Accuracy of Gaussian approximation in nonparametric Bernstein – von Mises Theorem[*]

Vladimir Spokoiny[†]

Weierstrass Institute and HU Berlin,
HSE University, Skoltech and IITP RAS,
Mohrenstr. 39, 10117 Berlin, Germany
`spokoiny@wias-berlin.de`

Maxim Panov

Skolkovo Institute
of Science and Technology
121205 Moscow, Russia,
`m.panov@skoltech.ru`

June 2, 2020

Abstract

The prominent Bernstein – von Mises (BvM) result claims that the posterior distribution after centering by the efficient estimator and standardizing by the square root of the total Fisher information is nearly standard normal. In particular, the prior completely washes out from the asymptotic posterior distribution. This fact is fundamental and justifies the Bayes approach from the frequentist viewpoint. In the nonparametric setup the situation changes dramatically and the impact of prior becomes essential even for the contraction of the posterior; see van der Vaart and van Zanten (2008), Bontemps (2011), Castillo and Nickl (2013, 2014) for different models like Gaussian regression or i.i.d. model in different weak topologies. This paper offers another non-asymptotic approach to studying the behavior of the posterior for a special but rather popular and useful class of statistical models and for Gaussian priors. First we derive tight finite sample bounds on posterior contraction in terms of the so-called effective dimension of the parameter space. Our main results describe the accuracy of Gaussian approximation of the posterior. In particular, we show that restricting to the class of all centrally symmetric credible sets around the penalized maximum likelihood estimator (pMLE) allows to get Gaussian approximation up to order n^{-1}. We also show that the posterior distribution mimics well the distribution of the pMLE and reduce the question of reliability of credible sets to

1

consistency of the pMLE-based confidence sets. The obtained results are specified for nonparametric log-density estimation and generalized regression.

Contents

1   Introduction

Bernstein – von Mises (BvM) Theorem is one of most prominent results in statistical inference. It 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 Bayesian 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 main questions to address by studying the behavior of a nonparametric Bayes procedure are

- concentration: find possibly small concentration sets of the posterior distribution;

- asymptotic normality or any other asymptotic approximation of the posterior;

- covering: whether one can use credible sets as frequentist confidence sets.

The classical versions of the BvM Theorem claim that the posterior concentrates on a root-n vicinity of the true parameter, after proper centering and scaling it is root-n standard normal, and credible sets can be well used in place of classical confidence sets. However, these results require a fixed finite dimensional parameter set, correct model specification, and large samples. We refer to van der Vaart and Wellner (1996); van der Vaart (1998) for a detailed historical overview.

Any extension of the BvM approach to the case of a large or infinite dimensional parameter space appears to be very involved, in particular, more involved than the expansions of the maximum likelihood estimate. The first problem is related to the posterior concentration. Such a result requires to bound the integral of the likelihood process in the complement of the local vicinity and this is a hard task in the nonparametric setup. The second problem is due to fact that a standard Gaussian measure on $I\!R^\infty$ is only defined in a weak sense. In particular, it does not concentrate on any ℓ_2 ball in $I\!R^\infty$. This makes it difficult to study asymptotically the total variation distance between the scaled posterior and the Gaussian law. We refer to Castillo and Nickl (2013, 2014), and Ghosal and van der Vaart (2017) for a more discussion. Our approach can be called *preasymptotic*: we fix the sample and study the distance between the posterior and an accompanying Gaussian distribution. Similar approach was recently used in Yano and Kato (2020) for a special ℓ_∞ topology and it applies to the case of a moderate or high parameter dimension p with the rate $\left(\log p\right)^a n^{-1/6}$. In our approach the parameter dimension can be arbitrary, however, the so called effective dimension has to be relatively small.

Restricting to centrally symmetric credible sets allows to get the accuracy of order n^{-1} for Gaussian approximation. One more crucial issue is an inconsistency problem: in some situations, Bayesian credible sets do not contain the true parameter with the probability close to one; cf. Cox (1993); Freedman (1999), or Kleijn and van der Vaart (2006, 2012). It appears that the posterior in the case of a Gaussian prior is nearly normal but its first two moments mimic the *penalized maximum likelihood estimator* (pMLE) with the quadratic penalization coming from the prior distribution. It is well known that the penalization yields some estimation bias. If the squared bias exceeds the variance of the penalized MLE, the Bayesian credible sets become unreliable.

The main results of this paper describe the properties of the posterior distribution for Gaussian priors in a high-dimensional or nonparametric setups. In particular, we establish a nonasymptotic upper bound on concentration and on the error of Gaussian approximation for the posterior in total variation distance in terms of efficient dimension of the problem. We also show that the latter bound can be dramatically improved when restricting to the class of centrally symmetric sets around pMLE. Our assumptions include two important conditions. The first one requires that the stochastic part of the log-likelihood is linear in the target parameter, while the second one is about concavity of the expected log-likelihood. These two conditions are automatically fulfilled in a number of popular models like Gaussian, Poissonian, Binary or Generalized Linear regression, log-density estimation, linear diffusion, etc. Under these assumptions we manage to state and prove our results in a concise way and avoid the machinery of the empirical process theory. A forthcoming paper Spokoiny (2019) explains how the approach and the results can be extended to much more general setups including nonlinear (generalized) regression and nonlinear inverse problems with noisy observations. The main contributions of the paper are *finite samples* results with *accuracy guarantees* including

- sharp bounds on concentration of pMLE and of the posterior distribution;
- Gaussian approximation of the posterior with an explicit error term for the total variation distance and for the class of centrally symmetric sets around pMLE;
- systematic use of an *effective dimension* in place of the total parameter dimension;
- addressing frequentist validity of Bayesian credible sets;
- specification of the results to log-density estimation and generalized regression.

The whole approach is *coordinate free*, we do not use any spectral decomposition and/or any basis representation for the target parameter and penalization. In this paper we suppose the prior to be given and do not address the question of prior selection. A

number of studies explain how an empirical or hierarchical Bayes approach can be used for building adaptive confidence sets; see e.g. Knapik et al. (2016); Nickl and Szabó (2016); Sniekers and van der Vaart (2015). We, however, indicate below how our approach can be used to reduce the original problem of Bayes model selection to the well studied Gaussian case using uniform Gaussian approximation; see Section 2.5.3.

The paper is structured as follows. Section 2 describes our setup, presents the main conditions and states the main results about the properties of the pMLE and of the posterior. Section 2.1 collects our conditions and main notations. The central notion of effective dimension is discussed in Section 2.2. Properties of the pMLE $\widetilde{\boldsymbol{\theta}}_G$ are described in Section 2.3. Section 2.4 presents our main results about Gaussian approximation of the posterior; see Theorem 2.9 and its Corollary 2.10. We also address the issues of contraction and coverage of Bayesian credible sets. Section 3.1 comments how the result can be applied to the case of the Bayesian nonparametric log-density estimation, while Section 3.2 discusses generalized regression estimation. Some more results are given in the Supplement. Section 2.5 presents some extensions including the use of posterior mean in place of the pMLE in the construction of credible sets, or the use of a general prior with a log-concave density in place of a Gaussian one. It also addresses the important question of prior impact. The proofs and auxiliary results are collected in the Appendix.

2   Nonparametric BvM Theorem

This section discusses an extension of the BvM result for a a class of models with a high-dimensional or infinite dimensional parameter set and for a Gaussian prior. Compared to existing literature, our results provide finite sample bounds on posterior concentration and on accuracy of Gaussian approximation for the posterior. Moreover, we show that the quality of Gaussian approximation can be gradually improved up to order n^{-1} if we only consider credible sets which are centrally symmetric around the pMLE.

Below $\mathbb{R}^p$ means a p-dimensional Euclidean space equipped with the norm $\| \cdot \|$, $p \leq \infty$. Scalar product in $\mathbb{R}^p$ is denoted by $\langle \cdot, \cdot \rangle$. For a linear operator B in $\mathbb{R}^p$, the norm $\|B\|$ means the largest eigenvalue of B.

First we specify our setup. Let $\boldsymbol{Y}$ denote the observed data and $\mathbb{P}$ mean their distribution. A general parametric assumption (PA) means that $\mathbb{P}$ belongs to infinite-dimensional family $(\mathbb{P}_{\boldsymbol{\theta}}, \boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^p)$ dominated by a measure $\boldsymbol{\mu}_0$. This family yields the log-likelihood function $L(\boldsymbol{\theta}) = L(\boldsymbol{Y}, \boldsymbol{\theta}) \overset{\text{def}}{=} \log \frac{d\mathbb{P}_{\boldsymbol{\theta}}}{d\boldsymbol{\mu}_0}(\boldsymbol{Y})$. The PA can be misspecified, so, in general, $L(\boldsymbol{\theta})$ is a *quasi log-likelihood*. The classical maximum likelihood principle

suggests to estimate $\boldsymbol{\theta}$ by maximizing the function $L(\boldsymbol{\theta})$:

$$\widetilde{\boldsymbol{\theta}} \stackrel{\text{def}}{=} \operatorname*{argmax}_{\boldsymbol{\theta} \in \Theta} L(\boldsymbol{\theta}). \tag{2.1}$$

If $I\!\!P \notin (I\!\!P_{\boldsymbol{\theta}})$, then the estimate $\widetilde{\boldsymbol{\theta}}$ from (2.1) is still meaningful and it appears to be an estimate of the value $\boldsymbol{\theta}^*$ defined by maximizing the expected value of $L(\boldsymbol{\theta})$ w.r.t. $I\!\!P$:

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

Such a value $\boldsymbol{\theta}^*$ is the true parameter under correct model specification and it can be viewed as the parameter of the best parametric fit in the general case. In the Bayes setup, the parameter $\boldsymbol{\vartheta}$ is a random element following a prior measure Π on the parameter set Θ. The posterior describes the conditional distribution of $\boldsymbol{\vartheta}$ given $\boldsymbol{Y}$ obtained by normalization of the product $\exp\{L(\boldsymbol{\theta})\}\Pi(d\boldsymbol{\theta})$. This relation is usually written as

$$\boldsymbol{\vartheta} \mid \boldsymbol{Y} \propto \exp\{L(\boldsymbol{\theta})\}\,\Pi(d\boldsymbol{\theta}).$$

Below we focus on the case of a Gaussian prior. Without loss of generality, a Gaussian prior $\Pi(\boldsymbol{\theta})$ will be assumed to be centered at zero. By G^{-2} we denote its covariance matrix, so that, $\Pi \sim \mathcal{N}(0, G^{-2})$. The main question studied below is to understand under which conditions on the prior covariance G^{-2} and the model, the BvM-type result holds and what is the error term in the BvM approximation. For a Gaussian likelihood, the posterior is Gaussian as well and its properties can be studied directly; see e.g. Bontemps (2011); Leahu (2011). For the case when the log-likelihood function is not quadratic in $\boldsymbol{\theta}$, the study is more involved. The posterior is obtained by normalizing the product density $\exp\{L_G(\boldsymbol{\theta})\}$ with

$$L_G(\boldsymbol{\theta}) = L(\boldsymbol{\theta}) - \left\|G\boldsymbol{\theta}\right\|^2/2\,,$$

where $\|\cdot\|$ is the Euclidean norm in $I\!\!R^p$. This expression arises in penalized maximum likelihood estimation, one can treat the prior term $\left\|G\boldsymbol{\theta}\right\|^2/2$ as roughness penalty. Define

$$\widetilde{\boldsymbol{\theta}}_G = \operatorname*{argmax}_{\boldsymbol{\theta} \in \Theta} L_G(\boldsymbol{\theta}), \qquad \boldsymbol{\theta}_G^* = \operatorname*{argmax}_{\boldsymbol{\theta} \in \Theta} I\!\!E L_G(\boldsymbol{\theta}).$$

2.1   Conditions

This section collects the conditions which are systematically used in the text. We mainly require that the stochastic part of the log-likelihood process $L(\boldsymbol{\theta})$ is linear in $\boldsymbol{\theta}$, while its expectation is a smooth concave function of $\boldsymbol{\theta}$. We also implicitly assume that the

parameter set Θ is an open subset of $I\!\!R^p$ where p is typically equal to infinity. The model and complexity reduction will be done via the the prior structure in terms of the so called effective dimension.

($\mathcal{L}$) *The set Θ is open and convex in $I\!\!R^p$. The function $I\!\!E L(\boldsymbol{\theta})$ is concave in $\boldsymbol{\theta} \in \Theta$.*

(E) *The stochastic component $\zeta(\boldsymbol{\theta}) = L(\boldsymbol{\theta}) - I\!\!E L(\boldsymbol{\theta})$ of the process $L(\boldsymbol{\theta})$ is linear in $\boldsymbol{\theta}$. We denote by $\nabla \zeta \equiv \nabla \zeta(\boldsymbol{\theta})$ its gradient and by $V^2 = \mathrm{Var}(\nabla \zeta)$ its covariance.*

(EV) *There exist a positive self-adjoint operator H with $H^2 \geq V^2$, and constant $\nu_0 \geq 1$ such that $\nabla \zeta$ fulfills*

$$\sup_{\boldsymbol{u} \in I\!\!R^p} \log I\!\!E \exp\left\{ \lambda \frac{\langle \boldsymbol{u}, H^{-1} \nabla \zeta \rangle}{\|\boldsymbol{u}\|} \right\} \leq \frac{\nu_0^2 \lambda^2}{2}.$$

Condition **(EV)** basically requires that the normalized score $\boldsymbol{\xi} = H^{-1} \nabla \zeta$ is a sub-Gaussian random vector. One can relax this condition to finite exponential moments for $|\lambda| \leq \mathsf{g}$ with g sufficiently large; see (A.6) of Section A.3. In fact, **(EV)** is only used to establish the deviation bounds for quadratic forms of $\nabla \zeta$; see e.g. (2.11). One can directly operate with the quantiles of the corresponding distribution. In the finite dimensional case $p < \infty$, one can often take $H = V$; see Section 3 for more examples.

Apart the basic conditions **($\mathcal{L}$)**, **(E)**, **(EV)** we need some local properties of the expected log-likelihood $I\!\!E L(\boldsymbol{\theta})$. Let Θ° be a local set. It is required that this set contains the concentration set $\mathcal{A}_G(\mathbf{r}_G)$ of the estimate $\widetilde{\boldsymbol{\theta}}_G$; see Proposition 2.4 below. Define

$$\mathbb{F}(\boldsymbol{\theta}) \stackrel{\mathrm{def}}{=} -\nabla^2 I\!\!E L(\boldsymbol{\theta}),$$

$$\mathbb{F}_G(\boldsymbol{\theta}) \stackrel{\mathrm{def}}{=} -\nabla^2 I\!\!E L_G(\boldsymbol{\theta}) = -\nabla^2 I\!\!E L(\boldsymbol{\theta}) + G^2 = \mathbb{F}(\boldsymbol{\theta}) + G^2,$$

(HG) For all $\boldsymbol{\theta} \in \Theta^\circ$, it holds $H^2 \leq \mathbb{F}_G(\boldsymbol{\theta})$ and $H \mathbb{F}_G^{-1}(\boldsymbol{\theta}) H$ is a trace operator:

$$\mathsf{p}_G(\boldsymbol{\theta}) \stackrel{\mathrm{def}}{=} \mathrm{tr}\left\{ H \mathbb{F}_G^{-1}(\boldsymbol{\theta}) H \right\} < \infty.$$

Also we require that the function $I\!\!E L(\boldsymbol{\theta})$ is four times differentiable on Θ°. Define for each $\boldsymbol{\theta} \in \Theta^\circ$, and any $\boldsymbol{u} \in I\!\!R^p$, the directional Gateaux derivative

$$\delta_k(\boldsymbol{\theta}, \boldsymbol{u}) \stackrel{\mathrm{def}}{=} \frac{1}{k!} \frac{d^k}{dt^k} I\!\!E L(\boldsymbol{\theta} + t\boldsymbol{u}) \bigg|_{t=0}, \quad k = 3, 4. \tag{2.2}$$

Clearly $\delta_k(\boldsymbol{\theta}, \boldsymbol{u})$ is proportional to $\|\boldsymbol{u}\|^k$. Later we need a uniform bound on $\delta_k(\boldsymbol{\theta}, \boldsymbol{u})$.

$(\mathcal{L}_0)$ *It holds with H from (EV) for $k = 3, 4$ and some $\mathtt{r}$ sufficiently large*

$$\tau_{k,H}(\mathtt{r}) \stackrel{\text{def}}{=} \sup_{\boldsymbol{\theta} \in \Theta^\circ} \sup_{\|H\boldsymbol{u}\| \leq \mathtt{r}} \mathtt{r}^{-k} \delta_k(\boldsymbol{\theta}, \boldsymbol{u}) < \infty. \tag{2.3}$$

In what follows we consider the situation with n independent observations. The resulted expected log-likelihood $I\!\!EL(\boldsymbol{\theta})$ is of order n as well as its derivatives. Also we will assume that $\|H^{-2}\| \leq \mathtt{C}/n$. Then for some fixed $\mathtt{C}_{3,\delta}, \mathtt{C}_{4,\delta}$

$$\tau_{k,H}(\mathtt{r}) \leq \mathtt{C}_{k,\delta} \, \mathtt{r}^{-k} n \, (\mathtt{r} n^{-1/2})^k = \mathtt{C}_{k,\delta} \, n^{1-k/2}, \qquad k = 3, 4. \tag{2.4}$$

2.2   Effective dimension

This section discusses the central notion of *effective dimension*. Define with H^2 from (EV) and $\mathbb{F}_G = \mathbb{F}_G(\boldsymbol{\theta}_G^*)$

$$B_G \stackrel{\text{def}}{=} H \mathbb{F}_G^{-1} H, \quad \mathtt{p}_G \stackrel{\text{def}}{=} \operatorname{tr} B_G, \qquad \lambda_G \stackrel{\text{def}}{=} \|B_G\|.$$

Here $\|B\|$ means the operator norm or the maximal eigenvalue of B. The values $\operatorname{tr} B_G$ and λ_G are important because they enter in the definition of the upper quantile function $z(B_G, \mathtt{x})$ for $\|\mathbb{F}_G^{-1/2} \nabla \zeta\|$; see (2.11) below. In Spokoiny (2017) the quantity $\mathtt{p}_G$ was called the effective dimension. Below we consider the *local effective dimension* $\mathtt{p}_G(\boldsymbol{\theta})$ at a point $\boldsymbol{\theta} \in \Theta^\circ$ given by

$$\mathtt{p}_G(\boldsymbol{\theta}) \stackrel{\text{def}}{=} \operatorname{tr}\{H^2 \mathbb{F}_G^{-1}(\boldsymbol{\theta})\} = \operatorname{tr}\{H^2 (\mathbb{F}(\boldsymbol{\theta}) + G^2)^{-1}\}.$$

Our results involve two particular values:

$$\mathtt{p}_G \stackrel{\text{def}}{=} \mathtt{p}_G(\boldsymbol{\theta}_G^*) = \mathtt{p}_G, \qquad \widetilde{\mathtt{p}}_G \stackrel{\text{def}}{=} \mathtt{p}_G(\widetilde{\boldsymbol{\theta}}_G).$$

Usually all the values $\mathtt{p}_G$ and $\mathtt{p}_G(\boldsymbol{\theta})$ for $\boldsymbol{\theta}$ close to $\boldsymbol{\theta}_G^*$ are close to each other; see (2.18) below. If $G^2 \equiv 0$ and $H^2 \approx \mathbb{F}(\boldsymbol{\theta}^*)$, one obviously gets $\mathtt{p}_G(\boldsymbol{\theta}) \asymp p$.

In what follows we will consider two non-trivial examples of priors. A *truncation prior* assumes that a growing sequence of nested linear subspaces $\mathbb{V}_m$ of dimension m is given, and the prior distribution is supported to a m-dimensional subspace $\mathbb{V}_m$ of $I\!\!R^p$. This formally corresponds to a covariance operator G_m^{-2} with $G_m^{-2}(I - \varPi_{\mathbb{V}_m}) = 0$, where $\varPi_{\mathbb{V}}$ means the projector on $\mathbb{V}$.

Lemma 2.1. *Suppose that $\mathbb{V}_m$ is a linear subspace of $I\!\!R^p$ with $\dim(\mathbb{V}_m) = m$ and the prior $\mathcal{N}(0, G_m^{-2})$ is supported to $\mathbb{V}_m$. Then $\mathtt{p}_G(\boldsymbol{\theta}) \leq m$. Moreover, if*

$$\|G_m \boldsymbol{u}\|^2 \leq g_m^2 \|\boldsymbol{u}\|^2, \qquad \boldsymbol{u} \in \mathbb{V}_m,$$

and if $\mathbb{F}(\boldsymbol{\theta})$ *fulfills*

$$\mathtt{C}_{1,\mathbb{F}}\, n\|\boldsymbol{u}\|^2 \leq \boldsymbol{u}^\top \mathbb{F}(\boldsymbol{\theta})\boldsymbol{u} \leq \mathtt{C}_{2,\mathbb{F}}\, n\|\boldsymbol{u}\|^2, \qquad \boldsymbol{u} \in \mathbb{V}_m, \tag{2.5}$$

and $H^2 \geq \mathbb{F}(\boldsymbol{\theta})$, *then*

$$\frac{\mathtt{C}_{1,\mathbb{F}}\, n}{\mathtt{C}_{2,\mathbb{F}}\, n + g_m^2}\, m \leq \mathtt{p}_G(\boldsymbol{\theta}) \leq m. \tag{2.6}$$

If $g_m^2 \ll n$, *then* $\mathtt{p}_G(\boldsymbol{\theta}) \asymp m$.

A *smoothing prior* corresponds to the situation when $\|G\boldsymbol{u}\|^2$ becomes large for $\boldsymbol{u} \notin \mathbb{V}_m$ and m large. Usually one assumes that $\mathbb{V}_m$ is spanned by the eigenvectors of G^2 corresponding to its smallest eigenvalues $g_1^2 \leq g_2^2 \leq \ldots \leq g_m^2$. One can write this condition in the form

$$\begin{aligned}
\|G\boldsymbol{u}\|^2 &\leq g_m^2\|\boldsymbol{u}\|^2, \qquad \boldsymbol{u} \in \mathbb{V}_m, \\
\|G\boldsymbol{u}\|^2 &\geq g_m^2\|\boldsymbol{u}\|^2, \qquad \boldsymbol{u} \in \mathbb{V}_m^c,
\end{aligned} \tag{2.7}$$

where $\mathbb{V}_m^c$ is the orthogonal complement of $\mathbb{V}_m$. Further we assume that g_j^2 grow polynomially yielding for some $\mathtt{C}_{1,g}, \mathtt{C}_{2,g}$ and each J

$$\mathtt{C}_{1,g} \leq \frac{1}{J g_J^{-2}} \sum_{j \geq J} g_j^{-2} \leq \mathtt{C}_{2,g}. \tag{2.8}$$

A typical example is given by $g_j^2 = w j^{2s}$ for $s > 1/2$ and some window parameter w.

Lemma 2.2. *Let the matrices* $\mathbb{F}(\boldsymbol{\theta})$ *and* H^2 *satisfy* (2.5) *and let the prior precision matrix* G^2 *follow* (2.7) *and* (2.8). *Define* m *by* $g_m^2 \approx n$. *Then*

$$\mathtt{C}_3\, m \leq \mathtt{p}_G(\boldsymbol{\theta}) \leq \mathtt{C}_4\, m, \tag{2.9}$$

where $\mathtt{C}_3$ *and* $\mathtt{C}_4$ *only depend on* $\mathtt{C}_{1,\mathbb{F}}, \mathtt{C}_{2,\mathbb{F}}$, *and* $\mathtt{C}_{1,g}, \mathtt{C}_{2,g}$. *In particular, if* $g_j^2 = w j^{2s}$ *with* $s > 1/2$, *then the effective dimension can be obtained as* $\mathtt{p}_G(\boldsymbol{\theta}) \asymp (n/w)^{1/(2s)}$.

The next technical result explains how the prior can be linked to "smoothness" of the unknown vector $\boldsymbol{\theta}$.

Lemma 2.3. *Let* $\mathbb{F}(\boldsymbol{\theta})$ *and* H^2 *satisfy* (2.5) *for all* $\boldsymbol{\theta} \in \Theta$. *Suppose that* Θ *is a subset of a Sobolev ball* $\mathcal{B}_s(1)$

$$\mathcal{B}_s(1) \stackrel{\text{def}}{=} \Big\{ \boldsymbol{\theta} = (\theta_j)\colon \sum_{j \geq 1} j^{2s}\theta_j^2 \leq 1 \Big\}.$$

If $G^2 = \mathrm{diag}\big(g_j^2\big)$ with $g_j^2 = wj^{2s}$ and $s > 1/2$, then $\|G\boldsymbol{\theta}\|^2 \leq w$. Define w by the bias-variance trade-off

$$(n/w)^{1/(2s)} \asymp w \quad or \quad w^{2s+1} \asymp n.$$

Then $w \asymp n^{1/(2s+1)}$, $\mathtt{p}_G(\boldsymbol{\theta}) \lesssim n^{1/(2s+1)}$ and $\|G\boldsymbol{\theta}\|^2 \lesssim n^{1/(2s+1)}$ for all $\boldsymbol{\theta} \in \mathcal{B}_s(1)$.

Here and below "$a \lesssim b$" means $a \leq \mathtt{C}b$ with an absolute constant $\mathtt{C}$. In what follows we implicitly assume that each value $\mathtt{p}_G(\boldsymbol{\theta})$ is much smaller than the full dimension p which can be even infinite. Most of our results requires $\mathtt{p}_G(\boldsymbol{\theta}) \ll n^{1/3}$, that is, $s > 1$ in case of Lemma 2.3.

2.3   Properties of the pMLE $\widetilde{\boldsymbol{\theta}}_G$

This section briefly reviews some properties of the penalized MLE $\widetilde{\boldsymbol{\theta}}_G = \mathrm{argmax}\, L_G(\boldsymbol{\theta})$. Our results are based on conditions $(\mathcal{L})$, (E), (EV), (HG), and $(\mathcal{L}_0)$ even if not mentioned explicitly. In particular, we systematically use that the stochastic term in the log-likelihood only linearly depends on $\boldsymbol{\theta}$ and that the expected log-likelihood is concave in $\boldsymbol{\theta}$. We state two results, the first one claims a kind of local concentration of the penalized MLE $\widetilde{\boldsymbol{\theta}}_G$, while the second one describes some useful expansions for the estimator $\widetilde{\boldsymbol{\theta}}_G$ and for the fitted log-likelihood $L_G(\widetilde{\boldsymbol{\theta}}_G)$. The presented results substantially improve similar statements in Spokoiny (2017).

Remind that $\boldsymbol{\theta}_G^* = \mathrm{arginf}_{\boldsymbol{\theta}}\, \mathbb{E}L_G(\boldsymbol{\theta})$ and $D_G^2 = \mathbb{F}(\boldsymbol{\theta}_G^*) + G^2$. Below we show that the penalized MLE $\widetilde{\boldsymbol{\theta}}_G$ concentrates with a high probability on the elliptic set

$$\mathcal{A}_G(\mathtt{r}) \stackrel{\mathrm{def}}{=} \big\{\boldsymbol{\theta}\colon \|D_G(\boldsymbol{\theta} - \boldsymbol{\theta}_G^*)\| \leq \mathtt{r}\big\} \tag{2.10}$$

under a proper choice of $\mathtt{r}$.

As the stochastic component of $L_G(\boldsymbol{\theta})$ is linear in $\boldsymbol{\theta}$, the gradient $\nabla\zeta = \nabla\big\{L_G(\boldsymbol{\theta}) - \mathbb{E}L_G(\boldsymbol{\theta})\big\}$ does not depend on $\boldsymbol{\theta}$. Under condition (EV), there exists a random set $\Omega(\mathtt{x})$ with $\mathbb{P}\big(\Omega(\mathtt{x})\big) \geq 1 - \mathtt{C}e^{-\mathtt{x}}$ such that on this set $\big\|D_G^{-1}\nabla\zeta\big\| \leq z(B_G, \mathtt{x})$:

$$\big\|D_G^{-1}\nabla\zeta\big\| \leq z(B_G, \mathtt{x}) \quad \text{on } \Omega(\mathtt{x}) \text{ with } \mathbb{P}\big(\Omega(\mathtt{x})\big) \geq 1 - \mathtt{C}e^{-\mathtt{x}}, \tag{2.11}$$

where $B_G = H\, D_G^{-2}\, H$ and $z(B_G, \mathtt{x})$ is given by (A.10); see Theorem A.4 with $\boldsymbol{\xi} = \nabla\zeta$. One can use the simplified bound

$$z(B_G, \mathtt{x}) \leq \sqrt{\mathrm{tr}(B_G)} + \sqrt{2\mathtt{x}\,\|B_G\|}. \tag{2.12}$$

It is worth mentioning that this deviation bound is the only place where the stochastic nature of the log-likelihood $L(\boldsymbol{\theta})$ is accounted for. In the rest, we only use the condition $(\boldsymbol{E})$ about linearity the stochastic component $\zeta(\boldsymbol{\theta})$ in $\boldsymbol{\theta}$.

Our first result describes the concentration properties of the penalized MLE $\widetilde{\boldsymbol{\theta}}_G$.

Proposition 2.4. *Assume* (2.11). *Let also* $\mathtt{r}_G$ *be such that* Θ° *contain the set* $\mathcal{A}_G(\mathtt{r}_G) \stackrel{\text{def}}{=} \{\boldsymbol{\theta}\colon \|D_G(\boldsymbol{\theta} - \boldsymbol{\theta}_G^*)\| \le \mathtt{r}_G\}$, *and with* $\tau_{3,H} = \tau_{3,H}(\mathtt{r}_G)$ *from* (2.3), *it holds*

$$3\mathtt{r}_G\,\tau_{3,H} \le \rho \le 1/2, \qquad (1-\rho)\mathtt{r}_G \ge z(B_G, \mathtt{x}). \tag{2.13}$$

Then on $\Omega(\mathtt{x})$, *the estimate* $\widetilde{\boldsymbol{\theta}}_G$ *belongs to this set* $\mathcal{A}_G(\mathtt{r}_G)$ *as well, that is,*

$$\left\|D_G\big(\widetilde{\boldsymbol{\theta}}_G - \boldsymbol{\theta}_G^*\big)\right\| \le \mathtt{r}_G. \tag{2.14}$$

Remark 2.1. In words, (2.14) means that $\widetilde{\boldsymbol{\theta}}_G$ belongs with a high probability to the vicinity $\mathcal{A}_G(\mathtt{r}_G)$ from (2.10) with $\mathtt{r}_G \le 2z(B_G, \mathtt{x})$. Under (2.4) $\tau_{3,H} \lesssim n^{-1/2}$ while $z^2(B_G, \mathtt{x}) \asymp \mathtt{p}_G = \operatorname{tr}(B_G)$; see (2.12) and examples in Section 3. Therefore, $\mathtt{r}_G\,\tau_{3,H} \asymp (\mathtt{p}_G/n)^{1/2}$, and conditions (2.13) require only that the value $\mathtt{p}_G$ is smaller in order than the sample size n, i.e. $\mathtt{p}_G \ll n$.

Due to the concentration result of Proposition 2.4, the estimate $\widetilde{\boldsymbol{\theta}}_G$ lies with a dominating probability in a local vicinity of the point $\boldsymbol{\theta}_G^*$. Now one can use a quadratic approximation for the penalized log-likelihood process $L_G(\boldsymbol{\theta})$ to establish an expansion for the penalized MLE $\widetilde{\boldsymbol{\theta}}_G$ and for the excess $L_G(\widetilde{\boldsymbol{\theta}}_G) - L_G(\boldsymbol{\theta}_G^*)$.

Theorem 2.5. *Under the conditions of Proposition* 2.4, *it holds on* $\Omega(\mathtt{x})$

$$\left\|D_G\big(\widetilde{\boldsymbol{\theta}}_G - \boldsymbol{\theta}_G^*\big) - D_G^{-1}\nabla\zeta\right\|^2 \le 4\mathtt{r}_G^3\,\tau_{3,H}. \tag{2.15}$$

$$\left|L_G(\widetilde{\boldsymbol{\theta}}_G) - L_G(\boldsymbol{\theta}_G^*) - \frac{1}{2}\big\|D_G^{-1}\nabla\zeta\big\|^2\right| \le \mathtt{r}_G^3\,\tau_{3,H}. \tag{2.16}$$

Also

$$\left|L_G(\widetilde{\boldsymbol{\theta}}_G) - L_G(\boldsymbol{\theta}_G^*) - \frac{1}{2}\big\|D_G(\widetilde{\boldsymbol{\theta}}_G - \boldsymbol{\theta}_G^*)\big\|^2\right| \le \mathtt{r}_G^3\,\tau_{3,H},$$

$$\sup_{\boldsymbol{\theta}\in\mathcal{A}_G(\mathtt{r}_G)}\left|L_G(\widetilde{\boldsymbol{\theta}}_G) - L_G(\boldsymbol{\theta}) - \frac{1}{2}\big\|\widetilde{D}_G(\widetilde{\boldsymbol{\theta}}_G - \boldsymbol{\theta})\big\|^2\right| \le \mathtt{r}_G^3\,\tau_{3,H}, \tag{2.17}$$

where the random matrix $\widetilde{D}_G^2 = \mathbb{F}_G(\widetilde{\boldsymbol{\theta}}_G)$ *fulfills on* $\Omega(\mathtt{x})$ *for some universal constant* $\mathtt{C}$

$$\left\|D_G^{-1}\big(\widetilde{D}_G^2 - D_G^2\big)D_G^{-1}\right\| \le \mathtt{C}\mathtt{r}_G\,\tau_{3,H}. \tag{2.18}$$

Remark 2.2. The results of Theorem 2.5 can be viewed as finite sample nonparametric analogs of classical asymptotic parametric results such as Fisher expansion for the MLE and Wilks phenomenon. In fact, all the mentioned classical results can be easily derived from (2.15) through (2.17) provided asymptotic normality of the normalized score $D_G^{-1} \nabla \zeta$; see Theorem 2.7 below. In particular, one can state root-n normality of the pMLE and a generalized chi-squared limit distribution for the excess $L_G(\widetilde{\boldsymbol{\theta}}_G) - L_G(\boldsymbol{\theta}_G^*)$.

Remark 2.3. Similarly to Proposition 2.4, the results of Theorem 2.5 are meaningful if $\mathtt{p}_G$ is significantly smaller than n.

The concentration set $\mathcal{A}_G(\mathtt{r}_G)$ becomes smaller when G^2 increases. In particular, if G^2 is large then $\widetilde{\boldsymbol{\theta}}_G$ concentrates on a small vicinity of $\boldsymbol{\theta}_G^*$. At the same time, penalization $\|G\boldsymbol{\theta}\|^2$ yields some estimation bias measured by $I\!E L_G(\boldsymbol{\theta}_G^*) - I\!E L_G(\boldsymbol{\theta}^*)$ and $\boldsymbol{\theta}_G^* - \boldsymbol{\theta}^*$. The bias is not critical if the true value $\boldsymbol{\theta}^*$ is "smooth", that is, $\|G\boldsymbol{\theta}^*\|^2$ is not too big. The next result makes these statements precise.

Proposition 2.6. *It holds*

$$I\!E L_G(\boldsymbol{\theta}_G^*) - I\!E L_G(\boldsymbol{\theta}^*) \leq \frac{1}{2}\|G\boldsymbol{\theta}^*\|^2.$$

If, in addition, $\|G\boldsymbol{\theta}^*\|^2 \leq \mathtt{r}_b^2/2$ for some $\mathtt{r}_b$ such that $\mathtt{r}_b\,\tau_{3,H}(\mathtt{r}_b) \leq 1/2$, then

$$\left| I\!E L_G(\boldsymbol{\theta}_G^*) - I\!E L_G(\boldsymbol{\theta}^*) - \left\|D_G(\boldsymbol{\theta}_G^* - \boldsymbol{\theta}^*)\right\|^2/2 \right| \leq \mathtt{r}_b^3\,\tau_{3,H}(\mathtt{r}_b),$$

$$\left\|D_G(\boldsymbol{\theta}^* - \boldsymbol{\theta}_G^*) - D_G^{-1}G^2\boldsymbol{\theta}^*\right\|^2 \leq 4\mathtt{r}_b^3\,\tau_{3,H}(\mathtt{r}_b). \tag{2.19}$$

Moreover, for any linear mapping Q in $I\!R^p$ it holds

$$\|Q(\boldsymbol{\theta}_G^* - \boldsymbol{\theta}^*)\| \leq \|QD_G^{-2}G^2\boldsymbol{\theta}^*\| + 2\sqrt{\|QD_G^{-2}Q^\top\|\,\mathtt{r}_b^3\,\tau_{3,H}(\mathtt{r}_b)}. \tag{2.20}$$

Finally, we can combine all the previous results together to bound the loss $\|Q(\widetilde{\boldsymbol{\theta}}_G - \boldsymbol{\theta}^*)\| = \|Q(\widetilde{\boldsymbol{\theta}}_G - \boldsymbol{\theta}^*)\| = \|Q(\widetilde{\boldsymbol{\theta}}_G - \boldsymbol{\theta}_G^*) + Q(\boldsymbol{\theta}_G - \boldsymbol{\theta}^*)\|$. One can apply $Q = \sqrt{\mathbb{F}(\boldsymbol{\theta}^*)}$ for prediction and $Q = \sqrt{n}\,I_p$ for estimation. The first bound can be viewed as analog of classical bias-variance decomposition of the loss $\|Q(\widetilde{\boldsymbol{\theta}}_G - \boldsymbol{\theta}^*)\|$. The second one is asymptotic and it corresponds to the case of "small bias" or "undersmoothing". Below $o(1)$ means a small asymptotically vanishing value, and "$a \lesssim b$" means $a \leq \mathtt{C}b$ with an absolute constant $\mathtt{C}$ that possibly depends on the constants from our conditions. We also ignore the small error term in (2.20).

Theorem 2.7. *Given Q with $Q^\top Q \leq D_G^2$, define $B_{Q|G} = QD_G^{-2}H^2D_G^{-2}Q^\top$.*
(1) On a random set $\Omega_1(\mathtt{x})$ with $I\!\!P\big(\Omega_1(\mathtt{x})\big) \leq 2\mathrm{e}^{-\mathtt{x}}$, it holds

$$\|Q(\widetilde{\boldsymbol{\theta}}_G - \boldsymbol{\theta}^*)\| \leq \|QD_G^{-2}G^2\boldsymbol{\theta}^*\| + z(B_{Q|G}, \mathtt{x}), \tag{2.21}$$

where $z(B, \mathtt{x}) \leq \sqrt{\mathrm{tr}\, B} + \sqrt{2\|B\|\,\mathtt{x}}$; see (2.12). Under the bias-variance trade-off

$$\|QD_G^{-2}G^2\boldsymbol{\theta}^*\|^2 \lesssim \mathrm{tr}\big(B_{Q|G}\big),$$

one obtains on $\Omega(\mathtt{x})$

$$\|Q(\widetilde{\boldsymbol{\theta}}_G - \boldsymbol{\theta}^*)\|^2 \lesssim \mathrm{tr}(B_{Q|G}) + \|B_{Q|G}\|\,\mathtt{x}. \tag{2.22}$$

(2) Let $\nabla\zeta$ be nearly normal in the sense that

$$\sup_{\boldsymbol{a}\in I\!\!R^p}\sup_{z>0}\Big|I\!\!P\big(\|QD_G^{-2}\nabla\zeta - \boldsymbol{a}\| \leq z\big) - I\!\!P\big(\|V_{Q|G}\boldsymbol{\gamma} - \boldsymbol{a}\| \leq z\big)\Big| = o(1) \tag{2.23}$$

with $\boldsymbol{\gamma} \in I\!\!R^p$ standard normal and $V_{Q|G}^2 = \mathrm{Var}\big(QD_G^{-2}\nabla\zeta\big) = QD_G^{-2}V^2D_G^{-2}Q^\top$. Assume also the "small bias" condition

$$\frac{\|QD_G^{-2}G^2\boldsymbol{\theta}^*\|^2}{\mathrm{tr}\big(V_{Q|G}^2\big)} = o(1). \tag{2.24}$$

Then it holds

$$\sup_{z>0}\Big|I\!\!P\big(\|Q(\widetilde{\boldsymbol{\theta}}_G - \boldsymbol{\theta}^*)\| \leq z\big) - I\!\!P\big(\|V_{Q|G}\boldsymbol{\gamma}\| \leq z\big)\Big| = o(1). \tag{2.25}$$

Remark 2.4. Given α, define z_α by $I\!\!P(\|V_{Q|G}\boldsymbol{\gamma}\| \geq z_\alpha) = \alpha$. It follows from (2.25) that $\mathcal{E}_{Q|G}(z_\alpha) \stackrel{\mathrm{def}}{=} \big\{\boldsymbol{\theta}\colon \|Q(\widetilde{\boldsymbol{\theta}}_G - \boldsymbol{\theta})\| \leq z_\alpha\big\}$ is asymptotically valid confidence set for $\boldsymbol{\theta}^*$.

Remark 2.5. The result (2.22) with $Q = D_G$ yields on $\Omega(\mathtt{x})$ a bound $\|Q(\widetilde{\boldsymbol{\theta}}_G - \boldsymbol{\theta}^*)\|^2 \lesssim \mathtt{p}_G$ as in Spokoiny (2017).

2.4 Posterior concentration and Gaussian approximation

Now we turn to the properties of the posterior $\boldsymbol{\vartheta}_G\,\big|\,\boldsymbol{Y}$. Our first result shows that the posterior concentrates on the elliptic set $\mathcal{E}_G(\mathtt{r}_0) = \big\{\boldsymbol{u}\colon \|H(\boldsymbol{u} - \widetilde{\boldsymbol{\theta}}_G)\| \leq \mathtt{r}_0\big\}$ for a proper value $\mathtt{r}_0 \geq \mathtt{C}\sqrt{\widetilde{\mathtt{p}}_G} + \mathtt{C}\sqrt{\mathtt{x}}$ for $\widetilde{\mathtt{p}}_G = \mathtt{p}_G(\widetilde{\boldsymbol{\theta}}_G)$. For this we bound from above the random quantity

$$\rho(\mathtt{r}_0) \stackrel{\mathrm{def}}{=} \frac{\int I\!\!I\big(\|H\boldsymbol{u}\| > \mathtt{r}_0\big)\exp\big\{L_G(\widetilde{\boldsymbol{\theta}}_G + \boldsymbol{u})\big\}d\boldsymbol{u}}{\int I\!\!I\big(\|H\boldsymbol{u}\| \leq \mathtt{r}_0\big)\exp\big\{L_G(\widetilde{\boldsymbol{\theta}}_G + \boldsymbol{u})\big\}d\boldsymbol{u}}. \tag{2.26}$$

Obviously $\mathbb{P}\big(\|H(\boldsymbol{\vartheta}_G - \widetilde{\boldsymbol{\theta}}_G)\| > \mathtt{r}_0 \,\big|\, \boldsymbol{Y}\big) \leq \rho(\mathtt{r}_0)$. Therefore, small values of $\rho(\mathtt{r}_0)$ indicate a concentration of $\boldsymbol{\vartheta}_G \,|\, \boldsymbol{Y}$ on the set $\mathcal{E}_G(\mathtt{r}_0)$.

Proposition 2.8. *Let conditions of Proposition 2.4 be satisfied. Assume that for some fixed values $\mathtt{r}_0$ and $\mathtt{x} > 0$, it holds*

$$\diamond(\mathtt{r}_0) \stackrel{\text{def}}{=} 4\mathtt{r}_0^6\,\tau_{3,H}^2 + 4\mathtt{r}_0^4\,\tau_{4,H} \leq 1/2, \qquad \mathtt{C}_0 \stackrel{\text{def}}{=} 1 - 3\mathtt{r}_0\,\tau_{3,H} \geq 1/2, \qquad (2.27)$$

and also

$$\mathtt{C}_0\mathtt{r}_0 \geq 2\sqrt{\mathtt{p}_G(\boldsymbol{\theta})} + \sqrt{\mathtt{x}}, \qquad \boldsymbol{\theta} \in \Theta^\circ. \qquad (2.28)$$

Then, on the random set $\Omega(\mathtt{x})$ from (2.11), $\rho(\mathtt{r}_0)$ from (2.26) fulfills with $\widetilde{\mathtt{p}}_G = \mathtt{p}_G(\widetilde{\boldsymbol{\theta}}_G)$

$$\rho(\mathtt{r}_0) \leq \frac{1}{1 - \diamond(\mathtt{r}_0)}\, \frac{\exp\big\{-(\widetilde{\mathtt{p}}_G + \mathtt{x})/2\big\}}{1 - \exp\big\{-(\widetilde{\mathtt{p}}_G + \mathtt{x})/2\big\}}. \qquad (2.29)$$

Remark 2.6. Conditions (2.27) and (2.28) can be spelled out as follows: the value $\mathtt{r}_0^2$ is of order at least the effective dimension $\mathtt{p}_G(\boldsymbol{\theta})$, while the values $\mathtt{r}_0^3\,\tau_{3,H}$ and $\mathtt{r}_0^4\,\tau_{4,H}$ should be small. If $\tau_{3,H} \asymp n^{-1/2}$, $\tau_{4,H} \asymp n^{-1}$, then (2.27) requires $\mathtt{p}_G^3(\boldsymbol{\theta}) \ll n$, $\boldsymbol{\theta} \in \Theta^\circ$.

Remark 2.7. It is of interest to compare the concentration result of Proposition 2.4 for the pMLE $\widetilde{\boldsymbol{\theta}}_G$ and the concentration bound (2.29) for the posterior $\boldsymbol{\vartheta}_G \,|\, \boldsymbol{Y}$. Suppose that $\mathtt{p}_G(\boldsymbol{\theta}) \asymp \mathtt{p}_G$ for $\boldsymbol{\theta} \in \Theta^\circ$. Then also $\mathtt{r}_0 \asymp \mathtt{r}_G$. Therefore, the concentration results for the penalized MLE $\widetilde{\boldsymbol{\theta}}_G$ and for the posterior $\boldsymbol{\vartheta}_G \,|\, \boldsymbol{Y}$ look similar, but there is one essential difference. The properly shifted MLE $\widetilde{\boldsymbol{\theta}}_G$ well concentrates on a rather small elliptic set $\big\{\boldsymbol{u}\colon \|D_G(\boldsymbol{u} - \boldsymbol{\theta}_G^*)\| \leq \mathtt{r}_G\big\}$ centered at $\widetilde{\boldsymbol{\theta}}_G$. In other words, $D_G\big(\widetilde{\boldsymbol{\theta}}_G - \boldsymbol{\theta}_G^*\big)$ belongs to the ball in $\mathbb{R}^p$ of radius $\mathtt{r}_G$ with a high probability. This holds true even if $p = \infty$. Our result of Proposition 2.8 claims concentration of $\boldsymbol{\vartheta}_G \,|\, \boldsymbol{Y}$ on a larger set $\mathcal{E}_G(\mathtt{r}_0)$, also with an elliptic shape, but centered at $\widetilde{\boldsymbol{\theta}}_G$, and with larger axes corresponding to H^{-1} instead of D_G^{-1}. Later we will see that $D_G\big(\boldsymbol{\vartheta}_G - \widetilde{\boldsymbol{\theta}}_G\big) \,|\, \boldsymbol{Y}$ is close to the standard Gaussian measure and it does not concentrate on a ball in $\mathbb{R}^\infty$ for any radius $\mathtt{r}$.

Our main result claims that the posterior can be well approximated by a Gaussian distribution $\mathcal{N}\big(\widetilde{\boldsymbol{\theta}}_G, \widetilde{D}_G^{-2}\big)$. By $\mathbb{P}'$ we denote a standard normal distribution of a random vector $\boldsymbol{\gamma} \in \mathbb{R}^p$ given $\widetilde{D}_G = D_G(\widetilde{\boldsymbol{\theta}}_G)$. We distinguish between the class $\mathcal{B}_s(\mathbb{R}^p)$ of centrally symmetric Borel sets and the class $\mathcal{B}(\mathbb{R}^p)$ of all Borel sets in $\mathbb{R}^p$.

Theorem 2.9. *Let the conditions of Proposition 2.8 hold and $\rho(\mathbf{r}_0)$ satisfy (2.29). It holds on the set $\Omega(\mathbf{x})$ from (2.11) for any centrally symmetric Borel set $A \in \mathcal{B}_s(\mathbb{R}^p)$*

$$\mathbb{P}\big(\boldsymbol{\vartheta}_G - \widetilde{\boldsymbol{\theta}}_G \in A \,\big|\, \boldsymbol{Y}\big) \geq \frac{1 - \Diamond(\mathbf{r}_0)}{\{1 + \Diamond(\mathbf{r}_0) + \rho(\mathbf{r}_0)\}} \, \mathbb{P}'\big(\widetilde{D}_G^{-1}\boldsymbol{\gamma} \in A\big) - \rho(\mathbf{r}_0) \,,$$

$$\mathbb{P}\big(\boldsymbol{\vartheta}_G - \widetilde{\boldsymbol{\theta}}_G \in A \,\big|\, \boldsymbol{Y}\big) \leq \frac{1 + \Diamond(\mathbf{r}_0)}{\{1 - \Diamond(\mathbf{r}_0)\}\big(1 - \mathrm{e}^{-\mathbf{x}}\big)} \, \mathbb{P}'\big(\widetilde{D}_G^{-1}\boldsymbol{\gamma} \in A\big) + \rho(\mathbf{r}_0) \,.$$

For any measurable set $A \in \mathcal{B}(\mathbb{R}^p)$, similar bounds hold with $\tau_{3,H}$ in place of $\Diamond(\mathbf{r}_0)$.

The first result of the theorem states for any symmetric set $A \in \mathcal{B}_s(\mathbb{R}^p)$

$$\left| \mathbb{P}\big(\boldsymbol{\vartheta}_G - \widetilde{\boldsymbol{\theta}}_G \in A \,\big|\, \boldsymbol{Y}\big) - \mathbb{P}'\big(\widetilde{D}_G^{-1}\boldsymbol{\gamma} \in A\big) \right| \lesssim \mathbb{P}'\big(\widetilde{D}_G^{-1}\boldsymbol{\gamma} \in A\big)\big\{\Diamond(\mathbf{r}_0) + \mathrm{e}^{-\mathbf{x}}\big\} + \rho(\mathbf{r}_0).$$

The second statement applies to any $A \in \mathcal{B}(\mathbb{R}^p)$ and hence, it bounds the distance in total variation between the posterior and its Gaussian approximation $\widetilde{D}_G^{-1}\boldsymbol{\gamma}$.

Corollary 2.10. *Suppose that $\mathbf{r}_0$ satisfies the conditions (2.27) and (2.28) with $\mathbf{x} = 2\log n$. It holds on $\Omega(\mathbf{x})$*

$$\sup_{A \in \mathcal{B}_s(\mathbb{R}^p)} \left| \mathbb{P}\big(\boldsymbol{\vartheta}_G - \widetilde{\boldsymbol{\theta}}_G \in A \,\big|\, \boldsymbol{Y}\big) - \mathbb{P}'\big(\widetilde{D}_G^{-1}\boldsymbol{\gamma} \in A\big) \right| \lesssim \Diamond(\mathbf{r}_0) + 1/n, \qquad (2.30)$$

$$\sup_{A \in \mathcal{B}(\mathbb{R}^p)} \left| \mathbb{P}\big(\boldsymbol{\vartheta}_G - \widetilde{\boldsymbol{\theta}}_G \in A \,\big|\, \boldsymbol{Y}\big) - \mathbb{P}'\big(\widetilde{D}_G^{-1}\boldsymbol{\gamma} \in A\big) \right| \lesssim \mathbf{r}_0^3\,\tau_{3,H} + 1/n. \qquad (2.31)$$

Comparison of two bounds of Corollary 2.10 reveals that the use of symmetric credible sets improves the accuracy of Gaussian approximation from $\mathbf{r}_0^3\,\tau_{3,H}$ to $\Diamond(\mathbf{r}_0) \asymp \mathbf{r}_0^6\,\tau_{3,H}^2 + \mathbf{r}_0^4\,\tau_{4,H}$. In particular, under (2.4) and (2.5), $\tau_{3,H} \asymp \mathbf{r}_0^3\,n^{-1/2}$ while $\Diamond(\mathbf{r}_0) \asymp \mathbf{r}_0^6/n$. The choice $\mathbf{x} = 2\log n$ and $\mathbf{r}_0 = \mathtt{C}\big(\sqrt{\mathtt{p}_G} + \sqrt{\log n}\big)$ yields $\rho(\mathbf{r}_0) \leq 1/n$, and the leading term in the error of Gaussian approximation (2.30) is $\Diamond(\mathbf{r}_0) \asymp \mathtt{p}_G^6/n$. The bound (2.31) in TV-distance ensures an error of order $\sqrt{\mathtt{p}_G^3/n}$.

Now we turn to *contraction property* describing the distance between the support of the posterior and the true value $\boldsymbol{\theta}^*$. For a given linear mapping from $\mathbb{R}^p$ satisfying $Q^\top Q \leq D_G^2$, we would like to describe a minimal radius $\mathbf{r}$ ensuring

$$\mathbb{P}\big(\|Q(\boldsymbol{\vartheta}_G - \boldsymbol{\theta}^*)\| > \mathbf{r} \,\big|\, \boldsymbol{Y}\big) = o(1).$$

Theorem 2.11. *Assume that $Q^\top Q \leq D_G^2$, $H^2 \lesssim D_G^2$, and*

$$\|QD_G^{-2}G^2\boldsymbol{\theta}^*\|^2 \lesssim \mathrm{tr}\big(QD_G^{-2}Q^\top\big). \qquad (2.32)$$

Then it holds on $\Omega(\mathtt{x})$ *for some fixed* $\mathtt{C}, \mathtt{C}_1, \mathtt{C}_2$

$$I\!\!P\left(\left\| Q(\boldsymbol{\vartheta}_G - \boldsymbol{\theta}^*) \right\|^2 \geq \mathtt{C}_1 \operatorname{tr}\left(QD_G^{-2}Q^\top \right) + \mathtt{C}_2 \log n \, \| QD_G^{-2}Q^\top \| \, \big| \, \boldsymbol{Y} \right) \leq \mathtt{C} \, n^{-1}. \quad (2.33)$$

Remark 2.8. If $Q = H$, then $\operatorname{tr}\left(QD_G^{-2}Q^\top \right) = \mathtt{p}_G$, and the contraction radius squared is proportional to the effective dimension. The relation (2.32) is called "bias-variance trade-off", and it means that the squared bias $\left\| Q(\boldsymbol{\theta}_G^* - \boldsymbol{\theta}^*) \right\|^2$ is not larger in order than the trace of the variance of the posterior $Q\boldsymbol{\vartheta}_G \, | \, \boldsymbol{Y}$. The bound (2.33) on posterior contraction is sharp in the sense that it cannot be improved even in the Gaussian case.

Our results help to address the issue of *frequentist coverage* of Bayesian credible sets. Corollary 2.10 suggests to consider credible sets of the form

$$\mathcal{E}_{Q|G}(\mathtt{r}) \stackrel{\text{def}}{=} \left\{ \boldsymbol{\theta} \colon \left\| Q(\widetilde{\boldsymbol{\theta}}_G - \boldsymbol{\theta}) \right\| \leq \mathtt{r} \right\},$$

where $Q^\top Q \leq D_G^2$ and $\mathtt{r} = \mathtt{r}_\alpha$ is fixed to ensure

$$I\!\!P'\left(\left\| Q\widetilde{D}_G^{-1}\boldsymbol{\gamma} \right\| > \mathtt{r}_\alpha \right) = \alpha, \qquad \boldsymbol{\gamma} \sim \mathcal{N}(0, I_p).$$

It appears that frequentist validity of such credible sets can be stated under the same "small bias" condition (2.24) as for pMLE-based confidence sets; see Theorem 2.7.

Theorem 2.12. *Suppose that* $V^2 \leq D_G^2$ *and moreover, the score* $\nabla\zeta$ *is nearly normal as in* (2.23). *Assume also "small bias" condition* (2.24). *Then*

$$I\!\!P\left(\boldsymbol{\theta}^* \notin \mathcal{E}_{Q|G}(\mathtt{r}_\alpha) \right) \leq \alpha + o(1). \quad (2.34)$$

In general, the coverage probability of $\mathcal{E}_{Q|G}(\mathtt{r}_\alpha)$ is larger than the nominal probability α. The reason is that the prior covariance $\widetilde{D}_G^{-2}$ is larger than the covariance $D_G^{-2}V^2D_G^{-2}$ of the pMLE $\widetilde{\boldsymbol{\theta}}_G$.

2.5   Some extensions

Here we list some possible straightforward extensions of the results presented above.

2.5.1   The use of posterior mean

The main results so far used the pMLE $\widetilde{\boldsymbol{\theta}}_G$ for centering the posterior. In practical applications, computing the pMLE or equivalently, MAP $\widetilde{\boldsymbol{\theta}}_G$ could be a hard task. A natural question here is whether one can instead use the posterior mean $\overline{\boldsymbol{\vartheta}}_G = I\!\!E(\boldsymbol{\vartheta}_G \, | \, \boldsymbol{Y})$ which can be efficiently computed by Monte Carlo simulations. One can easily show that

the posterior mean $\overline{\boldsymbol{\vartheta}}_G$ is very close to the pMLE. Unfortunately, our most advanced results on $1/n$ accuracy of Gaussian approximation of the posterior would not apply after centering by $\overline{\boldsymbol{\vartheta}}_G$ because the symmetricity arguments do not apply. This section states some sufficient conditions under which we still get the desirable $1/n$ accuracy even after centering by $\overline{\boldsymbol{\vartheta}}_G$.

We start with the bounds on the first two moments of the posterior. The results claim that posterior mean and variance are close to that of normal $\mathcal{N}\big(\widetilde{\boldsymbol{\theta}}_G, \widetilde{D}_G^{-2}\big)$. Let $\overline{\boldsymbol{\vartheta}}_G = \mathbb{E}\big(\boldsymbol{\vartheta}_G \,\big|\, \boldsymbol{Y}\big)$ and $\boldsymbol{a} = \overline{\boldsymbol{\vartheta}}_G - \widetilde{\boldsymbol{\theta}}_G$. Note that symmetricity arguments do not apply to the posterior mean, therefore one can expect an accuracy of order $\|\boldsymbol{a}\| \asymp \mathtt{r}_0^3 \, \delta_{3,H} \asymp \sqrt{\mathtt{p}_G^3/n}$. In the contrary, the posterior covariance can be estimated with a higher accuracy of order $\Diamond(\mathtt{r}_0)$, again by symmetricity arguments. The next results describes the moments of $Q\big(\boldsymbol{\vartheta}_G - \widetilde{\boldsymbol{\theta}}_G\big)$ for an arbitrary linear operator Q with $Q^\top Q \le D_G^2$.

Theorem 2.13. *Let the conditions of Theorem 2.8 hold. Let $(\mathcal{L}_0)$ be fulfilled with $\mathtt{r}_0$ satisfying (2.28) with $\sqrt{\mathtt{p}_G(\boldsymbol{\theta}) + 1}$ in place of $\sqrt{\mathtt{p}_G(\boldsymbol{\theta})}$, and $\Diamond(\mathtt{r}_0)$ be given by (2.27) and (2.28) with $\mathtt{x} = 2\log n$. Then $\overline{\boldsymbol{\vartheta}}_G = \mathbb{E}\big(\boldsymbol{\vartheta}_G \,\big|\, \boldsymbol{Y}\big)$ fulfills on the set $\Omega(\mathtt{x})$ for any linear operator Q*

$$\big\| Q\big(\overline{\boldsymbol{\vartheta}}_G - \widetilde{\boldsymbol{\theta}}_G\big) \big\| \lesssim \delta_3(\mathtt{r}_0)\sqrt{\widetilde{\mathtt{p}}_{Q|G}} + n^{-1}\widetilde{\mathtt{p}}_{Q|G}, \tag{2.35}$$

where $\widetilde{\mathtt{p}}_{Q|G} = \mathrm{tr}(Q\widetilde{D}_G^{-2}Q^\top)$. The posterior variance $\mathrm{Var}\big(\boldsymbol{\vartheta}_G \,\big|\, \boldsymbol{Y}\big)$ fulfills on $\Omega(\mathtt{x})$

$$\big\| I - \widetilde{D}_G \, \mathrm{Var}\big(\boldsymbol{\vartheta}_G \,\big|\, \boldsymbol{Y}\big)\widetilde{D}_G \big\| = \sup_{\boldsymbol{z}\in\mathbb{S}_p}\left| \mathbb{E}\Big[\big\langle \boldsymbol{z}, \widetilde{D}_G\big(\boldsymbol{\vartheta}_G - \widetilde{\boldsymbol{\theta}}_G\big)\big\rangle^2 \,\Big|\, \boldsymbol{Y}\Big] - 1 \right| \lesssim \Diamond(\mathtt{r}_0),$$

$$\left| \mathbb{E}\Big(\big\| Q\big(\boldsymbol{\vartheta}_G - \widetilde{\boldsymbol{\theta}}_G\big)\big\|^2 \,\Big|\, \boldsymbol{Y}\Big) - \widetilde{\mathtt{p}}_{Q|G} \right| \lesssim \Diamond(\mathtt{r}_0)\,\widetilde{\mathtt{p}}_{Q|G}.$$

Remark 2.9. The result (2.35) of Theorem 2.13 does not apply to $Q = \widetilde{D}_G$ for $p = \infty$ because $\widetilde{D}_G\big(\boldsymbol{\vartheta}_G - \widetilde{\boldsymbol{\theta}}_G\big)$ is nearly standard normal given $\boldsymbol{Y}$ and $\widetilde{\mathtt{p}}_{Q|G} = \infty$. However, it well applies to $Q = \widetilde{D}$ yielding

$$\big\| \widetilde{D}\big(\overline{\boldsymbol{\vartheta}}_G - \widetilde{\boldsymbol{\theta}}_G\big) \big\| \lesssim \delta_3(\mathtt{r}_0)\sqrt{\widetilde{\mathtt{p}}_G},$$

$$\mathbb{E}\Big(\big\| \widetilde{D}\big(\boldsymbol{\vartheta}_G - \widetilde{\boldsymbol{\theta}}_G\big)\big\|^2 \,\Big|\, \boldsymbol{Y}\Big) \approx \widetilde{\mathtt{p}}_G.$$

Another typical choice of Q is $Q = \Pi_m\widetilde{D}_G$, where Π_m is the projector on the first m eigenvectors of $\widetilde{D}_G$. Then $\widetilde{\mathtt{p}}_{Q|G} = m$.

Now we state the result about the Gaussian approximation of $\boldsymbol{\vartheta}_G - \overline{\boldsymbol{\vartheta}}_G \,\big|\, \boldsymbol{Y}$.

Theorem 2.14. *Let the conditions of Theorem 2.13 be satisfied, let $Q = Q\Pi$ where Π is a projector in $\mathbb{R}^p$. Then it holds on the set $\Omega(\mathbf{x})$ shown in Proposition 2.4*

$$\sup_{\mathbf{r} > 0} \left| I\!\!P\left(\|Q(\boldsymbol{\vartheta}_G - \overline{\boldsymbol{\vartheta}}_G)\| \leq \mathbf{r} \,|\, \boldsymbol{Y} \right) - I\!\!P'\left(\|Q\widetilde{D}_G^{-1}\boldsymbol{\gamma}\| \leq \mathbf{r} \right) \right| \lesssim \Diamond(\mathbf{r}_0)\sqrt{\widetilde{\mathbf{p}}_\Pi} + n^{-1}\widetilde{\mathbf{p}}_\Pi$$

with $\widetilde{\mathbf{p}}_\Pi \stackrel{\text{def}}{=} \mathrm{tr}\left(\Pi\widetilde{D}_G^2\Pi\widetilde{D}_G^{-2}\Pi \right)$.

Remark 2.10. If $\widetilde{D}_G\Pi = \Pi\widetilde{D}_G$, then $\widetilde{\mathbf{p}}_\Pi = \dim(\Pi)$. In particular, if $Q = \widetilde{D}_G\Pi_m$ for the eigenprojector Π_m as in Remark 2.9, then $\widetilde{\mathbf{p}}_G = \widetilde{\mathbf{p}}_{\Pi_m} = m$. One can see that the use of the posterior mean instead of the penalized MLE $\widetilde{\boldsymbol{\theta}}_G$ is justified under a bit stronger condition "$\Diamond(\mathbf{r}_0)\sqrt{\widetilde{\mathbf{p}}_\Pi} + n^{-1}\widetilde{\mathbf{p}}_\Pi$ is small" compared to "$\Diamond(\mathbf{r}_0)$ is small".

2.5.2   Prior comparison and prior impact

The classical BvM result claims that the prior impact asymptotically washes out, as the sample size increases. The posterior becomes close to the normal distribution with the same distribution as for the MLE $\widetilde{\boldsymbol{\theta}}$, namely, to $\mathcal{N}(\widetilde{\boldsymbol{\theta}}, D^{-2})$. It is well understood that a general BvM result is impossible in a infinite dimensional nonparametric set-up whatever sample size is. In this section we want to quantify the accuracy of the BvM approximation using the obtained bounds on the Gaussian approximation of the posterior and the results on Gaussian comparison. As in previous sections we show that restricting to the class of elliptic sets helps to improve the bounds. We slightly change the statement of the problem and consider it as a problem of prior impact. Let G^{-2} and G_1^{-2} be prior covariance matrices for two different priors. We want to compare their posteriors. A special case of interest is when G_1 is large and corresponds to a model of low complexity, which can be measured by the effective dimension. The second prior has a smaller precision matrix G^2 and hence a more complex model. By Theorem 2.9 the posterior $\boldsymbol{\vartheta}_G \,|\, \boldsymbol{Y}$ is nearly Gaussian $\mathcal{N}(\widetilde{\boldsymbol{\theta}}_G, \widetilde{D}_G^{-2})$. Similarly the posterior $\boldsymbol{\vartheta}_{G_1} \,|\, \boldsymbol{Y}$ is close to $\mathcal{N}(\widetilde{\boldsymbol{\theta}}_{G_1}, \widetilde{D}_{G_1}^{-2})$. We are interested if elliptic credible sets calibrated for the simple G_1-prior can be used for the more complex G-prior. The Gaussian approximation reduces this question to Gaussian comparison; see Götze et al. (2019) or Theorem A.1. Motivated by the above discussion we assume $G^2 \leq G_1^2$.

Theorem 2.15. *Let the conditions of Theorem 2.13 be satisfied for two priors $\mathcal{N}(0, G^{-2})$*

and $\mathcal{N}(0, G_1^{-2})$ with $G^2 \leq G_1^2$. Then it holds on a set $\Omega(\mathtt{x})$ with $I\!\!P\big(\Omega(\mathtt{x})\big) \geq 1 - 2/n$

$$
\sup_{\mathtt{r}} \left| I\!\!P\Big(\|Q(\boldsymbol{\vartheta}_G - \widetilde{\boldsymbol{\theta}}_{G_1})\| \leq \mathtt{r} \,\big|\, \boldsymbol{Y} \Big) - I\!\!P\Big(\|Q(\boldsymbol{\vartheta}_{G_1} - \widetilde{\boldsymbol{\theta}}_{G_1})\| \leq \mathtt{r} \,\big|\, \boldsymbol{Y} \Big) \right|
$$

$$
\lesssim \delta_3(\mathtt{r}_0) + n^{-1} + \frac{1}{\|Q\widetilde{D}_G^{-2}Q^\top\|_{\mathrm{Fr}}} \Big\{ \mathrm{tr}\big(Q(\widetilde{D}_G^{-2} - \widetilde{D}_{G_1}^{-2})Q^\top\big) + \big\|Q(\widetilde{\boldsymbol{\theta}}_G - \widetilde{\boldsymbol{\theta}}_{G_1})\big\|^2 \Big\}. \quad (2.36)
$$

Remark 2.11. The last term in the bound (2.36) comes from the Gaussian comparison result of Theorem A.1. It includes the "variance" part that relates two covariance operators $Q\widetilde{D}_G^{-2}Q^\top$ and $Q\widetilde{D}_{G_1}^{-2}Q^\top$, and the "squared bias" term $\|Q(\widetilde{\boldsymbol{\theta}}_G - \widetilde{\boldsymbol{\theta}}_{G_1})\|^2$. Applicability of the prior G_1 in place of G is justified under "small bias" condition $\|Q(\widetilde{\boldsymbol{\theta}}_G - \widetilde{\boldsymbol{\theta}}_{G_1})\|^2 \ll \|Q\widetilde{D}_G^{-2}Q^\top\|_{\mathrm{Fr}}$, and under the "variance" condition $\widetilde{\mathtt{p}}_G - \widetilde{\mathtt{p}}_{G_1} \ll \|Q\widetilde{D}_G^{-2}Q^\top\|_{\mathrm{Fr}}$.

2.5.3 A family of priors and uniform Gaussian approximation

Consider a more general situation when a family of Gaussian priors $\mathcal{N}(0, G_\varkappa^{-2})$, $\varkappa \in \mathcal{M}$, is given. For each of them, under appropriate conditions, one can state the Gaussian approximation result as in Corollary 2.10 or Theorem 2.9. The choice of a prior by empirical or full Bayes approaches requires to state this approximation uniformly in $\varkappa \in \mathcal{M}$. Surprisingly, in the contrary to the classical frequentist model selection, such a uniform approximation can be stated in a straightforward way. In fact, all the results about posterior distribution are stated conditionally on the data after restricting to the random set $\Omega(\mathtt{x})$ on which a deviation bound $\|D_G^{-1}\nabla\boldsymbol{\zeta}\| \leq z(B_{V|G}, \mathtt{x})$ holds. All we need is a uniformly in $\varkappa$ version of this bound. Note that the conditions $(\mathcal{L})$, $(\boldsymbol{E})$, $(\boldsymbol{EV})$, and $(\mathcal{L}_0)$ do not involve any prior. The only condition $(\boldsymbol{HG})$ has to be fulfilled uniformly in $G \in \{G_\varkappa\}$. In the case when the family $\{G_\varkappa\}$ contains the smallest covariance $G_{\min}^2 \leq G_\varkappa^2$, it suffices to check $(\boldsymbol{HG})$ for $G_{\min}^2$ only. This is trivially fulfilled for the family of smooth (s, w)-priors for different $s > 1$. For m-truncation priors, one can take any $m \leq \mathtt{C}n^{1/(2s_{\min}+1)}$.

The next result describes the uniform properties of the estimators $\widetilde{\boldsymbol{\theta}}_\varkappa$ and the posteriors $\boldsymbol{\vartheta}_\varkappa \,|\, \boldsymbol{Y}$. Everywhere we write the subindex $\varkappa$ in place of $G_\varkappa$. In particular, $D_\varkappa^2 = \mathbb{F}_\varkappa(\boldsymbol{\theta}_\varkappa^*) = \mathbb{F}(\boldsymbol{\theta}_\varkappa^*) + G_\varkappa^2$ with $\mathbb{F}(\boldsymbol{\theta}) = -\nabla^2 I\!\!E L(\boldsymbol{\theta})$.

Theorem 2.16. Let the conditions $(\mathcal{L})$, $(\boldsymbol{E})$, $(\boldsymbol{EV})$, and $(\mathcal{L}_0)$ be fulfilled. Let also $\{\mathcal{N}(0, G_\varkappa^{-2}), \varkappa \in \mathcal{M}\}$ be a family of Gaussian priors, and let $(\boldsymbol{HG})$ be satisfied uniformly in $\varkappa \in \mathcal{M}$. Let also there exist a random set $\Omega(\mathtt{x})$ with $I\!\!P\big(\Omega(\mathtt{x})\big) \geq 1 - \mathrm{e}^{-\mathtt{x}}$

and an upper function $z(B_{V|G_\varkappa}, \mathrm{x})$ *of* $\varkappa$ *such that it holds on* $\Omega(\mathrm{x})$

$$\|D_\varkappa^{-1} \nabla \boldsymbol{\zeta}\| \leq z(B_{V|G_\varkappa}, \mathrm{x}), \qquad \forall \varkappa \in \mathcal{M}.$$

Then all the statements of Proposition 2.4 through 2.15 are fulfilled on $\Omega(\mathrm{x})$ *uniformly in* $\varkappa \in \mathcal{M}$. *In particular, each of* $\widetilde{\boldsymbol{\theta}}_\varkappa$ *concentrates on the elliptic set*

$$\mathcal{A}_\varkappa(\mathtt{r}_\varkappa) = \big\{ \boldsymbol{\theta} \colon \|D_\varkappa(\boldsymbol{\theta} - \boldsymbol{\theta}_\varkappa^*)\| \leq \mathtt{r}_\varkappa \big\},$$

with the center at $\boldsymbol{\theta}_\varkappa^*$, *while the posterior* $\boldsymbol{\vartheta}_\varkappa \,\big|\, \boldsymbol{Y}$ *concentrates on a larger vicinity*

$$\mathcal{E}_\varkappa(\mathtt{r}_\varkappa) = \big\{ \boldsymbol{\theta} \colon \big\|H(\boldsymbol{\theta} - \widetilde{\boldsymbol{\theta}}_\varkappa)\big\| \leq \mathtt{r}_\varkappa \big\}$$

centered at $\widetilde{\boldsymbol{\theta}}_\varkappa$. *Each posterior* $\boldsymbol{\vartheta}_\varkappa \,\big|\, \boldsymbol{Y}$ *can be approximated by the Gaussian* $\mathcal{N}(\widetilde{\boldsymbol{\theta}}_\varkappa, \widetilde{D}_\varkappa^{-2})$, *the error of approximation is given by Theorem 2.9 or Corollary 2.10.*

This result follows from the fact that after restricting to the set $\Omega(\mathrm{x})$ we only operate with deterministic function $I\!\!E L(\boldsymbol{\theta})$ and use its local smoothness properties from $(\boldsymbol{\mathcal{L}_0})$.

Remark 2.12. The probabilistic bound $I\!\!P\big(\|D_\varkappa^{-1}\nabla\boldsymbol{\zeta}\| > z(B_{V|G_\varkappa}, \mathrm{x})\big) \leq \mathrm{e}^{-\mathrm{x}}$ follows from $(\boldsymbol{EV})$ for each $\varkappa \in \mathcal{M}$. If $\mathcal{M}$ is a finite set and $|\mathcal{M}|$ is its cardinality then a uniform version of this bound can be easily obtained by an increase of x to $\mathrm{x} + \log|\mathcal{M}|$.

2.5.4 Non-Gaussian priors

The assumption of a Gaussian prior was essential in the stated results. However, the applied technique allows to extend the approach to a broad class of non-Gaussian priors under some mild assumptions. For simplicity we restrict ourselves to the case of Gaussian likelihood $L(\boldsymbol{\theta})$ which is a quadratic function of $\boldsymbol{\theta}$, so that $-\nabla^2 L(\boldsymbol{\theta}) = \mathbb{F}$ for a symmetric operator $\mathbb{F}$ in $I\!\!R^p$. Now, let the prior has a log-concave density $\Pi(\boldsymbol{\theta})$ which we also assume to be a sufficiently smooth. More precisely, we suppose that

$$G^2(\boldsymbol{\theta}) \stackrel{\text{def}}{=} -\nabla^2 \log \Pi(\boldsymbol{\theta}) \geq 0.$$

Define

$$\mathbb{F}_G(\boldsymbol{\theta}) \stackrel{\text{def}}{=} \mathbb{F} + G^2(\boldsymbol{\theta})$$

and redefine $\delta_m(\boldsymbol{\theta}, \boldsymbol{u})$ in (2.2) using $\log \Pi(\boldsymbol{\theta})$ in place of $I\!\!E L(\boldsymbol{\theta})$. It is rather straightforward to see that with this exchange, all the previous results continue to apply without any change.

3 Examples

In this section we illustrate the general results of the Section 2 by applying to nonparametric density estimation and generalized regression. Log-density model is a popular example in statistical literature related to BvM Theorem and nonparametric Bayes study. We mention Castillo and Nickl (2014), Castillo and Rousseau (2015) among many others. Generalized regression model includes the logit model for binary response or classification problems, Poisson and Cox regression, several reliability models and so on. The related BvM results can be found e.g. in Castillo and Nickl (2014), Ghosal and van der Vaart (2017) and references therein. The results on Gaussian approximation of the posterior are typically asymptotic and do not provide any accuracy guarantees for this approximation. Our results are stated for finite samples and deliver the quantitative and tight bounds on the accuracy of this approximation in terms of effective dimension of the problem.

3.1 Nonparametric log-density estimation

Suppose we are given a random sample $X_1, \ldots, X_n$ in $\mathbb{R}^d$. The i.i.d. model assumption means that all these random variables are independent identically distributed from some measure P with a density $f(x)$ with respect to a σ-finite measure μ_0 in $\mathbb{R}^d$. This density function is the target of estimation. By definition, the function f is non-negative, measurable, and integrates to one: $\int f(x)\,\mu_0(dx) = 1$. Here and below, the integral $\int$ without limits means the integral over the whole space $\mathbb{R}^d$. If $f(\cdot)$ has a smaller support $\mathcal{X}$, one can restrict integration to this set. Below we parametrize the model by a linear decomposition of the log-density function. Let $\{\psi_j(x),\, j = 1, \ldots, p\}$ with $p \leq \infty$ be a collection of functions in $\mathbb{R}^d$ (a dictionary). For each $\boldsymbol{\theta} = (\theta_j) \in \mathbb{R}^p$, define

$$\ell(x, \boldsymbol{\theta}) \stackrel{\text{def}}{=} \sum_{j=1}^{p} \theta_j \psi_j(x) - \phi(\boldsymbol{\theta}) = \langle \Psi(x), \boldsymbol{\theta} \rangle - \phi(\boldsymbol{\theta}),$$

where $\phi(\boldsymbol{\theta})$ is given by

$$\phi(\boldsymbol{\theta}) \stackrel{\text{def}}{=} \log \int \exp\left\{ \sum_{j=1}^{p} \theta_j \psi_j(x) \right\} \mu_0(dx) = \log \int e^{\langle \Psi(x), \boldsymbol{\theta} \rangle} \mu_0(dx). \tag{3.1}$$

Here $\Psi(x)$ is a vector with components $\psi_j(x)$. Linear log-density modeling assumes

$$\log f(x) = \ell(x, \boldsymbol{\theta}^*) = \langle \Psi(x), \boldsymbol{\theta}^* \rangle - \phi(\boldsymbol{\theta}^*) \tag{3.2}$$

for some $\boldsymbol{\theta}^* \in \Theta \subseteq \mathbb{R}^p$. A nice feature of such representation is that the function $\log f(x)$ in the contrary to the density itself does not need to be non-negative. One more

important benefit of using the log-density is that the stochastic part of the corresponding log-likelihood is *linear* w.r.t. the parameter $\boldsymbol{\theta}$. The log-likelihood $L(\boldsymbol{\theta})$ reads as

$$L(\boldsymbol{\theta}) = \sum \ell(X_i, \boldsymbol{\theta}) = \sum \langle \Psi(X_i), \boldsymbol{\theta} \rangle - n\phi(\boldsymbol{\theta}) = \langle S, \boldsymbol{\theta} \rangle - n\phi(\boldsymbol{\theta}), \quad S = \sum \Psi(X_i),$$

where $\sum$ means $\sum_{i=1}^{n}$. For applying the general results of Section 2, it suffices to check the general conditions of Section 2.1 for the log-density model. First note that the generalized linear structure of the model yields automatically conditions $(\mathcal{L})$ and (E). Indeed, convexity of $\phi(\cdot)$ implies that $\mathbb{E}L(\boldsymbol{\theta}) = \langle \mathbb{E}S, \boldsymbol{\theta} \rangle - n\phi(\boldsymbol{\theta})$ is concave. Further, for the stochastic component $\zeta(\boldsymbol{\theta}) = L(\boldsymbol{\theta}) - \mathbb{E}L(\boldsymbol{\theta})$, it holds

$$\nabla \zeta(\boldsymbol{\theta}) = \nabla \zeta = S - \mathbb{E}S = \sum \big[\Psi(X_i) - \mathbb{E}\, \Psi(X_i) \big],$$

and (E) follows. Further, the representation $\mathbb{E}L(\boldsymbol{\theta}) = \langle \mathbb{E}S, \boldsymbol{\theta} \rangle - n\phi(\boldsymbol{\theta})$ implies

$$\mathbb{F}(\boldsymbol{\theta}) = -\nabla^2 \mathbb{E}L(\boldsymbol{\theta}) = n\nabla^2 \phi(\boldsymbol{\theta}).$$

To simplify our presentation, we assume that $X_1, \ldots, X_n$ are indeed i.i.d. This can be easily extended to non i.i.d. r.v.'s at cost of more complicated notations. Then

$$\mathbb{E}S = \sum \mathbb{E}\, \Psi(X_i) = n\, \mathbb{E}\Psi(X_1) = n\,\overline{\Psi}$$

with $\overline{\Psi} = \mathbb{E}\, \Psi(X_1)$. We further assume that the underlying density $f(x)$ can be represented in the form (3.2) for some parameter vector $\boldsymbol{\theta}^*$. It also holds

$$\boldsymbol{\theta}^* = \operatorname*{argmax}_{\boldsymbol{\theta} \in \Theta} \mathbb{E}L(\boldsymbol{\theta}) = \operatorname*{argmax}_{\boldsymbol{\theta} \in \Theta} \big\{ \langle \mathbb{E}S, \boldsymbol{\theta} \rangle - n\phi(\boldsymbol{\theta}) \big\} = \operatorname*{argmax}_{\boldsymbol{\theta} \in \Theta} \big\{ \langle \overline{\Psi}, \boldsymbol{\theta} \rangle - \phi(\boldsymbol{\theta}) \big\}.$$

Now we switch to the Bayesian framework and restrict ourselves to m-truncation or (s, w)-smoothing priors; see Section 2.2. In the second case, we still truncate the prior at the fixed level $m^* = n^{1/3}/\log n$. This means that the prior is supported to the subspace $\mathbb{R}^{m^*}$ of $\mathbb{R}^p$. For a given penalty operator G^2, the corresponding penalized MLE $\widetilde{\boldsymbol{\theta}}_G$, and the target $\boldsymbol{\theta}_G^*$ are

$$\widetilde{\boldsymbol{\theta}}_G = \operatorname*{argmax}_{\boldsymbol{\theta} \in \Theta} L_G(\boldsymbol{\theta}) = \operatorname*{argmax}_{\boldsymbol{\theta} \in \Theta} \Big\{ \langle \boldsymbol{\theta}, S \rangle - n\phi(\boldsymbol{\theta}) - \frac{1}{2} \|G\boldsymbol{\theta}\|^2 \Big\},$$

$$\boldsymbol{\theta}_G^* = \operatorname*{argmax}_{\boldsymbol{\theta} \in \Theta} \mathbb{E}L_G(\boldsymbol{\theta}) = \operatorname*{argmax}_{\boldsymbol{\theta} \in \Theta} \Big\{ \langle \boldsymbol{\theta}, \mathbb{E}S \rangle - n\phi(\boldsymbol{\theta}) - \frac{1}{2} \|G\boldsymbol{\theta}\|^2 \Big\}.$$

We write $D_G^2 = \mathbb{F}_G(\boldsymbol{\theta}_G^*)$ and $\mathtt{p}_G = \mathtt{p}_G(\boldsymbol{\theta}_G^*)$. Below we assume:

$(\boldsymbol{\Theta})$ $\Theta = \mathcal{B}_s(1)$ for $s > 1$ yielding $\sum_j \theta_j^2 j^{2s} \leq 1$ for all $\boldsymbol{\theta} = (\theta_j)$.

This condition is standard in of log-density estimation; cf. Castillo and Nickl (2014).

(ψ_j) Define $q_j^{-2} = j \log^2(j)$. Then

$$\sup_{x \in \mathfrak{X}} \sum_{j \geq 1} \psi_j^2(x) \, q_j^2 \leq \mathtt{C}_\psi^2. \tag{3.3}$$

One can check that this condition is fulfilled in two important special case: (1) all the basis functions $\psi_j(x)$ are uniformly bounded by a constant $\mathtt{C}_\psi$, e.g. Fourier or cosine basis; (2) $(\psi_j) = (\psi_{kl})_{k \geq 0, l \in \mathcal{I}_k}$ is a double indexed set of wavelet functions satisfying

$$\sup_{x \in \mathfrak{X}} \sum_{l \in \mathcal{I}_k} \psi_{kl}^2(x) \leq \mathtt{C} 2^k.$$

In what follows, for some small but fixed ϱ, we use Θ° of the form

$$\Theta^\circ = \Theta_\varrho = \left\{ \boldsymbol{\theta} = \boldsymbol{\theta}^* + \boldsymbol{u} \colon \langle \nabla^2 \phi(\boldsymbol{\theta}) \boldsymbol{u}, \boldsymbol{u} \rangle \leq \varrho^2 \right\}.$$

$(\nabla^2\phi)$ For some $\mathtt{C}_{\phi,1}, \mathtt{C}_{\phi,2} \geq 1$ and for all $\boldsymbol{\theta} \in \Theta^\circ$ and all $\boldsymbol{u} \in \mathbb{R}^{m^*}$

$$\mathtt{C}_{\phi,1}^{-2} \leq \|\boldsymbol{u}\|^{-2} \langle \nabla^2 \phi(\boldsymbol{\theta}) \boldsymbol{u}, \boldsymbol{u} \rangle \leq \mathtt{C}_{\phi,2}^2. \tag{3.4}$$

This is an identifiability condition ensuring that different features can be well identified from the data.

Define for any $\boldsymbol{\theta} \in \Theta$ a measure $P_{\boldsymbol{\theta}}$ by the relation:

$$\frac{dP_{\boldsymbol{\theta}}}{d\mu_0}(x) = \exp\{\langle \Psi(x), \boldsymbol{\theta} \rangle - \phi(\boldsymbol{\theta})\}.$$

The identity (3.1) ensures that $P_{\boldsymbol{\theta}}$ is a probabilistic measure. Moreover, due to (Θ) and (ψ_j) all such measures are equivalent in the sense the ratio $d\mathbb{P}_{\boldsymbol{\theta}}/dP_{\boldsymbol{\theta}^\circ}$ is bounded by a constant for all $\boldsymbol{\theta}, \boldsymbol{\theta}^\circ$ in Θ because $\sup_{x \in \mathfrak{X}} |\langle \boldsymbol{\theta} - \boldsymbol{\theta}^\circ, \Psi(x) \rangle| < \infty$.

$(\Psi\boldsymbol{u})$ *There exists a constant* $\mathtt{C}_f \geq 1$ *such that it holds for all* $\boldsymbol{\theta} \in \Theta^\circ$, *all* $\boldsymbol{u} \in \mathbb{R}^{m^*}$, *and* $k = 3, 4$

$$E_{\boldsymbol{\theta}} \big| \langle \Psi(X_1) - E_{\boldsymbol{\theta}} \Psi(X_1), \boldsymbol{u} \rangle \big|^k \leq \big\{ \mathtt{C}_f^2 E_{\boldsymbol{\theta}} \langle \Psi(X_1) - E_{\boldsymbol{\theta}} \Psi(X_1), \boldsymbol{u} \rangle^2 \big\}^{k/2}. \tag{3.5}$$

Remark 3.1. In fact, in view of (3.3), it suffices to check (3.4) and (3.5) for one point $\boldsymbol{\theta} \in \Theta^\circ$, in particular, for the true point $\boldsymbol{\theta}^*$ corresponding to the underlying measure P. Condition $(\Psi\boldsymbol{u})$ means that each measure $P_{\boldsymbol{\theta}}$ for $\boldsymbol{\theta}$ in a vicinity Θ° of $\boldsymbol{\theta}^*$ satisfies a kind of Khinchin's inequality which relates the fourth and the second directional moments

of $\Psi(X_1)$. Conditions like (3.5) are often used in high dimensional probability. It is automatically fulfilled for Gaussian measures and for the case when the individual features $\psi_j(x)$ of the vector $\Psi(x)$ are independent and have fourth moments.

Below by $\mathtt{C}_0, \mathtt{C}_1, \mathtt{C}_2, \ldots$ we denote some fixed constants which possibly depending on $\mathtt{C}_f$ and $\mathtt{C}_\phi$ from $(\boldsymbol{\Psi u})$ and $(\boldsymbol{\nabla^2 \phi})$.

Theorem 3.1. *Assume* $(\boldsymbol{\Theta})$, $(\boldsymbol{\psi_j})$, $(\boldsymbol{\nabla^2 \phi})$, *and* $(\boldsymbol{\Psi u})$. *Let* G^2 *correspond to a* m-*truncation prior with* $m \leq m^*$ *yielding* $\mathtt{p}_G \leq m$, *or to a* m^*-*truncated* (s, w)-*smooth prior with* $w \gg n^{1-2s/3}$ *and* $\mathtt{p}_G \lesssim (n/w)^{1/(2s)}$. *Define*

$$z_G \stackrel{\text{def}}{=} \sqrt{\mathtt{p}_G} + \sqrt{2 \log n}. \tag{3.6}$$

Let finally $\boldsymbol{\theta}_G^* \in \Theta^\circ$. *Then it holds for* $n \geq \mathtt{C}_0 z_G^2$ *on a set* Ω_n *with* $\mathbb{P}(\Omega_n) \geq 1 - 3/n$

$$\|D_G(\widetilde{\boldsymbol{\theta}}_G - \boldsymbol{\theta}_G^*)\| \leq 2z_G,$$

$$\left\| D_G(\widetilde{\boldsymbol{\theta}}_G - \boldsymbol{\theta}_G^*) - D_G^{-1} \nabla \zeta \right\| \leq \mathtt{C}_1 z_G^2 n^{-1/2}, \tag{3.7}$$

$$\left| L_G(\widetilde{\boldsymbol{\theta}}_G) - L_G(\boldsymbol{\theta}_G^*) - \frac{1}{2} \|D_G^{-1} \nabla \zeta\|^2 \right| \leq \mathtt{C}_1 z_G^3 n^{-1/2}.$$

Now we continue with the properties of the posterior $\boldsymbol{\vartheta}_G \,|\, \boldsymbol{X}$ for a Gaussian prior $\mathcal{N}(0, G^2)$. Remind $\widetilde{D}_G^2 = \mathbb{F}_G(\widetilde{\boldsymbol{\theta}}_G) = \mathbb{F}(\widetilde{\boldsymbol{\theta}}_G) + G^2$.

Theorem 3.2. *Suppose that the conditions of Theorem 3.1 hold. Then Theorem 2.8 through 2.12 continue to apply to the posterior* $\boldsymbol{\vartheta}_G \,|\, \boldsymbol{X}$. *In particular, on a set* Ω_n *with* $\mathbb{P}(\Omega_n) \geq 1 - 3/n$, *it holds with* z_G *from* (3.6)

$$\sup_{A \in \mathcal{B}_s(\mathbb{R}^p)} \left| \mathbb{P}(\boldsymbol{\vartheta}_G - \widetilde{\boldsymbol{\theta}}_G \in A \,|\, \boldsymbol{X}) - \mathbb{P}'(\widetilde{D}_G^{-1} \boldsymbol{\gamma} \in A) \right| \leq \mathtt{C}_4 z_G^6 n^{-1}, \tag{3.8}$$

$$\sup_{A \in \mathcal{B}(\mathbb{R}^p)} \left| \mathbb{P}(\boldsymbol{\vartheta}_G - \widetilde{\boldsymbol{\theta}}_G \in A \,|\, \boldsymbol{X}) - \mathbb{P}'(\widetilde{D}_G^{-1} \boldsymbol{\gamma} \in A) \right| \leq \mathtt{C}_4 z_G^3 n^{-1/2}.$$

Under a proper choice $m = w = n^{1/(2s+1)}$ of the prior parameters one can derive finite sample versions of the standard nonparametric *rate optimal* results about concentration of the pMLE and posterior contraction; cf Castillo and Nickl (2014), Castillo and Rousseau (2015). We use that $\widetilde{D}_G^2 \geq n \mathtt{C}_\phi^2 I_p$ and $n^{-1} z_G^2 \asymp n^{-1} m \asymp n^{-2s/(2s+1)}$. The error term in (3.8) is of order $n^{(2-2s)/(2s+1)} \to 0$ as $n \to \infty$ because $s > 1$.

Corollary 3.3. *Assume conditions of Theorem 3.1. Define $m = w = n^{1/(2s+1)}$. For the m-truncation prior or for the (s, w) smooth prior, it holds on the same set Ω_n*

$$\|\widetilde{\boldsymbol{\theta}}_G - \boldsymbol{\theta}^*\|^2 \leq \mathtt{C} n^{-2s/(2s+1)},$$

$$\mathbb{P}\Big(\|\boldsymbol{\vartheta}_G - \boldsymbol{\theta}^*\|^2 > \mathtt{C} n^{-2s/(2s+1)} \,\big|\, \boldsymbol{X}\Big) \leq n^{-1},$$

for n sufficiently large. Moreover, under "undersmoothing" $\boldsymbol{\theta}^* \in \mathcal{B}_{s^*}(1)$ with $s^* > s$, Bayesian credible sets $\mathcal{E}_G(\mathtt{r}_\alpha) = \big\{\boldsymbol{\theta} \colon \big\|\sqrt{n}(\widetilde{\boldsymbol{\theta}}_G - \boldsymbol{\theta})\big\| \leq \mathtt{r}_\alpha\big\}$ with $\mathtt{r}_\alpha$ satisfying $\mathbb{P}'\big(\sqrt{n}\big\|\widetilde{D}_G^{-1}\boldsymbol{\gamma}\big\| > \mathtt{r}_\alpha\big) = \alpha$, $\boldsymbol{\gamma} \sim \mathcal{N}(0, \boldsymbol{I}_p)$ are asymptotic valid; see (2.34).

3.2 Generalized regression

Now we discuss how the general results apply to generalized regression. Suppose we are given independent data $Y_1, \ldots, Y_n$ which follow the model

$$Y_i \sim P_{v_i} \in \mathcal{P}, \qquad i = 1, \ldots, n, \tag{3.9}$$

where $\mathcal{P} = (P_v, v \in \Upsilon)$ be a univariate exponential family with a canonical parameter. The latter means that $\mathcal{P}$ is dominated by a σ-finite measure μ and

$$\log \frac{dP_v}{d\mu}(y) = vy - \phi(v) + \ell(y)$$

for a convex function $\phi(v)$ of a univariate parameter v. A typical example is given by the logistic regression with binary observations Y_i. Then $\phi(v) = \log(1 + \mathrm{e}^v)$. The model (3.9) yields $\mathbb{E}Y_i = \phi'(v_i)$ and $\mathrm{Var}(Y_i) = \phi''(v_i)$. We, however, do not assume that the model is correct. The value v_i is just defined by the canonical link $\mathbb{E}Y_i = \phi'(v_i)$.

Generalized regression assumes that the v_i's in (3.9) are values of a function $f(X_i)$ at deterministic design points $X_1, \ldots, X_n$. A linear basis expansion $f(x) = \sum_j \theta_j \psi_j(x)$ leads to a generalized linear model

$$Y_i \sim P_{\langle \boldsymbol{\Psi}_i, \boldsymbol{\theta} \rangle} \tag{3.10}$$

with $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)^\top \in \mathbb{R}^p$ and $\boldsymbol{\Psi}_i = \big(\psi_1(X_i), \ldots, \psi_p(X_i)\big)^\top \in \mathcal{X} \subset \mathbb{R}^p$ for $p \leq \infty$. The model (3.10) yields

$$L(\boldsymbol{\theta}) = \sum Y_i \langle \boldsymbol{\Psi}_i, \boldsymbol{\theta} \rangle - \phi\big(\langle \boldsymbol{\Psi}_i, \boldsymbol{\theta} \rangle\big).$$

Here again $\sum = \sum_{i=1}^n$. The corresponding Fisher information operator reads as

$$\mathbb{F}(\boldsymbol{\theta}) = -\nabla^2 L(\boldsymbol{\theta}) = \sum \phi''\big(\langle \boldsymbol{\Psi}_i, \boldsymbol{\theta} \rangle\big) \boldsymbol{\Psi}_i \otimes \boldsymbol{\Psi}_i \geq 0,$$

because ϕ is strictly convex. This yields $(\mathcal{L})$. Define $\varepsilon_i = Y_i - I\!\!E Y_i$. Then the stochastic component of the log-likelihood is linear in $\boldsymbol{\theta}$ and (E) is fulfilled with

$$\zeta(\boldsymbol{\theta}) = L(\boldsymbol{\theta}) - I\!\!E L(\boldsymbol{\theta}) = \sum \varepsilon_i \langle \boldsymbol{\Psi}_i, \boldsymbol{\theta} \rangle, \qquad \nabla\zeta = \sum \varepsilon_i \boldsymbol{\Psi}_i.$$

We assume a number of regularity conditions similar to the log-density case.

$(\boldsymbol{\psi_\infty})$ It holds $\| \psi_j \, I\!\!I_\mathcal{X} \|_\infty \leq C_\psi \, j^{1/2}$ for $j \leq p$.

$(\boldsymbol{\Theta})$ $\Theta \subseteq \mathcal{B}_s(1)$ for $s > 1$. This condition and $(\boldsymbol{\psi_\infty})$ yield by the Cauchy-Schwarz inequality, for any $x \in \mathcal{X}$ and $\boldsymbol{\theta} \in \Theta$

$$\left| \langle \boldsymbol{\Psi}(x), \boldsymbol{\theta} \rangle \right| = \left(\sum_{j=1}^{p} \theta_j \psi_j(x) \right)^2 \leq \left(\sum_{j=1}^{p} \theta_j^2 j^{2s} \right) \left(\sum_{j=1}^{p} C_\psi^2 \, j^{-2s+1} \right) \leq \frac{C_\psi^2}{2s-2} \stackrel{\text{def}}{=} C_\Psi.$$

$(\boldsymbol{\phi^{(k)}})$ $\phi(\cdot)$ is a smooth function with a continues fourth derivative on the interval $[-C_\Psi, C_\Psi]$ and the second derivative $\phi''(\cdot)$ fulfills for some $C_{1,\phi} \leq 1 \leq C_{2,\phi}$

$$C_{1,\phi} \, \phi''(0) \leq \phi''(t) \leq C_{2,\phi} \, \phi''(0), \quad |t| \leq C_\Psi.$$

$(\boldsymbol{\varepsilon_i})$ There are $\varrho > 0$, $\sigma_{\max}$, and $\nu_0 \geq 1$ such that $\varepsilon_i = Y_i - I\!\!E Y_i$ satisfy with $\sigma_i^2 = I\!\!E \varepsilon_i^2$

$$\max_{i \leq n} \sigma_i^2 \leq \sigma_{\max}^2, \qquad \sup_{|\lambda| \leq \varrho} \max_{i \leq n} \log I\!\!E \exp(\lambda \varepsilon_i) \leq \frac{\nu_0^2 \lambda^2 \sigma_{\max}^2}{2}.$$

$(\boldsymbol{\Psi u})$ It holds for all $\boldsymbol{u} \in I\!\!R^{m^*}$ and for a fixed constant C_f

$$\frac{1}{n} \sum_{i=1}^{n} \langle \boldsymbol{\Psi}_i, \boldsymbol{u} \rangle^4 \leq \left\{ \frac{C_f^2}{n} \sum_{i=1}^{n} \langle \boldsymbol{\Psi}_i, \boldsymbol{u} \rangle^2 \right\}^2.$$

$(\mathbb{F})$ For some $C_\mathbb{F} > 0$

$$\inf_{\boldsymbol{u} \in I\!\!R^{m^*}} \frac{\langle \mathbb{F}(\boldsymbol{\theta}^*) \boldsymbol{u}, \boldsymbol{u} \rangle}{n \| \boldsymbol{u} \|^2} \geq C_\mathbb{F}^{-2}.$$

Further, define

$$V^2 \stackrel{\text{def}}{=} \mathrm{Var}(\nabla\zeta) = \sum \sigma_i^2 \, \boldsymbol{\Psi}_i \otimes \boldsymbol{\Psi}_i,$$

$$H^2 = \sum \sigma_{\max}^2 \, \boldsymbol{\Psi}_i \otimes \boldsymbol{\Psi}_i,$$

Under the correct model specification $Y_i \sim P_{\langle \boldsymbol{\Psi}_i, \boldsymbol{\theta}^* \rangle}$, it holds $\sigma_i^2 = \phi''(\langle \boldsymbol{\Psi}_i, \boldsymbol{\theta}^* \rangle)$ and $V^2 = \mathbb{F}(\boldsymbol{\theta}^*)$. Let us fix some Gaussian prior $\mathcal{N}(0, G^{-2})$. As previously, we focus on a m-truncation or m^*-truncated (s, w)-smooth prior with $m^* = n^{1/3}/\log n$. In the first case, $m \leq m^*$. For any $\boldsymbol{\theta} \in \Theta$, conditions $(\boldsymbol{\Theta})$ and $(\mathbb{F})$ imply

$$\mathrm{p}_G(\boldsymbol{\theta}) = \mathrm{tr}\{(\mathbb{F}(\boldsymbol{\theta}) + G^2)^{-1} H^2\} \lesssim m,$$

where $m = (n/w)^{1/(2s)}$ for the (s, w)-prior.

Theorem 3.4. *Suppose* (ψ_∞), $(\boldsymbol{\Theta})$, $(\phi^{(k)})$, (ε_i), $(\boldsymbol{\Psi}u)$, *and* $(\mathbb{F})$ *for the model* (3.10). *Then all the expansions in* (3.7) *for the penalized MLE* $\widetilde{\boldsymbol{\theta}}_G$ *and the properties of the posterior* $\boldsymbol{\vartheta}_G \,|\, \boldsymbol{Y}$ *listed in Theorem 3.2 continue to hold provided that* $m^3 \ll n$. *Under the prior choice with* $m = w = n^{1/(2s+1)}$, *the results of Corollary 3.3 hold as well.*

A   Tools

A.1   Gaussian comparison

Let $\mathbb{H}$ be a Hilbert space and $\Sigma_{\boldsymbol{\xi}}$ be a covariance operator of an arbitrary Gaussian random element in $\mathbb{H}$. By $\{\lambda_{k\boldsymbol{\xi}}\}_{k \geq 1}$ we denote the set of its eigenvalues arranged in the non-increasing order, i.e. $\lambda_{1\boldsymbol{\xi}} \geq \lambda_{2\boldsymbol{\xi}} \geq \ldots$, and let $\boldsymbol{\lambda}_{\boldsymbol{\xi}} \stackrel{\mathrm{def}}{=} \mathrm{diag}(\lambda_{j\boldsymbol{\xi}})_{j=1}^\infty$. Note that $\sum_{j=1}^\infty \lambda_{j\boldsymbol{\xi}} < \infty$. Introduce the following quantities

$$\Lambda_{k\boldsymbol{\xi}}^2 \stackrel{\mathrm{def}}{=} \sum_{j=k}^\infty \lambda_{j\boldsymbol{\xi}}^2, \quad k = 1, 2,$$

Theorem A.1 (Götze et al. (2019)). *Let* $\boldsymbol{\xi}$ *and* $\boldsymbol{\eta}$ *be Gaussian elements in* $\mathbb{H}$ *with zero mean and covariance operators* $\Sigma_{\boldsymbol{\xi}}$ *and* $\Sigma_{\boldsymbol{\eta}}$ *respectively. Then for any* $\boldsymbol{a} \in \mathbb{H}$

$$\sup_{x>0} |I\!P(\|\boldsymbol{\xi} - \boldsymbol{a}\| \leq x) - I\!P(\|\boldsymbol{\eta}\| \leq x)|$$

$$\lesssim \left(\frac{1}{(\Lambda_{1\boldsymbol{\xi}} \Lambda_{2\boldsymbol{\xi}})^{1/2}} + \frac{1}{(\Lambda_{1\boldsymbol{\eta}} \Lambda_{2\boldsymbol{\eta}})^{1/2}} \right) \left(\|\boldsymbol{\lambda}_{\boldsymbol{\xi}} - \boldsymbol{\lambda}_{\boldsymbol{\eta}}\|_1 + \|\boldsymbol{a}\|^2 \right).$$

Moreover, assume that

$$3\|\Sigma_{\boldsymbol{\xi}}\|^2 \leq \|\Sigma_{\boldsymbol{\xi}}\|_{\mathrm{Fr}}^2 \quad and \quad 3\|\Sigma_{\boldsymbol{\eta}}\|^2 \leq \|\Sigma_{\boldsymbol{\eta}}\|_{\mathrm{Fr}}^2.$$

Then for any $\boldsymbol{a} \in \mathbb{H}$

$$\sup_{x>0} |I\!P(\|\boldsymbol{\xi} - \boldsymbol{a}\| \leq x) - I\!P(\|\boldsymbol{\eta}\| \leq x)| \lesssim \left(\frac{1}{\|\Sigma_{\boldsymbol{\xi}}\|_{\mathrm{Fr}}} + \frac{1}{\|\Sigma_{\boldsymbol{\eta}}\|_{\mathrm{Fr}}} \right) \left(\|\boldsymbol{\lambda}_{\boldsymbol{\xi}} - \boldsymbol{\lambda}_{\boldsymbol{\eta}}\|_1 + \|\boldsymbol{a}\|^2 \right).$$

A.2 Deviation bounds for Gaussian quadratic forms

The next result explains the concentration effect of $\langle B\boldsymbol{\xi}, \boldsymbol{\xi} \rangle$ for a centered Gaussian vector $\boldsymbol{\xi} \sim \mathcal{N}(0, H^2)$ and a symmetric trace operator B in $\mathbb{R}^p$, $p \leq \infty$. We use a version from Laurent and Massart (2000). For completeness, we present a simple proof of the bound.

Theorem A.2. *Let* $\boldsymbol{\xi} \sim \mathcal{N}(0, H^2)$ *be a Gaussian element in* $\mathbb{R}^p$ *and* B *be symmetric non-negative such that* $W = H^{-1} B H$ *is a trace operator in* $\mathbb{R}^p$. *Then with* $\mathtt{p} = \operatorname{tr}(W)$, $\mathtt{v}^2 = \operatorname{tr}(W^2)$, *and* $\lambda = \|W\|$, *it holds for each* $\mathtt{x} \geq 0$

$$\mathbb{P}\Big(\langle B\boldsymbol{\xi}, \boldsymbol{\xi} \rangle > z^2(W, \mathtt{x}) \Big) \leq \mathrm{e}^{-\mathtt{x}},$$

$$z(W, \mathtt{x}) \stackrel{\text{def}}{=} \sqrt{ \mathtt{p} + 2\mathtt{v}\,\mathtt{x}^{1/2} + 2\lambda\mathtt{x} } \,. \tag{A.1}$$

It also implies

$$\mathbb{P}\big(\|B^{1/2}\boldsymbol{\xi}\| > \mathtt{p}^{1/2} + (2\lambda\mathtt{x})^{1/2} \big) \leq \mathrm{e}^{-\mathtt{x}}.$$

If B *is symmetric but non necessarily positive then*

$$\mathbb{P}\big(\big| \langle B\boldsymbol{\xi}, \boldsymbol{\xi} \rangle - \mathtt{p} \big| > 2\mathtt{v}\,\mathtt{x}^{1/2} + 2\lambda\mathtt{x} \big) \leq 2\mathrm{e}^{-\mathtt{x}}.$$

Proof. W.l.o.g. assume that $H = I_p$ and $\lambda = \|W\| = 1$. Then $\boldsymbol{\xi} = \boldsymbol{\gamma} \sim \mathcal{N}(0, I_p)$ is standard Gaussian. We apply the exponential Chebyshev inequality: with $\mu > 0$

$$\mathbb{P}\Big(\langle B\boldsymbol{\gamma}, \boldsymbol{\gamma} \rangle > z^2 \Big) \leq \mathbb{E} \exp\Big(\mu \langle B\boldsymbol{\gamma}, \boldsymbol{\gamma} \rangle / 2 - \mu z^2 / 2 \Big).$$

Given $\mathtt{x} > 0$, fix $\mu < 1$ by the equation

$$\frac{\mu}{1 - \mu} = \frac{2\sqrt{\mathtt{x}}}{\mathtt{v}} \quad \text{or} \quad \mu^{-1} = 1 + \frac{\mathtt{v}}{2\sqrt{\mathtt{x}}}\,. \tag{A.2}$$

Let λ_j be ordered eigenvalues of W, $1 = \lambda_1 \geq \lambda_2 \geq \lambda_3 \ldots$. It holds with $\mathtt{p} = \operatorname{tr} W$

$$\log \mathbb{E}\Big\{ \frac{\mu}{2}\big(\langle W\boldsymbol{\gamma}, \boldsymbol{\gamma} \rangle - \mathtt{p} \big) \Big\} = \log \det(I - \mu B)^{-1/2} - \frac{\mu\,\mathtt{p}}{2} = -\frac{1}{2} \sum_{j \leq p} \big\{ \log(1 - \mu\lambda_j) + \mu\lambda_j \big\}$$

$$= \frac{1}{2} \sum_{j=1}^{p} \sum_{k=2}^{\infty} \frac{(\mu\lambda_j)^k}{k} \leq \sum_{j=1}^{p} \frac{(\mu\lambda_j)^2}{4} \sum_{k=0}^{\infty} (\mu\lambda_j)^k \leq \sum_{j=1}^{p} \frac{\mu^2 \lambda_j^2}{4(1 - \mu)} = \frac{\mu^2 \mathtt{v}^2}{4(1 - \mu)}\,. \tag{A.3}$$

It remains to check that the choice μ by (A.2) and $z = z(W, \mathtt{x})$ yields

$$\frac{\mu^2 \mathtt{v}^2}{4(1 - \mu)} - \frac{\mu(z^2 - \mathtt{p})}{2} = \frac{\mu^2 \mathtt{v}^2}{4(1 - \mu)} - \mu\big(\mathtt{v}\sqrt{\mathtt{x}} + \mathtt{x} \big) = \mu\Big\{ \frac{\mathtt{v}\sqrt{\mathtt{x}}}{2} - \mathtt{v}\sqrt{\mathtt{x}} - \mathtt{x} \Big\} = -\mathtt{x} \tag{A.4}$$

as required. □

As a special case, we present a bound for the chi-squared distribution corresponding to $B = H^2 = I_p$, $p < \infty$. Then $\text{tr}(W) = p$, $\text{tr}(W^2) = p$ and $\lambda(W) = 1$.

Corollary A.3. *Let $\boldsymbol{\gamma}$ be a standard normal vector in $I\!\!R^p$. Then for any $\mathtt{x} > 0$*

$$I\!\!P\big(\|\boldsymbol{\gamma}\|^2 \geq p + 2\sqrt{p\,\mathtt{x}} + 2\mathtt{x}\big) \leq \mathrm{e}^{-\mathtt{x}},$$

$$I\!\!P\big(\|\boldsymbol{\gamma}\| \geq \sqrt{p} + \sqrt{2\mathtt{x}}\big) \leq \mathrm{e}^{-\mathtt{x}},$$

$$I\!\!P\big(\|\boldsymbol{\gamma}\|^2 \leq p - 2\sqrt{p\,\mathtt{x}}\big) \leq \mathrm{e}^{-\mathtt{x}}.$$

A.3 Deviation bounds for non-Gaussian quadratic forms

This section collects some probability bounds for non-Gaussian quadratic forms starting from the subgaussian case. Then we extend the result to the case of exponential tails. Let $\boldsymbol{\xi}$ be a random vector in $I\!\!R^p$, $p \leq \infty$ satisfying $I\!\!E\boldsymbol{\xi} = 0$. We suppose that there exists an operator H in $I\!\!R^p$ such that

$$\log I\!\!E \exp\big(\langle \boldsymbol{u}, H^{-1}\boldsymbol{\xi}\rangle\big) \leq \frac{\|\boldsymbol{u}\|^2}{2}, \qquad \boldsymbol{u} \in I\!\!R^p. \tag{A.5}$$

In the Gaussian case, one obviously takes $H^2 = \text{Var}(\boldsymbol{\xi})$. In general, $H^2 \geq \text{Var}(\boldsymbol{\xi})$. We consider a quadratic form $\langle B\boldsymbol{\xi}, \boldsymbol{\xi}\rangle$, where $\boldsymbol{\xi}$ satisfies (A.5) and B is a given symmetric non-negative operator in $I\!\!R^p$ such that $B \leq H^{-2}$ and $W = HBH$ is a trace operator:

$$\mathtt{p} = \text{tr}\big(W\big) < \infty.$$

Denote also

$$\mathtt{v}^2 \overset{\text{def}}{=} \text{tr}(W^2).$$

We show that under these conditions, the quadratic form $\langle B\boldsymbol{\xi}, \boldsymbol{\xi}\rangle$ follows the same deviation bound $I\!\!P\big(\langle B\boldsymbol{\xi}, \boldsymbol{\xi}\rangle \geq z^2(W, \mathtt{x})\big) \leq \mathrm{e}^{-\mathtt{x}}$ with $z^2(W, \mathtt{x})$ from (A.1) as in the Gaussian case.

Theorem A.4. *Suppose* (A.5). *Let* $\mathtt{p} = \text{tr}\,W < \infty$ *for* $W = HBH$. *Then*

$$I\!\!P\big(\langle B\boldsymbol{\xi}, \boldsymbol{\xi}\rangle > z(W, \mathtt{x})\big) \leq \mathrm{e}^{-\mathtt{x}}.$$

Proof. For any $\mu < 1$, we use the identity

$$I\!\!E \exp\big(\mu\langle B\boldsymbol{\xi}, \boldsymbol{\xi}\rangle/2\big) = I\!\!E I\!\!E_{\boldsymbol{\gamma} \sim \mathcal{N}(0,I)} \exp\big(\mu^{1/2}\langle HB^{1/2}\boldsymbol{\gamma}, H^{-1}\boldsymbol{\xi}\rangle\big)$$

Application of Fubini's theorem and (A.5) yields

$$\mathbb{E}\exp\big(\mu\langle B\boldsymbol{\xi},\boldsymbol{\xi}\rangle/2\big) \leq \exp\Big(\frac{\mu^2\operatorname{tr}W^2}{4(1-\mu)} + \frac{\mu\operatorname{tr}W}{2}\Big).$$

Further we proceed as in the Gaussian case.   □

Now we turn to the main case of light exponential tails of $\boldsymbol{\xi}$. Namely, we suppose that $\mathbb{E}\boldsymbol{\xi} = 0$ and for some fixed $\mathtt{g} > 0$

$$\log\mathbb{E}\exp\big(\langle\boldsymbol{u}, H^{-1}\boldsymbol{\xi}\rangle\big) \leq \frac{\|\boldsymbol{u}\|^2}{2}, \qquad \boldsymbol{u}\in\mathbb{R}^p, \|\boldsymbol{u}\|\leq\mathtt{g}, \tag{A.6}$$

for some self-adjoint operator H in $\mathbb{R}^p$, $H \geq I_p$. In fact, it suffices to assume that

$$\sup_{\|\boldsymbol{u}\|\leq\mathtt{g}}\mathbb{E}\exp\big(\langle\boldsymbol{u}, H^{-1}\boldsymbol{\xi}\rangle\big) \leq \mathtt{C}.$$

Then one can use the fact that existence of the exponential moment $\mathbb{E}\mathrm{e}^{\lambda_0\xi}$ for a centered random variable ξ and some fixed λ_0 implies that the moment generating function $f_\xi(\lambda) \stackrel{\text{def}}{=} \log\mathbb{E}\mathrm{e}^{\lambda\xi}$ is analytic in $\lambda \in (0,\lambda_0)$ with $f_\xi(0) = f'_\xi(0) = 0$ and hence, it can be well majorated by a quadratic function in a smaller interval $[0,\lambda_1]$ for $\lambda_1 < \lambda_0$; see Golubev and Spokoiny (2009).

Remind $W = HBH$. By normalization, one can easily reduce the study to the case

$$\|W\| = 1.$$

Let $\mathtt{p} = \operatorname{tr}(W)$, $\mathtt{v}^2 = \operatorname{tr}(W^2)$, and $\mu(\mathtt{x})$ be defined by $\mu(\mathtt{x}) = \big(1 + \frac{\mathtt{v}}{2\sqrt{\mathtt{x}}}\big)^{-1}$; see (A.2). Obviously $\mu(\mathtt{x})$ grows with $\mathtt{x}$. Define the value $\mathtt{x}_c$ by the equation

$$\frac{\mathtt{g} - \sqrt{\mathtt{p}\,\mu(\mathtt{x}_c)}}{\mu(\mathtt{x}_c)} = z(W,\mathtt{x}_c) + 1. \tag{A.7}$$

The left hand-side here decreases with $\mathtt{x}$, while the right hand-side is increasing in $\mathtt{x}$. Therefore, the solution exists and unique. Also denote $\mu_c = \mu(\mathtt{x}_c)$ and

$$\mathtt{g}_c = \mathtt{g} - \sqrt{\mathtt{p}\mu_c}. \tag{A.8}$$

Obviously $\mathtt{g} - \sqrt{\mathtt{p}} \leq \mathtt{g}_c \leq \mathtt{g}$. Moreover, as $z(W,\mathtt{x}) \leq \sqrt{\mathtt{p}} + \sqrt{2\mathtt{x}}$, we derive that

$$\sqrt{2\mathtt{x}_c} \geq \mathtt{g}_c/\mu_c - \sqrt{\mathtt{p}}.$$

Our results implicitly assume that $\mathtt{x}_c$ is large, that is, $\mathtt{g} \gg 2\sqrt{\mathtt{p}}$.

Theorem A.5. *Let* (A.6) *hold and let* B *be such that* $W = HBH$ *satisfies* $\|W\| = 1$ *and* $\mathtt{p} = \operatorname{tr}(W) < \infty$. *Define* $\mathtt{x}_c$ *by* (A.7) *and* $\mathtt{g}_c$ *by* (A.8), *and suppose* $\mathtt{g}_c \geq 1$. *Then for any* $\mathtt{x} > 0$

$$I\!P\big(\langle B\boldsymbol{\xi}, \boldsymbol{\xi}\rangle \geq z_c^2(W, \mathtt{x})\big) \leq 2e^{-\mathtt{x}} + e^{-\mathtt{x}_c}\, I\!I(\mathtt{x} < \mathtt{x}_c), \tag{A.9}$$

where $z_c(W, \mathtt{x})$ *is defined by*

$$z_c(W, \mathtt{x}) \stackrel{\text{def}}{=} \begin{cases} \sqrt{\mathtt{p} + 2\mathtt{v}\,\mathtt{x}^{1/2} + 2\mathtt{x}}, & \mathtt{x} \leq \mathtt{x}_c, \\ \mathtt{g}_c/\mu_c + 2(\mathtt{x} - \mathtt{x}_c)/\mathtt{g}_c, & \mathtt{x} > \mathtt{x}_c. \end{cases} \tag{A.10}$$

The value $z(W, \mathtt{x}) = \sqrt{\mathtt{p} + 2\mathtt{v}\,\mathtt{x}^{1/2} + 2\mathtt{x}}$ *can be upper bounded by* $\sqrt{\mathtt{p}} + \sqrt{2\mathtt{x}}$:

$$z_c(W, \mathtt{x}) \leq \begin{cases} \sqrt{\mathtt{p}} + \sqrt{2\mathtt{x}}, & \mathtt{x} \leq \mathtt{x}_c, \\ z_c + 2(\mathtt{x} - \mathtt{x}_c)/\mathtt{g}_c, & \mathtt{x} > \mathtt{x}_c. \end{cases}$$

Depending on the value $\mathtt{x}$, we have two types of tail behavior of the quadratic form $\|\boldsymbol{\xi}\|^2$. For $\mathtt{x} \leq \mathtt{x}_c$, we have essentially the same deviation bounds as in the Gaussian case with the extra-factor two in the deviation probability. For $\mathtt{x} > \mathtt{x}_c$, we switch to the special regime driven by the exponential moment condition (A.6). Usually $\mathtt{g}^2$ is a large number (of order n in the i.i.d. setup) and the second term in (A.9) can be simply ignored.

As a corollary, we state the result for the norm of $\boldsymbol{\xi} \in I\!R^p$ corresponding to the case $H^{-2} = B = I_p$ and $p < \infty$. Then

$$\mathtt{p} = p = \mathtt{v}^2 = p.$$

Corollary A.6. *Let* (A.6) *hold with* $H = I_p$ *and* $\mathtt{g} \geq 2\sqrt{p}$. *Then for each* $\mathtt{x} > 0$

$$I\!P\big(\|\boldsymbol{\xi}\| \geq z_c(p, \mathtt{x})\big) \leq 2e^{-\mathtt{x}} + e^{-\mathtt{x}_c}\, I\!I(\mathtt{x} < \mathtt{x}_c),$$

where $z_c(p, \mathtt{x})$ *is defined by*

$$z_c(p, \mathtt{x}) \stackrel{\text{def}}{=} \begin{cases} \big(p + 2\sqrt{p\,\mathtt{x}} + 2\mathtt{x}\big)^{1/2}, & \mathtt{x} \leq \mathtt{x}_c, \\ \mathtt{g}_c/\mu_c + 2\mathtt{g}_c^{-1}(\mathtt{x} - \mathtt{x}_c), & \mathtt{x} > \mathtt{x}_c. \end{cases}$$

Proof of Theorem A.5. First we consider the most interesting case $\mathtt{x} \leq \mathtt{x}_c$. We expect to get Gaussian type deviation bounds for such $\mathtt{x}$. The main tool for the proof is the following lemma.

Lemma A.7. *Let* $\mu \in (0,1)$ *and* $\mathfrak{z}(\mu) = \mathsf{g}/\mu - \sqrt{\mathsf{p}/\mu} > 0$. *Then* (A.6) *implies*

$$\mathbb{E} \exp\big(\mu\langle B\boldsymbol{\xi}, \boldsymbol{\xi}\rangle/2\big) \, \mathbb{I}\big(\|HB\boldsymbol{\xi}\| \leq \mathfrak{z}(\mu)\big) \leq 2 \exp\Big(\frac{\mu^2 \mathsf{v}^2}{4(1-\mu)} + \frac{\mu\,\mathsf{p}}{2}\Big). \qquad (A.11)$$

Proof. Let us fix for a moment some $\boldsymbol{\xi} \in \mathbb{R}^p$ and $\mu < 1$ and define

$$\boldsymbol{a} = H^{-1}\boldsymbol{\xi}, \qquad \Sigma = \mu W = \mu H B H.$$

Consider the Gaussian measure $\mathbb{P}_{\boldsymbol{a},\Sigma} = \mathcal{N}(\boldsymbol{a}, \Sigma^{-1})$, and let $\boldsymbol{U} \sim \mathcal{N}(0, \Sigma^{-1})$. By the Girsanov formula

$$\log \frac{d\mathbb{P}_{\boldsymbol{a},\Sigma}}{d\mathbb{P}_{0,\Sigma}}(\boldsymbol{u}) = \langle \Sigma\boldsymbol{a}, \boldsymbol{u}\rangle - \frac{1}{2}\langle \Sigma\boldsymbol{a}, \boldsymbol{a}\rangle$$

and for any set $A \in \mathbb{R}^p$

$$\mathbb{P}_{\boldsymbol{a},\Sigma}(A) = \mathbb{P}_{0,\Sigma}(A - \boldsymbol{a}) = \mathbb{E}_{0,\Sigma}\Big[\exp\Big\{\langle \Sigma\boldsymbol{U}, \boldsymbol{a}\rangle - \frac{1}{2}\langle \Sigma\boldsymbol{a}, \boldsymbol{a}\rangle\Big\} \mathbb{I}(A)\Big].$$

Now we select $A = \big\{\boldsymbol{u}\colon \|\Sigma\boldsymbol{u}\| \leq \mathsf{g}\big\}$. Under $\mathbb{P}_{0,\Sigma}$, one can represent $\Sigma\boldsymbol{U} = \Sigma^{1/2}\boldsymbol{\gamma} = \mu^{1/2}HB^{1/2}\boldsymbol{\gamma}$ with a standard Gaussian $\boldsymbol{\gamma}$. Therefore,

$$\mathbb{P}_{0,\Sigma}(A - \boldsymbol{a}) = \mathbb{P}_{\boldsymbol{\gamma}\sim\mathcal{N}(0,I)}\big(\|\Sigma^{1/2}(\boldsymbol{\gamma} - \Sigma^{1/2}\boldsymbol{a})\| \leq \mathsf{g}\big)$$

$$\geq \mathbb{P}_{\boldsymbol{\gamma}\sim\mathcal{N}(0,I)}\big(\|\Sigma^{1/2}\boldsymbol{\gamma}\| \leq \mathsf{g} - \|\Sigma\boldsymbol{a}\|\big).$$

We now use that $\mathbb{P}_{\boldsymbol{\gamma}\sim\mathcal{N}(0,I)}\big(\|\Sigma^{1/2}\boldsymbol{\gamma}\|^2 \leq \operatorname{tr}(\Sigma)\big) \geq 1/2$ with $\operatorname{tr}(\Sigma) = \mu \operatorname{tr}(W) = \mu\,\mathsf{p}$. Therefore, the condition $\|\Sigma\boldsymbol{a}\| + \sqrt{\mu\,\mathsf{p}} \leq \mathsf{g}$ implies in view of $\langle \Sigma\boldsymbol{a}, \boldsymbol{a}\rangle = \mu\langle B\boldsymbol{\xi}, \boldsymbol{\xi}\rangle$

$$1/2 \leq \mathbb{P}_{\boldsymbol{a},\Sigma}(A) = \mathbb{E}_{0,\Sigma}\Big[\exp\Big\{\langle \Sigma\boldsymbol{U}, H^{-1}\boldsymbol{\xi}\rangle - \mu\langle B\boldsymbol{\xi}, \boldsymbol{\xi}\rangle/2\Big\} \mathbb{I}(\|\Sigma\boldsymbol{U}\| \leq \mathsf{g})\Big]$$

or

$$\exp\big(\mu\langle B\boldsymbol{\xi}, \boldsymbol{\xi}\rangle/2\big) \, \mathbb{I}\big(\|\Sigma H^{-1}\boldsymbol{\xi}\| \leq \mathsf{g} - \sqrt{\mu\,\mathsf{p}}\big)$$
$$\leq 2\mathbb{E}_{0,\Sigma}\Big[\exp\Big\{\langle \Sigma\boldsymbol{U}, H^{-1}\boldsymbol{\xi}\rangle \, \mathbb{I}(\|\Sigma\boldsymbol{U}\| \leq \mathsf{g})\Big].$$

We now take the expectation of the each side of this equation w.r.t. $\boldsymbol{\xi}$, change the integration order, and use (A.6) yielding

$$\mathbb{E} \exp\big(\mu\langle B\boldsymbol{\xi}, \boldsymbol{\xi}\rangle/2\big) \, \mathbb{I}\big(\|\Sigma H^{-1}\boldsymbol{\xi}\| \leq \mathsf{g} - \sqrt{\mu\,\mathsf{p}}\big) \leq 2\mathbb{E}_{0,\Sigma} \exp\big(\|\Sigma\boldsymbol{U}\|^2/2\big)$$

$$= 2\mathbb{E}_{\boldsymbol{\gamma}\sim\mathcal{N}(0,I)} \exp\big(\mu\|W^{1/2}\boldsymbol{\gamma}\|^2/2\big) = 2\det\big(\boldsymbol{I} - \mu W\big)^{-1/2}.$$

We also use that for any $\mu > 0$

$$\log \det \left(I - \mu W\right)^{-1/2} - \frac{\mu \operatorname{tr} W}{2} \le \frac{\mu^2 \operatorname{tr} W^2}{4(1 - \mu)};$$

see (A.3), and the first statement follows in view of $\Sigma H^{-1} \boldsymbol{\xi} = \mu H B \boldsymbol{\xi}$. $\qquad \square$

The use of $\mu = \mu_{\mathbf{x}}$ from (A.2) in (A.11) yields similarly to the proof of Theorem A.2

$$I\!\!P\Big(\langle B\boldsymbol{\xi}, \boldsymbol{\xi}\rangle > z^2(W, \mathbf{x}), \ \|HB\boldsymbol{\xi}\| \le \mathfrak{z}(\mu_{\mathbf{x}})\Big) \le 2\mathrm{e}^{-\mathbf{x}}. \tag{A.12}$$

It remains to consider the probability of large deviation $I\!\!P\big(\|HB\boldsymbol{\xi}\| > \mathfrak{z}(\mu_{\mathbf{x}})\big)$.

Lemma A.8. *For any* $\mathbf{x}_c > 0$ *such that* $z(W, \mathbf{x}_c) + 1 \le \mathsf{g}_c/\mu_c$, *it holds with* $\mu_c = \big\{1 + \mathtt{v}/(2\sqrt{\mathbf{x}_c})\big\}^{-1}$ *and* $z_c = \mathfrak{z}(\mu_c) = \mathsf{g}/\mu_c - \sqrt{\mathsf{p}/\mu_c}$

$$I\!\!P\big(\|HB\boldsymbol{\xi}\| > z_c\big) \le I\!\!P\big(\langle B\boldsymbol{\xi}, \boldsymbol{\xi}\rangle > z_c^2\big) \le \mathrm{e}^{-\mathbf{x}_c}.$$

Proof. It follows due to (A.2) and (A.4) for any $\mu \le \mu_c$

$$\frac{\mu^2 \mathtt{v}^2}{4(1 - \mu)} + \frac{\mu \, \mathsf{p}}{2} \le \frac{\mu_c^2 \mathtt{v}^2}{4(1 - \mu_c)} + \frac{\mu_c \, \mathsf{p}}{2} \le \frac{\mu_c z^2(W, \mathbf{x}_c)}{2} - \mathbf{x}_c,$$

where the right hand-side does not depend on μ. We now define $\eta^2 = \langle B\boldsymbol{\xi}, \boldsymbol{\xi}\rangle$ and use that $\|HB\boldsymbol{\xi}\| \le \|B^{1/2}\boldsymbol{\xi}\| = \eta$. Then by (A.11)

$$I\!\!E \exp(\mu \eta^2/2) \, I\!\!I\big(\eta \le \mathfrak{z}(\mu)\big) \le \exp\big\{\mu_c z^2(W, \mathbf{x}_c)/2 - \mathbf{x}_c\big\}, \tag{A.13}$$

Define the inverse function $\mu(\mathfrak{z}) = \mathfrak{z}^{-1}(\mu)$. For any $\mathfrak{z} \ge z_c$, it follows from (A.13) with $\mu = \mu(\mathfrak{z})$

$$I\!\!E \exp\big\{\mu(\mathfrak{z})(\mathfrak{z} - 1)^2/2\big\} \, I\!\!I\big(\mathfrak{z} - 1 \le \eta \le \mathfrak{z}\big) \le \exp\big\{\mu_c z^2(W, \mathbf{x}_c)/2 - \mathbf{x}_c\big\}$$

yielding

$$I\!\!P\big(\mathfrak{z} - 1 \le \eta \le \mathfrak{z}\big) \le \exp\Big(-\mu(\mathfrak{z})\,(\mathfrak{z} - 1)^2/2 + \mu_c z^2(W, \mathbf{x}_c)/2 - \mathbf{x}_c\Big)$$

and hence,

$$I\!\!P\big(\eta > \mathfrak{z}\big) \le \int_{\mathfrak{z}}^{\infty} \exp\big\{-\mu(z)(z - 1)^2/2 + \mu_c z^2(W, \mathbf{x}_c)/2 - \mathbf{x}_c\big\} dz$$

Further, $\mu \mathfrak{z}(\mu) = \mathsf{g} - \sqrt{\mathsf{p}\mu}$ and

$$\mathsf{g}_c = \mu_c z_c \le \mu \mathfrak{z}(\mu) \le \mathsf{g}, \quad \mu \le \mu_c.$$

This implies the same bound for the inverse function:

$$\mathsf{g}_c \leq \mathfrak{z}\,\mu(\mathfrak{z}) \leq \mathsf{g}, \quad \mathfrak{z} \geq z_c,$$

and

$$\mathbb{P}(\eta > \mathfrak{z}) \leq \int_{\mathfrak{z}}^{\infty} \exp\{-\mu(z)(z^2/2 - z) + \mu_c z^2(W, \mathtt{x}_c)/2 - \mathtt{x}_c\}dz$$

$$\leq \int_{\mathfrak{z}}^{\infty} \exp\{-\mathsf{g}_c\, z/2 + \mathsf{g} + \mu_c\, z^2(W, \mathtt{x}_c)/2 - \mathtt{x}_c\}dz$$

$$\leq \frac{2}{\mathsf{g}_c} \exp\{-\mathsf{g}_c\, \mathfrak{z}/2 + \mathsf{g} + \mu_c\, z^2(W, \mathtt{x}_c)/2 - \mathtt{x}_c\}. \qquad (A.14)$$

Conditions $\mathsf{g}_c z_c = \mu_c^{-1}\mathsf{g}_c^2 \geq \mu_c\{z(W, \mathtt{x}_c) + 1\}^2$ and $\mathsf{g}_c \geq 1$ ensure that $\mathbb{P}(\eta > z_c) \leq e^{-\mathtt{x}_c}$. $\qquad \square$

Remind that $\mathtt{x}_c$ is the largest $\mathtt{x}$-value ensuring the condition $\mathsf{g}_c \geq z(W, \mathtt{x}_c) + 1$. We also use that for $\mathtt{x} \leq \mathtt{x}_c$, it holds $\mathfrak{z}(\mu_{\mathtt{x}}) \geq \mathfrak{z}(\mu_c) = z_c$. Therefore, by (A.12) and Lemma A.8

$$\mathbb{P}\big(\langle B\boldsymbol{\xi}, \boldsymbol{\xi}\rangle \geq z^2(W, \mathtt{x})\big) \leq \mathbb{P}\big(\langle B\boldsymbol{\xi}, \boldsymbol{\xi}\rangle \geq z^2(W, \mathtt{x}), \|HB\boldsymbol{\xi}\| \leq \mathfrak{z}(\mu_{\mathtt{x}})\big) + \mathbb{P}\big(\langle B\boldsymbol{\xi}, \boldsymbol{\xi}\rangle \geq z_c^2\big)$$

$$\leq 2e^{-\mathtt{x}} + e^{-\mathtt{x}_c}.$$

Finally we consider $\mathtt{x} > \mathtt{x}_c$. Applying (A.14) yields by $\mathfrak{z} \geq z_c$

$$\mathbb{P}(\eta > \mathfrak{z}) \leq \frac{2}{\mu_c\, z_c} \exp\{-\mu_c\, z_c^2/2 + \mathsf{g} + \mu_c\, z^2(W, \mathtt{x}_c)/2 - \mathtt{x}_c\} \exp\{-\mu_c\, z_c(\mathfrak{z} - z_c)/2\}$$

$$\leq e^{-\mathtt{x}_c} \exp\{-\mathsf{g}_c(\mathfrak{z} - z_c)/2\}.$$

The choice $\mathfrak{z}$ by

$$\mathsf{g}_c(\mathfrak{z} - z_c)/2 = \mathtt{x} - \mathtt{x}_c$$

ensures the desired bound.

$\qquad \square$

A.4   Taylor expansions

Here we collect some useful bounds for various Taylor-type expansions for a smooth function. Let f be a four time differentiable function on $\mathbb{R}^p$. Here $p \leq \infty$. By

$f^{(k)}(\boldsymbol{x}, \boldsymbol{u})$ we denote the m th directional derivative at $\boldsymbol{x}$:

$$f^{(k)}(\boldsymbol{x}, \boldsymbol{u}) \stackrel{\text{def}}{=} \frac{d^k}{dt^k} f(\boldsymbol{x} + t\boldsymbol{u})\Big|_{t=0}.$$

In particular, $f'(\boldsymbol{x}, \boldsymbol{u}) = \langle \nabla f(\boldsymbol{x}), \boldsymbol{u} \rangle$ and $f''(\boldsymbol{x}, \boldsymbol{u}) = \langle \nabla^2 f(\boldsymbol{x})\, \boldsymbol{u}, \boldsymbol{u} \rangle$. Below we assume that some open set $\mathcal{X} \subseteq \mathbb{R}^p$ is fixed, and, in addition, for each $\boldsymbol{x} \in \mathcal{X}$, and a centrally symmetric convex set $\mathcal{U}(\boldsymbol{x})$ are fixed and

$$\frac{1}{k!}\big|f^{(k)}(\boldsymbol{x}, \boldsymbol{u})\big| = \delta_k(\boldsymbol{x}, \boldsymbol{u}) \leq \delta_k, \quad \boldsymbol{x} \in \mathcal{X}, \boldsymbol{u} \in \mathcal{U}, \quad k = 3, 4 \tag{A.15}$$

for some constants δ_k depending on $\mathcal{X}$ and $\mathcal{U}$. All bounds will be given in terms of δ_3 and δ_4. The construction can be extended by making $\mathcal{U}$ dependent on $\boldsymbol{x} \in \mathcal{X}$ at cost of more complicated notation.

Lemma A.9. *Suppose* (A.15) *with* $\delta_k \leq 1$ *for* $k = 3, 4$. *Then for any point* $\boldsymbol{x} \in \mathcal{X}$

$$\left| \frac{1}{2} \left(e^{f(\boldsymbol{x}+\boldsymbol{u})-f(\boldsymbol{x})-f'(\boldsymbol{x},\boldsymbol{u})} + e^{f(\boldsymbol{x}-\boldsymbol{u})-f(\boldsymbol{x})+f'(\boldsymbol{x},\boldsymbol{u})} \right) - e^{f''(\boldsymbol{x},\boldsymbol{u})/2} \right|$$
$$\leq e^{f''(\boldsymbol{x},\boldsymbol{u})/2} \left(4\delta_3^2 + 4\delta_4 \right). \tag{A.16}$$

Furthermore,

$$\left| e^{f(\boldsymbol{x}+\boldsymbol{u})-f(\boldsymbol{x})-f'(\boldsymbol{x},\boldsymbol{u})} - e^{f''(\boldsymbol{x},\boldsymbol{u})/2} \right| \leq \delta_3\, e^{f''(\boldsymbol{x},\boldsymbol{u})/2}. \tag{A.17}$$

Proof. Taylor expansions of the forth order imply

$$f(\boldsymbol{x} + \boldsymbol{u}) - f(\boldsymbol{x}) - f'(\boldsymbol{x}, \boldsymbol{u}) - \frac{1}{2}f''(\boldsymbol{x}, \boldsymbol{u}) - \frac{1}{6}f^{(3)}(\boldsymbol{x}, \boldsymbol{u}) = \rho_1, \qquad |\rho_1| \leq \delta_4,$$

$$f(\boldsymbol{x} - \boldsymbol{u}) - f(\boldsymbol{x}) + f'(\boldsymbol{x}, \boldsymbol{u}) - \frac{1}{2}f''(\boldsymbol{x}, \boldsymbol{u}) + \frac{1}{6}f^{(3)}(\boldsymbol{x}, \boldsymbol{u}) = \rho_2, \qquad |\rho_2| \leq \delta_4.$$

Further, define $\varkappa = f^{(3)}(\boldsymbol{x}, \boldsymbol{u})/6$, so that $|\varkappa| \leq \delta_3 \leq 1$. Then

$$e^{f(\boldsymbol{x}+\boldsymbol{u})-f(\boldsymbol{x})-f'(\boldsymbol{x},\boldsymbol{u})} + e^{f(\boldsymbol{x}-\boldsymbol{u})-f(\boldsymbol{x})+f'(\boldsymbol{x},\boldsymbol{u})} - 2e^{f''(\boldsymbol{x},\boldsymbol{u})/2}$$

$$= e^{f''(\boldsymbol{x},\boldsymbol{u})/2} \left(e^{\varkappa+\rho_1} + e^{-\varkappa+\rho_2} - 2 \right).$$

The function

$$g(s) \stackrel{\text{def}}{=} \frac{1}{2}\exp\big(s\,\varkappa + \rho_1\big) + \frac{1}{2}\exp\big(-s\,\varkappa + \rho_2\big) - 1$$

fulfills

$$|g(0)| = \left| \frac{1}{2}e^{\rho_1} + \frac{1}{2}e^{\rho_2} - 1 \right| \le |\rho_1| + |\rho_2|,$$

$$|g'(0)| = \frac{1}{2}\left| \varkappa\left(e^{\rho_1} - e^{\rho_2}\right) \right| \le |\rho_1| + |\rho_2|$$

and for any $s \in [0,1]$ by simple algebra due to $|\varkappa| \le 1$ and $|\rho_k| \le 1$ for $m = 1, 2$

$$|g''(s)| = \frac{1}{2}\left| \varkappa^2 \left\{ \exp(s\,\varkappa + \rho_1) + \exp(-s\,\varkappa + \rho_2) \right\} \right|$$

$$\le \frac{|\varkappa|^2 e}{2}\left(e^{|\varkappa|} + e^{-|\varkappa|}\right) < 8|\varkappa|^2,$$

and thus

$$\left|g(1)\right| \le \sup_{s \in [0,1]} \left| g(0) + g'(0) + \frac{1}{2}g''(s) \right| \le 4|\varkappa|^2 + 2|\rho_1| + 2|\rho_2|,$$

and (A.16) follows. The bound (A.17) can be obtained in a similar way using the Taylor expansion of the third order. □

Now we study the modulus of continuity for the gradient $\nabla f(\boldsymbol{x})$ and the Hessian $\nabla^2 f(\boldsymbol{x})$.

Lemma A.10. *Suppose* (A.15) *with* $\delta_3 \le 1$. *Let* $\boldsymbol{x} \in \mathcal{X}$ *and* $\boldsymbol{u} \in \mathcal{U}$ *be such that* $\boldsymbol{x} + \boldsymbol{u} \in \mathcal{X}$. *Then, for any* $\boldsymbol{w} \in \mathcal{U}$

$$\left| \langle \boldsymbol{w}, \nabla f(\boldsymbol{x} + \boldsymbol{u}) - \nabla f(\boldsymbol{x}) - \nabla^2 f(\boldsymbol{x})\boldsymbol{u} \rangle \right| \le \mathtt{C}\delta_3,$$

$$\left| \langle \boldsymbol{w}, \{ \nabla^2 f(\boldsymbol{x} + \boldsymbol{u}) - \nabla^2 f(\boldsymbol{x}) \}\boldsymbol{w} \rangle \right| \le \mathtt{C}\delta_3. \tag{A.18}$$

Proof. Let us fix any $\boldsymbol{x}^\circ \in \mathcal{X}$ and $\boldsymbol{w}^\circ \in \mathcal{U}$ and define the function

$$g(t) \stackrel{\text{def}}{=} f(\boldsymbol{x}^\circ + t\boldsymbol{w}^\circ) + f(\boldsymbol{x}^\circ - t\boldsymbol{w}^\circ) - 2f(\boldsymbol{x}^\circ) - t^2 f''(\boldsymbol{x}^\circ, \boldsymbol{w}^\circ).$$

The Taylor expansion of the third order yields

$$\left| g(1) \right| = \left| f(\boldsymbol{x}^\circ + \boldsymbol{w}^\circ) + f(\boldsymbol{x}^\circ - \boldsymbol{w}^\circ) - 2f(\boldsymbol{x}^\circ) - f''(\boldsymbol{x}^\circ, \boldsymbol{w}^\circ) \right| \le 2\delta_3(\boldsymbol{x}^\circ, \boldsymbol{w}^\circ).$$

We apply this bound for $\boldsymbol{x}^\circ = \boldsymbol{x}$ and $\boldsymbol{x}^\circ = \boldsymbol{x} + \boldsymbol{u}$ and take the difference between them. This implies

$$\left| f''(\boldsymbol{x}, \boldsymbol{w}^\circ) - f''(\boldsymbol{x} + \boldsymbol{u}, \boldsymbol{w}^\circ) \right| \le \left| f(\boldsymbol{x} + \boldsymbol{w}^\circ) + f(\boldsymbol{x} - \boldsymbol{w}^\circ) - 2f(\boldsymbol{x}) \right.$$

$$\left. - f(\boldsymbol{x} + \boldsymbol{u} + \boldsymbol{w}^\circ) - f(\boldsymbol{x} + \boldsymbol{u} - \boldsymbol{w}^\circ) + 2f(\boldsymbol{x} + \boldsymbol{u}) \right| + 2\delta_3(\boldsymbol{x}, \boldsymbol{w}^\circ) + 2\delta_3(\boldsymbol{x} + \boldsymbol{u}, \boldsymbol{w}^\circ). \tag{A.19}$$

For given $\boldsymbol{x}, \boldsymbol{u}, \boldsymbol{w}$, and $\overline{\boldsymbol{x}} = \boldsymbol{x} + \boldsymbol{u}/2$, define

$$g(t) \stackrel{\text{def}}{=} f\big(\overline{\boldsymbol{x}} + t(\boldsymbol{u} + \boldsymbol{w})\big) - f\big(\overline{\boldsymbol{x}} - t(\boldsymbol{u} + \boldsymbol{w})\big)$$
$$+ f\big(\overline{\boldsymbol{x}} + t(\boldsymbol{u} - \boldsymbol{w})\big) - f\big(\overline{\boldsymbol{x}} - t(\boldsymbol{u} - \boldsymbol{w})\big) - 2f\big(\overline{\boldsymbol{x}} + t\boldsymbol{u}\big) + 2f\big(\overline{\boldsymbol{x}} - t\boldsymbol{u}\big).$$

It is straightforward to see that $g(0) = g'(0) = g''(0) = 0$. Moreover, in view of $\boldsymbol{u} \in \mathcal{U}$ and $(\boldsymbol{u} \pm \boldsymbol{w})/2 \in \mathcal{U}$, it holds $\delta_3(\overline{\boldsymbol{x}}, \boldsymbol{u}/2) = \delta_3(\overline{\boldsymbol{x}}, \boldsymbol{u})/8$ and for any $|t| \leq 1/2$

$$\frac{1}{6}\big|g^{(3)}(t)\big| \leq \frac{5\delta_3}{2}.$$

By Taylor expansion of the third order we derive

$$\big|g(1/2)\big| \leq \sup_{t \in [0,1]} \frac{1}{6}\big|g^{(3)}(t)\big| \leq \frac{5\delta_3}{2}.$$

Note that $g(1/2)$ is exactly the expression in the right hand-side of (A.19) with $\boldsymbol{w}^{\circ} = \boldsymbol{w}/2$. The use of $\delta_3(\boldsymbol{x}^{\circ}, \boldsymbol{w}^{\circ}) = \delta_3(\boldsymbol{x}^{\circ}, \boldsymbol{w})/8$ together with (A.19) yields (A.18) with $\mathtt{C} = 3$. □

Now we specify the result to the case of an elliptic set $\mathcal{U}$ of the form

$$\mathcal{U} = \big\{\boldsymbol{u} \colon \|Q\boldsymbol{u}\| \leq \mathtt{r}\big\} \tag{A.20}$$

for a positive invertible operator Q and $\mathtt{r} > 0$.

Lemma A.11. *Let* $\mathcal{U}$ *be given by* (A.20) *with* $Q > 0$, *and let* $\boldsymbol{x} \in \mathcal{X}$ *and* $\boldsymbol{u} \in \mathcal{U}$ *be such that* $\boldsymbol{x} + \boldsymbol{u} \in \mathcal{X}$. *Then*

$$\big\|Q^{-1}\big\{\nabla f(\boldsymbol{x} + \boldsymbol{u}) - \nabla f(\boldsymbol{x}) - \nabla^2 f(\boldsymbol{x})\boldsymbol{u}\big\}\big\| \leq \mathtt{C}\mathtt{r}^{-1}\delta_3, \tag{A.21}$$

$$\big\|Q^{-1}\big\{\nabla^2 f(\boldsymbol{x}) - \nabla^2 f(\boldsymbol{x} + \boldsymbol{u})\big\}Q^{-1}\big\| \leq \mathtt{C}\mathtt{r}^{-2}\delta_3.$$

Proof. For any $\boldsymbol{w} \in \mathcal{U}$, it holds by Lemma A.10

$$\big|\langle \boldsymbol{w}, \big\{\nabla^2 f(\boldsymbol{x} + \boldsymbol{u}) - \nabla^2 f(\boldsymbol{x})\big\}\boldsymbol{w}\rangle\big| = \big|\langle Q\boldsymbol{w}, Q^{-1}\big\{\nabla^2 f(\boldsymbol{x} + \boldsymbol{u}) - \nabla^2 f(\boldsymbol{x})\big\}Q^{-1}(Q\boldsymbol{w})\rangle\big| \leq \mathtt{C}\delta_3.$$

As this bound holds for all $\boldsymbol{w} \in \mathcal{U}$ with $\|Q\boldsymbol{w}\| \leq \mathtt{r}$, the result follows. □

The result of Lemma A.9 can be extended to the integral of $\mathrm{e}^{f(\boldsymbol{x}+\boldsymbol{u})}$ over $\boldsymbol{u} \in \mathcal{U}$.

Lemma A.12. *Let* $\mathcal{U}$ *be a subset in* $\mathbb{R}^p$. *Suppose* (A.15) *with* $\delta_k \leq 1$ *for* $k = 3, 4$. *Then for any point* $\boldsymbol{x} \in \mathcal{X}$ *and any centrally symmetric set* $A \subset \mathcal{U}$

$$\left|\int_A \mathrm{e}^{f(\boldsymbol{x}+\boldsymbol{u}) - f(\boldsymbol{x}) - f'(\boldsymbol{x}, \boldsymbol{u})}\, d\boldsymbol{u} - \int_A \mathrm{e}^{f''(\boldsymbol{x}, \boldsymbol{u})/2}\, d\boldsymbol{u}\right| \leq \diamondsuit \int_A \mathrm{e}^{f''(\boldsymbol{x}, \boldsymbol{u})/2}\, d\boldsymbol{u} \tag{A.22}$$

with $\diamondsuit = 4\delta_3^2 + 4\delta_4$ and for any vector $\boldsymbol{z}$

$$\left| \int_A \langle \boldsymbol{z}, \boldsymbol{u} \rangle^2 \mathrm{e}^{f(\boldsymbol{x}+\boldsymbol{u})-f(\boldsymbol{x})-f'(\boldsymbol{x},\boldsymbol{u})} \, d\boldsymbol{u} - \int_A \langle \boldsymbol{z}, \boldsymbol{u} \rangle^2 \mathrm{e}^{f''(\boldsymbol{x},\boldsymbol{u})/2} \, d\boldsymbol{u} \right|$$

$$\leq \diamondsuit \int_A \langle \boldsymbol{z}, \boldsymbol{u} \rangle^2 \mathrm{e}^{f''(\boldsymbol{x},\boldsymbol{u})/2} \, d\boldsymbol{u} . \tag{A.23}$$

If A is not centrally symmetric then

$$\left| \int_A \mathrm{e}^{f(\boldsymbol{x}+\boldsymbol{u})-f(\boldsymbol{x})-f'(\boldsymbol{x},\boldsymbol{u})} \, d\boldsymbol{u} - \int_A \mathrm{e}^{f''(\boldsymbol{x},\boldsymbol{u})/2} \, d\boldsymbol{u} \right| \leq \delta_3 \int_A \mathrm{e}^{f''(\boldsymbol{x},\boldsymbol{u})/2} \, d\boldsymbol{u} . \tag{A.24}$$

Proof. By symmetricity of $\mathcal{U}$, it holds

$$\int_A \mathrm{e}^{f(\boldsymbol{x}+\boldsymbol{u})-f(\boldsymbol{x})-f'(\boldsymbol{x},\boldsymbol{u})} \, d\boldsymbol{u} = \frac{1}{2} \int_A \left(\mathrm{e}^{f(\boldsymbol{x}+\boldsymbol{u})-f(\boldsymbol{x})-f'(\boldsymbol{x},\boldsymbol{u})} + \mathrm{e}^{f(\boldsymbol{x}-\boldsymbol{u})-f(\boldsymbol{x})+f'(\boldsymbol{x},\boldsymbol{u})} \right) d\boldsymbol{u},$$

and the first result is proved by (A.16). The same symmetricity arguments apply to (A.23). The final bound for any A follows from (A.17). □

The bound (A.25) can be specified to the case of a massive set $\mathcal{U}$. We assume that f is concave and $\mathsf{H}^2 \stackrel{\text{def}}{=} -\nabla^2 f(x) \geq 0$.

Lemma A.13. *Let $f(\cdot)$ be strictly concave with $\mathsf{H}^2 = -\nabla^2 f(\boldsymbol{x}) > 0$. Suppose (A.15) with $\delta_3 \leq 1$. For a linear operator Q, it holds*

$$\left\| \int_{\mathcal{U}} Q\boldsymbol{u} \, \mathrm{e}^{f(\boldsymbol{x}+\boldsymbol{u})-f(\boldsymbol{x})-f'(\boldsymbol{x},\boldsymbol{u})} \, d\boldsymbol{u} \right\| \leq \delta_3 \int_{\mathcal{U}} \|Q\boldsymbol{u}\| \mathrm{e}^{f''(\boldsymbol{x},\boldsymbol{u})/2} \, d\boldsymbol{u} , \tag{A.25}$$

Let also $\mathcal{U}$ be massive in the sense that

$$I\!P\left(\mathsf{H}^{-1}\boldsymbol{\gamma} \in \mathcal{U} \right) \geq 1/2 \tag{A.26}$$

with $\boldsymbol{\gamma}$ standard normal in $I\!R^p$. Then for any linear operator Q, it holds

$$\left\| \frac{\int_{\mathcal{U}} Q\boldsymbol{u} \, \mathrm{e}^{f(\boldsymbol{x}+\boldsymbol{u})-f(\boldsymbol{x})-f'(\boldsymbol{x},\boldsymbol{u})} \, d\boldsymbol{u}}{\int_{\mathcal{U}} \mathrm{e}^{f''(\boldsymbol{x},\boldsymbol{u})/2} \, d\boldsymbol{u}} \right\| \leq 2\delta_3 \, I\!E \|Q \, \mathsf{H}^{-1}\boldsymbol{\gamma}\| .$$

Proof. The bound (A.25) follows in a way similar to the proof of Lemma A.12 using (A.17) instead of (A.16). We apply (A.25) yielding in view of $f''(\boldsymbol{x},\boldsymbol{u}) = -\|\mathsf{H}\boldsymbol{u}\|^2$

$$\left\| \frac{\int_{\mathcal{U}} Q\boldsymbol{u} \, \mathrm{e}^{f(\boldsymbol{x}+\boldsymbol{u})-f(\boldsymbol{x})-f'(\boldsymbol{x},\boldsymbol{u})} \, d\boldsymbol{u}}{\int_{\mathcal{U}} \mathrm{e}^{f''(\boldsymbol{x},\boldsymbol{u})/2} \, d\boldsymbol{u}} \right\| \leq \delta_3 \frac{\int_{\mathcal{U}} \|Q\boldsymbol{u}\| \, \mathrm{e}^{f''(\boldsymbol{x},\boldsymbol{u})/2} \, d\boldsymbol{u}}{\int_{\mathcal{U}} \mathrm{e}^{f''(\boldsymbol{x},\boldsymbol{u})/2} \, d\boldsymbol{u}}$$

$$\leq 2\delta_3 \frac{\int_{\mathcal{U}} \|Q\boldsymbol{u}\| \, \mathrm{e}^{f''(\boldsymbol{x},\boldsymbol{u})/2} \, d\boldsymbol{u}}{\int \mathrm{e}^{f''(\boldsymbol{x},\boldsymbol{u})/2} \, d\boldsymbol{u}} \leq 2\delta_3 \, I\!E \|Q \, \mathsf{H}^{-1}\boldsymbol{\gamma}\|$$

as required. □

All the bounds presented above assume that $\mathcal{U}$ is a symmetric subset of $\mathbb{R}^p$, in particular, an ellipsoid centred at zero. Now we check what happens under a small departure from symmetricity.

Lemma A.14. *Let f be concave with $\mathsf{H}^2 = -\nabla^2 f(\boldsymbol{x})$. Let also $\mathcal{U}$ be centrally symmetric massive set; see* (A.26) *and let $\boldsymbol{a} \in \mathcal{U}$ be fixed and $\mathcal{U} + \boldsymbol{a} \subset \mathcal{E}(\mathsf{r}_0) = \{\boldsymbol{u} \colon \|\mathsf{H}\boldsymbol{u}\| \leq \mathsf{r}_0\}$. Suppose* (A.15) *with $\delta_k = \delta_k(\mathsf{r}_0) \leq 1$ for $k = 3, 4$. Then*

$$\left| \int_{\mathcal{U}} \mathrm{e}^{f(\boldsymbol{x}+\boldsymbol{u}+\boldsymbol{a})-f(\boldsymbol{x})-f'(\boldsymbol{x},\boldsymbol{u}+\boldsymbol{a})} \, d\boldsymbol{u} - \int_{\mathcal{U}} \mathrm{e}^{f''(\boldsymbol{x},\boldsymbol{u}+\boldsymbol{a})/2} \, d\boldsymbol{u} \right|$$

$$\lesssim \left\{ \Diamond(\mathsf{r}_0) + \left(\|\mathsf{H}\boldsymbol{a}\| + \|\mathsf{H}\boldsymbol{a}\|^2 \right) \delta_3(\mathsf{r}_0) \right\} \int_{\mathcal{U}} \mathrm{e}^{f''(\boldsymbol{x},\boldsymbol{u})/2} \, d\boldsymbol{u} .$$

Proof. Define

$$h(t) \stackrel{\mathrm{def}}{=} q_0 \int_{\mathcal{U}} \mathrm{e}^{f(\boldsymbol{x}+\boldsymbol{u}+t\boldsymbol{a})-f(\boldsymbol{x})-f'(\boldsymbol{x},\boldsymbol{u}+t\boldsymbol{a})} \, d\boldsymbol{u}$$

$$v(t) \stackrel{\mathrm{def}}{=} q_0 \int_{\mathcal{U}} \mathrm{e}^{f''(\boldsymbol{x},\boldsymbol{u}+t\boldsymbol{a})/2} \, d\boldsymbol{u}$$

with

$$q_0 \stackrel{\mathrm{def}}{=} \left(\int_{\mathcal{U}} \mathrm{e}^{f''(\boldsymbol{x},\boldsymbol{u})/2} \, d\boldsymbol{u} \right)^{-1} .$$

Then we have to bound the difference $h(t) - v(t)$ for $t \leq 1$. For this, we bound the first two derivatives of $h(t)$. By (A.22) Lemma A.12, it holds $|h(0) - v(0)| \leq \Diamond(\mathsf{r}_0)$. Further, in view of $f''(\boldsymbol{x}, \boldsymbol{u} + t\boldsymbol{a}) = -\|\mathsf{H}(\boldsymbol{u} + t\boldsymbol{a})\|^2$

$$h'(0) = q_0 \int_{\mathcal{U}} \langle \boldsymbol{a}, \nabla f(\boldsymbol{x} + \boldsymbol{u}) - \nabla f(\boldsymbol{x}) \rangle \, \mathrm{e}^{f(\boldsymbol{x}+\boldsymbol{u})-f(\boldsymbol{x})-f'(\boldsymbol{x},\boldsymbol{u})} \, d\boldsymbol{u} , \qquad (A.27)$$

$$v'(0) = q_0 \int_{\mathcal{U}} \langle \mathsf{H}\boldsymbol{a}, \mathsf{H}\boldsymbol{u} \rangle \, \mathrm{e}^{f''(\boldsymbol{x},\boldsymbol{u})/2} \, d\boldsymbol{u} = 0$$

because $\mathcal{U}$ is centrally symmetric. By (A.21) of Lemma A.11 and (A.22) of Lemma A.12

$$\left| \int_{\mathcal{U}} \langle \boldsymbol{a}, \nabla f(\boldsymbol{x} + \boldsymbol{u}) - \nabla f(\boldsymbol{x}) - \nabla^2 f(\boldsymbol{x})\boldsymbol{u} \rangle \, \mathrm{e}^{f(\boldsymbol{x}+\boldsymbol{u})-f(\boldsymbol{x})-f'(\boldsymbol{x},\boldsymbol{u})} \, d\boldsymbol{u} \right|$$

$$\leq \mathsf{C} \, \|\mathsf{H}\boldsymbol{a}\| \mathsf{r}_0^{-1} \delta_3(\mathsf{r}_0) \int_{\mathcal{U}} \mathrm{e}^{f(\boldsymbol{x}+\boldsymbol{u})-f(\boldsymbol{x})-f'(\boldsymbol{x},\boldsymbol{u})} \, d\boldsymbol{u}$$

$$\leq \mathsf{C} \, \|\mathsf{H}\boldsymbol{a}\| \mathsf{r}_0^{-1} \delta_3(\mathsf{r}_0) \left(1 + \Diamond(\mathsf{r}_0) \right) \int_{\mathcal{U}} \mathrm{e}^{f''(\boldsymbol{x},\boldsymbol{u})/2} \, d\boldsymbol{u} \qquad (A.28)$$

for $\Diamond(\mathtt{r}_0) = 4\delta_3^2(\mathtt{r}_0) + 4\delta_4(\mathtt{r}_0)$. Further we use $f''(\boldsymbol{x}, \boldsymbol{u}) = -\|\mathsf{H}\boldsymbol{u}\|^2$, $-\langle \boldsymbol{a}, \nabla^2 f(\boldsymbol{x})\boldsymbol{u} \rangle = \langle \mathsf{H}\boldsymbol{a}, \mathsf{H}\boldsymbol{u} \rangle$, and it follows by Lemma A.13 that

$$\left| q_0 \int_{\mathcal{U}} \langle \boldsymbol{a}, \nabla^2 f(\boldsymbol{x})\boldsymbol{u} \rangle \mathrm{e}^{f(\boldsymbol{x}+\boldsymbol{u})-f(\boldsymbol{x})-f'(\boldsymbol{x},\boldsymbol{u})} \, d\boldsymbol{u} \right| \leq 2\delta_3(\mathtt{r}_0) \, \mathit{I\!E} \big| \langle \mathsf{H}\boldsymbol{a}, \boldsymbol{\gamma} \rangle \big| \leq 2\delta_3(\mathtt{r}_0) \|\mathsf{H}\boldsymbol{a}\|.$$

Putting this together with (A.27), (A.28) yields

$$\big| h'(0) \big| \lesssim \delta_3(\mathtt{r}_0) \|\mathsf{H}\boldsymbol{a}\|.$$

For the second derivative,

$$h''(t) = q_0 \int_{\mathcal{U}} \Big\{ \big\langle \boldsymbol{a}, \nabla f(\boldsymbol{x}+\boldsymbol{u}+t\boldsymbol{a}) - \nabla f(\boldsymbol{x}) \big\rangle^2 + \big\langle \boldsymbol{a}, \nabla^2 f(\boldsymbol{x}+\boldsymbol{u}+t\boldsymbol{a})\boldsymbol{a} \big\rangle \Big\}$$
$$\times \mathrm{e}^{f(\boldsymbol{x}+\boldsymbol{u}+t\boldsymbol{a})-f(\boldsymbol{x})-f'(\boldsymbol{x},\boldsymbol{u}+t\boldsymbol{a})} \, d\boldsymbol{u} \, .$$

Similarly, by the use of $f''(\boldsymbol{x}, \boldsymbol{u}) = -\|\mathsf{H}\boldsymbol{u}\|^2$ we derive

$$v''(t) = q_0 \int_{\mathcal{U}} \Big\{ \big\langle \boldsymbol{a}, \mathsf{H}^2(\boldsymbol{u}+t\boldsymbol{a}) \big\rangle^2 - \big\langle \boldsymbol{a}, \mathsf{H}^2\boldsymbol{a} \big\rangle \Big\} \mathrm{e}^{-\|\mathsf{H}(\boldsymbol{u}+t\boldsymbol{a})\|^2/2} \, d\boldsymbol{u} \, .$$

Now by (A.25) of Lemma A.12

$$\big| \langle \boldsymbol{a}, \nabla^2 f(\boldsymbol{x}+\boldsymbol{u}+t\boldsymbol{a})\boldsymbol{a} \rangle - \langle \boldsymbol{a}, \nabla^2 f(\boldsymbol{x})\boldsymbol{a} \rangle \big|$$
$$= \big| \langle \mathsf{H}\boldsymbol{a}, \mathsf{H}^{-1}\{\nabla^2 f(\boldsymbol{x}+\boldsymbol{u}+t\boldsymbol{a}) - \nabla^2 f(\boldsymbol{x})\}\mathsf{H}^{-1}\mathsf{H}\boldsymbol{a} \rangle \big| \lesssim \|\mathsf{H}\boldsymbol{a}\|^2 \mathtt{r}_0^{-2} \delta_3(\mathtt{r}_0) \, .$$

Similarly

$$\big| \langle \boldsymbol{a}, \nabla f(\boldsymbol{x}+\boldsymbol{u}+t\boldsymbol{a}) - \nabla f(\boldsymbol{x}) + \mathsf{H}^2(\boldsymbol{u}+t\boldsymbol{a}) \rangle \big| \lesssim \|\mathsf{H}\boldsymbol{a}\| \mathtt{r}_0^{-1} \delta_3(\mathtt{r}_0)$$

and by $\|\mathsf{H}(\boldsymbol{u}+t\boldsymbol{a})\| \leq \mathtt{r}_0$ it holds $\big| \langle \boldsymbol{a}, \mathsf{H}^2(\boldsymbol{u}+t\boldsymbol{a}) \rangle \big| \leq \|\mathsf{H}\boldsymbol{a}\| \mathtt{r}_0$ and

$$\big\langle \boldsymbol{a}, \nabla f(\boldsymbol{x}+\boldsymbol{u}+t\boldsymbol{a}) - \nabla f(\boldsymbol{x}) \big\rangle^2 - \big\langle \boldsymbol{a}, \mathsf{H}^2(\boldsymbol{u}+t\boldsymbol{a}) \big\rangle^2$$
$$\lesssim \|\mathsf{H}\boldsymbol{a}\|^2 \, \mathtt{r}_0^{-2} \delta_3^2(\mathtt{r}_0) + 2\|\mathsf{H}\boldsymbol{a}\| \, \mathtt{r}_0^{-1} \delta_3(\mathtt{r}_0) \|\mathsf{H}\boldsymbol{a}\| \mathtt{r}_0$$
$$\lesssim \|\mathsf{H}\boldsymbol{a}\|^2 \, \delta_3(\mathtt{r}_0).$$

We conclude that

$$h''(t) = q_0 \int_{\mathcal{U}} \Big\{ \big\langle \boldsymbol{a}, \mathsf{H}^2(\boldsymbol{u}+t\boldsymbol{a}) \big\rangle^2 - \big\langle \boldsymbol{a}, \mathsf{H}^2\boldsymbol{a} \big\rangle + \tau(\boldsymbol{a}, \boldsymbol{u}, t) \Big\} \mathrm{e}^{f(\boldsymbol{x}+\boldsymbol{u}+t\boldsymbol{a})-f(\boldsymbol{x})-f'(\boldsymbol{x},\boldsymbol{u}+t\boldsymbol{a})} \, d\boldsymbol{u}$$

where $|\tau(\boldsymbol{a}, \boldsymbol{u}, t)| \lesssim \|\mathsf{H}\boldsymbol{a}\|^2 \delta_3(\mathtt{r}_0)$. Lemma A.9 helps to bound

$$\big| h''(t) - v''(t) \big| \lesssim \|\mathsf{H}\boldsymbol{a}\|^2 \delta_3(\mathtt{r}_0)$$

uniformly in $|t| \le 1$. This yields with some $\rho \in [0, 1]$

$$\left| h(1) - v(1) \right| \le \left| h(0) - v(0) \right| + \left| h'(0) - v'(0) \right| + \left| h''(\rho) - v''(\rho) \right|/2$$

$$\lesssim \Diamond(\mathbf{r}_0) + \left(\|\mathsf{H}\boldsymbol{a}\| + \|\mathsf{H}\boldsymbol{a}\|^2 \right) \delta_3(\mathbf{r}_0)$$

which completes the proof. □

A.5 Concavity and tail bounds

Let $f(\boldsymbol{x})$ be a function on $I\!\!R^p$. Previous results describe the local behavior of $f(\boldsymbol{x} + \boldsymbol{u})$ for $\boldsymbol{u} \in \mathcal{U}$ under local smoothness conditions. Now we derive some upper bounds on $f(\boldsymbol{x} + \boldsymbol{u})$ for $\boldsymbol{u}$ large using that f is concave. More precisely, we fix $\boldsymbol{x}$ and $\boldsymbol{u}$ and bound the values $f(\boldsymbol{x} + t\boldsymbol{u}) - f(\boldsymbol{x}) - t f'(\boldsymbol{x}, \boldsymbol{u})$ for $\boldsymbol{u} \in \mathcal{U}$ and large t.

Lemma A.15. *Suppose* (A.15) *with* $\delta_m \le 1$ *for* $m = 3, 4$. *Let* $\boldsymbol{x} + \mathcal{U} \subset \mathcal{X}$. *Let the function* $f(\boldsymbol{x} + t\boldsymbol{u})$ *be concave in* t. *Then it holds for any* $\boldsymbol{u} \in \mathcal{U}$ *and for* $t > 1$

$$f(\boldsymbol{x} + t\boldsymbol{u}) - f(\boldsymbol{x}) - \langle \nabla f(\boldsymbol{x}), \boldsymbol{u} \rangle t \le \left(t - \frac{1}{2} \right) \left\{ \langle \nabla^2 f(\boldsymbol{x}) \boldsymbol{u}, \boldsymbol{u} \rangle - 3\delta_3 \right\}. \qquad \text{(A.29)}$$

Proof. The Taylor expansion of the third order for $g(t) = f(\boldsymbol{x} + t\boldsymbol{u})$ at $t = 0$ yields

$$\left| g(1) - g(0) - g'(0) - \frac{1}{2} g''(0) \right| \le \delta_3.$$

Similarly one obtains

$$g'(1) - g'(0) = g'(1) - g'(0) - g''(0) + g''(0) \le g''(0) + 3\delta_3.$$

Concavity of $g(\cdot)$ implies

$$g(t) - g(1) \le (t - 1) g'(1).$$

We summarize that

$$g(t) - g(0) - t g'(0) = g(t) - g(1) - (t - 1) g'(1) + (t - 1) \left\{ g'(1) - g'(0) \right\} + g(1) - g(0) - g'(0)$$

$$\le (t - 1) \left\{ g''(0) + 3\delta_3 \right\} + \frac{1}{2} g''(0) + \delta_3$$

$$\le (t - 1/2) \left\{ g''(0) + 3\delta_3 \right\}.$$

This implies the assertion in view of $g''(0) = \langle \nabla^2 f(\boldsymbol{x}) \boldsymbol{u}, \boldsymbol{u} \rangle$. □

Now we specify the result of Lemma A.15 for the elliptic set $\mathcal{U}(\mathtt{r}_0)$ defined by the condition $-\langle \nabla^2 f(\boldsymbol{x})\boldsymbol{u}, \boldsymbol{u}\rangle \leq \mathtt{r}_0^2$. We write $\delta_3(\mathtt{r}_0)$ in place of $\delta_3(\mathcal{X}, \mathcal{U}(\mathtt{r}_0))$. We aim at bounding from above the value $f(\boldsymbol{x}+\boldsymbol{u}) - f(\boldsymbol{x})$ for $\boldsymbol{u}$ with $-\langle \nabla^2 f(\boldsymbol{x})\boldsymbol{u}, \boldsymbol{u}\rangle = \mathtt{r}^2 > \mathtt{r}_0^2$.

Lemma A.16. *Consider $\boldsymbol{x} \in \mathcal{X}$ and $\mathcal{U} = \mathcal{U}(\mathtt{r}_0) = \{\boldsymbol{u}\colon -\langle \nabla^2 f(\boldsymbol{x})\boldsymbol{u}, \boldsymbol{u}\rangle \leq \mathtt{r}_0^2\}$. Let $f(\boldsymbol{x}+\boldsymbol{u})$ be concave in $\boldsymbol{u}$. Then for any $\boldsymbol{u}$ with $-\langle \nabla^2 f(\boldsymbol{x})\boldsymbol{u}, \boldsymbol{u}\rangle = \mathtt{r}^2 > \mathtt{r}_0^2$, it holds*

$$f(\boldsymbol{x}+\boldsymbol{u}) - f(\boldsymbol{x}) - \langle \nabla f(\boldsymbol{x}), \boldsymbol{u}\rangle \leq -(\mathtt{r}\mathtt{r}_0 - \mathtt{r}_0^2/2)\{1 - 3\mathtt{r}_0^{-2}\delta_3(\mathtt{r}_0)\}. \qquad \text{(A.30)}$$

Proof. Define $t = \mathtt{r}/\mathtt{r}_0$ and $\boldsymbol{u}^\circ = \boldsymbol{u}\mathtt{r}_0/\mathtt{r}$, so that $-\langle \nabla^2 f(\boldsymbol{x})\boldsymbol{u}, \boldsymbol{u}\rangle = \mathtt{r}_0^2$ and $\boldsymbol{u}^\circ \in \mathcal{U}(\mathtt{r}_0)$. Then it holds by (A.29)

$$\begin{aligned}
f(\boldsymbol{x}+\boldsymbol{u}) - f(\boldsymbol{x}) - \langle \nabla f(\boldsymbol{x}), \boldsymbol{u}\rangle &= f(\boldsymbol{x}+t\boldsymbol{u}^\circ) - f(\boldsymbol{x}) - \langle \nabla f(\boldsymbol{x}), \boldsymbol{u}\rangle \\
&\leq -(\mathtt{r}/\mathtt{r}_0 - 1/2)\{\mathtt{r}_0^2 - 3\delta_3(\mathtt{r}_0)\} \\
&= -(\mathtt{r}\mathtt{r}_0 - \mathtt{r}_0^2/2)\{1 - 3\mathtt{r}_0^{-2}\delta_3(\mathtt{r}_0)\}
\end{aligned}$$

and the result follows. $\qquad\square$

The result is meaningful if $3\mathtt{r}_0^{-2}\delta_3(\mathtt{r}_0) < 1$. Then with $\mathtt{C}_0 = 1 - 3\mathtt{r}_0^{-2}\delta_3(\mathtt{r}_0)$, we obtain for any $\boldsymbol{u}$ with $-\langle \nabla^2 f(\boldsymbol{x})\boldsymbol{u}, \boldsymbol{u}\rangle = \mathtt{r}^2 > \mathtt{r}_0^2$

$$f(\boldsymbol{x}+\boldsymbol{u}) - f(\boldsymbol{x}) - \langle \nabla f(\boldsymbol{x}), \boldsymbol{u}\rangle \leq -\mathtt{C}_0(\mathtt{r}\mathtt{r}_0 - \mathtt{r}_0^2/2).$$

A.6   Gaussian integrals

Let $\mathcal{T}$ be a linear operator in $I\!\!R^p$, $p \leq \infty$, with $\|\mathcal{T}\|_{\mathrm{op}} \leq 1$. By $\mathcal{T}^\top$ we denote the adjoint operator for $\mathcal{T}$. Given positive $\mathtt{r}_0$ and $\mathtt{C}_0$, consider the following ratio

$$\frac{\int_{\|\mathcal{T}\boldsymbol{u}\|>\mathtt{r}_0} \exp\left(-\mathtt{C}_0\|\mathcal{T}\boldsymbol{u}\| + \frac{1}{2}\mathtt{C}_0\mathtt{r}_0^2 + \frac{1}{2}\|\mathcal{T}\boldsymbol{u}\|^2 - \frac{1}{2}\|\boldsymbol{u}\|^2\right)d\boldsymbol{u}}{\int_{\|\mathcal{T}\boldsymbol{u}\|\leq\mathtt{r}_0} \exp\left(-\frac{1}{2}\|\boldsymbol{u}\|^2\right)d\boldsymbol{u}}.$$

Obviously, one can rewrite this value as ratio of two expectations

$$\frac{I\!\!E\left\{\exp\left(-\mathtt{C}_0\mathtt{r}_0\|\mathcal{T}\boldsymbol{\gamma}\| + \frac{1}{2}\mathtt{C}_0\mathtt{r}_0^2 + \frac{1}{2}\|\mathcal{T}\boldsymbol{\gamma}\|^2\right) 1\!\!1\left(\|\mathcal{T}\boldsymbol{\gamma}\| > \mathtt{r}_0\right)\right\}}{I\!\!P\left(\|\mathcal{T}\boldsymbol{\gamma}\| \leq \mathtt{r}_0\right)},$$

where $\boldsymbol{\gamma} \sim \mathcal{N}(0, I_p)$. Note that without the linear term $-\mathtt{C}_0\|\mathcal{T}\boldsymbol{\gamma}\|$ in the exponent, the expectation in the numerator can be infinite. We aim at describing $\mathtt{r}_0$ and $\mathtt{C}_0$-values which ensure that the probability in denominator is close to one while the expectation in the numerator is small.

Lemma A.17. *Let $\mathcal{T}$ be a linear operator in $\mathbb{R}^p$ with $\|\mathcal{T}\|_{\mathrm{op}} \leq 1$. Define $\mathtt{p}_\tau = \mathrm{tr}(\mathcal{T}^\top \mathcal{T})$. For any $\mathtt{C}_0, \mathtt{r}_0$ with $1/2 < \mathtt{C}_0 \leq 1$ and $\mathtt{C}_0 \mathtt{r}_0 = 2\sqrt{\mathtt{p}_\tau} + \sqrt{\mathtt{x}}$ for $\mathtt{x} > 0$*

$$\mathbb{E}\left\{\exp\left(-\mathtt{C}_0 \mathtt{r}_0 \|\mathcal{T}\boldsymbol{\gamma}\| + \frac{\mathtt{C}_0 \mathtt{r}_0^2}{2} + \frac{1}{2}\|\mathcal{T}\boldsymbol{\gamma}\|^2\right) \mathrm{I\!I}(\|\mathcal{T}\boldsymbol{\gamma}\| > \mathtt{r}_0)\right\} \leq \mathtt{C}\mathrm{e}^{-(\mathtt{p}_\tau + \mathtt{x})/2}$$

and

$$\mathbb{P}\left(\|\mathcal{T}\boldsymbol{\gamma}\| \leq \mathtt{r}_0\right) \geq 1 - \exp\left\{-\frac{1}{2}(\mathtt{r}_0 - \sqrt{\mathtt{p}_\tau})^2\right\} \geq 1 - \mathrm{e}^{-(\mathtt{p}_\tau + \mathtt{x})/2}.$$

Remark A.1. The result applies even if the full dimension p is infinite and $\boldsymbol{\gamma}$ is a Gaussian element in a Hilbert space, provided that $\mathtt{p}_\tau = \mathrm{tr}(\mathcal{T}^\top \mathcal{T})$ is finite, that is, $\mathcal{T}^\top \mathcal{T}$ is a trace operator.

Proof. Define

$$\Phi(\mathtt{r}) \overset{\mathrm{def}}{=} \mathbb{P}\left(\|\mathcal{T}\boldsymbol{\gamma}\| \geq \mathtt{r}\right),$$

$$f(\mathtt{r}) \overset{\mathrm{def}}{=} \exp\left(-\mathtt{C}_0 \mathtt{r}_0 \mathtt{r} + \frac{\mathtt{C}_0 \mathtt{r}_0^2}{2} + \frac{\mathtt{r}^2}{2}\right).$$

Then

$$\mathbb{E}\left\{\exp\left(-\mathtt{C}_0 \mathtt{r}_0 \|\mathcal{T}\boldsymbol{\gamma}\| + \frac{\mathtt{C}_0 \mathtt{r}_0^2}{2} + \frac{1}{2}\|\mathcal{T}\boldsymbol{\gamma}\|^2\right) \mathrm{I\!I}(\|\mathcal{T}\boldsymbol{\gamma}\| > \mathtt{r}_0)\right\}$$

$$= -\int_{\mathtt{r}_0}^\infty f(\mathtt{r}) d\Phi(\mathtt{r}) = f(\mathtt{r}_0)\Phi(\mathtt{r}_0) + \int_{\mathtt{r}_0}^\infty f'(\mathtt{r})\Phi(\mathtt{r})\, d\mathtt{r}.$$

Now we use that $\Phi\left(\sqrt{\mathtt{p}_\tau} + \sqrt{2\mathtt{x}}\right) \leq \mathrm{e}^{-\mathtt{x}}$ for any $\mathtt{x} > 0$. This can be rewritten as

$$\Phi(\mathtt{r}) \leq \exp\left\{-\frac{1}{2}(\mathtt{r} - \sqrt{\mathtt{p}_\tau})^2\right\}$$

for $\mathtt{r} > \sqrt{\mathtt{p}_\tau}$. In particular, in view of $\mathtt{r}_0 \geq 2\sqrt{\mathtt{p}_\tau} + \sqrt{\mathtt{x}}$

$$f(\mathtt{r}_0)\Phi(\mathtt{r}_0) \leq \Phi(\mathtt{r}_0) \leq \exp\left\{-\frac{1}{2}(\mathtt{r}_0 - \sqrt{\mathtt{p}_\tau})^2\right\} \leq \exp\left\{-\frac{1}{2}(\mathtt{p}_\tau + \mathtt{x})\right\}.$$

Now we use that $f'(\mathtt{r}) = (\mathtt{r} - \mathtt{C}_0 \mathtt{r}_0)f(\mathtt{r})$ and

$$\int_{\mathtt{r}_0}^\infty f'(\mathtt{r})\Phi(\mathtt{r})\, d\mathtt{r} = \int_{\mathtt{r}_0}^\infty (\mathtt{r} - \mathtt{C}_0 \mathtt{r}_0) f(\mathtt{r})\Phi(\mathtt{r})\, d\mathtt{r}$$

$$\leq \int_{\mathtt{r}_0}^\infty (\mathtt{r} - \mathtt{C}_0 \mathtt{r}_0) \exp\left\{-\frac{1}{2}(\mathtt{r} - \sqrt{\mathtt{p}_\tau})^2 - \mathtt{C}_0 \mathtt{r}_0 \mathtt{r} + \frac{\mathtt{C}_0 \mathtt{r}_0^2}{2} + \frac{\mathtt{r}^2}{2}\right\} d\mathtt{r}$$

$$= \int_{\mathtt{r}_0}^\infty (\mathtt{r} - \mathtt{C}_0 \mathtt{r}_0) \exp\left\{-(\mathtt{C}_0 \mathtt{r}_0 - \sqrt{\mathtt{p}_\tau})\mathtt{r} + \frac{\mathtt{C}_0 \mathtt{r}_0^2}{2} - \frac{\mathtt{p}_\tau}{2}\right\} d\mathtt{r}$$

$$= \int_0^\infty (x + \mathtt{r}_0 - \mathtt{C}_0 \mathtt{r}_0) \exp\left\{-(\mathtt{C}_0 \mathtt{r}_0 - \sqrt{\mathtt{p}_\tau})(x + \mathtt{r}_0) + \frac{\mathtt{C}_0 \mathtt{r}_0^2}{2} - \frac{\mathtt{p}_\tau}{2}\right\} dx.$$

The use of $\int_0^\infty e^{-x} dx = \int_0^\infty x e^{-x} dx = 1$ yields

$$\int_{r_0}^\infty f'(r) \Phi(r) \, dr \leq \left(\frac{r_0 - C_0 r_0}{C_0 r_0 - \sqrt{p_\tau}} + \frac{1}{(C_0 r_0 - \sqrt{p_\tau})^2} \right) \exp\left\{ r_0 \sqrt{p_\tau} - \frac{C_0 r_0^2}{2} - \frac{p_\tau}{2} \right\}.$$

It remains to check that for $C_0 \in (1/2, 1)$ and $C_0 r_0 = 2\sqrt{p_\tau} + \sqrt{x}$

$$-r_0 \sqrt{p_\tau} + \frac{C_0 r_0^2}{2} + \frac{p_\tau}{2} \geq \frac{x + p_\tau}{2}.$$

The result follows. $\qquad\square$

Now we consider Gaussian integrals with an additional quadratic multiplier.

Lemma A.18. *Let $\mathcal{T}$ be a linear operator in $\mathbb{R}^p$ with $\|\mathcal{T}\|_{\mathrm{op}} \leq 1$. Let $z \in \mathbb{R}^\infty$ be a unit norm vector: $\|z\| = 1$. Define $p_\tau = \mathrm{tr}(\mathcal{T}^\top \mathcal{T})$. For any positive C_0, r_0 with $1/2 < C_0 \leq 1$ and $C_0 r_0 > 2\sqrt{p_\tau + 1} + \sqrt{x}$*

$$\mathbb{E}\left\{ |\langle z, \gamma \rangle|^2 \exp\left(-C_0 r_0 \|\mathcal{T}\gamma\| + \frac{C_0 r_0^2}{2} + \frac{1}{2}\|\mathcal{T}\gamma\|^2 \right) \mathbb{I}(\|\mathcal{T}\gamma\| > r_0) \right\}$$

$$\leq C e^{-(p_\tau + x)/2}. \tag{A.31}$$

Proof. Define $\mathcal{T}_z$ by $\mathcal{T}_z^\top \mathcal{T}_z = \mathcal{T}^\top \mathcal{T} + z \otimes z$. Obviously $\|\mathcal{T}_z \gamma\| \geq \|\mathcal{T}\gamma\|$, $|\langle z, \gamma \rangle| \leq \|\mathcal{T}_z \gamma\|$. Further, $r^2/2 - C_0 r_0 r$ grows in $r \geq r_0$ in view of $C_0 \leq 1$. Therefore,

$$|\langle z, \gamma \rangle|^2 \exp\left(-C_0 r_0 \|\mathcal{T}\gamma\| + \frac{C_0 r_0^2}{2} + \frac{1}{2}\|\mathcal{T}\gamma\|^2 \right) \mathbb{I}(\|\mathcal{T}\gamma\| > r_0)$$

$$\leq \|\mathcal{T}_z \gamma\|^2 \exp\left(-C_0 r_0 \|\mathcal{T}_z \gamma\| + \frac{C_0 r_0^2}{2} + \frac{1}{2}\|\mathcal{T}_z \gamma\|^2 \right) \mathbb{I}(\|\mathcal{T}_z \gamma\| > r_0)$$

Now we can follow the line of the proof of Lemma A.17. Consider

$$\Phi_\lambda(r) = \mathbb{P}\big(\|\mathcal{T}_z \gamma\| \geq r \big) \leq \exp\left\{ -\frac{1}{2}(r - \sqrt{p\lambda}) \right\},$$

$$f(r) \stackrel{\text{def}}{=} r^2 \exp\left(-C_0 r_0 r + \frac{C_0 r_0^2}{2} + \frac{r^2}{2} \right)$$

with $p_\lambda \stackrel{\text{def}}{=} \mathrm{tr}\, \mathcal{T}_z^\top \mathcal{T}_z = p + 1$. Then

$$\mathbb{E}\left\{ (\langle z, \gamma \rangle)^2 \exp\left(-C_0 r_0 \|\mathcal{T}\gamma\| + \frac{C_0 r_0^2}{2} + \frac{1}{2}\|\mathcal{T}\gamma\|^2 \right) \mathbb{I}(\|\mathcal{T}\gamma\| > r_0) \right\}$$

$$\leq -\int_{r_0}^\infty f(r) d\Phi_\lambda(r) = f(r_0)\Phi_\lambda(r_0) + \int_{r_0}^\infty f'(r)\Phi_\lambda(r) \, dr.$$

Now we can continue as in the proof of Lemma A.17. $\qquad\square$

The bound (A.31) can be easily extended to the case of a more general functional $Q\gamma$ in place of $\langle \lambda, \gamma \rangle$.

Lemma A.19. *Let $\mathcal{T}$ be a linear operator in $\mathbb{R}^p$ with $\|\mathcal{T}\|_{\mathrm{op}} \leq 1$ and $\mathtt{p}_\tau = \mathrm{tr}(\mathcal{T}^\top \mathcal{T}) < \infty$. Let A be a bounded linear operator with $\mathrm{tr}(A^\top A) < \infty$. For any positive $\mathtt{C}_0, \mathtt{r}_0$ with $1/2 < \mathtt{C}_0 \leq 1$ and $\mathtt{C}_0 \mathtt{r}_0 > 2\sqrt{\mathtt{p}_\tau + 1} + \sqrt{\mathtt{x}}$*

$$\mathbb{E}\Big\{\|A\gamma\|^2 \exp\Big(-\mathtt{C}_0 \mathtt{r}_0 \|\mathcal{T}\gamma\| + \frac{\mathtt{C}_0 \mathtt{r}_0^2}{2} + \frac{1}{2}\|\mathcal{T}\gamma\|^2\Big) \, \mathbb{I}\big(\|\mathcal{T}\gamma\| > \mathtt{r}_0\big)\Big\} \lesssim \mathrm{tr}(A^\top A)\mathrm{e}^{-(\mathtt{p}_\tau + \mathtt{x})/2}.$$

Proof. We use the Karhunen-Loeve decomposition of $A^\top A$:

$$\|A\gamma\|^2 = \sum_j \mu_j \langle z_j, \gamma \rangle^2$$

with orthogonal unit vectors z_j and $\sum_j \mu_j = \mathrm{tr}(A^\top A)$, and apply the result of Lemma A.18 to each term of this decomposition. $\qquad\square$

B Proofs of the main results

This section collects the proofs of the main theorems.

Proof of Proposition 2.4 The idea of the proof is to show that for each u with $\|D_G u\| = \mathtt{r}_G$, the derivative of the function $L_G(\boldsymbol{\theta}_G^* + t u)$ in t is negative for $|t| \geq 1$. This yields that the point of maximum of $L_G(\boldsymbol{\theta})$ cannot be outside of $\mathcal{A}_G(\mathtt{r}_G)$. Let us fix any u with $\|D_G u\| \leq \mathtt{r}_G$. We use the decomposition

$$L_G(\boldsymbol{\theta}_G^* + t u) - L_G(\boldsymbol{\theta}_G^*) = \langle \nabla\zeta, u \rangle t + \mathbb{E}L_G(\boldsymbol{\theta}_G^* + t u) - \mathbb{E}L_G(\boldsymbol{\theta}_G^*).$$

With $f(t) = \mathbb{E}L_G(\boldsymbol{\theta}_G^* + t u)$, it holds

$$\frac{d}{dt} L_G(\boldsymbol{\theta}_G^* + t u) = \langle \nabla\zeta, u \rangle + f'(t). \tag{B.1}$$

The bound (2.11) implies on $\Omega(\mathtt{x})$

$$\big|\langle \nabla\zeta, u \rangle\big| = \big|\langle D_G^{-1}\nabla\zeta, D_G u \rangle\big| \leq \mathtt{r}_G \, z(B_{V|G}, \mathtt{x}). \tag{B.2}$$

By definition of $\boldsymbol{\theta}_G^*$, it also holds $f'(0) = 0$. Condition $(\mathcal{L}_0)$ implies

$$\big|f'(t) - tf''(0)\big| = \big|f'(t) - f'(0) - tf''(0)\big| \leq 3t^2 \, \mathtt{r}_G^3 \, \tau_{3,H}.$$

For $t = 1$, we obtain

$$f'(1) \leq f''(0) + 3\mathtt{r}_G^3 \tau_{3,H} = -\langle D_G^2 \boldsymbol{u}, \boldsymbol{u}\rangle + 3\mathtt{r}_G^3 \tau_{3,H} = -\mathtt{r}_G^2 + 3\mathtt{r}_G^3 \tau_{3,H}.$$

If $3\tau_{3,H} \leq \rho$ for $\rho < 1$, then $f'(1) < 0$. Concavity of $f(t)$ and $f'(0) = 0$ imply that $f'(t)$ decreases in t for $t > 1$. Further, on $\Omega(\mathtt{x})$ by (B.2)

$$\frac{d}{dt}L_G(\boldsymbol{\theta}_G^* + t\boldsymbol{u})\big|_{t=1} \leq \langle \nabla\zeta, \boldsymbol{u}\rangle - \mathtt{r}_G^2 + 3\mathtt{r}_G^3 \tau_{3,H}$$

$$\leq \mathtt{r}_G \, z(B_G, \mathtt{x}) - \mathtt{r}_G^2 + 3\mathtt{r}_G^3 \tau_{3,H} \leq \mathtt{r}_G \, z(B_G, \mathtt{x}) - (1-\rho)\mathtt{r}_G^2 < 0$$

for $\mathtt{r}_G > (1-\rho)^{-1} z(B_G, \mathtt{x})$. As $\frac{d}{dt}L_G(\boldsymbol{\theta}_G^* + t\boldsymbol{u})$ decreases with $t \geq 1$ together with $f'(t)$ due to (B.1), the same applies to all such t. This implies the assertion.

Proof of Theorem 2.5   To show (2.17), we use that $\widetilde{\boldsymbol{\theta}}_G \in \mathcal{A}_G(\mathtt{r}_G)$ and $\nabla L_G(\widetilde{\boldsymbol{\theta}}_G) = 0$. Therefore,

$$L_G(\widetilde{\boldsymbol{\theta}}_G + \boldsymbol{u}) - L_G(\widetilde{\boldsymbol{\theta}}_G) = L_G(\widetilde{\boldsymbol{\theta}}_G + \boldsymbol{u}) - L_G(\widetilde{\boldsymbol{\theta}}_G) - \langle \nabla L_G(\widetilde{\boldsymbol{\theta}}_G), \boldsymbol{u}\rangle.$$

Let us fix any $\boldsymbol{\theta} \in \mathcal{A}_G(\mathtt{r}_G)$ and $\boldsymbol{u}$ with $\|D_G \boldsymbol{u}\| \leq \mathtt{r}_G$, and consider

$$f(t) = f(t, \boldsymbol{u}) \stackrel{\text{def}}{=} L_G(\boldsymbol{\theta} + t\boldsymbol{u}) - L_G(\boldsymbol{\theta}) - \langle \nabla L_G(\boldsymbol{\theta}), \boldsymbol{u}\rangle t.$$

As the stochastic term of $L(\boldsymbol{\theta})$ and thus, of $L_G(\boldsymbol{\theta})$ is linear in $\boldsymbol{\theta}$, it cancels in this expression, and it suffices to consider the deterministic part $I\!\!E L_G(\boldsymbol{\theta})$. Obviously $f(0) = 0$, $f'(0) = 0$. Moreover, $f''(0) = \langle \nabla^2 I\!\!E L_G(\boldsymbol{\theta}) \boldsymbol{u}, \boldsymbol{u}\rangle = -\langle D_G^2(\boldsymbol{\theta}) \boldsymbol{u}, \boldsymbol{u}\rangle < 0$. Taylor expansion of the third order implies

$$\left|f(1) - \frac{1}{2}f''(0)\right| \leq \left|\tau_{3,H}(\boldsymbol{\theta}', \boldsymbol{u})\right|, \quad \boldsymbol{\theta}' \in [\boldsymbol{\theta}, \boldsymbol{\theta} + \boldsymbol{u}].$$

In particular, for any $\boldsymbol{\theta} \in \mathcal{A}_G(\mathtt{r}_G)$

$$\left|I\!\!E L_G(\boldsymbol{\theta}_G^*) - I\!\!E L_G(\boldsymbol{\theta}) - \frac{1}{2}\left\|D_G(\boldsymbol{\theta} - \boldsymbol{\theta}_G^*)\right\|^2\right| \leq \mathtt{r}_G^3 \tau_{3,H}. \tag{B.3}$$

We now use that $\nabla L_G(\widetilde{\boldsymbol{\theta}}_G) = 0$ and by Proposition 2.4, $\boldsymbol{u} = \boldsymbol{\theta}_G^* - \widetilde{\boldsymbol{\theta}}_G$ fulfills $\|D_G \boldsymbol{u}\| \leq \mathtt{r}_G$ on $\Omega(\mathtt{x})$. Therefore, for $\boldsymbol{\theta} \in \mathcal{A}_G(\mathtt{r}_G)$

$$\left|L_G(\boldsymbol{\theta}) - L_G(\widetilde{\boldsymbol{\theta}}_G) - \frac{1}{2}\|\widetilde{D}_G(\boldsymbol{\theta} - \widetilde{\boldsymbol{\theta}}_G)\|^2\right|$$

$$= \left|L_G(\boldsymbol{\theta}) - L_G(\widetilde{\boldsymbol{\theta}}_G) - \langle \nabla L_G(\widetilde{\boldsymbol{\theta}}_G), \boldsymbol{\theta} - \widetilde{\boldsymbol{\theta}}_G\rangle - \frac{1}{2}\|\widetilde{D}_G(\boldsymbol{\theta} - \widetilde{\boldsymbol{\theta}}_G)\|^2\right| \leq \mathtt{r}_G^3 \tau_{3,H}.$$

The result (2.17) follows. Further, as $\widetilde{\boldsymbol{\theta}}_G \in \mathcal{A}_G(\mathbf{r}_G)$, it holds

$$L_G(\widetilde{\boldsymbol{\theta}}_G) - L_G(\boldsymbol{\theta}_G^*) - \frac{1}{2}\|D_G^{-1}\nabla\zeta\|^2 = \max_{\boldsymbol{\theta}\in\mathcal{A}_G(\mathbf{r}_G)}\left\{L_G(\boldsymbol{\theta}) - L_G(\boldsymbol{\theta}_G^*) - \frac{1}{2}\|D_G^{-1}\nabla\zeta\|^2\right\}$$

$$= \max_{\boldsymbol{\theta}\in\mathcal{A}_G(\mathbf{r}_G)}\left\{\langle\boldsymbol{\theta} - \boldsymbol{\theta}_G^*, \nabla\zeta\rangle + I\!\!EL_G(\boldsymbol{\theta}) - I\!\!EL_G(\boldsymbol{\theta}_G^*) - \frac{1}{2}\|D_G^{-1}\nabla\zeta\|^2\right\}$$

$$\leq \max_{\boldsymbol{\theta}\in\mathcal{A}_G(\mathbf{r}_G)}\left\{\langle D_G(\boldsymbol{\theta} - \boldsymbol{\theta}_G^*), D_G^{-1}\nabla\zeta\rangle - \frac{1}{2}\|D_G(\boldsymbol{\theta} - \boldsymbol{\theta}_G^*)\|^2 - \frac{1}{2}\|D_G^{-1}\nabla\zeta\|^2\right\} + \mathbf{r}_G^3\,\tau_{3,H}$$

$$\leq \max_{\boldsymbol{\theta}\in\mathcal{A}_G(\mathbf{r}_G)}\left\{-\frac{1}{2}\|D_G(\boldsymbol{\theta} - \boldsymbol{\theta}_G^*) - D_G^{-1}\nabla\zeta\|^2\right\} + \mathbf{r}_G^3\,\tau_{3,H} = \mathbf{r}_G^3\,\tau_{3,H}$$

and similarly $L_G(\widetilde{\boldsymbol{\theta}}_G) - L_G(\boldsymbol{\theta}_G^*) - \frac{1}{2}\|D_G^{-1}\nabla\zeta\|^2 \geq -\mathbf{r}_G^3\,\tau_{3,H}$. This two-sided bound yields as (2.15) as (2.16).

The last statement (2.18) of the theorem follows directly from Lemma A.11 with $Q = D_G$ and $f(\boldsymbol{\theta}) = I\!\!EL_G(\boldsymbol{\theta})$.

Proof of Proposition 2.6 and Theorem 2.7 The definition of $\boldsymbol{\theta}^*$ and $\boldsymbol{\theta}_G^*$ implies

$$I\!\!EL_G(\boldsymbol{\theta}_G^*) \geq I\!\!EL_G(\boldsymbol{\theta}^*), \qquad I\!\!EL(\boldsymbol{\theta}_G^*) \leq I\!\!EL(\boldsymbol{\theta}^*).$$

As $I\!\!EL_G(\boldsymbol{\theta}) = I\!\!EL(\boldsymbol{\theta}) - \|G\boldsymbol{\theta}\|^2/2$, it follows that

$$2I\!\!EL_G(\boldsymbol{\theta}_G^*) - 2I\!\!EL_G(\boldsymbol{\theta}^*) \leq \|G\boldsymbol{\theta}^*\|^2 - \|G\boldsymbol{\theta}_G^*\|^2 \leq \|G\boldsymbol{\theta}^*\|^2. \tag{B.4}$$

The bound (B.3) with $\boldsymbol{\theta} = \boldsymbol{\theta}^*$ implies the first statement of (2.19).

Further we show that $\|G\boldsymbol{\theta}^*\| \leq \mathbf{r}_b/2$ implies $\|D_G(\boldsymbol{\theta}_G^* - \boldsymbol{\theta}^*)\| \leq \mathbf{r}_b$. Indeed, suppose the opposite inequality. Define $\boldsymbol{u} = \mathbf{r}_b D_G(\boldsymbol{\theta}^* - \boldsymbol{\theta}_G^*)/\|D_G(\boldsymbol{\theta}_G^* - \boldsymbol{\theta}^*)\|$, so that $\|\boldsymbol{u}\| = \mathbf{r}_b$. The function $f(t) = I\!\!EL_G(\boldsymbol{\theta}_G^*) - I\!\!EL_G(\boldsymbol{\theta}_G^* + t\boldsymbol{u})$ is convex in t and $\boldsymbol{\theta}_G^* + t\boldsymbol{u} \in \Theta^\circ$ for $|t| \leq 1$. Using the approximation (B.3) for $\boldsymbol{\theta} = \boldsymbol{\theta}_G^* + \boldsymbol{u}$ implies

$$2I\!\!EL_G(\boldsymbol{\theta}_G^*) - 2I\!\!EL_G(\boldsymbol{\theta}_G^* + t\boldsymbol{u}) \geq \mathbf{r}_b^2 - \mathbf{r}_b^3\,\tau_{3,H}(\mathbf{r}_b) \geq \mathbf{r}_b^2/2$$

and concavity of $I\!\!EL_G(\boldsymbol{\theta})$ together with $\nabla I\!\!EL_G(\boldsymbol{\theta}_G^*) = 0$ implies for $t \geq 1$

$$I\!\!EL_G(\boldsymbol{\theta}_G^*) - I\!\!EL_G(\boldsymbol{\theta}_G^* + t\boldsymbol{u}) \geq \mathbf{r}_b^2/2.$$

This contradicts to the bounds (B.4) and $\|G\boldsymbol{\theta}^*\|^2 \leq \mathbf{r}_b^2/2$.

Now for any $\boldsymbol{\theta}$ with $\|D_G(\boldsymbol{\theta}_G^* - \boldsymbol{\theta})\| \leq \mathbf{r}_b$

$$\left|I\!\!EL_G(\boldsymbol{\theta}_G^*) - I\!\!EL_G(\boldsymbol{\theta}) - \frac{1}{2}\|D_G(\boldsymbol{\theta} - \boldsymbol{\theta}_G^*)\|^2\right| \leq \mathbf{r}_b^3\,\tau_{3,H}(\mathbf{r}_b). \tag{B.5}$$

Further we use that $\boldsymbol{\theta}^* = \operatorname{argmax} I\!\!E L(\boldsymbol{\theta})$ and $I\!\!E L_G(\boldsymbol{\theta}) = I\!\!E L(\boldsymbol{\theta}) - \|G\boldsymbol{\theta}\|^2/2$. By (B.5) in view of $\|D_G(\boldsymbol{\theta}_G^* - \boldsymbol{\theta}^*)\| \leq \mathtt{r}_b$ and $D_G^2 = \mathbb{F}(\boldsymbol{\theta}_G^*) + G^2 = D^2 + G^2$

$$
\begin{aligned}
I\!\!E L(\boldsymbol{\theta}^*) - I\!\!E L_G(\boldsymbol{\theta}_G^*) &= \max_{\boldsymbol{\theta} \in \mathcal{A}_G(\mathtt{r}_b)} \left\{ I\!\!E L_G(\boldsymbol{\theta}) + \frac{1}{2}\|G\boldsymbol{\theta}\|^2 - I\!\!E L_G(\boldsymbol{\theta}_G^*) \right\} \\
&\leq \max_{\boldsymbol{\theta} \in \mathcal{A}_G(\mathtt{r}_b)} \left\{ -\frac{1}{2}\|D_G(\boldsymbol{\theta} - \boldsymbol{\theta}_G^*)\|^2 + \frac{1}{2}\|G\boldsymbol{\theta}\|^2 \right\} + \mathtt{r}_b^3\, \tau_{3,H}(\mathtt{r}_b) \\
&= \max_{\boldsymbol{\theta} \in \mathcal{A}_G(\mathtt{r}_b)} \left\{ -\frac{1}{2}\|D\boldsymbol{\theta} - D^{-1}D_G^2\,\boldsymbol{\theta}_G^*\|^2 + \frac{1}{2}\|D^{-1}D_G^2\,\boldsymbol{\theta}_G^*\|^2 \right\} + \mathtt{r}_b^3\, \tau_{3,H}(\mathtt{r}_b).
\end{aligned}
$$

A similar inequality holds from below with opposite sign for $\tau_{3,H}$-term yielding for the maximizer $\boldsymbol{\theta}^*$ the bound

$$
\left\| D\boldsymbol{\theta}^* - D^{-1}D_G^2\,\boldsymbol{\theta}_G^* \right\|^2 \leq 4\mathtt{r}_b^3\, \tau_{3,H}(\mathtt{r}_b).
$$

Equivalently, using again $D_G^2 = D^2 + G^2$

$$
\left\| D^{-1}D_G^2(\boldsymbol{\theta}^* - \boldsymbol{\theta}_G^*) - D^{-1}G^2\boldsymbol{\theta}^* \right\|^2 \leq 4\mathtt{r}_b^3\, \tau_{3,H}(\mathtt{r}_b).
$$

As $D^2 \leq D_G^2$, this also implies

$$
\left\| D_G(\boldsymbol{\theta}^* - \boldsymbol{\theta}_G^*) - D_G^{-1}G^2\boldsymbol{\theta}^* \right\|^2 \leq 4\mathtt{r}_b^3\, \tau_{3,H}(\mathtt{r}_b).
$$

The statement (2.21) of Theorem 2.7 follows from the bound (2.20) on the bias $\|Q(\boldsymbol{\theta}_G^* - \boldsymbol{\theta}^*)\|$ and the deviation bound $\left\| QD_G^{-2}\nabla\zeta \right\| \leq z(B_{Q|G}, \mathtt{x})$ on a set of probability at least $1 - \mathrm{e}^{-\mathtt{x}}$ by the triangle inequality.

For the second statement of Theorem 2.7, we apply the decomposition

$$
\left\| Q(\widetilde{\boldsymbol{\theta}}_G - \boldsymbol{\theta}^*) \right\| = \left\| Q(\widetilde{\boldsymbol{\theta}}_G - \boldsymbol{\theta}_G^*) - \boldsymbol{a} \right\|
$$

with $\boldsymbol{a} = \boldsymbol{\theta}^* - \boldsymbol{\theta}_G^*$. Now the result follows from asymptotic normality (2.23) and Gaussian comparison result of Theorem A.1 in view of small bias condition (2.24).

Proof of Proposition 2.8   Let $\widetilde{\boldsymbol{\theta}}_G = \operatorname{argmax}_{\boldsymbol{\theta}} L_G(\boldsymbol{\theta})$ be the penalized MLE of the parameter $\boldsymbol{\theta}$. We aim at bounding from above the quantity

$$
\rho(\mathtt{r}_0) = \frac{\int_{\|H\boldsymbol{u}\| > \mathtt{r}_0} \exp\{L_G(\widetilde{\boldsymbol{\theta}}_G + \boldsymbol{u})\}d\boldsymbol{u}}{\int_{\|H\boldsymbol{u}\| \leq \mathtt{r}_0} \exp\{L_G(\widetilde{\boldsymbol{\theta}}_G + \boldsymbol{u})\}d\boldsymbol{u}}
$$

with H^2 from $(\boldsymbol{EV})$. We suppose in the proof that $p < \infty$. The general case can be obtained by taking a limit as $p \to \infty$.

Step 1 The use of $\nabla L_G(\widetilde{\boldsymbol{\theta}}_G) = 0$ allows to represent

$$
\begin{aligned}
\rho(\mathbf{r}_0) &= \frac{\int_{\|H\boldsymbol{u}\|>\mathbf{r}_0} \exp\{L_G(\widetilde{\boldsymbol{\theta}}_G + \boldsymbol{u}) - L_G(\widetilde{\boldsymbol{\theta}}_G)\} d\boldsymbol{u}}{\int_{\|H\boldsymbol{u}\|\leq\mathbf{r}_0} \exp\{L_G(\widetilde{\boldsymbol{\theta}}_G + \boldsymbol{u}) - L_G(\widetilde{\boldsymbol{\theta}}_G)\} d\boldsymbol{u}} \\
&= \frac{\int_{\|H\boldsymbol{u}\|>\mathbf{r}_0} \exp\{L_G(\widetilde{\boldsymbol{\theta}}_G + \boldsymbol{u}) - L_G(\widetilde{\boldsymbol{\theta}}_G) - \langle\nabla L_G(\widetilde{\boldsymbol{\theta}}_G), \boldsymbol{u}\rangle\} d\boldsymbol{u}}{\int_{\|H\boldsymbol{u}\|\leq\mathbf{r}_0} \exp\{L_G(\widetilde{\boldsymbol{\theta}}_G + \boldsymbol{u}) - L_G(\widetilde{\boldsymbol{\theta}}_G) - \langle\nabla L_G(\widetilde{\boldsymbol{\theta}}_G), \boldsymbol{u}\rangle\} d\boldsymbol{u}}.
\end{aligned}
$$

Now we study this expression for any possible value $\boldsymbol{\theta}$ from the concentration set of $\widetilde{\boldsymbol{\theta}}_G$. Consider $f(\boldsymbol{\theta}) = I\!\!E L_G(\boldsymbol{\theta})$. As the stochastic term of $L(\boldsymbol{\theta})$ and thus, of $L_G(\boldsymbol{\theta})$ is linear in $\boldsymbol{\theta}$, it holds

$$
L_G(\boldsymbol{\theta} + \boldsymbol{u}) - L_G(\boldsymbol{\theta}) - \langle\nabla L_G(\boldsymbol{\theta}), \boldsymbol{u}\rangle = f(\boldsymbol{\theta} + \boldsymbol{u}) - f(\boldsymbol{u}) - \langle\nabla f(\boldsymbol{\theta}), \boldsymbol{u}\rangle.
$$

Therefore, it suffices to bound uniformly in $\boldsymbol{\theta} \in \Theta^\circ$ the ratio

$$
\rho(\mathbf{r}_0, \boldsymbol{\theta}) \stackrel{\text{def}}{=} \frac{\int I\!\!I(\|H\boldsymbol{u}\| > \mathbf{r}_0) \exp\{f(\boldsymbol{\theta} + \boldsymbol{u}) - f(\boldsymbol{u}) - \langle\nabla f(\boldsymbol{\theta}), \boldsymbol{u}\rangle\} d\boldsymbol{u}}{\int I\!\!I(\|H\boldsymbol{u}\| \leq \mathbf{r}_0) \exp\{f(\boldsymbol{\theta} + \boldsymbol{u}) - f(\boldsymbol{u}) - \langle\nabla f(\boldsymbol{\theta}), \boldsymbol{u}\rangle\} d\boldsymbol{u}} \tag{B.6}
$$

Step 2 First we bound the denominator of $\rho(\mathbf{r}_0, \boldsymbol{\theta})$. Lemma A.12 yields

$$
\int_{\|H\boldsymbol{u}\|\leq\mathbf{r}_0} e^{f(\boldsymbol{\theta}+\boldsymbol{u})-f(\boldsymbol{u})-\langle\nabla f(\boldsymbol{\theta}),\boldsymbol{u}\rangle} d\boldsymbol{u} \geq (1 - \diamond(\mathbf{r}_0)) \int_{\|H\boldsymbol{u}\|\leq\mathbf{r}_0} e^{-\|D_G(\boldsymbol{\theta})\boldsymbol{u}\|^2/2} d\boldsymbol{u},
$$

$$
\int_{\|H\boldsymbol{u}\|\leq\mathbf{r}_0} e^{f(\boldsymbol{\theta}+\boldsymbol{u})-f(\boldsymbol{u})-\langle\nabla f(\boldsymbol{\theta}),\boldsymbol{u}\rangle} d\boldsymbol{u} \leq (1 + \diamond(\mathbf{r}_0)) \int_{\|H\boldsymbol{u}\|\leq\mathbf{r}_0} e^{-\|D_G(\boldsymbol{\theta})\boldsymbol{u}\|^2/2} d\boldsymbol{u},
$$

where $D_G^2(\boldsymbol{\theta}) = \mathbb{F}_G(\boldsymbol{\theta}) = -\nabla^2 f(\boldsymbol{\theta})$ and $\diamond(\mathbf{r}_0)$ is given by (2.27). Moreover, after a proper normalization, the integral $\int_{\|H\boldsymbol{u}\|\leq\mathbf{r}_0} \exp(-\|D_G(\boldsymbol{\theta})\boldsymbol{u}\|^2/2) d\boldsymbol{u}$ can be viewed as the probability of the Gaussian event. Namely

$$
\frac{\det D_G(\boldsymbol{\theta})}{(2\pi)^{p/2}} \int_{\|H\boldsymbol{u}\|\leq\mathbf{r}_0} \exp\left(-\frac{\|D_G(\boldsymbol{\theta})\boldsymbol{u}\|^2}{2}\right) d\boldsymbol{u} = I\!\!P(\|H D_G^{-1}(\boldsymbol{\theta})\boldsymbol{\gamma}\| \leq \mathbf{r}_0)
$$

for a standard normal $\boldsymbol{\gamma} \in I\!\!R^p$. The choice $\mathbf{r}_0 \geq \sqrt{\mathsf{p}_G(\boldsymbol{\theta})} + \sqrt{2\mathbf{x}}$ yields by Corollary A.3

$$
I\!\!P(\|H D_G^{-1}(\boldsymbol{\theta})\boldsymbol{\gamma}\| \leq \mathbf{r}_0) \geq 1 - e^{-\mathbf{x}}.
$$

If the error term $\diamond(\mathbf{r}_0)$ is small, we obtain a sharp bound for the integral in the denominator of $\rho(\mathbf{r}_0, \boldsymbol{\theta})$ from (B.6).

Step 3   Now we bound the integral on the exterior of $\mathcal{U}^\circ = \{\boldsymbol{u} \colon \|H\boldsymbol{u}\| \le \mathtt{r}_0\}$. Linearity of stochastic term in $L_G(\boldsymbol{\theta}) = L(\boldsymbol{\theta}) - \|G\boldsymbol{\theta}\|^2/2$ and quadraticity of the penalty imply

$$L_G(\boldsymbol{\theta} + \boldsymbol{u}) - L_G(\boldsymbol{\theta}) - \langle \nabla L_G(\boldsymbol{\theta}), \boldsymbol{u} \rangle = I\!\!E L(\boldsymbol{\theta} + \boldsymbol{u}) - I\!\!E L(\boldsymbol{\theta}) - \langle \nabla I\!\!E L(\boldsymbol{\theta}), \boldsymbol{u} \rangle - \frac{1}{2}\|G\boldsymbol{u}\|^2.$$

Now we apply Lemma A.16 with $f(\boldsymbol{\theta} + \boldsymbol{u}) = I\!\!E L(\boldsymbol{\theta} + \boldsymbol{u}) - \langle (I\!\!F(\boldsymbol{\theta}) + H^2)\boldsymbol{u}, \boldsymbol{u} \rangle/2$. This function is concave and it holds $-\langle \nabla^2 f(\boldsymbol{\theta})\boldsymbol{u}, \boldsymbol{u} \rangle = \|H\boldsymbol{u}\|^2$. The bound (A.30) yields for any $\boldsymbol{u}$ with $\|H\boldsymbol{u}\| = \mathtt{r} > \mathtt{r}_0$ and $\mathtt{C}_0 = 1 - 3\mathtt{r}_0\,\tau_{3,H} \ge 1/2$

$$L_G(\boldsymbol{\theta} + \boldsymbol{u}) - L_G(\boldsymbol{\theta}) - \langle \nabla L_G(\boldsymbol{\theta}), \boldsymbol{u} \rangle$$
$$= f(\boldsymbol{\theta} + \boldsymbol{u}) - f(\boldsymbol{\theta}) - \langle \nabla f(\boldsymbol{\theta}), \boldsymbol{u} \rangle - \langle (D_G^2 - H^2)\boldsymbol{u}, \boldsymbol{u} \rangle/2$$
$$\le -\mathtt{C}_0(\|H\boldsymbol{u}\|\mathtt{r}_0 - \mathtt{r}_0^2/2) - \langle (D_G^2 - H^2)\boldsymbol{u}, \boldsymbol{u} \rangle/2$$
$$= -\mathtt{C}_0(\|H\boldsymbol{u}\|\mathtt{r}_0 - \mathtt{r}_0^2/2) - \|D_{G|H}(\boldsymbol{\theta})\boldsymbol{u}\|^2/2,$$

where $D_{G|H}^2 = D_G^2 - H^2$. Now we can use the result about Gaussian integrals from Section A.5. With $\mathcal{T} = H D_{G|H}^{-1}(\boldsymbol{\theta})$, it holds by Lemma A.17

$$\frac{\det D_G(\boldsymbol{\theta})}{(2\pi)^{p/2}} \int I\!\!I\big(\|H\boldsymbol{u}\| > \mathtt{r}_0\big) \exp\big\{ L_G(\boldsymbol{\theta} + \boldsymbol{u}) - L_G(\boldsymbol{\theta}) - \langle \nabla L_G(\boldsymbol{\theta}), \boldsymbol{u} \rangle \big\}\, d\boldsymbol{u}$$
$$\le I\!\!E\Big\{ \exp\Big(-\mathtt{C}_0\mathtt{r}_0\|\mathcal{T}\boldsymbol{\gamma}\| + \frac{\mathtt{C}_0\mathtt{r}_0^2}{2} + \frac{1}{2}\|\mathcal{T}\boldsymbol{\gamma}\|^2 \Big) I\!\!I\big(\|\mathcal{T}\boldsymbol{\gamma}\| > \mathtt{r}_0\big) \Big\} \le \mathtt{C}\mathrm{e}^{-(\mathtt{p}_G(\boldsymbol{\theta})+\mathtt{x})/2}.$$

Putting together of Step 1 through Step 3 yields the statement about $\rho(\mathtt{r}_0)$.

Proof of Theorem 2.9 and Corollary 2.10   We proceed similarly to the proof of Theorem 2.8. Fix any centrally symmetric set A. First we restrict the posterior probability to the set $\mathcal{E}(\mathtt{r}_0) = \{\boldsymbol{u}\colon \|H\boldsymbol{u}\| \le \mathtt{r}_0\}$. Then we apply the quadratic approximation of the log-likelihood function $L(\boldsymbol{\theta})$. Denote $A(\mathtt{r}_0) = A \cap \mathcal{E}(\mathtt{r}_0)$. Obviously, $A(\mathtt{r}_0)$ is centrally symmetric as well. Further,

$$I\!\!P\big(\boldsymbol{\vartheta}_G - \widetilde{\boldsymbol{\theta}}_G \in A \,\big|\, \boldsymbol{Y}\big) = \frac{\int_A \exp\big\{ L_G(\widetilde{\boldsymbol{\theta}}_G + \boldsymbol{u}) \big\}\, d\boldsymbol{u}}{\int_{I\!\!R^p} \exp\big\{ L_G(\widetilde{\boldsymbol{\theta}}_G + \boldsymbol{u}) \big\}\, d\boldsymbol{u}}$$
$$\le \frac{\int_{A(\mathtt{r}_0)} \exp\big\{ L_G(\widetilde{\boldsymbol{\theta}}_G + \boldsymbol{u}) - L_G(\widetilde{\boldsymbol{\theta}}_G) - \langle \nabla L_G(\widetilde{\boldsymbol{\theta}}_G), \boldsymbol{u} \rangle \big\}\, d\boldsymbol{u}}{\int_{\|H\boldsymbol{u}\| \le \mathtt{r}_0} \exp\big\{ L_G(\widetilde{\boldsymbol{\theta}}_G + \boldsymbol{u}) - L_G(\widetilde{\boldsymbol{\theta}}_G) - \langle \nabla L_G(\widetilde{\boldsymbol{\theta}}_G), \boldsymbol{u} \rangle \big\}\, d\boldsymbol{u}} + \rho(\mathtt{r}_0).$$

Now we apply the bounds from the proof of Theorem 2.8 yielding the upper bound

$$\mathbb{P}(\boldsymbol{\vartheta}_G - \widetilde{\boldsymbol{\theta}}_G \in A \,|\, \boldsymbol{Y}) \leq \frac{\{1 + \Diamond(\mathtt{r}_0)\} \int_{A(\mathtt{r}_0)} \exp\{-\|\widetilde{D}_G \boldsymbol{u}\|^2/2\} d\boldsymbol{u}}{\{1 - \Diamond(\mathtt{r}_0)\} \int_{\|H\boldsymbol{u}\| \leq \mathtt{r}_0} \exp\{-\|\widetilde{D}_G \boldsymbol{u}\|^2/2\} d\boldsymbol{u}} + \rho(\mathtt{r}_0)$$

$$\leq \frac{\{1 + \Diamond(\mathtt{r}_0)\} \mathbb{P}(\widetilde{D}_G^{-1} \boldsymbol{\gamma} \in A)}{\{1 - \Diamond(\mathtt{r}_0)\} \mathbb{P}(\|H\widetilde{D}_G^{-1}\boldsymbol{\gamma}\| \leq \mathtt{r}_0)} + \rho(\mathtt{r}_0).$$

Now we prove the lower bound. It obviously holds

$$\mathbb{P}(\boldsymbol{\vartheta}_G - \widetilde{\boldsymbol{\theta}}_G \in A \,|\, \boldsymbol{Y}) = \frac{\int_A \exp\{L_G(\widetilde{\boldsymbol{\theta}}_G + \boldsymbol{u})\} d\boldsymbol{u}}{\int_{\mathbb{R}^p} \exp\{L_G(\widetilde{\boldsymbol{\theta}}_G + \boldsymbol{u})\} d\boldsymbol{u}}$$

$$\geq \frac{\int_{A(\mathtt{r}_0)} \exp\{L_G(\widetilde{\boldsymbol{\theta}}_G + \boldsymbol{u}) - L_G(\widetilde{\boldsymbol{\theta}}_G) - \langle \nabla L_G(\widetilde{\boldsymbol{\theta}}_G), \boldsymbol{u} \rangle\} d\boldsymbol{u}}{(\int_{\|H\boldsymbol{u}\| \leq \mathtt{r}_0} + \int_{\|H\boldsymbol{u}\| > \mathtt{r}_0}) \exp\{L_G(\widetilde{\boldsymbol{\theta}}_G + \boldsymbol{u}) - L_G(\widetilde{\boldsymbol{\theta}}_G) - \langle \nabla L_G(\widetilde{\boldsymbol{\theta}}_G), \boldsymbol{u} \rangle\} d\boldsymbol{u}}$$

and in a similar way as above

$$\mathbb{P}(\boldsymbol{\vartheta}_G - \widetilde{\boldsymbol{\theta}}_G \in A \,|\, \boldsymbol{Y}) \geq \frac{\{1 - \Diamond(\mathtt{r}_0)\} \mathbb{P}(\widetilde{D}_G^{-1}\boldsymbol{\gamma} \in A(\mathtt{r}_0))}{\{1 + \Diamond(\mathtt{r}_0)\} \mathbb{P}(\|H\widetilde{D}_G^{-1}\boldsymbol{\gamma}\| \leq \mathtt{r}_0) + \mathtt{C} e^{-(\widetilde{\mathtt{p}}_G + \mathtt{x})/2}}$$

$$\geq \frac{\{1 - \Diamond(\mathtt{r}_0)\}\{\mathbb{P}(\widetilde{D}_G^{-1}\boldsymbol{\gamma} \in A) - \rho(\mathtt{r}_0)\}}{\{1 + \Diamond(\mathtt{r}_0)\} \mathbb{P}(\|H\widetilde{D}_G^{-1}\boldsymbol{\gamma}\| \leq \mathtt{r}_0) + \mathtt{C} e^{-(\widetilde{\mathtt{p}}_G + \mathtt{x})/2}}.$$

For the case of an arbitrary possibly non-symmetric A, the proof is similar with the use of (A.24) instead of (A.23).

Proof of Theorem 2.11 and of Theorem 2.12 The difference $\boldsymbol{\vartheta}_G - \boldsymbol{\theta}^*$ can be decomposed as

$$\boldsymbol{\vartheta}_G - \boldsymbol{\theta}^* = (\boldsymbol{\vartheta}_G - \widetilde{\boldsymbol{\theta}}_G) + (\widetilde{\boldsymbol{\theta}}_G - \boldsymbol{\theta}^*) = (\boldsymbol{\vartheta}_G - \widetilde{\boldsymbol{\theta}}_G) + (\widetilde{\boldsymbol{\theta}}_G - \boldsymbol{\theta}_G^*) + (\boldsymbol{\theta}_G^* - \boldsymbol{\theta}^*).$$

Theorem 2.7 with $\mathtt{x} = \log n$ allows to bound with high probability

$$\|Q(\widetilde{\boldsymbol{\theta}}_G - \boldsymbol{\theta}^*)\|^2 \lesssim \|QD_G^{-2}G^2\boldsymbol{\theta}^*\|^2 + z^2(B_{Q|G}^2, \mathtt{x}) \leq \|QD_G^{-2}G^2\boldsymbol{\theta}^*\|^2 + \mathrm{tr}(B_{Q|G}^2) + \log n,$$

where $B_{Q|G}^2 = QD_G^{-2}H^2D_G^{-2}Q^\top$. Moreover, as $H^2 \lesssim D_G^2$, we bound $\mathrm{tr}(B_{Q|G}^2) \lesssim \mathrm{tr}(QD_G^{-2}Q^\top)$. Further, Theorem 2.9 yields on a random set $\Omega(\mathtt{x})$ for $\mathtt{x} = \log n$

$$\mathbb{P}(\|Q(\boldsymbol{\vartheta}_G - \widetilde{\boldsymbol{\theta}}_G)\| \geq \mathtt{r} \,|\, \boldsymbol{Y}) \lesssim \mathbb{P}'(\|Q\widetilde{D}_G^{-1}\boldsymbol{\gamma}\| \geq \mathtt{r}) + 1/n.$$

Now we apply Theorem A.2 with $\mathtt{r} = \mathtt{r}_Q = z(Q\widetilde{D}_G^{-2}Q^\top, \mathtt{x}) \leq \sqrt{\mathrm{tr}(Q\widetilde{D}_G^{-2}Q^\top)} + \sqrt{2\mathtt{x}}$ to the Gaussian quadratic form $\|Q\widetilde{D}_G^{-1}\boldsymbol{\gamma}\|^2$. The desired result (2.33) follows by (2.18) of Theorem 2.5 and by the bias bound (2.32).

To check (2.34), note first that by definition, it holds for the true parameter $\boldsymbol{\theta}^*$:

$$IP\big(\boldsymbol{\theta}^* \in \mathcal{A}_{Q|G}(\mathtt{r})\big) = IP\big(\|Q(\widetilde{\boldsymbol{\theta}}_G - \boldsymbol{\theta}^*)\| \leq \mathtt{r}\big).$$

The Fisher expansion (2.15) $\widetilde{\boldsymbol{\theta}}_G - \boldsymbol{\theta}_G^* \approx D_G^{-2}\nabla\zeta$ of Theorem 2.5 combined with the CLT $V^{-1}\nabla\zeta \xrightarrow{w} \boldsymbol{\gamma}$ for a standard normal $\boldsymbol{\gamma}$ reduces the latter question to Gaussian probability

$$IP\big(\|Q(\widetilde{\boldsymbol{\theta}}_G - \boldsymbol{\theta}^*)\| \leq \mathtt{r}\big) \approx IP\Big(\|Q(D_G^{-2}V\boldsymbol{\gamma} + \boldsymbol{\theta}_G^* - \boldsymbol{\theta}^*)\| \leq \mathtt{r}\Big).$$

By Gaussian comparison Theorem A.1, the impact of the bias $\boldsymbol{\theta}_G^* - \boldsymbol{\theta}^*$ is negligible under the undersmoothing condition $\|Q(\boldsymbol{\theta}_G^* - \boldsymbol{\theta}^*)\|^2 \ll \mathrm{tr}\big(QD_G^{-2}V^2 D_G^{-2}Q^\top\big)$. Combining with Theorem 2.7 yields in view of $D_G^{-2}V^2 D_G^{-2} \leq D_G^{-2}$

$$1 - \alpha = IP'\big(\|Q\widetilde{D}_G^{-1}\boldsymbol{\gamma}\| \leq \mathtt{r}_\alpha\big) \approx IP\big(\|QD_G^{-1}\boldsymbol{\gamma}\| \leq \mathtt{r}_\alpha\big) \leq IP\Big(\|QD_G^{-2}V\boldsymbol{\gamma}\| \leq \mathtt{r}_\alpha\Big)$$

$$\approx IP\big(\|Q(\widetilde{\boldsymbol{\theta}}_G - \boldsymbol{\theta}^*)\| \leq \mathtt{r}_\alpha\big),$$

that is, the credible set $\mathcal{A}_{Q|G}(\mathtt{r}_\alpha)$ is an asymptotically valid confidence set.

Proof of Theorems 3.1 through 3.4   It suffices to check the conditions of the general results from Section 2. We start with the log-density model. First we show $(\mathcal{L}_0)$. Remind that $IEL(\boldsymbol{\theta}) = n\{\langle\overline{\Psi}, \boldsymbol{\theta}\rangle - \phi(\boldsymbol{\theta})\}$ and $\mathbb{F}(\boldsymbol{\theta}) = -\nabla^2 IEL(\boldsymbol{\theta}) = n\nabla^2\phi(\boldsymbol{\theta})$. Fix any $\boldsymbol{\theta} \in \Theta^\circ$ and denote also $\Psi_{\boldsymbol{\theta}} = E_{\boldsymbol{\theta}}\Psi(X_1)$ and define for $\boldsymbol{u} \in I\!\!R^{m^*}$ with $\|H\boldsymbol{u}\| = \mathtt{r}$ and any t

$$q(t) \stackrel{\text{def}}{=} \int \exp\{\langle\Psi(x) - \Psi_{\boldsymbol{\theta}}, \boldsymbol{\theta} + t\boldsymbol{u}\rangle - \phi(\boldsymbol{\theta})\}\mu_0(dx) = \int \exp\{t\langle\Psi(x) - \Psi_{\boldsymbol{\theta}}, \boldsymbol{u}\rangle\}P_{\boldsymbol{\theta}}(dx),$$

$$q_k(t) \stackrel{\text{def}}{=} \frac{d^k q(t)}{dt^k} = \int \langle\Psi(x) - \Psi_{\boldsymbol{\theta}}, \boldsymbol{u}\rangle^k \exp\{t\langle\Psi(x) - \Psi_{\boldsymbol{\theta}}, \boldsymbol{u}\rangle\}P_{\boldsymbol{\theta}}(dx), \quad k \geq 1. \tag{B.7}$$

Due to (ψ_j), all these quantities are well defined. Moreover, $q(0) = 1$ and does not depend on $\boldsymbol{\theta}, \boldsymbol{u}$ while $q_1(0) = 0$. Also

$$q_2(0) = \langle\nabla^2\phi(\boldsymbol{\theta})\boldsymbol{u}, \boldsymbol{u}\rangle.$$

Further, define

$$h(t) \stackrel{\text{def}}{=} \log q(t) = \phi(\boldsymbol{\theta} + t\boldsymbol{u}) - \phi(\boldsymbol{\theta}) - t\langle\Psi_{\boldsymbol{\theta}}, \boldsymbol{u}\rangle.$$

Then

$$\delta_k(\boldsymbol{\theta}, \boldsymbol{u}) = -n\frac{1}{k!}\frac{d^k}{dt^k}\phi(\boldsymbol{\theta} + t\boldsymbol{u})\Big|_{t=0} = -n\frac{1}{k!}\frac{d^k}{dt^k}h(t)\Big|_{t=0}, \quad k = 3, 4,$$

Straightforward calculus yields

$$h^{(3)}(0) = -q_3(0) + 3q_2(0)\, q_1(0) - 2q_1^3(0)\,,$$

$$h^{(4)}(0) = -q_4(0) + 4q_3(0)\, q_1(0) + 3q_2^2(0) - 12q_2(0)\, q_1^2(0) + 6q_1^4(0)\,.$$

Now $(\boldsymbol{\Psi u})$ implies

$$\left| q_k(0) \right| \le \left\{ \mathtt{C}_f^2\, q_2(0) \right\}^{k/2} \le \left\{ \mathtt{C}_f^2\, \langle \nabla^2\phi(\boldsymbol{\theta})\boldsymbol{u}, \boldsymbol{u} \rangle \right\}^{k/2}, \quad k = 3, 4.$$

As $H^2 \ge n\nabla^2\phi(\boldsymbol{\theta})$ and $\|H\boldsymbol{u}\|^2 = \mathtt{r}^2$, this yields for some absolute constant $\mathtt{C}_3, \mathtt{C}_4$

$$\delta_3(\boldsymbol{\theta}, \boldsymbol{u}) \le \mathtt{C}_3\, \mathtt{C}_f^3\, n(\mathtt{r}^2/n)^{3/2}, \quad \delta_4(\boldsymbol{\theta}, \boldsymbol{u}) \le \mathtt{C}_4\, \mathtt{C}_f^4\, n(\mathtt{r}^2/n)^2.$$

Now we check $(\boldsymbol{EV})$ for $\nabla\zeta = S - I\!ES$ and $V^2 = n\,\nabla^2\phi(\boldsymbol{\theta}^*)$. Let $\|H\boldsymbol{u}\| = \lambda$. It holds from (3.1) due to the i.i.d. structure of the data in view of $\nabla\phi(\boldsymbol{\theta}^*) = \overline{\boldsymbol{\Psi}}$

$$\log I\!E \exp\big\{ \langle \nabla\zeta, \boldsymbol{u} \rangle \big\} = n \log E \exp\big\{ \langle \boldsymbol{\Psi}(X_1) - \overline{\boldsymbol{\Psi}}, \boldsymbol{u} \rangle \big\}$$

$$= n\big\{ \phi(\boldsymbol{\theta}^* + \boldsymbol{u}) - \langle \nabla\phi(\boldsymbol{\theta}^*), \boldsymbol{u} \rangle \big\} = \frac{n}{2}\big\langle \nabla^2\phi(\boldsymbol{\theta}^* + t\boldsymbol{u})\, \boldsymbol{u}, \boldsymbol{u} \big\rangle$$

for $t \in [0, 1]$. By $(\boldsymbol{\nabla^2\phi})$, $\|\boldsymbol{u}\| \le \|H^{-1}\|\,\|H\boldsymbol{u}\| \le \mathtt{C}_{\phi,1} n^{-1/2}\lambda$. By the Cauchy-Schwarz inequality and $(\boldsymbol{\psi_j})$

$$\left| \langle \boldsymbol{\Psi}(x), \boldsymbol{u} \rangle \right|^2 \le q_{m^*}^{-2}\|\boldsymbol{u}\|^2 \sum_{j=1}^{m^*} \psi_j^2(x)q_j^2 \le \mathtt{C}_{\phi,1}^2\, \mathtt{C}_\psi^2\, \frac{\lambda^2\, q_{m^*}^{-2}}{n} \le \mathtt{C}_{\phi,1}^2\, \mathtt{C}_\psi^2\, \frac{\lambda^2 \log^2(n)}{n^{2/3}}\,.$$

If the latter value is smaller than a constant $\mathtt{C}$ then by (B.7)

$$\big\langle \nabla^2\phi(\boldsymbol{\theta}^* + t\boldsymbol{u})\, \boldsymbol{u}, \boldsymbol{u} \big\rangle \le \mathrm{e}^{\mathtt{C}}\big\langle \nabla^2\phi(\boldsymbol{\theta}^*)\, \boldsymbol{u}, \boldsymbol{u} \big\rangle.$$

This yields $(\boldsymbol{EV})$ with $\mathtt{g} \lesssim n^{1/3}/\log(n) \lesssim m^*$. Theorem A.5 ensure for $z_G = \sqrt{\mathtt{p}_G} + \sqrt{2\log n}$ the probability bound (2.11) $\big\| D_G^{-1}\nabla\zeta \big\| \le z_G$ on a random set Ω_n with $I\!P(\Omega_n) \ge 1 - 3/n$. Now all the statements of Theorem 3.1 and 3.2 follow directly from the general results of Section 2.

Now we check the general conditions for the GLM starting with $(\boldsymbol{\mathcal{L}_0})$. Let $\boldsymbol{u} \in I\!R^{m^*}$ with $\|H\boldsymbol{u}\| \le \mathtt{r}$. It holds by $|\langle \boldsymbol{\Psi}_i, \boldsymbol{\theta} \rangle| \le \mathtt{C}_\Psi$, $(\boldsymbol{\phi^{(k)}})$, and $(\boldsymbol{\Psi u})$ for $k = 3, 4$

$$\left| \delta_k(\boldsymbol{\theta}, \boldsymbol{u}) \right| = \left| \sum \phi^{(k)}\big(\langle \boldsymbol{\Psi}_i, \boldsymbol{\theta} \rangle \big) \langle \boldsymbol{\Psi}_i, \boldsymbol{u} \rangle^k \right| \le \|\phi^{(k)}\|_\infty \left(\sum \langle \boldsymbol{\Psi}_i, \boldsymbol{u} \rangle^4 \right)^{k/4}$$

$$\le n^{k/2-1}\|\phi^{(k)}\|_\infty \left(\sum \langle \boldsymbol{\Psi}_i, \boldsymbol{u} \rangle^2 \right)^{k/2} \le \mathtt{C}\mathtt{r}^k n^{-k/2+1}$$

yielding (2.4). Further, by independence of the Y_i's,

$$\log \mathbb{E} \exp\{\lambda\langle\nabla\zeta, \boldsymbol{u}\rangle\} = \sum \log \mathbb{E} \exp\{\lambda\varepsilon_i\langle\Psi_i, \boldsymbol{u}\rangle\}.$$

Under (ψ_∞) and $(\mathbb{F})$, one can bound for any $\boldsymbol{u}$ with $\|H\boldsymbol{u}\| = 1$

$$\left|\langle\Psi_i, \boldsymbol{u}\rangle\right| \leq C_\psi \sqrt{m^*} \|\boldsymbol{u}\| \leq C_\psi C_\mathbb{F} \sqrt{m^*} \|H\boldsymbol{u}\| n^{-1/2} \leq C_\psi C_\mathbb{F} \sqrt{m^*} n^{-1/2}.$$

Thus, $\left|\lambda\langle\Psi_i, \boldsymbol{u}\rangle\right| \leq \lambda C_\mathbb{F} \sqrt{m^*} n^{-1/2}$, and, for $|\lambda| \leq \varrho/(C\sqrt{m^*} n^{-1/2})$, it follows by (ε_i)

$$\sum \log \mathbb{E} \exp\{\lambda\varepsilon_i\langle\Psi_i, \boldsymbol{u}\rangle\} \leq \sum \frac{\nu_0^2\lambda^2\sigma_i^2}{2}\langle\Psi_i, \boldsymbol{u}\rangle^2 \leq \frac{\nu_0^2\lambda^2}{2}\|H\boldsymbol{u}\|^2 = \frac{\nu_0^2\lambda^2}{2}$$

yielding $(\boldsymbol{EV})$ with $\lambda \asymp n^{1/3}$. We complete the proof as in the log-density case.

B.1   Proof of Theorem 2.13

It holds

$$\overline{\boldsymbol{\vartheta}}_G - \widetilde{\boldsymbol{\theta}}_G = \frac{\int\left(\boldsymbol{\theta} - \widetilde{\boldsymbol{\theta}}_G\right)\exp L_G(\boldsymbol{\theta})d\boldsymbol{\theta}}{\int \exp L_G(\boldsymbol{\theta})d\boldsymbol{\theta}}.$$

The use of $\nabla L_G(\widetilde{\boldsymbol{\theta}}_G) = 0$ helps to represent with $\mathcal{E}(\mathbf{r}_0) = \{\boldsymbol{u}\colon \|H\boldsymbol{u}\| \leq \mathbf{r}_0\}$

$$Q\left(\overline{\boldsymbol{\vartheta}}_G - \widetilde{\boldsymbol{\theta}}_G\right) = \frac{\left(\int_{\|H\boldsymbol{u}\|\leq\mathbf{r}_0} + \int_{\|H\boldsymbol{u}\|>\mathbf{r}_0}\right) Q\boldsymbol{u}\exp\{L_G(\widetilde{\boldsymbol{\theta}}_G + \boldsymbol{u}) - L_G(\widetilde{\boldsymbol{\theta}}_G) - \langle\nabla L_G(\widetilde{\boldsymbol{\theta}}_G), \boldsymbol{u}\rangle\}d\boldsymbol{u}}{\int \exp\{L_G(\widetilde{\boldsymbol{\theta}}_G + \boldsymbol{u}) - L_G(\widetilde{\boldsymbol{\theta}}_G) - \langle\nabla L_G(\widetilde{\boldsymbol{\theta}}_G), \boldsymbol{u}\rangle\}d\boldsymbol{u}}.$$

Now, with $f_{\boldsymbol{\theta}}(\boldsymbol{u}) = \mathbb{E}L_G(\boldsymbol{\theta}+\boldsymbol{u})$, define $f(\boldsymbol{u})$ by using $\boldsymbol{\theta} = \widetilde{\boldsymbol{\theta}}_G$, that is, $f(\boldsymbol{u}) = f_{\widetilde{\boldsymbol{\theta}}_G}(\boldsymbol{u})$. Linearity of the stochastic part of $L_G(\boldsymbol{\theta})$ implies

$$L_G(\widetilde{\boldsymbol{\theta}}_G + \boldsymbol{u}) - L_G(\widetilde{\boldsymbol{\theta}}_G) - \langle\nabla L_G(\widetilde{\boldsymbol{\theta}}_G), \boldsymbol{u}\rangle = f(\boldsymbol{u}) - f(0) - f'(0, \boldsymbol{u}),$$

and it holds

$$\|Q(\overline{\boldsymbol{\vartheta}}_G - \widetilde{\boldsymbol{\theta}}_G)\| \leq \rho_0(\mathbf{r}_0) + \rho_1(\mathbf{r}_0)$$

with

$$\rho_0(\mathbf{r}_0) \stackrel{\text{def}}{=} \left\|\frac{\int_{\|H\boldsymbol{u}\|\leq\mathbf{r}_0} Q\boldsymbol{u}\exp\{f(\boldsymbol{u}) - f(0) - f'(0, \boldsymbol{u})\}d\boldsymbol{u}}{\int_{\|H\boldsymbol{u}\|\leq\mathbf{r}_0}\exp\{f(\boldsymbol{u}) - f(0) - f'(0, \boldsymbol{u})\}d\boldsymbol{u}}\right\|,$$

$$\rho_1(\mathbf{r}_0) \stackrel{\text{def}}{=} \left\|\frac{\int_{\|H\boldsymbol{u}\|>\mathbf{r}_0} Q\boldsymbol{u}\exp\{f(\boldsymbol{u}) - f(0) - f'(0, \boldsymbol{u})\}d\boldsymbol{u}}{\int_{\|H\boldsymbol{u}\|\leq\mathbf{r}_0}\exp\{f(\boldsymbol{u}) - f(0) - f'(0, \boldsymbol{u})\}d\boldsymbol{u}}\right\|,$$

As $-\nabla^2 f(0) = \widetilde{D}_G^2$, Lemma A.13 and Theorem 2.8 imply

$$\rho_0(\mathbf{r}_0) \lesssim \frac{\delta_3 I\!\!E \|Q\widetilde{D}_G^{-1}\boldsymbol{\gamma}\|}{I\!\!P(\|H\widetilde{D}_G^{-1}\boldsymbol{\gamma}\| \le \mathbf{r}_0)} \lesssim \delta_3\sqrt{\widetilde{\mathsf{p}}_{Q|G}}$$

with $\widetilde{\mathsf{p}}_{Q|G} = \operatorname{tr}(Q\widetilde{D}_G^{-2}Q^\top)$. For bounding the term $\rho_1(\mathbf{r}_0)$, we apply the bound from Lemma A.16 and then Lemma A.17 and A.18 with $\mathcal{T} = H\widetilde{D}_G^{-1}$ and $\mathsf{p}_\tau = \operatorname{tr}(\mathcal{T}^\top\mathcal{T}) = \widetilde{\mathsf{p}}_G$. The use of $2\|Q\boldsymbol{u}\| \le 1 + \|Q\boldsymbol{u}\|^2$ yields on $\Omega(\mathbf{x})$

$$\rho_1(\mathbf{r}_0) \lesssim \widetilde{\mathsf{p}}_{Q|G} \exp\{-(\widetilde{\mathsf{p}}_G + \mathbf{x})/2\}. \tag{B.8}$$

The second moment of the expression $\langle \boldsymbol{u}, \widetilde{D}_G(\boldsymbol{\vartheta}_G - \widetilde{\boldsymbol{\theta}}_G)\rangle$ given $\boldsymbol{Y}$ and a unit vector $\boldsymbol{z}$ is evaluated similarly. One gets

$$I\!\!E\left[\langle \boldsymbol{z}, \widetilde{D}_G(\boldsymbol{\vartheta}_G - \widetilde{\boldsymbol{\theta}}_G)\rangle^2 \,\big|\, \boldsymbol{Y}\right] - 1 = \frac{\int \langle \boldsymbol{z}, \widetilde{D}_G(\boldsymbol{\vartheta}_G - \widetilde{\boldsymbol{\theta}}_G)\rangle^2 \exp L_G(\boldsymbol{\theta})\,d\boldsymbol{\theta}}{\int \exp L_G(\boldsymbol{\theta})\,d\boldsymbol{\theta}} - 1$$

$$= \frac{\left(\int_{\|H\boldsymbol{u}\| \le \mathbf{r}_0} + \int_{\|H\boldsymbol{u}\| > \mathbf{r}_0}\right)\left[\langle \boldsymbol{z}, \widetilde{D}_G\boldsymbol{u}\rangle^2 - 1\right] \exp\{f(\boldsymbol{u}) - f(0) - f'(0, \boldsymbol{u})\}d\boldsymbol{u}}{\int \exp\{f(\boldsymbol{u}) - f(0) - f'(0, \boldsymbol{u})\}d\boldsymbol{u}}$$

$$= \rho_2 + \rho_3.$$

Similarly to (B.8), one can get $|\rho_3| \le \widetilde{\mathsf{p}}_{Q|G} \exp\{-(\widetilde{\mathsf{p}}_G + \mathbf{x})/2\}$. For the value $|\rho_2|$, we use symmetricity of $\mathcal{U}^\circ = \mathcal{E}(\mathbf{r}_0) = \{\boldsymbol{u}\colon \|H\boldsymbol{u}\| \le \mathbf{r}_0\}$ and Lemma A.12 yielding $|\rho_2| \lesssim \Diamond$.

B.2 Proof of Theorem 2.14

We start with an extension of Theorem 2.8 to some non-symmetric sets A. Let $\boldsymbol{a}$ be a possibly random vector in $I\!\!R^p$. The next result assumes that $\boldsymbol{a}$ is sufficiently small so that $\|\widetilde{D}_G\boldsymbol{a}\| \le 1$.

Proposition B.1. *Let the conditions of Theorem 2.8 hold. Suppose that $\mathbf{r}_0$ satisfies the conditions (2.27) and (2.28) with $\mathbf{x} = 2\log n$. Let a random vector $\boldsymbol{a}$ satisfy $\|\widetilde{D}_G\boldsymbol{a}\| \le 1$ on the set $\Omega(\mathbf{x})$ from (2.11). Then for any set $A \in \mathcal{B}_s(I\!\!R^p)$, it holds on $\Omega(\mathbf{x})$*

$$\left| I\!\!P(\boldsymbol{\vartheta}_G - \widetilde{\boldsymbol{\theta}}_G - \boldsymbol{a} \in A \,|\, \boldsymbol{Y}) - I\!\!P'(\widetilde{D}_G^{-1}\boldsymbol{\gamma} - \boldsymbol{a} \in A) \right|$$

$$\lesssim \left\{\Diamond(\mathbf{r}_0) + \delta_3(\mathbf{r}_0)\|\widetilde{D}_G\boldsymbol{a}\|\right\} I\!\!P'(\widetilde{D}_G^{-1}\boldsymbol{\gamma} \in A) + 1/n.$$

This result can be proved in the same line as Theorem 2.9 using Lemma A.14. Taking into account the Gaussian approximation result from Corollary 2.10, we only have to

compare the posterior probability of $\|Q(\boldsymbol{\vartheta}_G - \overline{\boldsymbol{\vartheta}}_G)\| \leq \mathtt{r}$ with $I\!\!P'\big(\|Q\widetilde{D}_G^{-1}\boldsymbol{\gamma}\| \leq \mathtt{r}\big)$. Let $\boldsymbol{a}$ be defined as

$$\boldsymbol{a} = \widetilde{\boldsymbol{\theta}}_G - \overline{\boldsymbol{\vartheta}}_G.$$

As $Q = Q\Pi$ for a projector Π, it also holds with $\boldsymbol{a}_0 = \Pi\boldsymbol{a}$

$$\|Q(\boldsymbol{\vartheta}_G - \overline{\boldsymbol{\vartheta}}_G)\| = \|Q(\boldsymbol{\vartheta}_G - \widetilde{\boldsymbol{\theta}}_G - \boldsymbol{a}_0)\|$$

and

$$I\!\!P\big(\boldsymbol{\vartheta}_G - \overline{\boldsymbol{\vartheta}}_G \in \mathcal{E}_Q(\mathtt{r})\,\big|\,\boldsymbol{Y}\big) = I\!\!P\big(\boldsymbol{\vartheta}_G - \widetilde{\boldsymbol{\theta}}_G - \boldsymbol{a}_0 \in \mathcal{E}_Q(\mathtt{r})\,\big|\,\boldsymbol{Y}\big)$$

Now Theorem B.1 implies

$$\left| I\!\!P\big(\boldsymbol{\vartheta}_G - \widetilde{\boldsymbol{\theta}}_G - \boldsymbol{a}_0 \in \mathcal{E}_Q(\mathtt{r})\,\big|\,\boldsymbol{Y}\big) - I\!\!P'\big(\widetilde{D}_G^{-1}\boldsymbol{\gamma} - \boldsymbol{a}_0 \in \mathcal{E}_Q(\mathtt{r})\big) \right|$$
$$\lesssim \mathtt{c}\Big\{\diamond + \delta_3\|\widetilde{D}_G\boldsymbol{a}_0\| + n^{-1}\Big\}.$$

Theorem 2.13 yields that the norm of $\widetilde{D}_G\boldsymbol{a}_0$ can be bounded on $\Omega(\mathtt{x})$ as

$$\|\widetilde{D}_G\boldsymbol{a}_0\| = \|\widetilde{D}_G\Pi\boldsymbol{a}\| \lesssim \delta_3\sqrt{\widetilde{\mathtt{p}}_\Pi} + n^{-1}\widetilde{\mathtt{p}}_\Pi$$

with $\widetilde{\mathtt{p}}_\Pi = \mathrm{tr}(\Pi\widetilde{D}_G^{-2}\Pi\widetilde{D}_G^2\Pi)$. It remains to compare two Gaussian probabilities of $\|Q\widetilde{D}_G^{-1}\boldsymbol{\gamma}\| \leq \mathtt{r}$ and of $\|Q(\widetilde{D}_G^{-1}\boldsymbol{\gamma} - \boldsymbol{a}_0)\| \leq \mathtt{r}$. For this one can apply the Pinsker inequality. However, the Gaussian comparison result of Theorem A.1 provides a more precise bound in view of the elliptic shape of the considered credible sets:

$$\left| I\!\!P\big(\|Q(\widetilde{D}_G^{-1}\boldsymbol{\gamma} - \boldsymbol{a}_0)\| \leq \mathtt{r}\big) - I\!\!P\big(\|Q\widetilde{D}_G^{-1}\boldsymbol{\gamma}\| \leq \mathtt{r}\big) \right| \leq \frac{\|Q\boldsymbol{a}_0\|^2}{\|Q\widetilde{D}_G^{-2}Q^\top\|_{\mathrm{Fr}}}.$$

Now the assertion follows by one more application of Theorem 2.13 in view of $\widetilde{\mathtt{p}}_{Q|G} \leq \|Q\widetilde{D}_G^{-2}Q^\top\|_{\mathrm{Fr}}^2\,\widetilde{\mathtt{p}}_\Pi$.

B.3   Proof of Theorem 2.15

We assume that all the conditions are fulfilled for the smaller prior covariance G^2, and all error terms correspond to that prior. Corollary 2.10 implies on a set of probability at least $1 - 1/n$ for any measurable set A

$$\left| I\!\!P\big(\boldsymbol{\vartheta}_G - \widetilde{\boldsymbol{\theta}}_G \in A\,\big|\,\boldsymbol{Y}\big) - I\!\!P'\big(\widetilde{D}_G^{-1}\boldsymbol{\gamma} \in A\big) \right| \lesssim \delta_3 + n^{-1}.$$

Similarly, again on a set of probability at least $1 - 1/n$

$$\left| I\!P\big(\vartheta_{G_1} - \widetilde{\boldsymbol{\theta}}_{G_1} \in A \,\big|\, \boldsymbol{Y}\big) - I\!P'\big(\widetilde{D}_{G_1}^{-1}\boldsymbol{\gamma} \in A\big)\right| \lesssim \delta_3 + n^{-1}.$$

Define $\boldsymbol{a} \stackrel{\text{def}}{=} \widetilde{\boldsymbol{\theta}}_G - \widetilde{\boldsymbol{\theta}}_{G_1}$. The bound of Theorem A.1 yields

$$\left| I\!P'\big(\|Q(\widetilde{D}_G^{-1}\boldsymbol{\gamma} - \boldsymbol{a})\| \leq \mathtt{r}\big) - I\!P'\big(\|Q\widetilde{D}_{G_1}^{-1}\boldsymbol{\gamma}\| \leq \mathtt{r}\big)\right|$$

$$\lesssim \frac{1}{\|Q\widetilde{D}_G^{-2}Q^\top\|_{\mathrm{Fr}}}\big\{\mathrm{tr}\big(Q(\widetilde{D}_G^{-2} - \widetilde{D}_{G_1}^{-2})Q^\top\big) + \|Q\boldsymbol{a}\|^2\big\}.$$

Putting all bounds together completes the proof.

B.4   Proof of Lemma 2.1 through 2.3

The condition $G_m^{-2}\big(I - \mathit{\Pi}_{\mathbb{V}_m}\big) = 0$ effectively means that the prior is limited to the subspace $\mathbb{V}_m$ and moreover, its norm is bounded by g_m on this subspace. This implies (2.6) in view of (2.5).

Next we show the statement (2.9) of Lemma 2.2. For $J \geq 1$, by (2.5) and (2.7)

$$\mathrm{tr}\big[H^2\{\mathbb{F}(\boldsymbol{\theta}) + G^2\}^{-1}\mathit{\Pi}_{\mathbb{V}_J}\big] \leq \mathrm{tr}\,\mathit{\Pi}_{\mathbb{V}_J} = J,$$

$$\mathrm{tr}\big[H^2\{\mathbb{F}(\boldsymbol{\theta}) + G^2\}^{-1}\mathit{\Pi}_{\mathbb{V}_J}\big] \geq \mathrm{tr}\big[\mathbb{F}(\boldsymbol{\theta})\{\mathbb{F}(\boldsymbol{\theta}) + g_J^2 I\}^{-1}\mathit{\Pi}_{\mathbb{V}_J}\big]$$

$$\geq \frac{\mathtt{C}_{1,\mathbb{F}}\,n}{\mathtt{C}_{1,\mathbb{F}}\,n + g_J^2}\,\mathrm{tr}\,\mathit{\Pi}_{\mathbb{V}_J} = \frac{J\,\mathtt{C}_{1,\mathbb{F}}\,n}{\mathtt{C}_{1,\mathbb{F}}\,n + g_J^2}.$$

Similarly

$$\mathrm{tr}\big[H^2\{\mathbb{F}(\boldsymbol{\theta}) + G^2\}^{-1}(I - \mathit{\Pi}_{\mathbb{V}_J})\big] \leq \sum_{j \geq J} \frac{\mathtt{C}_{2,\mathbb{F}}\,n}{\mathtt{C}_{2,\mathbb{F}}\,n + g_j^2}.$$

Further, by (2.8)

$$\sum_{j \geq J} \frac{\mathtt{C}_{2,\mathbb{F}}\,n}{\mathtt{C}_{2,\mathbb{F}}\,n + g_j^2} \leq \mathtt{C}_{2,\mathbb{F}}\,n \sum_{j \geq J} \frac{1}{g_j^2} \leq \mathtt{C}_{2,\mathbb{F}}\,n\,\mathtt{C}_{2,g}\,J g_J^{-2}.$$

Therefore,

$$J\,\frac{\mathtt{C}_{1,\mathbb{F}}\,n}{\mathtt{C}_{1,\mathbb{F}}\,n + g_J^2} \leq \mathrm{tr}\big[H^2\{\mathbb{F}(\boldsymbol{\theta}) + G^2\}^{-1}\big] \leq J\big(1 + \mathtt{C}_{2,\mathbb{F}}\,\mathtt{C}_{2,g}\,n g_J^{-2}\big).$$

If J be such that $g_J^2 \asymp n$, then $\mathtt{p}_G(\boldsymbol{\theta}) \asymp J$.

Lemma 2.3 is a special case of Lemma 2.2 with $g_j^2 = w j^{2s}$.

References

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

Castillo, I. and Nickl, R. (2013). Nonparametric Bernstein–von Mises theorems in Gaussian white noise. *Ann. Statist.*, 41(4):1999–2028.

Castillo, I. and Nickl, R. (2014). On the Bernstein – von Mises phenomenon for nonparametric Bayes procedures. *Ann. Statist.*, 42(5):1941–1969.

Castillo, I. and Rousseau, J. (2015). A Bernstein–von Mises theorem for smooth functionals in semiparametric models. *Ann. Statist.*, 43(6):2353–2383.

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. and van der Vaart, A. (2017). *Fundamentals of nonparametric Bayesian inference*, volume 44 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge.

Golubev, Y. and Spokoiny, V. (2009). Exponential bounds for minimum contrast estimators. *Electron. J. Statist.*, 3:712–746.

Götze, F., Naumov, A., Spokoiny, V., and Ulyanov, V. (2019). Gaussian comparison and anti-concentration inequalities for norms of gaussian random elements. *Bernoulli*, 25(4A):2538–2563. arXiv:1708.08663.

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., Szabó, B. T., van der Vaart, A. W., and van Zanten, J. H. (2016). Bayes procedures for adaptive inference in inverse problems for the white noise model. *Probab. Theory Related Fields*, 164(3-4):771–813.

Laurent, B. and Massart, P. (2000). Adaptive estimation of a quadratic functional by model selection. *Ann. Statist.*, 28(5):1302–1338.

Leahu, H. (2011). On the Bernstein-von Mises phenomenon in the Gaussian white noise model. *Electronic J. Statist.*, 5:373–404.

Nickl, R. and Szabó, B. (2016). A sharp adaptive confidence ball for self-similar functions. *Stochastic Processes and their Applications*, 126(12):3913 – 3934. In Memoriam: Evarist Gine.

Sniekers, S. and van der Vaart, A. (2015). Adaptive Bayesian credible sets in regression with a Gaussian process prior. *Electron. J. Stat.*, 9(2):2475–2527.

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

Spokoiny, V. (2017). Penalized maximum likelihood estimation and effective dimension. *AIHP*, 53(1):389–429. arXiv:1205.0498.

Spokoiny, V. (2019). Bayesian inference for nonlinear inverse problems. https://arxiv.org/abs/1912.12694.

van der Vaart, A. (1998). *Asymptotic statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

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

van der Vaart, A. W. and van Zanten, J. H. (2008). Rates of contraction of posterior distributions based on gaussian process priors. *Ann. Statist.*, 36(3):1435–1463.

Yano, K. and Kato, K. (2020). On frequentist coverage errors of bayesian credible sets in moderately high dimensions. *Bernoulli*, 26(1):616–641.