^* \) while the LAN approach is essentially restricted to a root-n vicinity of \( \theta^* \). Another benefit of the bracketing approach is that it applies to an arbitrary parameter dimension and a possible model misspecification. The bracketing bound (2.1) requires some conditions which are listed in Section A of the Appendix. Similarly to the LAN theory, the bracketing result has a number of remarkable corollaries like the Wilks and Fisher Theorems; see Spokoiny (2012). Below we show that the BvM result is in some sense also a corollary of (2.1).

For making a precise statement, we have to specify the ingredients of the bracketing device. The most important one is a symmetric positive \( p \times p \)-matrix \( D_0^2 \). In typical situations, it can be defined as the negative Hessian of the expected log-likelihood: \( D_0^2 = -\nabla^2 \mathbb{E} L(\theta^*) \). Also one has to specify a radius \( r_0 \), and a non-negative constant \( \epsilon < 1 \).
which all are related to the conditions of Section A. The local vicinity $\Theta_0(x_0)$ of the central point $\theta^*$ is defined as $\Theta_0(x_0) = \{\theta : \|D_0(\theta - \theta^*)\| \leq r_0\}$. The bracketing result (2.1) can be stated for $\Theta_0 = \Theta_0(x_0)$ with

$$\mathbb{L}_\epsilon(\theta, \theta^*) \overset{\text{def}}{=} (\theta - \theta^*)^\top \nabla L(\theta^*) - \|D_\epsilon(\theta - \theta^*)\|^2/2$$

(2.2)

and

$$D_\epsilon^2 = (1 - \epsilon)D_0^2, \quad \xi_\epsilon \overset{\text{def}}{=} D_\epsilon^{-1} \nabla L(\theta^*),$$

(2.3)

and similarly for $\xi = -\epsilon$. The construction is slightly changed relative to Spokoiny (2012) (in fact, it is simplified) by using only one matrix $D_0^2$ while Spokoiny (2012) used two matrices $D_0^2$ and $V_0^2$. However, the construction are equivalent under the identifiability condition $(I)$; see Section A. The error terms $\diamond_\epsilon = \diamond_\epsilon(x_0)$ and $\diamond = \diamond_\epsilon(x_0)$ follow the probability bound given in Proposition 3.7 of Spokoiny (2012); see also Theorem 2.1 below. The bracketing bound (2.1) becomes useful if these errors are relatively small and can be neglected. The stochastic LAN condition of Kleijn and van der Vaart (2012) is similar in flavor but it stated in the asymptotic form with a fixed parameter set and a growing sample size.

The local bracketing result (2.1) has to be accompanied with a large deviation bound which ensures a small probability of the event $\tilde{\theta} \notin \Theta_0(x_0)$. Spokoiny (2012) explains how such a bound can be obtained by the upper function approach. An upper function $u(\theta)$ ensures (with a high probability) that $L(\theta, \theta^*) \leq -u(\theta)$ uniformly in $\theta \in \Theta \setminus \Theta_0(x_0)$. More precisely, given $x > 0$, there exist a random set $\Omega_n$ with $\mathbb{P}(\Omega_n) \geq 1 - e^{-x}$ and a function $u(\theta) = u(\theta, x)$ such that on $\Omega_n$, it holds

$$L(\theta, \theta^*) \leq -u(\theta), \quad \theta \in \Theta \setminus \Theta_0(x_0).$$

(2.4)

The further study will only use (2.1) and (2.4). We impose the same conditions as in Spokoiny (2012) and the statements (2.1) and (2.4) are Theorem 2.1 and Theorem 4.1 of that paper. However, at one point there is an essential difference. The results of Spokoiny (2012) are stated for just one finite sample and an arbitrary parameter dimension $p$. Below we aim to study the case of a growing parameter dimension $p$. This requires to introduce an asymptotic parameter denoted by $n \to \infty$ which can be viewed as the sample size. We assume that all considered objects depend on $n$ including the likelihood function, the parameter set $\Theta$ and its dimension $p$, as well as all the constants in our conditions. This especially concerns the constant $a$ which measures the model identifiability and model misspecification, and the functions $\delta(r)$ and $\omega(r)$ which are
measure the regularity of the considered parametric family. The primary goal of our study is to fix some sufficient conditions ensuring the BvM result.

Below by \( C \) is denoted a generic absolute constant, not necessarily the same. However, it is independent of the important constants like \( p, a, \) etc. We also suppose a growing sequence \( x_n \) with \( x_n \leq p_n \) to be fixed. For instance, one can fix \( x_n \) proportional to the dimension \( p_n : x_n = Cp_n \). By \( \Omega_n \) we denote a random event of a dominating probability such that

\[
\mathbb{P}(\Omega_n) \geq 1 - Ce^{-x_n}.
\]

The next theorem presents some sufficient conditions for (2.1) and (2.4) and it is proved in a bit more general form in Spokoiny (2012).

**Theorem 2.1.** Let \( p = p_n \to \infty \). Suppose the conditions \((ED_0), (ED_1), (L_0), \) and \((I)\) from Section A with \( r_0^2 \geq C(a^2 + 1)p \) for a fixed constant \( C \), where \( a^2 \) is the constant from condition \((L_0)\). Let also the constant \( \epsilon \) fulfill

\[
\epsilon \geq 3\nu_0 a^2 \omega(r_0) + \delta(r_0).
\]

Then \( L_\epsilon(\theta, \theta^*) \) and \( \underline{L}_\epsilon(\theta, \theta^*) \) from (2.2) and (2.3) fulfill

\[
L_\epsilon(\theta, \theta^*) - \Diamond_\epsilon(r_0) \leq L(\theta, \theta^*) \leq L_\epsilon(\theta, \theta^*) + \Diamond_\epsilon(r_0), \quad \theta \in \Theta_0(r_0),
\]

and the error terms \( \Diamond_\epsilon(r_0) \) and \( \Diamond_\underline{L}(r_0) \) fulfill on a random set \( \Omega_n \) of a dominating probability the bounds

\[
\Diamond_\epsilon(r_0) \leq C\epsilon p, \quad \Diamond_\underline{L}(r_0) \leq C\epsilon p.
\]

Moreover, the random vector \( \xi = D_0^{-1}\nabla L(\theta^*) \) fulfills on \( \Omega_n \)

\[
\|\xi\|^2 \leq CE\|\xi\|^2 \leq Ca^2 p.
\]

Furthermore, assume \((Er)\) and \((Lr)\) with \( b(r) \equiv b \) yielding

\[
-EL(\theta, \theta^*) \geq b\|D_0(\theta - \theta^*)\|^2
\]

for each \( \theta \in \Theta \setminus \Theta_0(r_0) \). Let also \( r_0^2 \geq C(a^2 \lor b^{-1})p \) and \( g(r) \geq Cb \) for all \( r \geq r_0 \); see \((Er)\). Then (2.4) holds on \( \Omega_n \) with \( u(\theta) = b\|D_0(\theta - \theta^*)\|^2/2 \).

The results (2.5) and (2.6) are stated in Spokoiny (2012) but for a slightly different definition of \( D_\epsilon^2 \) and \( D^2 \). Namely, for \( \epsilon = (\delta, \varrho) \), one defines \( D_\epsilon^2 = D_0^2(1 - \delta) - \varrho V_0^2 \).
and similarly for \( \varepsilon = -\varepsilon \). However, these two constructions are essentially equivalent due to the identifiability condition \((I)\). Indeed, \( a^2 D_0^2 \geq V_0^2 \) implies
\[
D_0^2 (1 - \delta) - \varrho V_0^2 \geq (1 - \varepsilon) D_0^2, \quad D_0^2 (1 + \delta) + \varrho V_0^2 \leq (1 + \varepsilon) D_0^2
\]
with \( \varepsilon = \delta + a^2 \varrho \).

In what follows we will assume that the conditions of this theorem are fulfilled and will use the statements \((2.5)\) and \((2.4)\) and the bounds \((2.6)\) and \((2.7)\). The results of Theorem 2.1 yield a number of important corollaries. We present one of them concerning an expansion of the qMLE \( \bar{\theta} \); see Corollary 3.4 in Spokoiny (2012).

**Corollary 2.2.** Suppose the conditions of Theorem 2.1 with \( \varepsilon \leq 1/2 \). Then
\[
\| D_0 (\bar{\theta} - \theta^*) - \xi_\varepsilon \|_2^2 \leq 2 \Delta_\varepsilon (r_0),
\]
where
\[
\Delta_\varepsilon (r_0) \dot{=} \diamond \varepsilon (r_0) + \diamond \xi_\varepsilon (r_0) + (\| \xi_\varepsilon \|^2 - \| \xi_\varepsilon \|^2)/2 \quad (2.8)
\]
and with \( q \dot{=} p + \| \xi \|^2 \)
\[
\Delta_\varepsilon (r_0) \leq C \varepsilon p + \varepsilon \| \xi \|^2 \leq C \varepsilon q.
\]
Moreover, it holds on \( \Omega_n \) with \( q^* = p + E\| \xi \|^2 \)
\[
\Delta_\varepsilon (r_0) \leq C \varepsilon q^*.
\]

The value \( \Delta_\varepsilon (r_0) \) called the *spread* can be viewed as the error induced by the bracketing device. The Wilks and Fisher results from Spokoiny (2012) apply for growing \( p \) under the condition “\( \varepsilon \) is small” because this yields \( \Delta_\varepsilon (r_0) \ll q \).

**2.2 A Bernstein - von Mises Theorem**

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. This information can be effectively used for building Bayesian credible sets and frequentist confidence sets with elliptic shape.

Let \( D_0^2 \) be the matrix from condition \((L_0)\). Define also
\[
\tilde{\theta} = \theta^* + D_0^{-2} \nabla L(\theta^*).
\]
The result of Corollary 2.2 implies the expansion of the MLE $\tilde{\theta}$ in the form $\tilde{\theta} \approx \hat{\theta}$. Introduce the posterior moments

$$\bar{\theta} \overset{\text{def}}{=} \mathbb{E}(\theta \mid Y), \quad \mathbb{S}^2 \overset{\text{def}}{=} \text{Cov}(\theta \mid Y) \overset{\text{def}}{=} \mathbb{E}\{(\theta - \bar{\theta})(\theta - \bar{\theta})^\top \mid Y\}.$$ 

This section presents a version of the BvM result in the considered nonasymptotic setup which claims that $\bar{\theta}$ is close to $\hat{\theta}$, $\mathbb{S}^2$ is nearly equal to $D_0^{-2}$, and $D_0(\theta - \hat{\theta})$ is nearly standard normal conditionally on $Y$.

**Theorem 2.3.** Suppose the conditions of Theorem 2.1 with $\epsilon \leq 1/2$. It holds on $\Omega_n$ with $q = p + \|\xi\|^2$

$$\|D_0(\theta - \hat{\theta})\|^2 \leq C\epsilon q,$$

$$\|I_p - D_0\mathbb{S}^2D_0\|_\infty \leq C\epsilon q. \quad (2.9)$$

Moreover, for any $\lambda \in \mathbb{R}^p$ with $\|\lambda\|^2 \leq p$

$$\left|\log\mathbb{E}\left[\exp\{\lambda^\top D_0(\theta - \hat{\theta})\} \mid Y\right] - \|\lambda\|^2/2\right| \leq C\Delta^*_\epsilon \leq C\epsilon q. \quad (2.10)$$

For any measurable set $A \subset \mathbb{R}^p$,

$$\mathbb{P}(D_0(\theta - \hat{\theta}) \in A \mid Y) \geq \exp(-C\epsilon q)\left\{\mathbb{P}(\gamma \in A) - C\epsilon q^{1/2}\right\} - Ce^{-\Delta^*_n},$$

$$\mathbb{P}(D_0(\theta - \hat{\theta}) \in A \mid Y) \leq \exp(C\epsilon q)\left\{\mathbb{P}(\gamma \in A) + C\epsilon q^{1/2}\right\} + Ce^{-\Delta^*_n}.$$ 

One can see that all statements of Theorem 2.3 require "$\epsilon q$ is small". Later we show that the results continue to hold if $\hat{\theta}$ is replaced by any efficient estimate $\tilde{\theta}$, e.g. by the MLE $\tilde{\theta}$, satisfying $\|D_0(\tilde{\theta} - \hat{\theta})\|^2 \leq C\epsilon q$ with a dominating probability.

### 2.3 Elliptic credible sets

This section discusses a possibility of building some Bayesian credible sets in the elliptic form motivated by the Gaussian approximation of the posterior. The BvM result ensures that the posterior can be well approximated by the normal law with the mean $\hat{\theta}$ and the covariance $D_0^{-2}$. In particular, if $z_\alpha$ is a proper quantile of the chi-squared distribution $\chi^2_p$, that is,

$$\mathbb{P}(\|\gamma\|^2 > z_\alpha) = \alpha$$

then the posterior probability of the set

$$C^\circ(z_\alpha) = \{\theta : \|D_0(\theta - \hat{\theta})\|^2 \leq z_\alpha\} \quad (2.11)$$
is close to $1 - \alpha$ up to an error term of order $\epsilon q^*$. Unfortunately, the quantities $\hat{\theta}$ and $D_0^{-2}$ are unknown and cannot be used for building the elliptic credible sets. A natural question is whether one can replace these values by some empirical counterparts without any substantial change of the posterior mass. Due to Lemma C.1 below, the use of the estimates $\hat{\theta}$ and $D_0^{-2}$ in place of $\hat{\theta}$ and $D_0^2$ does not significantly affect the posterior mass of the set $C^*(\bar{\theta})$ provided that

$$\|D_0(\bar{\theta} - \hat{\theta})\|^2 \leq C \epsilon q, \quad \text{tr}(D_0^{-1}\hat{D}^2D_0^{-1} - I)^2 \leq C \epsilon q. \quad (2.12)$$

Theorem 2.3 justifies the use of the posterior mean $\overline{\theta}$ in place of $\hat{\theta}$. The next important question is whether the posterior covariance $\mathcal{S}^2$ is a reasonable estimate of $D_0^{-2}$. Unfortunately, the result (2.9) only implies that

$$\text{tr}(D_0^{-1}\mathcal{S}^2D_0^{-1} - I)^2 \leq C(\epsilon q)^2 p.$$ \hfill (2.13)

This yields that the use of credible sets in the form

$$C(\bar{\theta}) = \{\theta : \|\mathcal{S}^{-1}(\theta - \overline{\theta})\|^2 \leq \bar{\theta}\}$$

is only justified if $\epsilon^2 p^3$ is small. If the dimension $p$ is fixed we only need $\epsilon$ small. If the dimension $p$ is large, the use of posterior covariance can be too rough.

Alternatively, one can use a plug-in estimator of the matrix $D_0^2$. Namely, if $D_0^2(\theta) \overset{\text{def}}{=} -\nabla^2 E_\theta L(\theta)$ is the total Fisher information matrix at the point $\theta$, then one can define $\hat{D}^2 = D^2(\hat{\theta})$, where $\hat{\theta}$ is a consistent estimator of $\theta^*$. Due to Theorem 2.3, the posterior mean $\overline{\theta}$ is a natural candidate for $\hat{\theta}$. The only condition to check is that $\text{tr}(D_0^{-1}D^2(\theta)D_0^{-1} - I)^2$ is small for all $\theta$ from the set $\Theta_0$ on which the estimator $\hat{\theta}$ concentrates with a dominating probability. However, this approach does not include the case of a model misspecification because the matrix $D^2(\theta)$ is computed under the parametric measure.

**Corollary 2.4.** Let $\mathcal{P} = \mathcal{P}_{\Theta^*}$ and the conditions of Theorem 2.1 be fulfilled. It holds on $\Omega_n$ for any $\bar{\theta} > 0$

$$\mathcal{P}^*(C(\bar{\theta}) \mid Y) \leq \exp(C \epsilon q) \{\mathcal{P}(\|\gamma\|^2 \leq \bar{\theta}) + C \epsilon q^{1/2}\} + Ce^{-x_n},$$

where

$$C(\bar{\theta}) \overset{\text{def}}{=} \{\theta : \|D(\overline{\theta})(\theta - \overline{\theta})\|^2 \leq \bar{\theta}\}.$$
2.4 Coverage probability and model misspecification

Here we briefly discuss the question of using the Bayesian credible sets like \( \mathcal{C}(z) \) or \( \overline{\mathcal{C}}(z) \) as frequentist confidence sets. To simplify the discussion, we consider the sets \( \mathcal{C}^\circ(z) \) supposing that reasonable estimators of \( D_0^2 \) and \( \hat{\theta} \) are available. The main question to answer is whether such sets can be used as frequentist confidence sets for the true parameter \( \theta^* \). This leads to study of the coverage probability \( \mathbb{P}(\theta^* \in \mathcal{C}^\circ(z)) \) of the set \( \mathcal{C}^\circ(z) \). A particular issue is whether the coverage probability of the set \( \mathcal{C}^\circ(z_\alpha) \) from (2.11) matches its posterior mass \( 1 - \alpha \). The result of Theorem 2.3 effectively means that the elliptic set \( \mathcal{C}^\circ(z_\alpha) \) is centered at the random point \( \hat{\theta} = \theta^* + D_0^{-1} \xi \) with \( \xi \overset{\text{def}}{=} D_0^{-1} \nabla L(\theta^*) \) and the definition (2.11) yields

\[
\mathbb{P}(\theta^* \notin \mathcal{C}^\circ(z_\alpha)) = \mathbb{P}(\|\xi\|^2 > z_\alpha).
\]

In the classical asymptotic setup with a correctly specified parametric model, it holds \( V_0^2 \overset{\text{def}}{=} \text{Var}(\nabla L(\theta^*)) = D_0^2 \) and \( \xi = D_0^{-1} \nabla L(\theta^*) \) is nearly standard normal. This implies that \( \mathbb{P}(\|\xi\|^2 > z_\alpha) \approx \alpha \) and \( \mathcal{C}^\circ(z_\alpha) \) is indeed an asymptotic \( 1 - \alpha \) confidence set. Moreover, Spokoiny and Zhilova (2013) argued that even for finite samples, the deviation probability \( \mathbb{P}(\|\xi\|^2 > z_\alpha) \) of the quadratic form \( \|\xi\|^2 \) behaves very similarly to the \( \chi^2 \)-deviation probability. This justifies the use of \( \mathcal{C}(z_\alpha) \) as a frequentist confidence set, at least for large samples. However, it is well recognized that the \( \chi^2 \) approximation for the tails of \( \|\xi\|^2 \) is not very accurate. The use of a confidence set \( \mathcal{C}^\circ(z) \) with a data dependent choice of the threshold \( z \) ensuring \( \mathbb{P}(\|\xi\|^2 > z) \approx \alpha \) is highly recommended. One can summarize as follows: the asymptotic construction and asymptotic results for Bayesian credible sets and frequentist confidence sets are similar. However, the nature of these results is very different. The Gaussian approximation of the posterior is based only on the smoothness properties of the log-likelihood function and it is stated conditionally on \( \xi \), while Gaussian approximation of the log-likelihood involves a central limit theorem for \( \xi \) and a \( \chi^2 \) approximation for \( \|\xi\|^2 \). As a result, in finite samples, one should be very careful with the use of Bayesian credible sets as frequentist confidence sets.

The situation is even worse if the parametric model can be misspecified. This can lead to two different problems. First, the matrices \( D_0^2 \) and \( V_0^2 \) can be different. Then the covariance matrix of the vector \( \xi \) follows the famous “sandwich” formula \( \text{Cov}(\xi) = D_0^{-1} V_0^2 D_0^{-1} \); see e.g. Kleijn and van der Vaart (2012) for a detailed historical overview of this issue in Bayesian context. The deviation probability \( \mathbb{P}(\|\xi\|^2 > z) \) is also driven by this matrix \( D_0^{-1} V_0^2 D_0^{-1} \), and it is nearly the probability \( \mathbb{P}(\|D_0^{-1} V_0 \gamma\|^2 > z) \) with a standard normal vector \( \gamma \sim \mathcal{N}(0, I_p) \); see Spokoiny and Zhilova (2013). Thus, the coverage probability of the set \( \mathcal{C}(z) \) is determined by the relation between two Gaussian
probabilities $\mathbb{P}(\|D_0^{-1}V_0\gamma\|^2 > \delta)$ and $\mathbb{P}(\|\gamma\|^2 > \delta)$. Kleijn and van der Vaart (2012) presented some examples showing that the coverage probability can be as larger as smaller than $1 - \alpha$ depending on the relation between $D_0^2$ and $V_0^2$.

Another issue related to model misspecification is a possible modeling bias. Suppose that the set $\Theta$ is a subset of a larger set $\Theta^*$ and the true parameter $\theta^*$ belongs to $\Theta^*$ but not necessarily to $\Theta$. A typical example is given by a finite dimensional sieve approximation of $\Theta^*$ by $\Theta$. Then the value $\theta^*$ defined as the best $\Theta$-fit via the optimization problem $\theta^* = \arg\max_{\theta} \mathbb{E}L(\theta)$ does not coincide with $\theta^*_0$. The modeling bias can naturally be measured by $\mathbb{E}_{\theta^*_0} \log(d\mathbb{P}_{\theta^*_0}/d\mathbb{P}_{\theta^*})$. If this value significantly exceeds the dimension $p$, the bias-variance trade-off is destroyed and the true parameter would lie outside the Bayesian credible set with a high probability. Some examples can be found in Cox (1993), Freedman (1999), Bontemps (2011), Knapik et al. (2011) among others.

2.5 Extension to a continuous prior

The previous results for a non-informative prior can be extended to the case of a general prior $\Pi(d\theta)$ with a density $\pi(\theta)$ which is uniformly continuous on the local set $\Theta_0(x_0)$. More precisely, let $\pi(\theta)$ satisfy

$$\sup_{\theta \in \Theta_0(x_0)} \left| \frac{\pi(\theta)}{\pi(\theta^*)} - 1 \right| \leq \alpha, \quad \sup_{\theta \in \Theta} \frac{\pi(\theta)}{\pi(\theta^*)} \leq C,$$

(2.13)

where $\alpha$ is a small constant while $C$ is any fixed constant. The second condition in (2.13) is not restrictive but it implicitly assumes that the true point $\theta^*$ belongs to the level set of the prior density. Then the results of Theorem B.1 through 2.3 continue to apply with an obvious correction of the error $\Delta^*_\epsilon$.

As an example, consider the case of a Gaussian prior $\Pi = N(0, G^{-2})$ with the density $\pi(\theta) \propto \exp\{-\|G\theta\|^2/2\}$. 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? The answer is given by the next theorem. For the ease of formulation, we fix $a^2 = 1$.

**Theorem 2.5.** Suppose the conditions of Theorem 2.1 with $a^2 = 1$. Let also $\Pi = N(0, G^{-2})$ be a Gaussian prior measure on $\mathbb{R}^p$ such that

$$\|G\theta^*\| \leq C, \quad \|D_0^{-1}G^2D_0^{-1}\|_\infty p_n \to 0,$$

(2.14)

as $n \to \infty$. Then all the results of Theorem 2.3 continue to hold.

Note that the conditions (2.13) effectively mean that the prior $\Pi$ does not significantly affect the posterior distribution. Similar conditions and results can be found in the literature for more specific models, see e.g. Bontemps (2011) for the Gaussian case.
2.6 Applications to the i.i.d. case

This section comments how the previously obtained general results can be linked to the classical asymptotic results in the statistical literature. More specifically, we briefly discuss how the BvM result can be applied to an i.i.d. experiment. Let \( Y = (Y_1, \ldots, Y_n) \) 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 \( (P_\theta) \) to be fulfilled. We admit that the parametric assumption \( P \in (P_\theta, \theta \in \Theta) \) can be misspecified and consider the asymptotic setup with \( n \) growing to infinity and simultaneously \( p = p_n \) growing to infinity. The bracketing bound and the large deviation result apply if the sample size \( n \) fulfills \( n \geq C p_n \) for a fixed constant \( C \). It appears that the BvM result requires a stronger condition. Indeed, in the regular i.i.d. case it holds \( \delta(r_0) \asymp r_0/\sqrt{n} \) and \( \omega(r_0) \asymp r_0/\sqrt{n} \). The radius \( r_0 \) should fulfill \( r_0^2 \geq C p_n \) to ensure the large deviation result. This yields

\[
e \geq \delta(r_0) + 3\eta_0 a^2 \omega(r_0) \geq C \sqrt{p_n/n}.
\]

Here we admit a model misspecification but assume that \( a^2 \) is an absolute constant independent of \( n \). The BvM requires the condition \( \epsilon p_n \) is small”, which effectively means that \( p_n^3/n \to 0 \) as \( n \to \infty \).

**Theorem 2.6.** Suppose the conditions of Theorem 5.1 in Spokoiny (2012). Let also \( p_n^3/n \to 0 \). Then the result of Theorem 2.3 holds with \( D_0^2 = n F_{\theta^*} \), where \( F_{\theta^*} \) is the Fisher information of \( (P_\theta) \) at \( \theta^* \).

Another constraint on the dimension growth \( p \) can be found in Ghosal (1999) for linear models; see the condition (2.6) \( p^{3/2}(\log p)^{1/2} \eta_n \to 0 \) there, in which \( \eta_n \) is of order \( (p/n)^{-1/2} \) in regular situations yielding a suboptimal constraint \( n^{-1}p^4 \log p \to 0 \). A forthcoming paper Panov and Spokoiny (2013) presents an example illustrating that the condition \( p_n^3/n \to 0 \) cannot be dropped or relaxed.

A Conditions

Below we collect the list of conditions which are systematically used in the text. The list is essentially the same as in Spokoiny (2012). The conditions are quite general and seem to be non-restrictive; see the discussion at the end of the section. The whole list can be split into local and global conditions.

A.1 Local conditions

Local conditions describe the properties of \( L(\theta) \) in a vicinity of the central point \( \theta^* \) from (1.2) defined as maximizer of the expected log-likelihood \( E L(\theta) \). Degree of locality
is determined by the value $r_0$. Our results require $r_0^2 \geq C q^*$ with $q^* \overset{\text{def}}{=} p + E \| \xi \|^2$.

The bracketing result (2.5) requires second order smoothness of the expected log-likelihood $E L(\theta)$. By definition, $L(\theta^*, \theta^*) \equiv 0$ and $\nabla E L(\theta^*) = 0$ because $\theta^*$ is the extreme point of $E L(\theta)$. Therefore, $-E L(\theta, \theta^*)$ can be approximated by a quadratic function of $\theta - \theta^*$ in the neighborhood of $\theta^*$. The next condition quantifies this quadratic approximation from above and from below.

\[ (L_0) \quad \text{There are a symmetric strictly positive-definite matrix } D_0^2 \text{ and for each } r \leq r_0 \text{ and a constant } \delta(r) \leq 1/2, \text{ such that it holds for each } \theta \text{ with } \| D_0(\theta - \theta^*) \| \leq r \]

\[ \left| -\frac{2E L(\theta, \theta^*)}{\| D_0(\theta - \theta^*) \|^2} - 1 \right| \leq \delta(r). \]

Usually $D_0^2$ is defined as the negative Hessian of $E L(\theta)$ at $\theta = \theta^*$: $D_0^2 = -\nabla^2 E L(\theta^*)$. If $L(\theta, \theta^*)$ is the log-likelihood ratio and $F = F_{\theta^*}$ then $-E L(\theta, \theta^*) = E_{\theta^*} \log(d F_{\theta^*}/d F_{\theta}) = K(F_{\theta^*}, F_{\theta})$, the Kullback-Leibler divergence between $F_{\theta^*}$ and $F_{\theta}$. Then condition $(L_0)$ follows from the usual regularity conditions on the family $(F_{\theta})$; cf. Ibragimov and Khas’minskij (1981). In the important special case of an i.i.d. model one can take $\delta(r) = \delta^* r/n^{1/2}$ for some constant $\delta^*$; see Section 5.1 of Spokoiny (2012).

The further conditions concern the stochastic component $\zeta(\theta)$ of $L(\theta)$:

\[ \zeta(\theta) \overset{\text{def}}{=} L(\theta) - E L(\theta). \]

Below we suppose that the random function $\zeta(\theta)$ is differentiable in $\theta$ and its gradient $\nabla \zeta(\theta) = \partial \zeta(\theta)/\partial \theta \in \mathbb{R}^q$ has some exponential moments. Our next condition describes the property of the gradient $\nabla \zeta(\theta^*)$ at the central point $\theta^*$.

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

\[ \sup_{\gamma \in \mathbb{R}^p} \log \mathbb{E} \exp \left\{ \frac{\lambda \nabla \zeta(\theta^*)}{\| V_0 \gamma \|} \right\} \leq \nu_0^2 \lambda^2 / 2. \]

The matrix $V_0^2$ shown in this condition determines the local variability of the stochastic component of $L(\theta)$. In typical situations, the matrix $V_0^2$ can be defined as the covariance matrix of the gradient vector $\nabla \zeta(\theta^*)$: $V_0^2 = \operatorname{Var}(\nabla \zeta(\theta^*)) = \operatorname{Var}(\nabla L(\theta^*))$. If $L(\theta)$ is the log-likelihood for a correctly specified model, then $\theta^*$ is the true parameter value and $V_0^2$ coincides with the total Fisher information matrix $D_0^2$. In the i.i.d. case it is equal to $n F$ where $n$ is the sample size while $F$ is the Fisher matrix of one observation. In the general situations the matrices $V_0^2$ and $D_0^2$ may be different. We only assume the identifiability condition which relates these matrices.
There is a constant \( a > 0 \) such that \( a^2 D_0^2 \geq V_0^2 \).

The constant \( a^2 \) here is important and it enters in the definition of \( \epsilon \) via the condition \( \epsilon \geq \delta(r_0) + 3\nu_0^2 a^2 \omega(r_0) \); see Theorem 2.1. We do not show this dependence explicitly to keep the notation short. In the asymptotic setup it may depend on the asymptotic parameter \( n \).

The final local condition concerns the increments of the gradient \( \nabla \zeta(\theta) \) in the local elliptic neighborhoods of \( \theta^* \) defined as \( \Theta_0^2(r) \defeq \{ \theta \in \Theta : \|D_0(\theta - \theta^*)\| \leq r \} \).

For each \( r \leq r_0 \), there exists a constant \( \omega(r) \leq 1/2 \) such that it holds for all \( \theta \in \Theta_0(r) \)

\[
\sup_{\gamma \in \mathbb{R}^p} \log \mathbb{E} \exp \left\{ \frac{\lambda \gamma^T \{ \nabla \zeta(\theta) - \nabla \zeta(\theta^*) \}}{\omega(\gamma)} \right\} \leq \nu_0^2 \lambda^2 / 2, \quad |\lambda| \leq g.
\]

A.2 Global conditions

The global conditions have to be fulfilled for all \( \theta \) lying beyond \( \Theta_0(r_0) \). We only impose one condition on the smoothness of the stochastic component of the process \( L(\theta) \) in terms of its gradient, and one identifiability condition in terms of the expectation \( \mathbb{E} L(\theta, \theta^*) \).

The first condition is similar to the local condition \( ED_0 \) and it requires some exponential moment of the gradient \( \nabla \zeta(\theta) \) for all \( \theta \in \Theta \). However, the constant \( g \) may be dependent of the radius \( r = \|D_0(\theta - \theta^*)\| \).

For any \( r \), there exists a value \( g(r) > 0 \) such that for all \( \lambda \leq g(r) \)

\[
\sup_{\theta \in \Theta_0(r)} \sup_{\gamma \in \mathbb{R}^p} \log \mathbb{E} \exp \left\{ \frac{\lambda \gamma^T \nabla \zeta(\theta)}{\|V_0\gamma\|} \right\} \leq \nu_0^2 \lambda^2 / 2.
\]

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

There is a function \( b(r) \) such that \( rb(r) \) monotonously increases in \( r \) and for each \( r \geq r_0 \)

\[
\inf_{\theta : \|D_0(\theta - \theta^*)\| = r} \frac{\mathbb{E} L(\theta, \theta^*)}{\|D_0(\theta - \theta^*)\|} \geq b(r) r^2.
\]

In our results it is assumed that \( b(r) \) is a constant, however, the condition in stated in the same form as in Spokoiny (2012). The results can be extended to the situation
when the function \( b(r) \to 0 \) as \( r \to \infty \) but \( rb(r) \to \infty \). Section 5.3 of Spokoiny (2012) presents an example of median regression where the use of \( b(r) \) is helpful.

The conditions involve some constants. We distinguish between important constants and technical ones. The impact of the important constants is shown in our results, the list includes \( \delta(r) \), \( \omega(r) \), \( r_0 \), \( a \). The other constants are technical. The constant \( \nu_0 \) is introduced for convenience only; it can be omitted by rescaling the matrix \( V_0 \). In the asymptotic setup it can usually be selected very close to one. The constant \( g \) should be large enough, our results require \( g^2 \geq Cq^* = (1 + a^2)p \). This is, however, not restrictive. For instance, in the i.i.d. case, \( g^2 \) is of order \( n \), see Section 5.1 in Spokoiny (2012). The same applies to \( g(r) \) from \((Er)\). For ease of presentation, we assume that \( b(r) \) from \((Lr)\) is bounded from below by an absolute constant.

We briefly comment how restrictive the imposed conditions are. Spokoiny (2012), Section 5.1, considered in details the i.i.d. case and presented some mild sufficient conditions on the parametric family which imply the above general conditions. Another class of examples is built by generalized linear models which includes the cases of Gaussian, Poissonian, binary, regression and exponential type models among others. Condition \((ED_0)\) requires some exponential moments of the observations (errors). Usually one only assumes some finite moments of the errors; cf. Ibragimov and Khas’minskij (1981), Chapter 2. Our condition is a bit more restrictive but it allows to obtain some finite sample bounds. The linearity of the log-likelihood w.r.t. the parameter automatically yields condition \((ED_1)\). Condition \((L_0)\) only requires some regularity of the considered parametric family and is not restrictive. Conditions \((Er)\) with \( g(r) \equiv g > 0 \) and \((Lr)\) with \( b(r) \equiv b > 0 \) are easy to verify if the parameter set \( \Theta \) is compact and the sample size \( n \) is sufficiently large. It suffices to check a usual identifiability condition that the value \( \mathbb{I} \mathbb{E} L(\theta, \theta^*) \) does not vanish for \( \theta \neq \theta^* \).

The regression and generalized regression models are included as well; cf. Ghosal (1999, 2000) or Kim (2006). If one uses a Gaussian likelihood functions, the regression errors have to fulfill some exponential moments conditions. If this condition is too restrictive and a more stable (robust) estimation procedure is desirable, one can apply the LAD-type contrast leading to median regression. Spokoiny (2012), Section 5.3, showed for the case of linear median regression that all the required conditions are fulfilled automatically if the sample size \( n \) exceeds \( Cp \) for a fixed constant \( C \). Spokoiny et al. (2013) applied this approach for local polynomial quantile regression. Burnaev et al. (2013) applied the approach to the problem of regression with Gaussian process where the unknown parameters enter in the likelihood in a rather complicated way. We conclude that the imposed conditions are quite general and can be verified for the majority of examples met in the literature on BvM.
B Some auxiliary results

This section collects some auxiliary results about the behavior of the posterior measures which might be of independent interest.

B.1 Local Gaussian approximation of the posterior. Upper bound

As the first step, we study $D_\epsilon(\vartheta - \theta_\epsilon)$, where

$$\theta_\epsilon = \theta^* + D_\epsilon^{-1} \xi_\epsilon = \theta^* + D_\epsilon^{-2} \nabla L(\theta^*)$$

and $\xi_\epsilon = D_\epsilon^{-1} \nabla L(\theta^*)$. For any nonnegative function $f$, the bracketing bound (2.5) yields:

$$\int_{\Theta_0(r_0)} \exp \{ L(\theta, \theta^*) \} f(D_\epsilon(\theta - \theta_\epsilon)) d\theta$$

$$\leq e^\Delta \epsilon \int_{\Theta_0(x_0)} \exp \{ L_\epsilon(\theta, \theta^*) \} f(D_\epsilon(\theta - \theta_\epsilon)) d\theta.$$

Similarly, with $\theta_\epsilon = \theta^* + D_\epsilon^{-1} \xi_\epsilon$

$$\int_{\Theta_0(x_0)} \exp \{ L(\theta, \theta^*) \} f(D_\epsilon(\theta - \theta_\epsilon)) d\theta$$

$$\geq e^{-\Delta \epsilon} \int_{\Theta_0(x_0)} \exp \{ L_\epsilon(\theta, \theta^*) \} f(D_\epsilon(\theta - \theta_\epsilon)) d\theta.$$

The main benefit of these bounds is that both $L_\epsilon(\theta, \theta^*)$ and $L_\epsilon(\theta, \theta^*)$ are quadratic in $\theta$. This enables to explicitly evaluate the posterior and to show that the posterior measure is nearly Gaussian. In what follows $\gamma$ is a standard normal vector in $\mathbb{R}^p$.

Theorem B.1. Suppose the conditions of Theorem 2.1. Then for any nonnegative function $f(\cdot)$ on $\mathbb{R}^p$, it holds on $\Omega_n$

$$\mathbb{E} \left[ f(D_\epsilon(\theta - \theta_\epsilon)) \mathbb{I}\{ \theta \in \Theta_0(x_0) \} \mid Y \right] \leq \exp \{ \Delta_\epsilon^+(x_0) \} \mathbb{E} f(\gamma), \quad (B.1)$$

where

$$\Delta_\epsilon^+(x_0) = \Delta_\epsilon(x_0) + \frac{p}{2} \log \left( \frac{1 + \epsilon}{1 - \epsilon} \right) + \nu(x_0),$$

with $\Delta_\epsilon(x_0)$ from (2.8), and for $\gamma \sim N(0, I_p)$

$$\nu(x_0) \overset{\text{def}}{=} - \log \mathbb{P}(\|\gamma + \xi\| \leq r_0 \mid Y). \quad (B.2)$$

Furthermore, on $\Omega_n$, it holds $\Delta_\epsilon^+(x_0) \leq C \epsilon q \leq C \epsilon q^*$ with $q = p + \|\xi\|^2$ and $q^* = \mathbb{E} q$. 
The condition $\epsilon q^* \to 0$ allows us to ignore the exp-factor in (B.1) and this result yields an upper bound $\mathbb{E}_f(\gamma)$ for the posterior expectation of $f(D_\epsilon(\vartheta - \theta_\epsilon))$ conditioned on $Y$ and on $\vartheta \in \Theta_0(r_0)$. The lower bound requires some additional results on the tails of the posterior.

It is convenient to introduce local conditional expectation: for a random variable $\eta$, define

$$\mathbb{I}_\varnothing^\circ \eta \overset{\text{def}}{=} \mathbb{E} \left[ \eta \mathbb{I} \{ \vartheta \in \Theta_0(r_0) \} \mid Y \right].$$

The bound (B.1) reads as

$$\mathbb{I}_\varnothing^\circ f(\lambda^\top D_\epsilon(\vartheta - \theta_\epsilon)) \leq \exp\{ \Delta_\epsilon^+(r_0) \} \mu_f(\gamma).$$

The next result considers some special cases with $f(u) = \exp(\lambda^\top u)$ and $f(u) = \mathbb{I}(u \in A)$ for a measurable subset $A \subset \mathbb{R}^p$.

**Corollary B.2.** For any $\lambda \in \mathbb{R}^p$

$$\log \mathbb{I}_\varnothing^\circ \exp\{ \lambda^\top D_\epsilon(\vartheta - \theta_\epsilon) \} \leq \| \lambda \|^2/2 + \Delta_\epsilon^+(r_0), \quad (B.3)$$

Moreover, if $\| \lambda \|^2 \leq p$, then it holds for $\hat{\theta} \overset{\text{def}}{=} \theta^* + D_0^{-2} \nabla L(\theta^*)$

$$\log \mathbb{I}_\varnothing^\circ \exp\{ \lambda^\top D_0(\vartheta - \hat{\theta}) \} \leq \| \lambda \|^2/2 + \Delta_\epsilon^\circ(r_0), \quad (B.4)$$

with

$$\Delta_\epsilon^\circ(r_0) \overset{\text{def}}{=} \Delta_\epsilon^+(r_0) + \epsilon p + 2\epsilon p^{1/2}\| \xi \|.$$

On $\Omega_n$ one obtains $\Delta_\epsilon^\circ(r_0) \leq C\epsilon q$.

Similarly to the previous result, the condition $\epsilon q^* \to 0$ ensures that the terms $\Delta_\epsilon^+(r_0)$ in (B.3) and $\Delta_\epsilon^\circ(r_0)$ in (B.4) are negligible.

The next corollary describes an upper bound for the posterior probability.

**Corollary B.3.** For any measurable set $A \subset \mathbb{R}^p$, it holds on $\Omega_n$ with $\delta_\epsilon \overset{\text{def}}{=} D_0(\hat{\theta} - \theta_\epsilon)$

$$\mathbb{P}^\circ(D_0(\vartheta - \hat{\theta}) \in A) \leq \exp\{ \Delta_\epsilon^+(r_0) \} \mathbb{P}(D_0D_\epsilon^{-1} \gamma + \delta_\epsilon \in A) \quad (B.5)$$

$$\leq \exp\{ \Delta_\epsilon^+(r_0) \} \{ \mathbb{P}(\gamma \in A) + C\epsilon q^{1/2} \}. \quad (B.6)$$

The next result describes the local concentration properties of the posterior. Namely, the centered and scaled posterior vector $\eta \overset{\text{def}}{=} D_\epsilon(\vartheta - \theta_\epsilon)$ concentrates on the coronary set $\{ u : \| u \|^2 - p \leq \sqrt{2p} \}$ with $\mathbb{P}^\circ$-probability of order $1 - 2e^{-x/4}$. 

Corollary B.4. Let $x \leq p/2$. Then

$$
\mathbb{P}^\circ \left( \| D_\varepsilon (\vartheta - \theta_\varepsilon) \|^2 - p > \sqrt{2px} \right) \leq \exp \left( -x/4 + \Delta^+_{\varepsilon} (r_0) \right), \quad (B.7)
$$

$$
\mathbb{P}^\circ \left( \| D_\varepsilon (\vartheta - \theta_\varepsilon) \|^2 - p > -\sqrt{2px} \right) \leq \exp \left( -x/2 + \Delta^+_{\varepsilon} (r_0) \right).
$$

B.2 Tail posterior probability and contraction

The next important step in our analysis is to check that $\vartheta$ concentrates in a small vicinity $\Theta_0 = \Theta_0 (r_0)$ of the central point $\theta^*$ with a properly selected $r_0$. The concentration properties of the posterior will be described by using the random quantity

$$
\rho (r_0) \overset{\text{def}}{=} \frac{\int_{\Theta_0} \exp \{ L(\theta, \theta^*) \} d\theta}{\int_{\Theta_0} \exp \{ L(\theta, \theta^*) \} d\theta}.
$$

Obviously $\mathbb{P} \{ \vartheta \notin \Theta_0 (r_0) \ | \ Y \} \leq \rho (r_0)$. Therefore, small values of $\rho (r_0)$ indicate a small posterior probability of the set $\Theta \setminus \Theta_0$.

**Theorem B.5.** Suppose the conditions of Theorem 2.1. Then it holds on $\Omega_n$

$$
\rho (r_0) \leq \exp \left\{ \Delta_2 (r_0) + \nu (r_0) \right\} \left( \frac{1 + \varepsilon}{b} \right)^{p/2} \mathbb{P} \left( \| \gamma \|^2 \geq br_0^2 \right),
$$

with $\nu (r_0)$ from (B.2). Similarly, for each $m \geq 0$

$$
\rho_m (r_0) \overset{\text{def}}{=} \mathbb{E} \left[ \| D_\varepsilon (\vartheta - \vartheta_\varepsilon) \|^m \mathbb{I} \{ \vartheta \notin \Theta_0 (r_0) \} \right] \leq \exp \left\{ \Delta_2 (r_0) + \nu (r_0) \right\} \left( \frac{1 + \varepsilon}{b} \right)^{p/2} \mathbb{E} \left[ \| \gamma \|^m \mathbb{I} \{ \| \gamma \|^2 \geq br_0^2 \} \right].
$$

This result yields simple sufficient conditions on the value $r_0$ which ensures the concentration of the posterior on $\Theta_0 (r_0)$.

**Corollary B.6.** Assume the conditions of Theorem B.5. Then the inequality $br_0^2 \geq C_p$ ensures

$$
\rho_m (r_0) \leq C e^{-x_n}, \quad m = 0, 1, 2.
$$

B.3 Local Gaussian approximation of the posterior. Lower bound

Now we present a local lower bound for the posterior probability. The reason for separating the upper and lower bounds is that the lower bound also requires a tail probability estimation; see (B.9).

**Theorem B.7.** Suppose the conditions of Theorem 2.1. Then for any nonnegative function $f(\cdot)$ on $\mathbb{R}^p$, it holds on $\Omega_n$

$$
\mathbb{E}^0 f (D_\varepsilon (\vartheta - \theta_\varepsilon)) \geq \exp \left\{ -\Delta^-_{\varepsilon} (r_0) \right\} \mathbb{E} \left\{ f (\gamma) \mathbb{I} \{ \| \gamma \| \leq C r_0 \} \right\}, \quad (B.10)
$$
where
\[ \Delta_\epsilon^{-}(r_0) = \Delta_\epsilon(r_0) + \frac{p}{2} \log \left( \frac{1 + \epsilon}{1 - \epsilon} \right) + \rho(r_0). \]

On \( \Omega_n \) one obtains \( \Delta_\epsilon^{-}(r_0) \leq C \epsilon q \).

As a corollary, we state the results for the distribution and moment generating functions of \( D_\epsilon(\theta - \theta_e) \). Here we assume \( x_n \leq p \). We also need an additional condition that the \( r_0^2 \geq C q^* \) for a fixed constant \( C \).

**Corollary B.8.** Let \( r_0^2 \geq C q^* \). For any \( \lambda \in \mathbb{R}^p \) with \( \| \lambda \|^2 \leq p \)

\[ \log \mathbb{E}^o \exp \{ \lambda^\top D_\epsilon(\theta - \theta_e) \} \geq \| \lambda \|^2 / 2 - \Delta_\epsilon^o(r_0) - e^{-x_n}. \]

Moreover, if \( \| \lambda \|^2 \leq p \), then

\[ \log \mathbb{E}^o \exp \{ \lambda^\top D_0(\theta - \tilde{\theta}) \} \geq \| \lambda \|^2 / 2 - \Delta_\epsilon^o(r_0), \]

where the error term \( \Delta_\epsilon^o(r_0) \) satisfy on \( \Omega_n \) the bound \( \Delta_\epsilon^o(r_0) \leq C \epsilon q \).

For any \( A \subset \mathbb{R}^p \), it holds on \( \Omega_n \) with \( \delta_e = D_0(\tilde{\theta} - \theta_e) \)

\[ \mathbb{P}^o(D_0(\theta - \tilde{\theta}) \in A) \geq \exp \{ \Delta_\epsilon^{-}(r_0) \} \mathbb{P}(D_0 D_\epsilon^{-1} \gamma + \delta_e \in \mathbb{A}) - e^{-x_n} \]
\[ \geq \exp \{ \Delta_\epsilon^{-}(r_0) \} \{ \mathbb{P}(\gamma \in \mathbb{A}) - C \epsilon q^{1/2} \} - e^{-x_n}. \]

### C Proofs

Here we collect the proofs of the main results.

**C.1 Proof of Theorem B.1**

We use that \( L_\epsilon(\theta, \theta^*) = \xi_e^\top D_\epsilon(\theta - \theta^*) - \| D_\epsilon(\theta - \theta^*) \|^2 / 2 \) is proportional to the density of a Gaussian distribution and similarly for \( L_\epsilon(\theta, \theta^*) \). More precisely, define

\[ m_\epsilon(\xi_e) \overset{\text{def}}{=} -\| \xi_e \|^2 / 2 + \log(\det D_\epsilon) - p \log(\sqrt{2\pi}). \tag{C.1} \]

Then

\[ m_\epsilon(\xi_e) + L_\epsilon(\theta, \theta^*) \]
\[ = -\| D_\epsilon(\theta - \theta_e) \|^2 / 2 + \log(\det D_\epsilon) - p \log(\sqrt{2\pi}) \tag{C.2} \]

is (conditionally on \( Y \)) the log-density of the normal law with the mean \( \theta_e = D_\epsilon^{-1} \xi_e + \theta^* \) and the covariance matrix \( D_\epsilon^{-2} \). Change of variables \( u = D_\epsilon(\theta - \theta_e) \) implies by (C.2)
for any nonnegative function $f$ that

$$\int_{\Theta_0} \exp\{L(\theta, \theta^*) + m_\epsilon(\xi_\epsilon)\} f(D_\epsilon(\theta - \theta_\epsilon)) d\theta \leq e^{\hat{\phi}_\epsilon(x_0)} \int \exp\{\mathbb{L}_\epsilon(\theta, \theta^*) + m_\epsilon(\xi_\epsilon)\} f(D_\epsilon(\theta - \theta_\epsilon)) d\theta$$

$$= e^{\hat{\phi}_\epsilon(x_0)} \int \phi(u) f(u) du$$

$$= e^{\hat{\phi}_\epsilon(x_0)} \mathbb{E} f(\gamma). \quad (C.3)$$

Similarly, if $m_\epsilon(\xi_\epsilon)$ is defined by (C.1) with $\epsilon$ in place of $\epsilon$, then the value $m_\epsilon(\xi_\epsilon) + \mathbb{L}_\epsilon(\theta, \theta^*)$ is (conditionally on $Y$) the density of the normal law with the mean $\theta_\epsilon = D_\epsilon^{-1}\xi_\epsilon + \theta^*$ and the covariance matrix $D_\epsilon^{-2}$. For any nonnegative function $f$, it follows by change of variables $u = D_\epsilon(\theta - \theta_\epsilon)$ and $D_0(\theta - \theta^*) = D_0D_\epsilon^{-1}(u + \xi_\epsilon)$ that

$$\int \exp\{L(\theta, \theta^*)\} f(D_\epsilon(\theta - \theta_\epsilon)) \mathbb{P}\{\|D_0(\theta - \theta^*)\| \leq x_0\} d\theta \geq \exp\{-\hat{\phi}_\epsilon(x_0) - m_\epsilon(\xi_\epsilon)\} \int \phi(u) f(u) \mathbb{P}\{\|D_0D_\epsilon^{-1}(u + \xi_\epsilon)\| \leq x_0\} du. \quad (C.4)$$

Further, by construction, $D_\epsilon^2 \geq D_0^2$ and $\|\xi_\epsilon\| \leq \|\xi\|$, yielding

$$\{D_\epsilon^{-1}(u + \xi_\epsilon) \in \Theta_0(x_0)\} = \{\|D_0D_\epsilon^{-1}(u + \xi_\epsilon)\| \leq x_0\} \supset \{\|u + \xi\| \leq x_0\}.$$ 

Hence, a special case of (C.4) with $f(u) \equiv 1$ implies by definition of $\nu(x_0)$:

$$\int_{\Theta_0(x_0)} \exp\{L(\theta, \theta^*)\} d\theta \geq \exp\{-\hat{\phi}_\epsilon(x_0) - m_\epsilon(\xi_\epsilon) - \nu(x_0)\}. \quad (C.5)$$

Now we are prepared to finalize the proof of the theorem. (C.3) and (C.5) imply

$$\int_{\Theta_0(x_0)} \exp\{L(\theta, \theta^*)\} f(D_\epsilon(\theta - \theta_\epsilon)) d\theta \leq \exp\{\hat{\phi}_\epsilon(x_0) + \hat{\phi}_\epsilon(x_0) + m_\epsilon(\xi_\epsilon) - m_\epsilon(\xi_\epsilon) + \nu(x_0)\} \mathbb{E} f(\gamma).$$

The definition of $m_\epsilon(\xi_\epsilon)$, $m_\epsilon(\xi_\epsilon)$, $D_\epsilon$, $D_\epsilon$, and $\Delta_\epsilon(x_0)$ imply

$$\hat{\phi}_\epsilon(x_0) + \hat{\phi}_\epsilon(x_0) + m_\epsilon(\xi_\epsilon) - m_\epsilon(\xi_\epsilon) = \Delta_\epsilon(x_0) + \frac{p}{2} \log\left(\frac{1 + \epsilon}{1 - \epsilon}\right)$$

and (B.1) follows. As $\|\xi\|^2 \leq C \alpha^2 p$ on $\Omega_n$ and $x_0^2 \geq C(1 + \alpha^2)p$, this implies for $\gamma \sim N(0, I_p)$ and $x_n \leq C p$

$$\nu(x_0) = - \log \mathbb{P}\{\|\gamma + \xi\| \leq x_0 \mid Y\} \leq - \log \mathbb{P}\{\|\gamma\|^2 \leq p + C x_n\} \leq e^{-x_n}. $$
C.2 Proof of Corollary B.2

The first fact is a direct implication of (B.1). For the second one we use that

\[ \exp\{ \lambda^\top D_0(\theta - \hat{\theta}) \} = \exp\{ \lambda_1^\top D_\epsilon(\theta - \theta_\epsilon) \} \exp\{ \lambda_2^\top D_0(\theta_\epsilon - \hat{\theta}) \} \]  
(C.6)

with \( \lambda_1 = (1 - \epsilon)^{-1/2} \lambda \). Furthermore, \( \|\lambda\|^2 \leq p \), hence

\[ \|\lambda_1\|^2 = \lambda_1^\top D_\epsilon^{-1} D_0^2 D_\epsilon^{-1} \lambda \leq (1 - \epsilon)^{-1} \|\lambda\|^2 \leq (1 + 2\epsilon) \|\lambda\|^2 \leq \|\lambda\|^2 + 2\epsilon p. \]

It remains to bound \( \lambda^\top D_0(\theta_\epsilon - \hat{\theta}) \) for \( \hat{\theta} = \theta^* + D_0^{-2} \nabla L(\theta^*) = \theta^* + D_0^{-1} \xi \) and \( \theta_\epsilon = \theta^* + D_\epsilon^{-1} \xi_\epsilon \). It holds by definition of \( D_\epsilon \) in view of \( \epsilon \leq 1/2 \)

\[ \|D_0(\theta_\epsilon - \hat{\theta})\| = \|D_0 D_\epsilon^{-1} \xi_\epsilon - \xi\| \]
\[ = \|(D_0 D_\epsilon^{-2} D_0 - I_p)\xi\| \leq \left( \frac{1}{1 - \epsilon} - 1 \right) \|\xi\| \leq 2\epsilon \|\xi\|. \]  
(C.7)

This yields the result by (B.3) and (C.6) in view of \( \|\lambda\|^2 \leq p \).

C.3 Proof of Corollary B.3

The first statement (B.5) follows from Theorem B.1 with \( f(u) = \Pi(D_0 D_\epsilon^{-1} \gamma + \delta_\epsilon \in A) \). Further, by (C.7), it holds on \( \Omega_n \) for \( \delta_\epsilon \overset{\text{def}}{=} D_0(\hat{\theta} - \theta_\epsilon) \)

\[ \|\delta_\epsilon\| \leq 2\epsilon \|\xi\|. \]  
(C.8)

For proving (B.6), we compute the Kullback–Leibler divergence between two multivariate normal distributions and apply Pinsker’s inequality. Let \( \gamma \) be standard normal in \( \mathbb{R}^p \), and \( \Pi_0 \) stand for its distribution. The random variable \( D_0 D_\epsilon^{-1} \gamma + \delta_\epsilon \) is normal with mean \( \delta_\epsilon \) and variance \( B_\epsilon^{-2} \overset{\text{def}}{=} D_0 D_\epsilon^{-2} D_0 \). Denote this distribution by \( \Pi_\epsilon \). Obviously \( B_\epsilon = D_0^{-1} D_\epsilon^2 D_0^{-1} = (1 - \epsilon) I_p \) and the definition of \( \epsilon \) implies \( \|B_\epsilon - I_p\|_\infty \leq \epsilon \). We use the following technical lemma.

**Lemma C.1.** Let \( \|B_\epsilon - I_p\|_\infty \leq \epsilon \leq 1/2 \). Then

\[ 2\mathcal{K}(\Pi_0, \Pi_\epsilon) = -2E_0 \log \frac{d\Pi_\epsilon}{d\Pi_0}(\gamma) \]
\[ \leq \text{tr}(B_\epsilon - I_p)^2 + (1 + \epsilon)\|\delta_\epsilon\|^2 = \epsilon^2 p + (1 + \epsilon)\|\delta_\epsilon\|^2. \]

**Proof.** It holds

\[ 2 \log \frac{d\Pi_\epsilon}{d\Pi_0}(\gamma) = -\log \det(B_\epsilon) - (\gamma - \delta_\epsilon)^\top B_\epsilon(\gamma - \delta_\epsilon) + \|\gamma\|^2 \]
and

\[ 2\mathcal{K}(\mathbb{P}_0, \mathbb{P}_\epsilon) = -2\mathbb{E}_0 \log \frac{d\mathbb{P}_\epsilon}{d\mathbb{P}_0}(\gamma) \]

\[ = \log \det(B_\epsilon) + \text{tr}(B_\epsilon - I_p) + \delta_\epsilon^\top B_\epsilon \delta_\epsilon. \]

Denote by \( a_j \) the \( j \)th eigenvalue of \( B_\epsilon - I_p \). Then \( \|B_\epsilon - I_p\|_\infty \leq \epsilon \leq 1/2 \) yields \(|a_j| \leq 1/2 \) and

\[ 2\mathcal{K}(\mathbb{P}_0, \mathbb{P}_\epsilon) = \delta_\epsilon^\top B_\epsilon \delta_\epsilon + \sum_{j=1}^p \left\{ a_j - \log(1 + a_j) \right\} \]

\[ \leq (1 + \epsilon)\|\delta_\epsilon\|^2 + \sum_{j=1}^p a_j^2 \]

\[ \leq (1 + \epsilon)\|\delta_\epsilon\|^2 + \text{tr}(B_\epsilon - I_p)^2 \leq (1 + \epsilon)\|\delta_\epsilon\|^2 + \epsilon^2 p. \]

as required. \( \square \)

This lemma and (C.8) imply by Pinsker’s inequality

\[ \|\mathbb{P}_0 - \mathbb{P}_\epsilon\|_{TV} \leq \sqrt{\frac{1}{2} \mathcal{K}(\mathbb{P}_0, \mathbb{P}_\epsilon)} \leq C \epsilon \sqrt{p + \|\xi\|^2} \leq C \epsilon q^{1/2}. \]

Equivalently, for any measurable set \( A \), it holds

\[ \mathbb{P}(D_0 D_\epsilon^{-1} \gamma + \delta_\epsilon \in A \mid Y) \leq \mathbb{P}(\gamma \in A) + C \epsilon q^{1/2}. \]

C.4 Proof of Corollary B.4

Denote \( \eta_\epsilon \overset{\text{def}}{=} D_\epsilon(\vartheta - \theta_\epsilon) \). The exponential Chebyshev inequality implies for any positive \( \lambda \leq 1/2 \) in view of \(- \log(1 - \lambda) \leq \lambda + \lambda^2\)

\[ \mathbb{P}^0(\|\eta_\epsilon\|^2 > p + \sqrt{2px}) \leq \exp \left\{ -\frac{\lambda}{2} \left( p + \sqrt{2px} \right) \right\} \mathbb{E}^0 \exp \left( \frac{\lambda}{2} \|\eta_\epsilon\|^2 \right) \]

\[ \leq \exp \left\{ -\frac{\lambda}{2} \left( p + \sqrt{2px} \right) + \Delta_\epsilon^+(x_0) \right\} \mathbb{E} \exp \left( \frac{\lambda}{2} \|\gamma\|^2 \right) \]

\[ = \exp \left\{ -\frac{\lambda}{2} \left( p + \sqrt{2px} \right) - \frac{p}{2} \log(1 - \lambda) + \Delta_\epsilon^+(x_0) \right\} \]

\[ \leq \exp \left\{ -\lambda \sqrt{px/2} + p \lambda^2/2 + \Delta_\epsilon^+(x_0) \right\} \]
The choice $\lambda = \sqrt{\frac{x}{2p}}$ yields the result (B.7). Similarly

$$P\left(\|\eta_\epsilon\|^2 - p < -\sqrt{2px}\right) \leq \exp\left\{-\frac{\lambda}{2}(\sqrt{2px} - p)\right\}\mathbb{E}\exp\left\{-\frac{\lambda}{2}\|\eta_\epsilon\|^2\right\}$$

$$\leq \exp\left\{-\frac{\lambda}{2}(\sqrt{2px} - p) + \Delta^+_\epsilon(r_0)\right\}\mathbb{E}\exp\left\{-\frac{\gamma}{2}\|\gamma\|^2\right\}$$

$$= \exp\left\{-\frac{\lambda}{2}(\sqrt{2px} - p) + \frac{p}{2}\log(1 + \lambda) + \Delta^+_\epsilon(r_0)\right\}$$

$$\leq \exp\left\{-\lambda\sqrt{px/2} + p\lambda^2/4 + \Delta^+_\epsilon(r_0)\right\}\mathbb{E}\exp\left\{\frac{\gamma}{2}\|\gamma\|^2\right\}.$$ 

Here the choice $\lambda = \sqrt{2x/p}$ does the job.

C.5 Proof of Theorem B.5

Define $u(\theta) = b\|D_0(\theta - \theta^*)\|^2/2$. Then it holds

$$-\mathbb{E}L(\theta, \theta^*) - u(\theta) \geq b\|D_0(\theta - \theta^*)\|^2/2.$$ 

Now we apply Corollary 4.2 of Spokoiny (2012) in which $b(\cdot)$ is replaced by $b/2$. This result ensures that $u(\theta)$ is an upper function for $L(\theta, \theta^*)$. Furthermore, by a change of variables, one obtains

$$\frac{b^{p/2} \det(D_0)}{(2\pi)^{p/2}} \int_{\Theta \cap \Theta_0} \exp\{-u(\theta)\} d\theta$$

$$\leq \frac{b^{p/2} \det(D_0)}{(2\pi)^{p/2}} \int_{\Theta \cap \Theta_0} \exp\{-b\|D_0(\theta - \theta^*)\|^2/2\} d\theta = P(\|\gamma\|^2 \geq br_0^2).$$ 

For the integral in the nominator of (B.8), it holds on $\Omega_n$ by (2.4) for any nonnegative function $f(\cdot)$

$$\int_{\Theta \cap \Theta_0} \exp\{L(\theta, \theta^*)\} f(\theta) d\theta \leq \int_{\Theta \cap \Theta_0} \exp\{-u(\theta)\} f(\theta) d\theta.$$ 

(C.9)

The bound (C.5) for the local integral $\int_{\Theta_0} \exp\{L(\theta, \theta^*)\} d\theta$ implies that

$$\rho(r_0) \leq \exp\{\Delta^+_\epsilon(r_0) + \nu(r_0) + m_\epsilon(\xi_\epsilon)\} \int_{\Theta \cap \Theta_0} \exp\{-u(\theta)\} d\theta.$$ 

Finally

$$\exp\{m_\epsilon(\xi_\epsilon)\} = \exp\{-\|\xi_\epsilon\|^2/2\} (2\pi)^{-p/2} \det(D_\epsilon) \leq (2\pi)^{-p/2} \det(D_0)$$

and the assertion follows by $\det(D_\epsilon) = (1 + \epsilon)^{p/2} \det(D_0)$. 
C.6 Proof of Theorem B.7

On the set $\Omega_n$, it holds by (C.3) with $f(\cdot) \equiv 1$, (2.4) and (C.9):

$$\int \exp\{L(\theta, \theta^*)\} d\theta \leq \int_{\Theta} \exp\{L(\theta, \theta^*)\} d\theta + \int_{\Theta^c} \exp\{L(\theta, \theta^*)\} d\theta$$

$$\leq \{1 + \rho(x_0)\} \int_{\Theta} \exp\{L(\theta, \theta^*)\} d\theta$$

$$\leq \{1 + \rho(x_0)\} \exp\{\delta(x_0) - m_\epsilon(\xi_\epsilon) + \nu(x_0)\}$$

$$\leq \exp\{\delta(x_0) - m_\epsilon(\xi_\epsilon) + \nu(x_0) + \rho(x_0)\}.$$

This and the bound (C.4) imply

$$\int_{\Theta} \exp\{L(\theta, \theta^*)\} f(D_{\xi_\epsilon}(\theta - \theta_\epsilon)) d\theta$$

$$\geq \frac{\exp\{-\delta(x_0) - m_\epsilon(\xi_\epsilon)\} \int \phi(u)f(u) \mathbb{I}\{u \in B_\epsilon(x_0)\} du}{\exp\{\delta(x_0) - m_\epsilon(\xi_\epsilon) + \nu(x_0) + \rho(x_0)\}}$$

$$\geq \exp\{-\Delta_\epsilon(x_0)\} \mathbb{E}[f(\gamma) \mathbb{I}\{\gamma \in B_\epsilon(x_0)\}]$$

where

$$B_\epsilon(x_0) \overset{\text{def}}{=} \{u \in \mathbb{R}^p : \|D_0 D_{\epsilon}(u + \xi_\epsilon)\| \leq x_0\}.$$ 

The bounds $\|\xi_\epsilon\|^2 = (1 + \epsilon)^{-1}\|\xi\|^2$, $D_{\xi_\epsilon}^2 = (1 + \epsilon)D_0^2$, imply for $x_0^2 \geq c q$

$$B_\epsilon(x_0) \supseteq \{u \in \mathbb{R}^p : \|u\|^2 \leq c p\}.$$

This yields (B.10).

C.7 Proof of Corollary B.8

The first result follows from Theorem B.7. The only important additional step is an evaluation of the integral $\mathbb{E}\{\exp(\lambda^T \gamma) \mathbb{I}(\|\gamma\| \leq r)\}$.

Lemma C.2. Let $\gamma \sim N(0, I_p)$, $\mu \in (0, 1)$. Then for any vector $\lambda \in \mathbb{R}^p$ with $\|\lambda\|^2 \leq p$

$$\log \mathbb{E}\{\exp(\lambda^T \gamma) \mathbb{I}(\|\gamma\| > r)\} \leq -\frac{1 - \mu}{2} r^2 + \frac{1}{2\mu} \|\lambda\|^2 + \frac{p}{2} \log(\mu^{-1}).$$

(C.10)

Proof. We use that for $\mu < 1$

$$\mathbb{E}\{\exp(\lambda^T \gamma) \mathbb{I}(\|\gamma\| > r)\} \leq e^{-(1-\mu) r^2/2} \mathbb{E}\{\exp(\lambda^T \gamma + (1 - \mu)\|\gamma\|^2/2)\}.$$
It holds
\[
\mathbb{E} \exp \{ \lambda^\top \gamma + (1 - \mu)\|\gamma\|^2/2 \} = (2\pi)^{-p/2} \int \exp \{ \lambda^\top \gamma - \mu\|\gamma\|^2/2 \} d\gamma \\
= \mu^{-p/2} \exp(\mu^{-1}\|\lambda\|^2/2)
\]
and (C.10) follows.

Now we apply this result with \( \mu = 1/2 \).

**Lemma C.3.** Let \( \gamma \sim \mathcal{N}(0, I_p) \), and let \( r^2 \geq 4(p + x) \). Then for any \( \lambda \in \mathbb{R}^p \) with \( \|\lambda\|^2 \leq p \)
\[
\mathbb{E} \left\{ \exp(\lambda^\top \gamma) \mathbb{I}(\|\gamma\| \leq r) \right\} \geq \exp(-\|\lambda\|^2/2)(1 - e^{-x}).
\]
(C.11)

**Proof.** The result (C.10) applied with \( \mu = 1/2 \) yields in view of \( \mathbb{E} \exp(\lambda^\top \gamma) = \exp(\|\lambda\|^2/2), \|\lambda\|^2 \leq p \), and \( 1 + \log(2) < 2 \) that
\[
e^{-\|\lambda\|^2/2} \mathbb{E} \left\{ \exp(\lambda^\top \gamma) \mathbb{I}(\|\gamma\| \leq r) \right\} \\
\geq 1 - \exp(-r^2/4 + p/2 + (p/2)\log(2)) \geq 1 - \exp(-x)
\]
and (C.11) follows.

The bound (C.11) yields the first assertion of Corollary B.8 in view of \( \log(1 - e^{-3x/2}) \geq -e^{-x} \) for \( x \geq 1 \). The second statement can be proved similarly to Corollary B.2, while the last statement is similar to Corollary B.3.

**C.8 Proof of Theorem 2.3**

The results are entirely based on our obtained statements from Corollaries B.2, B.3, and B.8. Below \( \Delta_*^\epsilon \equiv \max\{\Delta_{\epsilon}^\oplus, \Delta_{\epsilon}^\ominus\} \). As previously, we assume that \( p = p_n \to \infty \) but \( \epsilon p_n \to 0 \). We already know that \( \Delta_*^\epsilon \leq C_\epsilon p_n \) on a set of a dominating probability. Due to our previous results, it is convenient to decompose the r.v. \( \vartheta \) in the form
\[
\vartheta = \vartheta \mathbb{I}\{\vartheta \in \Theta_0(r_0)\} + \vartheta \mathbb{I}\{\vartheta \notin \Theta_0(r_0)\} = \vartheta^\oplus + \vartheta^\ominus.
\]
The large deviation results yields that the posterior distribution of the part \( \vartheta^\ominus \) is negligible provided a proper choice of \( r_0 \). Below we show that \( \vartheta^\oplus \) is nearly normal which yields the BvM result. Define
\[
\mathcal{S}_0^2 \equiv \text{Cov}^\oplus(\vartheta^\oplus) \equiv \mathbb{E}^\oplus\left\{ (\vartheta - \mathbb{E}^\ominus \vartheta)(\vartheta - \mathbb{E}^\ominus \vartheta)^\top \right\}.
\]
Define also the first two moments of \( \eta \) \( \overset{\text{def}}{=} D_0(\vartheta - \tilde{\theta}) \):

\[
\overline{\eta} \overset{\text{def}}{=} \mathbb{E} \eta, \quad S^2_0 \overset{\text{def}}{=} \mathbb{E} \{(\eta - \overline{\eta})(\eta - \overline{\eta})^\top\} = D_0 \mathbb{S}^2_0 D_0.
\]

Similarly to the proof of Corollaries B.2 and B.8 one derives for any unit vector \( u \in \mathbb{R}^p \)

\[
\exp(-\Delta) \leq \mathbb{E} |u^\top \eta|^2 \leq \exp(\Delta),
\]
with \( \Delta^- = \Delta^-_e \) and \( \Delta^+ = \Delta^+_e \). It suffices to show that (C.12) implies

\[
\| \overline{\eta} \|^2 \leq c \Delta^*, \quad \| S^2_0 - I_p \|_{\infty} \leq c \Delta^*
\]
with \( \Delta^* = \max\{ \Delta^+, \Delta^- \} \leq 1/2 \). Note now that

\[
\mathbb{E} |u^\top \eta|^2 = u^\top S^2_0 u + |u^\top \overline{\eta}|^2.
\]

Hence

\[
\exp(-\Delta) \leq u^\top S^2_0 u + |u^\top \overline{\eta}|^2 \leq \exp(\Delta^+).
\]

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

\[
u^\top S^2_0 u = \mathbb{E} |u^\top (\eta - \overline{\eta})|^2 \geq \exp(-\Delta^-) \mathbb{E} |u^\top (\gamma - \overline{\eta})|^2 = \exp(-\Delta^-)(1 + \| \overline{\eta} \|^2)
\]

yielding

\[
u^\top S^2_0 u \geq (1 + \| \overline{\eta} \|^2) \exp(-\Delta^-).
\]

This inequality contradicts (C.14) if \( \| \overline{\eta} \|^2 > 2 \Delta^* \) for \( \Delta^* \leq 1 \), and (C.13) follows.

The bound for the first moment implies

\[
\| D_0(\mathbb{E} \vartheta - \tilde{\theta}) \|^2 \leq c \Delta^*
\]
while the second bound yields \( \| D_0 \mathbb{S}^2_0 D_0 - I_p \|_{\infty} \leq c \Delta^* \). This completes the proof of (2.10).

### C.9 Proof of Corollary 2.4

Due to Lemma C.1, the bounds of the probability \( \mathbb{P}(D_0(\vartheta - \tilde{\theta}) \in A | Y) \) from Theorem 2.3 continue to apply if we replace the true Fisher matrix \( D^2_0 \) by any other matrix.
credible sets for growing parameter dimension

\( \hat{D}^2 \) satisfying (2.12). Hence, we only need to check this condition for \( \hat{D}^2 = D^2(\overline{\theta}) \). Condition \((L_0)\) yields that \( \|D_0^{-1}D^2(\theta)D_0^{-1} - I_p\|_\infty \leq \delta(r_0) \) for any \( \theta \in \Theta_0(r_0) \). Further, \( \overline{\theta} \in \Theta_0(r_0) \) on the set \( \Omega_n \). This implies

\[
tr(D_0^{-1}D^2(\overline{\theta})D_0^{-1} - I_p)^2 \leq C\delta^2(r_0)p \leq C\epsilon^2 p,
\]

and the result follows.

C.10 Proof of Theorem 2.5

First evaluate the ratio \( \pi(\theta)/\pi(\theta^*) \) for any \( \theta \in \Theta_0(r_0) \). It holds

\[
\log \frac{\pi(\theta)}{\pi(\theta^*)} = -\|G\theta\|^2/2 + \|G\theta^*\|^2/2 = (\theta - \theta^*)^\top G^2\theta^* - \|G(\theta - \theta^*)\|^2/2,
\]

It is obvious from the definition of \( \Theta_0(r_0) \) that for \( \theta \in \Theta_0(r_0) \)

\[
\|G(\theta - \theta^*)\|^2 = \|GD_0^{-1}D_0(\theta - \theta^*)\|^2 \leq \|D_0^{-1}G^2D_0^{-1}\|_\infty r_0^2.
\]

If we take \( r_0^2 = C p_n \), then by (2.14)

\[
\|G(\theta - \theta^*)\|^2 \leq \|D_0^{-1}G^2D_0^{-1}\|_\infty C p_n \to 0.
\]

Similarly

\[
|G(\theta - \theta^*)^\top G^2\theta^*| \leq \|G\theta^*\| \cdot \|G(\theta - \theta^*)\| \leq \|G\theta^*\| \cdot \|GD_0^{-1}\|_\infty r_0 \to 0.
\]

The result is proved.

References

Bickel, P. J. and Kleijn, B. J. K. (2012). The semiparametric Bernstein-von Mises theorem. *Ann. Statist.*, 40(1):206–237.

Bontemps, D. (2011). Bernstein-von Mises theorems for Gaussian regression with increasing number of regressors. *Ann. Statist.*, 39(5):2557–2584.

Boucheron, S. and Gassiat, E. (2009). A Bernstein-von Mises theorem for discrete probability distributions. *Electron. J. Stat.*, 3:114–148.

Bunke, O. and Milhaud, X. (1998). Asymptotic behavior of Bayes estimates under possibly incorrect models. *Ann. Statist.*, 26(2):617–644.

Burnaev, E., Spokoiny, V., and Zaytsev, A. (2013). Finite sample properties for gaussian fied regression. *Avtomatika i Telemekhanika*. To appear in russian.
Castillo, I. (2012). A semiparametric Bernstein-von Mises theorem for Gaussian process priors. *Probability Theory and Related Fields*, 152:53–99. 10.1007/s00440-010-0316-5.

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

Cox, D. D. (1993). An analysis of Bayesian inference for nonparametric regression. *Ann. Stat.*, 21(2):903–923.

Freedman, D. (1999). On the Bernstein-von Mises theorem with infinite-dimensional parameters. *Ann. Stat.*, 27(4):1119–1140.

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

Ghosal, S. (2000). Asymptotic normality of posterior distributions for exponential families when the number of parameters tends to infinity. *J. Multivariate Anal.*, 74(1):49–68.

Ghosal, S., Ghosh, J. K., and van der Vaart, A. W. (2000). Convergence rates of posterior distributions. *Ann. Statist.*, 28(2):500–531.

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

Kim, Y. (2006). The Bernstein-von Mises theorem for the proportional hazard model. *Ann. Statist.*, 34(4):1678–1700.

Kleijn, B. J. K. and van der Vaart, A. W. (2006). Misspecification in infinite-dimensional Bayesian statistics. *Ann. Statist.*, 34(2):837–877.

Kleijn, B. J. K. and van der Vaart, A. W. (2012). The Bernstein-von-Mises theorem under misspecification. *Electronic J. Statist.*, 6:354–381.

Knapik, B. T., van der Vaart, A. W., and van Zanten, J. H. (2011). Bayesian inverse problems with Gaussian priors. *Ann. Statist.*, 39(5):2626–2657.

Panov, M. and Spokoiny, V. (2013). Critical dimension in semiparametric Bernstein-von Mises Theorem. Manuscript.
Shen, X. (2002). Asymptotic normality of semiparametric and nonparametric posterior distributions. *J. Amer. Statist. Assoc.*, 97(457):222–235.

Spokoiny, V. (2012). Parametric estimation. Finite sample theory. *Ann. Statist.*, 40(6):2877–2909. arXiv:1111.3029.

Spokoiny, V., Wang, W., and Härdle, W. (2013). Local quantile regression (with rejoinder). *J. of Statistical Planing and Inference*, 143(7):1109–1129. ArXiv:1208.5384.

Spokoiny, V. and Zhilova, M. (2013). Sharp deviation bounds for quadratic forms. *Mathematical Methods of Statistics*, 22(2):100–113. arXiv:1302.1699; doi:10.3103/S1066530713020026.