Bayesian Analysis (2025) **20**, Number 1, pp. 1303–1330

Inexact Laplace Approximation and the Use of Posterior Mean in Bayesian Inference*

Vladimir Spokoiny†

Abstract. The prominent Bernstein – von Mises (BvM) Theorem claims a kind of approximation of the posterior distribution by a Gaussian one with the covariance close to the inverse of the total Fisher information matrix. A more general Laplace approximation result states a similar Gaussian approximation of the posterior with the parameters depending on the prior. These two results build a basis for Bayesian inference and uncertainty quantification in a rather general situation. Spokoiny and Panov (2021) offered a new look at this problem which allows to state rather strong results on the quality of Gaussian approximation in non-asymptotic and dimension free form assuming linearity and concavity of log-likelihood function which can be misspecified. The established results provide explicit non-asymptotic bounds on the quality of a Gaussian approximation of the posterior distribution in total variation distance in terms of the so called *effective dimension* $\mathtt{p}_G$ defined as interplay between information contained in the data and in the prior distribution. This paper substantially improves and further develops the results from Spokoiny and Panov (2021) using the recent progress on high dimensional Laplace approximation. We address the question of *effective and critical dimension* in Bayesian inference, the relations between Laplace approximation and Bernstein–von Mises Theorem, and, particularly, the use of *posterior mean* instead of *Maximum A Posteriori Probability* estimator in Bayesian inference. The results are illustrated for the case of log-density estimation.

MSC2020 subject classifications: Primary 62F15; secondary 60F25, 62C10.

Keywords: Gaussian approximation, effective dimension, posterior mean.

1 Introduction

The prominent Bernstein – von Mises (BvM) Theorem is one of the most fundamental and most mysterious results in Bayesian inference stating asymptotic normality of the posterior distribution. It is often considered as a Bayesian counterpart of the Fisher Theorem claiming asymptotic normality of the maximum likelihood estimator (MLE). The inverse of the Fisher information matrix describes the variance of the MLE in the Fisher Theorem and the variance of the posterior in the BvM Theorem. These two results build the basis for statistical inference and uncertainty quantification of majority of statistical procedures. Parametric BvM theory is well-understood Van der Vaart (2000). One of corollaries of the BvM result is the fact that the choice of the prior is unimportant, it does not show up in the limiting distribution and washes out of the

*Financial support by the German Research Foundation (DFG) through the Collaborative Research Center 1294 "Data assimilation" is gratefully acknowledged.

†Weierstrass Institute and Humboldt University Berlin, Mohrenstr. 39, 10117 Berlin, Germany; HSE and IITP RAS Moscow, spokoiny@wias-berlin.de

 https://doi.org/10.1214/23-BA1391

posterior distribution as the sample size increases. The situation changes dramatically when modern statistical problems with a high dimensional parameter space and limited sample size are considered. The prior becomes crucial, its choice is an important issue as for inference problems as for uncertainty quantification; see e.g. Knapik et al. (2016), Szabó et al. (2015), Rousseau and Szabo (2020) and references therein. Lu (2017) studied a parametric BvM theorem for nonlinear Bayesian inverse problems with an increasing number of parameters.

Consider a Bayesian inference problem for a pseudo log-likelihood function $L(\upsilon) = L(\boldsymbol{Y}, \upsilon)$ with data $\boldsymbol{Y}$, a parameter $\upsilon \in \mathbb{R}^p$, and a prior π on $\mathbb{R}^p$. This paper focuses on the case of a Gaussian prior $\pi_G \sim \mathcal{N}(\upsilon_0, G^{-2})$ with a symmetric positive definite covariance matrix G^{-2}. An extension to non-Gaussian priors is commented in Section D in the Supplement (Spokoiny (2023)); later (S2023). The posterior density $\pi_G(\cdot)$ of υ given $\boldsymbol{Y}$ can be written in the form

$$\upsilon_G \,\big|\, \boldsymbol{Y} \sim \pi_G(\upsilon) \propto \exp\bigl\{L(\upsilon) - \|G(\upsilon - \upsilon_0)\|^2/2\bigr\},$$

where the sign $\propto$ means equality up to a normalizing multiplicative constant. Assume that the penalized maximum likelihood estimator (pMLE) $\widetilde{\upsilon}_G$ is well defined:

$$\widetilde{\upsilon}_G = \operatorname*{argmax}_{\upsilon} \bigl\{L(\upsilon) - \|G(\upsilon - \upsilon_0)\|^2/2\bigr\}. \tag{1.1}$$

Clearly $\widetilde{\upsilon}_G$ is maximizer of $\pi_G(\upsilon)$ and it is often referred to as maximum a posteriori probability (MAP) estimator. Define also the penalized Fisher information matrix

$$\mathbb{F}_G(\upsilon) = -\nabla^2 \mathbb{E} L(\upsilon) + G^2,$$

where $\mathbb{E}$ means the expectation w.r.t. the underlying data distribution $\mathbb{P}$. An important step in understanding the impact of the prior is made by the results on Laplace's approximation claiming that the posterior distribution π_G is close to the Gaussian distribution $\mathcal{N}(\widetilde{\upsilon}_G, \mathbb{F}_G^{-1}(\widetilde{\upsilon}_G))$. The Bernstein–von Mises phenomenon formally corresponds to the non-penalized case $G^2 = 0$. A number of papers discuss the BvM phenomenon for nonlinear inverse problems; see e.g. Nickl (2020); Monard et al. (2019); Giordano and Kekkonen (2020), where the convergence is quantified in a distance that metrizes the weak convergence. Schillings et al. (2020) showed that the Laplace approximation error in Hellinger distance converges to zero in the order of the noise level. The recent paper Helin and Kretschmann (2022) provides a finite sample error of Laplace approximation for the total variation (TV) distance with an explicit dependence on the dimension and on the nonlinearity of the forward mapping for Bayesian inverse problems. A common drawback of all these and similar results is that the error bounds depend implicitly or explicitly of the dimension of the parameter space. If the dimension grows this dependence may become crucial, thus questioning the range of applicability of Laplace approximation and of BvM Theorem. Spokoiny (2017) discussed general properties of the pMLE $\widetilde{\upsilon}_G$ in terms of the so called *effective dimension* which can be small or moderate even if the true parameter dimension is large. Spokoiny and Panov (2021) established similar and even stronger results under an additional assumption of

linearity of the stochastic term $L(\upsilon) - \mathbb{E}L(\upsilon)$ in υ. Spokoiny (2019) explained how a non-linear inverse problem can be reduced to the stochastically linear case by extending the parameter space without significant increase of the effective dimension.

Another challenge of applying the BvM-type results is that the parameters of the approximating Gaussian distribution are defined through the pMLE $\widetilde{\upsilon}_G$ from (1.1). This is a high dimensional optimization problem for a random objective function. A closed form analytic solution is available only in very special and simple situations, otherwise it can only be obtained by high-tech optimization methods with some error. This leads to an open problem of justifying a Laplace approximation of the posterior with inexact parameters for Bayesian inference, Durmus and Moulines (2019). Finding $\widetilde{\upsilon}_G$ could be especially difficult if computing $L(\upsilon)$ and its gradient is costly. This leads to gradient free methods Nesterov and Spokoiny (2017) or Bayesian optimization approach Mockus (1989), Frazier (2018). Ma et al. (2019) argued that Monte Carlo Markov Chain (MCMC) sampling can be more efficient than high dimensional optimization. A particular issue for applying MCMC type methods for Bayesian inference is to justify the use of posterior mean instead of posterior mode.

This paper aims at addressing the mentioned challenges. Below the list of the most important achievements in the paper.

Effective dimension and dimension free guarantees for penalized MLE $\widetilde{\upsilon}_G$. The *effective dimension* $\mathtt{p}_G$ is defined by an interplay between the information delivered by the data and information contained in the penalty; see Spokoiny and Panov (2021) or Section 2.4 for more details. Section 2 establishes explicit *non-asymptotic* and *dimension free* results on the accuracy of the pMLE $\widetilde{\upsilon}_G$ including concentration, Fisher and Wilks expansions, loss and risk bounds under assumptions of stochastic linearity and concavity of the log-likelihood function. The bounds apply well under model misspecification and are stated under the same *critical dimension* condition $\mathtt{p}_G \ll n$, where n is the *effective sample size*. The use of self-concordance type conditions from Section A of the Supplement (S2023) on $f(\upsilon) = \mathbb{E}L(\upsilon)$ allows to obtain sharper and more transparent results than in Spokoiny and Panov (2021). Section C of the Supplement (S2023) explains a rate optimal choice of penalty/prior and derives usual minimax rate results over Sobolev smoothness classes from the obtained bounds.

Laplace's approximation of the posterior. Section 3 of the paper presents new results on concentration and Gaussian approximation of the posterior. It is important to note that the analysis of the posterior distribution requires very different analytic tools than used for the pMLE study. We make use of the recent progress in high dimensional Laplace approximation from Section E of the Supplement (S2023). It appears that some rather sharp bounds on posterior concentration and contraction can be obtained under the critical dimension condition $\mathtt{p}_G \ll n$ similar to the case of pMLE; see Proposition 3.1. This is a substantial improvement over Spokoiny and Panov (2021) where $\mathtt{p}_G^3 \ll n$ was assumed. The main results of Theorem 3.4 provide some bounds on the accuracy of Gaussian approximation of the posterior. These results still require $\mathtt{p}_G^3 \ll n$. For the *total variation* distance, the accuracy of approximation is of order $\sqrt{\mathtt{p}_G^3/n}$. It can be improved to $\mathtt{p}_G^3/n$ if we limit ourselves to the class of centrally symmetric sets.

The use of posterior mean in place of MAP. The result on Gaussian approximation of the posterior justifies the use of elliptic credible set centered at the MAP $\widetilde{\boldsymbol{\upsilon}}_G$. The possibility of using the posterior mean $\overline{\boldsymbol{\upsilon}}_G$ in place of the MAP (1.1) is an important challenging question answered by Theorem 3.10: this use is justified under the same condition $\mathtt{p}_G^3 \ll n$, however, only after restricting to the class of elliptic credible sets. The proof involves some recent advances in Gaussian comparison Götze et al. (2019).

Log-density estimation. Section 4 specifies the general results to the case of log-density estimation. In particular, we provide finite sample explicit and sharp bounds on posterior concentration and contraction under the condition $s_0 > 0$ on the smoothness degree of the density while Rousseau and Szabo (2017) required $s_0 > 1/2$.

2 Properties of the pMLE $\widetilde{\boldsymbol{\upsilon}}_G$

This section collects general results about concentration and expansion of the pMLE which substantially improve the bounds from Spokoiny and Panov (2021). We assume to be given a pseudo log-likelihood random function $L(\boldsymbol{\upsilon})$, $\boldsymbol{\upsilon} \in \Upsilon \subseteq \mathbb{R}^p$, $p < \infty$. Given a quadratic penalty $\|G\boldsymbol{\upsilon}\|^2/2$, define

$$L_G(\boldsymbol{\upsilon}) = L(\boldsymbol{\upsilon}) - \|G\boldsymbol{\upsilon}\|^2/2.$$

Typical examples of choosing G^2 are given in Section B.1 of the Supplement (S2023). Consider the penalized MLE $\widetilde{\boldsymbol{\upsilon}}_G$ and its population counterpart $\boldsymbol{\upsilon}_G^*$

$$\widetilde{\boldsymbol{\upsilon}}_G = \operatorname*{argmax}_{\boldsymbol{\upsilon}} L_G(\boldsymbol{\upsilon}), \qquad \boldsymbol{\upsilon}_G^* = \operatorname*{argmax}_{\boldsymbol{\upsilon}} \mathbb{E} L_G(\boldsymbol{\upsilon}).$$

The corresponding Fisher information matrix $\mathbb{F}_G(\boldsymbol{\upsilon})$ is given by

$$\mathbb{F}(\boldsymbol{\upsilon}) = -\nabla^2 \mathbb{E} L(\boldsymbol{\upsilon}), \qquad \mathbb{F}_G(\boldsymbol{\upsilon}) = -\nabla^2 \mathbb{E} L_G(\boldsymbol{\upsilon}) = \mathbb{F}(\boldsymbol{\upsilon}) + G^2.$$

We assume $\mathbb{F}_G(\boldsymbol{\upsilon})$ to be positive definite for all considered $\boldsymbol{\upsilon}$. By $D_G(\boldsymbol{\upsilon})$ we denote a positive symmetric matrix with $D_G^2(\boldsymbol{\upsilon}) = \mathbb{F}_G(\boldsymbol{\upsilon})$, and $\mathbb{F}_G = \mathbb{F}_G(\boldsymbol{\upsilon}_G^*)$, $D_G = \mathbb{F}_G^{1/2}$.

2.1 Conditions

Now we present our conditions. The most important one is about linearity of the stochastic component $\zeta(\boldsymbol{\upsilon}) = L(\boldsymbol{\upsilon}) - \mathbb{E} L(\boldsymbol{\upsilon}) = L_G(\boldsymbol{\upsilon}) - \mathbb{E} L_G(\boldsymbol{\upsilon})$.

($\boldsymbol{\zeta}$) *The stochastic component* $\zeta(\boldsymbol{\upsilon}) = L(\boldsymbol{\upsilon}) - \mathbb{E} L(\boldsymbol{\upsilon})$ *of the process* $L(\boldsymbol{\upsilon})$ *is linear in* $\boldsymbol{\upsilon}$. *We denote by* $\nabla\zeta \equiv \nabla\zeta(\boldsymbol{\upsilon}) \in \mathbb{R}^p$ *its gradient*.

Below we assume some concentration properties of the stochastic vector $\nabla\zeta$; see (F.43) of Theorem F.15 of the Supplement (S2023).

$(\nabla\zeta)$ *Let* $V^2 = \mathrm{Var}(\nabla\zeta)$ *and* $D_G^2 = D_G^2(\upsilon_G^*)$. *Then for any considered* $\mathtt{x} > 0$

$$\mathbb{P}\bigl(\|D_G^{-1}\nabla\zeta\| \geq \mathtt{r}_G(\mathtt{x})\bigr) \leq 3\mathrm{e}^{-\mathtt{x}}, \tag{2.1}$$

where for $\mathtt{p}_G = \mathrm{tr}(D_G^{-2}V^2)$ and $\lambda_G = \|D_G^{-1}V^2D_G^{-1}\|$

$$\mathtt{r}_G(\mathtt{x}) \stackrel{\mathrm{def}}{=} \sqrt{\mathtt{p}_G} + \sqrt{2\mathtt{x}\lambda_G}\,. \tag{2.2}$$

This condition can be effectively checked if the errors in the data exhibit sub-Gaussian or sub-exponential behaviour; see Section F.3 of the Supplement (S2023). The important value $\mathtt{p}_G = \mathrm{tr}(D_G^{-2}V^2)$ can be called the *effective dimension*; see Spokoiny (2017).

We also assume that the deterministic part $\mathbb{E}L_G(\upsilon)$ of the penalized log-likelihood is a concave function. It can be relaxed using localization; see Spokoiny (2019).

$(\mathcal{C}_G)$ Υ *is an open and convex set in* $\mathbb{R}^p$. *The function* $\mathbb{E}L_G(\upsilon)$ *is concave on* Υ.

In Section 3.1 we consider a stronger condition of semi-concavity of $\mathbb{E}L(\upsilon)$. Further we will also need some smoothness conditions on the function $f(\upsilon) = \mathbb{E}L(\upsilon)$. The class of models satisfying the conditions **(ζ)**, **$(\nabla\zeta)$** with a smooth function $f(\upsilon) = \mathbb{E}L(\upsilon)$ will be referred to as *stochastically linear smooth* (SLS). This class includes linear regression, generalized linear models (GLM) and log-density models; see Spokoiny and Panov (2021). However, this class is much larger. For instance, nonlinear regression and nonlinear inverse problems can be adapted to the SLS framework by an extension of the parameter space; see Spokoiny (2019).

2.2 Concentration of the pMLE $\widetilde{\upsilon}_G$

This section discusses some concentration properties of the pMLE $\widetilde{\upsilon}_G = \mathrm{argmax}_{\upsilon} L_G(\upsilon)$.

Given $\mathtt{x}$ and $\mathtt{r}_G = \mathtt{r}_G(\mathtt{x})$ from (2.2), define for some $\nu < 1$ the set $\mathcal{U}_G$ by

$$\mathcal{U}_G \stackrel{\mathrm{def}}{=} \bigl\{u\colon \|D_G u\| \leq \nu^{-1}\mathtt{r}_G\bigr\}. \tag{2.3}$$

The result of this section states the concentration properties of the pMLE $\widetilde{\upsilon}_G$ in the local vicinity $\mathcal{A}_G$ of υ_G^* of the form

$$\mathcal{A}_G \stackrel{\mathrm{def}}{=} \upsilon_G^* + \mathcal{U}_G = \bigl\{\upsilon = \upsilon_G^* + u\colon u \in \mathcal{U}_G\bigr\} \subseteq \Upsilon^\circ.$$

Local Gateaux-regularity of $f(\upsilon) = \mathbb{E}L(\upsilon)$ within $\mathcal{A}_G$ will be measured by the error of the second order Taylor approximation

$$\begin{aligned} \delta_3(\upsilon, u) &= f(\upsilon + u) - f(\upsilon) - \langle \nabla f(\upsilon), u\rangle - \frac{1}{2}\langle \nabla^2 f(\upsilon), u^{\otimes 2}\rangle, \\ \delta_3'(\upsilon, u) &= \langle \nabla f(\upsilon + u), u\rangle - \langle \nabla f(\upsilon), u\rangle - \langle \nabla^2 f(\upsilon), u^{\otimes 2}\rangle\,. \end{aligned} \tag{2.4}$$

More precisely, define

$$\omega_G \stackrel{\text{def}}{=} \sup_{\boldsymbol{u} \in \mathcal{U}_G} \frac{2|\delta_3(\boldsymbol{\upsilon}_G^*, \boldsymbol{u})|}{\|D_G \boldsymbol{u}\|^2}, \qquad \omega_G' \stackrel{\text{def}}{=} \sup_{\boldsymbol{u} \in \mathcal{U}_G} \frac{2|\delta_3'(\boldsymbol{\upsilon}_G^*, \boldsymbol{u})|}{\|D_G \boldsymbol{u}\|^2}. \tag{2.5}$$

The quantities ω_G and ω_G' can be effectively bounded under smoothness conditions $(\boldsymbol{\mathcal{T}_3})$ or $(\boldsymbol{\mathcal{S}_3})$ given in Section A of the Supplement (S2023). Under $(\boldsymbol{\mathcal{T}_3})$ at $\boldsymbol{\upsilon} = \boldsymbol{\upsilon}_G^*$ with $\mathsf{D}^2(\boldsymbol{\upsilon}_G^*) = D_G^2$ and $\mathtt{r} = \mathtt{r}_G$, by Lemma A.1 of the Supplement (S2023), it holds for a small constant τ_3

$$\omega_G' \le \tau_3 \nu^{-1} \mathtt{r}_G, \qquad \omega_G \le \tau_3 \nu^{-1} \mathtt{r}_G / 3.$$

Furthermore, under $(\boldsymbol{\mathcal{S}_3})$, the same bounds apply with $\tau_3 = \mathtt{c}_3 n^{-1/2}$; see Lemma A.2 of the Supplement (S2023).

Proposition 2.1. *Suppose* $(\boldsymbol{\zeta})$, $(\boldsymbol{\nabla\zeta})$, *and* $(\boldsymbol{\mathcal{C}_G})$. *Let also*

$$1 - \nu - \omega_G' \ge 0; \tag{2.6}$$

see (2.5) *and* (2.3). *Then* $\widetilde{\boldsymbol{\upsilon}}_G \in \mathcal{A}_G$ *on a set* $\Omega(\mathtt{x})$ *with* $\mathbb{P}\bigl(\Omega(\mathtt{x})\bigr) \ge 1 - 3\mathrm{e}^{-\mathtt{x}}$, *i.e.*

$$\|D_G(\widetilde{\boldsymbol{\upsilon}}_G - \boldsymbol{\upsilon}_G^*)\| \le \nu^{-1} \mathtt{r}_G. \tag{2.7}$$

Proof. By $(\boldsymbol{\nabla\zeta})$, on a the random set $\Omega(\mathtt{x})$ with $\mathbb{P}(\Omega(\mathtt{x})) \ge 1 - 3\mathrm{e}^{-\mathtt{x}}$, it holds $\|D_G^{-1} \nabla\zeta\| \le \mathtt{r}_G$. Now the result follows from Proposition A.7 with $f(\boldsymbol{\upsilon}) = \mathbb{E}L_G(\boldsymbol{\upsilon})$, $g(\boldsymbol{\upsilon}) = L_G(\boldsymbol{\upsilon})$, $\mathtt{r} = \nu^{-1}\mathtt{r}_G$, and $\boldsymbol{A} = \nabla\zeta$. □

Remark 2.1. The result (2.7) continues to apply with any matrix $\mathbb{D}$ in place of D_G provided that $\mathbb{D} \le D_G$ and $(\boldsymbol{\nabla\zeta})$ as well as (2.5), (2.6) hold after this change.

2.3 Fisher and Wilks expansions

This section presents some finite sample results about the behavior of the penalized MLE $\widetilde{\boldsymbol{\upsilon}}_G$ and the excess $L_G(\widetilde{\boldsymbol{\upsilon}}_G) - L_G(\boldsymbol{\upsilon}_G^*)$. Proposition 2.1 states the concentration properties of $\widetilde{\boldsymbol{\upsilon}}_G$ around $\boldsymbol{\upsilon}_G^*$. Now we show that this concentration can be used to establish a version of the Fisher expansion for the estimation error $\widetilde{\boldsymbol{\upsilon}}_G - \boldsymbol{\upsilon}_G^*$ and the Wilks expansion for the excess $L_G(\widetilde{\boldsymbol{\upsilon}}_G) - L_G(\boldsymbol{\upsilon}_G^*)$.

Theorem 2.2. *Assume the conditions of Proposition 2.1 with* $\nu = 2/3$. *Then on* $\Omega(\mathtt{x})$

$$2L_G(\widetilde{\boldsymbol{\upsilon}}_G) - 2L_G(\boldsymbol{\upsilon}_G^*) - \bigl\|D_G^{-1}\nabla\zeta\bigr\|^2 \le \frac{\omega_G}{1-\omega_G}\bigl\|D_G^{-1}\nabla\zeta\bigr\|^2,$$

$$2L_G(\widetilde{\boldsymbol{\upsilon}}_G) - 2L_G(\boldsymbol{\upsilon}_G^*) - \bigl\|D_G^{-1}\nabla\zeta\bigr\|^2 \ge -\omega_G\bigl\|D_G^{-1}\nabla\zeta\bigr\|^2.$$

Also

$$\begin{aligned} \bigl\| D_G(\widetilde{\boldsymbol{\upsilon}}_G - \boldsymbol{\upsilon}_G^*) - D_G^{-1}\nabla\zeta \bigr\|^2 &\le \frac{3\omega_G}{(1-\omega_G)^2} \bigl\| D_G^{-1}\nabla\zeta \bigr\|^2 , \\ \bigl\| D_G(\widetilde{\boldsymbol{\upsilon}}_G - \boldsymbol{\upsilon}_G^*) \bigr\| &\le \frac{1+\sqrt{2\omega_G}}{1-\omega_G} \bigl\| D_G^{-1}\nabla\zeta \bigr\| . \end{aligned} \tag{2.8}$$

Proof. The result follows from Proposition A.8 of the Supplement (S2023) similarly to Proposition 2.1. □

2.4 Effective sample size and critical dimension in pMLE

This section discusses the important question of the critical parameter dimension still ensuring the validity of the presented results. A very important feature of our results is their dimension free and coordinate free form. The true parametric dimension p can be very large, it does not show up in the error terms. Neither do we use any spectral decomposition or sequence space structure, in particular, we do not require that the Fisher information matrix $\mathbb{F}$ and the penalty matrix G^2 are diagonal or can be jointly diagonalized. The results are stated for the general data $\boldsymbol{Y}$ and a quasi log-likelihood function. In particular, we do not assume independent or progressively dependent observations and additive structure of the log-likelihood. The *effective sample size* n can be defined via the smallest eigenvalue of the matrix $\mathbb{F}_G = D_G^2 = -\nabla^2 \mathbb{E} L_G(\boldsymbol{\upsilon}_G^*)$:

$$n^{-1} \stackrel{\text{def}}{=} \|\mathbb{F}_G^{-1}\|.$$

Our results apply as long as this value is sufficiently small. In typical examples like regression or density modeling such defined value is closely related to the sample size of the data.

For the concentration result of Proposition 2.1 we need the basic conditions **(ζ)** and **($\mathcal{C}_G$)**. Further, **($\nabla\zeta$)** identifies the radius $\mathtt{r}_G$ of the local vicinity $\mathcal{A}_G$. The final critical condition is given by (2.6). Essentially it says that the values ω_G and ω_G' are significantly smaller than 1. Under **($\mathcal{S}_3$)**, $\omega_G' \le \mathtt{c}_3\, \nu^{-1} \mathtt{r}_G\, n^{-1/2}$; see Lemma A.2 of the Supplement (S2023). So, (2.6) means $\mathtt{r}_G^2 \ll n$. Moreover, definition (2.1) of $\mathtt{r}_G$ yields that $\mathtt{r}_G^2 \asymp \mathrm{tr}(D_G^{-2} V^2) = \mathtt{p}_G$, where $\mathtt{p}_G$ is the *effective dimension* of the problem. We conclude that the main properties of the pMLE $\widetilde{\boldsymbol{\upsilon}}_G$ are valid under the condition $\mathtt{p}_G \ll n$ meaning sufficiently many observations per effective number of parameters.

2.5 The use of $\widetilde{D}_G^2$ instead of D_G^2

The penalized information matrix $D_G^2 = D_G^2(\boldsymbol{\upsilon}_G^*) = -\nabla^2 \mathbb{E} L_G(\boldsymbol{\upsilon}_G^*)$ plays an important role in our results. In particular, D_G describes the shape of the concentration set $\mathcal{A}_G = \boldsymbol{\upsilon}_G^* + \mathcal{U}_G$. However, this matrix is not available as it involves the unknown point $\boldsymbol{\upsilon}_G^*$. If the matrix function $\mathbb{F}(\boldsymbol{\upsilon})$ is locally constant in $\mathcal{A}_G$, one can replace $\boldsymbol{\upsilon}_G^*$ with its estimate $\widetilde{\boldsymbol{\upsilon}}_G$. Variability of $\mathbb{F}(\boldsymbol{\upsilon})$, or, equivalently, $\mathbb{F}_G(\boldsymbol{\upsilon}) = \mathbb{F}(\boldsymbol{\upsilon}) + G^2$ can be

measured under the Fréchet smoothness of $f(\boldsymbol{\upsilon}) = I\!\!E L_G(\boldsymbol{\upsilon})$ by the value ω_G^+ from (A.4) of the Supplement (S2023) with $\boldsymbol{\upsilon} = \boldsymbol{\upsilon}_G^*$, $\mathsf{D}(\boldsymbol{\upsilon}) = D_G$, and $\mathtt{r} = \nu^{-1}\mathtt{r}_G$.

Proposition 2.3. *Assume the conditions of Proposition 2.1 and let $\omega_G^+ \le 1/2$; see (A.4). The random matrix $\widetilde{D}_G^2 = I\!\!F_G(\widetilde{\boldsymbol{\upsilon}}_G)$ fulfills on $\Omega(\mathtt{x})$ for any $\boldsymbol{u} \in I\!\!R^p$*

$$\begin{aligned} &\bigl\| D_G^{-1} \widetilde{D}_G^2 \, D_G^{-1} - I_p \bigr\| \le \omega_G^+, \qquad \bigl\| D_G \, \widetilde{D}_G^{-2} \, D_G - I_p \bigr\| \le \frac{\omega_G^+}{1 - \omega_G^+}, \\ &(1 - \omega_G^+) \, \|D_G \boldsymbol{u}\|^2 \le \|\widetilde{D}_G \boldsymbol{u}\|^2 \le (1 + \omega_G^+) \, \|D_G \boldsymbol{u}\|^2. \end{aligned} \tag{2.9}$$

Proof. The value $\widetilde{\boldsymbol{\upsilon}}_G - \boldsymbol{\upsilon}_G^*$ belongs to $\mathcal{U}_G$ on $\Omega(\mathtt{x})$ and (2.9) follows from (A.5). □

2.6 Smoothness and bias

Due to Proposition 2.1, the penalized MLE $\widetilde{\boldsymbol{\upsilon}}_G$ is in fact an estimator of the vector $\boldsymbol{\upsilon}_G^*$. However, $\boldsymbol{\upsilon}_G^*$ depends on penalization which introduces some bias. This section discusses whether one can use $\widetilde{\boldsymbol{\upsilon}}_G$ for estimating the underlying truth $\boldsymbol{\upsilon}^*$ defined as the maximizer of the expected log-likelihood: $\boldsymbol{\upsilon}^* = \operatorname{argmax}_{\boldsymbol{\upsilon}} I\!\!E L(\boldsymbol{\upsilon})$. First we describe the bias $\boldsymbol{b}_G = \boldsymbol{\upsilon}_G^* - \boldsymbol{\upsilon}^*$ induced by penalization. It is important to mention that the previous results about the properties of the pMLE $\widetilde{\boldsymbol{\upsilon}}_G$ require strong concavity of the expected log-likelihood function $I\!\!E L_G(\boldsymbol{\upsilon})$ at least in a vicinity of the point $\boldsymbol{\upsilon}_G^*$. In some sense, this strong concavity is automatically forced by the penalizing term in the definition of $\boldsymbol{\upsilon}_G^*$. However, the underlying truth $\boldsymbol{\upsilon}^* = \operatorname{argmax}_{\boldsymbol{\upsilon}} I\!\!E L(\boldsymbol{\upsilon})$ is the maximizer of the non-penalized expected log-likelihood, and the corresponding Hessian $I\!\!F(\boldsymbol{\upsilon}^*) = -\nabla^2 I\!\!E L(\boldsymbol{\upsilon}^*)$ can degenerate. This makes evaluation of the bias more involved. To bypass this situation, we assume later in this section that the Hessian $\nabla^2 I\!\!E L_G(\boldsymbol{\upsilon})$ cannot change much in a reasonably large vicinity of $\boldsymbol{\upsilon}^*$. This allows to establish an accurate quadratic approximation of $f(\boldsymbol{\upsilon})$ and to evaluate the bias $\boldsymbol{b}_G = \boldsymbol{\upsilon}_G^* - \boldsymbol{\upsilon}^*$.

Define $I\!\!D_G$ by $I\!\!D_G^2 = I\!\!F_G(\boldsymbol{\upsilon}^*)$; cf. $D_G^2 = I\!\!F_G(\boldsymbol{\upsilon}_G^*)$. Let also Q be a symmetric matrix satisfying $Q^2 \le I\!\!D_G^2$. Typical examples include $Q = D_G$, $Q = I\!\!D_G$, and $Q^2 = n I_p$. Later we bound the norm $\|Q \boldsymbol{b}_G\|$. Denote with $\nu = 2/3$

$$\begin{aligned} \mathtt{b}_G &\stackrel{\text{def}}{=} \|Q \, I\!\!D_G^{-2} \, G^2 \boldsymbol{\upsilon}^*\|, \\ \omega_G^* &\stackrel{\text{def}}{=} \sup_{\boldsymbol{u}:\, \|Q\boldsymbol{u}\| \le \nu^{-1}\mathtt{b}_G} \|I\!\!D_G^{-1} \, I\!\!F_G(\boldsymbol{\upsilon}^* + \boldsymbol{u}) \, I\!\!D_G^{-1} - I_p\|; \end{aligned} \tag{2.10}$$

cf. (A.4) and (A.5) of the Supplement (S2023) for $f(\boldsymbol{\upsilon}) = I\!\!E L_G(\boldsymbol{\upsilon})$. Note that the definition of ω_G^+ in Proposition 2.3 uses another $\mathtt{r} = \nu^{-1}\mathtt{r}_G$, therefore, different notation. Proposition A.11 of the Supplement (S2023) yields the following result.

Proposition 2.4. *Let $I\!\!D_G^2 = I\!\!F_G(\boldsymbol{\upsilon}^*)$, $\nu \le 2/3$, and $\mathtt{b}_G = \|Q \, I\!\!D_G^{-2} \, G^2 \boldsymbol{\upsilon}^*\|$. Let also*

$\omega_G^* \le 1/3$; *see* (2.10). *Then the bias* $\boldsymbol{b}_G = \boldsymbol{\upsilon}_G^* - \boldsymbol{\upsilon}^*$ *fulfills*

$$\|Q\,\boldsymbol{b}_G\| \le \frac{\mathsf{b}_G}{1-\omega_G^*} = \frac{1}{1-\omega_G^*}\,\|Q\,\mathbb{D}_G^{-2}\,G^2\boldsymbol{\upsilon}^*\|\,, \tag{2.11}$$

$$\|Q(\boldsymbol{b}_G + \mathbb{D}_G^{-2}G^2\boldsymbol{\upsilon}^*)\| \le \frac{\omega_G^*}{1-\omega_G^*}\,\mathsf{b}_G = \frac{\omega_G^*}{1-\omega_G^*}\,\|Q\,\mathbb{D}_G^{-2}\,G^2\boldsymbol{\upsilon}^*\|.$$

Corollary 2.5. *Assume the conditions of Proposition 2.4. Then*

$$\|\mathbb{D}_G\,\boldsymbol{b}_G\| \le \frac{1}{1-\omega_G^*}\,\|\mathbb{D}_G^{-1}\,G^2\boldsymbol{\upsilon}^*\|, \qquad \|\boldsymbol{b}_G\| \le \frac{1}{1-\omega_G^*}\,\|\mathbb{D}_G^{-2}\,G^2\boldsymbol{\upsilon}^*\|. \tag{2.12}$$

The same bounds apply with $D_G^2 = \mathbb{F}_G(\boldsymbol{\upsilon}_G^*)$ *in place of* $\mathbb{D}_G^2 = \mathbb{F}_G(\boldsymbol{\upsilon}^*)$.

This is a special cases of (2.11) with $Q = \mathbb{D}_G$ and $Q = I_p$. The last statement is due to Remark A.1 of the Supplement (S2023).

2.7 Loss and risk of the pMLE

Now we combine the previous results about the stochastic term $\widetilde{\boldsymbol{\upsilon}}_G - \boldsymbol{\upsilon}_G^*$ and the bias term $\boldsymbol{b}_G = \boldsymbol{\upsilon}_G^* - \boldsymbol{\upsilon}^*$ to obtain the sharp bounds on the loss and risk of the pMLE $\widetilde{\boldsymbol{\upsilon}}_G$.

Theorem 2.6. *Assume the conditions of Proposition 2.1 and 2.4. Then on* $\Omega(\mathtt{x})$ *with* $\mathbb{P}\bigl(\Omega(\mathtt{x})\bigr) \ge 1 - 3\mathrm{e}^{-\mathtt{x}}$, *it holds with* $\boldsymbol{\xi}_G = D_G^{-1}\,\nabla\zeta$, $\mathtt{r}_G$ *from* (2.2), *and* $n^{-1} = \|D_G^{-2}\|$

$$\|D_G\,(\widetilde{\boldsymbol{\upsilon}}_G - \boldsymbol{\upsilon}^*)\| \le \frac{1+\sqrt{2\omega_G}}{1-\omega_G}\|\boldsymbol{\xi}_G\| + \frac{\|D_G^{-1}G^2\boldsymbol{\upsilon}^*\|}{1-\omega_G^*} \le \frac{1+\sqrt{2\omega_G}}{1-\omega_G}\mathtt{r}_G + \frac{\|D_G^{-1}G^2\boldsymbol{\upsilon}^*\|}{1-\omega_G^*}\,, \tag{2.13}$$

$$\|\widetilde{\boldsymbol{\upsilon}}_G - \boldsymbol{\upsilon}^*\| \le \frac{1+\sqrt{2\omega_G}}{\sqrt{n}\,(1-\omega_G)}\|\boldsymbol{\xi}_G\| + \frac{\|D_G^{-2}G^2\boldsymbol{\upsilon}^*\|}{1-\omega_G^*} \le \frac{3\,\mathtt{r}_G}{\sqrt{n}} + 3\,\|D_G^{-2}G^2\boldsymbol{\upsilon}^*\|\,. \tag{2.14}$$

Proof. Let $\Omega(\mathtt{x})$ be the random set from $(\boldsymbol{\nabla\zeta})$ on which with $\|\boldsymbol{\xi}_G\| \le \mathtt{r}_G$. It follows from (2.8) of Theorem 2.2 that on $\Omega(\mathtt{x})$ with $\boldsymbol{b}_G = \boldsymbol{\upsilon}_G^* - \boldsymbol{\upsilon}^*$

$$\bigl\|D_G\bigl(\widetilde{\boldsymbol{\upsilon}}_G - \boldsymbol{\upsilon}^*\bigr) + D_G\boldsymbol{b}_G\bigr\| \le \frac{1+\sqrt{2\omega_G}}{1-\omega_G}\,\bigl\|\boldsymbol{\xi}_G\bigr\|\,.$$

This and (2.12) imply (2.13). □

Now we state the results about the risk of the pMLE $\widetilde{\boldsymbol{\upsilon}}_G$. To avoid technical burden, we fix a large $\mathtt{x}$, $\mathtt{r}_G = \mathtt{r}_G(\mathtt{x})$, and exclude an event $\{\|\boldsymbol{\xi}_G\| > \mathtt{r}_G\}$ having an exponentially small probability; see condition (2.1) of Proposition 2.1.

Theorem 2.7. *Assume the conditions of Proposition 2.1 and Proposition 2.4. Then for a set* $\Omega(\mathtt{x})$ *with* $I\!\!P\big(\Omega(\mathtt{x})\big) \geq 1 - 3\mathrm{e}^{-\mathtt{x}}$, *it holds with* $\mathtt{p}_G = \mathrm{tr}(D_G^{-2} V^2)$

$$\begin{aligned} \big\| I\!\!E \left\{ D_G(\widetilde{\boldsymbol{\upsilon}}_G - \boldsymbol{\upsilon}^*)\, \mathbb{1}\big(\Omega(\mathtt{x})\big) \right\} \big\| &\leq \frac{1}{1-\omega_G^*} \| D_G^{-1} G^2 \boldsymbol{\upsilon}^* \| + \frac{\sqrt{3\omega_G}}{1-\omega_G}\, I\!\!E \| \boldsymbol{\xi}_G \| + \mathtt{C}_1 \mathrm{e}^{-\mathtt{x}} \\ &\leq \frac{1}{1-\omega_G^*} \| D_G^{-1} G^2 \boldsymbol{\upsilon}^* \| + \frac{\sqrt{3\omega_G}}{1-\omega_G} \sqrt{\mathtt{p}_G} + \mathtt{C}_1 \mathrm{e}^{-\mathtt{x}} , \end{aligned} \tag{2.15}$$

and

$$\begin{aligned} & I\!\!E \big\{ \| D_G \left(\widetilde{\boldsymbol{\upsilon}}_G - \boldsymbol{\upsilon}^* \right) \|^2 \, \mathbb{1}\big(\Omega(\mathtt{x})\big) \big\} \\ & \quad \leq \left(1 + \frac{\sqrt{2\omega_G}}{1-\omega_G} \right)^2 \mathtt{p}_G + \left(\frac{1}{1-\omega_G^*} \| D_G^{-1} G^2 \boldsymbol{\upsilon}^* \| + \frac{\sqrt{3\omega_G}}{1-\omega_G} \sqrt{\mathtt{p}_G} + \mathtt{C}_1 \mathrm{e}^{-\mathtt{x}} \right)^2 . \end{aligned} \tag{2.16}$$

Remark 2.2. For ω_G^*, ω_G small, (2.16) yields classical bias-variance decomposition:

$$I\!\!E \big\{ \| D_G \left(\widetilde{\boldsymbol{\upsilon}}_G - \boldsymbol{\upsilon}^* \right) \|^2 \, \mathbb{1}\big(\Omega(\mathtt{x})\big) \big\} \leq \big(\mathtt{p}_G + \| D_G^{-1} G^2 \boldsymbol{\upsilon}^* \|^2 \big) \big\{ 1 + o(1) \big\}. \tag{2.17}$$

With $n^{-1} = \| D_G^{-2} \|$, we also obtain

$$I\!\!E \big\{ n \| \widetilde{\boldsymbol{\upsilon}}_G - \boldsymbol{\upsilon}^* \|^2 \, \mathbb{1}\big(\Omega(\mathtt{x})\big) \big\} \leq \big(\mathtt{p}_G + n \| D_G^{-2} G^2 \boldsymbol{\upsilon}^* \|^2 \big) \big\{ 1 + o(1) \big\}.$$

Moreover, under the small bias condition $\| D_G^{-2} G^2 \boldsymbol{\upsilon}^* \|^2 \ll \mathtt{p}_G / n$, the impact of the bias induced by penalization is negligible. The relation $\| D_G^{-2} G^2 \boldsymbol{\upsilon}^* \|^2 \asymp \mathtt{p}_G / n$ is usually referred to as "bias-variance trade-off". Our bound is sharp in the sense that even for the special case of a linear models, (2.17) becomes equality.

Proof. Below we denote $I\!\!E_{\mathtt{x}} \eta = I\!\!E \big\{ \eta \, \mathbb{1}\big(\Omega(\mathtt{x})\big) \big\}$ for any r.v. η. As $I\!\!E \boldsymbol{\xi}_G = 0$, we derive

$$I\!\!E_{\mathtt{x}} \big\{ D_G(\widetilde{\boldsymbol{\upsilon}}_G - \boldsymbol{\upsilon}^*) \big\} = I\!\!E_{\mathtt{x}} D_G(\widetilde{\boldsymbol{\upsilon}}_G - \boldsymbol{\upsilon}^* - \boldsymbol{\xi}_G) - I\!\!E \boldsymbol{\xi}_G \, \mathbb{1}\big(\Omega^c(\mathtt{x})\big).$$

For the first term we apply (2.8) and (2.12) yielding

$$\| I\!\!E_{\mathtt{x}} D_G(\widetilde{\boldsymbol{\upsilon}}_G - \boldsymbol{\upsilon}^* - \boldsymbol{\xi}_G) \| \leq \frac{\| D_G^{-1} G^2 \boldsymbol{\upsilon}^* \|}{1-\omega_G^*} + \frac{\sqrt{3\omega_G \, \mathtt{p}_G}}{1-\omega_G} .$$

To show (2.15), we also have to bound the tail moments of $\| \boldsymbol{\xi}_G \|$:

$$\big\| I\!\!E \boldsymbol{\xi}_G \, \mathbb{1}\big(\Omega^c(\mathtt{x})\big) \big\| \leq I\!\!E \| \boldsymbol{\xi}_G \| \, \mathbb{1}\big(\Omega^c(\mathtt{x})\big) \leq \mathrm{e}^{-\mathtt{x}/2}.$$

This can be easily done using deviation bounds for the quadratic form $\| \boldsymbol{\xi}_G \|^2$; see Theorem F.9 of the Supplement (S2023). Similarly one can bound the variance of $D_G \widetilde{\boldsymbol{\upsilon}}_G$.

With $\mathbb{B}_G = D_G^{-1} V^2 D_G^{-1}$

$$\begin{aligned}\operatorname{Var}_{\mathbf{x}}\bigl(D_G\,\widetilde{\boldsymbol{\upsilon}}_G\bigr) &\le \mathbb{E}_{\mathbf{x}}\bigl\{D_G\,(\widetilde{\boldsymbol{\upsilon}}_G - \boldsymbol{\upsilon}_G^*)\bigr\}\bigl\{D_G\,(\widetilde{\boldsymbol{\upsilon}}_G - \boldsymbol{\upsilon}_G^*)\bigr\}^\top \\ &\le \left(1 + \frac{\sqrt{2\omega_G}}{1-\omega_G}\right)^2 \mathbb{E}(\boldsymbol{\xi}_G\,\boldsymbol{\xi}_G^\top) = \left(1 + \frac{\sqrt{2\omega_G}}{1-\omega_G}\right)^2 \mathbb{B}_G\,.\end{aligned}$$

This yields for the quadratic risk $\mathbb{E}\|D_G\,(\widetilde{\boldsymbol{\upsilon}}_G - \boldsymbol{\upsilon}^*)\|^2$

$$\mathbb{E}_{\mathbf{x}}\,\|D_G\,(\widetilde{\boldsymbol{\upsilon}}_G - \boldsymbol{\upsilon}^*)\|^2 \le \operatorname{tr}\operatorname{Var}_{\mathbf{x}}\bigl\{D_G\,(\widetilde{\boldsymbol{\upsilon}}_G - \boldsymbol{\upsilon}^*)\bigr\} + \bigl\|\mathbb{E}_{\mathbf{x}}\,D_G(\widetilde{\boldsymbol{\upsilon}}_G - \boldsymbol{\upsilon}^*)\bigr\|^2$$

and (2.16) follows. □

3 Laplace approximation of the posterior

This section studies the properties of the posterior $\boldsymbol{\upsilon}_G \,\big|\, \boldsymbol{Y}$. Our main result states Gaussian approximation of the posterior by $\mathcal{N}(\widetilde{\boldsymbol{\upsilon}}_G, \widetilde{D}_G^{-2})$. More specifically, our aim is, for any bounded measurable function g, to compare the conditional moments of $g(\boldsymbol{\upsilon}_G - \widetilde{\boldsymbol{\upsilon}}_G)$ and of $g(\widetilde{D}_G^{-1}\boldsymbol{\gamma})$, where $\boldsymbol{\gamma}$ is standard normal conditionally on $\boldsymbol{Y}$. The use of $\nabla L_G(\widetilde{\boldsymbol{\upsilon}}_G) = 0$ yields

$$\begin{aligned}\mathbb{E}\bigl\{g(\boldsymbol{\upsilon}_G - \widetilde{\boldsymbol{\upsilon}}_G)\,\big|\,\boldsymbol{Y}\bigr\} &= \frac{\int g(\boldsymbol{u} - \widetilde{\boldsymbol{\upsilon}}_G)\,\mathrm{e}^{L_G(\boldsymbol{u})}\,d\boldsymbol{u}}{\int \mathrm{e}^{L_G(\boldsymbol{u})}\,d\boldsymbol{u}} = \frac{\int g(\boldsymbol{u})\,\mathrm{e}^{L_G(\widetilde{\boldsymbol{\upsilon}}_G+\boldsymbol{u}) - L_G(\widetilde{\boldsymbol{\upsilon}}_G)}d\boldsymbol{u}}{\int \mathrm{e}^{L_G(\widetilde{\boldsymbol{\upsilon}}_G+\boldsymbol{u}) - L_G(\widetilde{\boldsymbol{\upsilon}}_G)}d\boldsymbol{u}} \\ &= \frac{\int g(\boldsymbol{u})\,\exp\bigl\{L_G(\widetilde{\boldsymbol{\upsilon}}_G + \boldsymbol{u}) - L_G(\widetilde{\boldsymbol{\upsilon}}_G) - \bigl\langle \nabla L_G(\widetilde{\boldsymbol{\upsilon}}_G), \boldsymbol{u}\bigr\rangle\bigr\}d\boldsymbol{u}}{\int \exp\bigl\{L_G(\widetilde{\boldsymbol{\upsilon}}_G + \boldsymbol{u}) - L_G(\widetilde{\boldsymbol{\upsilon}}_G) - \bigl\langle \nabla L_G(\widetilde{\boldsymbol{\upsilon}}_G), \boldsymbol{u}\bigr\rangle\bigr\}d\boldsymbol{u}}\,. \qquad (3.1)\end{aligned}$$

Now consider the Bregman divergence of the expected log-likelihood $f_G(\boldsymbol{\upsilon}) = \mathbb{E}L_G(\boldsymbol{\upsilon})$

$$f_G(\boldsymbol{\upsilon}; \boldsymbol{u}) = f_G(\boldsymbol{\upsilon} + \boldsymbol{u}) - f_G(\boldsymbol{\upsilon}) - \bigl\langle \nabla f_G(\boldsymbol{\upsilon}), \boldsymbol{u}\bigr\rangle, \quad \boldsymbol{u} \in \mathbb{R}^p\,.$$

As the stochastic term of $L(\boldsymbol{\upsilon})$ and thus, of $L_G(\boldsymbol{\upsilon})$ is linear in $\boldsymbol{\upsilon}$, it holds for any $\boldsymbol{\upsilon}, \boldsymbol{u}$

$$L_G(\boldsymbol{\upsilon} + \boldsymbol{u}) - L_G(\boldsymbol{\upsilon}) - \bigl\langle \nabla L_G(\boldsymbol{\upsilon}), \boldsymbol{u}\bigr\rangle = f_G(\boldsymbol{\upsilon} + \boldsymbol{u}) - f_G(\boldsymbol{u}) - \bigl\langle \nabla f_G(\boldsymbol{\upsilon}), \boldsymbol{u}\bigr\rangle = f_G(\boldsymbol{\upsilon}; \boldsymbol{u}).$$

Given $\widetilde{\boldsymbol{\upsilon}}_G = \boldsymbol{\upsilon}$, we derive from (3.1)

$$\mathbb{E}\bigl\{g(\boldsymbol{\upsilon}_G - \widetilde{\boldsymbol{\upsilon}}_G)\,\big|\,\boldsymbol{Y}\bigr\} = \mathbb{E}\bigl\{g(\boldsymbol{\upsilon}_G - \widetilde{\boldsymbol{\upsilon}}_G)\,\big|\,\widetilde{\boldsymbol{\upsilon}}_G = \boldsymbol{\upsilon}\bigr\} = \frac{\int g(\boldsymbol{u})\,\mathrm{e}^{f_G(\boldsymbol{\upsilon};\boldsymbol{u})}\,d\boldsymbol{u}}{\int \mathrm{e}^{f_G(\boldsymbol{\upsilon};\boldsymbol{u})}\,d\boldsymbol{u}}\,. \qquad (3.2)$$

This basic identity will be systematically used below. Laplace's approximation means nothing but the use of the second order Taylor approximation of the function $f_G(\cdot)$ at $\boldsymbol{\upsilon}$. Namely, $f_G(\boldsymbol{\upsilon}; \boldsymbol{u}) \approx -\|D_G(\boldsymbol{\upsilon})\,\boldsymbol{u}\|^2/2$ and

$$\frac{\int g(\boldsymbol{u})\,\mathrm{e}^{f_G(\boldsymbol{\upsilon};\boldsymbol{u})}\,d\boldsymbol{u}}{\int \mathrm{e}^{f_G(\boldsymbol{\upsilon};\boldsymbol{u})}\,d\boldsymbol{u}} \approx \frac{\int g(\boldsymbol{u})\,\mathrm{e}^{-\|D_G(\boldsymbol{\upsilon})\,\boldsymbol{u}\|^2/2}\,d\boldsymbol{u}}{\int \mathrm{e}^{-\|D_G(\boldsymbol{\upsilon})\,\boldsymbol{u}\|^2/2}\,d\boldsymbol{u}}\,.$$

The analysis includes two major steps: posterior concentration and a Gaussian approximation of the posterior distribution.

3.1 Posterior concentration

We start with the important technical result describing the concentration sets of the posterior. In all our result, the value $\mathtt{x}$ is fixed to ensure that $\mathrm{e}^{-\mathtt{x}}$ is negligible.

Proposition 2.1 enables us to restrict the study to the case with $\widetilde{\boldsymbol{\upsilon}}_G \in \mathcal{A}_G$. To describe the concentration properties of the posterior we need a slightly stronger concavity condition on $I\!\!E L(\boldsymbol{\upsilon})$, concavity of $I\!\!E L_G(\boldsymbol{\upsilon})$ is not sufficient.

$(\mathcal{C})$ The function $I\!\!E L(\boldsymbol{\upsilon})$ is concave.

This condition can be relaxed to *weak concavity*.

$(\mathcal{C}_\circ)$ There exists $G_\circ^2 \le G^2$ such that for any $\boldsymbol{\upsilon} \in \mathcal{A}_G$, the function $2I\!\!E L(\boldsymbol{\upsilon}+\boldsymbol{u}) - \|G_\circ \boldsymbol{u}\|^2$ is concave in $\boldsymbol{u}$.

$(\mathcal{C})$ is a special case of $(\mathcal{C}_\circ)$ with $G_\circ = 0$. In what follow we assume $(\mathcal{C})$. However, all the results apply under $(\mathcal{C}_\circ)$ after replacing D^2 with $D_\circ^2 = D^2 + G_\circ^2$. Define

$$\mathtt{p}(\boldsymbol{\upsilon}) \stackrel{\text{def}}{=} \operatorname{tr}\bigl\{D^2(\boldsymbol{\upsilon}) D_G^{-2}(\boldsymbol{\upsilon})\bigr\}, \qquad \mathtt{r}(\boldsymbol{\upsilon}) \stackrel{\text{def}}{=} 2\sqrt{\mathtt{p}(\boldsymbol{\upsilon})} + \sqrt{2\mathtt{x}}\,; \tag{3.3}$$

cf. (2.2) for $\mathtt{p}_G$ and $\mathtt{r}_G$. This ensures with $\boldsymbol{\gamma}$ standard normal

$$I\!\!P\bigl(\|D(\boldsymbol{\upsilon})\, D_G^{-1}(\boldsymbol{\upsilon})\, \boldsymbol{\gamma}\| > \mathtt{r}(\boldsymbol{\upsilon})\bigr) \le \mathrm{e}^{-\mathtt{x}},$$

see (F.11) of Corollary F.6 of the Supplement (S2023). With some fixed $\nu \le 1$, e.g. $\nu = 2/3$, define for any $\boldsymbol{\upsilon} \in \mathcal{A}_G$

$$\mathcal{U}(\boldsymbol{\upsilon}) = \bigl\{\boldsymbol{u}\colon \|D(\boldsymbol{\upsilon})\, \boldsymbol{u}\| \le \nu^{-1}\mathtt{r}(\boldsymbol{\upsilon})\bigr\}. \tag{3.4}$$

With $f(\boldsymbol{\upsilon}) = I\!\!E L(\boldsymbol{\upsilon})$ and $\delta_3(\boldsymbol{\upsilon}, \boldsymbol{u})$ from (2.4), local smoothness of $f(\cdot)$ at $\boldsymbol{\upsilon}$ will be measured by the value $\omega(\boldsymbol{\upsilon})$:

$$\omega(\boldsymbol{\upsilon}) \stackrel{\text{def}}{=} \sup_{\boldsymbol{u} \in \mathcal{U}(\boldsymbol{\upsilon})} \frac{1}{\|D\boldsymbol{u}\|^2/2} \bigl|\delta_3(\boldsymbol{\upsilon}, \boldsymbol{u})\bigr|; \tag{3.5}$$

cf. (2.5). Under $(\boldsymbol{\mathcal{S}_3})$, it holds $\omega(\boldsymbol{\upsilon}) \le \nu^{-1}\mathtt{c}_3\, \mathtt{r}(\boldsymbol{\upsilon})\, n^{-1/2}/3$; see Lemma A.2 of the Supplement (S2023).

Proposition 3.1. *Suppose* (ζ), $(\nabla\zeta)$, *and* $(\mathcal{C})$. *Let also* $\mathtt{p}(\boldsymbol{\upsilon})$ *and* $\mathtt{r}(\boldsymbol{\upsilon})$ *be defined by* (3.3) *and* $\mathcal{U}(\boldsymbol{\upsilon})$ *by* (3.4). *If* $\omega(\boldsymbol{\upsilon})$ *from* (3.5) *satisfies*

$$\omega(\boldsymbol{\upsilon}) \le 1/3, \qquad \boldsymbol{\upsilon} \in \mathcal{A}_G, \tag{3.6}$$

then on $\Omega(\mathtt{x})$, *it holds with* $\widetilde{D} = D(\widetilde{\boldsymbol{\upsilon}}_G)$ *and* $\widetilde{\mathtt{r}} = \mathtt{r}(\widetilde{\boldsymbol{\upsilon}}_G)$

$$\mathbb{P}\Big(\boldsymbol{\upsilon}_G - \widetilde{\boldsymbol{\upsilon}}_G \notin \widetilde{\mathcal{U}} \,\Big|\, \boldsymbol{Y}\Big) = \mathbb{P}\Big(\|\widetilde{D}(\boldsymbol{\upsilon}_G - \widetilde{\boldsymbol{\upsilon}}_G)\| > \widetilde{\mathtt{r}} \,\Big|\, \boldsymbol{Y}\Big) \le \mathrm{e}^{-\mathtt{x}}. \tag{3.7}$$

Proof. Let us fix $\widetilde{\boldsymbol{\upsilon}}_G = \boldsymbol{\upsilon}$ and apply (3.2) with $g(\boldsymbol{u}) = \mathbb{1}\big(\|D(\boldsymbol{\upsilon})\boldsymbol{u}\| \notin \mathcal{U}(\boldsymbol{\upsilon})\big)$. Then it suffices to bound uniformly in $\boldsymbol{\upsilon} \in \mathcal{A}_G$ the ratio

$$\rho(\boldsymbol{\upsilon}) \stackrel{\mathrm{def}}{=} \frac{\int \mathbb{1}\big(D(\boldsymbol{\upsilon})\boldsymbol{u} \notin \mathcal{U}(\boldsymbol{\upsilon})\big)\, \mathrm{e}^{f_G(\boldsymbol{\upsilon};\boldsymbol{u})} d\boldsymbol{u}}{\int \mathrm{e}^{f_G(\boldsymbol{\upsilon};\boldsymbol{u})} d\boldsymbol{u}}.$$

Now (E.9) of Theorem E.1 of the Supplement (S2023) yields the result. □

3.2 Posterior contraction

Now we bring together all the previous results to bound the posterior deviations $\boldsymbol{\upsilon}_G - \boldsymbol{\upsilon}^*$. The difference $\boldsymbol{\upsilon}_G - \boldsymbol{\upsilon}^*$ can be decomposed as

$$\boldsymbol{\upsilon}_G - \boldsymbol{\upsilon}^* = \big(\boldsymbol{\upsilon}_G - \widetilde{\boldsymbol{\upsilon}}_G\big) + \big(\widetilde{\boldsymbol{\upsilon}}_G - \boldsymbol{\upsilon}^*\big). \tag{3.8}$$

Result (2.14) of Theorem 2.6 provides a deviation bound for $\|\widetilde{\boldsymbol{\upsilon}}_G - \boldsymbol{\upsilon}^*\|$ while Proposition 3.1 claims concentration of the posterior on the set $\big\{\|\widetilde{D}(\boldsymbol{\upsilon}_G - \widetilde{\boldsymbol{\upsilon}}_G)\| \le \widetilde{\mathtt{r}}\big\}$. We conclude by the following result.

Proposition 3.2. *Assume the conditions of Theorem 2.6 and Proposition 3.1 and let* $\|D^{-2}(\boldsymbol{\upsilon})\| \le n^{-1}$ *for* $\boldsymbol{\upsilon} \in \mathcal{A}_G$. *It holds on* $\Omega(\mathtt{x})$

$$\mathbb{P}\Big(\|\boldsymbol{\upsilon}_G - \boldsymbol{\upsilon}^*\| \ge \|\widetilde{\boldsymbol{\upsilon}}_G - \boldsymbol{\upsilon}^*\| + \nu^{-1}\widetilde{\mathtt{r}}/\sqrt{n} \,\Big|\, \boldsymbol{Y}\Big) \le 2\mathrm{e}^{-\mathtt{x}}, \tag{3.9}$$

and $\|\widetilde{\boldsymbol{\upsilon}}_G - \boldsymbol{\upsilon}^*\|$ *satisfies* (2.14), *while* $\widetilde{\mathtt{r}} \le (1 - \omega_G^+)^{-1/2}\, \mathtt{r}(\boldsymbol{\upsilon}_G^*)$.

Proof. Bound (3.9) follows from decomposition (3.8) and Proposition 3.1. Further, the use of (2.9) of Proposition 2.3 yields $\widetilde{\mathtt{r}} \le (1 - \omega_G^+)^{-1/2}\, \mathtt{r}(\boldsymbol{\upsilon}_G^*)$. □

The use of (2.14) of Theorem 2.6 implies that most of posterior mass is concentrated in the root-n vicinity of $\boldsymbol{\upsilon}^*$:

$$\mathbb{P}\Big(\|\boldsymbol{\upsilon}_G - \boldsymbol{\upsilon}^*\| \ge 3\|D_G^{-2}G^2\boldsymbol{\upsilon}^*\| + \frac{3}{\sqrt{n}}\big(\mathtt{r}_G + \widetilde{\mathtt{r}}\big) \,\Big|\, \boldsymbol{Y}\Big) \le 2\mathrm{e}^{-\mathtt{x}}.$$

A prior ensuring the bias-variance trade-off leads to the optimal contraction rate which corresponds to the optimal penalty choice in penalized maximum likelihood estimation.

3.3 Gaussian approximation of the posterior

This section presents our main results about the accuracy of Gaussian approximation of the posterior $\upsilon_G \,|\, \boldsymbol{Y}$ in the total variation distance. The use of self-concordance type conditions from Section A helps to obtain very accurate and precise finite sample guarantees, which gradually improve the bounds from Spokoiny and Panov (2021).

Let $\mathcal{B}(\mathbb{R}^p)$ be the σ-field of all Borel sets in $\mathbb{R}^p$, while $\mathcal{B}_s(\mathbb{R}^p)$ stands for all centrally symmetric sets from $\mathcal{B}(\mathbb{R}^p)$.

Theorem 3.3. *Assume* **(ζ)**, **(∇ζ)**, *and* **(C)**. *Furthermore, let*

$$\omega(\upsilon)\,\mathtt{p}(\upsilon) \leq 2/3, \qquad \upsilon \in \mathcal{A}_G\,;$$

cf. (3.6). *Then with*

$$\diamondsuit_2(\upsilon) = \frac{0.75\,\omega(\upsilon)\,\mathtt{p}(\upsilon)}{1-\omega(\upsilon)}$$

and $\widetilde{\diamondsuit} = \diamondsuit_2(\widetilde{\upsilon}_G)$, *it holds on* $\Omega(\mathtt{x})$ *with*

$$\sup_{A\in\mathcal{B}(\mathbb{R}^p)} \Big| I\!\!P\big(\upsilon_G - \widetilde{\upsilon}_G \in A \,\big|\, \boldsymbol{Y}\big) - I\!\!P'\big(\widetilde{D}_G^{-1}\boldsymbol{\gamma} \in A\big)\Big| \leq \frac{2(\widetilde{\diamondsuit} + \mathrm{e}^{-\mathtt{x}})}{1-\widetilde{\diamondsuit} - \mathrm{e}^{-\mathtt{x}}} \leq 4(\widetilde{\diamondsuit} + \mathrm{e}^{-\mathtt{x}}). \quad (3.10)$$

Here $I\!\!P'$ *means a standard Gaussian distribution of* $\boldsymbol{\gamma}$ *given* $\boldsymbol{Y}$.

Now we present more advanced bounds on the error of Gaussian approximation under conditions **(S₃)** and **(S₄)** (resp. **(T₃)** and **(T₄)**) from Section A of the Supplement (S2023)for $\Upsilon^\circ = \mathcal{A}_G$.

Theorem 3.4. *Assume* **(ζ)**, **(∇ζ)**, **(C)**, **(T₃)**, *and let* $\tau_3\,\nu^{-1}\mathtt{r}(\upsilon) \leq 3/4$ *for* $\mathtt{r}(\upsilon)$ *from* (3.3) *and all* $\upsilon \in \mathcal{A}_G$. *Then the concentration bound* (3.7) *holds. Moreover, let*

$$\tau_3\,\nu^{-1}\mathtt{r}(\upsilon)\,\mathtt{p}(\upsilon) \leq 2, \qquad \upsilon \in \mathcal{A}_G\,. \quad (3.11)$$

With $\omega(\upsilon) \stackrel{\mathrm{def}}{=} \tau_3\,\mathtt{r}(\upsilon)/3 \leq 1/4$, *define*

$$\diamondsuit_3(\upsilon) \stackrel{\mathrm{def}}{=} \frac{\tau_3}{4\{1-\omega(\upsilon)\}^{3/2}}\,\{\mathtt{p}(\upsilon)+1\}^{3/2}\,.$$

Then the result (3.10) *applies on* $\Omega(\mathtt{x})$ *with* $\widetilde{\diamondsuit} = \diamondsuit_3(\widetilde{\upsilon}_G)$. *Moreover, under* **(T₄)**

$$\sup_{A\in\mathcal{B}_s(\mathbb{R}^p)} \Big| I\!\!P\big(\upsilon_G - \widetilde{\upsilon}_G \in A \,\big|\, \boldsymbol{Y}\big) - I\!\!P'\big(\widetilde{D}_G^{-1}\boldsymbol{\gamma} \in A\big)\Big| \leq \frac{2(\widetilde{\diamondsuit}_4 + \mathrm{e}^{-\mathtt{x}})}{1-\widetilde{\diamondsuit}_4 - \mathrm{e}^{-\mathtt{x}}} \leq 4(\widetilde{\diamondsuit}_4 + \mathrm{e}^{-\mathtt{x}})$$

with $\widetilde{\diamondsuit}_4 = \diamondsuit_4(\widetilde{\upsilon}_G)$ *and*

$$\diamondsuit_4(\upsilon) \stackrel{\mathrm{def}}{=} \frac{1}{16\{1-\omega(\upsilon)\}^2}\Big[\tau_3^2\,\big\{\mathtt{p}(\upsilon)+2\big\}^3 + 2\tau_4\big\{\mathtt{p}(\upsilon)+1\big\}^2\Big].$$

The results continue to apply with $(\boldsymbol{\mathcal{S}_3})$ *(resp.* $(\boldsymbol{\mathcal{S}_4})$*) in place of* $(\boldsymbol{\mathcal{T}_3})$ *(resp.* $(\boldsymbol{\mathcal{T}_4})$*) and* $\mathtt{c}_3\, n^{-1/2}$ *(resp.* $\mathtt{c}_4\, n^{-1}$*) in place of* τ_3 *(resp.* τ_4*).*

Proof of Theorem 3.3 (resp. Theorem 3.4). Similarly to the proof of Proposition 3.1, we restrict ourselves to the event $\widetilde{\boldsymbol{\upsilon}}_G \in \mathcal{A}_G$. Then we fix any possible value $\boldsymbol{\upsilon} \in \mathcal{A}_G$ of $\widetilde{\boldsymbol{\upsilon}}_G$ and use (3.2) to represent the posterior probability of a set A in the form

$$\mathbb{P}\bigl(\boldsymbol{\upsilon}_G - \widetilde{\boldsymbol{\upsilon}}_G \in A \,\big|\, \boldsymbol{Y}\bigr) = \frac{\int_A \mathrm{e}^{f_G(\boldsymbol{\upsilon};\boldsymbol{u})}\, d\boldsymbol{u}}{\int \mathrm{e}^{f_G(\boldsymbol{\upsilon};\boldsymbol{u})}\, d\boldsymbol{u}}.$$

Now the result follows by Theorem E.1 (resp. Theorem E.2). □

Under self-concordance conditions $(\boldsymbol{\mathcal{S}_3})$ and $(\boldsymbol{\mathcal{S}_4})$, constraint (3.11) reads as

$$\sup_{\boldsymbol{\upsilon} \in \mathcal{A}_G} \frac{\mathtt{c}_3\, \nu^{-1} \mathtt{r}(\boldsymbol{\upsilon})\, \mathtt{p}(\boldsymbol{\upsilon})}{n^{1/2}} \le 2.$$

As $\omega(\boldsymbol{\upsilon}) \le 1/4$, Theorem 3.4 yields on $\Omega(\mathtt{x})$ with $\widetilde{\mathtt{p}} = \mathtt{p}(\widetilde{\boldsymbol{\upsilon}}_G)$

$$\sup_{A \in \mathcal{B}(\mathbb{R}^p)} \Bigl| \mathbb{P}\bigl(\boldsymbol{\upsilon}_G - \widetilde{\boldsymbol{\upsilon}}_G \in A \,\big|\, \boldsymbol{Y}\bigr) - \mathbb{P}'\bigl(\widetilde{D}_G^{-1} \boldsymbol{\gamma} \in A\bigr) \Bigr| \le 2\, \mathtt{c}_3 \sqrt{\frac{(\widetilde{\mathtt{p}}+1)^3}{n}} + 4\mathrm{e}^{-\mathtt{x}},$$

$$\sup_{A \in \mathcal{B}_s(\mathbb{R}^p)} \Bigl| \mathbb{P}\bigl(\boldsymbol{\upsilon}_G - \widetilde{\boldsymbol{\upsilon}}_G \in A \,\big|\, \boldsymbol{Y}\bigr) - \mathbb{P}'\bigl(\widetilde{D}_G^{-1} \boldsymbol{\gamma} \in A\bigr) \Bigr| \le \frac{\mathtt{c}_3^2\, (\widetilde{\mathtt{p}}+2)^3 + 2\mathtt{c}_4 (\widetilde{\mathtt{p}}+1)^2}{2n} + 4\mathrm{e}^{-\mathtt{x}}.$$

3.4 Critical dimension in Bayesian inference

Posterior concentration in Proposition 3.1 only requires $\omega(\boldsymbol{\upsilon}) \ll 1$ for all $\boldsymbol{\upsilon} \in \mathcal{A}_G$. Under $(\boldsymbol{\mathcal{S}_3})$, one can bound $\omega(\boldsymbol{\upsilon}) \asymp \sqrt{\mathtt{p}(\boldsymbol{\upsilon})/n}$ yielding the condition $\mathtt{p}(\boldsymbol{\upsilon}) \ll n$ on the critical dimension which is essentially the same as the condition $\mathtt{p}_G \ll n$ for the pMLE. This is an important finding and an essential improvement of Spokoiny and Panov (2021). The main result of Theorem 3.3 requires $\omega(\boldsymbol{\upsilon})\, \mathtt{p}(\boldsymbol{\upsilon}) \ll 1$ which is much stronger because of the multiplicative factor $\mathtt{p}(\boldsymbol{\upsilon})$. Under $(\boldsymbol{\mathcal{S}_3})$, the remainder $\diamondsuit_3$ is of order $\sqrt{\mathtt{p}^3(\boldsymbol{\upsilon})/n}$ while under $(\boldsymbol{\mathcal{S}_4})$, $\diamondsuit_4 \asymp \mathtt{p}^3(\boldsymbol{\upsilon})/n$, still requiring $\mathtt{p}^3(\boldsymbol{\upsilon}) \ll n$. In some cases, e.g. for additive structure of the log-likelihood, it can be relaxed. However, it seems that the $\mathtt{p}^3(\boldsymbol{\upsilon}) \ll n$ condition is inherent in the problem and cannot be relaxed in general situation. We guess that in the region $n^{1/3} \ll \mathtt{p}(\boldsymbol{\upsilon}) \ll n$, another non-Gaussian type of limiting behavior of the posterior is well possible.

3.5 Laplace approximation with inexact parameters

Our main result of Theorem 3.3 states an approximation of the posterior distribution by the Gaussian measure with parameters $\widetilde{\boldsymbol{\upsilon}}_G$ and $\widetilde{D}_G^{-2}$. However, the vector

$\widetilde{\boldsymbol{\upsilon}}_G = \operatorname{argmax}_{\boldsymbol{\upsilon}} L_G(\boldsymbol{\upsilon})$ is typically hard to compute, because it solves a high dimensional optimization problem. If $\widetilde{\boldsymbol{\upsilon}}_G$ and thus $\widetilde{D}_G = D_G(\widetilde{\boldsymbol{\upsilon}}_G)$ are not available, one would be interested to use something more simple in place of $\widetilde{\boldsymbol{\upsilon}}_G$. Suppose to be given a vector $\widehat{\boldsymbol{\upsilon}}_G$ close to $\widetilde{\boldsymbol{\upsilon}}_G$ and a matrix $\check{D}_G^2$ close to $\widetilde{D}_G^2$. A typical example to keep in mind corresponds to $\widehat{\boldsymbol{\upsilon}}$ being the numerically evaluated posterior mean and H^2 being the posterior covariance, also evaluated numerically. Below we aim at presenting some sufficient conditions that ensure a reasonable approximation of the posterior by $\mathcal{N}(\widehat{\boldsymbol{\upsilon}}, H^{-2})$ using general results on Gaussian comparison; see Section G of the Supplement (S2023). For this result we need all the conditions of Theorem 3.3 corresponding to the special case with $\widehat{\boldsymbol{\upsilon}} = \widetilde{\boldsymbol{\upsilon}}_G$ and $H^2 = \widetilde{D}_G^2$. We write $\widehat{D} = D(\widehat{\boldsymbol{\upsilon}})$. Also we restrict ourselves to the class $\mathcal{B}_{el}(\mathbb{R}^p)$ of elliptic sets A in $\mathbb{R}^p$ of the form

$$A = \bigl\{\boldsymbol{\upsilon} \in \mathbb{R}^p \colon \|Q(\boldsymbol{\upsilon} - \widehat{\boldsymbol{\upsilon}})\| \le \mathtt{r}\bigr\}$$

for some linear mapping $Q \colon \mathbb{R}^p \to \mathbb{R}^q$ and $\mathtt{r} > 0$. Given two symmetric q-matrices Σ_1, Σ_2 and a vector $\boldsymbol{a} \in \mathbb{R}^q$, define

$$d(\Sigma_1, \Sigma_2, \boldsymbol{a}) \stackrel{\text{def}}{=} \left(\frac{1}{\|\Sigma_1\|_{\mathrm{Fr}}} + \frac{1}{\|\Sigma_2\|_{\mathrm{Fr}}} \right) \left(\|\Sigma_1 - \Sigma_2\|_1 + \|\boldsymbol{a}\|^2 \right); \tag{3.12}$$

see Section G. Obviously, if $\Sigma_1 \ge \Sigma_2$, then $\|\Sigma_1 - \Sigma_2\|_1 = \operatorname{tr}(\Sigma_1 - \Sigma_2)$.

Theorem 3.5. *Suppose the conditions of Theorem 3.3 to be fulfilled. Let also $\widehat{\boldsymbol{\upsilon}} \in \mathcal{A}_G$. Given $Q \colon \mathbb{R}^p \to \mathbb{R}^q$, define $\Sigma_1 = Q\widetilde{D}_G^{-2}Q^\top$, $\Sigma_2 = QH^{-2}Q^\top$, $\boldsymbol{a} = Q(\widehat{\boldsymbol{\upsilon}} - \widetilde{\boldsymbol{\upsilon}}_G)$ and suppose $\|\Sigma_j\|^2 \le 3\|\Sigma_j\|_{\mathrm{Fr}}^2$ for $j = 1, 2$. Then*

$$\sup_{\mathtt{r}>0} \bigl| I\!\!P\bigl(\|Q(\boldsymbol{\upsilon}_G - \widehat{\boldsymbol{\upsilon}})\| \le \mathtt{r} \,\big|\, \boldsymbol{Y}\bigr) - I\!\!P'\bigl(\|QH^{-1}\boldsymbol{\gamma}\| \le \mathtt{r}\bigr) \bigr| \le \frac{2(\diamondsuit + \mathrm{e}^{-\mathtt{x}})}{1 - \diamondsuit - \mathrm{e}^{-\mathtt{x}}} + \mathtt{C} d(\Sigma_1, \Sigma_2, \boldsymbol{a}) ,$$

where $\diamondsuit$ is from Theorem 3.3, $d(\cdot)$ from (3.12), and $\mathtt{C}$ is an absolute constant.

Proof. Use Theorem E.9 with $f(\boldsymbol{\upsilon}) = I\!\!E L_G(\boldsymbol{\upsilon})$, $\boldsymbol{x}^* = \widetilde{\boldsymbol{\upsilon}}_G$, $\boldsymbol{x} = \widehat{\boldsymbol{\upsilon}}$, and $\mathtt{D} = \widetilde{D}_G$. □

The result is particularly transparent if $H = \widetilde{D}_G$ or, if these two matrices are sufficiently close. Theorem E.10 yields the following bound.

Corollary 3.6. *Under the conditions of Theorem 3.5, it holds on $\Omega(\mathtt{x})$*

$$\sup_{\mathtt{r}>0} \Bigl| I\!\!P\bigl(\|Q(\boldsymbol{\upsilon}_G - \widehat{\boldsymbol{\upsilon}})\| \le \mathtt{r} \,\big|\, \boldsymbol{Y}\bigr) - I\!\!P'\bigl(\|Q\widetilde{D}_G^{-1}\boldsymbol{\gamma}\| \le \mathtt{r}\bigr) \Bigr| \le \frac{2(\diamondsuit + \mathrm{e}^{-\mathtt{x}})}{1 - \diamondsuit - \mathrm{e}^{-\mathtt{x}}} + \frac{\mathtt{C}\|Q(\widehat{\boldsymbol{\upsilon}} - \widetilde{\boldsymbol{\upsilon}}_G)\|^2}{\|Q\widetilde{D}_G^{-2}Q^\top\|_{\mathrm{Fr}}} .$$

3.6 Laplace approximation and Bernstein–von Mises Theorem

The prominent Bernstein–von Mises (BvM) Theorem claims asymptotic normality of the posterior distribution with the mean corresponding to the standard MLE $\widetilde{\boldsymbol{\upsilon}} = \operatorname{argmax}_{\boldsymbol{\upsilon}} L(\boldsymbol{\upsilon})$ and the variance $\widetilde{D}^{-2} = D^{-2}(\widetilde{\boldsymbol{\upsilon}})$. In particular, the prior does not

show up in this result, its impact on the posterior distribution becomes negligible as the sample size n grows. In our setup, the situation is different. The main results of Theorems 3.3 through 3.4 state another Gaussian approximation of the posterior with the mean $\widetilde{\boldsymbol{\upsilon}}_G$ and the variance $\widetilde{D}_G^{-2}$ both depending on the prior covariance G^{-2}. This dependence is important because the accuracy of approximation is given in terms of $\mathtt{p} = \mathtt{p}(\widetilde{\boldsymbol{\upsilon}}_G)$ also depending on G^{-2}. It is of interest to describe a kind of phase transition from the classical BvM approximation by $\mathcal{N}(\widetilde{\boldsymbol{\upsilon}}, \widetilde{D}^{-2})$ to the prior-dependent Laplace approximation by $\mathcal{N}(\widetilde{\boldsymbol{\upsilon}}_G, \widetilde{D}_G^{-2})$. Intuitively it is clear that the prior impact can be measured by the relation between the model-based Fisher information matrix D^2 and the prior precision matrix G^2. The main result below confirms this intuitive guess, however, the result is not trivial and requires a careful treatment based on Theorem 3.3 and the Gaussian comparison technique mentioned in Section 3.5. Implicitly we assume that the conditions ensuring concentration of the MLE $\widetilde{\boldsymbol{\upsilon}}$ corresponding to $G^2 = 0$ to be fulfilled. In particular, we need that $D^2(\boldsymbol{\upsilon})$ is sufficiently large for all $\boldsymbol{\upsilon}$ in the vicinity of $\boldsymbol{\upsilon}^*$. The radius of this vicinity is given by the value $\mathtt{r}_{\mathtt{x}} = \sqrt{\operatorname{tr}(D^{-2}V^2)} + \sqrt{2\mathtt{x}}$; see (2.2) of **$(\boldsymbol{\nabla\zeta})$**. Under correct model specification, it holds $V^2 \le \mathtt{C} D^2$ and $\mathtt{r}_{\mathtt{x}}^2 \le \mathtt{C} p$.

Theorem 3.7. *Under the conditions of Theorem 3.3, it holds on* $\Omega(\mathtt{x})$

$$\sup_{\mathtt{r}>0} \Bigl| I\!\!P\bigl(\|Q(\boldsymbol{\upsilon}_G - \widetilde{\boldsymbol{\upsilon}})\| \le \mathtt{r} \,\big|\, \boldsymbol{Y}\bigr) - I\!\!P'\bigl(\|Q\widetilde{D}^{-1}\boldsymbol{\gamma}\| \le \mathtt{r}\bigr) \Bigr|$$
$$\le \frac{2(\diamondsuit + \mathrm{e}^{-\mathtt{x}})}{1 - \diamondsuit - \mathrm{e}^{-\mathtt{x}}} + \frac{\mathtt{C}\|Q(\widetilde{\boldsymbol{\upsilon}} - \widetilde{\boldsymbol{\upsilon}}_G)\|^2}{\|Q\widetilde{D}^{-2}Q^\top\|_{\mathrm{Fr}}} + \mathtt{C}\|\widetilde{D}_G^{-1}G^2\widetilde{D}_G^{-1}\| \, \frac{\operatorname{tr}(Q\widetilde{D}^{-2}Q^\top)}{\|Q\widetilde{D}^{-2}Q^\top\|_{\mathrm{Fr}}} .$$

Moreover, with $Q = \widetilde{D}$, $\mathtt{r}_{\mathtt{x}} = \sqrt{\operatorname{tr}(D^{-2}V^2)} + \sqrt{2\mathtt{x}} \le \mathtt{C}\sqrt{p}$,

$$\sup_{\mathtt{r}>0} \Bigl| I\!\!P\bigl(\|\widetilde{D}(\boldsymbol{\upsilon}_G - \widetilde{\boldsymbol{\upsilon}})\| \le \mathtt{r} \,\big|\, \boldsymbol{Y}\bigr) - I\!\!P'\bigl(\|\boldsymbol{\gamma}\| \le \mathtt{r}\bigr) \Bigr|$$
$$\lesssim \diamondsuit + \mathrm{e}^{-\mathtt{x}} + \|G\boldsymbol{\upsilon}^*\|^2/\sqrt{p} + \|D_G^{-1}G^2D_G^{-1}\|^2\sqrt{p} .$$

Proof. Theorem 3.5 yields in view of $\widetilde{D}^{-2} \ge \widetilde{D}_G^{-2}$

$$\sup_{\mathtt{r}>0} \Bigl| I\!\!P\bigl(\|Q(\boldsymbol{\upsilon}_G - \widetilde{\boldsymbol{\upsilon}})\| \le \mathtt{r} \,\big|\, \boldsymbol{Y}\bigr) - I\!\!P'\bigl(\|Q\widetilde{D}^{-1}\boldsymbol{\gamma}\| \le \mathtt{r}\bigr) \Bigr|$$
$$\le \frac{2(\diamondsuit + \mathrm{e}^{-\mathtt{x}})}{1 - \diamondsuit - \mathrm{e}^{-\mathtt{x}}} + \frac{\mathtt{C}\|Q(\widetilde{\boldsymbol{\upsilon}} - \widetilde{\boldsymbol{\upsilon}}_G)\|^2}{\|Q\widetilde{D}^{-2}Q^\top\|_{\mathrm{Fr}}} + \frac{\mathtt{C}\operatorname{tr}\{Q(\widetilde{D}^{-2} - \widetilde{D}_G^{-2})Q^\top\}}{\|Q\widetilde{D}^{-2}Q^\top\|_{\mathrm{Fr}}} .$$

The last term here can easily be bounded:

$$\operatorname{tr}\{Q(\widetilde{D}^{-2} - \widetilde{D}_G^{-2})Q^\top\} = \operatorname{tr}\{Q\widetilde{D}^{-1}(I_p - \widetilde{D}\widetilde{D}_G^{-2}\widetilde{D})\widetilde{D}^{-1}Q^\top\}$$
$$\le \|I_p - \widetilde{D}\widetilde{D}_G^{-2}\widetilde{D}\| \operatorname{tr}(Q\widetilde{D}^{-2}Q^\top) = \|\widetilde{D}_G^{-1}G^2\widetilde{D}_G^{-1}\| \operatorname{tr}(Q\widetilde{D}^{-2}Q^\top).$$

Here we used that

$$\|\widetilde{D}_G^{-1} G^2 \widetilde{D}_G^{-1}\| = \|\widetilde{D}_G^{-1}(\widetilde{D}_G^2 - \widetilde{D}^2)\widetilde{D}_G^{-1}\| = \|I_p - \widetilde{D}_G^{-1}\widetilde{D}^2\widetilde{D}_G^{-1}\| = \|I_p - \widetilde{D}\widetilde{D}_G^{-2}\widetilde{D}\|.$$

For $Q = \widetilde{D}$, we use that $\|Q\widetilde{D}^{-2}Q^\top\|_{\mathrm{Fr}} = \sqrt{p}$, $\mathrm{tr}(Q\widetilde{D}^{-2}Q^\top) = p$, and apply the Fisher expansion (2.8) of Theorem 2.2 to $\widetilde{\boldsymbol{\upsilon}}$ and $\widetilde{\boldsymbol{\upsilon}}_G$. On $\Omega(\mathtt{x})$, it holds $\|D^{-1}\nabla\| \le \mathtt{r}_\mathtt{x}$ and with ω from (2.5)

$$\begin{aligned}\|D(\widetilde{\boldsymbol{\upsilon}} - \widetilde{\boldsymbol{\upsilon}}_G)\| &\le \|D(\boldsymbol{\upsilon}^* - \boldsymbol{\upsilon}_G^*)\| + \|(I_p - DD_G^{-2}D)D^{-1}\nabla\| \\ &\quad + \|D(\widetilde{\boldsymbol{\upsilon}} - \boldsymbol{\upsilon}^*) - D^{-1}\nabla\| + \|D(\widetilde{\boldsymbol{\upsilon}}_G - \boldsymbol{\upsilon}_G^* - D_G^{-2}\nabla)\| \\ &\le \|D(\boldsymbol{\upsilon}^* - \boldsymbol{\upsilon}_G^*)\| + \mathtt{C}\|D_G^{-1}G^2D_G^{-1}\|\,\mathtt{r}_\mathtt{x} + \mathtt{C}\omega\,\mathtt{r}_\mathtt{x}.\end{aligned}$$

By (2.11) of Proposition 2.4

$$\|D(\boldsymbol{\upsilon}^* - \boldsymbol{\upsilon}_G^*)\|^2 \le \|DD_G^{-2}D^\top\|\,\|G\boldsymbol{\upsilon}^*\|^2 \le \|G\boldsymbol{\upsilon}^*\|^2.$$

As $\mathtt{r}_\mathtt{x}^2 \le \mathtt{C}p$, $\omega\sqrt{p} \le \mathtt{C}\diamondsuit_3$, $\widetilde{D}^2 \le 2D^2$, and $\widetilde{D}_G^{-2} \le 2D_G^{-2}$, the assertion follows. □

Remark 3.1. The use of BvM requires rather strong bounds on the penalizing matrix G^2 and the related bias $\|G\boldsymbol{\upsilon}^*\|$. We need the condition of "light penalization" $\|D_G^{-1}G^2D_G^{-1}\| \ll p^{-1/2}$ which is much stronger than $G^2 \ll D^2$. Similarly, the "light bias" condition $\|G\boldsymbol{\upsilon}^*\|^2 \ll p^{1/2}$ is more restrictive than the "small bias" or "undersmoothing" condition $\|G\boldsymbol{\upsilon}^*\|^2 \ll p$.

3.7 Posterior mean

This section addresses an important question of using the posterior mean in place of the MAP $\widetilde{\boldsymbol{\upsilon}}_G$ for Bayesian inference. Our main result justifies the use of the posterior mean in place of the MAP under the same critical dimension condition $\mathtt{p}(\boldsymbol{\upsilon}) \ll n^{1/3}$ which is required for the Gaussian approximation result. First we quantify the deviation of the posterior mean $\overline{\boldsymbol{\upsilon}}_G$ from $\widetilde{\boldsymbol{\upsilon}}_G$. Then we apply Corollary 3.6 to measure the impact of using $\overline{\boldsymbol{\upsilon}}_G$ in place of $\widetilde{\boldsymbol{\upsilon}}_G$. By definition

$$\overline{\boldsymbol{\upsilon}}_G - \widetilde{\boldsymbol{\upsilon}}_G \stackrel{\mathrm{def}}{=} \mathbb{E}\big(\boldsymbol{\upsilon}_G \,\big|\, \boldsymbol{Y}\big) - \widetilde{\boldsymbol{\upsilon}}_G = \frac{\int (\boldsymbol{\upsilon} - \widetilde{\boldsymbol{\upsilon}}_G)\,\mathrm{e}^{L_G(\boldsymbol{\upsilon})}\,d\boldsymbol{\upsilon}}{\int \mathrm{e}^{L_G(\boldsymbol{\upsilon})}\,d\boldsymbol{\upsilon}}.$$

More precisely, we consider a linear mapping $Q\colon \mathbb{R}^p \to \mathbb{R}^q$ and evaluate the value $\big\|Q(\overline{\boldsymbol{\upsilon}}_G - \widetilde{\boldsymbol{\upsilon}}_G)\big\|$. The choice of Q is important. In particular, we cannot take $Q = D_G$ because this choice makes the bound dependent and linearly growing with p.

Theorem 3.8. *Assume the conditions of Theorem 3.4 and let $Q^\top Q \le D^2(\boldsymbol{\upsilon})$ for all $\boldsymbol{\upsilon} \in \mathcal{A}_G$. Then it holds with some absolute constant $\mathtt{C}$*

$$\|Q(\overline{\boldsymbol{\upsilon}}_G - \widetilde{\boldsymbol{\upsilon}}_G)\| \le 2.4\,\mathtt{c}_3\,\|Q\widetilde{D}^{-2}Q^\top\|^{1/2}\,(\widetilde{\mathtt{p}} + 1)^{3/2}\,n^{-1/2} + \mathtt{C}\mathrm{e}^{-\mathtt{x}}.$$

Proof. One can apply the same trick as before: by $\nabla L_G(\widetilde{\boldsymbol{\upsilon}}_G) = 0$

$$Q(\overline{\boldsymbol{\upsilon}}_G - \widetilde{\boldsymbol{\upsilon}}_G) = \frac{\int Q\boldsymbol{u} \exp\{L_G(\widetilde{\boldsymbol{\upsilon}}_G + \boldsymbol{u}) - L_G(\widetilde{\boldsymbol{\upsilon}}_G) - \langle \nabla L_G(\widetilde{\boldsymbol{\upsilon}}_G), \boldsymbol{u}\rangle\} \, d\boldsymbol{u}}{\int \exp\{L_G(\widetilde{\boldsymbol{\upsilon}}_G + \boldsymbol{u}) - L_G(\widetilde{\boldsymbol{\upsilon}}_G) - \langle \nabla L_G(\widetilde{\boldsymbol{\upsilon}}_G), \boldsymbol{u}\rangle\} \, d\boldsymbol{u}}.$$

For any particular value $\widetilde{\boldsymbol{\upsilon}}_G = \boldsymbol{\upsilon}$, stochastic linearity allows to replace the Bregman divergence of the log-likelihood $L_G(\boldsymbol{\upsilon})$ by the similar one for the expected log-likelihood $f_G(\boldsymbol{\upsilon}) = \mathbb{E} L_G(\boldsymbol{\upsilon})$. This yields

$$\|Q(\overline{\boldsymbol{\upsilon}}_G - \widetilde{\boldsymbol{\upsilon}}_G)\| \leq \left\| \frac{\int Q\boldsymbol{u} \, \mathrm{e}^{f_G(\boldsymbol{\upsilon};\boldsymbol{u})} \, d\boldsymbol{u}}{\int \mathrm{e}^{f_G(\boldsymbol{\upsilon};\boldsymbol{u})} \, d\boldsymbol{u}} \right\|.$$

Now we may apply (E.16) of Theorem E.6 of the Supplement (S2023) with $\mathsf{D} = D_G(\boldsymbol{\upsilon})$. □

Corollary 3.9. *Assume the conditions of Theorem 3.4. Then*

$$\|\widetilde{D}(\overline{\boldsymbol{\upsilon}}_G - \widetilde{\boldsymbol{\upsilon}}_G)\| \leq 2.4 \, \mathtt{c}_3 \, (\widetilde{\mathtt{p}} + 1)^{3/2} n^{-1/2} + \mathtt{C} \mathrm{e}^{-\mathtt{x}}.$$

Now we put together the result of Theorem 3.8 and the accuracy bound from Theorem 3.5. To make the result more transparent, assume $Q = \widetilde{D}$ and $H = \widetilde{D}_G$.

Theorem 3.10. *Assume the conditions of Theorem 3.4. Then on* $\Omega(\mathtt{x})$

$$\sup_{\mathtt{r}>0} \left| I\!P\big(\|\widetilde{D}(\boldsymbol{\upsilon}_G - \overline{\upsilon}_G)\| \leq \mathtt{r} \,\big|\, \boldsymbol{Y}\big) - I\!P'\big(\|\widetilde{D}\widetilde{D}_G^{-1}\boldsymbol{\gamma}\| \leq \mathtt{r}\big) \right| \leq \mathtt{C}\Big(\frac{(\widetilde{\mathtt{p}} + 1)^{3/2}}{n^{1/2}} + \mathrm{e}^{-\mathtt{x}} \Big).$$

We conclude that the use of posterior mean in place of posterior mode is possible under the same condition $\widetilde{\mathtt{p}}^3 \ll n$. This is a non-trivial result based on recent progress in Gaussian probability from Götze et al. (2019).

4 Log-density estimation

Suppose we are given a random sample $X_1, \ldots, X_n$ in $I\!R^d$. The density model assumes that all these random variables are independent identically distributed from some measure P with a density $f(\boldsymbol{x})$ with respect to a σ-finite measure μ_0 in $I\!R^d$. This density function is the target of estimation. By definition, the function f is non-negative, measurable, and integrates to one: $\int f(\boldsymbol{x}) \, d\mu_0(\boldsymbol{x}) = 1$. Here and below, the integral $\int$ without limits means the integral over the whole space $I\!R^d$. 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 $\{\psi_j(\boldsymbol{x}), \, j = 1, \ldots, p\}$ with $p \leq \infty$ be a collection of functions in $I\!R^d$ (a dictionary). For each $\boldsymbol{\upsilon} = (\upsilon_j) \in I\!R^p$, define

$$\ell(\boldsymbol{x}, \boldsymbol{\upsilon}) \stackrel{\text{def}}{=} \upsilon_1 \psi_1(\boldsymbol{x}) + \ldots + \upsilon_p \psi_p(\boldsymbol{x}) - \phi(\boldsymbol{\upsilon}) = \big\langle \boldsymbol{\Psi}(\boldsymbol{x}), \boldsymbol{\upsilon} \big\rangle - \phi(\boldsymbol{\upsilon}),$$

where $\boldsymbol{\Psi}(\boldsymbol{x})$ is a vector with components $\psi_j(\boldsymbol{x})$ and $\phi(\boldsymbol{\upsilon})$ is given by

$$\phi(\boldsymbol{\upsilon}) \stackrel{\text{def}}{=} \log \int \mathrm{e}^{\langle \boldsymbol{\Psi}(\boldsymbol{x}), \boldsymbol{\upsilon} \rangle} \, d\mu_0(\boldsymbol{x}). \tag{4.1}$$

It is worth stressing that the data point $\boldsymbol{x}$ only enters in the linear term $\langle \boldsymbol{\Psi}(\boldsymbol{x}), \boldsymbol{\upsilon} \rangle$ of the log-likelihood $\ell(\boldsymbol{x}, \boldsymbol{\upsilon})$. The function $\phi(\boldsymbol{\upsilon})$ is entirely model-driven. Below we restrict $\boldsymbol{\upsilon}$ to a subset Υ in $\mathbb{R}^p$ such that $\phi(\boldsymbol{\upsilon})$ is well defined and the integral of $\mathrm{e}^{\langle \boldsymbol{\Psi}(\boldsymbol{x}), \boldsymbol{\upsilon} \rangle}$ is finite. Linear log-density modeling assumes

$$\log f(\boldsymbol{x}) = \ell(\boldsymbol{x}, \boldsymbol{\upsilon}^*) = \langle \boldsymbol{\Psi}(\boldsymbol{x}), \boldsymbol{\upsilon}^* \rangle - \phi(\boldsymbol{\upsilon}^*) \tag{4.2}$$

for some $\boldsymbol{\upsilon}^* \in \Upsilon \subseteq \mathbb{R}^p$. A nice feature of such representation is that the function $\log f(\boldsymbol{x})$ 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 $\boldsymbol{\upsilon}$. With $S = \sum_{i=1}^n \boldsymbol{\Psi}(X_i)$, for a given penalty operator G^2, the penalized log-likelihood $L_G(\boldsymbol{\upsilon})$ reads as

$$L_G(\boldsymbol{\upsilon}) = \sum_{i=1}^n \langle \boldsymbol{\Psi}(X_i), \boldsymbol{\upsilon} \rangle - n\phi(\boldsymbol{\upsilon}) - \frac{1}{2}\|G\boldsymbol{\upsilon}\|^2 = \langle S, \boldsymbol{\upsilon} \rangle - n\phi(\boldsymbol{\upsilon}) - \frac{1}{2}\|G\boldsymbol{\upsilon}\|^2.$$

The penalized MLE $\widetilde{\boldsymbol{\upsilon}}_G$ and its population counterpart $\boldsymbol{\upsilon}_G^*$ are defined as

$$\widetilde{\boldsymbol{\upsilon}}_G = \operatorname*{argmax}_{\boldsymbol{\upsilon} \in \Upsilon} L_G(\boldsymbol{\upsilon}), \qquad \boldsymbol{\upsilon}_G^* = \operatorname*{argmax}_{\boldsymbol{\upsilon} \in \Upsilon} \mathbb{E} L_G(\boldsymbol{\upsilon}).$$

4.1 Conditions

For applying the general results of Section 2 and Section 3, it suffices to check the general conditions of Section 2 for the log-density model. First note that the generalized linear structure of the model automatically yields conditions **(C)** and **(ζ)**. Indeed, convexity of $\phi(\cdot)$ implies that $\mathbb{E}L(\boldsymbol{\upsilon}) = \langle \mathbb{E}S, \boldsymbol{\upsilon} \rangle - n\phi(\boldsymbol{\upsilon})$ is concave. Further, for the stochastic component $\zeta(\boldsymbol{\upsilon}) = L(\boldsymbol{\upsilon}) - \mathbb{E}L(\boldsymbol{\upsilon})$, it holds

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

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

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

To simplify our presentation, we assume that $X_1, \ldots, X_n$ are indeed i.i.d. and the density $f(\boldsymbol{x})$ can be represented in the form (4.2) for some parameter vector $\boldsymbol{\upsilon}^*$. This can be easily extended to non i.i.d. case at cost of more complicated notations. Then

$$\boldsymbol{\upsilon}^* = \operatorname*{argmax}_{\boldsymbol{\upsilon} \in \Upsilon} \mathbb{E}L(\boldsymbol{\upsilon}) = \operatorname*{argmax}_{\boldsymbol{\upsilon} \in \Upsilon} \bigl\{ \langle \mathbb{E}S, \boldsymbol{\upsilon} \rangle - n\phi(\boldsymbol{\upsilon}) \bigr\} = \operatorname*{argmax}_{\boldsymbol{\upsilon} \in \Upsilon} \bigl\{ \langle \overline{\Psi}, \boldsymbol{\upsilon} \rangle - \phi(\boldsymbol{\upsilon}) \bigr\}, \tag{4.3}$$

where $\overline{\Psi} = I\!\!E \Psi(X_1)$ and $I\!\!E S = n\overline{\Psi}$. This yields the identity

$$\nabla \phi(\upsilon^*) = \overline{\Psi}.$$

Moreover, by (4.1), $\nabla^2 \phi(\upsilon^*) = \mathrm{Var}\{\Psi(X_1)\}$ and

$$V^2 = \mathrm{Var}(\nabla \zeta) = n \nabla^2 \phi(\upsilon^*) = I\!\!F(\upsilon^*). \tag{4.4}$$

Here we present our conditions. For any $\upsilon \in \Upsilon$ and $\varrho > 0$, define $\mathtt{m}(\upsilon)$ by $\mathtt{m}^2(\upsilon) = \nabla^2 \phi(\upsilon)$ and consider the corresponding balls in $I\!\!R^p$

$$\mathcal{B}_\varrho(\upsilon) \stackrel{\mathrm{def}}{=} \{u \in I\!\!R^p : \|\mathtt{m}(\upsilon) u\| \le \varrho\} = \{u \in I\!\!R^p : \langle \nabla^2 \phi(\upsilon), u^{\otimes 2} \rangle \le \varrho^2\}.$$

(f) $X_1, \ldots, X_n$ are i.i.d. from a density f satisfying $\log f(x) = \Psi(x)^\top \upsilon^* - \phi(\upsilon^*)$.

(Υ) The set Υ is open and convex, the value $\phi(\upsilon)$ from (4.1) is finite for all $\upsilon \in \Upsilon$, υ^* from (4.3) is an internal point in Υ such that $\mathcal{B}_{2\varrho}(\upsilon^*) \subset \Upsilon$ for a fixed $\varrho > 0$.

(ϕ) For the Bregman divergence $\phi(\upsilon; u) \stackrel{\mathrm{def}}{=} \phi(\upsilon + u) - \phi(\upsilon) - \langle \nabla \phi(\upsilon), u \rangle$, it holds

$$\sup_{\upsilon \in \mathcal{B}_\varrho(\upsilon^*)} \sup_{u \in \mathcal{B}_{2\varrho}(\upsilon)} \exp \phi(\upsilon; u) \le \mathtt{C}_\varrho. \tag{4.5}$$

Introduce a measure P_υ by the relation:

$$\frac{dP_\upsilon}{d\mu_0}(x) = \exp\{\langle \Psi(x), \upsilon \rangle - \phi(\upsilon)\}. \tag{4.6}$$

Identity (4.1) ensures that P_υ is a probabilistic measure. Moreover, under (4.2), the data generating measure $I\!\!P$ coincides with $P_{\upsilon^*}^{\otimes n}$.

(Ψ_4) *There are* $\mathtt{C}_{\Psi,3} \ge 0$ *and* $\mathtt{C}_{\Psi,4} \ge 3$ *such that for all* $\upsilon \in \mathcal{B}_\varrho(\upsilon^*)$ *and* $\gamma \in I\!\!R^p$

$$\begin{aligned} \left| E_\upsilon \langle \Psi(X_1) - E_\upsilon \Psi(X_1), \gamma \rangle^3 \right| &\le \mathtt{C}_{\Psi,3} E_\upsilon^{3/2} \langle \Psi(X_1) - E_\upsilon \Psi(X_1), \gamma \rangle^2, \\ E_\upsilon \langle \Psi(X_1) - E_\upsilon \Psi(X_1), \gamma \rangle^4 &\le \mathtt{C}_{\Psi,4} E_\upsilon^2 \langle \Psi(X_1) - E_\upsilon \Psi(X_1), \gamma \rangle^2. \end{aligned}$$

In fact, conditions **(ϕ)** and **(Ψ_4)** follow from **(Υ)** and can be considered as a kind of definition of important quantities $\mathtt{C}_\varrho$, $\mathtt{C}_{\Psi,3}$, and $\mathtt{C}_{\Psi,4}$ which will be used for describing the smoothness properties of $\phi(\upsilon)$. The matrix $\nabla^2 \phi(\upsilon)$ is supposed well conditioned for $\upsilon \in \mathcal{B}_\varrho(\upsilon^*)$.

($\nabla^2 \phi$) For the information matrix $\nabla^2 \phi(\upsilon)$, it holds with some $\mathtt{C}_I\!\!F \ge 1$

$$\mathtt{C}_I\!\!F^{-1} I_p \le \nabla^2 \phi(\upsilon) \le \mathtt{C}_I\!\!F I_p, \qquad \upsilon \in \mathcal{B}_\varrho(\upsilon^*). \tag{4.7}$$

Later we show that it suffices to check (4.7) at υ^*, then it will be fulfilled in $\mathcal{B}_\varrho(\upsilon^*)$ with a slightly larger constant $\mathtt{C}_I\!\!F$.

Check of conditions $(\mathcal{S}_3)$ and $(\mathcal{S}_4)$

Let P_{υ} be defined by (4.6). It is straightforward to check that $E_{\upsilon}\boldsymbol{\Psi}(X_1) = \nabla\phi(\upsilon)$ and $\mathrm{Var}_{\upsilon}(\boldsymbol{\Psi}(X_1)) = \nabla^2\phi(\upsilon)$. Further, if $\boldsymbol{u} \in \mathcal{B}_{\varrho}(\upsilon)$ and $\upsilon + \boldsymbol{u} \in \Upsilon$, then

$$\phi(\upsilon + \boldsymbol{u}) = \log E_0 \exp\{\langle \boldsymbol{\Psi}(X_1), \upsilon + \boldsymbol{u}\rangle\} = \log E_{\upsilon} \exp\bigl\{\bigl\langle \boldsymbol{\Psi}(X_1), \boldsymbol{u}\bigr\rangle + \phi(\upsilon)\bigr\}.$$

This yields in view of $E_{\upsilon}\boldsymbol{\Psi}(X_1) = \nabla\phi(\upsilon)$ that $\boldsymbol{\varepsilon} = \boldsymbol{\Psi}(X_1) - E_{\upsilon}\boldsymbol{\Psi}(X_1)$ fulfills

$$\begin{aligned} \log E_{\upsilon} \exp(\langle \boldsymbol{\varepsilon}, \boldsymbol{u}\rangle) &= \phi(\upsilon + \boldsymbol{u}) - \phi(\upsilon) - \langle E_{\upsilon}\boldsymbol{\Psi}(X_1), \boldsymbol{u}\rangle \\ &= \phi(\upsilon + \boldsymbol{u}) - \phi(\upsilon) - \langle \nabla\phi(\upsilon), \boldsymbol{u}\rangle. \end{aligned} \tag{4.8}$$

Lemma 4.1. *The function $\phi(\upsilon)$ satisfies for any $\upsilon \in \mathcal{B}_{\varrho}(\upsilon^*)$ and $\boldsymbol{\gamma} \in \mathbb{R}^p$*

$$|\langle \nabla^3\phi(\upsilon), \boldsymbol{\gamma}^{\otimes 3}\rangle| \le \mathtt{C}_{\Psi,3} \langle \nabla^2\phi(\upsilon), \boldsymbol{\gamma}^{\otimes 2}\rangle^{3/2}, \tag{4.9}$$

$$|\langle \nabla^4\phi(\upsilon), \boldsymbol{\gamma}^{\otimes 4}\rangle| \le (\mathtt{C}_{\Psi,4} - 3) \langle \nabla^2\phi(\upsilon), \boldsymbol{\gamma}^{\otimes 2}\rangle^2. \tag{4.10}$$

Moreover, for any $\boldsymbol{\gamma}_1, \boldsymbol{\gamma} \in \mathbb{R}^p$

$$|\langle \nabla^3\phi(\upsilon), \boldsymbol{\gamma}_1 \otimes \boldsymbol{\gamma}^{\otimes 2}\rangle| \le \sqrt{\mathtt{C}_{\Psi,4} \langle \nabla^2\phi(\upsilon), \boldsymbol{\gamma}_1^{\otimes 2}\rangle}\ \langle \nabla^2\phi(\upsilon), \boldsymbol{\gamma}^{\otimes 2}\rangle.$$

Proof. Denote $\boldsymbol{\varepsilon} = X_1 - E_{\upsilon}X_1$. By (4.8) with $\boldsymbol{u} = t\boldsymbol{\gamma}$ for t sufficiently small

$$\chi(t) \stackrel{\mathrm{def}}{=} \log E_{\upsilon} \exp(t\langle \boldsymbol{\varepsilon}, \boldsymbol{\gamma}\rangle) = \phi(\upsilon + t\boldsymbol{\gamma}) - \phi(\upsilon) - \langle \nabla\phi(\upsilon), t\boldsymbol{\gamma}\rangle,$$

and by $(\boldsymbol{\Psi}_4)$ with $\mathtt{C}_{\Psi,4} \ge 3$

$$\begin{aligned} \bigl|\chi^{(3)}(0)\bigr| &= \bigl|E_{\upsilon}\langle \boldsymbol{\varepsilon}, \boldsymbol{\gamma}\rangle^3\bigr| \le \mathtt{C}_{\Psi,3}\, E_{\upsilon}^{3/2}\langle \boldsymbol{\varepsilon}, \boldsymbol{\gamma}\rangle^2, \\ \bigl|\chi^{(4)}(0)\bigr| &= \bigl|E_{\upsilon}\langle \boldsymbol{\varepsilon}, \boldsymbol{\gamma}\rangle^4 - 3E_{\upsilon}^2\langle \boldsymbol{\varepsilon}, \boldsymbol{\gamma}\rangle^2\bigr| \le (\mathtt{C}_{\Psi,4} - 3)E_{\upsilon}^2\langle \boldsymbol{\varepsilon}, \boldsymbol{\gamma}\rangle^2. \end{aligned}$$

If $\boldsymbol{\gamma}_1 \ne \boldsymbol{\gamma}$ then we may proceed in a similar way with the bivariate function $\chi(t_1, t) = \log E_{\upsilon} \exp\bigl\{t_1\langle \boldsymbol{\varepsilon}, \boldsymbol{\gamma}_1\rangle + t\langle \boldsymbol{\varepsilon}, \boldsymbol{\gamma}\rangle\bigr\}$. Its mixed derivative at zero satisfies

$$\left|\frac{\partial^3}{\partial t_1 \partial t^2}\chi(0,0)\right| = \bigl|E_{\upsilon}\langle \boldsymbol{\varepsilon}, \boldsymbol{\gamma}_1\rangle \langle \boldsymbol{\varepsilon}, \boldsymbol{\gamma}\rangle^2\bigr| \le \bigl\{E_{\upsilon}\langle \boldsymbol{\varepsilon}, \boldsymbol{\gamma}_1\rangle^2\, E_{\upsilon}\langle \boldsymbol{\varepsilon}, \boldsymbol{\gamma}\rangle^4\bigr\}^{1/2}$$

and the result follows as well. □

Lemma 4.2. *If $\upsilon \in \mathcal{B}_{\varrho}(\upsilon^*)$ then with $\mathtt{c}_{\phi} = \sqrt{\mathtt{C}_{\Psi,4}\,\mathtt{C}_{\varrho}}$*

$$\sup_{\boldsymbol{u} \in \mathcal{B}_{\varrho}(\upsilon)} \sup_{\boldsymbol{\gamma} \in \mathbb{R}^p} \frac{\langle \nabla^2\phi(\upsilon + \boldsymbol{u}), \boldsymbol{\gamma}^{\otimes 2}\rangle}{\langle \nabla^2\phi(\upsilon), \boldsymbol{\gamma}^{\otimes 2}\rangle} \le \mathtt{c}_{\phi}. \tag{4.11}$$

Proof. Let $\langle \nabla^2 \phi(\upsilon), \boldsymbol{u}^{\otimes 2} \rangle \le \varrho^2$. By (4.8) with $\varepsilon = X_1 - E_\upsilon X_1$

$$\nabla^2 \phi(\upsilon + \boldsymbol{u}) = \nabla^2 \log E_\upsilon \mathrm{e}^{\langle \varepsilon, \boldsymbol{u} \rangle} = \frac{E_\upsilon \{ \varepsilon \varepsilon^\top \mathrm{e}^{\langle \varepsilon, \boldsymbol{u} \rangle} \}}{(E_\upsilon \mathrm{e}^{\langle \varepsilon, \boldsymbol{u} \rangle})^2} - \frac{E_\upsilon \{ \varepsilon \, \mathrm{e}^{\langle \varepsilon, \boldsymbol{u} \rangle} \} \, E_\upsilon \{ \varepsilon \, \mathrm{e}^{\langle \varepsilon, \boldsymbol{u} \rangle} \}^\top}{(E_\upsilon \mathrm{e}^{\langle \varepsilon, \boldsymbol{u} \rangle})^2}$$

and by (4.10) and (4.5) in view of $E_\upsilon \mathrm{e}^{\langle \varepsilon, \boldsymbol{u} \rangle} \ge 1$

$$\begin{aligned} \bigl\langle \nabla^2 \phi(\upsilon + \boldsymbol{u}), \gamma^{\otimes 2} \bigr\rangle &\le E_\upsilon \bigl\{ \langle \varepsilon, \gamma \rangle^2 \mathrm{e}^{\langle \varepsilon, \boldsymbol{u} \rangle} \bigr\} \\ &\le E_\upsilon^{1/2} \langle \varepsilon, \gamma \rangle^4 \; E_\upsilon^{1/2} \mathrm{e}^{2 \langle \varepsilon, \boldsymbol{u} \rangle} \le \sqrt{\mathtt{C}_{\Psi,4} \, \mathtt{C}_\varrho} \, \bigl\langle \nabla^2 \phi(\upsilon), \gamma^{\otimes 2} \bigr\rangle \end{aligned}$$

and the assertion follows. □

Lemma 4.3. *Let $\upsilon \in \mathcal{B}_\varrho(\upsilon^*)$ and $\mathtt{r} \le \varrho \sqrt{n}$. Then $f(\upsilon) = \mathbb{E}_{\upsilon^*} L(\upsilon)$ satisfies $(\boldsymbol{\mathcal{S}_3})$ and $(\boldsymbol{\mathcal{S}_4})$ with $h(\upsilon) = \langle \nabla \phi(\upsilon^*), \upsilon \rangle - \phi(\upsilon)$, $\mathbb{m}^2(\upsilon) = \nabla^2 \phi(\upsilon)$, and constants $\mathtt{c}_3$ and $\mathtt{c}_4$ depending on $\mathtt{C}_\varrho$, $\mathtt{C}_{\Psi,3}$, and $\mathtt{C}_{\Psi,4}$ only.*

Proof. Let $\upsilon \in \mathcal{B}_\varrho(\upsilon^*)$. For any $\boldsymbol{u}$ with $\| \mathbb{m}(\upsilon) \boldsymbol{u} \| \le \mathtt{r} / \sqrt{n} \le \varrho$, by (4.9) and (4.11)

$$\frac{| \langle \nabla^3 \phi(\upsilon + t\boldsymbol{u}), \boldsymbol{u}^{\otimes 3} \rangle |}{\| \mathbb{m}(\upsilon) \boldsymbol{u} \|^3} \le \frac{\mathtt{C}_{\Psi,3} \, \| \mathbb{m}(\upsilon + t\boldsymbol{u}) \boldsymbol{u} \|^3}{\| \mathbb{m}(\upsilon) \boldsymbol{u} \|^3} \le \mathtt{C}_{\Psi,3} \, \mathtt{c}_\phi^{3/2},$$

and $(\boldsymbol{\mathcal{S}_3})$ follows with $\mathtt{c}_3 = \mathtt{C}_{\Psi,3} \, \mathtt{c}_\phi^{3/2}$. The proof of $(\boldsymbol{\mathcal{S}_4})$ is similar. □

Check of $(\boldsymbol{\nabla \zeta})$

Now we check the deviation bound for $\nabla \zeta = \boldsymbol{S} - \mathbb{E} \boldsymbol{S}$ under $(\boldsymbol{f})$ and $(\boldsymbol{\Upsilon})$. I.i.d. structure of $\boldsymbol{S} = \sum_i X_i$ and (4.4) yield $\operatorname{Var}(\boldsymbol{S}) = V^2 = n \nabla^2 \phi(\upsilon^*)$. Further, for any $\boldsymbol{u} \in \mathcal{B}_\varrho(\upsilon^*)$, again by the i.i.d. assumption and by (4.8)

$$n^{-1} \log \mathbb{E}_{\upsilon^*} \exp \bigl\{ \langle \nabla \zeta, \boldsymbol{u} \rangle \bigr\} = \log \mathbb{E}_{\upsilon^*} \mathrm{e}^{\langle \varepsilon, \boldsymbol{u} \rangle} = \phi(\upsilon^* + \boldsymbol{u}) - \phi(\upsilon^*) - \langle \nabla \phi(\upsilon^*), \boldsymbol{u} \rangle.$$

Fix $\mathtt{r} \le \varrho n^{1/2}$ and consider all $\boldsymbol{u}$ with $n \langle \nabla^2 \phi(\upsilon^*), \boldsymbol{u}^{\otimes 2} \rangle \le \mathtt{r}^2$. Then by $(\boldsymbol{\mathcal{S}_3})$ and (A.3) of Lemma A.2 of the Supplement (S2023)

$$\phi(\upsilon^* + \boldsymbol{u}) - \phi(\upsilon^*) - \langle \nabla \phi(\upsilon^*), \boldsymbol{u} \rangle \le \frac{1 + \mathtt{c}_3 \, \mathtt{r} \, n^{-1/2}/3}{2} \langle \nabla^2 \phi(\upsilon^*), \boldsymbol{u}^{\otimes 2} \rangle \le \langle \nabla^2 \phi(\upsilon^*), \boldsymbol{u}^{\otimes 2} \rangle$$

provided that $\mathtt{c}_3 \mathtt{r} \le 3 n^{1/2}$. This implies (F.37) with $\mathtt{g} = \varrho \sqrt{n}$ and thus, the deviation bound (F.43) of Theorem F.15 implies $(\boldsymbol{\nabla \zeta})$ for $\varrho \sqrt{n}$ sufficiently large.

4.2 Smoothness constraints, bias-variance trade-off

To handle the bias term, we impose some smoothness conditions on the underlying density parameter $\boldsymbol{\upsilon}^*$; see Section C of the Supplement (S2023). We also limit ourselves to the penalty matrices G^2 which ensure a kind of bias-variance trade-off.

$(G_0\upsilon^*)$ $\|G_0\boldsymbol{\upsilon}^*\|^2 \le 1$ for some fixed G_0^2.

Later we follow the suggestion of Section C and apply the penalizing matrix $G^2 = wG_0^2$. Set $\mathbb{D}^2 = n\nabla^2\phi(\boldsymbol{\upsilon}^*)$, $\mathbb{D}_G^2 = \mathbb{D}^2 + G^2$. The particular value $w = w_*$ can be selected by the bias-variance relation (C.3). First we evaluate the bias term. This is important to ensure that the point $\boldsymbol{\upsilon}_G^*$ is still in the local vicinity $\mathcal{B}_\varrho(\boldsymbol{\upsilon}^*)$.

Proposition 4.4. *Assume* **(f)**, **(ϕ)**, **(Ψ_4)**, **(Υ)**, **$(G_0\upsilon^*)$**. *Let* $G^2 = wG_0^2$, $\mathbb{D}_G^2 = n\nabla^2\phi(\boldsymbol{\upsilon}^*) + G^2$, *and* $\nu^{-1}w^{1/2}n^{-1/2} \le \varrho$. *Then* $\boldsymbol{\upsilon}_G^* \in \mathcal{B}_\varrho(\boldsymbol{\upsilon}^*)$ *and*

$$\|\mathbb{D}_G(\boldsymbol{\upsilon}_G^* - \boldsymbol{\upsilon}^*)\| \le \nu^{-1}w^{1/2}. \tag{4.12}$$

Proof. We intend to apply Proposition 2.4 with $Q = \mathbb{D}$ and $\mathtt{r}_\circ = \nu^{-1}\|\mathbb{D}_G^{-1}G^2\boldsymbol{\upsilon}^*\|$ for $\nu = 2/3$. It holds by **$(G_0\upsilon^*)$** in view of $G^2 \le \mathbb{D}_G^2$ and $G^2 = wG_0^2$

$$\mathtt{r}_\circ = \nu^{-1}\|\mathbb{D}_G^{-1}G^2\boldsymbol{\upsilon}^*\| \le \nu^{-1}w^{1/2}\|G_0\boldsymbol{\upsilon}^*\| = \nu^{-1}w^{1/2}.$$

Further, $\nu^{-1}w^{1/2}n^{-1/2} \le \varrho$ ensures that the set $\{\boldsymbol{\upsilon}\colon \|\mathbb{D}(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^*)\| \le \mathtt{r}_\circ\}$ belongs to the ball $\mathcal{B}_\varrho(\boldsymbol{\upsilon}^*)$. Now Lemma 4.3 yields **$(\mathcal{S}_3)$** for all $\boldsymbol{\upsilon} \in \mathcal{B}_\varrho(\boldsymbol{\upsilon}^*)$. By Lemma A.3, it holds $\omega_G^* \le \mathtt{c}_3\mathtt{r}_\circ \le \mathtt{c}_3\nu^{-1}w^{1/2}n^{-1/2} \le 1/3$ for $n \ge n_0$. Now Proposition 2.4 yields (4.12) and $\boldsymbol{\upsilon}_G^* \in \mathcal{B}_\varrho(\boldsymbol{\upsilon}^*)$. Also by Lemma A.3, the matrix $D_G^2 = D_G^2(\boldsymbol{\upsilon}_G^*)$ satisfies $(1-\omega_G^*)\mathbb{D}_G^2 \le D_G^2 \le (1+\omega_G^*)\mathbb{D}_G^2$. □

For $G^2 = wG_0^2$, the effective dimension $\mathtt{p}_G$ is given by (C.2):

$$\mathtt{p}_G = \mathtt{p}(w) = \operatorname{tr}(\mathbb{D}^2\,\mathbb{D}_G^{-2}) = \operatorname{tr}\bigl\{\mathbb{D}^2(\mathbb{D}^2 + wG_0^2)^{-1}\bigr\}.$$

A particular value w_* is defined by the bias-variance relation reads $\mathtt{p}(w) \asymp w$; see (C.3). Fix some $\mathtt{x}$ and consider

$$\mathtt{r}_G = \mathtt{r}(w) = \sqrt{\mathtt{p}(w)} + \sqrt{2\mathtt{x}}.$$

Now expansion (2.13) of Theorem 2.6 applies provided that the concentration set $\|D_G(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}_G^*)\| \le \nu^{-1}\mathtt{r}_G$ is contained in $\mathcal{B}_\varrho(\boldsymbol{\upsilon}_G^*)$. Proposition 4.4 allows to use $\mathbb{D}_G$ in place of D_G in this condition. In our results, $\mathtt{C}$ stands for a fixed constant depending on the other constants in our conditions like s_0, $\mathtt{C}_{\mathbb{F}}$, $\mathtt{c}_\phi$, ϱ, and $\mathtt{C}_{\Psi,4}$.

Theorem 4.5. *Assume* **(f)**, **(ϕ)**, **(Ψ_4)**, **(Υ)**, **$(G_0\upsilon^*)$**, *and let* $G^2 = wG_0^2$. *If* $\nu^{-1}w^{1/2}n^{-1/2} \le \varrho$ *with* $\nu = 2/3$ *and* $\mathtt{r}(w)n^{-1/2} \le \varrho$, *then on* $\Omega(\mathtt{x})$

$$\|\mathbb{D}_G(\widetilde{\boldsymbol{\upsilon}}_G - \boldsymbol{\upsilon}^*)\| \le 2\nu^{-1}\mathtt{r}(w) + \nu^{-1}w^{1/2}.$$

If $\mathtt{r}(w_*) = \mathtt{C}_0 w_*^{1/2}$ *then it holds on* $\Omega(\mathtt{x})$ *for* $G^2 = w_* G_0^2$ *and* $\mathbb{D}^2_{w_*} = \mathbb{D}^2 + w_* G_0^2$

$$\|\mathbb{D}_{w_*}(\widetilde{\boldsymbol{\upsilon}}_{w_*} - \boldsymbol{\upsilon}^*)\| \leq \mathtt{C}\,\mathtt{C}_0\, w_*^{1/2}\,.$$

Under the same conditions one can specify the results of Section 2 including the Fisher–Wilks expansions of Theorem 2.2. Moreover, for a Gaussian prior $\mathcal{N}(0, G^{-2})$ with $G^2 = wG_0^2$, one can derive the results about concentration and contraction of the posterior $\boldsymbol{\upsilon}_G \,\big|\, \boldsymbol{X}$. Only Laplace's approximation of the posterior in Theorem 3.4 requires a stronger condition on critical dimension: $\mathtt{c}_3\,\nu^{-1}\mathtt{r}(w)\,\mathtt{p}(w) \leq 2n^{1/2}$.

4.3 Rate optimality under Sobolev smoothness

To state standard rate-optimal results and to compare our conclusions with the existing results in the literature, we consider the univariate case $d = 1$ and introduce the condition on Sobolev smoothness of the log-density $\log f(\boldsymbol{x})$.

$$\textbf{(}\boldsymbol{s_0, w_0}\textbf{)} \quad \boldsymbol{\upsilon}^* \in \mathcal{B}(s_0, w_0) = \bigl\{\boldsymbol{\upsilon} = (\upsilon_j)\colon \textstyle\sum_j \upsilon_j^2\, j^{2s_0} \leq w_0\bigr\} \text{ for } s_0 > 0\,.$$

The assumption $\boldsymbol{\upsilon}^* \in \mathcal{B}(s_0, w_0)$ is standard in log-density estimation; cf. Castillo and Nickl (2014) or Spokoiny and Panov (2021) with $s_0 > 1$. Rousseau and Szabo (2017) requires $s_0 > 1/2$ while our results below are valid for $s_0 > 0$. The only exception is the final Gaussian approximation result requiring $s_0 > 1$. We also assume that $w_0 \asymp 1$.

Now we state the results under the smoothness condition $\textbf{(}\boldsymbol{s_0, w_0}\textbf{)}$. The precision/penalization matrix G^2 is taken of the (s, w)-form: $G^2 = \operatorname{diag}(g_1^2, \ldots, g_p^2)$ with $g_j^2 = j^{2s}/w$. One can take any degree $s > 1/2$, $s \geq s_0$, we recommend a large value like $s = 4$ or $s = 5$. Only the factor w should be fixed carefully to get the optimal accuracy of estimation from the relation $(wn)^{1/(2s)} \approx (w_0 n)^{1/(2s_0+1)}$; see (C.12) or (C.13). Under $\textbf{(}\boldsymbol{\nabla^2\phi}\textbf{)}$, the corresponding effective dimension $\mathtt{p}_G$ and the Laplace effective dimension $\mathtt{p}(\boldsymbol{\upsilon})$ are determined by the index m for which $g_m^2 \approx n$. Alternatively one can use a m_0-truncation prior with $m_0 \approx (w_0 n)^{1/(2s_0+1)}$. Such a choice of the prior parameters is frequently used for nonparametric *rate optimal* results about concentration of the pMLE and posterior contraction; cf. Castillo and Nickl (2014), Castillo and Rousseau (2015), Rousseau and Szabo (2017). Note that the mentioned results require at least $s_0 > 1/2$, while our concentration and contraction results apply under $s_0 > 0$. Theorem C.4 and C.5 of the Supplement (S2023) yield the following results.

Theorem 4.6. *Assume* $\textbf{(}\boldsymbol{f}\textbf{)}$, $\textbf{(}\boldsymbol{\Upsilon}\textbf{)}$, $\textbf{(}\boldsymbol{\phi}\textbf{)}$, $\textbf{(}\boldsymbol{\Psi_4}\textbf{)}$, $\textbf{(}\boldsymbol{\nabla^2\phi}\textbf{)}$ *and* $\textbf{(}\boldsymbol{s_0, w_0}\textbf{)}$. *Fix* $s > 1/2$, $s \geq s_0$ *and define* $g_j^2 = w^{-1} j^{2s}$ *with* w *satisfying* $(wn)^{1/(2s)} \approx (w_0 n)^{1/(2s_0+1)}$; *see (C.12) or (C.13). Then on* $\Omega(\mathtt{x})$ *for* $n \geq n_0$,

$$\|\widetilde{\boldsymbol{\upsilon}}_G - \boldsymbol{\upsilon}^*\| \leq \mathtt{C} n^{-s_0/(2s_0+1)},$$

and the posterior measure $\boldsymbol{\upsilon}_G \,\big|\, \boldsymbol{X}$ *satisfies*

$$I\!P\Bigl(\|\boldsymbol{\upsilon}_G - \widetilde{\boldsymbol{\upsilon}}_G\| > \mathtt{C} n^{-s_0/(2s_0+1)} \,\Big|\, \boldsymbol{X}\Bigr) \leq \mathrm{e}^{-\mathtt{x}}.$$

Our main result about Gaussian approximation requires more smoothness of the log-density $\log f(\cdot)$. Namely, we require (s_0, w_0) with $s_0 > 1$.

Theorem 4.7. *Assume the conditions of Theorem 4.6 and* (s_0, w_0) *with* $s_0 > 1$. *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* $(w\,n)^{1/(2s)} = (w_0\,n)^{1/(2s_0+1)} = m_0$, *on* $\Omega(\mathtt{x})$ *for* $n \geq n_0$,

$$\begin{aligned} \sup_{A \in \mathcal{B}(\mathbb{R}^p)} \left| \mathbb{P}\big(\boldsymbol{\upsilon}_G - \widetilde{\boldsymbol{\upsilon}}_G \in A \,\big|\, \boldsymbol{X}\big) - \mathbb{P}'\big(\widetilde{D}_G^{-1}\boldsymbol{\gamma} \in A\big) \right| &\leq \mathtt{C}\, n^{(1-s_0)/(2s_0+1)}, \\ \sup_{A \in \mathcal{B}_s(\mathbb{R}^p)} \left| \mathbb{P}\big(\boldsymbol{\upsilon}_G - \widetilde{\boldsymbol{\upsilon}}_G \in A \,\big|\, \boldsymbol{X}\big) - \mathbb{P}'\big(\widetilde{D}_G^{-1}\boldsymbol{\gamma} \in A\big) \right| &\leq \mathtt{C}\, n^{(2-2s_0)/(2s_0+1)}. \end{aligned} \tag{4.13}$$

The error terms in (4.13) tend to zero as $n \to \infty$ because $s_0 > 1$.

Supplementary Material

Supplement to "Laplace approximation and the use of posterior mean in Bayesian inference" (DOI: 10.1214/23-BA1391SUPP; .pdf).
A. Local smoothness conditions. Introduces smoothness conditions used in the text and their implications.
B. Examples of priors. Typical examples of priors such as truncation and smooth priors.
C. Smooth priors and rate over Sobolev classes. Classical rate results for smooth priors.
D. Non-Gaussian priors. Explains how the results extend to non-Gaussian priors.
E. Dimension free bounds for Laplace approximation. Presents further results and the proofs of the results from Section 3 about Laplace approximation of the posterior.
F. Deviation bounds for quadratic forms. This section collects technical results about Gaussian and non-Gaussian quadratic forms.
G. Gaussian comparison. This section presents an important result about Gaussian comparison from Götze et al. (2019).

References

Castillo, I. and Nickl, R. (2014). On the Bernstein–von Mises phenomenon for nonparametric Bayes procedures. *Annals of Statistics*, 42(5):1941–1969. MR3262473. doi: https://doi.org/10.1214/14-AOS1246. 1327

Castillo, I. and Rousseau, J. (2015). A Bernstein–von Mises theorem for smooth functionals in semiparametric models. *Annals of Statistics*, 43(6):2353–2383. MR3405597. doi: https://doi.org/10.1214/15-AOS1336. 1327

Durmus, A. and Moulines, É. (2019). High-dimensional Bayesian inference via the unadjusted Langevin algorithm. *Bernoulli*, 25(4A):2854–2882. MR4003567. doi: https://doi.org/10.3150/18-BEJ1073. 1305

Frazier, P. I. (2018). A tutorial on Bayesian optimization. https://arxiv.org/1807.02811. 1305

Giordano, M. and Kekkonen, H. (2020). Bernstein–von Mises theorems and uncertainty quantification for linear inverse problems. *SIAM/ASA Journal on Uncertainty Quantification*, 8(1):342–373. MR4069334. doi: https://doi.org/10.1137/18M1226269. 1304

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. MR4003557. doi: https://doi.org/10.3150/18-BEJ1062. 1306, 1321, 1328

Helin, T. and Kretschmann, R. (2022). Non-asymptotic error estimates for the Laplace approximation in Bayesian inverse problems. *Numerische Mathematik*, 150(2). MR4382587. doi: https://doi.org/10.1007/s00211-021-01266-9. 1304

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. MR3477780. doi: https://doi.org/10.1007/s00440-015-0619-7. 1304

Lu, Y. (2017). On the Bernstein-von Mises theorem for high dimensional nonlinear Bayesian inverse problems. https://arxiv.org/1706.00289. 1304

Ma, Y., Chen, Y., Jin, C., Flammarion, N., and Jordan, M. I. (2019). Sampling can be faster than optimization. *Proceedings of the National Academy of Sciences*, 42(116). http://arxiv.org/1811.08413. MR4025861. doi: https://doi.org/10.1073/pnas.1820003116. 1305

Mockus, J. (1989). *Bayesian approach to global optimization. Theory and applications*, volume 37 of *Mathematics and Its Applications. Soviet Series* Dordrecht etc.: Kluwer Academic Publishers. MR1111483. doi: https://doi.org/10.1007/978-94-009-0909-0. 1305

Monard, F., Nickl, R., Paternain, G. P., et al. (2019). Efficient nonparametric Bayesian inference for X-ray transforms. *The Annals of Statistics*, 47(2):1113–1147. MR3909962. doi: https://doi.org/10.1214/18-AOS1708. 1304

Nesterov, Y. and Spokoiny, V. (2017). Random gradient-free minimization of convex functions. *Foundations of Computational Mathematics*, 17(2):527–566. MR3627456. doi: https://doi.org/10.1007/s10208-015-9296-2. 1305

Nickl, R. (2020). Bernstein–von Mises theorems for statistical inverse problems I: Schrödinger equation. *Journal of the European Mathematical Society*, 22:2697–2750. MR4118619. doi: https://doi.org/10.4171/JEMS/975. 1304

Rousseau, J. and Szabo, B. (2017). Asymptotic behaviour of the empirical Bayes posteriors associated to maximum marginal likelihood estimator. *The Annals of Statistics*, 45(2):833–865. MR3650402. doi: https://doi.org/10.1214/16-AOS1469. 1306, 1327

Rousseau, J. and Szabo, B. (2020). Asymptotic frequentist coverage properties of Bayesian credible sets for sieve priors. *The Annals of Statistics*, 48(4):2155–2179. MR4134790. doi: https://doi.org/10.1214/19-AOS1881. 1304

Schillings, C., Sprungk, B., and Wacker, P. (2020). On the convergence of the Laplace approximation and noise-level-robustness of Laplace-based Monte Carlo methods for Bayesian inverse problems. *Numerische Mathematik*, 145:915–971. MR4125981. doi: https://doi.org/10.1007/s00211-020-01131-1. 1304

Spokoiny, V. (2017). Penalized maximum likelihood estimation and effective dimension. *AIHP*, 53(1):389–429. arXiv:1205.0498. MR3606746. doi: https://doi.org/10.1214/15-AIHP720. 1304, 1307

Spokoiny, V. (2019). Bayesian inference for nonlinear inverse problems. https://arxiv.org/1912.12694. 1305, 1307

Spokoiny, V. (2023). "Supplementary Material for "Inexact Laplace Approximation and the Use of Posterior Mean in Bayesian Inference"." *Bayesian Analysis*. doi: https://doi.org/10.1214/23-BA1391SUPP. 1304

Spokoiny, V. and Panov, M. (2021). Accuracy of Gaussian approximation for high-dimensional posterior distributions. *Bernoulli*. in print. arXiv:1910.06028. 1303, 1304, 1305, 1306, 1307, 1316, 1317, 1327

Szabó, B., van der Vaart, A. W., and van Zanten, J. H. (2015). Frequentist coverage of adaptive nonparametric Bayesian credible sets. *The Annals of Statistics*, 43(4):1391–1428. MR3357861. doi: https://doi.org/10.1214/14-AOS1270. 1304

Van der Vaart, A. W. (2000). *Asymptotic Statistics*, volume 3. Cambridge university press. MR1652247. doi: https://doi.org/10.1017/CBO9780511802256. 1303