Submitted to Bernoulli

Accuracy of Gaussian approximation for high-dimensional posterior distributions[*]

Vladimir Spokoiny [1] Maxim Panov [2]

[1] *Weierstrass Institute and Humboldt University Berlin*[†]*, IITP RAS, and HSE University Moscow*
E-mail: spokoiny@wias-berlin.de
[2] *Skolkovo Institute of Science and Technology (Skoltech), Skolkovo Innovation Center, Building 3, 143026, Moscow , Russia E-mail:* m.panov@skoltech.ru

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. 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 derive tight finite sample bounds on posterior contraction in terms of the so-called effective dimension of the parameter space and address the question of frequentist reliability of Bayesian credible sets. The obtained results are specified for nonparametric log-density estimation and generalized regression.

Keywords: posterior, concentration, contraction Gaussian approximation

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) $\widetilde{\boldsymbol{\theta}}$ 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;
- coverage: whether one can use credible sets as frequentist confidence sets.

[*] The research was supported by the Russian Science Foundation grant 19-71-30020.

[†] Financial support by the German Research Foundation (DFG) through the Collaborative Research Center 1294 “Data assimilation” is gratefully acknowledged.

The classical versions of the BvM Theorem claim that the posterior concentrates on a root-n vicinity of the true parameter $\boldsymbol{\theta}^{*}$, after proper centering and scaling it is root-n standard normal, and credible sets can be well used as frequentist 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 $\mathbb{R}^{\infty}$ is only defined in a weak sense. In particular, it does not concentrate on any ℓ_2 ball in $\mathbb{R}^{\infty}$. This makes it difficult to study asymptotically the total variation distance between the scaled posterior and the Gaussian law. Even for case of a Gaussian likelihood model and a Gaussian prior when the posterior is exactly Gaussian and completely known, the BvM result is non-trivial and can be stated only after restricting to some parameter subspace; Bontemps (2011); Leahu (2011). We refer to Castillo and Nickl (2013, 2014), and Ghosal and van der Vaart (2017) for more discussion in a non-Gaussian situation and some asymptotic BvM results. Our approach can be called *preasymptotic*: we fix the sample and study the behavior of the centered posterior without any rescaling. Similar approach was recently used in Yano and Kato (2020) for proving frequentist validity of Bayesian credible sets with rectangle shape for linear models with unknown error variance in a moderate or high parameter dimension p. In our approach the parameter dimension can be arbitrary, however, the so called effective dimension has to be relatively small. It appears that the posterior in the case of a Gaussian prior $\mathcal{N}(0, G^{-2})$ is nearly normal $\mathcal{N}(\widetilde{\boldsymbol{\theta}}_G, \widetilde{\mathbb{F}}_G^{-1})$, where $\widetilde{\boldsymbol{\theta}}_G$ is the *penalized maximum likelihood estimator* (pMLE) for the quadratic penalization $\|G\boldsymbol{\theta}\|^2/2$, and $\widetilde{\mathbb{F}}_G = \mathbb{F}(\widetilde{\boldsymbol{\theta}}_G) + G^2$ with $\mathbb{F}(\boldsymbol{\theta})$ being the Fisher information operator at $\boldsymbol{\theta}$. We also establish nonasymptotic upper bounds on concentration and on the error of Gaussian approximation for the posterior in total variation distance in terms of efficient dimension of the problem. The latter bound can be dramatically improved when restricting to the class of centrally symmetric sets around pMLE $\widetilde{\boldsymbol{\theta}}_G$. The approximating Gaussian distribution explicitly depends on the prior precision matrix G^2 and this dependence does not vanish even if the sample size n grows to infinity.

Further we discuss the frequentist validity of Bayesian credible sets. In the contrary to Yano and Kato (2020) we focus on elliptic credible sets. A crucial issue is the bias induced by the prior/penalization. In some situation it can even lead to 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). However, under an undersmoothing condition on the bias we prove that an upper bound on frequentist coverage probability of credible sets.

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 regression, Generalized Linear Modeling, 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. The general results of the paper are illustrated by two specific examples: nonparametric log-density estimation and nonparametric generalized regression. 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 sample* 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 *dimension free* and *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 6.3 of the supplement Spokoiny and Panov (2021).

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 posterior. Section 2.1 collects our conditions and main notations. The central notion of effective dimension is discussed in Section 2.2. Section 2.3 discusses some important properties of the penalized MLE (pMLE) including concentration, Fisher and Wilks expansion, bias-variance decomposition and risk of the pMLE. Section 2.4 presents our main results about Gaussian approximation of the posterior; see Theorem 2.9 and its Corollary 2.10. Section 2.5 addresses the issues of a bias, contraction and frequentist 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. Section 6 of the supplement Spokoiny and Panov (2021) 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. The proofs and auxiliary results are collected in the Appendix.

2. Finite sample properties of the pMLE and posterior

This section discusses general properties of the posterior which can be viewed as 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 $I\!\!R^p$ means a p-dimensional Euclidean space equipped with the norm $\| \cdot \|$, $p \leqslant \infty$. Scalar product in $I\!\!R^p$ is denoted by $\langle \cdot, \cdot \rangle$. For a linear operator B in $I\!\!R^p$, the norm $\|B\|$ means the largest eigenvalue of B.

First we specify our setup. Let $\boldsymbol{Y}$ denote the observed data and $I\!\!P$ mean their distribution. A general parametric assumption (PA) means that $I\!\!P$ belongs to infinite-dimensional family $(I\!\!P_{\boldsymbol{\theta}}, \boldsymbol{\theta} \in \Theta \subseteq I\!\!R^p)$ dominated by a measure $\boldsymbol{\mu}_0$. This family yields the log-likelihood function $L(\boldsymbol{\theta}) = L(\boldsymbol{Y}, \boldsymbol{\theta}) \stackrel{\text{def}}{=} \log \frac{dI\!\!P_{\boldsymbol{\theta}}}{d\boldsymbol{\mu}_0}(\boldsymbol{Y})$. 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 $L(\boldsymbol{\theta})$:

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

If $I\!\!P \notin \bigl(I\!\!P_{\boldsymbol{\theta}} \bigr)$, then the estimate $\tilde{\boldsymbol{\theta}}$ 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 $\exp\bigl\{ L(\boldsymbol{\theta}) \bigr\} \Pi(d\boldsymbol{\theta})$. This relation is usually written as

$$\boldsymbol{\vartheta} \bigm| \boldsymbol{Y} \, \propto \, \exp\bigl\{ L(\boldsymbol{\theta}) \bigr\} \, \Pi(d\boldsymbol{\theta}).$$

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); Yano and Kato (2020). 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 $\exp\bigl\{ L_G(\boldsymbol{\theta}) \bigr\}$ with

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

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

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

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.

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

($\boldsymbol{E}$) *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* $V^{2} = \operatorname{Var}(\nabla \zeta)$ *its covariance.*

$(\boldsymbol{E}\mathsf{H})$ *There exist a positive self-adjoint operator* H *with* $\mathsf{H}^2 \geqslant V^2$, *and constant* $\nu_0 \geqslant 1$ *such that* $\nabla\zeta$ *fulfills*

$$\sup_{\boldsymbol{u} \in \mathbb{R}^p} \log \mathbb{E} \exp\biggl\{ \lambda \frac{\langle \boldsymbol{u}, \mathsf{H}^{-1} \nabla\zeta \rangle}{\| \boldsymbol{u} \|} \biggr\} \leqslant \frac{\nu_0^2 \lambda^2}{2} .$$

Condition $(\boldsymbol{E}\mathsf{H})$ basically requires that the normalized score $\boldsymbol{\xi} = \mathsf{H}^{-1} \nabla\zeta$ is a sub-Gaussian random vector. One can relax this condition to finite exponential moments for $|\lambda| \leqslant \mathtt{g}$ with $\mathtt{g}$ sufficiently large; see (8.6) of Section 8.3 of the supplement Spokoiny and Panov (2021). In fact, $(\boldsymbol{E}\mathsf{H})$ 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 $\mathsf{H} = V$; see Section 3 for more examples.

Apart the basic conditions $(\mathcal{L})$, $(\boldsymbol{E})$, $(\boldsymbol{E}\mathsf{H})$ we need some local properties of the expected log-likelihood $\mathbb{E} L(\boldsymbol{\theta})$. Let us consider some subset $\varTheta^{\circ}$ of the parameter set $\varTheta$ which we call a *local set*. It is required that this set contains the concentration set $\mathcal{A}_G(\mathtt{r}_G)$ of the estimate $\widetilde{\boldsymbol{\theta}}_G$; see Proposition 2.3 below. Define

$$\begin{aligned} \mathbb{F}(\boldsymbol{\theta}) &\stackrel{\operatorname{def}}{=} -\nabla^2 \mathbb{E} L(\boldsymbol{\theta}), \\ \mathbb{F}_G(\boldsymbol{\theta}) &\stackrel{\operatorname{def}}{=} -\nabla^2 \mathbb{E} L_G(\boldsymbol{\theta}) = -\nabla^2 \mathbb{E} L(\boldsymbol{\theta}) + G^2 = \mathbb{F}(\boldsymbol{\theta}) + G^2. \end{aligned}$$

The matrix $\mathbb{F}(\boldsymbol{\theta})$ is usually called the *Fisher information* at $\boldsymbol{\theta}$ for the model $(\mathbb{P}_{\boldsymbol{\theta}})$. The matrix $\mathbb{F}_G(\boldsymbol{\theta})$ sums up the information from the model and from the prior $\mathcal{N}(0, G^{-2})$.

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

$$\mathtt{p}_G(\boldsymbol{\theta}) \stackrel{\operatorname{def}}{=} \operatorname{tr}\bigl\{ \mathsf{H}\, \mathbb{F}_G^{-1}(\boldsymbol{\theta})\, \mathsf{H} \bigr\} < \infty.$$

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

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

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

$(\mathcal{L}_{\mathbf{0}})$ *It holds with* H *from* $(\boldsymbol{E}\mathsf{H})$ *for* $k = 3, 4$ *and some* $\mathtt{r}$ *sufficiently large*

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

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

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

2.2. Effective dimension and examples of priors

Let H^2 be from $(\boldsymbol{E}\mathsf{H})$. The *local effective dimension* $\mathtt{p}_G(\boldsymbol{\theta})$ at $\boldsymbol{\theta} \in \Theta$ is defined by

$$\mathtt{p}_G(\boldsymbol{\theta}) \stackrel{\mathrm{def}}{=} \operatorname{tr}\bigl\{\mathsf{H}^2 I\!\!F_G^{-1}(\boldsymbol{\theta})\bigr\} = \operatorname{tr}\bigl\{\mathsf{H}^2 \bigl(I\!\!F(\boldsymbol{\theta}) + G^2\bigr)^{-1}\bigr\}.$$

Our results involve two particular values:

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

Usually all the values $\mathtt{p}_G(\boldsymbol{\theta})$ are close to $\mathtt{p}_G(\boldsymbol{\theta}_G^*)$ for $\boldsymbol{\theta}$ close to $\boldsymbol{\theta}_G^*$. If $G^2 \equiv 0$ and $\mathsf{H}^2 \approx I\!\!F(\boldsymbol{\theta}^*)$, one obviously gets $\mathtt{p}_G(\boldsymbol{\theta}) \asymp p$. Spokoiny (2017) defined the effective dimension $\mathtt{p}_G$ as $\mathtt{p}_G = \operatorname{tr}\bigl(V^2 I\!\!F_G^{-1}\bigr)$ with $V^2 = \operatorname{Var}(\nabla\zeta)$. Two definitions coincide if $\mathsf{H}^2 = V^2$.

Examples of priors.

In what follows we will consider two non-trivial examples of Gaussian priors: truncation and smooth priors. To make the presentation clear, we impose some assumptions on the considered setup. Most of all are non-restrictive and can be extended to more general situations. Below we assume to be given a growing sequence of nested linear approximation subspaces $I\!\!V_1 \subset I\!\!V_2 \ldots$ in $I\!\!R^p$ of dimension $\dim(I\!\!V_m) = m$. Below $\Pi_{I\!\!V_m}$ is the projector on $I\!\!V_m$ and $I\!\!V_m^c$ is the orthogonal complement of $I\!\!V_m$. A *smooth prior* is described by a self-adjoint operator G such that $\|G\boldsymbol{u}\|/\|\boldsymbol{u}\|$ becomes large for $\boldsymbol{u} \in I\!\!V_m^c$ and m large. One can write this condition in the form

$$\begin{aligned} \|G\boldsymbol{u}\|^2 &\leqslant g_m^2 \|\boldsymbol{u}\|^2, \qquad \boldsymbol{u} \in I\!\!V_m, \\ \|G\boldsymbol{u}\|^2 &\geqslant g_m^2 \|\boldsymbol{u}\|^2, \qquad \boldsymbol{u} \in I\!\!V_m^c. \end{aligned} \tag{2.5}$$

Often one assumes that $I\!\!V_m$ is spanned by the eigenvectors of G^2 corresponding to its smallest eigenvalues $g_1^2 \leqslant g_2^2 \leqslant \ldots \leqslant g_m^2$. We only need (2.5). Further we assume that g_j^2 grow polynomially yielding for some $\mathtt{C}_{1,g}, \mathtt{C}_{2,g}$ and each J

$$\mathtt{C}_{1,g} \leqslant \frac{1}{J g_J^{-2}} \sum_{j \geqslant J} g_j^{-2} \leqslant \mathtt{C}_{2,g}. \tag{2.6}$$

A typical example is given by $G^2 = \operatorname{diag}(g_j^2)$ with $g_j^2 = w^{-1} j^{2s}$ for $s > 1/2$ and some window parameter w. Below we refer to this case as (s,w)-*smooth prior.*

A m-*truncation prior* assumes that the prior distribution is restricted to $I\!\!V_m$. This formally corresponds to a covariance operator G_m^{-2} with $G_m^{-2}\bigl(\boldsymbol{I} - \Pi_{I\!\!V_m}\bigr) = 0$. Equivalently, we set $g_{m+1} = g_{m+2} = \ldots = \infty$ in (2.5).

To evaluate the effective dimension, we assume that the matrices $I\!\!F(\boldsymbol{\theta})$ and H^2 satisfy for each $\boldsymbol{\theta} \in \Theta^\circ$, $m \geqslant 1$, and any $\boldsymbol{u} \in I\!\!V_m$

$$\begin{aligned} \mathtt{C}_{1,I\!\!F}\, n \|\boldsymbol{u}\|^2 &\leqslant \bigl\langle I\!\!F(\boldsymbol{\theta})\boldsymbol{u}, \boldsymbol{u} \bigr\rangle \leqslant \mathtt{C}_{2,I\!\!F}\, n \|\boldsymbol{u}\|^2, \\ \mathtt{C}_{1,I\!\!F}\, n \|\boldsymbol{u}\|^2 &\leqslant \bigl\langle \mathsf{H}^2 \boldsymbol{u}, \boldsymbol{u} \bigr\rangle \leqslant \mathtt{C}_{2,I\!\!F}\, n \|\boldsymbol{u}\|^2. \end{aligned} \tag{2.7}$$

This condition is standard for direct regression or density models; see Section 3. However, our main results do not require this condition.

Lemma 2.1. *Assume* (2.5), (2.6), *and* (2.7). *(1) For a* m *-truncation prior, suppose that* $g_m^2 \leqslant n$. *(2) For a smooth prior, define* m *by* $g_m^2 \approx n$. *Then*

$$\mathtt{C}_3 \, m \leqslant \mathtt{p}_G(\boldsymbol{\theta}) \, \leqslant \, \mathtt{C}_4 \, m, \qquad \boldsymbol{\theta} \in \varTheta^{\circ}, \tag{2.8}$$

where $\mathtt{C}_3$ *and* $\mathtt{C}_4$ *only depend on* $\mathtt{C}_{1,I\!\!F}, \mathtt{C}_{2,I\!\!F}$, *and* $\mathtt{C}_{1,g}, \mathtt{C}_{2,g}$.
(3) For a (s,w) *-smooth prior, the effective dimension is given by* $\mathtt{p}_G(\boldsymbol{\theta}) \asymp (nw)^{1/(2s)}$.

Here and below " $a \asymp b$ " means $a \leqslant \mathtt{C} b$ and $b \leqslant \mathtt{C} a$ with an absolute constant $\mathtt{C}$.

Remark 2.1. Later we will see that under (2.7), a (s,w) -smooth prior yields nearly the same behavior of the posterior as m -truncation prior with $m \asymp (wn)^{1/(2s)}$.

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, $s > 3/2$ for w fixed.

The next result explains how the prior can be linked to "smoothness" of the unknown vector $\boldsymbol{\theta}$. Suppose that the unknown vector $\boldsymbol{\theta}$ belongs to a Sobolev ball $\mathcal{B}(s_0, w_0)$:

$$\mathcal{B}(s_0, w_0) \stackrel{\text{def}}{=} \Bigl\{ \boldsymbol{\theta} = (\theta_j) \colon \sum_{j \geqslant 1} j^{2 s_0} \theta_j^2 \leqslant w_0 \Bigr\}.$$

We present sufficient conditions ensuring that the prior is concentrated on $\varTheta = \mathcal{B}(s_0, w_0)$.

Lemma 2.2. *Let* $\varTheta = \mathcal{B}(s_0, w_0)$, $\mathtt{C}_{s_0} = \sum_j j^{2 s_0} g_j^{-2} < \infty$, $\lambda_{s_0} = \max_j j^{s_0} g_j^{-1}$. *Then*

$$\sqrt{w_0} \, \geqslant \, \sqrt{\mathtt{C}_{s_0}} + \sqrt{2 \lambda_{s_0} \mathtt{x}} \tag{2.9}$$

ensures for $\boldsymbol{\vartheta}_G \sim \mathcal{N}(0, G^{-2})$

$$I\!\!P\bigl(\boldsymbol{\vartheta}_G \in \varTheta \bigr) \, \geqslant \, 1 - 2 \mathrm{e}^{-\mathtt{x}}.$$

For a (s,w) *-smooth prior with* $s > s_0 + 1/2$, *the condition* (2.9) *follows from*

$$\sqrt{w_0 / w} \, \geqslant \, (2s - 2s_0 - 1)^{-1/2} + \sqrt{2 \mathtt{x}}.$$

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

This section presents some useful properties of the pMLE $\widetilde{\boldsymbol{\theta}}_G$ including bias-variance decomposition an asymptotic normality. First we present two concentration bounds for the penalized MLE $\widetilde{\boldsymbol{\theta}}_G = \operatorname{argmax} L_G(\boldsymbol{\theta})$ and for the posterior distribution. Our results are based on conditions $(\mathcal{L})$, $(\boldsymbol{E})$, $(\boldsymbol{E}\mathbf{H})$, $(\mathbf{H}\boldsymbol{G})$, and $(\mathcal{L}_{\mathbf{0}})$ 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 $\boldsymbol{\theta}$. The presented results substantially improve similar statements in Spokoiny (2017).

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

$$\mathcal{A}_G(\mathtt{r}) \stackrel{\text{def}}{=} \bigl\{ \boldsymbol{\theta} \colon \| D_G (\boldsymbol{\theta} - \boldsymbol{\theta}_G^*) \| \leqslant \mathtt{r} \bigr\} \tag{2.10}$$

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

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

$$\|D_G^{-1}\nabla\zeta\| \leqslant z(B_G, \mathtt{x}) \quad \text{on } \Omega(\mathtt{x}) \text{ with } \mathbb{P}\bigl(\Omega(\mathtt{x})\bigr) \geqslant 1 - \mathtt{C}\mathrm{e}^{-\mathtt{x}}, \tag{2.11}$$

where $B_G = \mathsf{H}\, D_G^{-2}\, \mathsf{H}$ and $z(B_G, \mathtt{x})$ is given by (8.10); see Theorem 8.5 of the supplement Spokoiny and Panov (2021) with $\boldsymbol{\xi} = \nabla\zeta$. One can use the simplified bound

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

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 $\boldsymbol{\theta}$.

Proposition 2.3. *Assume (2.11). Let $\mathtt{r}_G$ be such that with $\tau_{3,\mathsf{H}}(\mathtt{r}_G)$ from (2.3)*

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

Then on $\Omega(\mathtt{x})$, the estimate $\widetilde{\boldsymbol{\theta}}_G$ belongs to the set $\mathcal{A}_G(\mathtt{r}_G) \overset{\text{def}}{=} \bigl\{\boldsymbol{\theta}\colon \|D_G(\boldsymbol{\theta} - \boldsymbol{\theta}_G^*)\| \leqslant \mathtt{r}_G\bigr\}$:

$$\bigl\|D_G\bigl(\widetilde{\boldsymbol{\theta}}_G - \boldsymbol{\theta}_G^*\bigr)\bigr\| \leqslant \mathtt{r}_G\,. \tag{2.14}$$

Remark 2.2. 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 $\mathtt{r}_G \leqslant 2z(B_G, \mathtt{x})$. Under (2.4) $\tau_{3,\mathsf{H}}(\mathtt{r}_G) \lesssim n^{-1/2}$ while $z^2(B_G, \mathtt{x}) \asymp \mathtt{p}_G = \operatorname{tr}(B_G)$; see (2.12) and examples in Section 3. Therefore, $\mathtt{r}_G\,\tau_{3,\mathsf{H}}(\mathtt{r}_G) \asymp (\mathtt{p}_G/n)^{1/2}$, 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 2.3, 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 $L_G(\widetilde{\boldsymbol{\theta}}_G) - L_G(\boldsymbol{\theta}_G^*)$.

Proposition 2.4. *Under the conditions of Proposition 2.3, it holds on $\Omega(\mathtt{x})$*

$$\bigl\|D_G\bigl(\widetilde{\boldsymbol{\theta}}_G - \boldsymbol{\theta}_G^*\bigr) - D_G^{-1}\nabla\zeta\bigr\|^2 \leqslant 4\mathtt{r}_G^3\,\tau_{3,\mathsf{H}}\,, \tag{2.15}$$

$$\Bigl|L_G(\widetilde{\boldsymbol{\theta}}_G) - L_G(\boldsymbol{\theta}_G^*) - \frac{1}{2}\bigl\|D_G^{-1}\nabla\zeta\bigr\|^2\Bigr| \leqslant \mathtt{r}_G^3\,\tau_{3,\mathsf{H}}\,, \tag{2.16}$$

with $\tau_{3,\mathsf{H}}$ from (2.3). Also

$$\Bigl|L_G(\widetilde{\boldsymbol{\theta}}_G) - L_G(\boldsymbol{\theta}_G^*) - \frac{1}{2}\bigl\|D_G(\widetilde{\boldsymbol{\theta}}_G - \boldsymbol{\theta}_G^*)\bigr\|^2\Bigr| \leqslant \mathtt{r}_G^3\,\tau_{3,\mathsf{H}},$$

$$\sup_{\boldsymbol{\theta}\in\mathcal{A}_G(\mathtt{r}_G)}\Bigl|L_G(\widetilde{\boldsymbol{\theta}}_G) - L_G(\boldsymbol{\theta}) - \frac{1}{2}\bigl\|\widetilde{D}_G(\widetilde{\boldsymbol{\theta}}_G - \boldsymbol{\theta})\bigr\|^2\Bigr| \leqslant \mathtt{r}_G^3\,\tau_{3,\mathsf{H}}, \tag{2.17}$$

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

$$\bigl\| D_G^{-1} \bigl(\widetilde{D}_G^2 - D_G^2 \bigr) D_G^{-1} \bigr\| \leqslant \mathtt{C} \mathtt{r}_G \, \tau_{3,\mathsf{H}} . \tag{2.18}$$

Remark 2.3. The results of Proposition 2.4 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 2.6 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 2.4. Similarly to Proposition 2.3, the results of Proposition 2.4 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 $\bigl\| G\boldsymbol{\theta} \bigr\|^2$ yields some estimation bias measured by $\mathbb{E} L_G(\boldsymbol{\theta}_G^*) - \mathbb{E} L_G(\boldsymbol{\theta}^*)$ 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 2.5. *It holds*

$$\mathbb{E} L_G(\boldsymbol{\theta}_G^*) - \mathbb{E} L_G(\boldsymbol{\theta}^*) \leqslant \| G\boldsymbol{\theta}^* \|^2 / 2.$$

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

$$\begin{aligned} \Bigl| \mathbb{E} L_G(\boldsymbol{\theta}_G^*) - \mathbb{E} L_G(\boldsymbol{\theta}^*) - \bigl\| D_G(\boldsymbol{\theta}_G^* - \boldsymbol{\theta}^*) \bigr\|^2/2 \Bigr| &\leqslant \delta_3(\mathtt{r}_b), \\ \bigl\| D_G(\boldsymbol{\theta}^* - \boldsymbol{\theta}_G^*) - D_G^{-1} G^2 \boldsymbol{\theta}^* \bigr\|^2 &\leqslant 4\delta_3(\mathtt{r}_b) \end{aligned} \tag{2.19}$$

with $\delta_3(\mathtt{r}_b) = \mathtt{r}_b^3 \, \tau_{3,\mathsf{H}}(\mathtt{r}_b)$. *Moreover, for any linear mapping* Q *in* $\mathbb{R}^p$ *it holds*

$$\| Q(\boldsymbol{\theta}_G^* - \boldsymbol{\theta}^*) \| \leqslant \| Q D_G^{-2} G^2 \boldsymbol{\theta}^* \| + 2\sqrt{\| Q D_G^{-2} Q^\top \| \, \delta_3(\mathtt{r}_b)} . \tag{2.20}$$

Now we can combine all the previous results together to bound the loss $\| Q(\widetilde{\boldsymbol{\theta}}_G - \boldsymbol{\theta}^*) \| = \| Q(\widetilde{\boldsymbol{\theta}}_G - \boldsymbol{\theta}_G^*) + Q(\boldsymbol{\theta}_G^* - \boldsymbol{\theta}^*) \|$. One can apply $Q = \sqrt{\mathbb{F}(\boldsymbol{\theta}^*)}$ for prediction and $Q = \sqrt{n} \, \boldsymbol{I}_p$ for estimation. We state two results. The first bound can be viewed as analog of classical bias-variance decomposition of the loss $\bigl\| Q(\widetilde{\boldsymbol{\theta}}_G - \boldsymbol{\theta}^*) \bigr\|$. 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. We assume $\mathtt{x} = \log n$, where n means the sample size tending to ∞. Also "$a \lesssim b$" means $a \leqslant \mathtt{C} b$ 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 2.6. *Assume the conditions of Proposition 2.3 with* $\mathtt{x} = \log n$. *Given* Q *with* $Q^\top Q \leqslant D_G^2$, *define* $B_{Q|G} = Q D_G^{-2} \mathsf{H}^2 D_G^{-2} Q^\top$.
(1) On a random set $\Omega_1(\mathtt{x})$ *with* $\mathbb{P}\bigl(\Omega_1(\mathtt{x})\bigr) \leqslant 2\mathrm{e}^{-\mathtt{x}}$, *it holds*

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

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

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

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

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

(2) Let $\nabla\zeta$ *be nearly normal in the sense that with* $V^2 = \operatorname{Var}(\nabla\zeta)$

$$\sup_{\boldsymbol{a}\in I\!\!R^p}\sup_{z>0}\Bigl| I\!\!P\bigl(\|QD_G^{-2}\nabla\zeta - \boldsymbol{a}\| \leqslant z\bigr) - I\!\!P\bigl(\|V_{Q|G}\boldsymbol{\gamma} - \boldsymbol{a}\| \leqslant z\bigr)\Bigr| = o(1), \tag{2.23}$$

where $\boldsymbol{\gamma} \in I\!\!R^p$ *standard normal and* $V_{Q|G}^2 = \operatorname{Var}\bigl(QD_G^{-2}\nabla\zeta\bigr) = QD_G^{-2}V^2D_G^{-2}Q^\top$ *. Assume also a "small bias" condition*

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

Then it holds

$$\sup_{z>0}\Bigl| I\!\!P\bigl(\|Q(\widetilde{\boldsymbol{\theta}}_G - \boldsymbol{\theta}^*)\| \leqslant z\bigr) - I\!\!P\bigl(\|V_{Q|G}\boldsymbol{\gamma}\| \leqslant z\bigr)\Bigr| = o(1). \tag{2.25}$$

Remark 2.5. Given α, define z_α by $I\!\!P(\|V_{Q|G}\boldsymbol{\gamma}\| \geqslant z_\alpha) = \alpha$. It follows from (2.25) that $\mathcal{E}_{Q|G}(z_\alpha) \stackrel{\text{def}}{=} \bigl\{\boldsymbol{\theta}\colon \|Q(\widetilde{\boldsymbol{\theta}}_G - \boldsymbol{\theta})\| \leqslant z_\alpha\bigr\}$ is asymptotically valid confidence set for $\boldsymbol{\theta}^*$.

Remark 2.6. 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).

Proof. The statement (2.21) of Theorem 2.6 follows from the bound (2.20) on the bias $\|Q(\boldsymbol{\theta}_G^* - \boldsymbol{\theta}^*)\|$ and the deviation bound $\|QD_G^{-2}\nabla\zeta\| \leqslant z(B_{Q|G},\mathtt{x})$ on a set of probability at least $1 - \mathrm{e}^{-\mathtt{x}}$ by the triangle inequality. For the second statement of Theorem 2.6, we apply the decomposition

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

with $\boldsymbol{a} = \boldsymbol{\theta}^* - \boldsymbol{\theta}_G^*$. Now the result follows from asymptotic normality (2.23) and Gaussian comparison result of Theorem 8.1 in view of small bias condition (2.24). □

Examples of priors

Here we consider the case of m-truncation or (s,w)-smooth priors.

Corollary 2.7. *Assume the conditions of Proposition* 2.3 *and* (2.7)*. Let* $\boldsymbol{\theta}^* \in \mathcal{B}(s_0,w_0)$. *Set* $\mathtt{x} = \log n$ *and consider (1) a* m*-truncation prior with* $m = m_0 \stackrel{\text{def}}{=} (w_0\, n)^{1/(2s_0+1)}$ *; (2) a* (s,w)*-smooth* *prior with* $s > s_0 + 1/2$ *and* $(w\,n)^{1/(2s)} = (w_0\,n)^{1/(2s_0+1)} = m_0$. *Then on a random set* Ω_n *with* $\mathbb{P}\bigl(\Omega_n\bigr) \geqslant 1 - 1/n$, *it holds for any* $Q \leqslant \boldsymbol{I}$

$$\bigl\| Q(\widetilde{\boldsymbol{\theta}}_G - \boldsymbol{\theta}^*) \bigr\|^2 \leqslant \mathtt{C}_1 w_0^{1/(2s_0+1)} n^{-2s_0/(2s_0+1)} + \mathtt{C}_2 n^{-1} \log n$$

for some fixed $\mathtt{C}_1, \mathtt{C}_2$. *Let, in addition, asymptotic normality condition* (2.35) *hold and*

$$(1) \quad \frac{(w_0\,n)^{1/(2s_0+1)}}{m} = o(1); \qquad (2) \quad \frac{(w_0\,n)^{1/(2s_0+1)}}{(w\,n)^{1/(2s)}} = o(1). \tag{2.26}$$

Then the statement (2.25) *holds as well.*

2.4. Gaussian approximation of posterior distributions

Theorem 2.9 below presents the main result of the paper about Gaussian approximation of the posterior with an explicit error term.

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 $\mathcal{E}_G(\mathtt{r}_0) = \bigl\{\boldsymbol{\theta} \colon \|\mathsf{H}(\boldsymbol{\theta} - \widetilde{\boldsymbol{\theta}}_G)\| \leqslant \mathtt{r}_0\bigr\}$ for a proper value $\mathtt{r}_0 \geqslant \mathtt{C}\sqrt{\widetilde{\mathtt{p}}_G} + \mathtt{C}\sqrt{\mathtt{x}}$ for $\widetilde{\mathtt{p}}_G = \mathtt{p}_G(\widetilde{\boldsymbol{\theta}}_G)$. For this we bound from above the quantity

$$\rho(\mathtt{r}_0) \stackrel{\text{def}}{=} \frac{\int \mathbb{1}\bigl(\|\mathsf{H}\boldsymbol{u}\| > \mathtt{r}_0\bigr) \exp\bigl\{L_G(\widetilde{\boldsymbol{\theta}}_G + \boldsymbol{u})\bigr\} d\boldsymbol{u}}{\int \mathbb{1}\bigl(\|\mathsf{H}\boldsymbol{u}\| \leqslant \mathtt{r}_0\bigr) \exp\bigl\{L_G(\widetilde{\boldsymbol{\theta}}_G + \boldsymbol{u})\bigr\} d\boldsymbol{u}} . \tag{2.27}$$

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

Proposition 2.8. *Assume* (2.11). *Assume that for some fixed values* $\mathtt{r}_0$ *and* $\mathtt{x} > 0$, *it holds with* $\delta_3(\mathtt{r}_0) = \mathtt{r}_0^3 \tau_{3,\mathsf{H}}(\mathtt{r}_0)$ *and* $\delta_4(\mathtt{r}_0) = \mathtt{r}_0^4 \tau_{4,\mathsf{H}}(\mathtt{r}_0)$

$$\diamondsuit(\mathtt{r}_0) \stackrel{\text{def}}{=} 4\delta_3^2(\mathtt{r}_0) + 4\delta_4(\mathtt{r}_0) \leqslant 1/2, \qquad \mathtt{C}_0 \stackrel{\text{def}}{=} 1 - 3\mathtt{r}_0\,\tau_{3,\mathsf{H}}(\mathtt{r}_0) \geqslant 1/2, \tag{2.28}$$

and also

$$\mathtt{C}_0 \mathtt{r}_0 \geqslant 2\sqrt{\mathtt{p}_G(\boldsymbol{\theta})} + \sqrt{\mathtt{x}}\,, \qquad \boldsymbol{\theta} \in \Theta^{\circ} . \tag{2.29}$$

Then, on the random set $\Omega(\mathtt{x})$ *from* (2.11), $\rho(\mathtt{r}_0)$ *from* (2.27) *fulfills with* $\widetilde{\mathtt{p}}_G = \mathtt{p}_G(\widetilde{\boldsymbol{\theta}}_G)$

$$\rho(\mathtt{r}_0) \leqslant \frac{1}{1 - \diamondsuit(\mathtt{r}_0)} \, \frac{\exp\bigl\{-(\widetilde{\mathtt{p}}_G + \mathtt{x})/2\bigr\}}{1 - \exp\{-(\widetilde{\mathtt{p}}_G + \mathtt{x})/2\}} . \tag{2.30}$$

Remark 2.7. Conditions (2.28) and (2.29) can be spelled out as follows: the value $\mathtt{r}_0^2$ should be larger than $\mathtt{C}\mathtt{p}_G(\boldsymbol{\theta})$, while the values $\delta_3(\mathtt{r}_0)$ and $\delta_4(\mathtt{r}_0)$ should be small. If $\tau_{3,\mathsf{H}}(\mathtt{r}_0) \asymp n^{-1/2}$, $\tau_{4,\mathsf{H}}(\mathtt{r}_0) \asymp n^{-1}$, then (2.28) requires $\mathtt{p}_G^3(\boldsymbol{\theta}) \ll n$, $\boldsymbol{\theta} \in \Theta^{\circ}$.

Remark 2.8. Let us compare the concentration result of Proposition 2.3 for the pMLE $\widetilde{\boldsymbol{\theta}}_G$ and the concentration bound (2.30) for the posterior $\boldsymbol{\vartheta}_G \,\big|\, \boldsymbol{Y}$. 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 $\boldsymbol{\vartheta}_G \,\big|\, \boldsymbol{Y}$ look similar, but there is one essential difference. The properly shifted MLE $\widetilde{\boldsymbol{\theta}}_G$ well concentrates on a rather small elliptic set $\mathcal{A}_G(\mathtt{r}_G)$ from (2.10) centered at $\boldsymbol{\theta}^*_G$. In other words, $D_G\bigl(\widetilde{\boldsymbol{\theta}}_G - \boldsymbol{\theta}^*_G\bigr)$ belongs to the ball in $\mathbb{R}^p$ of radius $\mathtt{r}_G$ with a high probability. This holds true even if $p = \infty$. The result of Proposition 2.8 claims concentration of $\boldsymbol{\vartheta}_G \,\big|\, \boldsymbol{Y}$ 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 $D_G(\boldsymbol{\vartheta}_G - \widetilde{\boldsymbol{\theta}}_G) \,\big|\, \boldsymbol{Y}$ 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 $\mathcal{N}\bigl(\widetilde{\boldsymbol{\theta}}_G, \widetilde{D}_G^{-2}\bigr)$ with $\widetilde{D}_G^2 = \mathbb{F}(\widetilde{\boldsymbol{\theta}}_G) + G^2$. By $\mathbb{P}'$ we denote a standard normal distribution of a random vector $\boldsymbol{\gamma} \in \mathbb{R}^p$ given $\widetilde{D}_G$. 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. It holds on the set $\Omega(\mathtt{x})$ from (2.11) for any centrally symmetric Borel set $A \in \mathcal{B}_s(\mathbb{R}^p)$*

$$
\begin{aligned}
\mathbb{P}\bigl(\boldsymbol{\vartheta}_G - \widetilde{\boldsymbol{\theta}}_G \in A \,\big|\, \boldsymbol{Y}\bigr) &\geqslant \frac{1 - \diamondsuit(\mathtt{r}_0)}{\bigl\{1 + \diamondsuit(\mathtt{r}_0) + \rho(\mathtt{r}_0)\bigr\}} \, \mathbb{P}'\bigl(\widetilde{D}_G^{-1}\boldsymbol{\gamma} \in A\bigr) - \rho(\mathtt{r}_0), \\
\mathbb{P}\bigl(\boldsymbol{\vartheta}_G - \widetilde{\boldsymbol{\theta}}_G \in A \,\big|\, \boldsymbol{Y}\bigr) &\leqslant \frac{1 + \diamondsuit(\mathtt{r}_0)}{\bigl\{1 - \diamondsuit(\mathtt{r}_0)\bigr\}\bigl(1 - \mathrm{e}^{-\mathtt{x}}\bigr)} \mathbb{P}'\bigl(\widetilde{D}_G^{-1}\boldsymbol{\gamma} \in A\bigr) + \rho(\mathtt{r}_0).
\end{aligned}
$$

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

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

$$
\Bigl| \mathbb{P}\bigl(\boldsymbol{\vartheta}_G - \widetilde{\boldsymbol{\theta}}_G \in A \,\big|\, \boldsymbol{Y}\bigr) - \mathbb{P}'\bigl(\widetilde{D}_G^{-1}\boldsymbol{\gamma} \in A\bigr) \Bigr| \lesssim \mathbb{P}'\bigl(\widetilde{D}_G^{-1}\boldsymbol{\gamma} \in A\bigr)\bigl\{\diamondsuit(\mathtt{r}_0) + \mathrm{e}^{-\mathtt{x}}\bigr\} + \rho(\mathtt{r}_0).
$$

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 $\mathtt{r}_0$ satisfies the conditions (2.28) and (2.29) with $\mathtt{x} = 2\log n$. It holds on $\Omega(\mathtt{x})$*

$$
\sup_{A \in \mathcal{B}_s(\mathbb{R}^p)} \Bigl| \mathbb{P}\bigl(\boldsymbol{\vartheta}_G - \widetilde{\boldsymbol{\theta}}_G \in A \,\big|\, \boldsymbol{Y}\bigr) - \mathbb{P}'\bigl(\widetilde{D}_G^{-1}\boldsymbol{\gamma} \in A\bigr) \Bigr| \lesssim \diamondsuit(\mathtt{r}_0) + 1/n, \tag{2.31}
$$

$$
\sup_{A \in \mathcal{B}(\mathbb{R}^p)} \Bigl| \mathbb{P}\bigl(\boldsymbol{\vartheta}_G - \widetilde{\boldsymbol{\theta}}_G \in A \,\big|\, \boldsymbol{Y}\bigr) - \mathbb{P}'\bigl(\widetilde{D}_G^{-1}\boldsymbol{\gamma} \in A\bigr) \Bigr| \lesssim \delta_3(\mathtt{r}_0) + 1/n. \tag{2.32}
$$

Remark 2.9. Comparison of two bounds of Corollary 2.10 reveals that the use of symmetric credible sets improves the accuracy of Gaussian approximation from $\delta_3(\mathtt{r}_0) = \mathtt{r}_0^3 \tau_{3,\mathsf{H}}$ to $\diamondsuit(\mathtt{r}_0) \asymp 4\delta_3^2(\mathtt{r}_0) + \delta_4(\mathtt{r}_0)$. In particular, under (2.4) and (2.7), $\tau_{3,\mathsf{H}}(\mathtt{r}_0) \asymp n^{-1/2}$ while $\diamondsuit(\mathtt{r}_0) \asymp \mathtt{r}_0^6/n$. The choice $\mathtt{x} =$

$2\log n$ and $\mathtt{r}_0 = \mathtt{C}\bigl(\sqrt{\mathtt{p}_G} + \sqrt{\log n}\bigr)$ yields that the leading term in the error of Gaussian approximation (2.31) is $\diamondsuit(\mathtt{r}_0) \asymp \mathtt{p}_G^6/n$. The bound (2.32) in TV-distance ensures an error of order $\delta_3(\mathtt{r}_0) \asymp \sqrt{\mathtt{p}_G^3/n}$.

Remark 2.10. The approximating Gaussian distribution $\mathcal{N}(\widetilde{\boldsymbol{\theta}}_G, \widetilde{D}_G^{-2})$ of the posterior depends on the prior through the penalized MLE $\widetilde{\boldsymbol{\theta}}_G$ and the matrix $\widetilde{D}_G^2 = I\!\!F(\widetilde{\boldsymbol{\theta}}_G) + G^2$. In the contrary to the parametric situation, the impact of the prior is essential and does not vanish as the sample size n increases. The random matrix $\widetilde{D}_G^2$ is close to D_G^2 on a set of dominating probability; see Proposition 2.4.

2.5. Posterior contraction and frequentist coverage probability

This section presents more results about of the posterior $\boldsymbol{\vartheta}_G \,\big|\, \boldsymbol{Y}$. The main result describes the frequentist coverage probability of Bayesian credible sets. We systematically consider the posterior of $Q\boldsymbol{\vartheta}_G$ for a given linear mapping $Q\colon I\!\!R^p \to I\!\!R^q$. This includes the case of estimating the whole vector $\boldsymbol{\theta}$ for $Q = I\!\!I_p$, a subvector of $\boldsymbol{\theta}$ for Q being a projector on a parameter subspace, or estimation of linear functionals for $q = 1$. We aim at describing the distance between the support of the posterior and the true value $\boldsymbol{\theta}^*$. For a given linear mapping from $I\!\!R^p$ satisfying $Q^\top Q \leqslant D_G^2$, we would like to describe a minimal radius $\mathtt{r}$ ensuring

$$I\!\!P\bigl(\|Q(\boldsymbol{\vartheta}_G - \boldsymbol{\theta}^*)\| > \mathtt{r} \,\big|\, \boldsymbol{Y}\bigr) \;=\; o(1).$$

Theorem 2.11. *Assume the conditions of Proposition 2.8. Let also $Q^\top Q \leqslant D_G^2$, $\mathsf{H}^2 \leqslant D_G^2$, and the "bias-variance" relation hold*

$$\|Q D_G^{-2} G^2 \boldsymbol{\theta}^*\|^2 \;\leqslant\; \mathtt{C}\operatorname{tr}\bigl(Q D_G^{-2} Q^\top\bigr) \tag{2.33}$$

for some fixed constant $\mathtt{C}$. Then for $\mathtt{x} = \log n$ on $\Omega(\mathtt{x})$ with some fixed $\mathtt{C}_1, \mathtt{C}_2, \mathtt{C}_3$

$$I\!\!P\Bigl(\bigl\|Q(\boldsymbol{\vartheta}_G - \boldsymbol{\theta}^*)\bigr\|^2 \geqslant \mathtt{C}_1 \operatorname{tr}\bigl(Q D_G^{-2} Q^\top\bigr) + \mathtt{C}_2 \log n \, \|Q D_G^{-2} Q^\top\| \;\Big|\, \boldsymbol{Y}\Bigr) \;\leqslant\; \mathtt{C}_3\, n^{-1}. \tag{2.34}$$

Remark 2.11. If $Q = \mathsf{H}$, then $\operatorname{tr}\bigl(Q D_G^{-2} Q^\top\bigr) = \mathtt{p}_G$ and the contraction radius is of order $\sqrt{\mathtt{p}_G} + \sqrt{\log n}$. The relation (2.33) is called "bias-variance trade-off", and it means that the squared bias $\bigl\|Q\bigl(\boldsymbol{\theta}_G^* - \boldsymbol{\theta}^*\bigr)\bigr\|^2$ is not larger in order than the trace of the variance of the posterior $Q\boldsymbol{\vartheta}_G \,\big|\, \boldsymbol{Y}$; see Proposition 2.5. The bound (2.34) on posterior contraction is sharp in the sense that it cannot be improved even in the Gaussian case.

Here we address the issue of *frequentist coverage* of Bayesian credible sets. Consider credible sets of the form

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

where $\widetilde{\mathtt{r}} = \widetilde{\mathtt{r}}_\alpha$ is a random radius ensuring

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

Frequentist validity of such credible sets means that $\mathcal{E}_{Q|G}(\tilde{\mathtt{r}}_\alpha)$ can be used as a confidence set ensuring nominal coverage level: $\mathbb{P}\bigl(\boldsymbol{\theta}^* \notin \mathcal{E}_{Q|G}(\tilde{\mathtt{r}}_\alpha)\bigr) \approx \alpha$. The posterior distribution is nearly normal with mean $\widetilde{\boldsymbol{\theta}}_G$. Therefore, frequentist validity of $\mathcal{E}_{Q|G}(\tilde{\mathtt{r}}_\alpha)$ can be reduced to questions of asymptotic normality of $\widetilde{\boldsymbol{\theta}}_G$ and to "small bias" condition which ensures that $\mathbb{E}\widetilde{\boldsymbol{\theta}}_G \approx \boldsymbol{\theta}^*$. We state the result in asymptotic form assuming that the sample size n tends to infinity.

Theorem 2.12. *Assume the conditions of Proposition 2.8. Let $\nabla\zeta$ be nearly normal in the sense that with $V^2 = \operatorname{Var}(\nabla\zeta)$*

$$\sup_{\boldsymbol{a}\in\mathbb{R}^p}\sup_{z>0}\Bigl|\mathbb{P}\bigl(\|QD_G^{-2}\nabla\zeta - \boldsymbol{a}\| \leqslant z\bigr) - \mathbb{P}\bigl(\|V_{Q|G}\boldsymbol{\gamma} - \boldsymbol{a}\| \leqslant z\bigr)\Bigr| = o(1), \tag{2.35}$$

where $\boldsymbol{\gamma} \in \mathbb{R}^p$ standard normal and $V_{Q|G}^2 = \operatorname{Var}\bigl(QD_G^{-2}\nabla\zeta\bigr) = QD_G^{-2}V^2D_G^{-2}Q^\top$. Under a "small bias" condition

$$\frac{\|QD_G^{-2}G^2\boldsymbol{\theta}^*\|^2}{\operatorname{tr}\bigl(V_{Q|G}^2\bigr)} \leqslant o(1), \tag{2.36}$$

it holds

$$\mathbb{P}\bigl(\boldsymbol{\theta}^* \notin \mathcal{E}_{Q|G}(\tilde{\mathtt{r}}_\alpha)\bigr) \leqslant \alpha + o(1). \tag{2.37}$$

Remark 2.12. Classical Bernstein - von Mises results particularly yield that Bayesian credible sets can be used as valid and asymptotically accurate frequentist confidence sets. One of the reasons behind this statement is that the impact of the prior is asymptotically vanishing in the posterior distribution. In the nonparametric setup of this paper we can only state a lower bound on the coverage probability of $\mathcal{E}_{Q|G}(\tilde{\mathtt{r}}_\alpha)$: it is not significantly smaller than the nominal probability $1-\alpha$. The reason is that the posterior covariance $Q\widetilde{D}_G^{-2}Q^\top$ is in general larger than the covariance $V_{Q|G}^2 = QD_G^{-2}V^2D_G^{-2}Q^\top$ of $Q\widetilde{\boldsymbol{\theta}}_G$; see (2.23) of Theorem 2.6. The only exception corresponds to the case of truncation priors; cf. Yano and Kato (2020). That paper also evaluates the error term $o(1)$ in (2.37) for linear models and rectangle credible sets in term of sample size n and the dimension p which can be moderately large. To state a similar explicit error bound we additionally need to quantify the accuracy in the CLT result (2.35).

Now we illustrate the *bias-variance trade-off* (2.33) or the small bias condition (2.36) for the case $Q \leqslant \boldsymbol{I}_p$ and for a m-truncation or (s,w)-smooth priors.

Corollary 2.13. *Assume the conditions of Proposition 2.8 and (2.7). Let $\boldsymbol{\theta}^* \in \mathcal{B}(s_0,w_0)$. Set $\mathtt{x} = \log n$ and consider (1) a m-truncation prior with $m = m_0 \overset{\text{def}}{=} (w_0\,n)^{1/(2s_0+1)}$; (2) a (s,w)-smooth prior with $s > s_0 + 1/2$ and $(w\,n)^{1/(2s)} = (w_0\,n)^{1/(2s_0+1)} = m_0$. Then on a random set Ω_n with $\mathbb{P}\bigl(\Omega_n\bigr) \geqslant 1 - 1/n$, it holds for any $Q \leqslant \boldsymbol{I}_p$*

$$\mathbb{P}\Bigl(\bigl\|Q(\boldsymbol{\vartheta}_G - \boldsymbol{\theta}^*)\bigr\|^2 \geqslant \mathtt{C}_1 w_0^{1/(2s_0+1)} n^{-2s_0/(2s_0+1)} + \mathtt{C}_2 n^{-1}\log n \,\Bigm|\, \boldsymbol{Y}\Bigr) \leqslant \mathtt{C}\, n^{-1} \tag{2.38}$$

for some fixed $\mathtt{C}, \mathtt{C}_1, \mathtt{C}_2$. Let also the asymptotic normality condition (2.35) and undersmoothing condition (2.26) hold. Then the statement (2.37) applies as well.

Remark 2.13. Truncation at $m = m_0 = (w_0 n)^{1/(2s_0+1)}$ is referred to as rate optimal choice yielding the rate optimal accuracy of estimation $n^{-2s_0/(2s_0+1)}$. The choice $m \gg m_0$ is called *undersmoothing* and it ensures that the stochastic part $Q(\boldsymbol{\vartheta}_G - \boldsymbol{\theta}_G^*)$ of the loss $Q(\boldsymbol{\vartheta}_G - \boldsymbol{\theta}^*)$ dominates and the bias term $Q(\boldsymbol{\theta}_G^* - \boldsymbol{\theta}^*)$ is negligible. This enables us to state the result of Teorem 2.12 about frequentist validity of Bayesian credible sets. For a (s, w)-smooth prior, the only important parameter is the index m for which $g_m^2 = w^{-1} m^{2s} \approx n$. Under the relation $m = m_0$, we again obtain the optimal contraction rate as in (2.38). For any $s > s_0 + 1/2$, one can fix the value $w = m_0^{2s}/n$ ensuring an accurate contraction rate. Selecting $w \gg m_0^{2s}/n$ yields undersmoothing (2.26) and thus, (2.37). By choosing a proper projector Q, one can derive the results similar to Bontemps (2011) and Leahu (2011) for non-Gaussian case.

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 $p \leqslant \infty$ be a collection of functions in $\mathbb{R}^d$ (a dictionary). For each $\boldsymbol{\theta} = (\theta_j) \in \mathbb{R}^p$, define

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

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

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

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

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

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

$$L(\boldsymbol{\theta}) = \sum_{i=1}^{n} \ell(X_i, \boldsymbol{\theta}) = \sum_{i=1}^{n} \bigl\langle \Psi(X_i), \boldsymbol{\theta} \bigr\rangle - n\phi(\boldsymbol{\theta}) = \langle S, \boldsymbol{\theta} \rangle - n\phi(\boldsymbol{\theta}), \quad S = \sum_{i=1}^{n} \Psi(X_i) .$$

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 automatically yields conditions $(\mathcal{L})$ and $(\boldsymbol{E})$. 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

$$\nabla\zeta(\boldsymbol{\theta}) \; = \; \nabla\zeta = S - \mathbb{E}S = \sum_{i=1}^{n} \bigl[\Psi(X_i) - \mathbb{E}\Psi(X_i) \bigr],$$

and $(\boldsymbol{E})$ 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

$$\mathbb{E}S \; = \; \sum_{i=1}^{n} \mathbb{E}\Psi(X_i) = n\, \mathbb{E}\Psi(X_1) = n\overline{\Psi}$$

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

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

Now we switch to the Bayesian framework and restrict ourselves to m-truncation or (s, w)-smooth priors; see Section 2.2. For a given penalty operator G^2, the corresponding penalized MLE $\widetilde{\boldsymbol{\theta}}_G$, and the target $\boldsymbol{\theta}^{*}_G$ are

$$\begin{aligned} \widetilde{\boldsymbol{\theta}}_G \; &= \; \operatorname*{argmax}_{\boldsymbol{\theta} \in \Theta} L_G(\boldsymbol{\theta}) = \operatorname*{argmax}_{\boldsymbol{\theta} \in \Theta} \Bigl\{ \langle \boldsymbol{\theta}, S \rangle - n\phi(\boldsymbol{\theta}) - \frac{1}{2} \| G\boldsymbol{\theta} \|^2 \Bigr\}, \\ \boldsymbol{\theta}^{*}_G \; &= \; \operatorname*{argmax}_{\boldsymbol{\theta} \in \Theta} \mathbb{E}L_G(\boldsymbol{\theta}) = \operatorname*{argmax}_{\boldsymbol{\theta} \in \Theta} \Bigl\{ \langle \boldsymbol{\theta}, \mathbb{E}S \rangle - n\phi(\boldsymbol{\theta}) - \frac{1}{2} \| G\boldsymbol{\theta} \|^2 \Bigr\}. \end{aligned}$$

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

$(\boldsymbol{\Theta})$ $\Theta = \mathcal{B}_{s_0}(w_0)$ for $s_0 > 1$ yielding $\sum_j \theta_j^2 j^{2s_0} \leqslant w_0$ for all $\boldsymbol{\theta} = (\theta_j)$.

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

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

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

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) $(\psi_j) = \bigl(\psi_{kl}\bigr)_{k \geqslant 0, l \in \mathcal{I}_k}$ is a double indexed set of wavelet functions satisfying

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

In what follows, for some small but fixed ϱ, we use $\varTheta^\circ$ of the form

$$\varTheta^\circ = \varTheta_\varrho \,=\, \bigl\{ \boldsymbol{\theta} = \boldsymbol{\theta}^* + \boldsymbol{u} \colon \bigl\langle \nabla^2 \phi(\boldsymbol{\theta}) \boldsymbol{u}, \boldsymbol{u} \bigr\rangle \leqslant \varrho^2 \bigr\} .$$

$(\boldsymbol{\nabla^2 \phi})$ For some $\mathtt{C}_{\phi,1}, \mathtt{C}_{\phi,2} \geqslant 1$ and for all $\boldsymbol{\theta} \in \varTheta^\circ$ and all $\boldsymbol{u} \in I\!\!R^p$

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

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

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

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

The identity (3.1) ensures that $P_{\boldsymbol{\theta}}$ is a probabilistic measure. Moreover, due to $(\boldsymbol{\varTheta})$ and $(\boldsymbol{\psi_j})$ all such measures are equivalent in the sense the ratio $dI\!\!P_{\boldsymbol{\theta}} / dP_{\boldsymbol{\theta}^\circ}$ is bounded by a constant for all $\boldsymbol{\theta}, \boldsymbol{\theta}^\circ$ in $\varTheta$ because $\sup_{x \in \mathcal{X}} |\langle \boldsymbol{\theta} - \boldsymbol{\theta}^\circ, \varPsi(x) \rangle| < \infty$.

$(\boldsymbol{\varPsi u})$ *There exists a constant $\mathtt{C}_f \geqslant 1$ such that it holds for all $\boldsymbol{\theta} \in \varTheta^\circ$, all $\boldsymbol{u} \in I\!\!R^p$, and $k = 3, 4$*

$$E_{\boldsymbol{\theta}} \bigl| \bigl\langle \varPsi(X_1) - E_{\boldsymbol{\theta}} \varPsi(X_1), \boldsymbol{u} \bigr\rangle \bigr|^k \leqslant \bigl\{ \mathtt{C}_f^2 E_{\boldsymbol{\theta}} \langle \varPsi(X_1) - E_{\boldsymbol{\theta}} \varPsi(X_1), \boldsymbol{u} \rangle^2 \bigr\}^{k/2} . \tag{3.5}$$

Remark 3.1. In fact, in view of (3.3), it suffices to check (3.4) and (3.5) for one point $\boldsymbol{\theta} \in \varTheta^\circ$, in particular, for the true point $\boldsymbol{\theta}^*$ corresponding to the underlying measure P. Condition $(\boldsymbol{\varPsi u})$ means that each measure $P_{\boldsymbol{\theta}}$ for $\boldsymbol{\theta}$ in a vicinity $\varTheta^\circ$ of $\boldsymbol{\theta}^*$ satisfies a kind of Khinchin's inequality which relates the fourth and the second directional moments of $\varPsi(X_1)$. 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 $\varPsi(x)$ 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 $(\boldsymbol{\varPsi u})$ and $(\boldsymbol{\nabla^2 \phi})$. Remind $\widetilde{D}_G^2 = I\!\!F_G(\widetilde{\boldsymbol{\theta}}_G) = I\!\!F(\widetilde{\boldsymbol{\theta}}_G) + G^2$.

Theorem 3.1. *Assume $(\boldsymbol{\varTheta})$, $(\boldsymbol{\psi_j})$, $(\boldsymbol{\nabla^2 \phi})$, and $(\boldsymbol{\varPsi u})$. Consider a m-truncation prior yielding $\mathtt{p}_G \lesssim m$, or to a (s, w)-smooth prior with $\mathtt{p}_G \lesssim (n/w)^{1/(2s)}$. Define*

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

Let finally $\boldsymbol{\theta}_G^* \in \Theta^\circ$. *Then all the general results of Section 2 continue to apply to the posterior* $\boldsymbol{\vartheta}_G \,\big|\, \boldsymbol{X}$. *In particular, on a set* Ω_n *with* $\mathbb{P}\bigl(\Omega_n\bigr) \geqslant 1 - 3/n$, *it holds*

$$\sup_{A \in \mathcal{B}_s(\mathbb{R}^p)} \Bigl| \mathbb{P}\bigl(\boldsymbol{\vartheta}_G - \widetilde{\boldsymbol{\theta}}_G \in A \,\big|\, \boldsymbol{X}\bigr) - \mathbb{P}'\bigl(\widetilde{D}_G^{-1}\boldsymbol{\gamma} \in A\bigr) \Bigr| \leqslant \mathtt{C}\, z_G^6\, n^{-1}, \tag{3.6}$$

$$\sup_{A \in \mathcal{B}(\mathbb{R}^p)} \Bigl| \mathbb{P}\bigl(\boldsymbol{\vartheta}_G - \widetilde{\boldsymbol{\theta}}_G \in A \,\big|\, \boldsymbol{X}\bigr) - \mathbb{P}'\bigl(\widetilde{D}_G^{-1}\boldsymbol{\gamma} \in A\bigr) \Bigr| \leqslant \mathtt{C}\, z_G^3\, n^{-1/2}.$$

Under a proper choice $m = m_0 = (w_0 n)^{1/(2s_0+1)}$ or $(w\,n)^{1/(2s)} = (w_0\,n)^{1/(2s_0+1)} = m_0$ 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 $\widetilde{D}_G^2 \geqslant n\,\mathtt{C}_\phi^2 \boldsymbol{I}_p$ and $n^{-1} z_G^2 \asymp n^{-1} m_0 \asymp w_0^{1/(2s_0+1)} n^{-2s_0/(2s_0+1)}$. The error term in (3.6) is of order $n^{(2-2s_0)/(2s_0+1)} \to 0$ as $n \to \infty$ because $s_0 > 1$.

Theorem 3.2. *Assume conditions of Theorem 3.1. Define* $m_0 = (w_0 n)^{1/(2s_0+1)}$. *For the* m*-truncation prior with* $m = m_0$ *or for the* (s,w)*-smooth prior with* $(w\,n)^{1/(2s)} = (w_0\,n)^{1/(2s_0+1)} = m_0$, *it holds on the same set* Ω_n

$$\mathbb{P}\Bigl(\|\boldsymbol{\vartheta}_G - \boldsymbol{\theta}^*\|^2 > \mathtt{C}\, w_0^{1/(2s_0+1)} n^{-2s_0/(2s_0+1)} \,\big|\, \boldsymbol{X} \Bigr) \leqslant n^{-1},$$

for n *sufficiently large. Moreover, under "undersmoothing" choice* $m_0/m = o(1)$ *or* $m_0 (w\,n)^{-1/(2s)} = o(1)$ *as in (2.26), Bayesian credible sets* $\mathcal{E}_G(\widetilde{\mathtt{r}}_\alpha) = \bigl\{\boldsymbol{\theta} \colon \|\widetilde{\boldsymbol{\theta}}_G - \boldsymbol{\theta}\| \leqslant \widetilde{\mathtt{r}}_\alpha\bigr\}$ *with* $\widetilde{\mathtt{r}}_\alpha$ *satisfying* $\mathbb{P}'\bigl(\|\widetilde{D}_G^{-1}\boldsymbol{\gamma}\| > \widetilde{\mathtt{r}}_\alpha\bigr) = \alpha$ *are asymptotically valid; see (2.37).*

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

$$Y_i \sim P_{\upsilon_i} \in \mathcal{P}, \qquad i = 1, \ldots, n, \tag{3.7}$$

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

$$\log \frac{dP_\upsilon}{d\mu}(y) = \upsilon y - \phi(\upsilon) + \ell(y)$$

for a convex function $\phi(\upsilon)$ of a univariate parameter υ. A typical example is given by the logistic regression with binary observations Y_i. Then $\phi(\upsilon) = \log(1 + \mathrm{e}^\upsilon)$. The model (3.7) yields $\mathbb{E} Y_i = \phi'(\upsilon_i)$ and $\operatorname{Var}(Y_i) = \phi''(\upsilon_i)$. We, however, do not assume that the model is correct. The value υ_i is just defined by the canonical link $\mathbb{E} Y_i = \phi'(\upsilon_i)$. Generalized regression assumes that the υ_i's in (3.7) 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

$$Y_i \sim P_{\langle \boldsymbol{\Psi}_i, \boldsymbol{\theta} \rangle} \tag{3.8}$$

with $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)^\top \in I\!\!R^p$ and $\boldsymbol{\Psi}_i = \bigl(\psi_1(X_i), \ldots, \psi_p(X_i)\bigr)^\top \in \mathcal{X} \subset I\!\!R^p$ for $p \leqslant \infty$. The model (3.8) yields

$$L(\boldsymbol{\theta}) \; = \; \sum_{i=1}^{n} Y_i \langle \boldsymbol{\Psi}_i, \boldsymbol{\theta} \rangle - \phi\bigl(\langle \boldsymbol{\Psi}_i, \boldsymbol{\theta} \rangle\bigr).$$

The corresponding Fisher information operator reads as

$$I\!\!F(\boldsymbol{\theta}) = -\nabla^2 L(\boldsymbol{\theta}) \; = \; \sum_{i=1}^{n} \phi''\bigl(\langle \boldsymbol{\Psi}_i, \boldsymbol{\theta} \rangle\bigr) \, \boldsymbol{\Psi}_i \otimes \boldsymbol{\Psi}_i \geqslant 0,$$

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

$$\zeta(\boldsymbol{\theta}) \; = \; L(\boldsymbol{\theta}) - I\!\!E L(\boldsymbol{\theta}) = \sum_{i=1}^{n} \varepsilon_i \langle \boldsymbol{\Psi}_i, \boldsymbol{\theta} \rangle, \qquad \nabla \zeta = \sum_{i=1}^{n} \varepsilon_i \boldsymbol{\Psi}_i \, .$$

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

$(\boldsymbol{\psi}_\infty)$ It holds $\| \psi_j \, I\!\!1_{\mathcal{X}} \|_\infty \leqslant \mathtt{C}_\psi \, j^{1/2}$ for $j \leqslant p$.

$(\boldsymbol{\Theta})$ $\Theta \subseteq \mathcal{B}_{s_0}(w_0)$ for $s_0 > 1$. This condition and $(\boldsymbol{\psi}_\infty)$ yield by the Cauchy-Schwarz inequality, for any $x \in \mathcal{X}$ and $\boldsymbol{\theta} \in \Theta$

$$\bigl| \langle \boldsymbol{\Psi}(x), \boldsymbol{\theta} \rangle \bigr|^2 = \biggl(\sum_{j=1}^{p} \theta_j \psi_j(x) \biggr)^2 \; \leqslant \; \biggl(\sum_{j=1}^{p} \theta_j^2 j^{2s_0} \biggr) \biggl(\sum_{j=1}^{p} \mathtt{C}_\psi^2 \, j^{-2s_0+1} \biggr) \leqslant \frac{\mathtt{C}_\psi^2 w_0}{2s_0 - 2} \overset{\text{def}}{=} \mathtt{C}_\Psi^2 \, .$$

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

$$\phi''(t) \; \leqslant \; \mathtt{C}_\phi \, \phi''(0), \quad |t| \leqslant \mathtt{C}_\Psi \, .$$

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

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

$(\boldsymbol{\Psi u})$ It holds for all $\boldsymbol{u} \in I\!\!R^{m_0}$ and for a fixed constant $\mathtt{C}_f$

$$\frac{1}{n} \sum_{i=1}^{n} \langle \boldsymbol{\Psi}_i, \boldsymbol{u} \rangle^4 \; \leqslant \; \biggl\{ \frac{\mathtt{C}_f^2}{n} \sum_{i=1}^{n} \langle \boldsymbol{\Psi}_i, \boldsymbol{u} \rangle^2 \biggr\}^2 .$$

$(I\!\!F)$ For some $\mathtt{C}_{I\!\!F} > 0$

$$\inf_{\boldsymbol{u} \in I\!\!R^{m_0}} \frac{\bigl\langle I\!\!F(\boldsymbol{\theta}^*) \boldsymbol{u}, \boldsymbol{u} \bigr\rangle}{n \|\boldsymbol{u}\|^2} \; \geqslant \; \mathtt{C}_{I\!\!F}^{-2} \, .$$

Further, define

$$V^2 \overset{\text{def}}{=} \operatorname{Var}\bigl(\nabla\zeta\bigr) = \sum_{i=1}^{n} \sigma_i^2 \boldsymbol{\Psi}_i \otimes \boldsymbol{\Psi}_i \, , \qquad H^2 = \sum_{i=1}^{n} \sigma_{\max}^2 \boldsymbol{\Psi}_i \otimes \boldsymbol{\Psi}_i \, ,$$

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 $V^2 = I\!\!F(\boldsymbol{\theta}^*)$. Let us fix some Gaussian prior $\mathcal{N}(0, G^{-2})$. As previously, we focus on a m-truncation or (s,w)-smooth prior.

Theorem 3.3. *Suppose* (ψ_∞), $(\boldsymbol{\Theta})$, $(\boldsymbol{\phi}^{(\boldsymbol{k})})$, $(\boldsymbol{\varepsilon_i})$, $(\boldsymbol{\Psi u})$, *and* $(I\!\!F)$ *for the model* (3.8)*. Then all the properties of the posterior* $\boldsymbol{\vartheta}_G \,\big|\, \boldsymbol{Y}$ *listed in Theorem* 3.1 *or Theorem* 3.2 *continue to hold.*

4. Proofs of the main results

This section collects the proofs of the main theorems.

4.1. Proof of Proposition 2.3

The idea of the proof is to show that for each $\boldsymbol{u}$ with $\|D_G \boldsymbol{u}\| = \mathtt{r}_G$, the derivative of the function $L_G(\boldsymbol{\theta}_G^* + t\boldsymbol{u})$ in t is negative for $|t| \geqslant 1$. 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 $\boldsymbol{u}$ with $\|D_G \boldsymbol{u}\| \leqslant \mathtt{r}_G$. We use the decomposition

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

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

$$\frac{d}{dt} L_G(\boldsymbol{\theta}_G^* + t\boldsymbol{u}) \; = \; \bigl\langle \nabla\zeta, \boldsymbol{u} \bigr\rangle + f'(t). \tag{4.1}$$

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

$$\bigl|\bigl\langle \nabla\zeta, \boldsymbol{u} \bigr\rangle\bigr| \; = \; \bigl|\langle D_G^{-1} \nabla\zeta, D_G \boldsymbol{u} \rangle\bigr| \leqslant \mathtt{r}_G \, z(B_{V|G}, \mathtt{x}). \tag{4.2}$$

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

$$\bigl| f'(t) - t f''(0) \bigr| \; = \; \bigl| f'(t) - f'(0) - t f''(0) \bigr| \leqslant 3t^2 \, \mathtt{r}_G^3 \, \tau_{3,\mathsf{H}}.$$

For $t = 1$, we obtain

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

If $3\tau_{3,\mathsf{H}} \leqslant \rho$ 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 (4.2)

$$\begin{aligned} \frac{d}{dt} L_G(\boldsymbol{\theta}_G^* + t\boldsymbol{u}) \bigr|_{t=1} \; & \leqslant \; \bigl\langle \nabla\zeta, \boldsymbol{u} \bigr\rangle - \mathtt{r}_G^2 + 3\mathtt{r}_G^3 \tau_{3,\mathsf{H}} \\ & \leqslant \; \mathtt{r}_G \, z(B_G, \mathtt{x}) - \mathtt{r}_G^2 + 3\mathtt{r}_G^3 \tau_{3,\mathsf{H}} \leqslant \mathtt{r}_G \, z(B_G, \mathtt{x}) - (1 - \rho)\mathtt{r}_G^2 < 0 \end{aligned}$$

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 $t \geqslant 1$ together with $f'(t)$ due to (4.1), the same applies to all such t. This implies the assertion.

4.2. Proof of Proposition 2.4

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 $\| D_G \boldsymbol{u} \| \leqslant \mathtt{r}_G$, 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

$$\bigl| f(1) - \frac{1}{2} f''(0) \bigr| \; \leqslant \; \bigl| \delta_3(\boldsymbol{\theta}', \boldsymbol{u}) \bigr|, \quad \boldsymbol{\theta}' \in [\boldsymbol{\theta}, \boldsymbol{\theta} + \boldsymbol{u}].$$

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

$$\Bigl| I\!\!E L_G(\boldsymbol{\theta}_G^*) - I\!\!E L_G(\boldsymbol{\theta}) - \| D_G(\boldsymbol{\theta} - \boldsymbol{\theta}_G^*) \|^2 / 2 \Bigr| \; \leqslant \; \mathtt{r}_G^3 \, \tau_{3,\mathsf{H}}. \tag{4.3}$$

We now use that $\nabla L_G(\widetilde{\boldsymbol{\theta}}_G) = 0$ and by Proposition 2.3, $\boldsymbol{u} = \boldsymbol{\theta}_G^* - \widetilde{\boldsymbol{\theta}}_G$ fulfills $\| D_G \boldsymbol{u} \| \leqslant \mathtt{r}_G$ on $\Omega(\mathtt{x})$. Therefore, for $\boldsymbol{\theta} \in \mathcal{A}_G(\mathtt{r}_G)$

$$\begin{aligned} & \Bigl| L_G(\boldsymbol{\theta}) - L_G(\widetilde{\boldsymbol{\theta}}_G) - \| \widetilde{D}_G(\boldsymbol{\theta} - \widetilde{\boldsymbol{\theta}}_G) \|^2 / 2 \Bigr| \\ & \quad = \Bigl| L_G(\boldsymbol{\theta}) - L_G(\widetilde{\boldsymbol{\theta}}_G) - \langle \nabla L_G(\widetilde{\boldsymbol{\theta}}_G), \boldsymbol{\theta} - \widetilde{\boldsymbol{\theta}}_G \rangle - \| \widetilde{D}_G(\boldsymbol{\theta} - \widetilde{\boldsymbol{\theta}}_G) \|^2 / 2 \Bigr| \leqslant \mathtt{r}_G^3 \, \tau_{3,\mathsf{H}}. \end{aligned}$$

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

$$\begin{aligned} L_G(\widetilde{\boldsymbol{\theta}}_G) - L_G(\boldsymbol{\theta}_G^*) - \frac{1}{2} \| D_G^{-1} \nabla \zeta \|^2 &= \max_{\boldsymbol{\theta} \in \mathcal{A}_G(\mathtt{r}_G)} \Bigl\{ L_G(\boldsymbol{\theta}) - L_G(\boldsymbol{\theta}_G^*) - \frac{1}{2} \| D_G^{-1} \nabla \zeta \|^2 \Bigr\} \\ &= \max_{\boldsymbol{\theta} \in \mathcal{A}_G(\mathtt{r}_G)} \Bigl\{ \bigl\langle \boldsymbol{\theta} - \boldsymbol{\theta}_G^*, \nabla \zeta \bigr\rangle + I\!\!E L_G(\boldsymbol{\theta}) - I\!\!E L_G(\boldsymbol{\theta}_G^*) - \frac{1}{2} \| D_G^{-1} \nabla \zeta \|^2 \Bigr\} \\ &\leqslant \max_{\boldsymbol{\theta} \in \mathcal{A}_G(\mathtt{r}_G)} \Bigl\{ \bigl\langle D_G(\boldsymbol{\theta} - \boldsymbol{\theta}_G^*), D_G^{-1} \nabla \zeta \bigr\rangle - \frac{1}{2} \| D_G(\boldsymbol{\theta} - \boldsymbol{\theta}_G^*) \|^2 - \frac{1}{2} \| D_G^{-1} \nabla \zeta \|^2 \Bigr\} + \mathtt{r}_G^3 \, \tau_{3,\mathsf{H}} \\ &\leqslant \max_{\boldsymbol{\theta} \in \mathcal{A}_G(\mathtt{r}_G)} \Bigl\{ -\frac{1}{2} \| D_G(\boldsymbol{\theta} - \boldsymbol{\theta}_G^*) - D_G^{-1} \nabla \zeta \|^2 \Bigr\} + \mathtt{r}_G^3 \, \tau_{3,\mathsf{H}} = \mathtt{r}_G^3 \, \tau_{3,\mathsf{H}} \end{aligned}$$

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

The statement (2.18) follows from Lemma 7.3 with $Q = D_G$ and $f(\boldsymbol{\theta}) = I\!\!E L_G(\boldsymbol{\theta})$.

4.3. Proof of Proposition 2.5

The definition of $\boldsymbol{\theta}^{*}$ and $\boldsymbol{\theta}^{*}_{G}$ implies

$$\mathbb{E}L_{G}(\boldsymbol{\theta}^{*}_{G}) \geqslant \mathbb{E}L_{G}(\boldsymbol{\theta}^{*}), \qquad \mathbb{E}L(\boldsymbol{\theta}^{*}_{G}) \leqslant \mathbb{E}L(\boldsymbol{\theta}^{*}).$$

As $\mathbb{E}L_{G}(\boldsymbol{\theta}) = \mathbb{E}L(\boldsymbol{\theta}) - \|G\boldsymbol{\theta}\|^{2}/2$, it follows that

$$2\mathbb{E}L_{G}(\boldsymbol{\theta}^{*}_{G}) - 2\mathbb{E}L_{G}(\boldsymbol{\theta}^{*}) \leqslant \bigl\|G\boldsymbol{\theta}^{*}\bigr\|^{2} - \bigl\|G\boldsymbol{\theta}^{*}_{G}\bigr\|^{2} \leqslant \bigl\|G\boldsymbol{\theta}^{*}\bigr\|^{2}. \tag{4.4}$$

The bound (4.3) with $\boldsymbol{\theta} = \boldsymbol{\theta}^{*}$ implies the first statement of (2.19).

Further we show that $\|G\boldsymbol{\theta}^{*}\| \leqslant \mathtt{r}_{b}/2$ implies $\|D_{G}(\boldsymbol{\theta}^{*}_{G} - \boldsymbol{\theta}^{*})\| \leqslant \mathtt{r}_{b}$. Indeed, suppose the opposite inequality. Define $\boldsymbol{u} = \mathtt{r}_{b} D_{G}(\boldsymbol{\theta}^{*} - \boldsymbol{\theta}^{*}_{G}) / \|D_{G}(\boldsymbol{\theta}^{*}_{G} - \boldsymbol{\theta}^{*})\|$, so that $\|\boldsymbol{u}\| = \mathtt{r}_{b}$. The function $f(t) = \mathbb{E}L_{G}(\boldsymbol{\theta}^{*}_{G}) - \mathbb{E}L_{G}(\boldsymbol{\theta}^{*}_{G} + t\boldsymbol{u})$ is convex in t and $\boldsymbol{\theta}^{*}_{G} + t\boldsymbol{u} \in \varTheta^{\circ}$ for $|t| \leqslant 1$. Using the approximation (4.3) for $\boldsymbol{\theta} = \boldsymbol{\theta}^{*}_{G} + \boldsymbol{u}$ implies

$$2\mathbb{E}L_{G}(\boldsymbol{\theta}^{*}_{G}) - 2\mathbb{E}L_{G}(\boldsymbol{\theta}^{*}_{G} + t\boldsymbol{u}) \geqslant \mathtt{r}_{b}^{2} - \mathtt{r}_{b}^{3}\tau_{3,\mathsf{H}}(\mathtt{r}_{b}) \geqslant \mathtt{r}_{b}^{2}/2$$

and concavity of $\mathbb{E}L_{G}(\boldsymbol{\theta})$ together with $\nabla\mathbb{E}L_{G}(\boldsymbol{\theta}^{*}_{G}) = 0$ implies for $t \geqslant 1$

$$\mathbb{E}L_{G}(\boldsymbol{\theta}^{*}_{G}) - \mathbb{E}L_{G}(\boldsymbol{\theta}^{*}_{G} + t\boldsymbol{u}) \geqslant \mathtt{r}_{b}^{2}/2.$$

This contradicts to the bounds (4.4) and $\|G\boldsymbol{\theta}^{*}\|^{2} \leqslant \mathtt{r}_{b}^{2}/2$.

Now for any $\boldsymbol{\theta}$ with $\|D_{G}(\boldsymbol{\theta}^{*}_{G} - \boldsymbol{\theta})\| \leqslant \mathtt{r}_{b}$

$$\Bigl|\mathbb{E}L_{G}(\boldsymbol{\theta}^{*}_{G}) - \mathbb{E}L_{G}(\boldsymbol{\theta}) - \frac{1}{2}\bigl\|D_{G}(\boldsymbol{\theta} - \boldsymbol{\theta}^{*}_{G})\bigr\|^{2}\Bigr| \leqslant \mathtt{r}_{b}^{3}\tau_{3,\mathsf{H}}(\mathtt{r}_{b}). \tag{4.5}$$

Further we use that $\boldsymbol{\theta}^{*} = \operatorname{argmax} \mathbb{E}L(\boldsymbol{\theta})$ and $\mathbb{E}L_{G}(\boldsymbol{\theta}) = \mathbb{E}L(\boldsymbol{\theta}) - \|G\boldsymbol{\theta}\|^{2}/2$. By (4.5) in view of $\|D_{G}(\boldsymbol{\theta}^{*}_{G} - \boldsymbol{\theta}^{*})\| \leqslant \mathtt{r}_{b}$ and $D_{G}^{2} = \mathbb{F}(\boldsymbol{\theta}^{*}_{G}) + G^{2} = D^{2} + G^{2}$

$$\begin{aligned}
\mathbb{E}L(\boldsymbol{\theta}^{*}) - \mathbb{E}L_{G}(\boldsymbol{\theta}^{*}_{G}) &= \max_{\boldsymbol{\theta} \in \mathcal{A}_{G}(\mathtt{r}_{b})} \Bigl\{\mathbb{E}L_{G}(\boldsymbol{\theta}) + \frac{1}{2}\|G\boldsymbol{\theta}\|^{2} - \mathbb{E}L_{G}(\boldsymbol{\theta}^{*}_{G})\Bigr\} \\
&\leqslant \max_{\boldsymbol{\theta} \in \mathcal{A}_{G}(\mathtt{r}_{b})} \Bigl\{-\frac{1}{2}\bigl\|D_{G}(\boldsymbol{\theta} - \boldsymbol{\theta}^{*}_{G})\bigr\|^{2} + \frac{1}{2}\|G\boldsymbol{\theta}\|^{2}\Bigr\} + \mathtt{r}_{b}^{3}\tau_{3,\mathsf{H}}(\mathtt{r}_{b}) \\
&= \max_{\boldsymbol{\theta} \in \mathcal{A}_{G}(\mathtt{r}_{b})} \Bigl\{-\frac{1}{2}\bigl\|D\boldsymbol{\theta} - D^{-1}D_{G}^{2}\boldsymbol{\theta}^{*}_{G}\bigr\|^{2} + \frac{1}{2}\|D^{-1}D_{G}^{2}\boldsymbol{\theta}^{*}_{G}\|^{2}\Bigr\} + \mathtt{r}_{b}^{3}\tau_{3,\mathsf{H}}(\mathtt{r}_{b}).
\end{aligned}$$

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

$$\bigl\|D\boldsymbol{\theta}^{*} - D^{-1}D_{G}^{2}\boldsymbol{\theta}^{*}_{G}\bigr\|^{2} \leqslant 4\mathtt{r}_{b}^{3}\tau_{3,\mathsf{H}}(\mathtt{r}_{b}).$$

Equivalently, using again $D_{G}^{2} = D^{2} + G^{2}$

$$\bigl\|D^{-1}D_{G}^{2}(\boldsymbol{\theta}^{*} - \boldsymbol{\theta}^{*}_{G}) - D^{-1}G^{2}\boldsymbol{\theta}^{*}\bigr\|^{2} \leqslant 4\mathtt{r}_{b}^{3}\tau_{3,\mathsf{H}}(\mathtt{r}_{b}).$$

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

$$\bigl\| D_G(\boldsymbol{\theta}^* - \boldsymbol{\theta}_G^*) - D_G^{-1} G^2 \boldsymbol{\theta}^* \bigr\|^2 \leqslant 4 \mathtt{r}_b^3 \tau_{3,\mathsf{H}}(\mathtt{r}_b).$$

This completes the proof.

4.4. 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

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

with H^2 from $(\boldsymbol{E}\mathsf{H})$. 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

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

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

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

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

$$\rho(\mathtt{r}_0, \boldsymbol{\theta}) \stackrel{\text{def}}{=} \frac{\int \mathbb{1}\bigl(\|\mathsf{H}\boldsymbol{u}\| > \mathtt{r}_0 \bigr) \exp\bigl\{ f(\boldsymbol{\theta} + \boldsymbol{u}) - f(\boldsymbol{\theta}) - \bigl\langle \nabla f(\boldsymbol{\theta}), \boldsymbol{u} \bigr\rangle \bigr\} d\boldsymbol{u}}{\int \mathbb{1}\bigl(\|\mathsf{H}\boldsymbol{u}\| \leqslant \mathtt{r}_0 \bigr) \exp\bigl\{ f(\boldsymbol{\theta} + \boldsymbol{u}) - f(\boldsymbol{\theta}) - \bigl\langle \nabla f(\boldsymbol{\theta}), \boldsymbol{u} \bigr\rangle \bigr\} d\boldsymbol{u}} . \tag{4.6}$$

Step 2. First we bound the denominator of $\rho(\mathtt{r}_0, \boldsymbol{\theta})$. Lemma 7.4 of the supplement Spokoiny and Panov (2021) yields

$$\begin{aligned} \int_{\|\mathsf{H}\boldsymbol{u}\| \leqslant \mathtt{r}_0} \mathrm{e}^{f(\boldsymbol{\theta}+\boldsymbol{u}) - f(\boldsymbol{\theta}) - \langle \nabla f(\boldsymbol{\theta}), \boldsymbol{u} \rangle} d\boldsymbol{u} &\geqslant \bigl(1 - \diamondsuit(\mathtt{r}_0) \bigr) \int_{\|\mathsf{H}\boldsymbol{u}\| \leqslant \mathtt{r}_0} \mathrm{e}^{-\|D_G(\boldsymbol{\theta})\boldsymbol{u}\|^2/2} d\boldsymbol{u}, \\ \int_{\|\mathsf{H}\boldsymbol{u}\| \leqslant \mathtt{r}_0} \mathrm{e}^{f(\boldsymbol{\theta}+\boldsymbol{u}) - f(\boldsymbol{\theta}) - \langle \nabla f(\boldsymbol{\theta}), \boldsymbol{u} \rangle} d\boldsymbol{u} &\leqslant \bigl(1 + \diamondsuit(\mathtt{r}_0) \bigr) \int_{\|\mathsf{H}\boldsymbol{u}\| \leqslant \mathtt{r}_0} \mathrm{e}^{-\|D_G(\boldsymbol{\theta})\boldsymbol{u}\|^2/2} d\boldsymbol{u}, \end{aligned}$$

where $D_G^2(\boldsymbol{\theta}) = \mathbb{F}_G(\boldsymbol{\theta}) = -\nabla^2 f(\boldsymbol{\theta})$ and $\diamondsuit(\mathtt{r}_0)$ is given by (2.28). Moreover, after a proper normalization, the integral $\int_{\|\mathsf{H}\boldsymbol{u}\| \leqslant \mathtt{r}_0} \exp\Bigl(-\|D_G(\boldsymbol{\theta})\boldsymbol{u}\|^2/2 \Bigr) d\boldsymbol{u}$ can be viewed as the probability of the

Gaussian event. Namely

$$\frac{\det D_{G}(\boldsymbol{\theta})}{(2\pi)^{p/2}} \int_{\|\mathsf{H}\boldsymbol{u}\| \leqslant \mathtt{r}_0} \exp\Bigl(- \frac{\|D_{G}(\boldsymbol{\theta}) \boldsymbol{u}\|^{2}}{2} \Bigr) \, d\boldsymbol{u} \; = \; \mathbb{P}\bigl(\|\mathsf{H} D_{G}^{-1}(\boldsymbol{\theta}) \boldsymbol{\gamma}\| \leqslant \mathtt{r}_0 \bigr)$$

for a standard normal $\boldsymbol{\gamma} \in \mathbb{R}^{p}$. The choice $\mathtt{r}_0 \geqslant \sqrt{\mathtt{p}_G(\boldsymbol{\theta})} + \sqrt{2\mathtt{x}}$ yields by Corollary 8.3 of the supplement Spokoiny and Panov (2021)

$$\mathbb{P}\bigl(\|\mathsf{H} D_{G}^{-1}(\boldsymbol{\theta}) \boldsymbol{\gamma}\| \leqslant \mathtt{r}_0 \bigr) \;\geqslant\; 1 - \mathrm{e}^{-\mathtt{x}}.$$

If the error term $\diamondsuit(\mathtt{r}_0)$ is small, we obtain a sharp bound for the integral in the denominator of $\rho(\mathtt{r}_0, \boldsymbol{\theta})$ from (4.6).

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

$$L_{G}(\boldsymbol{\theta} + \boldsymbol{u}) - L_{G}(\boldsymbol{\theta}) - \bigl\langle \nabla L_{G}(\boldsymbol{\theta}), \boldsymbol{u} \bigr\rangle \;=\; \mathbb{E} L(\boldsymbol{\theta} + \boldsymbol{u}) - \mathbb{E} L(\boldsymbol{\theta}) - \bigl\langle \nabla \mathbb{E} L(\boldsymbol{\theta}), \boldsymbol{u} \bigr\rangle - \frac{1}{2} \|G\boldsymbol{u}\|^{2} .$$

Now we apply Lemma 7.8 of the supplement Spokoiny and Panov (2021) with $f(\boldsymbol{\theta} + \boldsymbol{u}) = \mathbb{E} L(\boldsymbol{\theta} + \boldsymbol{u}) + \bigl\langle (\mathbb{F}(\boldsymbol{\theta}) - \mathsf{H}^{2}) \boldsymbol{u}, \boldsymbol{u} \bigr\rangle / 2$. This function is concave and it holds $-\bigl\langle \nabla^{2} f(\boldsymbol{\theta}) \boldsymbol{u}, \boldsymbol{u} \bigr\rangle = \|\mathsf{H}\boldsymbol{u}\|^{2}$. The bound (7.16) of that lemma yields for any $\boldsymbol{u}$ with $\|\mathsf{H}\boldsymbol{u}\| = \mathtt{r} > \mathtt{r}_0$ and $\mathtt{C}_0 = 1 - 3\mathtt{r}_0 \tau_{3,\mathsf{H}} \geqslant 1/2$

$$\begin{aligned} L_{G}(\boldsymbol{\theta} + \boldsymbol{u}) - L_{G}(\boldsymbol{\theta}) - \bigl\langle \nabla L_{G}(\boldsymbol{\theta}), \boldsymbol{u} \bigr\rangle &= f(\boldsymbol{\theta} + \boldsymbol{u}) - f(\boldsymbol{\theta}) - \bigl\langle \nabla f(\boldsymbol{\theta}), \boldsymbol{u} \bigr\rangle - \bigl\langle (D_{G}^{2} - \mathsf{H}^{2}) \boldsymbol{u}, \boldsymbol{u} \bigr\rangle / 2 \\ &\leqslant -\mathtt{C}_0 (\|\mathsf{H}\boldsymbol{u}\| \mathtt{r}_0 - \mathtt{r}_0^{2}/2) - \bigl\langle (D_{G}^{2} - \mathsf{H}^{2}) \boldsymbol{u}, \boldsymbol{u} \bigr\rangle / 2 = -\mathtt{C}_0 (\|\mathsf{H}\boldsymbol{u}\| \mathtt{r}_0 - \mathtt{r}_0^{2}/2) - \|D_{G|\mathsf{H}}(\boldsymbol{\theta}) \boldsymbol{u}\|^{2}/2, \end{aligned}$$

where $D_{G|\mathsf{H}}^{2} = D_{G}^{2} - \mathsf{H}^{2}$. Now we can use the result about Gaussian integrals from Section 7.2 of the supplement Spokoiny and Panov (2021). With $\mathcal{T} = \mathsf{H} D_{G|\mathsf{H}}^{-1}(\boldsymbol{\theta})$, it holds by Lemma 7.9 of the supplement Spokoiny and Panov (2021)

$$\begin{aligned} &\frac{\det D_{G}(\boldsymbol{\theta})}{(2\pi)^{p/2}} \int \mathbb{I}\bigl(\|\mathsf{H}\boldsymbol{u}\| > \mathtt{r}_0 \bigr) \exp\bigl\{ L_{G}(\boldsymbol{\theta} + \boldsymbol{u}) - L_{G}(\boldsymbol{\theta}) - \bigl\langle \nabla L_{G}(\boldsymbol{\theta}), \boldsymbol{u} \bigr\rangle \bigr\} \, d\boldsymbol{u} \\ &\qquad \leqslant \mathbb{E} \Bigl\{ \exp\Bigl(-\mathtt{C}_0 \mathtt{r}_0 \|\mathcal{T}\boldsymbol{\gamma}\| + \frac{\mathtt{C}_0 \mathtt{r}_0^{2}}{2} + \frac{1}{2} \|\mathcal{T}\boldsymbol{\gamma}\|^{2} \Bigr) \, \mathbb{I}\bigl(\|\mathcal{T}\boldsymbol{\gamma}\| > \mathtt{r}_0 \bigr) \Bigr\} \leqslant \mathtt{C} \, \mathrm{e}^{-(\mathtt{p}_G(\boldsymbol{\theta}) + \mathtt{x})/2}. \end{aligned}$$

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

4.5. Proof of Theorem 2.9 and Corollary 2.10

We proceed similarly to the proof of Proposition 2.8. Fix any centrally symmetric set A. First we restrict the posterior probability to the set $\mathcal{E}(\mathtt{r}_0) = \{ \boldsymbol{u} \colon \|\mathsf{H}\boldsymbol{u}\| \leqslant \mathtt{r}_0 \}$. 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,

$$\begin{aligned}
I\!\!P\bigl(\boldsymbol{\vartheta}_G - \widetilde{\boldsymbol{\theta}}_G \in A \bigm| \boldsymbol{Y}\bigr) &= \frac{\int_A \exp\bigl\{L_G(\widetilde{\boldsymbol{\theta}}_G + \boldsymbol{u})\bigr\} d\boldsymbol{u}}{\int_{I\!\!R^p} \exp\bigl\{L_G(\widetilde{\boldsymbol{\theta}}_G + \boldsymbol{u})\bigr\} d\boldsymbol{u}} \\
&\leqslant \frac{\int_{A(\mathtt{r}_0)} \exp\bigl\{L_G(\widetilde{\boldsymbol{\theta}}_G + \boldsymbol{u}) - L_G(\widetilde{\boldsymbol{\theta}}_G) - \bigl\langle \nabla L_G(\widetilde{\boldsymbol{\theta}}_G), \boldsymbol{u} \bigr\rangle\bigr\} d\boldsymbol{u}}{\int_{\|\mathsf{H}\boldsymbol{u}\| \leqslant \mathtt{r}_0} \exp\bigl\{L_G(\widetilde{\boldsymbol{\theta}}_G + \boldsymbol{u}) - L_G(\widetilde{\boldsymbol{\theta}}_G) - \bigl\langle \nabla L_G(\widetilde{\boldsymbol{\theta}}_G), \boldsymbol{u} \bigr\rangle\bigr\} d\boldsymbol{u}} + \rho(\mathtt{r}_0).
\end{aligned}$$

The estimates from the proof of Theorem 2.8 yield the upper bound

$$\begin{aligned}
I\!\!P\bigl(\boldsymbol{\vartheta}_G - \widetilde{\boldsymbol{\theta}}_G \in A \bigm| \boldsymbol{Y}\bigr) &\leqslant \frac{\bigl\{1 + \diamondsuit(\mathtt{r}_0)\bigr\} \int_{A(\mathtt{r}_0)} \exp\bigl\{-\|\widetilde{D}_G \boldsymbol{u}\|^2/2\bigr\} d\boldsymbol{u}}{\bigl\{1 - \diamondsuit(\mathtt{r}_0)\bigr\} \int_{\|\mathsf{H}\boldsymbol{u}\| \leqslant \mathtt{r}_0} \exp\bigl\{-\|\widetilde{D}_G \boldsymbol{u}\|^2/2\bigr\} d\boldsymbol{u}} + \rho(\mathtt{r}_0) \\
&\leqslant \frac{\bigl\{1 + \diamondsuit(\mathtt{r}_0)\bigr\} I\!\!P\bigl(\widetilde{D}_G^{-1} \boldsymbol{\gamma} \in A\bigr)}{\bigl\{1 - \diamondsuit(\mathtt{r}_0)\bigr\} I\!\!P\bigl(\|\mathsf{H}\widetilde{D}_G^{-1} \boldsymbol{\gamma}\| \leqslant \mathtt{r}_0\bigr)} + \rho(\mathtt{r}_0).
\end{aligned}$$

Now we prove the lower bound. It obviously holds

$$\begin{aligned}
& I\!\!P\bigl(\boldsymbol{\vartheta}_G - \widetilde{\boldsymbol{\theta}}_G \in A \bigm| \boldsymbol{Y}\bigr) = \frac{\int_A \exp\bigl\{L_G(\widetilde{\boldsymbol{\theta}}_G + \boldsymbol{u})\bigr\} d\boldsymbol{u}}{\int_{I\!\!R^p} \exp\bigl\{L_G(\widetilde{\boldsymbol{\theta}}_G + \boldsymbol{u})\bigr\} d\boldsymbol{u}} \\
&\quad \geqslant \frac{\int_{A(\mathtt{r}_0)} \exp\bigl\{L_G(\widetilde{\boldsymbol{\theta}}_G + \boldsymbol{u}) - L_G(\widetilde{\boldsymbol{\theta}}_G) - \bigl\langle \nabla L_G(\widetilde{\boldsymbol{\theta}}_G), \boldsymbol{u} \bigr\rangle\bigr\} d\boldsymbol{u}}{\bigl(\int_{\|\mathsf{H}\boldsymbol{u}\| \leqslant \mathtt{r}_0} + \int_{\|\mathsf{H}\boldsymbol{u}\| > \mathtt{r}_0}\bigr) \exp\bigl\{L_G(\widetilde{\boldsymbol{\theta}}_G + \boldsymbol{u}) - L_G(\widetilde{\boldsymbol{\theta}}_G) - \bigl\langle \nabla L_G(\widetilde{\boldsymbol{\theta}}_G), \boldsymbol{u} \bigr\rangle\bigr\} d\boldsymbol{u}}
\end{aligned}$$

and in a similar way as above with $\diamondsuit = \diamondsuit(\mathtt{r}_0)$ and $\rho(\mathtt{r}_0) = \rho$

$$I\!\!P\bigl(\boldsymbol{\vartheta}_G - \widetilde{\boldsymbol{\theta}}_G \in A \bigm| \boldsymbol{Y}\bigr) \geqslant \frac{\{1 - \diamondsuit\} I\!\!P\bigl(\widetilde{D}_G^{-1} \boldsymbol{\gamma} \in A(\mathtt{r}_0)\bigr)}{\{1 + \diamondsuit\} I\!\!P\bigl(\|\mathsf{H}\widetilde{D}_G^{-1} \boldsymbol{\gamma}\| \leqslant \mathtt{r}_0\bigr) + \rho} \geqslant \frac{\{1 - \diamondsuit\}\bigl\{I\!\!P\bigl(\widetilde{D}_G^{-1} \boldsymbol{\gamma} \in A\bigr) - \rho\bigr\}}{\{1 + \diamondsuit\} I\!\!P\bigl(\|\mathsf{H}\widetilde{D}_G^{-1} \boldsymbol{\gamma}\| \leqslant \mathtt{r}_0\bigr) + \rho}.$$

For the case of an arbitrary possibly non-symmetric A, the proof is similar with the use of (7.10) instead of (7.9) of the supplement Spokoiny and Panov (2021).

4.6. Proof of Theorem 2.11 and of Theorem 2.12

The difference $\boldsymbol{\vartheta}_G - \boldsymbol{\theta}^*$ can be decomposed as

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

Theorem 2.6 with $\mathtt{x} = \log n$ allows to bound with high probability

$$\bigl\|Q(\widetilde{\boldsymbol{\theta}}_G - \boldsymbol{\theta}^*)\bigr\|^2 \lesssim \|Q D_G^{-2} G^2 \boldsymbol{\theta}^*\|^2 + z^2(B_{Q|G}^2, \mathtt{x}) \leqslant \|Q D_G^{-2} G^2 \boldsymbol{\theta}^*\|^2 + \operatorname{tr}(B_{Q|G}^2) + \log n,$$

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

$$I\!\!P\bigl(\|Q(\boldsymbol{\vartheta}_G - \widetilde{\boldsymbol{\theta}}_G)\| \geqslant \mathtt{r} \bigm| \boldsymbol{Y}\bigr) \lesssim I\!\!P'\bigl(\|Q\widetilde{D}_G^{-1} \boldsymbol{\gamma}\| \geqslant \mathtt{r}\bigr) + 1/n.$$

Now we apply Theorem 8.2 of the supplement Spokoiny and Panov (2021) with $\mathtt{r} = \mathtt{r}_Q = z(Q\widetilde{D}_G^{-2}Q^\top, \mathtt{x}) \leqslant \sqrt{\operatorname{tr}(Q\widetilde{D}_G^{-2}Q^\top)} + \sqrt{2\mathtt{x}}$ to the Gaussian quadratic form $\|Q\widetilde{D}_G^{-1}\boldsymbol{\gamma}\|^2$. The desired result (2.34) follows by (2.18) of Proposition 2.4 and by the bias bound (2.33).

To check (2.37) note that by definition, it holds for the true parameter $\boldsymbol{\theta}^*$:

$$\mathbb{P}\bigl(\boldsymbol{\theta}^* \in \mathcal{A}_{Q|G}(\mathtt{r})\bigr) \; = \; \mathbb{P}\bigl(\|Q(\widetilde{\boldsymbol{\theta}}_G - \boldsymbol{\theta}^*)\| \leqslant \mathtt{r}\bigr).$$

The Fisher expansion (2.15) $\widetilde{\boldsymbol{\theta}}_G - \boldsymbol{\theta}_G^* \approx D_G^{-2}\nabla\zeta$ of Proposition 2.4 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

$$\mathbb{P}\bigl(\|Q(\widetilde{\boldsymbol{\theta}}_G - \boldsymbol{\theta}^*)\| \leqslant \mathtt{r}\bigr) \; \approx \; \mathbb{P}\Bigl(\bigl\|Q\bigl(D_G^{-2}V\boldsymbol{\gamma} + \boldsymbol{\theta}_G^* - \boldsymbol{\theta}^*\bigr)\bigr\| \leqslant \mathtt{r}\Bigr).$$

By Gaussian comparison Theorem 8.1 of the supplement Spokoiny and Panov (2021), the impact of the bias $\boldsymbol{\theta}_G^* - \boldsymbol{\theta}^*$ is negligible under the undersmoothing condition $\|Q(\boldsymbol{\theta}_G^* - \boldsymbol{\theta}^*)\|^2 \ll \operatorname{tr}\bigl(QD_G^{-2}V^2D_G^{-2}Q^\top\bigr)$. Combining with Theorem 2.6 yields in view of $D_G^{-2}V^2D_G^{-2} \leqslant D_G^{-2}$

$$\begin{aligned} 1-\alpha \; &= \; \mathbb{P}'\bigl(\|Q\widetilde{D}_G^{-1}\boldsymbol{\gamma}\| \leqslant \mathtt{r}_\alpha\bigr) \approx \mathbb{P}\bigl(\|QD_G^{-1}\boldsymbol{\gamma}\| \leqslant \mathtt{r}_\alpha\bigr) \leqslant \mathbb{P}\Bigl(\|QD_G^{-2}V\boldsymbol{\gamma}\| \leqslant \mathtt{r}_\alpha\Bigr) \\ &\approx \; \mathbb{P}\bigl(\|Q(\widetilde{\boldsymbol{\theta}}_G - \boldsymbol{\theta}^*)\| \leqslant \mathtt{r}_\alpha\bigr), \end{aligned}$$

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

4.7. Proof of Corollary 2.13

It suffices to consider the case of $Q = \boldsymbol{I}$. For both results, we only have to evaluate the bias $\|D_G^{-2}G^2\boldsymbol{\theta}^*\|^2$ and the variance $\operatorname{tr}\bigl(D_G^{-2}V^2D_G^{-2}\bigr)$.

Lemma 4.1. *Suppose* (2.7) *with* $V^2 = \operatorname{Var}(\nabla\zeta)$ *in place of* H^2. *Let* $\boldsymbol{\theta}^* \in \mathcal{B}(s_0, w_0)$. *For a* (s,w)*-smooth prior with* $s \geqslant s_0/2$, *fix* $m = (w\,n)^{1/(2s)}$. *It holds*

$$\begin{gathered} \|D_G^{-2}G^2\boldsymbol{\theta}^*\|^2 \; \leqslant \; \mathtt{C}\,w_0 m^{-2s_0} = \mathtt{C}\,w_0(w\,n)^{-2s_0/(2s)}, \\ \operatorname{tr}\bigl(D_G^{-2}\bigr) \; \geqslant \; \operatorname{tr}\bigl(D_G^{-2}V^2D_G^{-2}\bigr) \geqslant \mathtt{C}\,m/n. \end{gathered}$$

Proof. We use that $\|G_0\boldsymbol{\theta}^*\|^2 \leqslant 1$ with $G_0^2 = \operatorname{diag}\bigl(w_0^{-1}j^{2s_0}\bigr)$. Therefore,

$$\|D_G^{-2}G^2\boldsymbol{\theta}^*\|^2 \; \leqslant \; \|D_G^{-2}G^2G_0^{-1}\| \times \|G_0\boldsymbol{\theta}^*\|^2 \leqslant \|D_G^{-2}G^2G_0^{-1}\|.$$

Further, the choice $m \approx (w\,n)^{1/(2s)}$ ensures the relation $n \approx g_m^2 = w^{-1}m^{2s}$. For any $\boldsymbol{u} \in \mathbb{V}_m$ with $\|\boldsymbol{u}\| = 1$, it holds by (2.7) $\|D_G^{-2}\boldsymbol{u}\| \leqslant \mathtt{C}n^{-1}$ and by $2s \geqslant s_0$, the ratio g_j^4/j^{2s_0} grows with j, so that

$$\|D_G^{-2}G^2G_0^{-1}\boldsymbol{u}\|^2 \; \leqslant \; \mathtt{C}n^{-2}g_m^4 w_0 m^{-2s_0} \leqslant \mathtt{C}\,w_0 m^{-2s_0}.$$

Similarly, for $\boldsymbol{u} \in \mathbb{V}_m^c$ with $\|\boldsymbol{u}\| = 1$

$$\|D_G^{-2}G^2G_0^{-1}\boldsymbol{u}\|^2 \; \leqslant \; \|G_0^{-1}\boldsymbol{u}\|^2 \leqslant \mathtt{C}\,w_0 m^{-2s_0}$$

and the bound for the bias follows. The second statement can be proved similarly, cf. the proof of Lemma 2.2 below. □

Now the bias-variance relation (2.33) is satisfied for $m = m_0$, and (2.38) follows from Theorem 2.11. For $m \gg m_0$, we can apply Theorem 2.12.

4.8. Proof of Theorems 3.1 through 3.3

It suffices to check the conditions of the general results from Section 2. We start with the log-density model. First we show $(\mathcal{L}_{\mathbf{0}})$. Remind that $I\!\!E L(\boldsymbol{\theta}) = n\bigl\{\langle \overline{\Psi}, \boldsymbol{\theta}\rangle - \phi(\boldsymbol{\theta})\bigr\}$ and $I\!\!F(\boldsymbol{\theta}) = -\nabla^2 I\!\!E L(\boldsymbol{\theta}) = n\nabla^2\phi(\boldsymbol{\theta})$. Fix any $\boldsymbol{\theta} \in \Theta^{\circ}$ and denote also $\Psi_{\boldsymbol{\theta}} = E_{\boldsymbol{\theta}}\Psi(X_1)$ and define for $\boldsymbol{u} \in I\!\!R^{m_0}$ with $\|\mathsf{H}\boldsymbol{u}\| = \mathtt{r}$ and any t

$$
\begin{aligned}
q(t) &\stackrel{\text{def}}{=} \int \exp\bigl\{\bigl\langle \Psi(x) - \Psi_{\boldsymbol{\theta}}, \boldsymbol{\theta} + t\boldsymbol{u}\bigr\rangle - \phi(\boldsymbol{\theta})\bigr\}\mu_0(dx) = \int \exp\bigl\{t\bigl\langle \Psi(x) - \Psi_{\boldsymbol{\theta}}, \boldsymbol{u}\bigr\rangle\bigr\} P_{\boldsymbol{\theta}}(dx), \\
q_k(t) &\stackrel{\text{def}}{=} \frac{d^k q(t)}{dt^k} = \int \bigl\langle \Psi(x) - \Psi_{\boldsymbol{\theta}}, \boldsymbol{u}\bigr\rangle^k \exp\bigl\{t\bigl\langle \Psi(x) - \Psi_{\boldsymbol{\theta}}, \boldsymbol{u}\bigr\rangle\bigr\} P_{\boldsymbol{\theta}}(dx), \quad k \geqslant 1. \qquad (4.7)
\end{aligned}
$$

Due to $(\boldsymbol{\psi_j})$, 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

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

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

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

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

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

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

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

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

$$
\begin{aligned}
\log I\!\!E \exp\bigl\{\langle \nabla\zeta, \boldsymbol{u}\rangle\bigr\} &= n \log E \exp\bigl\{\langle \Psi(X_1) - \overline{\Psi}, \boldsymbol{u}\rangle\bigr\} \\
&= n\bigl\{\phi(\boldsymbol{\theta}^* + \boldsymbol{u}) - \langle \nabla\phi(\boldsymbol{\theta}^*), \boldsymbol{u}\rangle\bigr\} = \frac{n}{2}\bigl\langle \nabla^2\phi(\boldsymbol{\theta}^* + t\boldsymbol{u})\, \boldsymbol{u}, \boldsymbol{u}\bigr\rangle, \quad t \in [0, 1].
\end{aligned}
$$

By $(\boldsymbol{\nabla^2\phi})$, $\|\boldsymbol{u}\| \leqslant \|\mathsf{H}^{-1}\| \, \|\mathsf{H}\boldsymbol{u}\| \leqslant \mathtt{C}_{\phi,1} n^{-1/2}\lambda$. By the Cauchy-Schwarz inequality and $(\boldsymbol{\psi_j})$

$$\bigl|\bigl\langle \Psi(x), \boldsymbol{u} \bigr\rangle\bigr|^2 \leqslant q_{m_0}^{-2} \|\boldsymbol{u}\|^2 \sum_{j=1}^{m_0} \psi_j^2(x) q_j^2 \leqslant \mathtt{C}_{\phi,1}^2 \, \mathtt{C}_{\psi}^2 \, \frac{\lambda^2 q_{m_0}^{-2}}{n} \leqslant \mathtt{C}_{\phi,1}^2 \, \mathtt{C}_{\psi}^2 \, \frac{\lambda^2 \log^2(n)}{n^{2/3}} \, .$$

If the latter value is smaller than a constant $\mathtt{C}$ then by (4.7)

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

This yields $(\boldsymbol{E}\mathsf{H})$ with $\mathtt{g} \lesssim n^{1/3}/\log(n) \lesssim m_0$. Theorem 8.5 of the supplement Spokoiny and Panov (2021) ensure for $z_G = \sqrt{\mathtt{p}_G} + \sqrt{2\log n}$ the probability bound (2.11) $\bigl\| D_G^{-1} \nabla\zeta \bigr\| \leqslant z_G$ on a random set Ω_n with $\mathbb{P}\bigl(\Omega_n\bigr) \geqslant 1 - 3/n$. 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 $\boldsymbol{u} \in \mathbb{R}^{m_0}$ with $\|\mathsf{H}\boldsymbol{u}\| \leqslant \mathtt{r}$. It holds by $|\langle \Psi_i, \boldsymbol{\theta} \rangle| \leqslant \mathtt{C}_{\Psi}$, $(\boldsymbol{\phi^{(k)}})$, and $(\boldsymbol{\Psi u})$ for $k = 3, 4$

$$\begin{aligned} \bigl|\delta_k(\boldsymbol{\theta}, \boldsymbol{u})\bigr| &= \Bigl| \sum \phi^{(k)}\bigl(\langle \Psi_i, \boldsymbol{\theta} \rangle\bigr) \langle \Psi_i, \boldsymbol{u} \rangle^k \Bigr| \leqslant \|\phi^{(k)}\|_\infty \Bigl(\sum \langle \Psi_i, \boldsymbol{u} \rangle^4 \Bigr)^{k/4} \\ &\leqslant n^{k/2-1} \|\phi^{(k)}\|_\infty \Bigl(\sum \langle \Psi_i, \boldsymbol{u} \rangle^2 \Bigr)^{k/2} \leqslant \mathtt{C} \, \mathtt{r}^k n^{-k/2+1} \end{aligned}$$

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

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

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

$$\bigl|\langle \Psi_i, \boldsymbol{u} \rangle\bigr| \leqslant \mathtt{C}_{\psi} \sqrt{m_0} \, \|\boldsymbol{u}\| \leqslant \mathtt{C}_{\psi} \, \mathtt{C}_{\mathbb{F}} \sqrt{m_0} \, \|\mathsf{H}\boldsymbol{u}\| n^{-1/2} \leqslant \mathtt{C}_{\psi} \, \mathtt{C}_{\mathbb{F}} \sqrt{m_0} \, n^{-1/2} \, .$$

Thus, $\bigl|\lambda \langle \Psi_i, \boldsymbol{u} \rangle\bigr| \leqslant \lambda \mathtt{C}_{\mathbb{F}} \sqrt{m_0} \, n^{-1/2}$, and, for $|\lambda| \leqslant \varrho / \bigl(\mathtt{C} \sqrt{m_0} \, n^{-1/2}\bigr)$, it follows by $(\boldsymbol{\varepsilon_i})$

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

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

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). Large ball probabilities, gaussian comparison and anti-concentration. *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. (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.

Spokoiny, V. and Panov, M. (2021). Supplement to Accuracy of Gaussian approximation for high-dimensional posterior distributions.

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.