The Annals of Statistics
2014, Vol. 42, No. 5, 1850–1878
DOI: 10.1214/14-AOS1239

THE BERNSTEIN–VON MISES THEOREM AND NONREGULAR MODELS[1]

BY NATALIA A. BOCHKINA AND PETER J. GREEN

University of Edinburgh and Maxwell Institute, and University of Technology, Sydney and University of Bristol

We study the asymptotic behaviour of the posterior distribution in a broad class of statistical models where the "true" solution occurs on the boundary of the parameter space. We show that in this case Bayesian inference is consistent, and that the posterior distribution has not only Gaussian components as in the case of regular models (the Bernstein–von Mises theorem) but also has Gamma distribution components whose form depends on the behaviour of the prior distribution near the boundary and have a faster rate of convergence. We also demonstrate a remarkable property of Bayesian inference, that for some models, there appears to be no bound on efficiency of estimating the unknown parameter if it is on the boundary of the parameter space. We illustrate the results on a problem from emission tomography.

1. Introduction. The asymptotic behaviour of Bayesian methods has been a long-standing topic of interest, including approximation of the posterior distribution and questions that are important from a frequentist point of view, such as consistency, efficiency and coverage of Bayesian credible regions. For instance, for correctly specified regular finite-dimensional models with n independent observations, these properties are captured by the Bernstein–von Mises theorem that implies that the posterior distribution can be approximated in a $1/\sqrt{n}$ neighbourhood of the true value of the parameter by a Gaussian distribution with variance determined by the Fisher information. More generally, the Bernstein–von Mises theorem holds for dependent observations if the likelihood satisfies local asymptotic normality (LAN) conditions [LeCam (1953), Le Cam and Yang (1990)]. A total variation distance version of the theorem was derived by van der Vaart (1998). This theorem implies that the prior has no asymptotic influence on the posterior, that posterior inference is consistent and efficient in the frequentist sense, and that posterior credible regions are asymptotically the same as frequentist ones.

Received July 2013; revised May 2014.
[1]Both authors acknowledge financial support for research visits provided by the EPSRC-funded SuSTaIn programme at Bristol University.
MSC2010 subject classifications. 62F12, 62F15.
Key words and phrases. Approximate posterior, Bayesian inference, Bernstein–von Mises theorem, boundary, nonregular, posterior concentration, SPECT, tomography, total variation distance, variance estimation in mixed models.

One of the key assumptions of the Bernstein–von Mises (BvM) theorem is that the "true" value of the parameter is an interior point of the parameter space. However, for many problems, including our motivating example of a Poisson inverse problem in tomography, and, more generally for the class of models we consider, this assumption of the BvM theorem does not hold. For the tomography example, the unknown parameter is a vector of tracer concentrations, which are nonnegative and can be zero.

The situation where the unknown parameter can be on the boundary of the parameter support has been addressed in the frequentist literature by studying the asymptotic distribution of the maximum likelihood estimator [Moran (1971), Self and Liang (1987), among others]; however it has been studied very little under the Bayesian approach. Dudley and Haughton (2002) investigated the asymptotic behaviour of the posterior probability of the unknown parameter belonging to a half-space $\mathcal{H}$ for a regular correctly specified model, where they found that if the true value of the parameter belongs to the complement of $\mathcal{H}$, then the posterior probability of half-space $\mathcal{H}$ goes to zero much faster, namely at least at rate n rather than at the standard parametric rate $\sqrt{n}$ (n here is the sample size), and there is an exponential upper bound on this posterior probability. Also, Erkanli (1994) gave a formula for calculating the expectation of a smooth functional of a 3-dimensional posterior distribution where the unknown parameter is on a smooth boundary.

In this paper, we extend the Bernstein–von Mises theorem by relaxing the assumption that the "true" value of the parameter is interior to the parameter space, in a finite-dimensional setting. We consider a broad class of probability distributions for the data and allow the prior distribution to be improper and to have zero or infinite density on the boundary. A key model assumption is that the "true" value of the parameter minimises a generalised Kullback–Leibler distance. There is no assumption of any finite moments. We will show that for these models the consequences of relaxing this assumption are twofold: firstly, the convergence is faster, at least at rate n, if the "true" parameter is on the boundary, and secondly, the limit of the posterior distribution has non-Gaussian components.

We motivate our study by presenting in Section 2 an inverse problem from medical imaging; Section 3 establishes the class of models we study. In Section 4 we state the result on the local behaviour of the posterior distribution in a neighbourhood of the limit that is formulated as a modified Bernstein–von Mises theorem, discuss the assumptions, and give a nonasymptotic version of the result. In Section 5 we illustrate the application of our analogue of the BvM theorem for various examples including the problem of variance estimation in mixed effects models, and discuss the choice of prior distribution. We discuss issues in using the approximation of the posterior distribution in practice and apply it to data from the motivating example in Section 6. We conclude with a discussion. All proofs are deferred to the Appendix.

2. Motivating example.

2.1. *Single photon emission computed tomography.* Single photon emission computed tomography (SPECT) is a medical imaging technique in which a radioactively-labelled tracer, known to concentrate in the tissue to be imaged, is introduced into the subject. Emitted particles are detected in a device called a gamma camera, forming an array of counts. Tomographic reconstruction is the process of inferring the spatial pattern of concentration of the tracer in the tissue from these counts. The Poisson linear model

$$\mathcal{T} Y_i|\theta \sim \text{Poisson}(\mathcal{T} A_i\theta), \qquad i = 1, \dots, n, \text{ independently}, \tag{1}$$

comes close to reality for the SPECT problem (there are some dead-time effects and other artifacts in recording). Here $\theta = \{\theta_j\}$, $j = 1, 2, \dots, p$ represents the spatial distribution of the tracer, typically discretised on a grid, with $\theta_j \geq 0$ for all j, $Y = \{Y_i\}$ the array of the rate of detected photons per time unit, also discretised by the recording process, and $\mathcal{T}$ is the exposure time for photon detection. The $n \times p$ array $A = (A_{ij})$ with rows A_i quantifies the emission, transmission, attenuation, decay, and recording process; A_{ij} is the mean number of photons recorded at i per unit concentration per unit time at pixel/voxel j, and is nonnegative. In some methods of reconstruction, elements of the matrix A are modelled as discretised values of the Radon transform.

Since Poisson distributions form an exponential family, this model can be seen as a generalised linear model [Nelder and Wedderburn (1972)], with identity link function and dispersion $1/\mathcal{T}$; see also Example 1 in Section 3.2.

We formalise the notion of small-noise limit for this Poisson model in a practically-relevant way, by supposing that the exposure time for photon detection becomes large, that is, letting $\mathcal{T} \to \infty$.

The "true image" $\theta^\star$ in emission tomography corresponds to a physical reality, the discretised spatial distribution of concentration of the tracer. Since this is nonnegative, we impose the constraints $\theta \in \Theta = [0, \infty)^p \subset \mathbb{R}^p$.

Unless p is too large, that is, the spatial resolution of θ is too fine, the matrix A is normally of full rank p, and hence the inverse problem is well posed (although it may be ill-conditioned); see Johnstone and Silverman (1990) for eigenvalues of the Radon transform.

See Green (1990) for further detail about this model, and an approach based on EM estimation for MAP reconstruction of θ, in a Bayesian formulation in which spatial smoothness of the solution is promoted by using a pairwise difference Markov random field prior.

2.2. *Prior distribution.* From the beginning of Bayesian image analysis [Besag (1986), Geman and Geman (1984)], use has been made of Markov random fields as prior distributions for image scenes that express generic, qualitative beliefs about smoothness, yet do not rule out abrupt changes for real discontinuities (e.g., at tissue type boundaries in the case of medical imaging).

The prior distribution we consider for the SPECT model is a log cosh pairwise-interaction Markov random field [Green (1990)],

$$p(\theta) \propto \exp\biggl(-\frac{\zeta(1+\zeta)}{2\gamma^2}\sum_{j\sim j'}\log\cosh\biggl(\frac{\theta_j-\theta_{j'}}{\zeta}\biggr)\biggr), \qquad \theta\in\Theta, \tag{2}$$

where $j \sim j'$ stands for j and j' being neighbouring pixels. In this paper the parameters ζ and γ are considered to be fixed.

This model has some attractive properties. While giving less penalty to large abrupt changes in θ compared to the Gaussian, it remains log-concave. It bridges the extremes $\zeta \to \infty$, the Gaussian pairwise-interaction prior, and $\zeta = 0$, the corresponding Laplace pairwise-interaction model, sometimes called the "median prior."

This distribution is improper since it is invariant to perturbing θ by an arbitrary additive constant, but leads to a proper posterior distribution as long as $\sum_j A_{ij} \neq 0$ for some i.

2.3. *Nonstandard features of the SPECT model.* The Bayesian model for SPECT has three nonstandard features: (a) the true image $\theta^\star$ can lie on the boundary of the parameter space $[0,\infty)^p$; (b) if $A_i\theta^\star = 0$ for some i, then the distribution of the corresponding Y_i degenerates to a point mass at 0; (c) the prior distribution is not proper.

In the next section we formulate a model that includes the Bayesian SPECT model as a particular case. The approximate behaviour of the posterior distribution of θ for large $\mathcal{T}$ is investigated in Section 6.

3. Model formulation.

3.1. *Likelihood.* We now list assumptions on the distribution of the observable responses Y, taking values in $\mathcal{Y} \subseteq \mathbb{R}^n$; it has density (with respect to Lebesgue or counting measure) denoted by $p_\sigma(y|\theta)$ for $\theta \in \Theta \subset \mathbb{R}^p$. These assumptions are expressed in terms of the scaled log-likelihood defined by

$$\ell_{y,\sigma}(\theta) = \sigma^2 \log p_\sigma(y|\theta).$$

As we shall see from the assumptions, σ is related to the level of noise, and we are interested in the case where σ is small. We assume that the "true" value of the unknown parameter that generated the data is $\theta^\star \in \Theta$, and denote the true probability measure of Y by $\mathbb{P}_{\theta^\star,\sigma}$. Below, where it does not lead to ambiguity, we will omit the index σ to simplify the notation and will write $\ell_y(\theta)$ and $\mathbb{P}_{\theta^\star}$.

ASSUMPTION M. (1) For $Y \sim p_\sigma(y|\theta^\star)$, there exists a deterministic function $\ell^\star(\theta): \Theta \to \mathbb{R}$ such that for all $\theta \in \Theta$,

$$\forall \varepsilon > 0 \qquad \mathbb{P}_{\theta^\star,\sigma}\bigl(\bigl|\ell_{Y(\omega)}(\theta) - \ell^\star(\theta)\bigr| > \varepsilon\bigr) \to 0 \qquad \text{as } \sigma \to 0.$$

(2) The function $\ell^\star(\theta)$ has a unique maximum over Θ at $\theta = \theta^\star$.

Further assumptions on $\ell_{y,\sigma}(\theta)$ are given in Section 4.1.

Assumption M is satisfied for a wide class of models, in particular for models with independent identically distributed (i.i.d.) observations with $\sigma^2 = 1/n$ and for distributions from the exponential family in canonical form with dispersion $\sigma^2 \to 0$, that are discussed below.

The function $\ell^\star(\theta)$, defined in Assumption M(1), can be viewed as the limit of the negative Kullback–Leibler ($\mathcal{KL}$) distance, rescaled by σ^2, between distributions with densities $p(\cdot|\theta)$ and $p(\cdot|\theta^\star)$, that was used, for instance, in Petrone, Rousseau and Scricciolo (2012) and Barron, Schervish and Wasserman (1999). For i.i.d. models, $\ell^\star(\theta)$ is the negative Kullback–Leibler distance based on a single observation, and for generalised linear models $\ell^\star(\theta)$ is the log-likelihood for "noise-free" data. Assumption M(2) states that this generalised Kullback–Leibler distance is minimised at the "true" value $\theta^\star$, as holds for the usual $\mathcal{K}L$ distance. Assumption M(2) has been used by other authors, for instance, in the context of hidden Markov models by Douc et al. (2011) where it was called the *identifiability assumption*, and a finite sample analogue of this assumption was used in the context of a misspecified model by Spokoiny (2012). This assumption holds for some models where the parameter set Θ is not open and thus where the true value of the parameter $\theta^\star$ can be on the boundary of Θ; see Example 1. These assumptions are satisfied for the tomography model discussed in Section 2 where the unknown tracer image $\theta^\star$ can have zero intensity values in some pixels, as shown in Section 3.2.

Next we show that Assumption M is satisfied for two important classes of models, generalised linear models, and i.i.d. models, including the case when $\theta^\star$ is on the boundary of Θ.

3.2. *Generalised linear models.* In the generalised linear models of Nelder and Wedderburn (1972), an important class of nonlinear statistical regression problems, responses y_i, $i = 1, 2, \dots, n$ are drawn independently from a one-parameter exponential family of distributions in canonical form, with density or probability function

$$p_\sigma(y|\eta) = \exp\left(\sum_{i=1}^{n}\left[\frac{y_i b(\eta_i) - c(\eta_i)}{\sigma^2} + d(y_i, \sigma)\right]\right),$$

using the mean parameterisation, for appropriate functions b, c, and d characterising the particular distribution family. The parameter σ^2 is a common dispersion parameter shared by all responses. Assuming that functions $b(\cdot)$ and $c(\cdot)$ are twice differentiable, the expectation of this distribution is $\mathbb{E}(Y_i) = \eta_i = c'(\eta_i)/b'(\eta_i)$, and the variance is $\mathrm{Var}(Y_i) = \sigma^2[c''(\eta_i)b'(\eta_i) - c'(\eta_i)b''(\eta_i)]/[b'(\eta_i)]^3$. This implies that the random variable Y converges in probability to a finite deterministic limit $y^\star = \mathbb{E}Y$ as $\sigma \to 0$ and that the dispersion σ^2 is related to the noise level of the observations.

Firstly consider the case $\theta = \eta$. Then, $\ell_Y(\theta)$ is linear in Y, and hence it converges to $\ell_{y^\star}(\theta)$ in probability as $\sigma \to 0$. Therefore, Assumption M(1) is satisfied with $\ell^\star(\theta) = \ell_{y^\star}(\theta)$. If $\nabla \ell^\star(\theta^\star) = 0$ and the Hessian, which is diagonal, has negative entries, then $\theta^\star$ uniquely maximises $\ell^\star(\theta)$; that is, Assumption M(2) is satisfied. If $\theta^\star$ is on the boundary and the gradient is nonzero, see Examples 1 and 2 below.

Now consider a generalised linear model with $\eta = A\theta$ and matrix A such that $A^T A$ is of full rank, that is, such that the likelihood is identifiable with respect to parameter θ. In this case, Assumption M holds with $\theta^\star = (A^T A)^{-1} A^T y^\star$. The tomography example given in Section 2 belongs to this class of models, with $\sigma^2 = \mathcal{T}^{-1}$, $b(\eta_i) = \log \eta_i$, $c(\eta_i) = \eta_i$, and $\Theta = [0, \infty)^p$.

Now we show that Assumption M(2) is satisfied when $\theta^\star$ is on the boundary of Θ for some distributions from the exponential family.

EXAMPLE 1. Consider the Poisson distribution $Y/\sigma^2 \sim \mathrm{Poisson}(\eta/\sigma^2)$ with $\eta \geq 0$. The scaled log-likelihood for η is $\ell_y(\eta) = y \log \eta - \eta$. If Y is generated with $\eta = 0$, then we observe $y = 0$ with probability 1, so in this case the scaled log-likelihood for η is always $-\eta$, which is maximised over $\eta \geq 0$ at $\eta = 0$, that is, the true value of η.

EXAMPLE 2. For the Binomial distribution $Y \sim \mathrm{Bin}(n, \eta)$, the scaled log-likelihood for $\eta \in [0, 1]$ is $\ell_y(\eta) = [y \log(\eta) + (n - y)\log(1 - \eta)]/n$. If the true value of η is 1, then $\mathbb{P}(Y = n) = 1$ and the scaled log-likelihood for η is $\ell_y(\eta) = \log(\eta)$, which is maximised over $[0, 1]$ at $\eta = 1$, so that again we recover the true value, and Assumption M(2) is satisfied for this model.

The same holds for the other boundary point $\eta = 0$, and also for multinomial and negative binomial distributions.

3.3. *I.I.D. models.* Let $Y_1, \dots, Y_n$ be independent identically distributed random variables where the density or probability mass function of Y_i is $p(y_i|\theta) = C_{y_i} \exp\{\ell_{y_i}(\theta)\}$, with unknown parameter $\theta \in \Theta \subset \mathbb{R}^p$ where p is finite and independent of n. Here, $\sigma^2 = 1/n$ and $\ell_y(\theta) = n^{-1} \sum_{i=1}^n \ell_{y_i}(\theta)$. In this case, $\ell_{Y_i}(\theta)$ are i.i.d. random variables, so, as $n \to \infty$, Assumption M(1) is satisfied under the conditions of the weak law of large numbers for the random variable $\ell_{Y_i}(\theta)$, for all θ, which implies that there exists $\ell^\star(\theta)$ such that $\ell_Y(\theta)$ converges in probability to $\ell^\star(\theta)$ as $n \to \infty$. If $\mathbb{E}[\ell_Y(\theta)]$ exists for all $\theta \in \Theta$, then $\ell^\star(\theta) = \mathbb{E}[\ell_{Y_i}(\theta)]$, equal to the negative Kullback–Leibler distance between the distributions with densities

$p(\cdot|\theta)$ and $p(\cdot|\theta^\star)$, and then Assumption M(2) holds. For instance, it is easy to check that Assumption M is satisfied for i.i.d. Cauchy random variables Y_i with $\ell_{Y_i}(\theta) = \log(1 + (Y_i - \theta)^2)$ and $\theta \in \Theta \subseteq \mathbb{R}$.

3.4. *Bayesian formulation.* We adopt a Bayesian paradigm, using a σ-finite prior measure $\pi(d\theta)$ on Θ. Thus the posterior distribution satisfies

$$\pi(d\theta|y) \propto \exp\bigl(\ell_y(\theta)/\sigma^2\bigr)\pi(d\theta), \qquad \theta \in \Theta. \tag{3}$$

Here we do not assume that the prior distribution is proper, nor do we assume that its density is bounded away from 0 and infinity on the boundary of Θ; see Assumption P in Section 4.1.

4. The analogue of the Bernstein–von Mises theorem.

4.1. *Notation and assumptions.* We shall use the default norms $\|z\| = \|z\|_2$ for both vectors and matrices. If the appropriate derivatives exists, define the gradient $\nabla f(\theta)$ of a function f on Θ as a vector of partial derivatives (one-sided if θ is on the boundary of Θ), and $\nabla^2 f(\theta)$ is a matrix of second derivatives of f (again, one-sided if θ is on the boundary of Θ). We use notation θ_S to define the vector $(\theta_j, j \in S)$ for $S \subset \{1, \dots, p\}$, a convention which also applies to the gradient ∇, that is, $\nabla_S f(\theta) = (\nabla_j f(\theta), j \in S)$. We denote a submatrix Σ indexed by subsets S, J by $\Sigma_{S,J} = (\Sigma_{ij}, i \in S, j \in J)$; this also applies to the matrix of second derivatives, so we can write $\nabla^2_{S,J} f(\theta)$ to denote the corresponding submatrix.

We use $A\mathcal{X} + x_0 = \{Ax + x_0, x \in \mathcal{X}\}$ to denote the image of an affine transformation of the set $\mathcal{X}$ given matrix A and vector x_0.

The limit of the posterior distribution has a different character in different directions, and we need to partition the index set $\{1, 2, \dots, p\}$ of θ accordingly. Let

$$S_0 = \bigl\{j : \nabla_j \ell^\star(\theta^\star) = 0\bigr\} \quad \text{and} \quad S_1 = \bigl\{j : \nabla_j \ell^\star(\theta^\star) \neq 0\bigr\},$$

with dimensions p_0 and $p_1 = p - p_0$, respectively. We partition S_0 further:

$$S_0^\star = \bigl\{j : \nabla_j \ell^\star(\theta^\star) = 0 \text{ and } \theta_j^\star = 0\bigr\} \quad \text{and}$$
$$S_0 \setminus S_0^\star = \bigl\{j : \nabla_j \ell^\star(\theta^\star) = 0 \text{ and } \theta_j^\star \neq 0\bigr\},$$

with dimensions $p_0^\star$ and $p_0 - p_0^\star$ where $\theta_j^\star = 0$ corresponds to $\theta^\star$ being on the boundary of Θ; see Assumption B(1) below.

We then introduce a permutation of coordinates of θ, defined by any matrix U that maps $S_0 \setminus S_0^\star$ to the first $(p_0 - p_0^\star)$ coordinates, $S_0^\star$ to the next $p_0^\star$, and S_1 to the last p_1. The first p_0 rows of U will be denoted U_0 and the remainder U_1. We denote the index set $\{p_0 - p_0^\star + 1, \dots, p_0\}$ by $T_0^\star$ which is the image of $S_0^\star$ under the map defined by U. Note that $\theta_j^\star = 0$ for all $j \in S_0^\star \cup S_1$ (for $j \in S_1$, this is given by Lemma 1 below), so this set describes the coordinates of $\theta^\star$ that lie on the boundary; in the case of $S_0^\star$ the gradient is also zero in this direction.

We introduce the notation $\mathcal{PTN}_{p_0}(a_0, \Omega_{00}^{-1}, p_0^\star, \alpha_0)$ for a polynomially-tilted multivariate Gaussian distribution truncated to $\mathcal{V}_0 = \mathbb{R}^{p_0 - p_0^\star} \times \mathbb{R}_+^{p_0^\star}$, for which the corresponding measure of any measurable $\mathcal{B} \subset \mathcal{V}_0$ is defined by

$$\mathcal{PTN}_{p_0}(\mathcal{B}; a_0, \Omega_{00}^{-1}, p_0^\star, \alpha_0) = \frac{\int_{\mathcal{B}} \prod_{j \in T_0^\star} x_j^{\alpha_{0,j}-1} e^{-(x-a_0)^T \Omega_{00}(x-a_0)/2}\, dx}{\int_{\mathcal{V}_0} \prod_{j \in T_0^\star} x_j^{\alpha_{0,j}-1} e^{-(x-a_0)^T \Omega_{00}(x-a_0)/2}\, dx}, \tag{4}$$

where $a_0 \in \mathbb{R}^{p_0}$, Ω_{00} is a $p_0 \times p_0$ positive definite matrix, and $\alpha_0 = (\alpha_{0,j})_{j \in T_0^\star} \in (0, \infty)^{p_0^\star}$. α_0 could also be interpreted as a p_0-dimensional vector whose first $p_0 - p_0^\star$ coordinates are irrelevant. Note that this distribution is Gaussian if $p_0^\star = 0$, and truncated Gaussian if $p_0^\star \neq 0$ and $\alpha_{0,j} = 1$ for all j.

For $\alpha, a > 0$, $\Gamma(\alpha, a)$ denotes the Gamma distribution with density $p(x) = a^\alpha x^{\alpha-1} e^{-ax}/\Gamma(\alpha)$, $x > 0$, and $\Gamma(dx; \alpha, a)$ the corresponding probability measure.

In addition to Assumption M (Section 3.1), we make the following assumptions. They make use of the following neighbourhoods of $\theta^\star$:

$$\Theta^\star(\delta) = \{\theta \in \Theta : U(\theta - \theta^\star) \in B_{2,p_0}(0, \delta_0) \times B_{\infty,p_1}(0, \delta_1)\}, \tag{5}$$

where $\delta = (\delta_0, \delta_1)$, $\delta_0, \delta_1 > 0$ and $B_{q,s}(z_0, r) = \{z \in \mathbb{R}^s : \|z - z_0\|_q < r\}$.

ASSUMPTION B (On boundary of Θ, $\partial\Theta$). (1) $\Theta \subseteq [0, \infty)^p$ and $\Theta \cap \partial\Theta \subseteq \bigcup_{j=1}^p \{\theta \in \Theta : \theta_j = 0\}$.

(2) $U(\Theta - \theta^\star) \supseteq (-c_0, c_0)^{p_0 - p_0^\star} \times [0, c_0)^{p_0^\star} \times [0, c_1)^{p_1}$ for some $c_0, c_1 > 0$.

ASSUMPTION S (Smoothness in θ). There exist $\delta_0, \delta_1 > 0$ depending on σ such that:

(1) $\delta_0 \to 0$, $\delta_1 \to 0$, $\delta_0/\sigma \to \infty$, $\delta_1/\sigma^2 \to \infty$ as $\sigma \to 0$.

(2) For all $\theta \in \Theta^\star(\delta)$, $\nabla\ell^\star(\theta)$, $\nabla\ell_Y(\theta)$ and $\nabla^2_{S_0,S_0}\ell_Y(\theta)$ exist $\mathbb{P}_{\theta^\star,\sigma}$-almost everywhere, for small enough σ.

(3) For any $\varepsilon > 0$,

$$\mathbb{P}_{\theta^\star,\sigma}\Big(\sup_{\theta \in \Theta^\star(\delta)} \|\nabla\ell_{Y(\omega)}(\theta) - \nabla\ell^\star(\theta^\star)\|_\infty > \varepsilon\Big) \to 0 \qquad \text{as } \sigma \to 0.$$

(4) $\mathbb{P}_{\theta^\star,\sigma}(\|\sigma^{-1}\nabla_{S_0}\ell_{Y(\omega)}(\theta^\star)\| < \infty) = 1$ for small enough σ.

(5) There exists a $p_0 \times p_0$ positive definite matrix Ω_{00} such that

$$\forall\varepsilon > 0 \qquad \mathbb{P}_{\theta^\star,\sigma}\Big(\sup_{\theta \in \Theta^\star(\delta)} \|\nabla^2_{S_0,S_0}\ell_{Y(\omega)}(\theta) + \Omega_{00}\| > \varepsilon\Big) \to 0 \qquad \text{as } \sigma \to 0.$$

ASSUMPTION P (On the prior distribution). The σ-finite measure $\pi(d\theta)$ on Θ satisfies the following conditions:

(1) $\int_\Theta e^{\ell_y(\theta)/\sigma^2}\pi(d\theta) < \infty$ for $\mathbb{P}_{\theta^\star,\sigma}$-almost all $y \in \mathcal{Y}$, for small enough σ.

(2) For $\theta \in \Theta^\star(\delta)$, there exists $p(\theta) \ge 0$ such that $\pi(d\theta) = p(\theta)\,d\theta$.

(3) There exist $C_\pi > 0$ and $\boldsymbol{\alpha}_j > 0$ for $j \in S_1 \cup S_0^\star$, independent of σ, and there exists $\Delta_\pi = \Delta_\pi(\delta) \ge 0$, such that $\Delta_\pi \to 0$ as $\sigma \to 0$ and for $\theta \in \Theta^\star(\delta)$,

$$C_\pi(1-\Delta_\pi) \le p(\theta) \times \prod_{j \in S_1 \cup S_0^\star} \theta_j^{-(\boldsymbol{\alpha}_j - 1)} \le C_\pi(1+\Delta_\pi).$$

Denote $\alpha_0 = \boldsymbol{\alpha}_{S_0^\star}$, $\alpha_1 = \boldsymbol{\alpha}_{S_1}$.

ASSUMPTION L. Assume $\mathbb{P}_{\theta^\star,\sigma}(\Delta_0(\delta) \to 0) \to 1$ as $\sigma \to 0$, where

$$\Delta_0(\delta) = \sigma^{-p_0 - \sum_{j \in T_0^\star}(\alpha_{0,j}-1) - 2\sum_{j=1}^{p_1}\alpha_{1,j}} \int_{\Theta \setminus \Theta^\star(\delta)} e^{(\ell_Y(\theta) - \ell_Y(\theta^\star))/\sigma^2}\pi(d\theta). \tag{6}$$

Assumption L implies consistency of the posterior distribution at a certain rate, and it can be written as $\pi(\Theta^\star(\delta)|Y) = 1 + O_{\mathbb{P}_{\theta^\star,\sigma}}(1)$ as $\sigma \to 0$. Consistency of the posterior is a necessary assumption for the Bernstein–von Mises theorem [van der Vaart (1998), Theorem 10.1]. Under Assumption M, Assumption L holds if the following condition is satisfied:

$$\sigma^{-p_0 - \sum_{j \in T_0^\star}(\alpha_{0,j}-1) - 2\sum_{j=1}^{p_1}\alpha_{1,j}} \int_{\Theta \setminus \Theta^\star(\delta)} e^{-h(\theta)/\sigma^2}\pi(d\theta) \to 0 \qquad \text{as } \sigma \to 0, \tag{7}$$

where the function $h(\theta)$ is such that

$$\ell^\star(\theta) - \ell^\star(\theta^\star) \le -h(\theta) \qquad \text{for all } \theta \in \Theta \setminus \Theta^\star(\delta).$$

Under Assumption B, the complement of the polar cone of the set $\Theta - \theta^\star$ coincides with $\Theta - \theta^\star$ in a small enough neighbourhood of 0; this is essential for the analytic arguments of the paper. This property holds for other polyhedral boundaries; for affine transformations of the positive orthant this is trivial, while in general it relies on the fact that $\sigma \to 0$. For a set Θ that does not satisfy these conditions, the support of the posterior distribution in the limit may depend on the complement of the polar cone of $\Theta - \theta^\star$; see also Shapiro (2000).

In Assumption S, we assume uniform convergence in probability of the derivatives of the scaled log-likelihood at $\theta^\star$ as σ tends to 0, and that the score function of θ_{S_0} converges to 0 at rate σ^{-1}.

In Assumption P, we assume that the posterior distribution is proper but we do not assume that the prior measure itself is proper. Neither do we assume that $p(\theta)$ is finite and bounded away from 0 on the boundary of the parameter space, that is, that $\boldsymbol{\alpha}_j = 1$ for all j, which is the assumption of the BvM theorem. In particular, the log cosh Markov random field prior distribution that was discussed in Section 2 for the motivating example, satisfies these conditions with $\boldsymbol{\alpha}_j = 1$ for all $j \in S_1 \cup S_0^\star$. Other improper priors such as the Jeffreys prior for a Poisson likelihood, as well as the conjugate Gamma prior and Beta prior conjugate to a binomial likelihood, satisfy this assumption; see examples in Section 5.

4.2. *The main result*. Before presenting the main result, we state two preliminary lemmas. Firstly, we show that the elements $\theta^\star_{S_1}$ are on the boundary of Θ, and secondly, we study properties of the derivatives of $\ell^\star(\theta)$.

LEMMA 1. *If Assumption M in Section 3.1 and Assumption B in Section 4.1 hold, then $\theta^\star_{S_1}=0$ and vector $\nabla_{S_1}\ell^\star(\theta^\star)$ has negative coordinates.*

If also for any $\varepsilon>0$, $\mathbb{P}_{\theta^\star,\sigma}(\|\nabla^2_{S_0,S_0}\ell_Y(\theta^\star)-\nabla^2_{S_0,S_0}\ell^\star(\theta^\star)\|>\varepsilon)\to 0$ as $\sigma\to 0$, then the matrix $\Omega_{00}=-\nabla^2_{S_0,S_0}\ell^\star(\theta^\star)$ is positive semi-definite.

This lemma follows from standard optimality conditions [e.g., Proposition 2.1.2 in Bertsekas (2003)].

Define the following scaling transform $\mathcal{S}=\mathcal{S}_\sigma:\Theta-\theta^\star\to\mathbb{R}^{p_0}\times\mathbb{R}^{p_1}_+$:

$$\mathcal{S}(\theta-\theta^\star)=D_\sigma^{-1}U(\theta-\theta^\star), \tag{8}$$

where $D_\sigma=\operatorname{diag}(\sigma I_{p_0},\sigma^2 I_{p_1})$ and $U=(U_0^T:U_1^T)^T$ is defined in Section 4.1. The two subsets of coordinates are scaled differently; we are considering $(\theta_{S_0}-\theta^\star_{S_0})/\sigma$ and $(\theta_{S_1}-\theta^\star_{S_1})/\sigma^2$. In the next lemma we study the image of $\Theta^\star(\delta)$ defined by (5) under this transformation, in the limit.

LEMMA 2. *Let Assumption B in Section 4.1 hold, and take δ_0 and δ_1 such that $\delta_0\le c_0$, $\delta_1\le c_1$, $\delta_0/\sigma\to\infty$, and $\delta_1/\sigma^2\to\infty$ as $\sigma\to 0$. Then,*

$$\liminf_{\sigma\to 0}\mathcal{S}_\sigma\big(\Theta^\star(\delta)-\theta^\star\big)=\mathbb{R}^{p_0-p_0^\star}\times\mathbb{R}_+^{p_0^\star+p_1},$$

the lim inf *being in the sense of Shapiro (2000).*

The proof of the lemma is given in Appendix A.2.

The limit of the posterior distribution is described by the following parameters: $\alpha_0=\boldsymbol{\alpha}_{S_0^\star}$ and $\alpha_1=\boldsymbol{\alpha}_{S_1}$ defined in Assumption P, Ω_{00} defined in Assumption S, and $a_0(\omega)$ and a_1 defined by

$$a_0(\omega)=\sigma^{-1}\Omega_{00}^{-1}\nabla_{S_0}\ell_{Y(\omega)}(\theta^\star),\qquad a_1=-\nabla_{S_1}\ell^\star(\theta^\star). \tag{9}$$

The vector a_1 has positive coordinates, which follows from Lemma 1. The matrix $\sigma^{-2}\Omega_{00}$ is an analogue of the Fisher information for θ_{S_0}.

In the theorem below, which is an analogue of the Bernstein–von Mises theorem, we claim that under the stated assumptions, the posterior distribution of $\mathcal{S}(\theta-\theta^\star)$, $\mathbb{P}_{\mathcal{S}(\theta-\theta^\star)|Y}$, converges to a finite limit.

THEOREM 1. *Consider the Bayesian model defined in Section 3 under Assumption M and such that Assumptions B, S, P and L hold.*

Define a random probability measure on $\mathcal{V}_0\times\mathbb{R}^{p_1}_+$, with $v=(v_0,v_1)$:

$$\mu^\star(\omega)(dv)=\mathcal{PTN}_{p_0}\big(dv_0;a_0(\omega),\Omega_{00}^{-1},p_0^\star,\alpha_0\big)\times\Gamma_{p_1}(dv_1;\alpha_1,a_1),$$

where $\mathcal{V}_0 = \mathbb{R}^{p_0 - p_0^\star} \times \mathbb{R}_+^{p_0^\star}$, $\mathcal{PTN}_{p_0}(dv_0; a_0, \Omega_{00}^{-1}, p_0^\star, \alpha_0)$ *is the polynomially-tilted truncated Gaussian distribution defined by* (4), *and* $\Gamma_{p_1}(\cdot; \alpha_1, a_1)$ *is the probability measure of a* p_1*-dimensional vector* ξ *with independent coordinates* $\xi_i \sim \Gamma(\alpha_{1,i}, a_{1,i})$.

Then, with transform $\mathcal{S}$ *defined by* (8), *as* $\sigma \to 0$,

$$\forall \varepsilon > 0 \qquad \mathbb{P}_{\theta^\star,\sigma}\big(\|\mathbb{P}_{\mathcal{S}(\theta-\theta^\star)|Y} - \mu^\star\|_{\mathrm{TV}} > \varepsilon\big) \to 0.$$

The proof is given in Appendix A.1. If $\theta^\star$ is an interior point, then $p_1 = p_0^\star = 0$, the additional factor in the definition of $\mu^\star$ disappears, and the limit is Gaussian, as in the classical Bernstein–von Mises theorem.

Assumptions M and S imply that the log-likelihood can be approximated quadratically with respect to the parameter θ_{S_0} (which includes $\theta_{S_0^\star}$ where the "true" parameter is on the boundary of the parameter space) but not with respect to θ_{S_1}. This is related to the LAN property [Le Cam and Yang (1990)]. In particular, the rate of convergence for $\theta_{S_0^\star}$ is still σ^{-1}, and the limit of the rescaled posterior is truncated Gaussian, possibly modified by the behaviour of the local prior density on the boundary, whereas for θ_{S_1} the rate of convergence is faster (σ^{-2} instead of σ^{-1}), θ_{S_1} is asymptotically independent of θ_{S_0} given data, and its limiting distribution is Gamma. See examples in Section 5.

We shall see in Section 5 that in a number of models parameter components on the boundary can only be either all regular or all nonregular. However, in the motivating SPECT example, both types of boundary behaviour can occur. Hence the chosen prior, satisfying Assumption P with $\boldsymbol{\alpha}_j = 1$ for all $j \in S_1 \cup S_0^\star$, results in asymptotically efficient inference for the regular parameters.

REMARK 1. The key property of the posterior distribution, when the true parameter is on the boundary, is that the gradient of the log-likelihood at this point does not vanish asymptotically. Thus in some directions the leading term at the Taylor expansion of log posterior density is linear rather than quadratic, as would be the case when $\theta^\star$ is an interior point. If the local prior density at $\theta^\star$ is bounded away from 0 and infinity, then the limit of the posterior in these directions is an exponential distribution; if the local prior density has an additional polynomial term in a neighbourhood of $\theta_j^\star = 0$, then the limit is a Gamma distribution.

If the prior density behaves like a positive constant on the boundary or the "regular" part of the parameter is not on the boundary, then the limiting distribution $\mu^\star(\omega)$ has a simple form.

COROLLARY 1. *Assume that Assumption* P *is satisfied with* $\alpha_{0,j} = 1$ *for* $j \in T_0^\star$, *or the set* $T_0^\star$ *is empty* (*i.e.*, $p_0^\star = 0$). *Then, under the conditions of Theorem* 1, *the limiting probability measure* $\mu^\star(\omega)$ *on* $\mathcal{V}_0 \times \mathbb{R}_+^{p_1}$ *is defined by*

$$\mu^\star(\omega)(dv) = \mathcal{TN}_{p_0}\big(dv_0; a_0(\omega), \Omega_{00}^{-1}\big) \times \bigotimes_{i=1}^{p_1} \Gamma(dv_{1,i}; \alpha_{1,i}, a_{1,i}),$$

where $\mathcal{TN}_{p_0}(dv_0; a_0(\omega), \Omega_{00}^{-1})$ *denotes the Gaussian distribution truncated to* $\mathcal{V}_0$ *and normalised to be a probability measure.*

In particular, if the prior distribution behaves as a constant in a neighbourhood of $\theta^\star$ $(\alpha_{1,j} = 1$ *for all* $j)$, *then the limit of* θ_{S_1}/σ^2 *is multivariate exponential.*

4.3. *Efficiency of inference for "nonregular" parameters.* We can see that for $\theta_{S_0 \setminus S_0^\star}$ the standard Bernstein–von Mises theorem holds under the assumption that the prior density in the neighbourhood of $\theta_{S_0 \setminus S_0^\star}$ is bounded away from 0 and infinity, a standard assumption of the BvM theorem. Thus inference for $\theta_{S_0 \setminus S_0^\star}$ is asymptotically independent of the prior and is asymptotically equivalent to efficient frequentist inference.

However, inference for θ_{S_1} is different. The first key difference is that there is no need to require a similar assumption on the prior distribution: even if the local prior density tends to infinity or zero (both at a polynomial rate) on the boundary, for i.i.d. observations with $\sigma^2 = 1/n$, Bayesian inference is still consistent, at a rate faster than the parametric $\sqrt{n}$ rate. The second difference is that the limit of the rescaled and recentred posterior distribution for $\theta^\star_{S_1}$ is not random (i.e., does not depend on ω). These two properties lead to the third important difference which is the formulation of efficiency of the estimation procedure for these "nonregular" parameters. This point is elaborated below.

Consider the case where $p = 1$ and $\theta^\star$ is on the boundary (i.e., $\theta^\star = 0$) with $\nabla \ell^\star(\theta^\star) \neq 0$. If the prior density at $\theta^\star$ is not bounded away from 0 and infinity, the limit of the posterior distribution depends on the behaviour of the prior distribution on the boundary via exponent α ($\boldsymbol{\alpha}_j$ with $j = 1$). This exponent is a construct of the statistician and does not depend on the data or its model and can be chosen freely. If $\alpha > 1$, then the prior density at the true value $\theta^\star$ is 0, and if $\alpha < 1$, the local prior density of θ tends to infinity as $\theta \to \theta^\star$. The length of the asymptotic posterior credible interval for θ decreases to 0 as $\alpha \to 0$ (see Examples 3 and 4 in Section 5); hence it is possible to recover the true value on the boundary as precisely as desired, up to the error of approximation of the posterior distribution by its limit (an upper bound on that is presented in Proposition 1). Note that for the Poisson and Binomial distributions discussed in Examples 3 and 4, the Jeffreys prior satisfies Assumption P with $\alpha = 1/2$. This property raises questions about the formulation of efficiency in this case, as, from a theoretical perspective, there appears to be no lower bound on the length of the credible interval as in the regular case.

4.4. *Nonasymptotic upper bound.* We also state a nonasymptotic bound on the distance between the posterior distribution of the rescaled parameter and its limit.

PROPOSITION 1. *Assume that the following conditions hold for* δ_0 *and* δ_1 *and for some* $\delta_{*0} > 0$, $\delta_{*1} > 0$ *that may depend on* δ_0 *and* δ_1:

$$\begin{aligned} &\delta_{*1} < a_{\min}, \qquad \delta_{*0} < \lambda_{\min}(\Omega_{00}), \qquad \delta_0 < \|\theta^\star_{S_0}\|, \\ &\delta_0 \leq c_0, \qquad \delta_1 \leq c_1, \end{aligned} \tag{10}$$

where $a_{\min} = \min_j a_{1,j}$, $\lambda_{\min}(\Omega_{00})$ *is the smallest eigenvalue of* Ω_{00}, *and* c_0, c_1 *are constants from Assumption* B. *Let the assumptions of Theorem* 1 *hold, and define the following events*:

$$\mathcal{A}_0 = \Big\{\omega\colon \sup_{\theta\in\Theta^\star(\delta)} \|\nabla^2_{S_0,S_0}\ell_{Y(\omega)}(\theta) + \Omega_{00}\| \le \delta_{*0}\Big\},$$
$$\mathcal{A}_1 = \Big\{\omega\colon \sup_{\theta\in\Theta^\star(\delta)} \|\nabla_{S_1}\ell_{Y(\omega)}(\theta) + a_1\|_\infty \le \delta_{*1}\Big\}. \tag{11}$$

Then, *on* $\mathcal{A} = \mathcal{A}_0 \cap \mathcal{A}_1 \cap \{\|a_0(\omega)\| < \delta_0/\sigma\}$,

$$\begin{aligned}
&\|\mathbb{P}_{\mathcal{S}(\theta-\theta^\star)|Y} - \mu^\star\|_{\mathrm{TV}} \\
&\quad\le 2\max\Big\{C_1\delta_{*1},\, p_1 \max_j \Gamma\Big(\Big(\frac{a_{1,j}\delta_1}{\sigma^2}, \infty\Big); \alpha_{1,j}, 1\Big)\Big\} \\
&\qquad + 2\max\Big\{C_0\delta_{*0},\, C_{\alpha_0}\Gamma\Big(\Big(\frac{\lambda_{\min}(\Omega_{00})}{2}\Big[\frac{\delta_0}{\sigma} - \|a_0(\omega)\|\Big]^2, \infty\Big); \frac{p_{\alpha 0}}{2}, 1\Big)\Big\} \\
&\qquad + C_2\Delta_\pi + C_\Delta\Delta_0(\delta),
\end{aligned} \tag{12}$$

where $p_{\alpha 0} = p_0 + \sum_{j\in T_0^\star}(\alpha_{0,j} - 1)$ *and the constants are defined in the proof. If also* $\delta_{*0} \to 0$ *and* $\delta_{*1} \to 0$ *as* $\sigma \to 0$, *then the upper bound in* (12) *tends to* 0.

The proof is given in Appendix A.1. Note that under the assumptions of Theorem 1, $\mathbb{P}_{\theta^\star,\sigma}(\mathcal{A}) \to 1$ as $\delta_{*0} \to 0$ and $\delta_{*1} \to 0$. For the upper bound of the total variation to be small in practical applications, the dimensions p_k should not be too large compared to the corresponding rate, the smallest eigenvalue of the precision matrix Ω_{00} cannot be too small, that is, that $\lambda_{\min}(\Omega_{00})\delta_0^2/\sigma^2$ should be large, and that the combination of parameters $(\alpha_{1,j}, a_{1,j})$ should be such that value δ_1/σ^2 is far in the tail of all corresponding Gamma distributions. If $\alpha_{1,j} = 1$ for all j, this requires that the smallest value $a_{\min}$ of the parameter a_1 should not be too small, that is, $a_{\min}\delta_1/\sigma^2$ should be large.

It is interesting to note that, for each $k = 0, 1$, if $\delta_{*k} \asymp \delta_k$, which holds in many cases, the value of δ_k minimising the local upper bound (the first two lines of the upper bound) coincides with the upper bound of the Ky Fan distance between the posterior distribution of θ_{S_k} and its limit, a point mass at $\theta^\star_{S_k}$. These are $\delta_0 = C_{\Omega_{00}}\sigma\sqrt{\log(1/\sigma)}$ and $\delta_1 = C_{a_1}\sigma^2\log(1/\sigma)$ [Bochkina (2013)].

5. Examples. We now give examples where the asymptotic posterior distribution differs from Gaussian. We start with a rule to verify Assumption L which applies to exponential family distributions that we consider below.

LEMMA 3. *Take* $\delta_0, \delta_1 > 0$ *such that* $\delta_0, \delta_1 \to 0$, *and assume that for any* $\theta \in \Theta \setminus \Theta^\star(\delta)$,

$$\ell_Y(\theta) - \ell_Y(\theta^\star) \le -C_{\delta 0}\sum_{j\in S_0}|\theta_j - \theta_j^\star| - C_{\delta 1}\sum_{j\in S_1}|\theta_j - \theta_j^\star|$$

for some $C_{\delta 0}, C_{\delta 1} > 0$ *with probability close to* 1 *for small enough* σ, *and that there exist* $\alpha_j > 0$, $j = 1, \ldots, p$, *and* $C_{\pi 0} > 0$ *such that for all* $\theta \in \Theta$,

$$\frac{\pi(d\theta)}{d\theta} \le C_{\pi 0} \prod_{j \in S_0 : |\theta_j| < \delta_0/\sqrt{p_0}} \theta_j^{\boldsymbol{\alpha}_j - 1} \prod_{j \in S_1 : \theta_j < \delta_1} \theta_j^{\boldsymbol{\alpha}_j - 1}.$$

If $C_{\delta 0}\delta_0/\sigma^2 \to \infty$ *and* $C_{\delta 1}\delta_1/\sigma^2 \to \infty$, *then* $\Delta_0(\delta) \to 0$ *as* $\sigma \to 0$ *with probability* 1, *that is*, *Assumption* L *is satisfied*.

The proof is given in Appendix A.2.

EXAMPLE 3 (Poisson likelihood). Consider $Y_i \sim \text{Poisson}(\theta)$ independently for $i = 1, \ldots, n$, where the true value is $\theta^\star = 0$. In this case, $\sigma^2 = 1/n$ and $\mathbb{P}(Y_i = 0) = 1$. Consider an improper prior for θ with density $p(\theta) = \theta^{\alpha-1}$ with some $\alpha > 0$; the case $\alpha = 1/2$ corresponds to the Jeffreys prior for parameter θ. In this case, the exact posterior distribution for θ is $\Gamma(\alpha, n)$, that is, $n\theta|Y \sim \Gamma(\alpha, 1)$ which agrees with Theorem 1, and the exact 95% credible interval for θ is $[0, \gamma_\alpha(0.05)/n]$ where $\gamma_\alpha(0.05)$ is the 95% percentile of the $\Gamma(\alpha, 1)$ distribution. For $\alpha = 1/2$, the credible interval is $[0, 1.92/n]$, and for $\alpha = 0.05$, it is $[0, 0.27/n]$. By decreasing α to 0, we can construct a credible interval of arbitrarily small length for fixed n, even for $n = 1$.

EXAMPLE 4 (Binomial distribution). Consider the problem of estimating the unknown probabilities of Binomial distributions $Y_i \sim \text{Bin}(n_i, \theta_i)$ independently, $i = 1, \ldots, p$, for $\theta_i \in [0, 1]$, where the true value $\theta_i^\star$ of some θ_i is 0. We assume that all $\theta_i^\star < 1$ (if $\theta_i^\star = 1$ for some i, consider $n_i - Y_i$ as data and $1 - \theta_i^\star$ as the corresponding parameter). We study the limit of the posterior distribution for large n_i for all $i = 1, \ldots, p$ such that $n_i/n \to \omega_i \in (0, 1)$ where $n = \sum_{i=1}^p n_i$, and p is fixed. This situation is not covered by the standard BvM theorem. Consider a conjugate Beta prior $\theta_i \sim B(\alpha, 1)$ independently, with some fixed $\alpha > 0$. In this case, $\sigma^2 = 1/n$ and, as $n \to \infty$,

$$\ell^\star(\theta) = \lim_{n\to\infty} \ell_Y(\theta) = \sum_{i=1}^p \omega_i \big[\theta_i^\star \log(\theta_i) - \big(1 - \theta_i^\star\big) \log(1 - \theta_i)\big].$$

If $\theta_i^\star = 0$, the corresponding summand in $\ell^\star(\theta)$ is $-\omega_i \log(1 - \theta_i)$ which is defined for $\theta_i \in [0, 1)$, and then $\nabla_i \ell^\star(\theta) = \omega_i/(1 - \theta_i)$. In this case, $S_0^\star$ is always empty, $\nabla_i \ell^\star(\theta^\star) = 0$ for $\theta_i^\star \neq 0$ and $\nabla_i \ell^\star(\theta^\star) = -\omega_i$ for $\theta_i^\star = 0$. Assumption M was verified in Example 2, and it is easy to check that Assumptions S, P and L are satisfied (e.g., for $p = 1$ and $\theta_i^\star = 0$, conditions of Lemma 3 hold with $C_{\delta_1} = 1$ and $C_{\pi 0} = \alpha$). Therefore, $\Omega_{00} = \text{diag}(\omega_i/\{\theta_i^\star(1 - \theta_i^\star)\}, i \in S_0)$, $a_1 = (\omega_i, i \in S_1)$, and $a_{0,i} = (Y_i - n_i\theta_i^\star)/(\sqrt{n}\omega_i)$. Theorem 1 implies the following asymptotic approximation of the posterior distribution of $(\theta_{S_0}, \theta_{S_1})$:

$$(\theta_{S_0}, \theta_{S_1})|Y \sim \mathcal{N}_{p_0}\left(\theta_{S_0}^\star + \frac{a_0}{\sqrt{n}}, \frac{1}{n}\Omega_{00}^{-1}\right) \times \Gamma_{p_1}(\alpha, na_1).$$

Similarly to the Poisson likelihood case (Example 3), for α close to 0, the approximate credible intervals for θ_i, $i \in S_1$, are small. This is easy to see from the marginal $100(1-\beta)\%$ credible intervals which are $[0, \gamma_\alpha(\beta)/(n\omega_i)]$.

EXAMPLE 5 (Mixed effects model). Consider a model studied by Vu and Zhou (1997): $Y_{ij}|\beta_i \sim \mathcal{N}(\mu + \beta_i, \tau^2)$ where $\beta_i \sim \mathcal{N}(0, \theta)$ independently, for $i = 1, \dots, n$ and $j = 1, \dots, m$. Here there are n classes with m elements in each, and the parameter of interest is the contribution of the classes that is characterised by the parameter $\theta \in \Theta = [0, \infty)$, where the value $\theta = 0$ corresponds to the absence of the random effects β_i. We study the asymptotic concentration of the posterior distribution of θ when the number of classes n grows while the number of class elements m remains fixed. We consider a prior distribution for θ with density $p(\theta) \propto \theta^{\alpha-1} e^{-b\theta}$ for $\alpha > 0$ and $b \geq 0$, which includes a case of improper prior distributions when $b = 0$. Note that the inverse Gamma prior with density $p(\theta) \propto \theta^{-\alpha-1} e^{-b/\theta}$ potentially leads to very slow convergence, since it has a root of infinite order at 0.

We start with the case μ and τ known, so without loss of generality we fix $\mu = 0$ and $\tau = 1$. After integrating out β_i we have that $\overline{Y}_i = m^{-1} \sum_{i=1}^m Y_{ij} \sim \mathcal{N}(0, \theta^\star + 1/m)$, independently, where $\theta^\star$ is the true value of the parameter θ. If $\theta^\star > 0$, then the model is regular and the posterior distribution of θ is asymptotically Gaussian. Now we consider the case $\theta^\star = 0$. Using the marginal likelihood of $\bar{y}_i$ given θ and taking $\sigma^2 = 1/n$, we have

$$\ell^\star(\theta) = \lim_{n \to \infty} \ell_Y(\theta) = -\frac{1}{2m(\theta + 1/m)} - \frac{1}{2} \log(\theta + 1/m)$$

since $\mathbb{E}\overline{Y}_i^2 = \theta^\star + 1/m = 1/m$, and Assumption M is satisfied with $\nabla \ell^\star(\theta^\star) = 0$. It is easy to check that Assumptions B, S, P and L are satisfied, and $\nabla^2 \ell^\star(\theta^\star) = -m^2/2$. Thus, by Theorem 1, the approximate posterior distribution of $\sqrt{n}\theta$ has density

$$p_{\theta\sqrt{n}}(x|y) \approx C_{\alpha,m,a_0} x^{\alpha-1} e^{-(x-a_0)^2/m^2}, \qquad x \geq 0$$

with $a_0 = (m/(2\sqrt{n}))(n^{-1} \sum_{i=1}^n m\overline{Y}_i^2 - 1)$. It is easy to show that the Cramer–Rao lower bound on the variance of estimators of θ applies here, even in the case $\theta^\star = 0$. Thus, using a prior with $\alpha < 1$ (i.e., introducing a bias towards 0) would lead to superefficiency, that is, loss of efficiency for $\theta^\star \neq 0$. In the case $\alpha = 1$ the posterior distribution is Gaussian with the same mean and variance as in the BvM theorem but truncated to $\theta \geq 0$. The length of the credible interval for θ in this case is smaller than in the case where $\theta^\star$ is an interior point.

Now consider the case where parameters (μ, τ^2, θ) are estimated jointly with a continuous prior for (μ, τ^2) whose density is finite and positive at the true value $(\mu^\star, \tau^\star)$. Then

$$-\ell^\star(\mu, \tau^2, \theta) = \frac{(m-1)\tau^{\star 2}}{2\tau^2} + \frac{(\mu - \mu^\star)^2 + \tau^{\star 2}/m}{2\tau^2(\theta + 1/m)} + \frac{\log(\theta + 1/m)}{2} + \frac{m \log(\tau^2)}{2},$$

since $\mathbb{E}(\overline{Y_i}-\mu)^2 = \tau^{\star 2}/m + (\mu-\mu^\star)^2$ and $\mathbb{E}\sum_{j=1}^m (Y_{ij}-\overline{Y_i})^2 = (m-1)\tau^{\star 2}$. The function $\ell^\star(\mu, \tau^2, \theta)$ is maximised at $\mu = \mu^\star$, $\tau = \tau^\star$ and $\theta = \theta^\star = 0$, with zero gradient and the negative matrix of the second order derivatives Ω_{00} and its inverse (the covariance matrix) being

$$\Omega_{00} = \begin{pmatrix} \frac{m}{\tau^{\star 2}} & 0 & 0 \\ 0 & \frac{m}{2\tau^{\star 4}} & \frac{m}{2\tau^{\star 2}} \\ 0 & \frac{m}{2\tau^{\star 2}} & \frac{m^2}{2} \end{pmatrix},$$

$$\Omega_{00}^{-1} = \begin{pmatrix} \frac{\tau^{\star 2}}{m} & 0 & 0 \\ 0 & \frac{2\tau^{\star 4}}{m-1} & -\frac{2\tau^{\star 2}}{m(m-1)} \\ 0 & -\frac{2\tau^{\star 2}}{m(m-1)} & \frac{2}{m(m-1)} \end{pmatrix}.$$

If $\alpha = 1$, then the approximate joint posterior distribution of $\sqrt{n}(\theta - \theta^\star, \mu - \mu^\star, \tau^2 - \tau^{\star 2})$ is Gaussian truncated to $\theta - \theta^\star = \theta \geq 0$ with bias as given in Theorem 1 and the covariance matrix Ω_{00}^{-1} given above. Note that θ and τ^2 are asymptotically correlated, with correlation $-m^{-1/2}$.

6. Asymptotic behaviour of the posterior distribution for SPECT.

6.1. *Approximation of the posterior distribution.* Consider the SPECT model defined in Section 2, in which $\theta^\star$ has some zero coordinates. The assumptions of Theorem 1 were verified in Examples 1 and 3 (Assumptions M, B, S), and the log cosh Markov random field prior distribution satisfies Assumption P with $\alpha_j = 1$ for all j. Assumption L also holds, since the conditions of Lemma 3 are satisfied for independent Poisson random variables with $C_{\delta 0} = 0.5\delta_0(\delta_0 + \sqrt{p_0}y^\star_{\min})^{-1}y^\star_{\min}$, $C_{\delta 1} = \min_j(a_{1,j})$, where $y^\star_{\min} = \min_{j:y^\star_j>0} y^\star_j$ with $y^\star = A\theta^\star$ for small enough $\sigma = 1/\sqrt{\mathcal{T}}$, due to the inequality $\log(1+x) - x \leq -xb/(b+1)$ for $x > b > 0$.

For this model, $\nabla\ell^\star(\theta^\star) = -\sum_{i:y^\star_i=0} A_i^T$, which is nonzero if $Z = \{i \in \{1,\dots,n\} : y^\star_i = 0\}$ is not empty. Hence, nonregularity arises from the elements where there are no detected photons ($y^\star_i = 0$) and the likelihood degenerates: $\mathbb{P}_{\theta^\star}(Y_i = 0) = 1$ for $i \in Z$ but, since $A_i \neq 0$, this gives us information about those θ_j where $A_{ij} \neq 0$, that is, on $S_1 = \{j : \theta^\star_j = 0 \text{ and } \sum_{i\in Z} A_{ij} \neq 0\}$.

The limiting distribution of θ_{S_1}/σ^2 is exponential with parameter $a_1 = \sum_{i\in Z} A^T_{iS_1}$. The parameter $(\theta_{S_0} - \theta^\star_{S_0})/\sigma$ has approximately a truncated Gaussian distribution with parameters

$$\Omega_{00} = A^T_{\overline{Z},S_0}\operatorname{diag}(1/[y^\star]_{\overline{Z}})A_{\overline{Z},S_0}, \qquad a_0 = \Omega_{00}^{-1}A^T_{\overline{Z},S_0}\widetilde{Y}/\sigma,$$

where $\widetilde{Y}$ is a vector with coordinates $Y_i/y_i^\star - 1$ for $i \in \overline{Z}$. Truncation takes place for parameters $\theta_{S_0^\star}$ with $S_0^\star = \{j : \theta_j^\star = 0 \text{ and } \sum_{i \in Z} A_{ij} = 0\}$.

If the vector of Poisson means $y^\star = A\theta^\star$ has only positive coordinates (Z is empty), the model is regular, and the posterior distribution of $(\theta - \theta^\star)/\sigma$ is approximately truncated Gaussian.

6.2. *Practical implications of the approximate posterior.* We will briefly discuss some practical implications of Theorem 1. Well-developed methods for SPECT reconstruction using our model, using Markov chain Monte Carlo computation, deliver both approximate, simulation-consistent, posterior means and variances; see Weir (1997) for a fully Bayesian reconstruction. The theorem provides valuable knowledge which can enrich the interpretation of such results, enabling approximate probabilistic inference.

Inferential questions of real interest, including (a) quantitative inference about amounts of radio-labelled tracer within specified regions of interest, or (b) tests for significance of apparent hot- or cold-spots, can be answered using approximate posteriors for *linear combinations* $w^T\theta$ of parameters, and are particularly amenable to treatment. Specifically, suppose that for any nonempty set of pixels $R \subseteq \{1, 2, \dots, p\}$, w^R denotes the vector with elements $w_j^R = 1/|R|$ for $j \in R$, 0 otherwise. Then to deal with case (a) we can take $w = w^R$ to deliver $w^T\theta$ as the average concentration of tracer in region R, and for case (b) take $w = w^{R_1} - w^{R_2}$ for the difference in average concentration between regions R_1 and R_2.

To construct an approximation of the posterior distribution, we require estimates of unknown parameters. We use the marginal posterior modes estimate $\hat{\theta}$, $\hat{\theta}_i = \operatorname{argmax} p(\theta_i|y)$, instead of $\theta^\star$, $\hat{y} = A\hat{\theta}$ instead of $y^\star$,

$$\widehat{S}_1 = \{j : \nabla_j \ell_y(\hat{\theta}) < 0\}, \qquad \widehat{Z} = \{i : \hat{y}_i = 0\}.$$

A more robust way to estimate S_1 would be to use $\widehat{S}_{1,\varepsilon} = \{j : \nabla_j \ell_y(\hat{\theta}) < -\varepsilon\}$ for some small enough $\varepsilon > 0$; however, sensitivity to the choice of ε would need to be investigated. Then, the approximate posterior of $z = (\theta - \hat{\theta})$ is

$$\phi(z) = \prod_{j \in \widehat{S}_1} [\hat{a}_j/(2\sigma^2)](2\pi\sigma^2)^{-p_0/2}[\det(\widehat{\Omega})]^{1/2} \exp\{-z_{\widehat{S}_0}^T \widehat{\Omega} z_{\widehat{S}_0}/(2\sigma^2) - z_{\widehat{S}}^T \hat{a}/\sigma^2\},$$

where $\widehat{\Omega} = \sum_{i \notin \widehat{Z}} y_i/[\hat{y}_i]^2 A_{i,\widehat{S}_0} A_{i,\widehat{S}_0}^T$ and $\hat{a} = \sum_{i \in \widehat{Z}} A_{i,\widehat{S}_1}^T$.

6.3. *Finite sample performance.* We briefly discuss the extent to which the approximation in Theorem 1 holds true for data on the scale of a real SPECT study. A formal assessment of this would entail a major study beyond the scope of this paper, so we present selected results from analysis of two data sets based on a SPECT scan of the pelvis of a human subject.

In the first experiment, the matrix A was constructed according to the model in Green (1990) and Weir (1997), capturing geometry, attenuation, and radioactive decay for a setup consisting of 64 projections from a 2-dimensional slice through the patient, each projection yielding an array of 52 photon counts, on a spatial resolution of 0.57 cm. The data set was obtained from Bristol Royal Infirmary; the total photon count was 45,652; individual counts ranged from 0 to 85, averaging 13.7. Reconstruction was performed on a 48×48 square grid of 0.64 cm pixels, using the log cosh prior with hyperparameters fixed at $\gamma = 25$ and $\zeta = 8$, obtained using a simple MCMC sampler. We employed 20,000 sweeps of a deterministic-raster-scan single-pixel random walk Metropolis sampler on a square-root scale for θ, chosen to avoid extremes in acceptance rate at high- and low-spots in the image.

Figure 1 shows selected aspects of this analysis; see caption for details. Our tentative conclusion is that the marginal posterior distributions for individual pixels θ_j do appear to be approximately Gaussian in high-spots and approximately exponential in low-spots, consistent with the theoretical limits presented in Theorem 1.

A second experiment was focussed on a more precise and quantitative assessment of the approximation to the posterior derived in the previous section. The setup is the same as in the first experiment, except at half the resolution, so that reconstruction was on a 24×24 grid of 1.28 cm pixels. The corresponding A matrix is now better-conditioned, and p is only 576, so that manipulation of the matrices is entirely tractable. Synthetic data was generated using this A and a "ground truth" obtained from an approximate MAP reconstruction from the same real data set as used above, yielding photon counts between 0 and 243, totalling 138,310. 50,000 sweeps of the MCMC sampler were used, with prior settings $\gamma = 200$, $\zeta = 8$.

Figure 2 displays the agreement between the elements of $\hat{a}$ and the reciprocals of the MCMC-computed posterior means of θ, for pixels in $\widehat{S}_1$, and also that between the diagonal elements of $\widehat{\Omega}^{-1}$ and the posterior variances of θ for pixels in $\widehat{S}_0$.

Figure 3 displays two bivariate posterior marginals, computed by MCMC, and the corresponding approximations. In the left panel, one component is in $\widehat{S}_1$ and

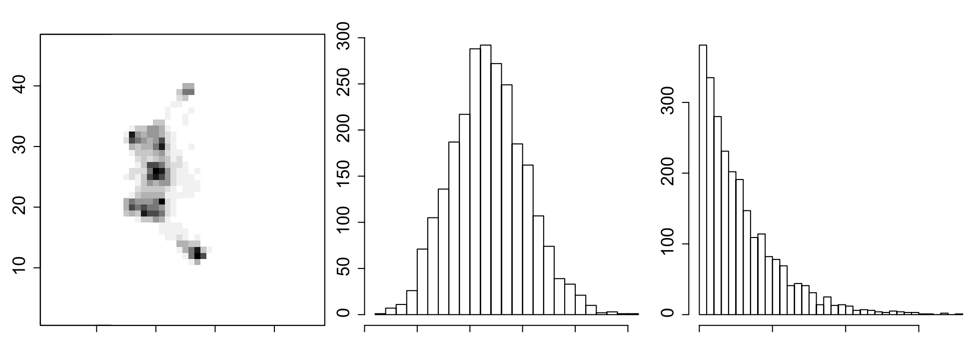

FIG. 1. *Analysis of real SPECT data*: *posterior mean reconstruction as a grey-scale image*, *histogram of marginal posterior for a high-spot pixel* (*row* 12, *column* 28), *and the same for a low-spot pixel* (*row* 12, *column* 31).

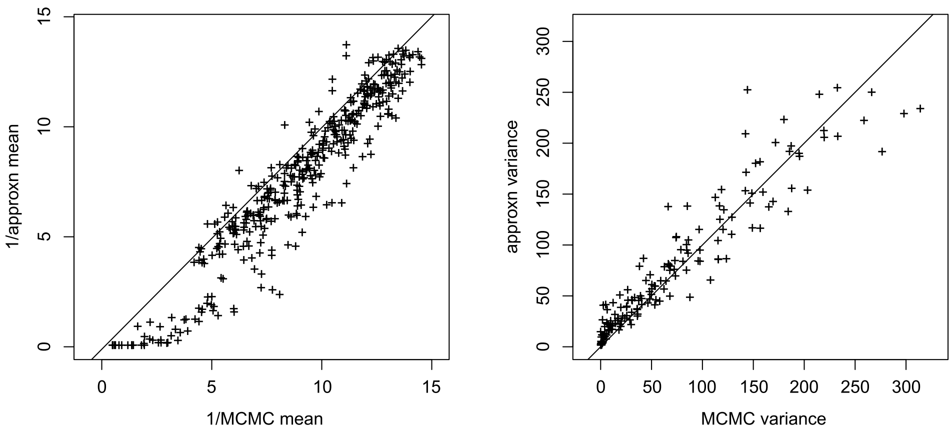

FIG. 2. *Agreement between* (*left panel*) *the elements of* $\hat{a}$ *and the reciprocals of the MCMC-computed posterior means of* θ, *for pixels in* $\widehat{S}_1$, *and also that between* (*right panel*) *the diagonal elements of* $\widehat{\Omega}^{-1}$ *and the posterior variances of* θ *for pixels in* $\widehat{S}_0$.

one in $\widehat{S}_0$, so the approximation is Gaussian/exponential; on the right both components are from $\widehat{S}_0$, so we have a bivariate Gaussian.

We conclude that for this realistic/modest-scale SPECT reconstruction problem, the small-variance asymptotics of this paper provide a good approximation to the posterior, even for $\sigma^2 = 1$.

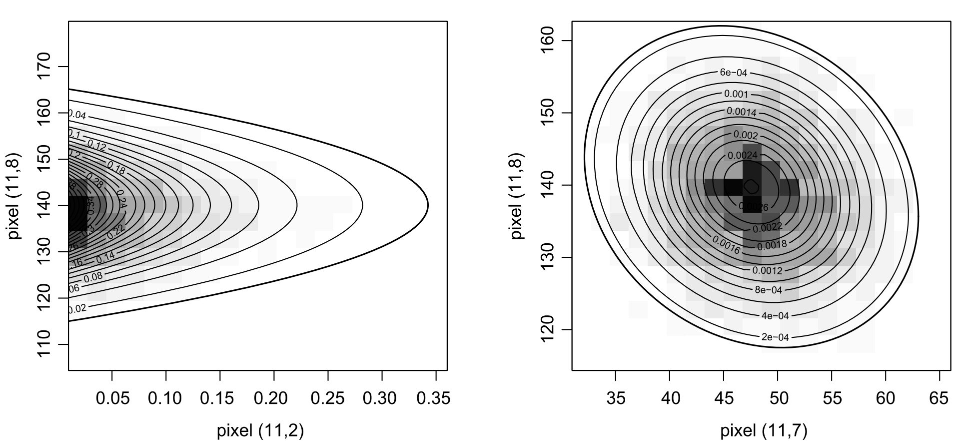

FIG. 3. *Two bivariate marginals of the posterior, as computed by MCMC* (*grey-scale image*), *and the corresponding approximations* (*contours*). *In the left panel, one pixel is in* $\widehat{S}_1$ *and one in* $\widehat{S}_0$, *so the approximation is Gaussian/exponential*; *in the right panel both pixels are from* $\widehat{S}_0$, *so we have a bivariate Gaussian. The outermost contour represents the* 95% *HPD credible region based on the approximation.*

7. Discussion. When the posterior distribution concentrates on the boundary, we have shown that the classic Bernstein–von Mises theorem does not hold for all components. There are two different types of non-Gaussian component: one, with the same parametric rate of convergence, is a truncated Gaussian or a polynomially tilted modification of this if the prior density is not bounded away from zero and infinity on the boundary, and the second is a Gamma, with a faster rate of convergence. An interesting property of the components of the second type is that they are not subject to a lower bound on efficiency, unlike the "regular" and the first-type boundary components. Under some models with this property, at least part of the data is observed exactly, so perhaps it should not be an unexpected phenomenon; see examples of Poisson and Binomial likelihoods in Section 5. This property is quite remarkable: in principle, it allows the recovery of the unknown parameter on the boundary with an arbitrarily small precision (particularly in the case there is no approximation error), by choosing an appropriate prior distribution, without losing asymptotic efficiency if the parameter is not on the boundary. This property is related to convergence in finitely-many steps of the projected gradient method for a sharp minimum for a noise-free function [Polyak (1983), Theorem 1, page 182; thanks to Alexandre Tsybakov for bringing this to our attention].

A related but different problem involves a nonregular model where the density of the observations has one or more jumps at a point that depends on the unknown parameter, for example, $Y_i \sim U[0, \theta]$, $i = 1, \dots, n$, independently. This type of problem has been extensively studied from both frequentist and Bayesian perspectives [Chernozhukov and Hong (2004), Ghosal, Ghosh and Samanta (1995), Ghosal and Samanta (1995), Ghosh, Ghosal and Samanta (1994), Hirano and Porter (2003), Ibragimov and Has'minskiĭ (1981)]. In the problem treated in this paper, the rate of convergence of the posterior distribution of the unknown nonregular parameter as a function of n is the same as in this case where the unknown parameter controls the positions of jumps, faster than the standard parametric rate. However, there is a crucial difference: in the former case, the posterior distribution has a data-dependent random shift, whereas in the latter case there is no such shift.

The nonasymptotic version of the main result shows that other parameters of the model can also affect convergence in practice, such as the smallest eigenvalues of the precision matrices in the $\mathcal{PTN}$ part of the limit and the smallest parameter of the scale of the Gamma distributions.

It is easy to verify that Theorem 1 derived here applies also to misspecified models, with $\mathbb{P}_{\theta^\star,\sigma}$ being replaced by the true distribution of $\mathbf{Y}$ and $\theta^\star$ defined as the unique maximum of $\ell^\star(\theta)$ as in Assumption M. This will be discussed elsewhere.

An interesting direction for future work is to study both the behaviour of the posterior distribution, and the question of optimal prior specification, in a framework where the spatial resolution is infinitely refined, placing smoothness class constraints on $\theta^\star$.

APPENDIX: PROOFS

A.1. Proof of the main result. We start with a lemma.

LEMMA 4. *Consider the function $\ell_Y(\theta)$ defined in Section 3.1 and assume that Assumptions M, B and S hold. Then, on the event $\mathcal{A}_0 \cap \mathcal{A}_1$ defined by (11) with some $\delta_{*0}, \delta_{*1} > 0$, for $\theta \in \Theta^\star(\delta)$,*

$$\ell_Y(\theta) - \ell_Y(\theta^\star) \\ \geq (\theta_{S_0} - \theta^\star_{S_0})^T \nabla_{S_0} \ell_Y(\theta^\star) - (\theta_{S_0} - \theta^\star_{S_0})^T \widetilde{\Omega}_{00}(\theta_{S_0} - \theta^\star_{S_0})/2 - \tilde{a}^T \theta_{S_1},$$

$$\ell_Y(\theta) - \ell_Y(\theta^\star) \\ \leq (\theta_{S_0} - \theta^\star_{S_0})^T \nabla_{S_0} \ell_Y(\theta^\star) - (\theta_{S_0} - \theta^\star_{S_0})^T \overline{\Omega}_{00}(\theta_{S_0} - \theta^\star_{S_0})/2 - \bar{a}^T \theta_{S_1},$$

where

$$\widetilde{\Omega}_{00} = \Omega_{00} + \delta_{*0} I_{p_0}, \qquad \overline{\Omega}_{00} = \Omega_{00} - \delta_{*0} I_{p_0},$$
$$\tilde{a} = a_1 + \delta_{*1} \mathbf{1}_{p_1}, \qquad \bar{a} = a_1 - \delta_{*1} \mathbf{1}_{p_1}.$$

Here $\mathbf{1}_{p_1} = (1, \dots, 1)^T$—a vector of length p_1, and I_{p_0} is $p_0 \times p_0$ identity matrix.

PROOF. Applying the Taylor expansion of $\ell_Y(\theta)$ as a function of θ_{S_1} at point $\theta^\star_{S_1}$, and then expanding $\ell_Y(\tilde{\theta})$ where $\tilde{\theta}_{S_0} = \theta_{S_0}$ and $\tilde{\theta}_{S_1} = \theta^\star_{S_1}$, as a function of θ_{S_0} at point $\theta^\star_{S_0}$, for some $\theta_{c0}, \theta_{c1} \in \Theta^\star(\delta)$, we have

$$\ell_Y(\theta) - \ell_Y(\theta^\star) = (\theta_{S_1} - \theta^\star_{S_1})^T \nabla_{S_1} \ell_Y(\theta_{c1}) + (\theta_{S_0} - \theta^\star_{S_0})^T \nabla_{S_0} \ell_Y(\theta) \\ + (\theta_{S_0} - \theta^\star_{S_0})^T \nabla_{S_0, S_0} \ell_Y(\theta_{c0})(\theta_{S_0} - \theta^\star_{S_0})/2.$$

Applying the bounds defining events $\mathcal{A}_0$ and $\mathcal{A}_1$ to $\nabla_{S_1} \ell_Y(\theta_{c1})$ and $\nabla_{S_0,S_0} \ell_Y(\theta_{c0})$, and using that $\theta_{S_1} - \theta^\star_{S_1} = \theta_{S_1}$ is a vector with nonnegative components, we have

$$\ell_Y(\theta) - \ell_Y(\theta^\star) \leq (\theta_{S_1} - \theta^\star_{S_1})^T [-a_1 + \delta_{*1} 1_{|S_1|}] + (\theta_{S_0} - \theta^\star_{S_0})^T \nabla_{S_0} \ell_Y(\theta) \\ + (\theta_{S_0} - \theta^\star_{S_0})^T [-\Omega_{00} + \delta_{*0} I_{|S_0|}](\theta_{S_0} - \theta^\star_{S_0})/2,$$

and hence the first statement of the lemma. Applying the inequalities on the events $\mathcal{A}_k$ as lower bounds, we obtain the second statement of the lemma. □

PROOF OF THEOREM 1. Denote $v = (v_0^T, v_1^T)^T = D^{-1} U(\theta - \theta^\star)$ where $v_0 = (\theta_{S_0} - \theta^\star_{S_0})/\sigma$ and $v_1 = (\theta_{S_1} - \theta^\star_{S_1})/\sigma^2$; the Jacobian of this change of variables is $\sigma^{p_0 + 2p_1}$. The image of $\Theta^\star(\delta)$ under this transform is

$$B_R = B_2(0, R_0) \times [0, R_1)^{p_1} \cap D_\sigma^{-1} U(\Theta - \theta^\star),$$

with $R_0 = \delta_0/\sigma$ and $R_1 = \delta_1/\sigma^2$. Under Assumptions B and S, the conditions of Lemma 2 hold, which implies that if $\|\theta^\star_{S_0}\| \ge \delta_0$ and $\delta_k \le c_k$, $B_R = [B_{2,p_0}(0, R_0) \cap \mathcal{V}_0] \times [0, R_1]^{p_1}$ where $\mathcal{V}_0 = \mathbb{R}^{p_0 - p_0^\star} \times \mathbb{R}_+^{p_0^\star}$, and the set B_R becomes $\mathcal{V}^\star = \mathcal{V}_0 \times \mathbb{R}_+^{p_1}$ as $\sigma \to 0$.

The triangle inequality for the total variation norm gives

$$
\begin{aligned}
&\|\mathbb{P}_{\mathcal{S}(\theta-\theta^\star)|Y} - \mu^\star\|_{\mathrm{TV}} \\
&\quad \le \|\mathbb{P}_{\mathcal{S}(\theta-\theta^\star)|Y}\mathbf{1}_{B_R} - \mu^\star\mathbf{1}_{B_R}\|_{\mathrm{TV}} \\
&\qquad + \|\mu^\star\mathbf{1}_{B_R} - \mu^\star\|_{\mathrm{TV}} + \|\mathbb{P}_{\mathcal{S}(\theta-\theta^\star)|Y}\mathbf{1}_{B_R} - \mathbb{P}_{\mathcal{S}(\theta-\theta^\star)|Y}\|_{\mathrm{TV}},
\end{aligned}
\tag{13}
$$

where the balls B_R are defined above. Here $\mu\mathbf{1}_{B_R}$ is a probability measure μ truncated to B_R and normalised to be a probability measure. If the measure μ_1 is absolutely continuous with respect to measure μ_2, with density f, the total variation norm can be written as

$$
\|\mu_1 - \mu_2\|_{\mathrm{TV}} = 2\int_\Theta (f-1)_+ \, d\mu_2,
$$

where $(x)_+ = \max(x, 0)$ [van der Vaart (1998)]. This can be used in each of the summands in the upper bound (13).

In this proof we will use $\alpha = (\alpha_0, \alpha_1)$, for simplicity of notation.

Define the measure $\mu(dv; a_1, \alpha, b, \Sigma)$ for $v = (v_0^T, v_1^T)^T$, $v_0 \in \mathbb{R}^{p_0 - p_0^\star} \times [0,\infty)^{p_0^\star}$ and $v_1 \in [0,\infty)^{p_1}$, by

$$
\frac{\mu(dv; a_1, \alpha, b, \Sigma)}{dv} = \prod_{j \in T_0^\star \cup T_1} v_j^{\alpha_j - 1} e^{-a_1^T v_1 - v_0^T \Sigma v_0/2 + v_0^T b},
\tag{14}
$$

where $T_0^\star = \{p_0 - p_0^\star + 1, \dots, p_0\}$, $T_1 = \{p_0 + 1, \dots, p\}$, $a_1 \in (0,\infty)^{p_1}$, $b \in \mathbb{R}^{p_0}$, $\alpha = (\alpha_j)_{j \in T_0^\star \cup T_1} \in (0,\infty)^{p_0^\star + p_1}$, and Σ is a $p_0 \times p_0$ positive definite matrix.

We start with the first term in (13). By Lemma 4, on the event $\mathcal{A}_0 \cap \mathcal{A}_1$ defined by (11), for any measurable $\mathcal{B} \subseteq \Theta^\star(\delta)$, with $\mathcal{B}_v = D_\sigma^{-1} U(\mathcal{B} - \theta^\star) \subseteq B_R$, we have

$$
\begin{aligned}
&\int_{\mathcal{B}} \exp\{[\ell_Y(\theta) - \ell_Y(\theta^\star)]/\sigma^2\}\pi(d\theta) \\
&\quad \ge J_\sigma C_\pi (1 - \Delta_\pi) \\
&\qquad \times \int_{\mathcal{B}_v} \prod_{j \in T_0^\star \cup T_1} v_j^{\alpha_j - 1} \exp\{v_0^T \nabla_{S_0} \ell_Y(\theta^\star)/\sigma - \|\widetilde{\Omega}_{00}^{1/2} v_0\|^2/2 - \tilde{a}^T v_1\}\, dv \\
&\quad = J_\sigma C_\pi (1 - \Delta_\pi) \mu(\mathcal{B}_v; \tilde{a}, \alpha, \nabla_{S_0}\ell_Y(\theta^\star)/\sigma, \widetilde{\Omega}_{00}),
\end{aligned}
$$

where $J_\sigma = \sigma^{p_0 - p_0^\star + \sum_{j \in T_0^\star} \alpha_{0,j} + 2\sum_{j=1}^{p_1} \alpha_{1,j}}$, and the measure $\mu(dv; a_1, \alpha, b, \Sigma)$ is defined by (14). Similarly, using Lemma 4, we obtain an upper bound on the event $\mathcal{A}_0 \cap \mathcal{A}_1$,

$$\begin{aligned}
&\int_{\mathcal{B}} \exp\{[\ell_Y(\theta) - \ell_Y(\theta^\star)]/\sigma^2\}\pi(d\theta) \\
&\quad \le J_\sigma C_\pi (1+\Delta_\pi) \\
&\qquad \times \int_{\mathcal{B}_v} \prod_{j \in T_0^\star \cup T_1} v_j^{\alpha_j - 1} \exp\{v_0^T \nabla_{S_0} \ell_Y(\theta^\star)/\sigma - \|\overline{\Omega}_{00}^{1/2} v_0\|^2/2 - \bar{a}^T v_1\}\, dv \\
&\quad = J_\sigma C_\pi (1+\Delta_\pi) \mu(\mathcal{B}_v; \bar{a}, \alpha, \nabla_{S_0}\ell_Y(\theta^\star)/\sigma, \overline{\Omega}_{00}).
\end{aligned}$$

To simplify the notation, denote $a_0 = \Omega_{00}^{-1} \nabla_{S_0} \ell_Y(\theta^\star)/\sigma$ and

$$\begin{aligned}
\bar{\mu}(dv) &= \mu(dv; \bar{a}, \alpha, \Omega_{00} a_0, \overline{\Omega}_{00}), \\
\tilde{\mu}(dv) &= \mu(dv; \tilde{a}, \alpha, \Omega_{00} a_0, \widetilde{\Omega}_{00}).
\end{aligned}$$

The measure $\tilde{\mu}$ is finite since $a_0 = \nabla_{S_0} \ell_Y(\theta^\star)/\sigma$ is finite with high probability due to Assumption S(4), and all its other parameters are positive or positive definite. The measure $\bar{\mu}$ is finite if $\delta_{*1} < \min_j a_{1,j}$ and $\delta_{*0} < \lambda_{\min}(\Omega_{00})$. These conditions hold if δ_{*0}, δ_{*1} are small enough which is possible due to Assumption S.

For $\mathcal{B}_v = \mathcal{B}_1 \times B_\infty(0, r_1)$ for some $\mathcal{B}_1 \subset \mathcal{V}_0$ and $r_1 \in (0, R_1)$, we have

$$\mu(\mathcal{V}^\star; a_1, \alpha, b, \Sigma) = \prod_{i=1}^{p_1} [a_{1,i}^{-\alpha_{1,i}} \Gamma(\alpha_{1,i})] \int_{\mathcal{V}_0} \prod_{j \in T_0^\star} v_{0,j}^{\alpha_{0,j}-1} e^{-v_0^T \Sigma v_0/2 + v_0^T b}\, dv_0,$$

$$\frac{\mu(\mathcal{B}_v; a_1, \alpha, b, \Sigma)}{\mu(\mathcal{V}^\star; a_1, \alpha, b, \Sigma)} = \mathcal{PTN}_{p_0}(\mathcal{B}_1; \Sigma^{-1} b, \Sigma^{-1}, p_0^\star, \alpha_0) \prod_{j=1}^{p_1} \Gamma((0, r_1); \alpha_{1,j}, a_{1,j}),$$

where the probability measure $\mathcal{PTN}_{p_0}(\cdot; b, \Omega_{00}^{-1}, p_0^\star, \alpha_0)$ is defined by (4), and $\Gamma(\cdot; \alpha_{1,j}, a_{1,j})$ is the probability measure associated with distribution $\Gamma(\alpha_{1,j}, a_{1,j})$.

Hence, the posterior density of $\mathcal{S}(\theta - \theta^\star)$ normalised by the posterior measure of B_R, is bounded on $\mathcal{A}_0 \cap \mathcal{A}_1$ by

$$\frac{1-\Delta_\pi}{1+\Delta_\pi} \frac{\tilde{\mu}(dv)}{\bar{\mu}(B_R)} \le \frac{dp(\mathcal{S}(\theta-\theta^\star)|Y)}{p(B_R|Y)} \le \frac{\bar{\mu}(dv)}{\tilde{\mu}(B_R)} \frac{1+\Delta_\pi}{1-\Delta_\pi}.$$

Therefore, the first term in (13) is bounded on $\mathcal{A}_0 \cap \mathcal{A}_1$ by

$$\begin{aligned}
&\|\mathbb{P}_{\mathcal{S}(\theta-\theta^\star)|Y} \mathbf{1}_{B_R} - \mu^\star \mathbf{1}_{B_R}\|_{\mathrm{TV}} \\
&\quad \le 2 \int_{B_R} \left[\frac{\mathbb{P}(dv|Y)\mu^\star(B_R)}{\mathbb{P}(B_R|Y)\mu^\star(dv)} - 1\right]_+ \frac{\mu^\star(dv)}{\mu^\star(B_R)} \\
&\quad \le 2 \int_{B_R} \left[\frac{\bar{\mu}(dv)}{\tilde{\mu}(B_R)} \frac{\mu^\star(B_R)}{\mu^\star(dv)} \frac{(1+\Delta_\pi)}{(1-\Delta_\pi)} - 1\right]_+ \frac{\mu^\star(dv)}{\mu^\star(B_R)}.
\end{aligned}$$

Define $\mu_0(dv) = \mu(dv; a_1, \alpha, \Omega_{00}a_0, \Omega_{00})$. Then

$$\frac{\mu^\star(dv)}{\mu^\star(B_R)} = \frac{\mu_0(dv)}{\mu_0(B_R)}$$

and

$$\frac{\bar{\mu}(dv)}{\mu_0(dv)} = \exp\{\delta_{*1}\mathbf{1}^T v_1 + \delta_{*0}\|v_0\|^2/2\},$$

which implies

$$\begin{aligned}
&\frac{\bar{\mu}(dv)}{\mu_0(dv)}\frac{\mu_0(B_R)}{\tilde{\mu}(B_R)} \\
&\quad = \exp\{\delta_{*0}\|v_0\|^2/2 + \delta_{*1}\mathbf{1}^T v_1\} \\
&\qquad \times \left(\int_{B_R} \prod_{i \in T_0^\star \cup T_1} v_i^{\alpha_i - 1} \exp\{-a_1^T v_1\} \exp\{-\|\Omega_{00}^{1/2} v_0\|^2/2 + v_0^T \Omega_{00} a_0\}\, dv\right) \\
&\qquad \Big/ \left(\int_{B_R} \prod_{i \in T_0^\star \cup T_1} v_i^{\alpha_i - 1} \exp\{-(a_1 + \delta_{*1}\mathbf{1})^T v_1 - \|\tilde{\Omega}_{00}^{1/2} v_0\|^2/2 \right. \\
&\qquad\qquad \left. + v_0^T \Omega_{00} a_0\}\, dv\right).
\end{aligned}$$

To show that this expression is greater than 1, it is sufficient to show that for any $\mathcal{B} \subseteq \{v_0 : (v_0^T, v_1^T)^T \in B_R\}$, the following expression is positive:

$$\begin{aligned}
&\int_{\mathcal{B}} \prod_{i \in T_0^\star} w_i^{\alpha_i - 1} e^{w^T \Omega_{00} a_0 - \|\Omega_{00}^{1/2} w\|^2/2}\, dw - \int_{\mathcal{B}} \prod_{i \in T_0^\star} w_i^{\alpha_i - 1} e^{w^T \Omega_{00} a_0 - \|\tilde{\Omega}_{00}^{1/2} w\|^2/2}\, dw \\
&\quad = \int_{\mathcal{B}} \prod_{i \in T_0^\star} w_i^{\alpha_i - 1} e^{-\|\tilde{\Omega}_{00}^{1/2} w\|^2/2 + w^T \Omega_{00} a_0} [\exp\{\delta_{*0}\|w\|^2/2\} - 1]\, dw > 0
\end{aligned}$$

which is the case. Thus, on $\mathcal{A}_0 \cap \mathcal{A}_1$, $(\bar{\mu}(dv)/\mu_0(dv))(\mu_0(B_R)/\tilde{\mu}(B_R)) \geq 1$ and hence

$$\begin{aligned}
&\|\mathbb{P}_{\mathcal{S}(\theta - \theta^\star)|Y}\mathbf{1}_{B_R} - \mu^\star \mathbf{1}_{B_R}\|_{\mathrm{TV}} \\
&\quad \leq 2\int_{B_R}\left[\frac{\bar{\mu}(dv)}{\tilde{\mu}(B_R)}\frac{\mu^\star(B_R)}{\mu^\star(dv)}\frac{(1+\Delta_\pi)}{(1-\Delta_\pi)} - 1\right]\frac{\mu^\star(dv)}{\mu^\star(B_R)} \\
&\quad = 2\left[\frac{\bar{\mu}(B_R)}{\tilde{\mu}(B_R)}\frac{(1+\Delta_\pi)}{(1-\Delta_\pi)} - 1\right] \\
&\quad = 2\frac{\bar{\mu}(B_R) - \tilde{\mu}(B_R)}{\tilde{\mu}(B_R)}\frac{(1+\Delta_\pi)}{(1-\Delta_\pi)} + 2\left[\frac{(1+\Delta_\pi)}{(1-\Delta_\pi)} - 1\right].
\end{aligned}$$

The difference of measures $\bar{\mu}(B_R) - \tilde{\mu}(B_R)$ is bounded by

$$
\begin{aligned}
&\int_{B_R} \prod_{i\in T_0^\star\cup T_1} v_i^{\alpha_i-1} e^{-v_0^T\widetilde{\Omega}_{00}v_0/2+v_0^T\Omega_{00}a_0-\tilde{a}v_1}\big[e^{\delta_{*0}\|v_0\|^2/2+\delta_{*1}1_{p_1}^T v_1}-1\big]\,dv\\
&\quad\le \int_{B_R} \prod_{i\in T_0^\star\cup T_1} v_i^{\alpha_i-1}\big[\delta_{*0}\|v_0\|^2/2+\delta_{*1}1_{p_1}^T v_1\big]e^{-v_0^T\overline{\Omega}_{00}v_0/2+v_0^T\Omega_{00}a_0-\bar{a}v_1}\,dv\\
&\quad\le \Bigg[\delta_{*0}E_\Phi+\delta_{*1}\sum_{j=1}^{p_1}(\alpha_{1,j}/\bar{a}_j)\Bigg]\bar{\mu}(\mathcal{V}^\star)
\end{aligned}
$$

due to the inequality $e^x - 1 \le xe^x$ for $x > 0$, where E_Φ is defined by

$$
E_\Phi = 0.5\int_{\mathcal{V}_0}\|w\|^2\mathcal{PTN}_{p_0}\big(dw;\overline{\Omega}_{00}^{-1}\Omega_{00}a_0,\overline{\Omega}_{00}^{-1},p_0^\star,\alpha_0\big), \tag{15}
$$

which is finite. Therefore,

$$
\begin{aligned}
&\|\mathbb{P}_{\mathcal{S}(\theta-\theta^\star)|Y}\mathbf{1}_{B_R}-\mu^\star\mathbf{1}_{B_R}\|_{\mathrm{TV}}\\
&\qquad\le \frac{2\bar{\mu}(\mathcal{V}^\star)}{\tilde{\mu}(B_R)}\frac{(1+\Delta_\pi)}{(1-\Delta_\pi)}\Bigg[\delta_{*0}E_\Phi+\delta_{*1}\sum_{j=1}^{p_1}\frac{\alpha_{1,j}}{\bar{a}_j}\Bigg]+\frac{4\Delta_\pi}{1-\Delta_\pi},
\end{aligned}
$$

which goes to zero since $\delta_{*k}\to 0$ and $\Delta_\pi\to 0$ as $\sigma\to 0$. For small σ and hence large R_0 and R_1, the ratios $\bar{\mu}(\mathcal{V}^\star)/\tilde{\mu}(\mathcal{V}^\star)$ and

$$
\frac{\tilde{\mu}(B_R)}{\tilde{\mu}(\mathcal{V}^\star)} = \mathcal{PTN}_{p_0}\big(B_2(0,R_0);\widetilde{\Omega}_{00}^{-1}\Omega_{00}a_0,\widetilde{\Omega}_{00}^{-1},p_0^\star,\alpha_0\big)\prod_{j=1}^{p_1}\Gamma\big((0,R_1);\alpha_{1,j},\tilde{a}_j\big)
$$

are close to 1. Therefore, $\|\mathbb{P}_{\mathcal{S}(\theta-\theta^\star)|Y}\mathbf{1}_{B_R}-\mu^\star\mathbf{1}_{B_R}\|_{\mathrm{TV}}\to 0$ as $\sigma\to 0$.

The second term in (13) is bounded by $\|\mu^\star-\mu^\star\mathbf{1}_{B_R}\|_{\mathrm{TV}}\le 2\mu^\star(\overline{B_R})\to 0$ as $R_0, R_1\to\infty$, since the set B_R converges to $\mathcal{V}^\star$ by Lemma 2.

The third term in (13) is bounded by

$$
\|\mathbb{P}_{(\mathcal{S}(\theta-\theta^\star)|Y)}\mathbf{1}_{B_R}-\mathbb{P}_{\mathcal{S}(\theta-\theta^\star)|Y}\|_{\mathrm{TV}}\le 2\mathbb{P}_{\mathcal{S}(\theta-\theta^\star)|Y}(\overline{B_R})\le\frac{2\Delta_0(\delta)}{C_\pi(1-\Delta_\pi)\tilde{\mu}(B_R)},
$$

where $\Delta_0(\delta)$ is defined by (6). By Assumption L, with probability $\to 1$, $\Delta_0(\delta)\to 0$ as $\sigma\to 0$; also, $\tilde{\mu}(B_R)\to\mu_0(\mathcal{V}^\star)>0$.

Combining these bounds, we have that on $\mathcal{A}_0\cap\mathcal{A}_1$,

$$
\begin{aligned}
&\|\mathbb{P}_{\mathcal{S}(\theta-\theta^\star)|Y}-\mu^\star\|_{\mathrm{TV}}\\
&\quad\le 2\mu^\star(\overline{B_R})+2\big[C_\pi(1-\Delta_\pi)\tilde{\mu}(B_R)\big]^{-1}\Delta_0(\delta)\\
&\qquad+2\frac{\bar{\mu}(\mathcal{V}^\star)}{\tilde{\mu}(B_R)}\frac{(1+\Delta_\pi)}{(1-\Delta_\pi)}\Bigg[\delta_{*0}E_\Phi+\delta_{*1}\sum_{j=1}^{p_1}(\alpha_{1,j}/\bar{a}_j)\Bigg]+\frac{4\Delta_\pi}{(1-\Delta_\pi)}\to 0
\end{aligned}
$$

and $\mathbb{P}_{\theta^\star,\sigma}(\mathcal{A}_0 \cap \mathcal{A}_1) \to 1$ as $\sigma \to 0$ due to Assumption S, which gives the statement of the theorem. $\square$

PROOF OF PROPOSITION 1. In the proof of Theorem 1, we derived the following upper bound on event $\mathcal{A}$:

$$\|\mathbb{P}_{\mathcal{S}(\theta-\theta^\star)|Y} - \mu^\star\|_{\mathrm{TV}} \leq 2\mu^\star(\overline{B_R}) + C_\Delta \Delta_0(\delta) + 2C_0\delta_{*0} + 2C_1\delta_{*1} + C_2\Delta_\pi,$$

where $C_\Delta = 2[C_\pi(1-\Delta_\pi)\tilde{\mu}(B_R)]^{-1}$, $C_2 = 4/(1-\Delta_\pi)$, $C_0 = C_A E_\Phi$ with E_Φ defined by (15), $C_1 = C_A \sum_{j=1}^{p_1} \alpha_{1,j}/(a_{1,j} - \delta_{*1})$ and with $B_{R,0} = B_{2,p_0}(0, R_0) \cap \mathcal{V}_0$,

$$C_A = \frac{\bar{\mu}(\mathcal{V}^\star)}{\tilde{\mu}(B_R)} \frac{(1+\Delta_\pi)}{(1-\Delta_\pi)} = \frac{\bar{\mu}_{p_0}(\mathcal{V}_0)}{\tilde{\mu}_{p_0}(B_{R,0})} \prod_{j=1}^{p_1} \left[\frac{a_{1,j}+\delta_{*1}}{a_{1,j}-\delta_{*1}}\right]^{\alpha_{1,j}} \frac{(1+\Delta_\pi)}{(1-\Delta_\pi)},$$

where $\mu_{p_0}(\mathcal{B}_0) = \int_{\mathcal{B}_0 \times [0,\infty)^{p_1}} \mu(dv)$ for a measure μ, $\mathcal{B}_0 \subset \mathcal{V}_0$. If $S_0^\star = \varnothing$,

$$E_\Phi = \|\overline{\Omega}_{00}^{-1}\Omega_{00}a_0\|^2/2 + \operatorname{trace}(\overline{\Omega}_{00}^{-1})/2,$$

$$\frac{\bar{\mu}_{p_0}(\mathcal{V}_0)}{\tilde{\mu}_{p_0}(B_{R,0})} = \frac{e^{\delta_{*0}a_0^T\Omega_{00}\overline{\Omega}_{00}^{-1}\widetilde{\Omega}_{00}^{-1}\Omega_{00}a_0}[\det(\overline{\Omega}_{00}^{-1}\widetilde{\Omega}_{00})]^{1/2}}{\mathcal{TN}(B_{R,0}; \widetilde{\Omega}_{00}^{-1}\Omega_{00}a_0, \widetilde{\Omega}_{00}^{-1})}.$$

We bound the term $\mu^\star(\overline{B_R})$ by

$$\begin{aligned}\mu^\star(\overline{B_R}) &= 1 - \mu_{p_0}^\star(B_{R,0}) \prod_{j=1}^{p_1} \Gamma\left(\left(0, \frac{\delta_1}{\sigma^2}\right); \alpha_{1,j}, a_{1,j}\right) \\ &\leq \mu_{p_0}^\star(\overline{B_{R,0}}) + 1 - \prod_{j=1}^{p_1} \Gamma\left(\left(0, \frac{\delta_1}{\sigma^2}\right); \alpha_{1,j}, a_{1,j}\right)\end{aligned}$$

using the inequality $1 - xy \leq 1 - x + 1 - y$ for $x, y \in (0,1)$. We can also use

$$\begin{aligned}1 - \prod_{j=1}^{p_1} \Gamma\left(\left(0, \frac{\delta_1}{\sigma^2}\right); \alpha_{1,j}, a_{1,j}\right) &\leq p_1\left[1 - \min_j \Gamma\left((0, \delta_1/\sigma^2); \alpha_{1,j}, a_{1,j}\right)\right] \\ &= p_1 \max_j \Gamma\left((\delta_1/\sigma^2, \infty); \alpha_{1,j}, a_{1,j}\right),\end{aligned}$$

and, changing to polar coordinates and denoting $p_{\alpha 0} = p_0 + \sum_{j \in T_0^\star}(\alpha_{0,j} - 1)$ and $W = \{w \in \mathbb{R}^{p_0} : \|w\|_2^2 = 1, w_j > 0 \text{ for } j \in T_0^\star\}$, we have

$$\begin{aligned}\mu_{p_0}^\star(\overline{B_{R,0}}) &\leq \mu_0(\mathcal{V}^\star) \int_{R_0}^\infty r^{p_{\alpha 0}-1} e^{-\lambda_{\min}(\Omega_{00})(r-\|a_0\|)^2/2}\, dr \int_W \prod_{j \in T_0^\star} w_j^{\alpha_{0,j}-1}\, dw \\ &\leq C_{\alpha 0}\Gamma\left(\left((\delta_0/\sigma - \|a_0\|)^2/2, \infty\right); p_{\alpha 0}/2, \lambda_{\min}(\Omega_{00})\right)\end{aligned}$$

under the assumption that $R_0 = \delta_0/\sigma > \|a_0(\omega)\|$ where

$$C_{\alpha 0} = \mu_0(\mathcal{V}^\star) 2^{-p_0^\star + 1.5 p_{\alpha 0}} [\lambda_{\min}(\Omega_{00})]^{p_{\alpha 0}/2} \pi^{(p_0 - p_0^\star)/2} \prod_{i \in T_0^\star} \Gamma(\alpha_{0,i}/2).$$

Collecting conditions on δ_k used in the proof of Theorem 1, we have conditions (10). Thus, we have the required inequality on the event $\mathcal{A}$. $\square$

A.2. Auxiliary results.

PROOF OF LEMMA 2. Due to Assumption B and the fact that $\theta^\star_{S_0^\star \cup S_1} = 0$, the set $D^{-1}U(\Theta^\star(\delta) - \theta^\star)$ contains

$$\begin{aligned} &B_{2,p_0}\left(0, \frac{\delta_0}{\sigma}\right) \times B_{\infty, p_1}\left(0, \frac{\delta_1}{\sigma^2}\right) \cap \left(-\frac{c_0}{\sigma}, \frac{c_0}{\sigma}\right)^{p_0 - p_0^\star} \times \left[0, \frac{c_0}{\sigma}\right)^{p_0^\star} \times \left[0, \frac{c_1}{\sigma^2}\right)^{p_1} \\ &\quad = \{v : v \in B_{2,p_0}(0, \delta_0/\sigma) \text{ and } v_{T_0^\star} \ge 0\} \times [0, \delta_1/\sigma^2)^{p_1}, \end{aligned}$$

where $T_0^\star = \{p_0 - p_0^\star + 1, \ldots, p_0\}$. These sets monotonically increase to $\mathcal{V}^\star = \mathbb{R}^{p_0 - p_0^\star} \times \mathbb{R}_+^{p_0^\star + p_1}$ as $\sigma \to 0$ due to the assumption $\delta_0/\sigma \to \infty$ and $\delta_1/\sigma^2 \to \infty$; this implies the statement of the lemma. $\square$

PROOF OF LEMMA 3. Under the assumptions of the lemma, for small enough σ, with $\tilde{\delta}_0 = \delta_0/\sqrt{p_0}$, we have that

$$\begin{aligned} &\frac{1}{C_{\pi 0}(\delta)} \int_{\Theta \setminus \Theta^\star(\delta)} e^{(\ell_y(\theta) - \ell_y(\theta^\star))/\sigma^2} \pi(d\theta) \\ &\quad \le \sum_{j \in S_0} \int_{\tilde{\delta}_0}^\infty e^{-C_{\delta 0} v_j/\sigma^2} \, dv_j \\ &\qquad + \sum_{j \in S_0 \setminus S_0^\star} \int_0^{\theta_j^\star - \tilde{\delta}_0} \theta_j^{\boldsymbol{\alpha}_j - 1} e^{-C_{\delta 0}|\theta_j - \theta_j^\star|/\sigma^2} \, d\theta_j \\ &\qquad + \sum_{j \in S_1} \int_{\delta_1}^\infty e^{-C_{\delta 1} v_j/\sigma^2} \, dv_j \\ &\quad \le \sum_{j \in S_0 \setminus S_0^\star} \sigma^{\boldsymbol{\alpha}_j} e^{-C_{\delta 0}(\theta_j^\star - \sigma)/\sigma^2} + \frac{p_0 \sigma^2}{C_{\delta 0}} e^{-C_\delta \delta_0/[\sqrt{p_0}\sigma^2]} + \frac{p_1 \sigma^2}{C_{\delta 1}} e^{-C_{\delta 1}\delta_1/\sigma^2} \\ &\qquad + \sum_{j \in S_0 \setminus S_0^\star} [\sigma^{\boldsymbol{\alpha}_j - 1} I(\boldsymbol{\alpha}_j < 1) + \theta_j^{\star \boldsymbol{\alpha}_j - 1} I(\boldsymbol{\alpha}_j \ge 1)] \frac{\sigma^2}{C_{\delta 0}} e^{-C_{\delta 0}\tilde{\delta}_0/\sigma^2} \\ &\quad \le C[\sigma^{\min_j(\boldsymbol{\alpha}_j)} + \sigma] e^{-C_{\delta 0}\delta_0/[\sqrt{p_0}\sigma^2]} + p_1 e^{-C_{\delta 1}\delta_1/\sigma^2} \sigma^2 / C_{\delta 1} \end{aligned}$$

for a constant C. This implies that, with $J_\sigma = \sigma^{-\sum_{j\in S_0}\boldsymbol{\alpha}_j - 2\sum_{j\in S_1}\boldsymbol{\alpha}_j}$,

$$\Delta_0(\delta) \le C_{\pi 0}(\delta) J_\sigma \left[C\left[\sigma^{\min_j(\boldsymbol{\alpha}_j)} + \sigma\right] e^{-C_{\delta 0}\delta_0/[\sqrt{p_0}\sigma^2]} + \frac{p_1\sigma^2}{C_{\delta 1}} e^{-C_{\delta 1}\delta_1/\sigma^2}\right] \to 0$$

as $\sigma \to 0$ under the assumptions of the lemma. $\square$

REFERENCES

BARRON, A., SCHERVISH, M. J. and WASSERMAN, L. (1999). The consistency of posterior distributions in nonparametric problems. *Ann. Statist.* **27** 536–561. MR1714718

BERTSEKAS, D. P. (2003). *Convex Analysis and Optimization*. Athena Scientific and Tsinghua Univ. Press, Belmont, MA.

BESAG, J. (1986). On the statistical analysis of dirty pictures. *J. Roy. Statist. Soc. Ser. B* **48** 259–302. MR0876840

BOCHKINA, N. (2013). Consistency of the posterior distribution in generalized linear inverse problems. *Inverse Problems* **29** 095010, 43. MR3094485

CHERNOZHUKOV, V. and HONG, H. (2004). Likelihood estimation and inference in a class of nonregular econometric models. *Econometrica* **72** 1445–1480. MR2077489

DOUC, R., MOULINES, E., OLSSON, J. and VAN HANDEL, R. (2011). Consistency of the maximum likelihood estimator for general hidden Markov models. *Ann. Statist.* **39** 474–513. MR2797854

DUDLEY, R. M. and HAUGHTON, D. (2002). Asymptotic normality with small relative errors of posterior probabilities of half-spaces. *Ann. Statist.* **30** 1311–1344. MR1936321

ERKANLI, A. (1994). Laplace approximations for posterior expectations when the mode occurs at the boundary of the parameter space. *J. Amer. Statist. Assoc.* **89** 250–258. MR1266297

GEMAN, S. and GEMAN, D. (1984). Stochastic relaxation, gibbs distributions, and the Bayesian restoration of images. *IEEE Trans. Pattern Anal. Mach. Intell.* **6** 721–741.

GHOSAL, S., GHOSH, J. K. and SAMANTA, T. (1995). On convergence of posterior distributions. *Ann. Statist.* **23** 2145–2152. MR1389869

GHOSAL, S. and SAMANTA, T. (1995). Asymptotic behaviour of Bayes estimates and posterior distributions in multiparameter nonregular cases. *Math. Methods Statist.* **4** 361–388. MR1372011

GHOSH, J. K., GHOSAL, S. and SAMANTA, T. (1994). Stability and convergence of the posterior in non-regular problems. In *Statistical Decision Theory and Related Topics, V* (*West Lafayette, IN*, 1992) (S. S. Gupta and J. O. Berger, eds.) 183–199. Springer, New York. MR1286304

GREEN, P. J. (1990). Bayesian reconstructions from emission tomography data using a modified EM algorithm. *IEEE Trans. Med. Imag.* **9** 84–93.

HIRANO, K. and PORTER, J. R. (2003). Asymptotic efficiency in parametric structural models with parameter-dependent support. *Econometrica* **71** 1307–1338. MR2000249

IBRAGIMOV, I. A. and HAS'MINSKIĬ, R. Z. (1981). *Statistical Estimation: Asymptotic Theory*. Springer, New York. MR0620321

JOHNSTONE, I. M. and SILVERMAN, B. W. (1990). Speed of estimation in positron emission tomography and related inverse problems. *Ann. Statist.* **18** 251–280. MR1041393

LECAM, L. (1953). On some asymptotic properties of maximum likelihood estimates and related Bayes' estimates. *Univ. California Publ. Statist.* **1** 277–329. MR0054913

LE CAM, L. and YANG, G. L. (1990). *Asymptotics in Statistics: Some Basic Concepts*. Springer, New York. MR1066869

MORAN, P. A. P. (1971). Maximum-likelihood estimation in non-standard conditions. *Proc. Cambridge Philos. Soc.* **70** 441–450. MR0290493

NELDER, J. A. and WEDDERBURN, R. W. M. (1972). Generalized linear models. *J. R. Stat. Soc. A, General* **135** 370–384.

PETRONE, S., ROUSSEAU, J. and SCRICCIOLO, C. (2012). Bayes and empirical Bayes: Do they merge? *Biometrika* **99** 1–21.

POLYAK, B. T. (1983). *Introduction to Optimization* (*Vvedenie v Optimizatsiyu*, *in Russian*). Nauka, Moscow. MR0719196

SELF, S. G. and LIANG, K.-Y. (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *J. Amer. Statist. Assoc.* **82** 605–610. MR0898365

SHAPIRO, A. (2000). On the asymptotics of constrained local M-estimators. *Ann. Statist.* **28** 948–960. MR1792795

SPOKOINY, V. (2012). Parametric estimation. Finite sample theory. *Ann. Statist.* **40** 2877–2909. MR3097963

VAN DER VAART, A. W. (1998). *Asymptotic Statistics*. Cambridge Univ. Press, Cambridge. MR1652247

VU, H. T. V. and ZHOU, S. (1997). Generalization of likelihood ratio tests under nonstandard conditions. *Ann. Statist.* **25** 897–916. MR1439327

WEIR, I. S. (1997). Fully Bayesian reconstructions from single-photon emission computed tomography data. *J. Amer. Statist. Assoc.* **92** 49–60.

SCHOOL OF MATHEMATICS
UNIVERSITY OF EDINBURGH
EDINBURGH EH9 3JZ
UNITED KINGDOM
E-MAIL: N.Bochkina@ed.ac.uk

SCHOOL OF MATHEMATICS
UNIVERSITY OF BRISTOL
BRISTOL BS8 1TW
UNITED KINGDOM
E-MAIL: P.J.Green@bristol.ac.uk