Submitted to the Annals of Statistics

THE BERNSTEIN–VON MISES THEOREM FOR NON-REGULAR GENERALISED LINEAR INVERSE PROBLEMS

By Natalia Bochkina[†,¶] and Peter J. Green[§,¶]

University of Edinburgh[†] and Maxwell Institute[†], University of Bristol[§] and University of Technology, Sydney[§]

We consider a broad class of nonlinear statistical inverse problems from a Bayesian perspective. This provides a flexible and interpretable framework for their analysis, but it is important to understand the relationship between the chosen Bayesian model and the resulting solution, especially in the ill-posed case where in the absence of prior information the solution is not unique.

Following earlier work about consistency of the posterior distribution of the reconstruction, we obtain approximations to the posterior distribution in the form of a Bernstein–von Mises theorem for nonregular Bayesian models. Emission tomography is taken as a canonical example for study, but our results hold for a wider class of generalised linear models with constraints.

1. Introduction. Inverse problems are almost ubiquitous in applied science and technology, and because of the need for rigorous analysis to characterise such problems, derive numerical solutions and assess their performance, not to mention intrinsic mathematical interest, they have long been the subject of intense mathematical study. In the corresponding 'direct problem', (macroscopic, global) observational data are predicted from the (microscopic, local) model parameters of the system. In the inverse problem conclusions about model parameters are inferred from data.

This paper is a contribution to the theory of inverse problems from a Bayesian perspective. Motivated by important problems in tomographic reconstruction, taken as a canonical example, we consider asymptotic properties of Bayesian procedures in the small-noise limit, for a class of models that we call generalised linear inverse problems.

1.1. *Inverse problems from a Bayesian perspective.* Inverse problems encountered in nature are commonly ill-posed: their solutions fail to satisfy at

[¶]Both authors acknowledge financial support for research visits provided by the EPSRC-funded SuSTaIn programme at Bristol University.

Keywords and phrases: approximate posterior, Bayesian inference, ill-posed inverse problem, non-regular model, SPECT, tomography, total variation distance

least one of the three desiderata of existing, being unique, and being stable. Thus, the focus is not on a unique solution $x \in \mathbb{R}^p$ of

$$\mathcal{A}(x) = y_{\text{exact}}, \tag{1}$$

for given function $\mathcal{A}$ and data vector $y_{\text{exact}} \in \mathbb{R}^n$, but rather on the corresponding set of solutions. Even when the solution x to (1) exists and is unique for each possible y_{exact}, lack of stability means that the solution can be extremely sensitive to small errors, either in the observations or in computations. To circumvent this, the inverse problem is typically regularised, that is, re-formulated to include additional criteria, such as smoothness of the solution:

$$x^\star = \text{argmin}_{\mathcal{A}(x)=y_{\text{exact}}} \text{pen}(x),$$

where $\text{pen}(x)$ is a suitable scalar penalty function. If the inverse problem is ill-posed, the regularised solution $x^\star$ may differ from the actual value x_{true} that generated the data $y_{\text{exact}} = \mathcal{A}(x_{\text{true}})$.

If the data is observed with error, for example if we observe y modelled as a random variable with p.d.f. or p.m.f. $p(y \mid y_{\text{exact}})$ then, allowing for the possibility of lack of existence or uniqueness, the likelihood is penalised, and a commonly considered solution of the inverse problem is that maximising the penalised likelihood, that is,

$$\hat{x} = \text{argmin}_x \left[-\log p(y \mid \mathcal{A}(x)) + \lambda \, \text{pen}(x) \right], \tag{2}$$

with λ a positive constant controlling the trade-off between accuracy and smoothness.

Penalised least squares was one of the first approaches of this kind in inverse problems. While often natural, this corresponds to a Gaussian likelihood, which may not always be appropriate. For instance, Dupé *et al.* (2011) study inverse problems where the observations are counts and where a Poisson likelihood would have been more appropriate.

We now discuss the penalisation. Smoothness, or other 'regular' behaviour of the solution to an inverse problem, is a prior assumption on the unknown x, information about the model parameters known or assumed before the data are observed. To use such information thus accepts that the required solution must combine data with prior information. In a statistical context it is then natural to follow the Bayesian paradigm.

From this perspective, the solution to (2) is immediately recognisable as the maximum a posteriori (MAP) estimate of x, the mode of its posterior distribution in a Bayesian model with likelihood $p(y \mid \mathcal{A}(x))$, and in which the prior distribution of x has density proportional to $\exp\{-\lambda \text{pen}(x)\}$.

However, the Bayesian perspective brings more than a different characterisation of a familiar numerical solution. Formulating a statistical inverse problem as one of inference in a Bayesian model has great appeal, notably for what this brings in terms of coherence, the interpretability of regularisation penalties, the integration of all uncertainties, and the principled way in which the set-up can be elaborated to encompass broader features of the context, such as measurement error, indirect observation, etc. The Bayesian formulation comes close to the way that most scientists intuitively regard the inferential task, and in principle allows the free use of subject knowledge in probabilistic model building (see, for instance, Rover *et al.* (2007) and Davis *et al.* (1995)). For an interesting philosophical view on inverse problems, falsification, and the role of Bayesian argument, see Tarantola (2006).

Mathematical analysis of nonlinear inverse problem (1) is typically far more difficult and technical than for the linear case $\mathcal{A}(x) = Ax$. However, a modest generalisation is enough to formulate and analyse a broad range of nonlinear statistical inverse problems of considerable practical importance. The model class we consider – that of generalised linear inverse problems – is formally defined in Section 3.

1.2. *Convergence of the posterior distribution.* The main mathematical focus in inverse problems concerns how well the true solution can be recovered in the presence of noise, as the size of that noise goes to zero. In the case of a Bayesian analysis, the focus is on the small-variance asymptotic behaviour of the posterior distribution of x.

For a differentiable identifiable likelihood and prior distribution positive and continuous at the "true" value of the parameter, the posterior distribution is asymptotically Gaussian in the case where the "true" parameter is an interior point of the parameter space. This result is known as the Bernstein–von Mises theorem. van der Vaart (1998) gives a total variation distance version of the theorem, adapted from Le Cam (1953) and Le Cam and Yang (1990), under mild additional assumptions on the error model. The theorem implies that, under the above conditions, 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.

However, for our motivating example of the Poisson inverse problem, and, more generally for the class of models we consider, the assumptions of the theorem do not hold. Firstly, the assumption of identifiability of the likelihood may not hold if the inverse problem is ill-posed. Secondly, the assumption that the "true" value of the parameter is interior to the parameter space

may not be satisfied. For the tomography example, the unknown parameter is the vector of image intensities, which are nonnegative and can be zero, corresponding to holes in an organ for example.

Investigating the literature showed that there has been little study of the asymptotics of the posterior distribution for so called *nonregular models*, i.e. models where the above assumptions are not satisfied. Theoretical foundations for the study of the models where the Bernstein–von Mises theorem's assumption of the existence of the first derivative of the loglikelihood is violated were laid by Ibragimov and Has'minskij (1981), for the case of a one-dimensional parameter. These authors considered two types of densities, those with jumps and those with singularities. They gave expressions for distributions approximating the posterior, differing from the Gaussian in both cases, and for the rates of contraction of the posterior distribution (and hence for the correct order rescaling of the parameter) that also differs from the $1/\sqrt{n}$ obtained under the regular assumptions. There were further developments in this area by Ghosal and Samanta (1995), Ghosh *et al.* (1994), and Ghosal *et al.* (1995); these extend the results of Ibragimov and Has'minskij (1981) to i.i.d. models with a regular nuisance parameter (Ghosal and Samanta 1995), where the joint approximating distribution asymptotically factorises into the approximating distribution for the "nonregular" one-dimensional parameter and a Gaussian distribution for the regular nuisance parameter. For densities with jumps, when the limit exists it is a shifted exponential distribution for the recentred "nonregular" parameter, with the rate of contraction $1/n$; for densities with singularities the limiting distribution is more complex.

Under the conditions of Ibragimov and Has'minskij (1981), Ghosal *et al.* (1995) proved the existence of the limit of the posterior for the appropriately centred and rescaled p-dimensional parameter, without specifying the limit explicitly in a general setting. Ghosh *et al.* (1994) characterised the limit of the posterior distribution (for i.i.d. observations) in the particular case where the posterior distribution can be "properly centred". Such a setting applies to the regular case, to densities with jumps or singularities, as considered in Ibragimov and Has'minskij (1981), who showed that for densities with jumps the limit of the posterior does not always exist. Chernozhukov and Hong (2004) considered a class of nonlinear regression models with additive errors, where the error density has a jump at 0, that arise in econometric applications. The authors showed that the limit of the appropriately rescaled posterior distribution was a product of shifted multivariate exponential distributions for the "nonregular" parameter and a Gaussian distribution for the "regular" nuisance parameter of the distribution of errors. A particular

case of this model where the error density has support on the non-negative semiline was considered by Hirano and Porter (2003).

In this paper, we extend the Bernstein–von Mises theorem in two directions, by relaxing both the assumption of identifiability of the likelihood and the assumption that the "true" value of the parameter is interior to the parameter space. We consider a broad class of probability distributions for the data, generalised linear inverse problems, allowing the likelihood to be unidentifiable, and a broad class of prior distributions. We allow linear constraints on the solution of the inverse problem and allow the solution of the exact linear inverse problem to be on the boundary.

We will show that for these models a consequence of relaxing these two assumptions is that the limit of the posterior distribution, as well as the rate of convergence, depend on the choice of the prior distribution and that the limiting distribution is a product of Gaussian and exponential in different directions. We identify the directions in parameter space where the posterior distribution contracts at different rates. We also show how to derive approximations for Bayesian estimators for a given loss function, how to study asymptotic distribution of functionals of the parameter, and how these can be used in practice.

We motivate our study by presenting in Section 2 a nonlinear inverse problem important in medical imaging, and Section 3 establishes the class of models we study. In Section 4 we study geometry of the parameter space determined by the posterior distribution, with an illustration for the linear inverse problems with Gaussian likelihood and a Gaussian prior. In Section 5 we study local behaviour of the posterior distribution in a neighbourhood of the limit that is formulated as a version of Bernstein–von Mises theorem that is illustrated on the motivating example in Section 6. We conclude with a discussion. All proofs are deferred to the Appendix.

2. Motivation. In this section we consider an important example motivating the class of models studied in this paper, generalised linear inverse problems.

2.1. *Single photon emission computed tomography.* Single photon emission computed tomography (SPECT) is a medical imaging technique in which a radioactively-labelled substance, 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 radioactive isotope in the tissue from these counts.

The Poisson linear model

$$y_t \sim \text{Poisson}(A_t x) \tag{3}$$

independently for different t, is close to reality for the SPECT problem (there are some dead-time effects and other artifacts in recording). Here x represents the spatial distribution of the isotope, typically discretised on a grid, $x = \{x_s\}$, and y the array of detected photons, also discretised $y = \{y_t\}$ by the recording process. The array $A = (a_{ts})$ quantifies the emission, transmission, attenuation, decay and recording process; a_{ts} is the mean number of photons recorded at t per unit concentration at pixel/voxel s.

See Green (1990) for further detail about the model, and an approach based on EM estimation for MAP reconstruction of x, in a Bayesian formulation in which spatial smoothness of the solution is promoted by using a pairwise difference Markov random field prior. Later, Weir (1997) investigated fully Bayesian reconstruction.

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 since A is ill-posed we can call this a generalised linear inverse problem.

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 is extended by a factor $\mathcal{T}$, and then consider the *rate* of detection of photons, letting $\mathcal{T} \to \infty$. Thus the data-generation model becomes

$$\mathcal{T} Y_t | x_{\text{true}} \sim \text{Poisson}(\mathcal{T} A_t x_{\text{true}}),$$

independently, for $t = 1, 2, \ldots, n$.

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

In common with much of the literature, we will concentrate here on Markov random field prior distributions. The 'true image' x_{true} in emission tomography corresponds to a physical reality, the discretised spatial distribution of concentration of a radioactive isotope. Of course, this is non-negative, so we impose the constraints $x \geq 0$ (interpreted componentwise), written $x \in \mathcal{X} = [0, \infty)^p \subset \mathbb{R}^p$.

The first prior model we consider is Gaussian, apart from possible truncation by the constraint,

$$p(x) \propto \exp\left\{-\frac{1}{2\gamma^2}||x - x_0||_B^2\right\}, \quad x \in \mathcal{X},$$

where $||u||_B^2 = u^T B u$ and B is a non-negative definite matrix. An important special case is where $x_0 = 0$ and B satisfies $B_{ss'} = 1$ if s and s' are neighbouring pixels (written $s \sim s'$), otherwise $B_{ss'} = 0$. Then we have $||x - x_0||_B^2 = \sum_{s\sim s'}(x_s - x_{s'})^2$, a *pairwise-interaction* model. In this and other important cases B is singular.

A second prior model is a log cosh pairwise-interaction Markov random field (Green 1990):

$$p(x) \propto \exp\left(-\frac{\delta(1+\delta)}{2\gamma^2}\sum_{s\sim s'}\log\cosh((x_s - x_{s'})/\delta)\right), \quad x \in \mathcal{X}. \tag{4}$$

Here the parameter δ is considered to be fixed.

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

These distributions are improper since they are invariant to perturbing x by an arbitrary additive constant, but lead to proper posterior distributions, save in exceptional pathological circumstances.

3. Model formulation.

3.1. *Generalised linear inverse problems.* Motivated by the emission tomography example, we formulate a general class of inverse problems with similar properties that we call generalised linear inverse problems (GLIP).

We assume that the joint density of the observable responses Y taking values in $\mathcal{Y} \subset \mathbb{R}^n$ (with respect to Lebesgue or counting measure) can be written

$$p(y|\,x) = F(y, Ax, \tau) = C_{y,\tau}\exp\left\{-\frac{1}{\tau}\tilde{f}_y(Ax)\right\}, \quad y \in \mathcal{Y} \tag{5}$$

for some $n \times p$ matrix A. The key feature of these models is that the distribution depends on $x \in \mathcal{X}$ only via Ax, where τ is a scalar dispersion parameter; in the Gaussian model, τ is the variance σ^2. The observed data

are generated from this distribution, with $x = x_{\text{true}}$, and we aim to recover x_{true} as $\tau \to 0$.

We assume a continuous bijective link function $G : \mathcal{Y} \to \mathbb{R}^n$ and write $G(y_{\text{exact}}) = Ax_{\text{true}}$. (In generalised linear models – see Example 3 below – commonly G has identical component functions.)

We make the following assumptions about the error distribution:

1. If $Y \sim F(y, G(\mu_0), \tau)$, then $Y \overset{\mathbb{P}_{x_{\text{true}}}}{\to} \mu_0$ as $\tau \to 0$, for all $\mu_0 \in G^{-1}(A\mathcal{X})$.
2. For all $\mu_0 \in G^{-1}(A\mathcal{X})$, $\tilde{f}_{\mu_0}(\eta)$ has a unique minimum over $A\mathcal{X}$ at $\eta = G(\mu_0)$.

Assumption 1 states that τ is not only the dispersion parameter in the model but also that the distribution of Y contracts to its expected value as $\tau \to 0$. Assumption 2 establishes identifiability of the likelihood with respect to the linear predictor $\eta = Ax$. It is sufficient to assume that these conditions hold for $\mu_0 = y_{\text{exact}}$ where y_{exact} is the "exact" data defined in the Introduction.

A particular case of such models is a linear inverse problem with independent observations, where all A is independent of n and $\tau = 1/n$.

For example, Assumption 1 is not satisfied for the Cauchy distribution (or indeed any distribution with polynomial decay and with scale depending on τ) since the density cannot be cast in the form (5) for any choice of τ. Assumption 1 is satisfied for the power exponential (Subbotin) distributions $F(y, \mu, \sigma) = C_{\sigma,\beta} \exp\{-[(y-\mu)^2]^{\beta/2}/\sigma^\beta\}$ ($\beta > 0$), with $\tau = \sigma^\beta$ and $\tilde{f}_y(\mu) = [(y-\mu)^2]^{\beta/2}$.

Assumption 1 is satisfied by generalised linear models.

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

$$p(y_t; \mu_t, \tau) = \exp\left(\frac{y_t b(\mu_t) - c(\mu_t)}{\tau} + d(y_t, \tau)\right),$$

using the mean parameterisation, for appropriate functions b, c and d characterising the particular distribution family. The parameter τ is a common dispersion parameter shared by all responses. The expectation of this distribution is $\mathbb{E}(y_t; \mu_t, \tau) = \mu_t = c'(\mu_t)/b'(\mu_t)$. Both assumptions are satisfied for this example.

The tomography example given in Section 2 belongs to this class of models, with $\tau = \mathcal{T}^{-1}$, $b(\mu_t) = \log \mu_t$, $c(\mu_t) = \mu_t$, $\mu_t = A_t x$ and $\mathcal{X} = [0, \infty)^p$.

As the link function G is continuous and monotonic, we could consider a linear inverse problem $Ax = \tilde{y}_{\text{exact}}$ where $\tilde{y}_{\text{exact}} = G(y_{\text{exact}})$, $\tilde{Y} = G(Y)$ and $\tilde{\mathcal{Y}} = G(\mathcal{Y})$. Hence, to simplify the notation, we assume below that the link function is the identity.

3.2. *Bayesian formulation of GLIP.* We adopt a Bayesian paradigm, using a prior distribution with density given by

$$p(x) \propto \exp(-g(x)/\gamma^2), \quad x \in \mathcal{X}, \tag{6}$$

where γ^2 is a scalar dispersion parameter for the prior, that may depend functionally on τ; we relate this to the data dispersion parameter τ by $\gamma^2 = \tau/\nu$, and express most of our results below in terms of τ and ν. Thus the posterior distribution satisfies

$$p(x|\, y) \propto \exp(-[\tilde{f}_y(Ax) + \nu\, g(x)]/\tau), \quad x \in \mathcal{X}, \tag{7}$$

where $\tilde{f}_y(Ax)$ was defined by (5).

Denote $f_y(x) = \tilde{f}_y(Ax)$ and $h_y(x) = f_y(x) + \nu\, g(x)$, so that $p(x|\, y) \propto e^{-h_y(x)/\tau}$.

We will assume throughout this paper that $\mathcal{X} = [0, \infty)^p$. We could assume that the parameter x is restricted to an arbitrary convex polyhedron; this could be reduced to $[0, \infty)^p$ by a linear change of variables, and indeed some of the ideas we discuss would hold true for more general subsets of $\mathbb{R}^p$.

We shall also assume that y_{exact} is either an interior or a lower boundary point of $A\mathcal{X}$. Otherwise, if $y_{\text{exact}\,j}$ is an upper boundary point, one can replace $A_{j,}$ with $-A_{j,}$ for the corresponding j. We also assume that matrix A has no zero rows or columns.

We shall use the default norms $||z|| = ||z||_2$ for both vectors and matrices.

The limiting statements are given in terms of $\sigma = \sqrt{\tau}$.

4. Geometrical perspective. In this paper, we study inference for x given observed y, in the limit as a noise parameter $\tau = \sigma^2$ (in the SPECT example, $1/\mathcal{T}$) goes to 0. We generally assume an identity link function, so that y becomes concentrated on Ax_{true} as $\sigma^2 \to 0$.

Because of the ill-posed/ill-conditioned character of the problem, we cannot expect consistency in inference about x_{true} based on the likelihood alone. Even as $\sigma^2 \to 0$, so that y converges to 'exact data' $y_{\text{exact}} = Ax_{\text{true}}$, we will not be able to distinguish between $\{x : Ax = Ax_{\text{true}}\}$.

One of the roles of the prior in the Bayesian approach is to resolve this ambiguity (as well as generally improve reconstruction through 'regularisation', even without $\sigma^2 \to 0$). We recall the 'physical' constraint in the

SPECT problem, that x is componentwise non-negative, that is, $x \in \mathcal{X}$, since it quantifies the isotope concentration.

Insight into the interplay between the possibly ill-posed likelihood and the possibly degenerate prior, and the role of the constraint $x \in \mathcal{X}$ can be obtained from a geometrical view of the problem.

4.1. *Gaussian likelihood and prior.* In this section, we focus on the Gaussian prior $p(x) \propto \exp(-1/(2\gamma^2)||x - x_0||_B^2)$ and Gaussian likelihood $y|x \sim \mathcal{N}(Ax, \sigma^2 I)$.

In the limit as $\sigma^2 \to 0$, we are interested in solutions of $Ax = y_{\text{exact}}$, where $y_{\text{exact}} = Ax_{\text{true}}$, under the influence of the prior $p(x) \propto \exp(-1/(2\gamma^2)||x - x_0||_B^2)$. To obtain convergence to a degenerate limit, we will need $\gamma^2 \to 0$ as well (though, as shown by Hofinger and Pikkarainen (2007) for the case $B = I$, at a slower rate than σ^2).

Thus the posterior is proportional to

$$\exp(-1/(2\sigma^2)||y - Ax||^2 - 1/(2\gamma^2)||x - x_0||_B^2) \text{ subject to } x \in \mathcal{X}.$$

Let us first ignore any constraint on x. By standard manipulations, we can write this posterior as

$$x|y \sim \mathcal{N}\left((A^TA + \nu B)^{-1}(A^Ty + \nu B x_0), \sigma^2(A^TA + \nu B)^{-1}\right), \tag{8}$$

assuming the inverse matrix exists. A rank condition is needed to ensure this, so that the information from the likelihood and prior together define a proper posterior.

PROPOSITION 1. *Suppose that A is a real $n \times p$ matrix, and B a real symmetric non-negative definite $p \times p$ matrix, both possibly of deficient rank. Suppose also that the $p \times 2p$ block matrix $[B : A^TA]$ has full rank p (or equivalently, the rows are linearly independent). Then for all $\nu > 0$, $A^TA + \nu B$ is nonsingular.*

It follows that there exists a nonsingular real matrix P, not necessarily orthogonal, such that P^TBP, P^TA^TAP, and $P^T(A^TA + \nu B)P$ (for all $\nu > 0$) are all diagonal.

Furthermore, there exist well-defined finite non-negative definite matrices C and D with ranks $p - q$ and q respectively, where $q = rank(A)$, such that $\nu(A^TA + \nu B)^{-1} = C + D\nu + o(\nu)$ as $\nu \to 0$.

The last part of the proposition gives us a full description of the posterior variance matrix as $\sigma^2 \to 0$, $\gamma^2 \to 0$ while $\nu = \sigma^2/\gamma^2 \to 0$. In summary, the posterior distribution is Gaussian, with variance scaling differently in

different directions. If q is the rank of A, then asymptotically the variance has q eigenvalues scaling like σ^2 and the remaining $(p-q)$ like the (larger) γ^2. Geometrically, contours of equal posterior density are concentric ellipsoids in $\mathbb{R}^p$.

As $\sigma^2 \to 0$ and $\gamma^2 \to 0$ in such a way that $\nu = \sigma^2/\gamma^2 \to 0$, the posterior converges to the point

$$x^\star = \operatorname{argmin}_{x \in \mathcal{X}: Ax = y_{\text{exact}}} ||x - x_0||_B^2, \tag{9}$$

a point that is uniquely determined under the conditions of Proposition 1.

4.2. *Constrained Gaussian model and KKT theory.* When $\mathcal{X}$ is a proper subset of $\mathbb{R}^p$, the ellipsoidal contours are truncated by the constraints $x \in \mathcal{X}$. In the case of interest in SPECT, where we have simply componentwise non-negativity constraints, the ellipsoids are truncated into the non-negative orthant. As σ^2 and γ^2 become small, there are clear qualitative differences in the impact of this truncation according to whether the centre $(A^TA + \nu B)^{-1}(A^Ty + \nu Bx_0)$ of the ellipsoids lies in the interior of the orthant, on its boundary, or outside it.

Equation (9) is a quadratic programming problem, and could be solved numerically by standard software.

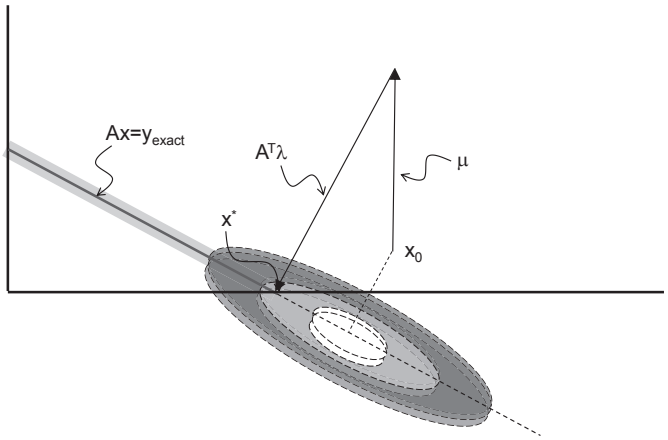

FIG 1. *Illustrating the geometry in the case $p = 2$, $n = 1$, with $B = I$. Contours of posterior when $\gamma^2 > \sigma^2 > 0$.*

We can get a theoretical handle on the solution through Karush–Kuhn–Tucker theory (Kuhn and Tucker 1951). In the non-negativity constrained case, $\mathcal{X} = [0, \infty)^p$, to minimise $||x - x_0||_B^2$ subject to $x \geq 0$ and $Ax = y_{\text{exact}}$

it is necessary and sufficient to find $(x^\star, \mu, \lambda) \in \mathbb{R}^p \times \mathbb{R}^p \times \mathbb{R}^n$ such that

$$\begin{gathered} B(x^\star - x_0) - \mu + A^T\lambda = 0 \\ x^\star \geq 0, \qquad Ax^\star = y_{\text{exact}}, \qquad \mu \geq 0 \\ \text{for all } s, \mu_s = 0 \text{ or } x^\star_s = 0 \end{gathered}$$

The feasible set $\mathcal{X}^\star = \{x \in \mathcal{X} : Ax = y_{\text{exact}}\}$ is closed and convex, and $x^\star$ may be an interior point, or satisfy one or more of the constraints $x_s = 0$.

In the case where all entries of A are non-negative (in accordance with physical reality), and for each s there is at least one t with $A_{ts} > 0$ (and if not, then x_s is unidentifiable, so might as well be omitted from the model), $\mathcal{X}^\star$ is a bounded polyhedron (or polytope). Otherwise, $\mathcal{X}^\star$ may be unbounded.

If γ^2 remains bounded away from 0 as $\sigma^2 \to 0$, then, in the limit, the posterior has support $\mathcal{X}^\star$.

If $x^\star$ is an interior point of $\mathcal{X}^\star$, there exists a neighbourhood of $x^\star$ that lies inside $\mathcal{X}$, and hence, on this neighbourhood, the posterior distribution is not truncated. In this case, as σ^2 and $\gamma^2 \to 0$, the posterior distribution behaves as in the unconstrained case. If $x^\star$ lies on the boundary of $\mathcal{X}^\star$, there are two possibilities: either the unconstrained minimum is achieved at $x^\star$, or outside $\mathcal{X}$. In the first case it is easy to see that the posterior distribution of x recentred by $x^\star$ is "half–Gaussian" (a multivariate Gaussian distribution centred at 0 and truncated at 0 where $x^\star$ is on the boundary). Thus, in a neighbourhood of $x^\star$, the posterior distribution has similar concentration ellipsoids as in the unconstrained case, but truncated at $x^\star$ in some directions (these directions will be defined precisely in Section 4.3). However, in the second case, where the unconstrained solution to the optimisation problem lies outside $\mathcal{X}$, for small σ and γ, the posterior distribution no longer exhibits Gaussian behaviour in the directions orthogonal to the boundary (where $x^\star$ is on the boundary). This is essentially a consequence of the tail behaviour of the Gaussian: if $\xi \sim \mathcal{N}(0,1)$ and $x > 0$,

$$\lim_{t\to\infty} \mathbb{P}(\xi > t + x/t \mid \xi > t) = e^{-x}.$$

The precise formulation of the limit of the posterior distribution is given in the Section 5, in a more general case.

4.3. *Geometry in a general constrained case.* The form of (9) strongly suggests that analogous properties for the limit of the posterior should hold in a much broader class of models. Provided that $\sigma^2 \to 0$ and $\gamma^2 \to 0$ in such a way that $\nu = \sigma^2/\gamma^2 \to 0$, we would expect similar limiting behaviour under the assumptions in Section 3.

In a general setting, more delicate, analytic, arguments will be needed to quantify the convergence precisely. However, for regular problems, the broad qualitative features of the solution for the Gaussian–Gaussian case (Section 4.1) continue to hold: the posterior becomes increasingly concentrated near the hyperplane $\{x : Ax = y_{\text{exact}}\}$, with its variation about this hyperplane controlled by τ, while the variance parallel to the hyperplane is of order γ^2. The effect of the truncation onto $x \in \mathcal{X}$ depends on whether in the absence of the constraint, the maximum of the posterior would lie in the interior of $\mathcal{X}$, on its boundary, or outside it.

Now we describe the local geometry of the posterior distribution around the point $x^\star$, under the assumption that functions $\tilde{f}_{y_{\text{exact}}}$ and g are differentiable, relaxing the assumption that $x^\star$ is an interior point by allowing it to lie on the boundary of $\mathcal{X}$. Such a model is nonregular.

Throughout, we use $\nabla_i = \frac{\partial}{\partial x_i}$ as the derivative operator, and $\nabla = (\nabla_1, \ldots, \nabla_p)^T$ as the gradient. Similarly, ∇_{ij} and ∇_{ijk} are operators of the second and third derivatives, with $\nabla^2 = (\nabla_{ij})$ being the matrix of second derivatives.

In the limit of zero noise, the Bayesian analysis has solved two optimisation problems:

$$\begin{aligned} \mathcal{X}^\star &= \arg \min_{x \in [0,\infty)^p,\, Ax = y_{\text{exact}}} f_{y_{\text{exact}}}(x), \\ x^\star &= \arg \min_{x \in \mathcal{X}^\star} g(x). \end{aligned} \tag{10}$$

We assume that the prior distribution is such that $x^\star$ is a unique solution. Denoting $\eta = Ax$, the first problem can be reformulated as follows:

$$y_{\text{exact}} = \arg \min_{\eta \in A\mathcal{X}} \tilde{f}_{y_{\text{exact}}}(\eta). \tag{11}$$

This condition is the identifiability of the likelihood with respect to $\eta = Ax$. The second expression is the definition of $x^\star$, the point where the posterior distribution concentrates, which depends on the prior distribution.

Now we use the Karush–Kuhn–Tucker (KKT) theory to study the local geometry of the solution. If the solution of the optimisation problem $x^\star$ is an interior point of $\mathcal{X}^\star = \{x \in \mathcal{X} : Ax = y_{\text{exact}}\}$, then

$$0 = \left(\frac{\partial}{\partial z_i} g(x^\star + (I - P_{A^T})z)|_{z=0} \right)_{i=1}^{p} = (I - P_{A^T}) \nabla g(x^\star), \tag{12}$$

where P_{A^T} is the projection on the range of A. However, if $x^\star$ is on the boundary, the gradient $\nabla g(x^\star)$ may not be zero. This corresponds to the

maximum in the unconstrained case lying outside $\mathcal{X}$. In this case, the KKT conditions are:

$$\nabla_j \tilde{f}_{y_{\text{exact}}}(Ax^\star) \geq 0 \quad \& \quad [Ax^\star]_j \, \nabla_j \tilde{f}_{y_{\text{exact}}}(Ax^\star) = 0, \quad j = 1, \dots, n, \tag{13}$$

$$\nabla g(x^\star) = A^T \lambda + \zeta \quad \& \quad \zeta_i x_i^\star = 0, \quad i = 1, \dots, p \tag{14}$$

for some $\lambda \in \mathbb{R}^n$ and $\zeta \in [0, \infty)^p$.

Define the sets of the nonregular boundary components of y_{exact} and of $x^\star$ by

$$S = \{i \in 1, 2, \dots, p : \zeta_i > 0\}, \qquad Z = \{j \in 1, 2, \dots, n : \nabla_j \tilde{f}_{y_{\text{exact}}}(Ax^\star) > 0\},$$
$$S^* = \{i \in 1, 2, \dots, p : x_i^\star = 0\}, \qquad Z^* = \{j \in 1, 2, \dots, n : [Ax^\star]_j = 0\}.$$

By the KKT conditions, $Z \subseteq Z^*$ and $S \subseteq S^*$. If $S \neq S^*$ or $Z \neq Z^*$, the corresponding minimum is achieved on the boundary and the gradient is zero.

In the small noise limit, we will show that the posterior distribution exhibits different types of behaviour on 4 subsets $\mathcal{W}_0, \mathcal{W}_1, \mathcal{W}_2, \mathcal{W}_3$ of $\mathbb{R}^p$ such that $\mathcal{X} - x^\star = \{z = x - x^\star,\, x \in \mathcal{X}\} = \mathcal{W}_0 \oplus \mathcal{W}_1 \oplus \mathcal{W}_2 \oplus \mathcal{W}_3$. These four subsets are determined by $p \times p_k$ matrices V_k of rank p_k: $\mathcal{W}_k = \{\sum_{j=1}^{p_k} [V_k]_{j,} \alpha_j,\, \alpha \in \mathbb{R}^{p_k}\}$ for $k = 0, 1$ and $\mathcal{W}_k = \{\sum_{j=1}^{p_k} [V_k]_{j,} \alpha_j,\, \alpha \in \mathbb{R}_+^{p_k}\}$ for $k = 2, 3$, and p_k are their dimensions, where the matrices V_k satisfy the following conditions:

$$A_{Z,} V_0 = 0, \quad AV_1 = 0, \quad AV_3 = 0$$
$$V_0^T A_{Z^c,}^T A_{Z^c,} V_0 \quad \text{and} \quad V_1^T V_1 \quad \text{are positive definite,}$$
$$V_1{}^T \zeta = 0, \quad V_3{}^T \zeta \ \text{ is a vector with positive coordinates } (\zeta_S \in \mathbb{R}_+^{|S|}),$$
$$V_2^T \nabla f_{y_{\text{exact}}}(x^\star) \ \text{ is a vector with positive coordinates.} \tag{15}$$

where V_0 and V_1 are the matrices of the largest size satisfying the conditions. These conditions imply that

$$p_0 = \operatorname{rank}(A) - \operatorname{rank}(A_{Z,}), \quad p_1 = p - \operatorname{rank}(A^T : \nabla g(x^\star)), \tag{16}$$
$$p_2 = \operatorname{rank}(A_{Z,}), \quad p_3 = \operatorname{rank}(A^T : \nabla g(x^\star)) - \operatorname{rank}(A).$$

Note that p_3 can be either 0 or 1. The four subsets can be characterised as follows:

a) $\mathcal{W}_0$: likelihood is identifiable, projection of $\nabla f_{y_{\text{exact}}}(x^\star)$ on $\mathcal{W}_0$ is 0;
b) $\mathcal{W}_1$: likelihood is not identifiable, projection of $\nabla g(x^\star)$ on $\mathcal{W}_1$ is 0;
c) cone $\mathcal{W}_2$: likelihood is identifiable, projection of $\nabla f_{y_{\text{exact}}}(x^\star)$ on $\mathcal{W}_2$ is nonzero (projection of $x^\star$ on $\mathcal{W}_2$ is on the boundary of $\mathcal{W}_2$);

d) cone $\mathcal{W}_3$: likelihood is not identifiable, projection of $\nabla g(x^\star)$ on $\mathcal{W}_3$ is nonzero (projection of $x^\star$ on $\mathcal{W}_3$ is on the boundary of $\mathcal{W}_3$).

The four matrices V_k define a transform from x to $w = (w_0^T, w_1^T, w_2^T, w_3^T)^T \in \mathbb{R}^{p_0+p_1} \times \mathbb{R}_+^{p_2+p_3}$:

$$x = x^\star + \sum_{k=0}^{3} V_k w_k. \tag{17}$$

LEMMA 1. *If (V_k) satisfy conditions* (15)*, then the matrix $V = (V_0 : V_1 : V_2 : V_3)$ has full rank.*

Now we propose a way to construct the matrices V and V^{-1}.

DEFINITION 1. *Define Z_2 to be a subset of Z such that rank$(A_{Z,}) =$ rank$(A_{Z_2,}) = |Z_2|$, so that for every $j \in Z$, $A_{j,}^T$ can be written as a linear combination of vectors $(A_{j,}^T,\ j \in Z_2)$ with nonnegative coefficients: $A_{Z,} = \alpha A_{Z_2,}$ where $\alpha \in [0,\infty)^{|Z|\times|Z_2|}$.*

Define $Z_0 \subset Z^c$ such that $(A_{j,},\ j \in Z_0)$ are linearly independent and $(A_{j,},\ j \in Z_2 \cup Z_0)$ are linearly independent. In particular, $p_0 = |Z_0| =$ rank$(A) - |Z_2|$.

The subset Z_2 exists by Caratheodory's theorem (p. 37 of Bertsekas (2006)).

Here is one way of constructing the matrix $U = V^{-1}$. Let $V_1 \in \mathbb{R}^{p\times p_1}$ satisfy

$$\begin{aligned}(A^T : \nabla g(x^\star))^T V_1 = 0, \quad V_1^T V_1 = I_{p_1},\\ \operatorname{rank}(V_1) = p - \operatorname{rank}(A^T : \nabla g(x^\star))\end{aligned} \tag{18}$$

such that the matrix $[V_1]_{S^c,}$ is of highest possible rank. Take $U = (U_0^T : U_1^T : U_2^T : U_3^T)^T$ with

$$U_0 = A_{Z_0,}, \quad U_1 = V_1^T, \quad U_2 = A_{Z_2,}, \quad U_3 = \zeta^T. \tag{19}$$

In Proposition 2 below, we give the explicit expression for $V = U^{-1}$, and study the image of $\mathcal{X} - x^\star$ under the transformation U, rescaled as follows.

Define the following scaling transform $\mathcal{S} = \mathcal{S}_{\sigma,\gamma}\colon \mathcal{X} - x^\star \to \mathbb{R}^{p_0+p_1} \times \mathbb{R}_+^{p_2+p_3}$:

$$\mathcal{S}(x - x^\star) = D_{\sigma,\gamma}^{-1} V^{-1}(x - x^\star), \tag{20}$$

where $D_{\sigma,\gamma} = \operatorname{diag}(\sigma I_{p_0}, \gamma I_{p_1}, \sigma^2 I_{p_2}, \gamma^2 I_{p_3})$ and $V = (V_0 : V_1 : V_2 : V_3)$ is defined by (15). This corresponds to rescaling each of the four subsets independently. As we shall see in Theorem 1, this is the appropriate scaling for the posterior distribution.

PROPOSITION 2. *Let the matrix U be defined by* (19) *with V_1 satisfying* (18). *Then,*

1) *the matrix $V = U^{-1}$ satisfies conditions* (15),
2) *the matrices V_0, V_2 and V_3 are defined by*

$$V_3 = \tilde{\zeta}/||\tilde{\zeta}||^2, \quad V_k = (I - \tilde{P}_{m,\zeta})A^T_{Z_k,}(A_{Z_k,}(I - \tilde{P}_{m,\zeta})A^T_{Z_k,})^{-1}, \tag{21}$$

for $(k, m) \in \{(0, 2), (2, 0)\}$, with projections on the range of $A^T_{Z_0,}$ (P_0), on the range of $(I - P_0)A^T_{Z_2,}$ ($\tilde{P}_{2,0}$) and on the range of $(I - P_\zeta)A^T_{Z_k,}$ ($\tilde{P}_{k,\zeta}$):

$$P_0 = P_{A^T_{Z_0,}}, \quad \tilde{P}_{2,0} = (I - P_0)A^T_{Z_2,}(A_{Z_2,}(I - P_0)A^T_{Z_2,})^{-1}A_{Z_2,}(I - P_0),$$
$$P_\zeta = \zeta\zeta^T/||\zeta||^2, \quad \tilde{P}_{k,\zeta} = (I - P_\zeta)A^T_{Z_k,}[A_{Z_k,}(I - P_\zeta)A_{Z_k,}]^{-1}A_{Z_k,}(I - P_\zeta)$$
$$\tilde{\zeta} = (I - \tilde{P}_{2,0})(I - P_0)\zeta$$

(if $p_k = 0$, then the projection matrix that uses the corresponding U_k is zero);

3) *if $Z = Z^*$ and $S = S^*$ and $|S| = p_2 + p_3$, the linear transform $\mathcal{S} = D^{-1}_{\sigma,\gamma}V^{-1}$ maps $\mathcal{X} - x^\star$ onto $\mathbb{R}^{p_0+p_1} \times \mathbb{R}^{p_2+p_3}_+$ under conditions* (29);
4) *more generally, under conditions* (29), *the linear transform $\mathcal{S} = D^{-1}_{\sigma,\gamma}V^{-1}$ maps $\mathcal{X} - x^\star$ onto $\mathcal{V}^\star \subset \mathbb{R}^{p_0+p_1} \times \mathbb{R}^{p_2+p_3}_+$ defined by*

$$\mathcal{V}^\star = \begin{cases} \{(v_0, v_1, v_2, v_3) : [V_k]_{S_k,}v_k \geq 0,\ k = 0, 1, 3\} & \text{if} \quad c = 0, \\ \{(v_0, v_1, v_2, v_3) : c[V_0]_{S_{03},}v_0 + [V_3]_{S_{03},}v_3 \geq 0 \,\&\, [V_1]_{S_1,}v_1 \geq 0\} & \text{if} \quad c > 0, \end{cases}$$

where the inequalities are component-wise, $S_{03} = S_0 \cup S_3$ and

$$\begin{aligned} S_1 &= \{\ell \in S^* : [V_1]_{\ell,} \neq 0\}, \\ S_3 &= \{\ell \in S^* : [V_1]_{\ell,} = 0 \,\&\, [V_3]_\ell \neq 0\}, \\ S_0 &= \{\ell \in S^* : [V_1]_{\ell,} = 0 \,\&\, [V_3]_\ell = 0 \,\&\, [V_0]_{\ell,} \neq 0\}. \end{aligned} \tag{22}$$

In particular, $|S_0| \leq |Z_0 \cap Z^*|$ and S_1 has at most $s = |S^*| - p_2 - p_3$ constraints.

In Section 5 we show that the posterior distribution has different asymptotic behaviour in these four sets of directions. Now we shall look at some examples.

EXAMPLE 2. *Consider the Poisson likelihood with identity link: $Y/\sigma^2 \sim Pois(Ax/\sigma^2)$ ($n = 1$, $p = 2$), with $A = (1, 1)$. We take $x_{\text{true}} = (1, 0)^T$, so that $y^\star = Ax_{\text{true}} = 1$. The linear inverse problem $Ax = y^\star$, i.e. $x_1 + x_2 = 1$, subject to constraints $x_1, x_2 \geq 0$, is ill-posed. To resolve the ambiguity, we use the penalty $||x - x_0||_2$, with two different x_0.*

1. $x_0 = (4,2)^T$. *Then* $x^\star = (1,0)^T$. *In this case,* $p_2 = 0$ *since for* $\eta = y^\star = 1$, $\nabla_\eta \tilde{f}_{y^\star}(\eta) = -1/\eta + 1 = 0$. *Thus,* $p_0 = rank(A) = 1$, *and we take* $U_0 = A = (1,1)$. *The null space of* A *is* $\{\alpha(1,-1)^T, \alpha \in \mathbb{R}\}$. *The gradient of the negative log prior at* $x^\star$ *(up to a factor* $1/\gamma^2$*) is* $-(3,2)^T = -3\,A^T + \zeta$ *where* $\zeta = (0,1)^T$. *The gradient is not orthogonal to the null space, hence* $p_3 = 1$ *and thus* $p_1 = 0$, *with* $U_3 = (0,1)^T$. *The corresponding* V_0 *and* V_3 *are* $V_0 = (1,0)^T$ *and* $V_3 = (-1,1)^T$. *The conditions of item 3) in Proposition 2 are satisfied hence the image of* $\mathcal{X} - x^\star$ *under* $\mathcal{S}$ *in the limit is* $\mathbb{R} \times \mathbb{R}_+$.
2. $x_0 = (3,3)^T$. *Then* $x^\star = (0.5, 0.5)^T$. *Since* $y^\star$ *is unchanged, we again have* $p_2 = 0$. *We can take the same* $U_0 = (1,1)^T$. *The gradient of* g *at* $x^\star$ *here is* $x^\star - x_0 = -2.5\,(1,1)^T$ *and is orthogonal to the null space of* A, $\{\alpha(-1,1)^T, \alpha \in \mathbb{R}\}$. *Therefore, we have* $V_1 = U_1^T = \sqrt{0.5}(-1,1)^T$. *Since the kernel is one-dimensional,* $p_3 = 0$. *Here* $V_0 = 0.5(1,1)^T$. *Here we are again under conditions of item 3) in Proposition 2, hence* $\mathcal{V}^\star = \mathbb{R}^2$.

Further examples can be found in Sections 5.4 and 6.1.

5. Analogue of the Bernstein–von Mises theorem.

5.1. *Assumptions on the likelihood and the prior.* In addition to assuming that we have a GLIP model with $\tau = \sigma^2$, we make the four main assumptions that the posterior distribution is proper, that the log likelihood and log prior density have bounded third order derivatives with respect to x, that the log likelihood and its first two derivatives are continuous with respect to y, and that the posterior distribution is concentrated in a neighbourhood of $x^\star$.

Assumption P.

We assume that the prior distribution is such that the posterior distribution is proper:

$$\exists \sigma_0 > 0: \quad \forall \sigma \leqslant \sigma_0, \quad \int_{\mathcal{X}} e^{-h_y(x)/\sigma^2} dx < \infty \quad \text{for } \mathbb{P}_{y_{\text{exact}}} \text{ almost all} \quad y \in \mathcal{Y}.$$

Assumption S (smoothness in x).

There exist $\delta_k > 0$, $k = 0,1,2,3$, such that there exist uniformly bounded derivatives up to third order: $\exists f_y^{(m)}$, $\exists g^{(m)}$ on $B_\delta(x^\star)$ for $\mathbb{P}_{y_{\text{exact}}}$ almost all $y \in \mathcal{Y}$, $m = 1,2,3$, and $\exists C_{f3}$, $C_{g3} < \infty$ such that for all $x \in B_\delta(x^\star)$, with probability $\to 1$ as $\sigma \to 0$ and all $1 \leqslant i,j,k \leqslant p$,

$$|\nabla_{ijk} f_y(x)| \leqslant C_{f3}, \qquad |\nabla_{ijk} g(x)| \leqslant C_{g3}, \tag{23}$$

where $B_\delta(x^\star) = \{x \in \mathcal{X} : V^{-1}(x - x^\star) \in B_\delta\}$ and $B_\delta = B_2(0,\delta_0) \times B_2(0,\delta_1) \times B_\infty(0,\delta_2) \times B_\infty(0,\delta_3)$.

Assumption C (continuity in y).

The derivatives of $f_Y(x^\star)$ converge to the corresponding derivatives of $f_{y_{\text{exact}}}(x^\star)$ with probability $\to 1$ as $\sigma \to 0$, i.e. for all $1 \leqslant j_1, \ldots, j_d \leqslant p$ with $d = 0, 1, 2$,

$$\nabla_{j_1,\ldots,j_d} f_Y(x^\star) - \nabla_{j_1,\ldots,j_d} f_{y_{\text{exact}}}(x^\star) \to 0 \quad \text{as} \quad \sigma \to 0. \tag{24}$$

These assumptions are satisfied if $\exists \nabla^d_{\mu_0} f_{\mu_0}(x)$ for $d = 1, 2, 3$ and these derivatives are bounded for $\mathbb{P}_{y_{\text{exact}}}$ almost all $\mu_0 \in \mathcal{Y}$.

Assumption L.

For $\delta_k > 0$ defined in Assumption S,

$$\mathbb{P}(\Delta_0(\delta) \to 0) \to 1 \quad \text{as} \quad \sigma \to 0, \tag{25}$$

where

$$\Delta_0(\delta) \;=\; \sigma^{-p_0-2p_2}\gamma^{-\tilde{p}_1-2\tilde{p}_3} \int_{\mathcal{X}\setminus B_\delta(x^\star)} e^{-(h_y(x)-h_y(x^\star))/\sigma^2}\, dx, \tag{26}$$

where $\tilde{p}_k$ is the non-degenerate dimension of $U_k(\mathcal{X} - x^\star)$, $k = 0, \ldots, 3$. For $k = 0$ and $k = 2$, $\tilde{p}_k = p_k$. This assumption implies that for small σ, the posterior distribution is concentrated on $B_\delta(x^\star)$.

5.2. *Notation.* The limiting behaviour of the posterior distribution is characterised by the matrices of second derivatives:

$$\begin{aligned} V_y(x) &= \nabla^2 \tilde{f}_y(Ax), \qquad B(x) = \nabla^2 g(x), \\ H_y(x) &= \nabla^2 h_y(x) = A^T V_y(x) A + \nu B(x), \\ \Omega_{00} &= V_0^T \nabla^2 f_{y_{\text{exact}}}(x^\star) V_0 = V_0^T A^T V_{y_{\text{exact}}}(x^\star) A V_0, \end{aligned}$$

and by the following projections of the gradients:

$$\begin{gathered} a = V_2^T \nabla f_{y_{\text{exact}}}(x^\star), \quad a_y = V_2^T [\nabla f_y(x^\star) + \nu \nabla g(x^\star)], \\ b = V_3^T \nabla g(x^\star), \end{gathered} \tag{27}$$

where $H_{00} = V_0^T H_y(x^\star) V_0$, and $B_{ij} = V_i^T B V_j$, $i, j = 0, 1, 2, 3$. By the conditions (15), a is a vector with positive coordinates, $b > 0$ if $p_3 > 0$, and Ω_{00} and B_{11} are positive definite. For U and V considered in Proposition 2, $b = 1$ if $p_3 > 0$. The smallest components of the vectors a_y and a will be denoted by $a_{y,\min} = \min_i a_{y,i}$ and $a_{\min} = \min_i a_i$ respectively.

5.3. *The Bernstein–von Mises theorem.* In the theorem below, which is an analogue of the Bernstein-von Mises theorem, we show that the posterior distribution converges to a finite limit, after the rescaling and the change of variables defined in Section 4.3. This can be used to approximate the posterior distribution in practice, for small values of σ, and to study asymptotic properties of Bayes estimates.

THEOREM 1. *Consider the Bayesian GLIP model defined in Section 3 such that matrix A has no zero rows or columns, and let Assumptions P, S, C and L on f_y and g stated in Section 5.1 hold. Assume also that, as $\sigma \to 0$,*

$$\frac{\sigma}{\gamma} \to 0, \quad \gamma \to 0, \quad c = \lim_{\sigma \to 0} \frac{\sigma}{\gamma^2} < \infty. \tag{28}$$

For V_k, $k = 0, \dots, 3$, satisfying conditions (15), we assume that $B_{00} - B_{01}B_{11}^{-1}B_{10}$ is positive semi-definite, and that the following limit exists for all ω:

$$a_0(\omega) = \Omega_{00}^{-1} \left[\lim_{\sigma \to 0} [\sigma^{-1} V_0^T \nabla f_{Y(\omega)}(x^\star)] + c V_0^T \nabla g(x^\star) \right].$$

Define a random probability measure on $\mathcal{V}^\star = \lim_{\sigma \to 0} \mathcal{S}(\mathcal{X} - x^\star) \subseteq \mathbb{R}^{p_0+p_1} \times \mathbb{R}_+^{p_2+p_3}$:

$$\mu^\star(\omega) = \{ \mathcal{N}_{p_0} \left(a_0(\omega), \Omega_{00}^{-1} \right) \times \mathcal{N}_{p_1} \left(0, B_{11}^{-1} \right) \times Exp_{p_2}(a) \times Exp_{p_3}(b) \} \, \mathbf{1}_{\mathcal{V}^\star},$$

where transform $\mathcal{S}$ is defined by (20), $Exp_m(v)$ is the distribution of an m-dimensional vector ξ with independent coordinates $\xi_i \sim Exp(v_i)$, and $\mathbf{1}_{\mathcal{V}^\star}$ is the indicator function of set $\mathcal{V}^\star$. If support $\mathcal{V}^\star$ degenerates to a manifold of a smaller dimension, then $\mu^\star(\omega)$ is normalised be a probability measure on this manifold, and it has mass 1 on the degenerate part of $\mathbb{R}^p$.

Then, as $\sigma \to 0$,

$$||\mathbb{P}_{\mathcal{S}(x-x^\star)|Y} - \mu^\star||_{TV} \overset{\mathbb{P}_{x_{\text{true}}}}{\to} 0.$$

If $p_k = 0$ then the corresponding factor in the definition of $\mu^\star$ disappears. In particular, if $x^\star$ is an interior point, the limit is Gaussian distribution with no constraint on its support. If the likelihood is also identifiable, the statement is the classical Bernstein-von Mises theorem.

If $x^\star$ or $Ax^\star$ is on the boundary of its corresponding parameter space, then there is at least one direction where the posterior distribution converges faster than the convergence around an interior point. If the plane $Ax = y_{\text{exact}}$ does not intersect the boundary at the right angles, i.e. if the positivity constraints on the coordinates imply nontrivial constraints after

the change of variables, it is possible to have degeneration of the support of the posterior distribution of $\mathcal{S}(x-x^\star)$ due to mutually exclusive constraints (see Examples 4 and 5).

Assumption of finiteness of $a_0(\omega)$ implies that $V_0^T(\nabla f_{Y(\omega)}(x^\star)-\nabla f_{y_{\text{exact}}}(x^\star))$ not only converges to zero (Assumption C) but also that it converges at rate σ or faster. This holds in the case where Y_j are independent and their distribution belongs to the exponential family, since $\tilde{f}_Y(\eta) = -\sum_{j=1}^n b(\eta_j)a(Y_j) - \sum_{j=1}^n c(\eta_j)$, and its variance is proportional to σ^2.

The assumption of the existence and boundness of the third derivatives of f_y and g can be relaxed. It is sufficient to assume that the supremum of the absolute value of each component of $V_0^T\nabla^2 f_y(x^\star)V_0$, $V_1^T\nabla^2 g(x^\star)V_1$, $V_2^T\nabla f_y(x^\star)$ and $V_3^T\nabla g(x^\star)$ on $B_\delta(x^\star)$ converges to zero as $\delta\to 0$ at the appropriate rate, with probability 1.

We will also state a nonasymptotic bound on the distance between the posterior distribution of the rescaled parameter and its limit.

PROPOSITION 3. *Take $\delta_k > 0$, $k = 0,\dots,3$, satisfying the following conditions*

$$\max[0.5\nu||B_{00}||/\kappa_A, \sigma||a_0(\omega)||] < \delta_0 < \min[0.2\lambda_{\min}(\Omega_{00})/\kappa_A, ||U_0x^\star||],$$
$$\delta_1 < \min[\lambda_{\min}(B_{11})/\kappa_B, ||U_1x^\star||], \quad \delta_k \le ||U_kx^\star||_\infty,\ k=2,3,$$
$$0.5\nu||V_2^T\nabla g(x^\star)|| < \max_{k=0,1,2}[\delta_k c_{2,k}] < 0.2a_{\min}/3,$$
$$\max_{k=0,1,2,3}[\delta_k c_{3,k}] < b_{\min}/4,$$

where constants c_{mk} are defined in the proof, and the inequality $\delta_0 > \sigma||a_0(\omega)||$ holds with high probability.

Define the following events

$$\begin{aligned}
\mathcal{A}_1(\delta) &= \{\omega:\ ||V_2^T[\nabla f_{Y(\omega)}(x^\star) - \nabla f_{y_{\text{exact}}}(x^\star)]||_\infty \le M_1 \max_k \delta_k\},\\
\mathcal{A}_2(\delta) &= \{\omega:\ ||V_0^T[\nabla^2 f_{Y(\omega)}(x^\star) - \nabla^2 f_{y_{\text{exact}}}(x^\star)]V_0|| \le M_2\delta_0\},\\
\mathcal{A}_3 &= \left\{\omega:\ ||\Omega_{00}a_0(\omega) - cV_0^T\nabla g(x^\star) - \sigma^{-1}V_0^T\nabla f_{Y(\omega)}(x^\star)|| \le \rho\right\},
\end{aligned}$$

for some $\rho\to 0$ as $\sigma\to 0$ and positive constants M_1 and M_2.

Then, under the assumptions of Theorem 1, on $\mathcal{A} = \mathcal{A}_1 \cap \mathcal{A}_2 \cap \mathcal{A}_3$,

$$\begin{aligned}
||\mathbb{P}_{\mathcal{S}(x-x^\star)|Y} - \mu^\star||_{TV} \quad \leq \quad & 2p_0 \max\left\{C_0\delta_0,\, 1 - \Gamma\left(\frac{\lambda_{\min}(\Omega_{00})(\delta_0/\sigma - ||a_0||)^2}{2} \mid \frac{p_0}{2}\right)\right\} \\
+ \quad & 2\tilde{p}_1 \max\left\{C_1\delta_1,\, 1 - \Gamma\left(\frac{\lambda_{\min}(B_{11})\delta_1^2}{2\gamma^2} \mid \frac{p_1}{2}\right)\right\} \\
+ \quad & 2p_2 \max\left\{C_2\delta_2,\, \exp\{-a_{\min}\delta_2/\sigma^2\}\right\} \\
+ \quad & 2\tilde{p}_3 \max\left\{C_3\delta_3,\, \exp\{-b\delta_3/\gamma^2\}\right\} \\
+ \quad & C_B\left\{C_4\left[\delta_0/\gamma + \delta_2/\gamma + \delta_3/\gamma\right]^{|S_1|} + C_5[\delta_2/\gamma^2]^{m_5}\right\} \\
+ \quad & C_\Delta \Delta_0(\delta),
\end{aligned}$$

where $\tilde{p}_k$ *is the dimension of the non-degenerate part of* $U_k(\mathcal{X} - x^\star)$, $k = 1, 3$, *and the constants are defined explicitly in the proof.*

The upper bound implies that for the total variation to be small in practical applications, the dimensions p_k should not be too large compared to the corresponding rate, and that the smallest values of the parameters $a_{\min}$, b and the smallest eigenvalues of the precision matrices Ω_{00} and B_{11} cannot be too small, namely cannot be smaller than the corresponding rate.

It is interesting to note that the smallest δ_k satisfying the local constraints (given in the first four lines of the upper bound) coincide with an upper bound on the Ky Fan distance between the posterior distribution and its limit $\delta_{x^\star}$ on the corresponding subspace/cone of the parameter space (Bochkina 2012). Thus, it appears that the Ky Fan distance determines the radius of the largest ball centred at $x^\star$ where the concentration of the posterior distribution can take place.

5.4. *Examples of degenerate support.* It follows from Theorem 1 and Proposition 2, that in some cases the limit of the posterior distribution of $\mathcal{S}(x - x^\star)$ can degenerate due to degeneration of its support.

In a first example we consider the degeneracy where the solution of the unconstrained optimisation problem coincides with that of the constrained problem and occurs on on the boundary of the parameter space.

EXAMPLE 3. *Consider the Gaussian likelihood with the identity link:* $Y \sim N(Ax, \sigma^2)$ *(*$n = 1$*), with* $A = (1, 1)$. *We take* $x_{\text{true}} = (0.8, 0.2)^T$, *so that* $y^\star = Ax_{\text{true}} = 1$. *The linear inverse problem* $Ax = y^\star$, *i.e.* $x_1 + x_2 = 1$, *subject to constraint* $x_1, x_2 \geq 0$, *is ill-posed. To resolve the ambiguity, we use the penalty* $||x - x_0||_2$, *with* $x_0 = (3, 2)^T$. *Then the solution to the*

constrained optimisation problem is $x^\star = (1,0)^T$*, the same as the solution to the unconstrained problem.*

The gradient $\nabla \tilde{f}_{y_{\text{exact}}}(Ax^\star) = -y_{\text{exact}}/(Ax^\star) + 1 = 0$ *implies* $\mathcal{W}_2$ *is empty and* $p_2 = 0$*. The gradient of* g *at* $x^\star$ *is* $x^\star - x_0 = -(2,2)^T = -2\,A^T + \zeta$ *where* $\zeta = (0,0)^T$*. Thus,* $\mathcal{W}_3$ *is also empty even though* $x^\star$ *occurs on the boundary.*

Therefore, we have that $\mathcal{W}_1$ *and* $\mathcal{W}_0$ *are nonempty, with* $U_0 = (1,1)$ *and* $U_1 = V_1^T = \sqrt{0.5}(-1,1)^T$*, since the kernel of* A *is* $\{\alpha(1,-1)^T, \alpha \in \mathbb{R}\}$*. The corresponding* V_0 *is* $V_0 = 0.5(1,1)^T$*. By statement 4) in Proposition 2, the image of* $\mathcal{X} - x^\star$ *under* $\mathcal{S}$ *is* $\mathbb{R} \times \mathbb{R}_+$*, since the only constraint is* $[V_1]_S v_1 = \sqrt{0.5} v_1 \geq 0$*, i.e.* $v_1 \geq 0$*. Hence,* $\mu^\star$ *is normal distribution* $\mathcal{N}(Z(\omega), 1) \times \mathcal{N}(0,1)$ *truncated to* $\mathcal{V}^\star = \mathbb{R} \times \mathbb{R}_+$ *where* $Z(\omega) = (Y(\omega) - 1)/\sigma \sim \mathcal{N}(0,1)$*.*

Another case where the support can degenerate is where the kernel of A intersects the constrained parameter space at a single point. We give two examples.

Example 4. *Consider a linear inverse problem with* $A = (1,1)$ *and* $x_{\text{true}} = (0,0)^T$*. Then* $y^\star = Ax_{\text{true}} = 0$*. Under the Poisson error model with the identity link, the true distribution of data is degenerate:* $\mathbb{P}_{\text{true}}(Y = 0) = 1$*. Even though ill-posed, the linear inverse problem* $Ax = 0$*, i.e.* $x_1 + x_2 = 1$*, subject to constraints* $x_1, x_2 \geq 0$*, has a solution* $x^\star = (0,0)^T$ *that is unique due to the constraints, for any penalty.*

Since $y_{\text{exact}} = 0$ *and* $\nabla \tilde{f}_{y_{\text{exact}}}(Ax^\star) = 1$*, we have* $Z = \{1\}$ *and thus* $U_2 = (1,1)$ *and* $p_0 = 0$*. Take penalty* $||x - x_0||_2$ *with* $x_0 = (\alpha, \beta)^T \in (0,\infty)^2$*. The kernel of* A *is* $\{a(1,-1)^T,\ a \in \mathbb{R}\}$*,* $\nabla g(x^\star) = -x_0$*.*

If $\alpha = \beta$*,* $(1,-1)\nabla g(x^\star) = 0$ *and hence* $V_1 = U_1^T = \sqrt{0.5}(1,-1)^T$ *and* $p_3 = 0$*. By Proposition 2,* $V_2 = 0.5(1,1)^T$ *and the constraints are* $\sqrt{0.5} v_1 \geq 0$ *and* $-\sqrt{0.5} v_1 \geq 0$ *due to* $S_1 = \{1,2\}$*, that imply that* v_1 *must be zero.*

If $\alpha \neq \beta$*, then* $p_1 = 0$ *and* $U_3 = ((\beta - \alpha)_+, (\alpha - \beta)_+)$*. Taking* $\beta = 1$ *and* $\alpha = 2$*, we have* $U_3 = (0,1)$*,* $V_2 = (1,0)^T$ *and* $V_3 = (-1,1)^T$*. By statement 4) of Proposition 2, the constraints are* $-v_3 \geq 0$ *and* $v_3 \geq 0$*, implying* $v_3 = 0$*.*

Hence, in both cases, the limit of the posterior distribution of $\mathcal{S}(x - x^\star)$ *is Exp*(1) *for the first component, and the second component is 0 with probability 1 with* $\mathcal{V}^\star = \mathbb{R}_+ \times \{0\}$*.*

Now we consider a higher-dimensional example of this phenomenon.

Example 5. *Take* $A = \begin{pmatrix} 1\,1\,1\,1 \\ 1\,1\,0\,0 \end{pmatrix}$ *and* $x_{\text{true}} = (0,0,1,1)^T$*. Then* $y^\star = Ax_{\text{true}} = (2,0)^T$*. The likelihood is Poisson with the identity link:* $Y_i/\sigma^2 \sim Pois(A_i x/\sigma^2)$*. The linear inverse problem* $Ax = y^\star$*, i.e.* $x_1 + x_2 = 0$ *and*

$\sum_{i=1}^4 x_i = 2$, *subject to constraints* $x_i \geq 0\,\forall i$, *is ill-posed. The constraints are equivalent to* $x_3 + x_4 = 2$, x_3, $x_4 \geq 0$, $x_1 = x_2 = 0$. *To resolve the ambiguity, we use the penalty* $||x - x_0||_2$ *with* $x_0 = (1,3,4,1)^T$. *Then* $x^\star = (0,0,2,0)^T$.

To construct U_2, *find* $\nabla_{\eta_i} \tilde{f}_{y^\star}(\eta) = -y_i^\star/\eta_i + 1$, *thus* $\nabla_\eta \tilde{f}_{y^\star}(y^\star) = (0,1)^T$ *and* $Z = \{2\}$. *Thus,* $U_2 = (1,1,0,0)$ *and* $U_0 = (1,1,1,1)$.

The null space of A *is* $\{\alpha(1,-1,0,0)^T + \beta(0,0,1,-1)^T,\ \alpha,\beta \in \mathbb{R}\}$. *The gradient of* g *at* $x^\star$, $x^\star - x_0 = (-1,-3,-2,-1)^T$, *is orthogonal to* $\alpha(1,-1,0,0)^T + \beta(0,0,1,-1)^T$ – *a direction in the null space of* A – *if* $\beta = 2\alpha$. *Thus,* $V_1 = \sqrt{0.1}(1,-1,2,-2)^T$ *is the direction of* $\mathcal{W}_1$. *A vector* ζ *that satisfies* (14) *is* $\zeta = (2,0,0,1)^T$ *implying* $S = \{1,4\}$, *and* $U_3 = (2,0,0,1)$. *This implies that* $U_1 = \sqrt{0.1}(1,-1,2,-2)$ *and*

$$V_0 = 0.2(-1,1,3,2)^T, \quad V_2 = 0.1(3,7,-4,-6)^T, \quad V_3 = 0.2(2,-2,-1,1)^T.$$

Now we check the constraints using Proposition 2. We have $s = |S^*| - p_2 - p_3 = 1$, *and the constraints are* $[V_1]_\ell v_1 \geq 0$ *for* $\ell \in S_1 = S^* = \{1,2,4\}$, *i.e. we must have* $v_1 = 0$.

Hence, $\mu^\star = \mathcal{N}(\sqrt{2}Z(\omega) - 4c, 2) \times \delta_{\{0\}} \times Exp(1) \times Exp(1)$ *truncated to* $\mathcal{V}^\star = \mathbb{R} \times \{0\} \times \mathbb{R}_+^2$ *where* $Z(\omega) = \lim_{\sigma\to 0}(Y_1(\omega))/\sqrt{2} - 1) \sim \mathcal{N}(0,1)$ *in probability.*

5.5. *Approximate Bayes estimators.* Now we apply the approximation of the posterior distribution stated in Theorem 1 to approximate the distribution of Bayes estimators. We use a similar approach to that of Chernozhukov and Hong (2004), by approximating the distribution of recentered and rescaled Bayes estimates with the distribution of the corresponding Bayes estimates obtained under the limiting distribution, for a wide class of loss functions. The approach of Chernozhukov and Hong (2004) relies on Theorem I.10.2 in Ibragimov and Has'minskij (1981) that implies this result from an analogue of Theorem 1, under some additional conditions.

Assumption Q. We consider loss functions $Q : \mathbb{R}^p \to [0,\infty)$ satisfying the following properties:

1. $Q(z) \geq 0$, $Q(z) = 0$ if and only if $z = 0$, and Q is convex;
2. $Q(z)$ is dominated by a polynomial in $||z||_\infty$ as $||z||_\infty \to \infty$.

This loss function is applicable to the modified parameter $v = \mathcal{S}(x - x^\star)$, and the corresponding Bayes estimate of v is

$$\hat{v}_Q = \arg \inf_{v \in \mathcal{S}(\mathcal{X} - x^\star)} \int_{\mathcal{S}(\mathcal{X}-x^\star)} Q(v - v') p_{\mathcal{S}(x-x^\star)|y}(v') dv',$$

where $p_{\mathcal{S}(x-x^\star)|y}$ is the density of the posterior distribution of $\mathcal{S}(x - x^\star)$ with respect to Lebesgue measure. Since $\mathcal{S}$ is a bijection, the corresponding Bayes

estimate of x is

$$\hat{x}_Q = \arg\inf_{x\in\mathcal{X}} \int_{\mathcal{X}} Q(\mathcal{S}(x-x'))p(x' \mid y)dx',$$

and $\hat{v}_Q = \mathcal{S}(\hat{x}_Q - x^\star)$. Thus, the corresponding loss function for x is $\tilde{Q}(x) = Q(\mathcal{S}x)$.

In the next theorem we state the asymptotic distribution of $\hat{x}_Q$ as $\sigma \to 0$. If $c = \lim_{\sigma\to 0} \sigma/\gamma^2 = 0$ and there is no degeneracy of the support (the conditions of item 3) in Proposition 2 are satisfied), then $\mathcal{V}^\star$ factorises to $\mathcal{V}^\star = \otimes_{k=0}^3 \mathcal{V}_k$ where $\mathcal{V}_k \subseteq \mathbb{R}^{p_k}$ for $k = 0, 1$ and $\mathcal{V}_k \subseteq \mathbb{R}_+^{p_k}$ for $k = 2, 3$. If $c > 0$ and there is no degeneracy, then the factorisation is $\mathcal{V}_{03} \times \mathcal{V}_1 \times \mathcal{V}_2$, up to a permutation of the coordinates, since in this case the scales of $\mathcal{V}_0$ and $\mathcal{V}_3$, σ and γ^2, are of the same order and there are joint constraints on v_0 and v_3 (Proposition 2). To simplify the statement, let $\mathcal{S}_k$ be the map from $\mathcal{X} - x^\star$ to $\mathcal{V}_k^\star$ ($\mathcal{S}_0 = \sigma U_0$, $\mathcal{S}_1 = \gamma U_1$, $\mathcal{S}_2 = \sigma^2 U_2$, $\mathcal{S}_3 = \gamma^2 U_3$) and $\mu_k^\star$ be the marginal distribution on $\mathcal{V}_k$. If there is no degeneracy of the support, the marginal distribution is Gaussian on $\mathcal{V}_0$ and $\mathcal{V}_1$ and exponential on $\mathcal{V}_2$ and $\mathcal{V}_3$ (Theorem 1).

THEOREM 2. *Suppose that conditions of Theorem 1 hold, and that the loss function Q satisfies Assumption Q. Then, for $\hat{x}_Q$ defined above,*

(1) $\mathcal{S}(\hat{x}_Q - x^\star) \xrightarrow{d} v_Q^\star$ *as* $\sigma \to 0$, *where*

$$v_Q^\star(\omega) = \arg\inf_{v\in\mathcal{V}^\star} \int_{\mathcal{V}^\star} Q(v-v')d\mu^\star(v', \omega),$$

and $\mu^\star(\cdot, \omega)$ is the limit of the rescaled posterior distribution defined in Theorem 1;

(2) if $Q(v) = \sum_{k=0}^3 Q_k(v_k)$ *and* $\mathcal{V}^\star = \otimes_{k=0}^3 \mathcal{V}_k$ *where* $v = (v_0, v_1, v_2, v_3)$, $v_k \in \mathcal{V}_k$, $k = 0, 1, 2, 3$, *then*

$$\mathcal{S}_k(\hat{x}_Q - x^\star) \xrightarrow{d} v_{Q,k}^\star(\omega) = \arg\inf_{v_k\in\mathcal{V}_k} \int_{\mathcal{V}_k} Q_k(v_k - v_k')d\mu_k^\star(v_k', \omega)dv_k'.$$

Here $\xrightarrow{d}$ denotes convergence in distribution.

This theorem establishes consistency, rates of convergence, and limit distributions of Bayes estimates.

For example, for the quadratic loss function $Q(z) = ||z||_2^2$, $v_{Q,k}^\star$ is the mean of the corresponding limiting distribution, i.e. $v_{Q,0}^\star(\omega) = a_0(\omega)$, $v_{Q,1}^\star = 0$, $v_{Q,2}^\star$ is the vector $(1/a_1, \ldots, 1/a_{p_2})$ and $v_{Q,3}^\star$ is the vector $(1/b_1, \ldots, 1/b_{p_3})$.

To obtain the MAP estimate, we take $Q(v) = \sum_{i=1}^{p} I(|v_i| \leq \epsilon)/\epsilon$ – an approximation of the Dirac delta function that satisfies the above assumption, obtain the corresponding Bayes estimate and take the limit $\epsilon \to 0$, independently of σ. Thus, it can be shown that the rescaled and recentred MAP estimate $\mathcal{S}(\hat{x}_{\delta_0} - x^\star)$ converges in distribution to the mode of the density of $\mu^\star(\omega)$.

6. The Bernstein–von Mises theorem for SPECT.

6.1. *Approximation of the posterior distribution.* Consider the SPECT model defined in Section 2, and allow some coordinates of y_{exact} to be zero. This model is nonregular, since $\mathbb{P}_{y_{\text{exact}\,j}}(Y_j = 0) = 1$ for $j \in Z$; hence, with probability 1, $\nabla_j \tilde{f}_Y(Ax^\star) = 1 = \nabla_j \tilde{f}_{y_{\text{exact}}}(Ax^\star)$, and, since we assume that matrix A has no zero rows or columns,

$$\nabla f_{y_{\text{exact}}}(x^\star) = -\sum_{j:\, y_{\text{exact}\,j} \neq 0} y_{\text{exact}\,j} A_j/(A_j x^\star) + \sum_{j=1}^{n} A_j = \sum_{j \in Z} A_j \neq 0.$$

The nonregularity arises from the elements where there is no data ($y_{\text{exact}\,j} = 0$) but, since $A_j \neq 0$, it gives us information about those x_i where $A_{ji} \neq 0$.

The assumptions of Theorem 1 are satisfied since the derivatives of the log posterior are bounded up to order 3 (Assumption S), Assumption C holds due to convergence in probability $\nabla_j \tilde{f}_Y(Ax^\star) = 1 - Y_j/y_{\text{exact}\,j} \to 0 = \nabla_j \tilde{f}_{y_{\text{exact}}}(Ax^\star)$ as $\sigma \to 0$ for $j \notin Z$, and the convergence assumption is true for the second order derivatives and the functions as well. Assumption L is satisfied since both Poisson likelihood and the log cosh prior have exponential tails for large values of their arguments, the appropriate rescaling is either of the same order as $D_{\sigma,\gamma}^{-1}$ (leading to the integral over the complement of a ball with radius tending to infinity) or smaller (leading to an integral with a factor tending to zero), hence the integral outside of the ball vanishes.

The matrices U and V determining the change of variables can be taken as given by (19) and in Proposition 2 respectively. Then, if $p_3 > 0$, $b = 1$ and the parameter of the other exponential distribution is given by $a = V_2^T A_{Z,}^T \mathbf{1}_Z = a_y$.

For the log cosh prior with density defined by (4), the prior precision matrix $B(x)$ has the following non-zero entries:

$$\begin{aligned} B_{ss}(x) &= \frac{2(1+\delta)}{\delta} \sum_{s' \sim s} \left[1 + e^{2(x_s - x_{s'})/\delta}\right]^{-2}, \\ B_{ss'}(x) &= -\frac{2(1+\delta)}{\delta} \left[1 + e^{2(x_s - x_{s'})/\delta}\right]^{-2}, \quad \text{if} \quad s' \sim s. \end{aligned}$$

Note that B is singular, as the prior is improper (x is a priori identified only up an additive constant). The matrix $[A^TA : B]$ is of full rank since the null space of B, $\{v = \alpha(1,\dots,1)^T, \alpha \in \mathbb{R}\}$, is one-dimensional and $A(1,\dots,1)^T \neq 0$ since A does not have zero rows, and all of its elements are nonnegative. The precision matrices are given by

$$B_{11} = V_1^T B(x^\star) V_1, \qquad \Omega_{00} = V_0^T A_{Z,}^T \mathrm{diag}(1/[y_{\mathrm{exact}}]_Z) A_{Z,} V_0.$$

$$a_0 = \Omega_{00}^{-1} V_0^T A_{Z^c,}^T \xi + c\Omega_{00}^{-1} V_0^T \nabla g(x^\star) \sim \mathcal{N}\left(c\Omega_{00}^{-1} V_0^T \nabla g(x^\star), \Omega_{00}^{-1}\right),$$

where $\xi_j \sim \mathcal{N}(0,1)$ for $j \in Z^c$ is the limit of $\sqrt{y_{\mathrm{exact}}}(Y_j/y_{\mathrm{exact}\,j} - 1)/\sigma$ as $\sigma \to 0$, ξ_j are independent due to independence of Y_j, and $\xi = (\xi_1,\dots,\xi_n)^T$. The random variable a_0 is centred at zero (and hence the posterior distribution of $U_0(x - x^\star)/\sigma$ is not affected by the prior) if $c = \lim_{\sigma\to 0} \frac{\sigma}{\gamma^2} = 0$.

6.2. *Ill-posed-ness and unidentifiability.* The results in Bochkina (2012) prove that, under a subset of our stated conditions, the posterior degenerates to the point $x^\star$ in the small-variance limit. The proof assumes that the data model $p(y|x)$ is correctly specified; in subsequent work we intend to relax this. However, in respect of the other key model component – the prior for x – no such assumption of correctness is made either in Bochkina (2012) or the present paper. The prior is regarded as the invention of the analyst, who, of course, has no knowledge of x_{true}, the true value of x. Yet the point $x^\star$ does depend on the prior, and so we need to understand the impact of the prior on the difference between x_{true} and $x^\star$.

The point $x^\star$ is defined (Section 4.3) as the point maximising the prior density $p(x)$ subject to the non-negativity constraints $x \in [0,\infty)^p$ and the constraint that the model fits the exact data: $Ax^\star = y_{\mathrm{exact}} = Ax_{\mathrm{true}}$. Thus $x^\star$ agrees with x_{true} perfectly in directions orthogonal to the model hyperplane $Ax = y_{\mathrm{exact}}$, i.e. $P_{A^T}(x^\star - x_{\mathrm{true}}) = 0$.

In contrast, parallel to this hyperplane, there is no information in the data about x_{true}, so $x^\star$ is determined solely by the prior; there is no reason for $(I - P_{A^T})(x^\star - x_{\mathrm{true}})$ to be small. For example, for the Gaussian prior $p(x) \propto \exp(-1/(2\gamma^2)||x - x_0||_B^2)$, discussed in Section 4.1, if the unconstrained maximum $\lim_{\nu\to 0}(A^TA + \nu B)^{-1}(A^TAx_{\mathrm{true}} + \nu B x_0)$ satisfies the non-negativity constraints, then $x^\star$ is given explicitly by this expression, and so equals x_{true} if and only if $(I - P_{A^T})x_0 = (I - P_{A^T})x_{\mathrm{true}}$.

6.3. *Practical implications of the approximate posterior.* In this section, we briefly discuss some practical implications of Theorem 1. On a realistic scale where p is of the order 2^{12}–2^{16} we cannot hope to construct and

manipulate $p \times p$ matrices such as V; this seems to rule out any practical use. However, there are well-developed methods for SPECT reconstruction using our model, using Markov chain Monte Carlo computation, delivering not only approximate, simulation-consistent, posterior means, but also variances; see Weir (1997). In this context, the theorem provides valuable knowledge which can enrich the interpretation of numerical 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 posterior distributions for *linear combinations* $\lambda^T x$ of elements of x, and are particularly amenable to treatment in this way. More specifically, if for any non-empty set of pixels $R \subseteq \{1, 2, \ldots, p\}$, α^R denotes the vector with elements $\alpha_j^R = 1/|R|$ for $j \in R$, 0 otherwise, then to deal with case (a) we can take $\lambda = \alpha^R$ to deliver $\lambda^T x$ as the average concentration of tracer in region R, and for case (b) take $\lambda = \alpha^{R_1} - \alpha^{R_2}$ to give the difference in average concentration in region R_1 compared to R_2. To avoid bias in such inferences arising from the lack of identifiability caused by ill-posed-ness, it is important to check that λ lies in the row space of A.

The approach we propose would exploit analyis of MCMC output of $\lambda^T x$, together with a numerical MAP estimate $\hat{x} = \text{argmax}_x p(x|y)$, calculated using an EM-based algorithm (Green 1990). This can be used to identify $x^\star$ and hence S, thus partitioning elements of x into those that are asymptotically normally or exponentially distributed; we anticipate $p_0 \gg p_1, p_2, p_3$ in practical situations. It will commonly be the case that $j \notin S$ for all j such that $\lambda_j \neq 0$, in which case the theorem tells us that $\lambda^T x$ is asymptotically normally distributed, and its mean and variance, computed by MCMC, can be directly used to specify the approximating normal distribution and answer the inferential question posed. In the contrary case where $\lambda_j \neq 0$ for some $j \in S$, a little work is needed to separate the asymptotically normal and exponential components of x contributing to $\lambda^T x$. The simplest example of this would be a test for $x_j = 0$ for a j such that $\hat{x}_j = 0$ and so $j \in S$. For such j, x_j is asymptotically exponentially distributed *a posteriori*, with a parameter that can be readily estimated from the MCMC output.

6.4. *Finite sample performance.* Finally, 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 instead we present selected results from an analysis of synthetic data based on a real SPECT scan of the pelvic region

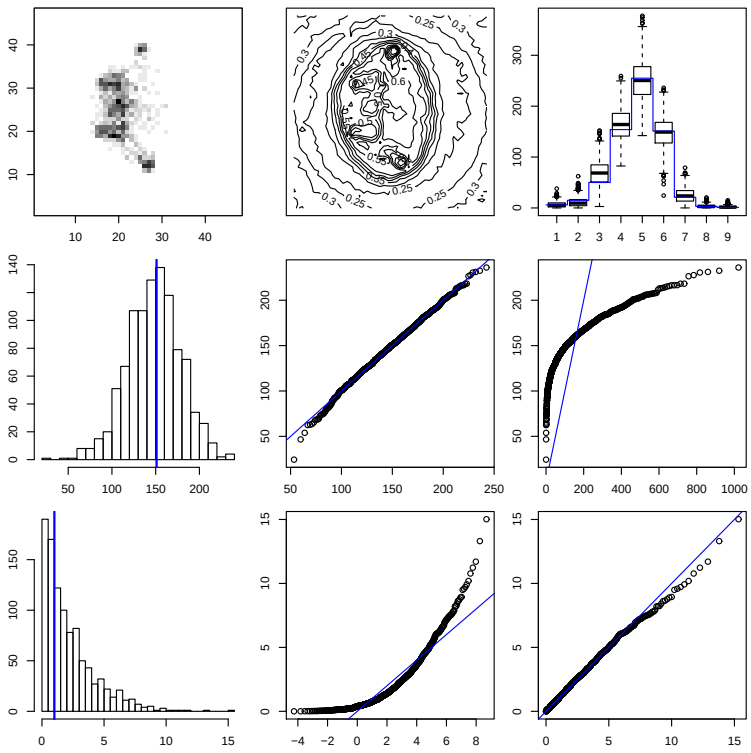

FIG 2. *First row: a sample from the posterior, contour map of acceptance rates, marginal posteriors for a section of 9 consecutive pixels through the image (row 12, columns 23 to 31), showing ground truth in blue. Second row: histogram of marginal posterior for a high-spot pixel (row 12, column 28), with truth indicated by blue line, and the same shown as a QQ plot against the normal and exponential distributions respectively. Third row, the same but for a low-spot pixel (row 12, column 31).*

of a human subject.

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, corresponding to a spatial resolution of 0.57cm. Synthetic data was generated using this A and a 'ground truth' obtained from an approximate MAP reconstruction from real data. The total photon count was 62953; individual counts ranged from 0 to 114, averaging 18.9. Reconstruction was performed on a 48×48 square grid, with pixel size 0.64cm, using the log cosh prior with hyperparameters fixed at $\gamma = 25$ and $\delta = 8$, was obtained using a simple MCMC sampler. We employed 20000 sweeps of a deterministic-raster-scan single-pixel random walk Metropolis sampler on a square-root scale for x, chosen to avoid extremes in acceptance rate at high- and low-spots in the image.

Figure 2 shows selected aspects of this analysis; see caption for details. Our tentative conclusion from this is that the marginal posterior distributions for individual pixels x_j do appear to be approximately normal in high-spots and approximately exponential in low-spots, consistent with the theoretical limits presented in Theorem 1.

7. Discussion. When the posterior distribution concentrates on the boundary, we have showed that the classic Bernstein–von Mises theorem, stating the limit of the posterior distribution recentred and rescaled by $\sqrt{n}$ for n independent random variables, does not hold. Instead, the limit differs in two respects, in directions towards the boundary: the limiting distribution is an exponential, and the appropriate scale is n, i.e. the convergence is faster. Parallel to the boundary, however, the classic version of Bernstein–von Mises theorem is applicable. Our results also extend the Bernstein–von Mises theorem to the case of non-iid observations and of a non-identifiable likelihood, for models that belong to the GLIP class.

These are examples of nonregular problems, differing from those considered by Ghosal and Samanta (1995), Ghosh *et al.* (1994) and Chernozhukov and Hong (2004). In their case, the density of the errors has a jump whose location depends on the unknown parameter; this is similar to a change point problem, whereas in our case there is a degeneration of the likelihood. This difference is reflected in the limiting distribution, that in the former case is shifted by a random variable that depends on data, whereas in our case there is no shift and no dependence of the exponential distribution on the data.

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 Gaussian part of the limit and the smallest parameter of each of the exponential distributions.

There are interesting questions, beyond the scope of this paper, concerning the appropriateness of different prior formulations (as assessed from a frequentist perspective). Within a subjectivist Bayesian paradigm, real prior information is necessary for an informed choice.

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 x_{true}.

APPENDIX A: PROOFS

A.1. Proofs, Section 4.

PROOF OF PROPOSITION 1. For arbitrary $\nu > 0$, suppose $x \in \mathbb{R}^p$ is such that $(A^TA + \nu B)x = 0_p$, the zero vector in $\mathbb{R}^p$. We have to show that $x = 0_p$. But $(A^TA + \nu B)x = 0_p$ implies $x^T(A^TA + \nu B)x = 0$, and so by non-negative-definiteness of B and A^TA, $x^TBx = 0 = x^TA^TAx$. But then $Bx = 0_p = A^TAx$, and so $x^T[B : A^TA] = 0_{2p}^T$. By the assumed full rank of this matrix, x must be 0_p.

Now fix $\nu_0 > 0$. By Theorem 2 of Searle (1982), page 313, (with his A replaced by $A^TA + \nu_0 B$), there exists a nonsingular real matrix P, not necessarily orthogonal, such that $P^T(A^TA + \nu_0 B)P = I$ and P^TBP is the diagonal matrix Λ of the solutions for λ to $|B - \lambda(A^TA + \nu_0 B)| = 0$ (which all satisfy $0 \le \lambda \le \nu_0^{-1}$). But then $P^TA^TAP = I - \nu_0\Lambda$ and for any $\nu > 0$, $P^T(A^TA + \nu B)P = I + (\nu - \nu_0)\Lambda$, both of which are of course also diagonal.

The matrix P can depend on the choice of ν_0, but evidently always diagonalises A^TA, B and any linear combination. Also, Λ depends on ν_0, but since $|B - \lambda(A^TA + \nu_0 B)| = (1 - \lambda\nu_0)^p|B - \alpha A^TA|$ where $\alpha = \lambda/(1 - \lambda\nu_0)$, the (diagonal) elements in Λ are $\lambda_i = \alpha_i/(1 + \alpha_i\nu_0)$ where $\{\alpha_i\}$ are the solutions to $|B - \alpha A^TA| = 0$ (possibly some $\alpha_i = +\infty$). So $P^TBP = \mathrm{diag}(\alpha_i/(1 + \alpha_i\nu_0))$, $P^TA^TAP = \mathrm{diag}(1/(1 + \alpha_i\nu_0))$ and for any ν, $P^T(A^TA + \nu B)P = \mathrm{diag}((1 + \alpha_i\nu)/(1 + \alpha_i\nu_0))$. For the final assertions, note that $\nu(A^TA + \nu B)^{-1} = P\mathrm{diag}(\nu(1 + \alpha_i\nu_0)/(1 + \alpha_i\nu))P^T$, which converges to $P\mathrm{diag}(\delta_i)P^T = C$, say, where $\delta_i = \nu_0$ if $\alpha_i = +\infty$, and 0 otherwise. Further, we can estimate the difference: $\nu(A^TA + \nu B)^{-1} - C = P\mathrm{diag}(\nu(1 + \alpha_i\nu_0)/(1 + \alpha_i\nu) - \delta_i)P^T = \nu P\mathrm{diag}(\phi_i)P^T + O(\nu^2)$, where $\phi_i = 0$ if $\alpha_i = +\infty$ and otherwise $\phi_i = 1 + \alpha_i\nu_0$.

Transformation by P scales and skews the result, but in a way independent of ν, so the limiting behaviour of $\nu(A^TA + \nu B)^{-1}$ follows from the facts that the diagonal terms corresponding to $\alpha_i = +\infty$ have finite positive limits and the remaining ones scale as ν. We see from $P^TA^TAP = \mathrm{diag}(1/(1 + \alpha_i\nu_0))$ that the number of α_i not equal to $+\infty$ is just the rank of A^TA, i.e. $\mathrm{rank}(A)$. Thus $\nu(A^TA + \nu B)^{-1} = C + D\nu + o(\nu)$ as $\nu \to 0$ as required. □

PROOF OF LEMMA 1. The matrix V is of full rank if

$$\sum_{k=0}^{3} V_k w_k = 0, \quad w_0 \in \mathbb{R}^{p_0},\ w_1 \in \mathbb{R}^{p_1},\ w_2 \in \mathbb{R}_+^{p_2},\ w_3 \in \mathbb{R}_+^{p_3}$$

implies $w_k = 0$ for all k.

Multiply the above expression by $[\nabla_Z \tilde{f}_{y_{\mathrm{exact}}}(x^\star)]^T A_Z$, we have that

$$0 = [\nabla_Z \tilde{f}_{y_{\mathrm{exact}}}(x^\star)]^T A_Z V_2 w_2$$

Since vector $V_2^T A_Z^T \nabla_Z \tilde{f}_{y_{\text{exact}}}(x^\star)$ has positive coordinates and w_2 has non-negative coordinates, the condition holds only if $w_2 = 0$.

Multiplying the above expression by $V_0^T A_{Z,}^T A_{Z,}$ and use $w_2 = 0$, we have that

$$0 = V_0^T A_{Z,}^T A_{Z,}[V_0 w_0 + V_2 w_2] = V_0^T A_{Z,}^T A_{Z,} V_0 w_0$$

Since matrix $V_0^T A_{Z,}^T A_{Z,} V_0$ is of full rank, the above condition implies $w_0 = 0$.

Now we multiply the condition by $\nabla g(x^\star)^T$ and use $w_0 = 0$ and $w_2 = 0$:

$$0 = \sum_{k=0}^{3} \nabla g(x^\star)^T V_k w_k = \nabla g(x^\star)^T V_1 w_1 + \nabla g(x^\star)^T V_3 w_3 = \zeta^T V_1 w_1 + \zeta^T V_3 w_3 = \zeta^T V_3 w_3.$$

Vector $V_3^T \zeta$ has positive entries therefore this condition implies $w_3 = 0$.

Now we multiply the condition by $[V_1]_{S^c,}^T J$ with $J_{,S^c} = I_{|S^c|}$ and $J_{,S} = 0$, and use $w_0 = 0$, $w_2 = 0$ and $w_3 = 0$:

$$0 = \sum_{k=0}^{3} [V_1]_{S^c,}^T J V_k w_k = [V_1]_{S^c,}^T [V_1]_{S^c,} w_1.$$

Since matrix $[V_1]_{S^c,}^T [V_1]_{S^c,}$ is positive definite, this condition implies $w_1 = 0$.

Thus, we showed that the $p \times p$ matrix V has p linearly independent columns, thus it is of full rank.

□

Proof of Proposition 2. Conditions (15) state that $p \times p_1$ matrix V_1 consists of p_1 linearly independent columns that satisfy $AV_1 = 0$ and $\zeta^T V_1 = 0$. If $\zeta = 0$ (i.e. $\nabla g(x^\star)$ is in the image space of A^T) then V_1 consists of the $p_1 = p - \text{rank}(A)$ vectors that form a basis of the null space of A. If $\zeta \neq 0$, then ζ is linearly independent of the columns of A and V_1 consists of the $p_1 = p - \text{rank}(A^T : \zeta) = p - \text{rank}(A) - 1$ vectors that form a basis of the null space of $(A^T : \zeta)^T$.

Therefore, if $\zeta = 0$, $p_3 = 0$, otherwise $p_3 = 1$ and V_3 is a vector in the null space of A that satisfies $V_3^T \zeta > 0$.

1. Introduce the change of variables $w = U(x - x^\star)$, where $U = V^{-1}$ is defined in the statement of the lemma. Condition $UV = I$ is equivalent to $U_k V_k = I_{p_k}$ and $U_k V_j = 0$ for $k \neq j$, i.e.

$$\begin{gathered} A_{Z_0,} V_0 = I_{p_0}, \quad A_{Z_2,} V_2 = I_{p_2}, \quad U_1 V_1 = I_{p_1}, \quad \zeta^T V_3 = 1_{p_3}, \\ AV_1 = 0, \quad AV_3 = 0, \quad A_{Z_0,} V_2 = 0, \quad A_{Z_2,} V_0 = 0, \\ \zeta^T V_2 = 0, \quad \zeta^T V_0 = 0, \quad \zeta^T V_1 = 0, \\ U_1 V_k = 0, \; k \neq 1. \end{gathered} \tag{29}$$

Thus, to show that all conditions (15) are satisfied, we need to show that

$$V_0^T A_{Z^c,}^T A_{Z^c,} V_0 \quad \text{is positive definite,}$$
$$V_2^T A_{Z,}^T \nabla_Z \tilde{f}_{y_{\text{exact}}}(Ax^\star) \text{ is a vector with positive coordinates.}$$

Recall that $\nabla_j \tilde{f}_{y_{\text{exact}}}(x^\star) > 0$ for all $j \in Z$.

If the matrix $A_{Z,}$ is of full rank, then $Z_2 = Z$ and $V_2^T A_{Z,}^T = I_{p_2}$, hence the latter condition is satisfied. Otherwise, by Caratheodory's theorem (p.37 of Bertsekas (2006)), $\exists Z_2 \subset Z$ such that vectors $\{A_{j,},\ j \in Z_2\}$ are linearly independent and define the cone $\left\{w = \sum_{j\in Z} A_{j,}^T \mu_j;\ \mu_j \geq 0\right\}$, i.e. for any $j \in Z$, $A_{j,}$ can be written a linear combination of vectors $A_{j,}$, $j \in Z_2$ with nonnegative coefficients, in particular, for $Z_{22} = Z \setminus Z_2$, $A_{Z_{22},} = \beta A_{Z_2,}$ with $\beta_{ij} \geq 0$, $i = 1, \ldots, |Z_{22}|$, $j = 1, \ldots, |Z_2|$. Then, vector $V_2^T A_Z^T \nabla_Z \tilde{f}_{y_{\text{exact}}}(x^\star)$ has positive coordinates

$$V_2^T A_Z^T \nabla_Z \tilde{f}_{y_{\text{exact}}}(x^\star) = \nabla_{Z_2} \tilde{f}_{y_{\text{exact}}}(x^\star) + \beta^T \nabla_{Z_{22}} \tilde{f}_{y_{\text{exact}}}(x^\star) > 0,$$

where the inequality is componentwise.

By definition of Z_0, there exist matrices α_0 and α_2 such that $A_{Z^c,} = \alpha_0 A_{Z_0,} + \alpha_2 A_{Z_2,}$, in particular $|Z^c| \times p_0$ matrix α_0 is of full rank. Therefore, $A_{Z^c,} V_0 = \alpha_0$ and the matrix $V_0^T A_{Z^c,}^T A_{Z^c,} V_0 = \alpha_0^T \alpha_0$ is of full rank, hence the first condition is also satisfied.

Columns of matrix $(A_{Z_0,}^T, A_{Z_2,}^T, V_1, \zeta)$ are linearly independent and span $\mathbb{R}^p$, hence, matrix U_1^T can be written as a linear combination of the columns of this matrix,

$$U_1 = \delta_{k0} A_{Z_0,} + \delta_{k2} A_{Z_2,} + \delta_{k1} V_1^T + \delta_{k3} \zeta^T.$$

Conditions (30) imply that $U_1 = V_1^T$.

2. Columns of matrix $(A_{Z_0,}^T : A_{Z_2,}^T : V_1 : \zeta)$ are linearly independent and span $\mathbb{R}^p$, hence, any matrix V_k^T can be written as a linear combination of the columns of this matrix, i.e. for $k = 0, 2, 3$,

$$V_k = A_{Z_0,}^T \delta_{k0} + A_{Z_2,}^T \delta_{k2} + V_1 \delta_{k1} + \zeta \delta_{k3},$$

and the same holds for U_1^T. By the conditions (30), multiplying the expression for V_3 by $A_{Z_0,}$, $A_{Z_2,}$, V_1^T and ζ^T implies that $\delta_{31} = 0$, $\delta_{33} = 1/||\zeta||^2$, $\delta_{3m} = -(A_{Z_m,} A_{Z_m,}^T)^{-1} A_{Z_m,} \zeta \delta_{33}$ for $m = 0, 2$, implying $V_3 = \tilde{\zeta}/||\tilde{\zeta}||^2$, where

$$P_0 = P_{A_{Z_0,}^T}, \quad \tilde{P}_{2,0} = (I - P_0) A_{Z_2,}^T (A_{Z_2,}(I - P_0) A_{Z_2,}^T)^{-1} A_{Z_2,}(I - P_0), \quad \tilde{\zeta} = (I - \tilde{P}_{2,0})(I - P_0)\zeta.$$

For $(k,m) \in \{(0,2),(2,0)\}$, multiplying the expression

$$V_k \quad = \quad A^T_{Z_k,}\delta_{kk} + A^T_{Z_m,}\delta_{km} + V_1\delta_{k1} + \zeta\delta_{k3}$$

by $A_{Z_k,}$, $A_{Z_m,}$, V_1^T and ζ^T and using conditions (30), we obtain

$$V_k = (I - \tilde{P}_{m,\zeta})A^T_{Z_k,}(A_{Z_k,}(I - \tilde{P}_{m,\zeta})A^T_{Z_k,})^{-1},$$

where $P_\zeta = \zeta\zeta^T/||\zeta||^2$ and $\tilde{P}_{k,\zeta} = (I - P_\zeta)A^T_{Z_k,}[A_{Z_k,}(I - P_\zeta)A_{Z_k,}]^{-1}A_{Z_k,}(I - P_\zeta)$.

3, 4. Now we study the image of map $\mathcal{S}$.

Since $[y_{\text{exact}}]_{Z^*} = 0$ and $x^\star_j \neq 0$ for all $j \in S^{*c}$, we must have $A_{Z^*,S^{*c}} = 0$ since

$$0 = [y_{\text{exact}}]_{Z^*} = A_{Z^*,S^*}x^\star_{S^*} + A_{Z^*,S^{*c}}x^\star_{S^{*c}} = A_{Z^*,S^{*c}}x^\star_{S^{*c}}.$$

The values of $U_2(x - x^\star) = A_{Z_2,}(x - x^\star) = A_{Z_2,S^*}x_{S^*}$ are positive since $A_{Z_2,S^{*c}} = 0$ and $A_{Z_2,S^*}x^\star_{S^*} = 0$ is the lower boundary point of $A\mathcal{X}$. The KKT conditions imply that for $j \in Z^*$, $[Ax^\star]_j$ is a boundary point of $A\mathcal{X}$, and therefore $A_{j,}(x - x^\star) = A_{j,}x - [Ax^\star]_j \geq 0$, since we assumed that $[Ax^\star]_j$ is a lower boundary point of $A\mathcal{X}$.

If $p_3 \neq 0$, $U_3(x - x^\star) = \zeta^T(x - x^\star) = \zeta^T_{S^*}x_{S^*}$. If $S \neq S^*$ and S is empty, then $\zeta^T_{S^*} = 0$ and hence the image of $U_3(x - x^\star)$ is zero, a single point. If S is not empty, then the image of $U_3(x - x^\star)$ is $[0,\infty)$ since $x_j \geq 0$ and $\zeta_j > 0$ for $j \in S$.

If we can choose Z_0 such that $Z_0 \cap (Z^* \setminus Z) = \emptyset$ and $S^{*c} \neq \emptyset$, values of $U_0(x - x^\star) = A_{Z_0,}(x - x^\star)$ can be both positive and negative for $x \in \mathcal{X}$ since $A_{Z_0,j} \neq 0$ for all $j \in S^{*c}$ (otherwise matrix A would have a zero column).

If $Z \neq Z^*$ and $Z_0 \cap Z^* \neq \emptyset$, then for $j \in Z^* \cap Z_0$, $A_{j,S^{*c}} = 0$ and the values of $[U_0]_{j,}(x - x^\star) = A_{j,S^*}x_{S^*}$ can only be nonnegative for $x \in \mathcal{X}$, since we assumed that $A\mathcal{X}$ has only lower boundary points, that could only be zeroes. Denote $Z^*_0 = Z^* \cap Z_0$ and $z = |Z^*_0|$.

Values of $U_1(x - x^\star)$ can be both positive and negative if $[U_1]_{i,S^{*c}}$, that is, $[V_1]_{S^{*c},i}$, is nonzero for all $i \in 1,\ldots,p_1$. This is equivalent to any solution v to the equation $(A^T : \zeta)^T v = 0$ satisfying $v_{S^{*c}} \neq 0$. The condition is equivalent to

$$0 = [(A_{,S^*}v_{S^*})^T : v^T_{S^*}\zeta_{S^*}]^T + [(A_{,S^{*c}}v_{S^{*c}})^T : 0]^T.$$

Since $A_{Z^*,S^{*c}} = 0$, v_{S^*} is a solution of $(A^T_{Z^*,S^*} : \zeta_{S^*})^T v_{S^*} = 0$. The number of linearly independent nonzero solutions v_{S^*} is $|S^*| - \text{rank}((A^T_{Z^*,S^*} : \zeta_{S^*})) = |S^*| - p_2 - p_3 - z \geq 0$. The remaining condition is $A_{,S^{*c}}v_{S^{*c}} =$

$-A_{,S^*}v_{S^*}$. Note that since $A_{Z_0,}$ are linearly independent vectors that are not in the range of $A_{Z,}^T$ and $A_{Z^*,S^{*c}} = 0$, this condition is equivalent to $A_{Z_0\setminus Z_0^*,S^{*c}}v_{S^{*c}} = 0$ (using column elimination), and also $\operatorname{rank}(A_{Z_0,S^{*c}}) + \operatorname{rank}(A_{Z_2,S^*}) = \operatorname{rank}(A_{Z_0\cap Z_2,}) = \operatorname{rank}(A)$ and hence $\operatorname{rank}(A_{Z_0,S^{*c}}) = \operatorname{rank}(A_{Z_0\setminus Z_0^*,S^{*c}}) = p_0$. Hence, the number of linearly independent nonzero solutions $v_{S^{*c}}$ of the above equation is $|S^{*c}| - \operatorname{rank}(A_{Z_0\setminus Z_0^*,S^{*c}}) = p - |S^*| - p_0$.

Thus, if the number of linearly independent nonzero solutions $v_{S^{*c}}$ $p - |S^*| - p_0$ is equal to p_1, i.e. if $|S^*| = p_2 + p_3$, then the range of $U_1(x - x^\star)$ includes both positive and negative values. If $p - |S^*| - p_0 < p_1$ (i.e. $|S^*| > p_2 + p_3 + z$), then, $s = |S^*| - p_2 - p_3$ rows of $[U_1]_{,S^{*c}}$ are zero (say, the last s rows) (thus, $[U_1]_{1:s,S^{*c}} = 0$ and $[V_1]_{S^{*c},1:s} = 0$). Then, for $\ell \in 1:s$,

$$[w_1]_\ell = [U_1]_{\ell,}(x-x^\star) = [U_1]_{\ell,S^*}x_{S^*}+[U_1]_{\ell,S^{*c}}(x_{S^{*c}}-x^\star_{S^{*c}}) = [U_1]_{\ell,S^*}x_{S^*} = [V_1]^T_{S^*,\ell}\, x_{S^*}.$$

Then, if $[V_1]_{S^*,\ell}$ includes both positive and negative values, the range of $[w_1]_\ell$ includes both positive and negative values. If $[V_1]_{S^*,\ell}$ has only positive or only negative values, then the range is either nonnegative or nonpositive.

Now we also need to check whether there is a constraint on v_k arising from the constraints on x.

Constraints: $x_\ell - x^\star_\ell = x_\ell = \sigma[V_0]_{\ell,}\, v_0 + \gamma[V_1]_{\ell,}\, v_1 + \sigma^2[V_2]_{\ell,}\, v_2 + \gamma^2[V_3]_\ell\, v_3 \geq 0$ for $\ell \in S^*$. In the limit, the dominating order is γ, hence for $\ell \in S^*$ such that $[V_1]_{\ell,} \neq 0$, the constraints imply $[V_1]_{\ell,}v_1 = [V_1]_{\ell,1:s}[v_1]_{1:s} \geq 0$.

For $\ell \in S^*$ such that $[V_1]_{\ell,} = 0$, the dominating order is γ^2, if $c = \lim \sigma/\gamma^2 = 0$, and γ^2 and σ if $c > 0$. If $c = 0$, for $\ell \in S^*$ such that $[V_3]_\ell \neq 0$, the constraints become $[V_3]_\ell v_3 \geq 0$. Thus, if all nonzero values of $[V_3]_\ell$ for $\ell \in S^*$ are positive, the constraint is $v_3 \geq 0$. However, if $[V_3]_{S^*}$ has both positive and negative values, the constraint implies w_3 must be zero, thus we have the degeneracy of the support of v_3. If $c > 0$, the same holds for $\ell \in S^*$ such that $[V_3]_\ell \neq 0$ and $[V_0]_{\ell,} = 0$.

If $c > 0$, for $\ell \in S^*$ such that $[V_3]_\ell \neq 0$, the constraints become $[V_3]_\ell v_3 + c[V_0]_{\ell,}v_0 \geq 0$.

Thus, if $c = 0$, then $\mathcal{V}^\star$ is defined by

$$\mathcal{V}^\star = \{(v_0, v_1, v_2, v_3) \in \mathbb{R}^{p_0+p_1-s} \times \mathbb{R}_+^{s+p_2+p_3} : [V_k]_{S_k,}v_k \geq 0,\ k = 0, \dots, 3\}$$

where the inequalities are component-wise. If $c > 0$, then $\mathcal{V}^\star$ is defined by

$$\mathcal{V}^\star = \{(v_0, v_1, v_2, v_3) \in \mathbb{R}^{p_0+p_1-s}\times\mathbb{R}_+^{s+p_2+p_3} : c[V_0]_{S_{03},}v_0+[V_3]_{S_{03},}v_3 \geq 0\,\&\,[V_k]_{S_k,}v_k \geq 0,\ k = 1, 2\}$$

where $S_{03} = S_0 \cup S_3$ and the inequalities are component-wise.

In particular, the inequality on v_1 is only on the last s components:

$$[V_1]_{S_1,(p_1-s+1):p_1}[v_1]_{(p_1-s+1):p_1} \geq 0.$$

□

A.2. Upper and lower bounds on the log posterior density. We give two lemmas that provide random and nonrandom upper and lower bounds on the log of the posterior density.

LEMMA 2. *Let* $B_\delta = B_2(0,\delta_0)\times B_2(0,\delta_1)\times B_\infty(0,\delta_2)\times B_\infty(0,\delta_3)$, $B_\delta(x^\star) = \{x\in\mathcal{X} : V^{-1}(x-x^\star)\in B_\delta$, *and denote* $\delta_+ = ||V||_\infty[\delta_0\sqrt{p_0}+\delta_1\sqrt{p_1}+\delta_2+\delta_3]$. *Denote also*

$$\begin{aligned}
\kappa_A &= \frac{2p}{3}(C_{f3}+2\nu C_{g3}), \quad \kappa_B = \frac{4p}{3}C_{g3},\\
\delta_a &= \delta_0||B_{02}||_{2,\infty}+\delta_1||B_{12}||_{2,\infty}+\frac{\delta_2}{2}||B_{22}||_{\infty,\infty}+\delta_2\delta_+\frac{q_2^2\kappa_A}{2p},\\
\delta_b &= \delta_0||B_{03}||_{2,\infty}+\delta_1||B_{13}||_{2,\infty}+\delta_2||B_{23}||_{\infty,\infty}+\frac{\delta_3}{2}||B_{33}||_{\infty,\infty}+\delta_+\delta_3\frac{q_3^2\kappa_B}{2p},
\end{aligned}$$

where $H_{ij} = V_iHV_j^T$ *and* $B_{ij} = V_iBV_j^T$, $i,j\in\{0,1,2,3\}$, $q_k = ||V_k||_{\infty,\infty}$ *for* $k=2,3$.

1. Upper bound. *Then, for* $x\in B_\delta(x^\star)$, *we have the following upper bound:*

$$\begin{aligned}
[h_y(x)-h_y(x^\star)]/\sigma^2 &\le (a_y+\delta_a\mathbf{1})^Tv_2+(b+\delta_b\mathbf{1})^Tv_3+||\widetilde{H}_{00}^{1/2}(v_0-\widetilde{H}_{00}^{-1}\nabla h_y(x^\star)/\sigma)||^2/2\\
&+ ||\widetilde{B}_{11}^{1/2}v_1||_2^2/2-\frac{1}{2\sigma^2}||\widetilde{H}_{00}^{-1/2}\nabla h_y(x^\star)||^2+||B_{10}||\delta_0\delta_1/\sigma^2,
\end{aligned}$$

where $\widetilde{B}_{11} = B_{11}+\delta_1\kappa_BI_{p_1}$, $\quad\widetilde{H}_{00} = H_{00}+\delta_0\kappa_AI_{p_0}$.

2. Lower bound. *For* $x\in B_\delta(x^\star)$ *and small enough* δ_k *and* ν, *we have the following lower bound:*

$$\begin{aligned}
[h_y(x)-h_y(x^\star)]/\sigma^2 &\ge (a_y-\delta_a\mathbf{1})^Tv_2+(b-\delta_b\mathbf{1})^Tv_3+||\bar{H}_{00}^{1/2}\left(v_0-\bar{H}_{00}^{-1}\nabla h_y(x^\star)/\sigma\right)||^2/2\\
&+ ||\bar{B}_{11}^{1/2}v_1||_2^2/2-\frac{1}{2\sigma^2}||\bar{H}_{00}^{-1/2}\nabla h_y(x^\star)||^2-||B_{10}||\delta_0\delta_1/\sigma^2,
\end{aligned}$$

where $\bar{B}_{11} = B_{11}-\delta_1\kappa_BI_{p_1}$, $\quad\bar{H}_{00} = H_{00}-\delta_0\kappa_AI_{p_0}$.

PROOF. Approximate $h_y(x)$ by a quadratic function using Taylor decomposition in a neighbourhood of $x^\star$:

$$h_y(x) = h_y(x^\star)+[\nabla h_y(x^\star)]^T(x-x^\star)+\frac{1}{2}(x-x^\star)^TH(x-x^\star)+\Delta_{00}(x).$$

We start with looking at the gradient:

$$\nabla h_y(x^\star) = P_{A^T}(\nabla f_y(x^\star)+\nu\nabla g(x^\star))+\nu(I-P_{A^T})\nabla g(x^\star).$$

Using the properties of the local geometry described in Section 4.3 and the representation $x = x^\star + \sum_{k=0}^{3} V_k w_k$, we have

$$\begin{aligned}(x-x^\star)^T \nabla f_{y_{\text{exact}}}(x^\star) &= [\sum_{k=0}^{3} V_k w_k]^T \nabla f_{y_{\text{exact}}}(x^\star) = w_2^T V_2^T A_{Z,} \nabla_Z \tilde{f}_{y_{\text{exact}}}(Ax^\star) = a^T w_2, \\ (x-x^\star)^T \nabla f_y(x^\star) &= w_2^T V_2 A_{Z,} \nabla_Z \tilde{f}_y(Ax^\star) + w_0^T V_0 \nabla f_y(x^\star),\end{aligned}$$

due to $V_0^T \nabla f_{y_{\text{exact}}}(x^\star) = 0$, $AV_1 = 0$ and $AV_3 = 0$. Also,

$$(x-x^\star)^T \nabla g(x^\star) = \sum_{k=0}^{3} w_k^T V_k^T \nabla g(x^\star) = w_0^T V_0^T \nabla g(x^\star) + w_2^T V_2^T \nabla g(x^\star) + w_3^T V_3^T \nabla g(x^\star),$$

due to $V_1^T \nabla g(x^\star) = 0$. Combining these expressions together, we have

$$(x-x^\star)^T \nabla h_y(x^\star) = w_0^T V_0^T \nabla h_y(x^\star) + w_2^T a_y + \nu w_3^T b.$$

Now we bound Δ_{00} for $w = V^{-1}(x - x^\star) \in B_\delta$ using Taylor decomposition of $h_y(x)$: $\exists x_c \in \langle x, x^\star \rangle$:

$$\begin{aligned}|\Delta_{00}(\delta)| &= \frac{1}{6} \left| \sum_{ijk} \nabla_{ijk} h_y(x_c)(x_i - x_i^\star)(x_j - x_j^\star)(x_k - x_k^\star) \right| \\ &= \frac{1}{6} \left| \sum_i (x_i - x_i^\star) \frac{\partial}{\partial z_i} \left[(x-x^\star)^T \nabla^2 h_y(z)(x-x^\star) \right]_{z=x_c} \right|\end{aligned}$$

Note that

$$\begin{aligned}(x-x^\star)^T \nabla^2 h_y(z)(x-x^\star) &= \sum_{k,j=0}^{3} w_k^T V_k^T \nabla^2 h_y(z) V_j w_j \\ &= \sum_{k,j=0,2} w_k^T V_k^T \nabla^2 f_y(z) V_j w_j + \nu \sum_{k,j=0}^{3} w_k^T V_k^T \nabla^2 g(z) V_j w_j.\end{aligned}$$

Differentiating with respect to z and bounding the third derivatives of f_y and g using Assumption S, we have that for every i, with high probability,

$$|w_k^T V_k^T \nabla_i \nabla^2 f_y(z) V_j w_j| \leq C_{f3} \, ||V_j w_j||_1 \, ||V_k w_k||_1.$$

Then, we can use inequalities $||V_k w_k||_1 \leq \sqrt{p} ||V_k w_k||_2 \leq \sqrt{p} ||w_k||_2$ for $k = 0, 1$, and $||V_k w_k||_1 \leq ||w_k||_1 ||V_k||_{\infty,\infty}$ for $k = 2, 3$. Denote $q_k = ||V_k||_{\infty,\infty}$ for

$k = 2, 3$. Then,

$$\begin{aligned}
|(x-x^\star)^T\nabla_i\nabla^2 h_y(z)(x-x^\star)| &\leq C_{f3}\sum_{k,j=0,2}||V_k w_k||_1\,||V_j w_j||_1 + \nu C_{g3}\sum_{k,j=0}^{3}||V_k w_k||_1\,||V_j w_j||_1 \\
&\leq 2C_{f3}\left[||V_0 w_0||_1^2 + ||V_2 w_2||_1^2\right] + 4\nu C_{g3}\sum_{k=0}^{3}||V_k w_k||_1^2 \\
&\leq 2p(C_{f3}+2\nu C_{g3})||w_0||_2^2 + 2q_2^2(C_{f3}+2\nu C_{g3})||w_2||_1^2 \\
&+ 4\nu p C_{g3}||w_1||_2^2 + 4\nu C_{g3}q_3^2||w_3||_1^2.
\end{aligned}$$

Therefore, using the constants κ_A and κ_B defined in the lemma, we have

$$\begin{aligned}
|\Delta_{00}(\delta)| &\leq \frac{1}{6}||x-x^\star||_1\max_i|(x-x^\star)^T\nabla_i\nabla^2 h_y(z)(x-x^\star)| \\
&\leq \frac{1}{6}\delta_+\left[(p||w_0||_2^2+\delta_2 q_2^2||w_2||_1)(2C_{f3}+4\nu C_{g3}) + 4\nu C_{g3}[p||w_1||_2^2+\delta_3 q_3^2||w_3||_1]\right], \\
&\leq \frac{\delta_+}{2}\left[\kappa_A||w_0||_2^2+\kappa_B\nu||w_1||_2^2+\kappa_A q_2^2\delta_2||w_2||_1/p+\nu\kappa_B q_3^2\delta_3||w_3||_1/p\right],
\end{aligned}$$

since $||x-x^\star||_1 = ||Vw||_1 \leq ||V||_\infty||w||_1 \leq ||V||_\infty[\sqrt{p_0}\delta_0+\sqrt{p_1}\delta_1+\delta_2+\delta_3] = \delta_+$.

1. *The upper bound.* Making the change of variables $v = \mathcal{S}(x-x^\star)$, we have

$$\begin{aligned}
[h_y(x)-h_y(x^\star)]/\sigma^2 &\leq a_y^T v_2 + b^T v_3 + \frac{1}{2}v_0^T H_{00}v_0 - v_0^T\nabla h_y(x^\star)/\sigma + \frac{1}{2}v_1^T B_{11}v_1 \\
&+ \sqrt{\nu}v_0^T B_{01}v_1 + \delta_+[\kappa_A||v_0||^2+\kappa_B||v_1||^2]/2 \\
&+ [\sigma v_0^T V_0 + \gamma v_1^T V_1]\,BV_3^T v_3 + \frac{\gamma^2}{2}v_3^T B_{33}v_3 \\
&+ [\sigma v_0^T V_0 + \gamma v_1^T V_1 + \gamma^2 v_2^T V_2]\,BV_3^T v_3 + \frac{\sigma^2}{2}v_2^T B_{22}v_2 \\
&+ \frac{\delta_+}{2}\left[\kappa_B q_3^2\delta_3||v_3||_1/p + \kappa_A q_2^2\delta_2||v_2||_1/p\right] \\
&\leq (b+\delta_b\mathbf{1})^T v_3 + (a_y+\delta_a\mathbf{1})^T v_2 \\
&+ \frac{1}{2}(v_1+\sqrt{\nu}\tilde{B}_{11}^{-1}B_{10}v_0)^T\tilde{B}_{11}(v_1+\sqrt{\nu}\tilde{B}_{11}^{-1}B_{10}v_0) \\
&+ \frac{1}{2}(v_0-\widetilde{H}_{00}^{-1}\nabla h_y(x^\star)/\sigma)^T\widetilde{H}_{00}(v_0-\widetilde{H}_{00}^{-1}\nabla h_y(x^\star)/\sigma) \\
&- \frac{1}{2\sigma^2}||\widetilde{H}_{00}^{-1/2}\nabla h_y(x^\star)||^2,
\end{aligned}$$

since $H_{jk} = \nu B_{jk}$ if at least one of j, k is 1 or 3.

Therefore, an upper bound is given by

$$\begin{aligned}[h_y(x)-h_y(x^\star)]/\sigma^2 &\le (b+\mathbf{1}\delta_b)^T v_3 + (a_y+\mathbf{1}\delta_a)_y^T v_2 + ||\widetilde{H}_{00}^{1/2}(v_0 - \widetilde{H}_{00}^{-1}\nabla h_y(x^\star)/\sigma)||^2/2 \\ &+ ||\widetilde{B}_{11}^{1/2}(v_1+\sqrt{\nu}\widetilde{B}_{11}^{-1}B_{10}v_0)||_2^2/2 - \frac{1}{2\sigma^2}||\widetilde{H}_{00}^{-1/2}\nabla h_y(x^\star)||^2.\end{aligned}$$

2. *The lower bound.* A similar argument leads to the following lower bound:

$$\begin{aligned}[h_y(x)-h_y(x^\star)]/\sigma^2 &\ge (b-\delta_b\mathbf{1})^T v_3 + (a_y-\delta_a\mathbf{1})^T v_2 + ||\bar{H}_{00}^{1/2}\left(v_0 - \bar{H}_{00}^{-1}\nabla h_y(x^\star)/\sigma\right)||^2/2 \\ &+ ||\bar{B}_{11}^{1/2}(v_1+\sqrt{\nu}\bar{B}_{11}^{-1}B_{10}v_0)||_2^2/2 - \frac{1}{2\sigma^2}||\bar{H}_{00}^{-1/2}\nabla h_y(x^\star)||^2.\end{aligned}$$

□

LEMMA 3. *In the notation of Lemma 4, introduce the following events*

$$\mathcal{A}_1 = \{\omega : ||V_2^T[\nabla f_{Y(\omega)}(x^\star) - \nabla f_{y_{\text{exact}}}(x^\star)]||_\infty \le 2\delta_a\}, \tag{30}$$

$$\mathcal{A}_2 = \{\omega : ||V_0^T[\nabla^2 f_{Y(\omega)}(x^\star) - \nabla^2 f_{y_{\text{exact}}}(x^\star)]V_0|| \le 2\kappa_A\delta_0\}. \tag{31}$$

Then, on $\mathcal{A}_1 \cap \mathcal{A}_2$, *for* $x \in B_\delta(x^\star)$,

$$\begin{aligned}[h_y(x)-h_y(x^\star)]/\sigma^2 &\le (a+\delta_A\mathbf{1})^T v_2 + (b+\delta_b\mathbf{1})^T v_3 + v_0^T\widetilde{\Omega}_{00}v_0/2 - v_0^T\nabla h_y(x^\star)/\sigma \\ &+ ||\widetilde{B}_{11}^{1/2}v_1||_2^2/2 + ||B_{10}||\delta_0\delta_1/\sigma^2,\end{aligned}$$

$$\begin{aligned}[h_y(x)-h_y(x^\star)]/\sigma^2 &\ge (a-\delta_A\mathbf{1})^T v_2 + (b-\delta_b\mathbf{1})^T v_3 + v_0^T\bar{\Omega}_{00}v_0/2 - v_0^T\nabla h_y(x^\star)/\sigma \\ &- ||\bar{B}_{11}^{1/2}v_1||_2^2/2 - ||B_{10}||\delta_0\delta_1/\sigma^2,\end{aligned}$$

where

$$\begin{aligned}\delta_A &= 3\delta_a + \nu||V_2^T\nabla g(x^\star)||, \\ \widetilde{\Omega}_{00} &= \Omega_{00} + (3\kappa_A\delta_0 + \nu\lambda_{\max}(B_{00}))I, \\ \bar{\Omega}_{00} &= \Omega_{00} - (3\kappa_A\delta_0 - \nu\lambda_{\min}(B_{00}))I.\end{aligned}$$

PROOF. There are two random leading terms in the expressions of the upper and the lower bounds, a_y and H_{00}. The lower bound has positive (or positive definite) coefficients if

$$\delta_0 > 0.5\kappa_A^{-1}||V_0^T[\nabla^2 f_y(x^\star) - \nabla^2 f_{y_{\text{exact}}}(x^\star)]V_0||, \quad \delta_0 > 0.5\nu\kappa_A^{-1}||B_{00}||$$
$$\delta_a > 0.5||V_2^T[\nabla f_y(x^\star) - \nabla f_{y_{\text{exact}}}(x^\star)]||, \quad \delta_a > 0.5\nu||V_2^T\nabla g(x^\star)||.$$

The random part of these conditions is satisfied on $\mathcal{A}_1 \cap \mathcal{A}_2$.

On $\mathcal{A}_1$,

$$a_y - a = V_2^T[\nabla f_y(x^\star) + \nu \nabla g(x^\star)] - V_2^T \nabla f_{y_{\text{exact}}}(x^\star) \le (2\delta_a + \nu ||V_2^T \nabla g(x^\star)||)\mathbf{1}$$

where the inequality is componentwise, and similarly $a_y - a \ge -(2\delta_a + \nu||V_2^T \nabla g(x^\star)||)\mathbf{1}$.

On event $\mathcal{A}_2$, Weyl's inequality implies

$$\begin{aligned}
\lambda_k(H_{00}) &\ge \lambda_k(V_0^T A^T \nabla^2 \tilde{f}_y(x^\star) A V_0) + \nu \lambda_{\min}(B_{00}) \ge \lambda_k(\Omega_{00}) - 2\kappa_A \delta_0 + \nu \lambda_{\min}(B_{00}), \\
\lambda_k(H_{00}) &\le \lambda_k(\Omega_{00}) + 2\kappa_A \delta_0 + \nu \lambda_{\max}(B_{00}),
\end{aligned}$$

where $\lambda_k(M)$ is the kth largest eigenvalue of matrix M. Applying these inequalities to the bounds in Lemma 4, we obtain the statement of the lemma.

□

A.3. Proofs, Section 5.3.

Proof of Theorem 1. Consider a neighbourhood of $x^\star$, $B_\delta(x^\star) = (x^\star + VB_\delta) \cap \mathcal{X}$, where $B_\delta = B_2(0,\delta_0) \times B_2(0,\delta_1) \times B_\infty(0,\delta_2) \times B_\infty(0,\delta_3)$. Denote $v = D_{\sigma,\gamma}^{-1} V^{-1}(x - x^\star)$, with the Jacobian of this change of variables being $J = \sigma^{-p_0 - 2p_2} \gamma^{-\tilde{p}_1 - 2\tilde{p}_3} / \det(V)$.

Denote ρ_k the smallest rate such that for some $M_k > 0$, as $\sigma \to 0$,

$$\begin{aligned}
\mathbb{P}\{|f_{Y(\omega)}(x^\star) - f_{y_{\text{exact}}}(x^\star)]| > M_0 \rho_0\} &\to 0, \\
\mathbb{P}\{||V_2^T[\nabla f_{Y(\omega)}(x^\star) - \nabla f_{y_{\text{exact}}}(x^\star)]||_\infty > M_1 \rho_1\} &\to 0, \\
\mathbb{P}\{||V_0^T[\nabla^2 f_{Y(\omega)}(x^\star) - \nabla^2 f_{y_{\text{exact}}}(x^\star)]|| > M_2 \rho_2\} &\to 0.
\end{aligned}$$

Due to Assumption C, $\rho_k \to 0$ as $\sigma \to 0$.

For the rescaled parameter, we use the corresponding neighbourhood B_R and its limit $B_R^\star$ defined by

$$\begin{aligned}
B_R &= [B_2(0,R_0) \times B_2(0,R_1) \times [0,R_2]^{p_2} \times [0,R_3]^{p_3}] \cap D^{-1} V^{-1}(\mathcal{X} - x^\star), \\
B_R^\star &= [B_2(0,R_0) \times B_2(0,R_1) \times [0,R_2]^{p_2} \times [0,R_3]^{p_3}] \cap \mathcal{V}^\star
\end{aligned}$$

where

$$R_0 = \delta_0/\sigma, \quad R_1 = \delta_1/\gamma, \quad R_2 = \delta_2/\sigma^2, \quad R_3 = \delta_3/\gamma^2.$$

We assume that δ_k are such that $\delta_k \to 0$ and $R_k \to \infty$. In addition, we will need conditions

$$R_0 = o(\gamma/\sigma), \quad R_2 = o(\gamma/\sigma^2), \quad R_3 = o(1/\gamma), \quad R_2 = o(1/\sigma),$$

and $R_2 = o(\gamma^2/\sigma^2)$ if $c > 0$. These conditions hold, for instance, with $R_k = C_k[-\log\sigma]^{a_k}$ for some positive constants C_k and a_k. Under these conditions, $C(\delta)$ defined in Lemma 6 tends to 0 as $\sigma \to 0$, that will be used below.

The triangle inequality for the total variation norm gives us

$$\begin{aligned} ||\mathbb{P}_{\mathcal{S}(x-x^\star)|Y} - \mu^\star||_{TV} &\leq ||\mathbb{P}_{\mathcal{S}(x-x^\star)|Y}\mathbf{1}_{B_R^\star} - \mu^\star\mathbf{1}_{B_R^\star}||_{TV} \\ (32) \qquad &+ ||\mu^\star\mathbf{1}_{B_R^\star} - \mu^\star||_{TV} + ||\mathbb{P}_{\mathcal{S}(x-x^\star)|Y}\mathbf{1}_{B_R^\star} - \mathbb{P}_{\mathcal{S}(x-x^\star)|Y}||_{TV}, \end{aligned}$$

where the balls $B_R^\star$ are defined above. Here $\mu\mathbf{1}_{B_R^\star}$ is a probability measure μ truncated to $B_R^\star$ and normalised to be a probability measure.

If measures μ_1, μ_2 are absolutely continuous with respect to some measure μ with densities f and g respectively, then the total variation norm can also be written as

$$||\mu_1 - \mu_2||_{TV} = 2\int_{\mathcal{X}} (f-g)_+ d\mu,$$

where $(x)_+ = \max(x, 0)$ (van der Vaart 1998). In each of the summands in the upper bound (33), the first measure is absolutely continuous with respect to the second one, so we will use this expression to evaluate the total variation norm.

We start with the distance between the truncations of the rescaled posterior distribution and the limit on $B_R^\star$. Introduce additional notation:

$$\begin{aligned} \widetilde{a} &= a + \delta_A\mathbf{1}, \quad \widetilde{b} = b + \delta_b\mathbf{1}, \\ \bar{a} &= a - \delta_A\mathbf{1}, \quad \bar{b} = b - \delta_b\mathbf{1}, \end{aligned}$$

and $\mu(\cdot; x, \Omega, B, a, b)$ is a measure on $\mathcal{V} = \mathbb{R}^{p_0+p_1} \times \mathbb{R}_+^{p_2+p_3}$, such that for $z = (z_1, z_2) \in \mathcal{V}$,

$$\mu(dz; \alpha, \Sigma, \beta) = \exp\left\{-||\Sigma^{1/2}z_1||^2/2 + z_1^T\alpha - \beta^T z_2\right\} dz.$$

This measure is finite if matrix Σ is positive definite, α is finite and all components of vector β are positive. If $\mathcal{B}_v = \mathcal{V} = \mathbb{R}^{p_0+p_1} \times \mathbb{R}_+^{p_2+p_3}$ or $\mathcal{B}_v = \mathcal{B}_1 \times B_\infty(0, r_2)$, we have

$$\begin{aligned} \mu(\mathcal{V}; \alpha, \Sigma, \beta) &= \prod_{i=1}^{p_2+p_3} \beta_i^{-1} \left[\det(\Sigma)\right]^{-1/2} (2\pi)^{(p_0+p_1)/2} \exp\{\alpha^T\Sigma^{-1}\alpha/2\}, \\ \mu(\mathcal{B}_1 \times B_\infty(0, r_2); \alpha, \Sigma, \beta) &= \mu(\mathcal{V}; \alpha, \Sigma, \beta)\, \Phi(\mathcal{B}_1; \Sigma^{-1}\alpha, \Sigma^{-1}) \prod_{i=1}^{p_2+p_3} [1 - \exp\{-\beta_i r_2\}], \end{aligned}$$

where $\Phi(\mathcal{B}; a, Q)$ is the measure of $\mathcal{B}$ under the Gaussian distribution centred at a with covariance matrix Q.

If $\mathcal{B}_v$ degenerates to a manifold of a smaller dimension, then we slightly abuse the notation and assume that μ has mass 1 on the degenerate part of $\mathcal{B}_v$, i.e. we replace the Lebesque measure in the definition of μ with the counting measure on the degenerate part.

By Lemma 5, on $\mathcal{A}_1 \cap \mathcal{A}_2$ for any $\mathcal{B}_x \subseteq B_\delta(x^\star)$, with $\mathcal{B}_v = D^{-1}V^{-1}(\mathcal{B}_x - x^\star) \subseteq B_R$, we have

$$\begin{aligned}
&\int_{\mathcal{B}_x} \exp\left\{-[h_y(x) - h_y(x^\star)]/\sigma^2\right\} dx \geq J \int_{\mathcal{B}_v} \exp\left\{-\widetilde{b}^T v_3 - \widetilde{a}^T v_2\right\} \\
&\times \exp\left\{-||\widetilde{\Omega}_{00}^{1/2} v_0||^2/2 + v_0^T \nabla h_y(x^\star)/\sigma - ||\widetilde{B}_{11}^{1/2} v_1||_2^2/2 - \sqrt{\nu} v_0^T B_{01} v_1\right\} dv \\
= \;& J\,\mu(\mathcal{B}_v; \alpha_y, \widetilde{\Sigma}, \widetilde{\beta}),
\end{aligned}$$

where $\alpha_y = ([\nabla h_y(x^\star)]^T/\sigma, 0)^T$, $\widetilde{\beta} = (\widetilde{a}^T, \widetilde{b}^T)^T$, $\widetilde{\Sigma} = \begin{pmatrix} \widetilde{\Omega}_{00} & \sqrt{\nu} B_{01} \\ \sqrt{\nu} B_{10} & \widetilde{B}_{11} \end{pmatrix}$.

Similarly, using Lemma 5, we obtain an upper bound:

$$\begin{aligned}
&\int_{\mathcal{B}_x} \exp\left\{-[h_y(x) - h_y(x^\star)]/\sigma^2\right\} dx \leq J \int_{\mathcal{B}_v} \exp\left\{-\bar{a}^T v_2 - \bar{b}^T v_3\right\} \\
&\times \exp\left\{-||\bar{\Omega}_{00}^{1/2} v_0||^2/2 + v_0^T \nabla h_y(x^\star)/\sigma - ||\bar{B}_{11}^{1/2} v_1||^2/2 - \sqrt{\nu} v_0^T B_{01} v_1\right\} dv \\
= \;& J\mu(\mathcal{B}_v; \alpha_y, \bar{\Sigma}, \bar{\beta}),
\end{aligned}$$

where $\bar{\beta} = (\bar{a}^T, \bar{b}^T)^T$ and $\bar{\Sigma} = \begin{pmatrix} \bar{\Omega}_{00} & \sqrt{\nu} B_{01} \\ \sqrt{\nu} B_{10} & \bar{B}_{11} \end{pmatrix}$.

To simplify the notation, denote

$$\bar{\mu}(dv) = \mu(dv; \alpha_y, \bar{\Sigma}, \bar{\beta}), \quad \widetilde{\mu}(dv) = \mu(dv; \alpha_y, \widetilde{\Sigma}, \widetilde{\beta}).$$

Define event $\mathcal{A}_3$:

$$\mathcal{A}_3 = \left\{\omega : ||\Omega_{00} a_0(\omega) - \sigma/\gamma^2\, V_0^T \nabla g(x^\star) - \sigma^{-1} V_0^T \nabla f_{Y(\omega)}(x^\star)||_\infty \leq \rho\right\}$$

where ρ is the smallest value such that $\mathbb{P}(\mathcal{A}_3) \to 0$ as $\sigma \to 0$. On $\mathcal{A}_3$, $||\nabla h_{Y(\omega)}(x^\star)/\sigma - \Omega_{00} a_0(\omega)||_\infty \leq \rho$. Therefore, on $\mathcal{A}_3$, measure $\widetilde{\mu}$ is finite since α_y is finite, and all other parameters are positive or positive definite. Measure $\bar{\mu}$ is finite on $\mathcal{A}_3$ if $\delta_b < b_{\min}$, $\delta_A < a_{\min}$, $\delta_1 < \lambda_{\min}(B_{11})/\kappa_B$ and $\delta_0 < \lambda_{\min}(\Omega_{00})/(3\kappa_A)$.

Hence, the posterior density of $\mathcal{S}(x - x^\star)$ normalised by the posterior measure of $B_R^\star$ is bounded on $\mathcal{A}_1 \cap \mathcal{A}_2$ by

$$\frac{\widetilde{\mu}(dv)}{\bar{\mu}(B_R^\star)} \leq \frac{d\,p(\mathcal{S}(x - x^\star) \mid Y)}{p(B_R^\star \mid Y)} \leq \frac{\bar{\mu}(dv)}{\widetilde{\mu}(B_R^\star)}.$$

Therefore, the total variation distance between the rescaled posterior distribution and its limit, both truncated to $B_R^\star$, is bounded on $\mathcal{A}_1 \cap \mathcal{A}_2$ by

$$\begin{aligned} ||\mathbb{P}_{\mathcal{S}(x-x^\star)|Y}\mathbf{1}_{B_R^\star} - \mu^\star \mathbf{1}_{B_R^\star}||_{TV} &\leq 2\int_{B_R^\star}\left[\frac{\mathbb{P}(dv \mid Y)\mu^\star(B_R^\star)}{\mathbb{P}(B_R^\star \mid Y)\mu^\star(dv)} - 1\right]_+ \frac{\mu^\star(dv)}{\mu^\star(B_R^\star)} \\ &\leq 2\int_{B_R^\star}\left[\frac{\bar{\mu}(dv)}{\widetilde{\mu}(B_R^\star)}\frac{\mu^\star(B_R^\star)}{\mu^\star(dv)} - 1\right]_+ \frac{\mu^\star(dv)}{\mu^\star(B_R^\star)} \end{aligned}$$

Now, $\frac{\mu^\star(dv)}{\mu^\star(B_R^\star)} = \frac{\mu(dv;\Sigma^\star,\alpha^\star,\beta^\star)}{\mu(B_R^\star;\Sigma^\star,\alpha^\star,\beta^\star)}$ where $\alpha^\star = (a_0^T\Omega_{00}, 0)^T$, $\beta^\star = (a^T, b^T)^T$, $\Sigma^\star = \mathrm{diag}(\Omega_{00}, B_{11})$. Denote $\mu_0(dv) = \mu(dv;\Sigma^\star,\alpha^\star,\beta^\star)$. Then, with $v_{01} = (v_0^T, v_1^T)^T$,

$$\frac{\bar{\mu}(dv)}{\mu_0(dv)} = \exp\{\delta_A \mathbf{1}^T v_2 + \delta_b \mathbf{1}^T v_3 + v_{01}^T(\Sigma^\star - \bar{\Sigma})v_{01}/2 + v_{01}^T(\alpha_y - \alpha^\star)\},$$

which implies

$$\begin{aligned} \frac{\bar{\mu}(dv)}{\mu_0(dv)}\frac{\mu_0(B_R^\star)}{\widetilde{\mu}(B_R^\star)} &= \exp\{(v_{01} + (\Sigma^\star - \bar{\Sigma})^{-1}(\alpha_y - \alpha^\star))^T(\Sigma^\star - \bar{\Sigma})(v_{01} + (\Sigma^\star - \bar{\Sigma})^{-1}(\alpha_y - \alpha^\star))/2\} \\ &\times \exp\{\delta_A \mathbf{1}^T v_2 + \delta_b \mathbf{1}^T v_3 - (\alpha_y - \alpha^\star)^T(\Sigma^\star - \bar{\Sigma})^{-1}(\alpha_y - \alpha^\star)/2\} \\ &\times \frac{\int_{B_R^\star}\exp\{-a^T v_2 - b^T v_3\}\exp\{-||\Sigma^{\star\,1/2}v_{01}||^2/2 + v_{01}^T\alpha^\star\}dv}{\int_{B_R^\star}\exp\{-(a+\delta_A\mathbf{1})^T v_2 - (b+\delta_b\mathbf{1})^T v_3\}\exp\{-||\widetilde{\Sigma}^{1/2}v_{01}||^2/2 + v_{01}^T\alpha_y\}dv}. \end{aligned}$$

To show that this expression is greater than 1, it is sufficient to show that for any $\mathcal{B} \subseteq \{v_{01} : (v_{01}, v_2, v_3) \in B_R^\star\}$, the following expression is positive:

$$\begin{aligned} & e^{-(\alpha_y-\alpha^\star)^T(\Sigma^\star-\bar{\Sigma})^{-1}(\alpha_y-\alpha^\star)/2}\int_{\mathcal{B}} e^{-||\Sigma^{\star\,1/2}z||^2/2 + z^T\alpha^\star\}dz - \int_{\mathcal{B}}\exp\{-||\widetilde{\Sigma}^{1/2}z||^2/2 + z^T\alpha_y}dz \\ = & \int_{\mathcal{B}} e^{-||\widetilde{\Sigma}^{1/2}z||^2/2 + z^T\alpha_y}\left[e^{(\alpha_y-\alpha^\star)^T[(\bar{\Sigma}-\Sigma^\star)^{-1} - (\widetilde{\Sigma}-\Sigma^\star)^{-1}](\alpha_y-\alpha^\star)/2}\times\right. \\ & \left.\times\, e^{z^T(\widetilde{\Sigma}-\Sigma^\star)z/2 + z^T(\alpha^\star-\alpha_y) + (\alpha^\star-\alpha_y)^T(\widetilde{\Sigma}-\Sigma^\star)^{-1}(\alpha^\star-\alpha_y)/2} - 1\right]dz > 0 \end{aligned}$$

which is indeed the case since

$$\begin{aligned} \widetilde{\Sigma} - \Sigma^\star &= \mathrm{diag}\left([3\kappa_A\delta_0 + \nu\lambda_{\max}(B_{00})]I_{p_0},\, \delta_1\kappa_B I_{\tilde{p}_1}\right), \\ (\widetilde{\Sigma} - \Sigma^\star) - (\bar{\Sigma} - \Sigma^\star) &= \widetilde{\Sigma} - \bar{\Sigma} = \mathrm{diag}\left([6\kappa_A\delta_0 + \nu(\lambda_{\max}(B_{00}) - \lambda_{\min}(B_{00}))]I_{p_0},\, 2\delta_1\kappa_B I_{\tilde{p}_1}\right) \end{aligned}$$

are positive definite matrices.

Therefore, on $\mathcal{A}_1 \cap \mathcal{A}_2$, $\frac{\bar{\mu}(dv)}{\mu_0(dv)}\frac{\mu_0(B_R^\star)}{\widetilde{\mu}(B_R^\star)} \geq 1$ and hence

$$||\mathbb{P}_{\mathcal{S}(x-x^\star)|Y}\mathbf{1}_{B_R^\star} - \mu^\star\mathbf{1}_{B_R^\star}||_{TV} \leq 2\int_{B_R^\star}\left[\frac{\bar{\mu}(dv)}{\widetilde{\mu}(B_R^\star)}\frac{\mu^\star(B_R^\star)}{\mu^\star(dv)} - 1\right]\frac{\mu^\star(dv)}{\mu^\star(B_R^\star)} = 2\left[\frac{\bar{\mu}(B_R^\star)}{\widetilde{\mu}(B_R^\star)} - 1\right].$$

Since $\bar{\mu}(dv)/dv \geq \widetilde{\mu}(dv)/dv$ for all $v \in B_R$, $\bar{\mu}(B_R^\star) - \widetilde{\mu}(B_R^\star) \leq \bar{\mu}(B_R) - \widetilde{\mu}(B_R)$, and thus

$$\frac{\bar{\mu}(B_R^\star)}{\widetilde{\mu}(B_R^\star)} - 1 \quad \leq \quad \frac{\bar{\mu}(B_R) - \widetilde{\mu}(B_R)}{\widetilde{\mu}(B_R^\star)} = \frac{\bar{\mu}(B_R) - \widetilde{\mu}(B_R)}{\widetilde{\mu}(B_R^\star)}.$$

The difference of measures of B_R is bounded by

$$\begin{aligned}
\bar{\mu}(B_R) - \widetilde{\mu}(B_R) &= \int_{B_R} e^{-z_1^T\tilde{\Sigma}z_1/2 + z_1^T\alpha_y - \widetilde{\beta}^T z_2} \left[exp\left\{ z_1^T(\tilde{\Sigma} - \bar{\Sigma})z_1/2 + (\widetilde{\beta} - \bar{\beta})^T z_2 \right\} - 1 \right] dz \\
&\leq \int_{B_R} \left[z_1^T(\tilde{\Sigma} - \bar{\Sigma})z_1/2 + (\widetilde{\beta} - \bar{\beta})^T z_2 \right] e^{-z_1^T\bar{\Sigma}z_1/2 + z_1^T\alpha_y - \bar{\beta}^T z_2} dz \\
&\leq (2\pi)^{(p_0+\tilde{p}_1)/2} [\det(\bar{\Sigma})]^{-1/2} e^{\alpha_y^T\bar{\Sigma}^{-1}\alpha_y/2} \prod \bar{\beta}_i^{-1} \\
&\quad \times \left[||(\tilde{\Sigma} - \bar{\Sigma})^{1/2}\bar{\Sigma}^{-1}\alpha_y||^2 + \mathrm{trace}(\bar{\Sigma}^{-1}(\tilde{\Sigma} - \bar{\Sigma})) + \frac{2\delta_A p_2}{\min_i \bar{a}_i} + \frac{2\delta_b \tilde{p}_3}{\min_i \bar{b}_i} \right] \\
&\leq (2\pi)^{(p_0+\tilde{p}_1)/2} [\det(\bar{\Sigma})]^{-1/2} e^{\alpha_y^T\bar{\Sigma}^{-1}\alpha_y/2} \prod \bar{\beta}_i^{-1} \\
&\quad \times \left[\delta_{01}[||\bar{\Sigma}^{-1}\alpha_y||^2 + \mathrm{trace}(\bar{\Sigma}^{-1})] + \frac{2\delta_A p_2}{\min_i \bar{a}_i} + \frac{2\delta_b \tilde{p}_3}{\min_i \bar{b}_i} \right]
\end{aligned}$$

due to inequality $e^x - 1 \leq xe^x$ for $x > 0$, where

$$\delta_{01} = \max([6\kappa_A\delta_0 + \nu(\lambda_{\max}(B_{00}) - \lambda_{\min}(B_{00}))],\, 2\delta_1\kappa_B).$$

Therefore,

$$\begin{aligned}
||\mathbb{P}_{\mathcal{S}(x-x^\star)|Y}\mathbf{1}_{B_R^\star} - \mu^\star\mathbf{1}_{B_R^\star}||_{TV} &\leq 2[\widetilde{\mu}(B_R^\star)]^{-1}(2\pi)^{(p_0+\tilde{p}_1)/2}[\det(\bar{\Sigma})]^{-1/2} e^{\alpha_y^T\bar{\Sigma}^{-1}\alpha_y/2} \prod \bar{\beta}_i^{-1} \\
&\quad \times \left[\delta_{01}[||\bar{\Sigma}^{-1}\alpha_y||^2 + \mathrm{trace}(\bar{\Sigma}^{-1})] + \frac{2\delta_A p_2}{a_{\min} - \delta_A} + \frac{2\delta_b \tilde{p}_3}{b_{\min} - \delta_b} \right]
\end{aligned}$$

which goes to zero since $\delta_k \to 0$ as $\sigma \to 0$.

The total variation distance between the posterior distribution truncated to B_R and to $B_R^\star$ is bounded by

$$\begin{aligned}
||\mathbb{P}_{\mathcal{S}(x-x^\star)|Y}\mathbf{1}_{B_R^\star} - \mathbb{P}_{\mathcal{S}(x-x^\star)|Y}\mathbf{1}_{B_R}||_{TV} &\leq 2[\mathbb{P}_{\mathcal{S}(x-x^\star)|Y}\mathbf{1}_{B_R}](B_R \setminus B_R^\star) \\
&\leq \frac{2\mu_L(B_R \setminus B_R^\star)\max_{v\in B_R}[\bar{\mu}(dv)/dv]}{\widetilde{\mu}(B_R)},
\end{aligned}$$

where $\mu_L(B)$ is the Lebesgue measure of B. By Lemma 6, $\mu_L(B_R \setminus B_R^\star) \to 0$ as $\sigma \to 0$.

For the measure $\mu(dz;\alpha,\Sigma,\beta)$, $\max_{z\in\mathcal{V}}[\mu(dz;\alpha,\Sigma,\beta)/dz] = \exp\{\alpha^T\Sigma^{-1}\alpha/2\}$. Therefore,

$$\frac{\max_{v\in B_R}[\bar{\mu}(dv)/dv]}{\widetilde{\mu}(B_R)} \le e^{\alpha_y^T\widetilde{\Sigma}^{-1}(\widetilde{\Sigma}-\bar{\Sigma})\bar{\Sigma}^{-1}\alpha_y/2}\prod_{i=1}^{p_2+\tilde{p}_3}\widetilde{\beta}_i\left[\det(\widetilde{\Sigma})\right]^{1/2}(2\pi)^{-(p_0+\tilde{p}_1)/2}\left[\frac{\widetilde{\mu}(B_R)}{\widetilde{\mu}(\mathcal{V})}\right]^{-1}.$$

The difference $\widetilde{\Sigma}-\bar{\Sigma} = O(\max(\delta_0,\delta_1)) \to 0$, on $\mathcal{A}_3$, $\bar{\Sigma}^{-1}\alpha_y$ and $\widetilde{\Sigma}^{-1}\alpha_y$ are bounded as well as $\widetilde{\beta}$. Now we need to show that the last factor is bounded.

Choose σ small enough so that $B_R = B(0,R_0)\times B(0,R_1)\times[0,R_2]^{p_2}\times[0,R_3]^{p_3}$. This condition is satisfied if $||U_kx^\star|| \ge \delta_k$ for $k=0,1$ and $||U_kx^\star||_\infty \ge \delta_k$ for $k=2,3$. If the degeneration of the support takes place, the degenerate dimensions are excluded. Due to $B(0,\sqrt{R_0^2+R_1^2}) \subset B(0,R_0)\times B(0,R_1)$ and to inequality $1-\Phi(B(0,R);\alpha,\Sigma^{-1}) \le 1-\Gamma\left(\frac{(R-||\alpha||)^2}{2||\Sigma||} \mid \frac{m}{2}\right)$, we have

$$\begin{aligned}\frac{\widetilde{\mu}(B_R)}{\widetilde{\mu}(\mathcal{V})} &\ge \Gamma\left(\frac{\lambda_{\min}(\widetilde{\Sigma})(\sqrt{R_0^2+R_1^2}-||\widetilde{\Sigma}^{-1}\alpha_y||)^2}{2} \mid \frac{p_0+\tilde{p}_1}{2}\right)\\ &\quad\times\prod_{i=1}^{p_2}[1-\exp\{-\widetilde{a}_iR_2\}]\prod_{i=1}^{\tilde{p}_3}[1-\exp\{-\widetilde{b}_iR_3\}]\end{aligned}$$

which is close to 1 for large R_k, $k=0,1,2,3$. Therefore, $||\mathbb{P}_{\mathcal{S}(x-x^\star)|Y}\mathbf{1}_{B_R^\star} - \mathbb{P}_{\mathcal{S}(x-x^\star)|Y}\mathbf{1}_{B_R}||_{TV} \to 0$ as $\sigma\to 0$.

The total variation distance between the limit measure and its truncation to $B_R^\star$ is bounded by

$$\begin{aligned}||\mu^\star - \mu^\star\mathbf{1}_{B_R^\star}||_{TV} &\le 2\mu^\star(\bar{B}_R^\star) = 2\mu^\star(\bar{B}_R) + 2\mu^\star(B_R\setminus B_R^\star)\\ &\le 2\mu_L(B_R\setminus B_R^\star)\,[\det(\Sigma^\star)]^{1/2}\,(2\pi)^{-(p_0+\tilde{p}_1)/2}\prod_{i=1}^{p_2+\tilde{p}_3}\beta_i^\star\\ &+ 2-2\,\Gamma\left(\frac{\lambda_{\min}(\Omega_{00})(R_0-||a_0||)^2}{2} \mid \frac{p_0}{2}\right)\times\Gamma\left(\frac{\lambda_{\min}(B_{11})R_1^2}{2} \mid \frac{\tilde{p}_1}{2}\right)\times\\ &\quad\times\prod_{i=1}^{p_2}[1-\exp\{-a_iR_2\}]\prod_{i=1}^{\tilde{p}_3}[1-\exp\{-b_iR_3\}] \to 0\end{aligned}$$

as $\sigma\to 0$, since $R_k\to\infty$ and $\mu_L(B_R\setminus B_R^\star)\to 0$ as $\sigma\to 0$ by Lemma 6 where $\bar{B}_R^\star$ is the complement of $B_R^\star$.

The total variation distance between the posterior distribution and its

truncation to B_R is bounded by

$$\begin{aligned}||\mathbb{P}_{(\mathcal{S}(x-x^\star)|Y)}\mathbf{1}_{B_R} - \mathbb{P}_{(\mathcal{S}(x-x^\star)|Y)}||_{TV} &\leq 2\mathbb{P}_{(\mathcal{S}(x-x^\star)|Y)}(\bar{B_R}) \\ &= 2\frac{\int_{\mathcal{X}\setminus B_\delta(x^\star)} \exp\{-(h_y(x)-h_y(x^\star))/\sigma^2\}\,dx}{\int_{\mathcal{X}} \exp\{-(h_y(x)-h_y(x^\star))/\sigma^2\}\,dx} \\ &\leq \frac{2\det(V)[\widetilde{\mu}(B_R)]^{-1}\Delta_0(\delta)}{1+\det(V)[\widetilde{\mu}(B_R)]^{-1}\Delta_0(\delta)},\end{aligned}$$

where

$$\Delta_0(\delta) = \sigma^{p_0+2p_2}\gamma^{\tilde{p}_1+2\tilde{p}_3}\int_{\mathcal{X}\setminus B_\delta(x^\star)} \exp\{-[h_y(x)-h_y(x^\star)]/\sigma^2\}\,dx.$$

By Assumption L, with probability $\to 0$, $\Delta_0(\delta)\to 0$ as $\sigma\to 0$, and $\widetilde{\mu}(B_R)\to \mu_0(B_R)>0$.

Combining these bounds, we have that

$$\begin{aligned}||\mathbb{P}_{\mathcal{S}(x-x^\star)|Y} - \mu^\star||_{TV} &\leq 2\mu^\star(\bar{B_R}) + 2\det(V)\Delta_0(\delta)/\widetilde{\mu}(B_R) \\ &+ 2[\widetilde{\mu}(B_R)]^{-1}(2\pi)^{(p_0+\tilde{p}_1)/2}[\det(\bar{\Sigma})]^{-1/2}e^{\alpha_y^T\bar{\Sigma}^{-1}\alpha_y/2}\prod\bar{\beta}_i^{-1} \\ &\quad\times\left[\delta_{01}[||\bar{\Sigma}^{-1}\alpha_y||^2+\mathrm{trace}(\bar{\Sigma}^{-1})] + \frac{2\delta_A p_2}{a_{\min}-\delta_A} + \frac{2\delta_b\tilde{p}_3}{b_{\min}-\delta_b}\right] \\ &+ 2\mu_L(B_R\setminus B_R^\star)\left[\max_{v\in B_R}[\mu^\star(dv)/dv] + \frac{\max_{v\in B_R}[\bar{\mu}(dv)/dv]}{\widetilde{\mu}(B_R)}\right] \overset{\mathbb{P}_{x_{\text{true}}}}{\to} 0\end{aligned}$$

as $\sigma\to 0$, which gives the statement of the theorem.

□

LEMMA 4. *Under conditions of Theorem 1,* $\mu_L(D^{-1}V^{-1}(B_\delta(x^\star)-x^\star)\setminus\mathcal{V}^\star)\leq C(\delta)\det(V^{-1})$ *where*

$$C(\delta) = [||V_0||_{\infty,2}\delta_0/\gamma + ||V_2||_{\infty,\infty}\delta_2/\gamma + ||V_3||_\infty\delta_3/\gamma]^{|S_1|} + \begin{cases}[||V_2||_{\infty,\infty}\,\delta_2/\sigma]^{|S_0|}, & c=0,\\ [||V_2||_{\infty,\infty}\,\delta_2/\gamma^2]^{|S_0|+|S_3|}, & c>0,\end{cases}$$

sets S_k are defined in Proposition 2 and μ_L is the Lebesgue measure.

PROOF. We study the constraints on v_k under $D^{-1}V^{-1}(B_\delta(x^\star)-x^\star)\setminus\mathcal{V}^\star$ using notation from the proof of Theorem 1.

We find the Lebesgue measure of $VB_R\setminus(V\mathcal{V}^\star)$, then the Lebesgue measure of $B_R\setminus\mathcal{V}^\star$ is $\det(V^{-1})\mu_L(VB_R\setminus(V\mathcal{V}^\star))$. For $\ell\in S_1$,

$$[V_1]_\ell\, v_1 \geq -\sigma/\gamma[V_0]_\ell\, v_0 - \sigma^2/\gamma[V_2]_\ell\, v_2 - \gamma[V_3]_\ell\, v_3 \quad \textit{and} \quad [V_1]_\ell\, v_1 < 0.$$

which is a subset of $-R_0\sigma/\gamma||[V_0]_{\ell,}||_2-\sigma^2R_2/\gamma||[V_2]_{\ell,}||_\infty-\gamma R_3[V_3]_\ell \le [V_1]_{\ell,}\, v_1 < 0$. This set could be empty, or its Lebesgue measure could be up to $[R_0\sigma/\gamma||V_0||_{\infty,2}+\sigma^2R_2/\gamma||V_2||_{\infty,\infty}+\gamma R_3[V_3]_\infty]^{|S_1|}$. If R_0 is chosen in such a way that $R_0\sigma/\gamma \to 0$, $\sigma^2R_2/\gamma \to 0$ and $\gamma R_3 \to 0$ (which is possible), the measure of this set tends to 0.

For $\ell \in S_3$ and $c = \lim \sigma/\gamma^2 = 0$, $0 > [V_3]_\ell\, v_3 \ge -[V_0]_{\ell,}\, v_0\sigma/\gamma^2$. This set is empty, since either $p_3 = 0$, or $p_3 = 1$ and $v_3 \ge 0$ and $[V_3]_\ell > 0$.

For $\ell \in S_{03}$ and $c > 0$, $0 > [V_3]_\ell\, v_3 + [V_0]_{\ell,}\, v_0 \ge -[V_2]_{\ell,}\, v_2\sigma^2/\gamma^2 \ge -||V_2||_{\infty,\infty}\, R_2\sigma^2/\gamma^2$. This set is either empty, or its Lebesgue measure is at most $[||V_2||_{\infty,\infty}\, R_2\sigma^2/\gamma^2]^{|S_0|+|S_3|}$ provided $R_2\sigma^2/\gamma^2 \to 0$.

For $\ell \in S_0$ and $c = \lim \sigma/\gamma^2 = 0$, $0 > [V_0]_{\ell,}\, v_0 \ge -[V_2]_{\ell,}\, v_2\sigma \ge -||V_2||_{\infty,\infty}\, R_2\sigma$. This set is either empty, or its Lebesgue measure is at most $[||V_2||_{\infty,\infty}\, R_2\sigma]^{|S_0|}$ provided $R_2\sigma \to 0$.

For $\ell \in S^{*c}$ such that $[V_1]_{\ell,} \ne 0$, the constraints are

$$[V_1]_{\ell,}v_1 \ge -x^\star_\ell/\gamma + \sigma/\gamma[V_0]_{\ell,}v_0 \ge -x^\star_\ell/\gamma + \sigma/\gamma[V_0]_{\ell,}v_0,$$

and $[V_0]_{\ell,}v_0 \ge -x^\star_\ell/\sigma$ if $[V_1]_{\ell,} = 0$. Since there is no constraints on $[V_1]_{\ell,}v_1$ and $V_0]_{\ell,}v_0$ for $\ell \in S^{*c}$, the difference $V_{S^{*c},}B_R \setminus (V\mathcal{V}^\star)$ is empty.

Therefore, the Lebesgue measure of $VB_R \setminus (V\mathcal{V}^\star)$ is at most

$$[R_0\sigma/\gamma||V_0||_{\infty,2}+\sigma^2R_2/\gamma||V_2||_{\infty,\infty}+\gamma R_3||V_3||_\infty]^{|S_1|}+[||V_2||_{\infty,\infty}\, R_2\sigma^2/\gamma^2]^{|S_0|+|S_3|}$$

if $c > 0$, and is at most

$$[R_0\sigma/\gamma||V_0||_{\infty,2} + \sigma^2R_2/\gamma||V_2||_{\infty,\infty} + \gamma R_3||V_3||_\infty]^{|S_1|} + [||V_2||_{\infty,\infty}\, R_2\sigma]^{|S_0|}$$

if $c = 0$.

□

Proof of Proposition 3. Collecting the non-asymptotic conditions on δ_k in the proofs of Theorem 1 and Proposition 2, we have

$$\begin{gathered}
0.5\nu||B_{00}||/\kappa_A < \delta_0 < 0.2\lambda_{\min}(\Omega_{00})/\kappa_A,\\
\delta_k \le ||U_kx^\star||,\ k = 0, 1; \quad \delta_k \le ||U_kx^\star||_\infty,\ k = 2, 3,\\
0.5\nu||V_2^T\nabla g(x^\star)|| < \delta_a < 0.2a_{\min},\\
\delta_1 < \lambda_{\min}(B_{11})/\kappa_B, \quad \delta_b < b_{\min},
\end{gathered}$$

and inequality $\delta_0 > \sigma||a_0(\omega)||$ should hold with high probability.

These assumptions imply that $\delta_A \le 5\delta_a$ and

$$||\widetilde{\Omega}_{00} - \bar{\Omega}_{00}|| = 6\kappa_A\delta_0 + \nu(\lambda_{\max}(B_{00}) - \lambda_{\min}(B_{00})) \le 10\kappa_A\delta_0.$$

Collecting the upper bounds on the total variation distance from the proof of Theorem 1 and using the upper bound on $\mu_L(B_R \setminus B_R^\star)$ given in Lemma 6, on $\mathcal{A}_1 \cap \mathcal{A}_2 \cap \mathcal{A}_3$, we have

$$
\begin{aligned}
&||\mathbb{P}_{\mathcal{S}(x-x^\star)|Y} - \mu^\star||_{TV} \leq 2\Delta_0(\delta)\det(V)\left[\det(\widetilde{\Sigma})\right]^{1/2}(2\pi)^{-(p_0+p_1)/2}e^{\alpha_y^T\widetilde{\Sigma}^{-1}\alpha_y/2}\prod_{i=1}^{p_2+p_3}\widetilde{\beta}_i \\
+\;& 2C(\delta)\det(V^{-1})(2\pi)^{-(p_0+p_1)/2}\prod_{i=1}^{p_2+p_3}\beta_i^\star\left[\det(\Sigma^\star)\right]^{1/2}\times \\
&\times\left[1+\frac{e^{\alpha_y^T\widetilde{\Sigma}^{-1}(\widetilde{\Sigma}-\bar{\Sigma})\bar{\Sigma}^{-1}\alpha_y/2}\prod_{i=1}^{p_2+p_3}(\widetilde{\beta}_i/\beta_i^\star)\left[\det(\widetilde{\Sigma})/\det(\Sigma^\star)\right]^{1/2}}{\Gamma\left(\frac{\lambda_{\min}(\widetilde{\Sigma})(\sqrt{R_0^2+R_1^2}-||\widetilde{\Sigma}^{-1}\alpha_y||)^2}{2}\mid\frac{p_0+p_1}{2}\right)\times\prod_{i=1}^{p_2}[1-e^{-\widetilde{a}_iR_2}]\prod_{i=1}^{p_3}[1-e^{-\widetilde{b}_iR_3}]}\right] \\
+\;& 2-2\,\Gamma\left(\frac{\lambda_{\min}(\Omega_{00})(R_0-||a_0||)^2}{2}\mid\frac{p_0}{2}\right)\Gamma\left(\frac{\lambda_{\min}(B_{11})R_1^2}{2}\mid\frac{p_1}{2}\right)\prod_{i=1}^{p_2}[1-e^{-a_iR_2}]\prod_{i=1}^{p_3}[1-e^{-b_iR_3}] \\
+\;& 2[\widetilde{\mu}(B_R)]^{-1}(2\pi)^{(p_0+\tilde{p}_1)/2}[\det(\bar{\Sigma})]^{-1/2}e^{\alpha_y^T\bar{\Sigma}^{-1}\alpha_y/2}\prod\bar{\beta}_i^{-1} \\
&\times\left[\delta_{01}[||\bar{\Sigma}^{-1}\alpha_y||^2+\mathrm{trace}(\bar{\Sigma}^{-1})]+\frac{2\delta_Ap_2}{a_{\min}-\delta_A}+\frac{2\delta_b\tilde{p}_3}{b_{\min}-\delta_b}\right],
\end{aligned}
$$

and on $\mathcal{A}_3$,

$$
\begin{aligned}
||\widetilde{\Sigma}^{-1}\alpha_y|| &\leq ||\widetilde{\Sigma}^{-1}\Sigma^\star||\,||\Sigma^{\star\,-1}\alpha_y|| \leq ||a_0(\omega)|| + ||\Omega_{00}^{-1}||\rho, \\
\alpha_y^T\widetilde{\Sigma}^{-1}(\widetilde{\Sigma}-\bar{\Sigma})\bar{\Sigma}^{-1}\alpha_y &\leq \delta_{01}||\bar{\Sigma}^{-1}\Sigma^\star||\,(||a_0(\omega)|| + ||\Omega_{00}^{-1}||\rho)^2, \\
\alpha_y^T\widetilde{\Sigma}^{-1}\alpha_y &\geq \lambda_{\min}(\Sigma^\star\widetilde{\Sigma}^{-1}\Sigma^\star)(||a_0(\omega)|| - ||\Omega_{00}^{-1}||\rho)^2,
\end{aligned}
$$

where $\delta_{01} = \max(10\kappa_A\delta_0,\, 2\delta_1\kappa_B)$.

Using inequalities $1-(1-x)(1-z) \leq x+z$, $e^x - 1 \leq xe^x$, $(1+x)^m - 1 \leq mxe^{mx}$ for $x, z > 0$, we have on $\mathcal{A}_1 \cap \mathcal{A}_2 \cap \mathcal{A}_3$, we have

$$
\begin{aligned}
&||\mathbb{P}_{\mathcal{S}(x-x^\star)|Y} - \mu^\star||_{TV} \leq C_\Delta\Delta_0(\delta) + C_BC(\delta) + 2p_2e^{-a_{\min}R_2} + 2p_3e^{-b_{\min}R_3} \\
+\;& 2\left(1-\Gamma\left(\frac{\lambda_{\min}(\Omega_{00})(R_0-||a_0||)^2}{2}\mid\frac{p_0}{2}\right)\right)+2\left(1-\Gamma\left(\frac{\lambda_{\min}(B_{11})R_1^2}{2}\mid\frac{p_1}{2}\right)\right) \\
+\;& 2C_A\left[\delta_{01}[||\bar{\Sigma}^{-1}\alpha_y||^2+\mathrm{trace}(\bar{\Sigma}^{-1})]+\frac{10\delta_ap_2}{a_{\min}-5\delta_a}+\frac{2\delta_b\tilde{p}_3}{b_{\min}-\delta_b}\right],
\end{aligned}
$$

where

$$
\begin{aligned}
C_\Delta &= 2\det(V)\left[\det(\widetilde{\Sigma})\right]^{1/2}(2\pi)^{-(p_0+p_1)/2}e^{||\Sigma^\star\widetilde{\Sigma}^{-1}\Sigma^\star||(||a_0(\omega)||-||\Omega_{00}^{-1}||\rho)^2/2}\prod_{i=1}^{p_2+p_3}\widetilde{\beta}_i,\\
C_B &= 2\det(V^{-1})(2\pi)^{-(p_0+p_1)/2}\prod_{i=1}^{p_2+p_3}\beta_i^\star\left[\det(\Sigma^\star)\right]^{1/2}\times\\
&\times\left[1+\frac{e^{\delta_{01}||\bar{\Sigma}^{-1}\Sigma^\star||\,(||a_0(\omega)||+||\Omega_{00}^{-1}||\rho)^2/2}\prod_{i=1}^{p_2+p_3}(\widetilde{\beta}_i/\beta_i^\star)\left[\det(\widetilde{\Sigma})/\det(\Sigma^\star)\right]^{1/2}}{\Gamma\left(\frac{\lambda_{\min}(\widetilde{\Sigma})(\sqrt{R_0^2+R_1^2}-||a_0(\omega)||-||\Omega_{00}^{-1}||\rho)^2}{2}\mid\frac{p_0+p_1}{2}\right)\times\prod_{i=1}^{p_2}[1-e^{-\widetilde{a}_iR_2}]\prod_{i=1}^{p_3}[1-e^{-\widetilde{b}_iR_3}]}\right],\\
C_A &= [\widetilde{\mu}(B_R)]^{-1}(2\pi)^{(p_0+\tilde{p}_1)/2}[\det(\bar{\Sigma})]^{-1/2}\prod\bar{\beta}_i^{-1}\exp\{||\Sigma^\star\bar{\Sigma}^{-1}\Sigma^\star||\,(||a_0(\omega)||+||\Omega_{00}^{-1}||\rho)^2/2\}.
\end{aligned}
$$

Denoting

$$
\begin{aligned}
C_0 &= C_A\left[5\kappa_A[||\bar{\Sigma}^{-1}\Sigma^\star||^2(||a_0(\omega)||+||\Omega_{00}^{-1}||\rho)^2+\mathrm{trace}(\bar{\Sigma}^{-1})]+\frac{10p_2||B_{02}||_{2,\infty}}{a_{\min}-5\delta_a}+\frac{2\tilde{p}_3||B_{03}||_{2,\infty}}{b_{\min}-\delta_b}\right]\\
C_1 &= C_A\left[\kappa_B[||\bar{\Sigma}^{-1}\Sigma^\star||^2(||a_0(\omega)||+||\Omega_{00}^{-1}||\rho)^2+\mathrm{trace}(\bar{\Sigma}^{-1})]+\frac{10p_2||B_{12}||_{2,\infty}}{a_{\min}-5\delta_a}+\frac{2\tilde{p}_3||B_{13}||_{2,\infty}}{b_{\min}-\delta_b}\right],\\
C_2 &= C_A\left[\frac{5[||B_{22}||_{\infty,\infty}+\delta_+q_2^2\kappa_A/p]}{a_{\min}-5\delta_a}+\frac{\tilde{p}_3||B_{23}||_{\infty,\infty}/p_2}{b_{\min}-\delta_b}\right],\\
C_3 &= \frac{C_A(||B_{33}||_{\infty,\infty}+\delta_+q_3^2\kappa_B/p)}{b_{\min}-\delta_b},\\
C_4 &= \max[||V_0||_{\infty,2},||V_2||_{\infty,\infty},||V_3||_\infty]^{|S_1|},
\end{aligned}
$$

$m_5 = |S_0|+|S_3|$ if $c=0$ and $m_5=|S_0|$ if $c>0$, $C_5 = ||V_2||_{\infty,\infty}^{m_5}$, and grouping the terms for each δ_k, we have the statement of the proposition.

Assumptions $0.5\nu||V_2^T\nabla g(x^\star)||<\delta_a<0.2a_{\min}$, $\delta_b<b_{\min}$ are satisfied if

$$
\begin{gathered}
0.5\nu||V_2^T\nabla g(x^\star)||<\max_{k=0,1,2}[\delta_kc_{2,k}]<0.2a_{\min}/3,\\
\max_{k=0,1,2,3}[\delta_kc_{3,k}]<b_{\min}/4,
\end{gathered}
$$

where $c_{3,2}=||B_{23}||_{\infty,\infty}$ and

$$
\begin{aligned}
&c_{2,k}=||B_{k2}||_{2,\infty}\text{ for }k=0,1,\quad c_{2,2}=0.5[||B_{22}||_{\infty,\infty}+\delta_+q_2^2\kappa_A/p],\\
&c_{3,k}=||B_{k3}||_{2,\infty}\text{ for }k=0,1,\quad c_{33}=0.5[||B_{33}||_{\infty,\infty}+\delta_+q_3^2\kappa_B/p].
\end{aligned}
$$

□

A.4. Proof of Theorem on Bayes estimates.

PROOF OF THEOREM 2. The limit Bayes estimate $v_Q^\star$ is measurable by the Jennrich's measurability theorem since it minimises the objective function that is continuous in data and parameters.

To prove the theorem, we follow Chernozhukov and Hong (2004) and apply Theorem I.10.2 of Ibragimov and Has'minskij (1981) (p. 107), which allows one to obtain the limit distribution of the Bayes estimates provided the following conditions on the penalised likelihood ratio process $\ell_\tau(v) = \exp\{-(h_y(x^\star + \mathcal{S}^{-1}v) - h_y(x^\star))/\tau\}$ are satisfied. The results of Theorem I.10.2 of Ibragimov and Has'minskij (1981) and the auxiliary lemmas apply since the factor $\exp\{-(g(x^\star + \mathcal{S}^{-1}v) - g(x^\star))/\gamma^2\}$ is non-random and is bounded on a compact neighbourhood of 0.

1. Hölder continuity of $\ell_\tau^{1/2}(v)$ in the mean square, and the exponential bound on the expected tail of $\ell_\tau(v)$. The first condition is that for any compact $K \subset \mathcal{X}$ $\exists C_1, C_2$ that depend on K such that for any $v, v' \in K$, $||v||_\infty, ||v'||_\infty \le R$,

$$\mathbb{E}|\ell_\tau(v)^{1/2} - \ell_\tau(v')^{1/2}|^2 \le C_1(1 + R^{C_2})||v - v'||_\infty^\alpha$$

for some $\alpha \in (0, 1]$.

The second condition is that any compact $K \subset \mathcal{X}$ $\exists q_\tau(z) : [0, \infty) \to (0, \infty)$ such that for any fixed τ, $q_\tau(z)$ increases to infinity as z increases to ∞, and for any $N \in \mathbb{N}$, $\lim_{\tau\to 0,\, z\to\infty} z^N e^{-q_\tau(z)} = 0$, so that for all for all $v \in \mathcal{S}^{-1}(\mathcal{X} - x^\star)$,

$$\mathbb{E}\ell_\tau(v)^{1/2} \le e^{-q_\tau(||v||_\infty)}.$$

These conditions are checked below.

2. Finite-dimensional convergence of $\ell_\tau(v)$ to the density of $\mu^\star$ is satisfied (Theorem 1).

3. The limit Bayes problem,

$$v_Q^\star(\omega) = \arg \inf_{v \in \mathbb{R}^{p_0+p_1} \times \mathbb{R}_+^{p_2+p_3}} \int_{\mathbb{R}^{p_0+p_1} \times \mathbb{R}_+^{p_2+p_3}} Q(v - v') d\mu^\star(v', \omega),$$

is uniquely solved by a random vector $v_Q^\star$. This condition is satisfied since Q is convex with a unique minimum, and $\mu^\star$ is a proper probability measure. In fact, this weaker condition on Q can replace the convexity condition.

4. Conditions on the loss functions Q are satisfied.

Now we check the first condition, using notation defined in the proof of Theorem 1. By Lemma 5, for $v \in B_R$, on $\mathcal{A} = \mathcal{A}_1 \cap \mathcal{A}_2 \cap \mathcal{A}_3$,

$$\begin{aligned}
\ell_\tau(v)^{1/2} &= \exp\{-(h_Y(x^\star + \mathcal{S}^{-1}v) - h_Y(x^\star))/(2\tau)\} \\
&\leq \exp\{-(a_{\min} - \delta_A)||v_2||_1/2 - (b_{\min} - \delta_b)||v_3||_1/2 \\
&- \lambda_{\min}(\bar{\Omega}_{00})||v_0||_2^2/4 - \lambda_{\min}(\bar{B}_{11})||v_1||_2^2/4 \\
&+ ||v_0||_2 \left(||\Omega_{00}a_0(\omega)|| + \rho)/2 + \sqrt{\nu}||B_{10}||_{1,1}\, ||v_1||_\infty\, ||v_0||_\infty/2\right\}.
\end{aligned}$$

Then,

$$\mathbb{E}\ell_\tau(v)^{1/2} \leq \exp\{-q_\tau(||v||_\infty)\}\mathbb{P}(\mathcal{A}) + 1 - \mathbb{P}(\mathcal{A}) = \exp\{-q_\tau(||v||_\infty)\}(1 + o(1)),$$

due to $\mathbb{P}(\mathcal{A}) = \mathbb{P}(\mathcal{A}_1 \cap \mathcal{A}_2 \cap \mathcal{A}_3) \to 1$ as $\tau \to 0$, where

$$\begin{aligned}
q_\tau(z) &= [a_{\min} - \delta_A + b_{\min} - \delta_b - \sqrt{p_0}(||\Omega_{00}a_0(\omega)|| + \rho)]z/2 \\
&+ [\lambda_{\min}(\bar{\Omega}_{00}) + \lambda_{\min}(\bar{B}_{11}) - 2\sqrt{\nu}||B_{10}||_{1,1}]z^2/4
\end{aligned}$$

satisfies the required conditions for τ small enough.

The second part: by Lemma 5, using both upper and lower bounds on the log posterior, for $v, v' \in K \subset B_R$ for a compact K, on $\mathcal{A} = \mathcal{A}_1 \cap \mathcal{A}_2 \cap \mathcal{A}_3$, we have

$$\begin{aligned}
\log\left(\ell_\tau(v)/\ell_\tau(v')\right) &\leq -(a - \delta_A\mathbf{1})^T v_2 + (a + \delta_A\mathbf{1})^T v_2' - (b - \delta_b\mathbf{1})^T v_3 + (b - \delta_b\mathbf{1})^T v_3' \\
&- v_0^T\bar{\Omega}_{00}v_0/2 + v_0'^T\widetilde{\Omega}_{00}v_0'/2 - v_1^T\bar{B}_{11}v_1/2 + v_1'^T\tilde{B}_{11}v_1'/2 \\
&+ v_0^T\bar{H}_{00}x_0 - v_0'^T\tilde{H}_{00}x_0 + \sqrt{\nu}v_1^T B_{10}v_0 - \sqrt{\nu}v_1'^T B_{10}v_0' \\
&\leq ||a||_1\, ||v_2 - v_2'||_\infty + \delta_A||v_2 + v_2'||_1 + ||b||_1\, ||v_3 - v_3'||_\infty + \delta_b||v_3 + v_3'||_1 \\
&- (v_0 + v_0')^T\bar{\Omega}_{00}(v_0 - v_0')/2 + v_0'^T(\tilde{\Omega}_{00} - \bar{\Omega}_{00})v_0'/2 \\
&- (v_1 + v_1')^T\bar{B}_{11}(v_1 - v_1')/2 + v_1'^T(\tilde{B}_{11} - \bar{B}_{11})v_1'/2 \\
&+ (v_0 - v_0')^T\nabla h_y(x^\star) + \sqrt{\nu}(v_1 - v_1')^T B_{10}v_0 + \sqrt{\nu}v_1'^T B_{10}(v_0 - v_0').
\end{aligned}$$

Thus, we can write

$$\ell_\tau(v)^{1/2}/\ell_\tau(v')^{1/2} \leq \exp\left\{c_0(1 + R_0 + R_1)||v - v'||_\infty + \sum_{k=0}^{3} c_k\delta_k R_k\right\}.$$

Taking δ_k such that $\delta_k \to 0$ and $\delta_k R_k \to 0$, the last four terms in the exponent tend to 0 as $\sigma \to 0$. Therefore, using inequality $e^x - 1 \leq xe^x$ for $x \geq 0$, for $k = 0, \ldots, 3$, on $\mathcal{A}$,

$$\begin{aligned}
|\ell_\tau(v)^{1/2} - \ell_\tau(v')^{1/2}| &\leq \ell_\tau(v')^{1/2}\left(c_0(1 + R_0 + R_1)||v - v'||_\infty + o(1)\right) \\
&\leq \left(c_0(1 + R_0 + R_1)||v - v'||_\infty + o(1)\right)\exp\{-q_\tau(||v'||_\infty)\}.
\end{aligned}$$

This implies that $\mathbb{E}|\ell_\tau(v)^{1/2} - \ell_\tau(v')^{1/2}| \le C||v - v'||_\infty + o(1)$. Thus, conditions 1-4 of Theorem I.10.2 of Ibragimov and Has'minskij (1981) are satisfied and hence $\mathcal{S}(\hat{x}_Q - x^\star) \xrightarrow{d} v^\star_Q$.

□

A.5. Auxiliary results.

LEMMA 5. *Let S be a nonempty subset of $\{1:p\}$ and denote $\mathbb{R}_S = \{x \in \mathbb{R}^p : x_i \ge 0 \,\forall\, i \in S\}$.*

Then, a linear map $\mathbb{R}_S \to \mathbb{R}_S$ defined by matrix V is a bijection if and only if $V_{S,\pi(S)} = diag(a_1, \ldots, a_{|S|})$ for some $a_k > 0$ and $V_{S,S^c} = 0$ for some permutation π of S.

Note that this statement also holds for V^{-1}.

PROOF OF LEMMA 7. We need to show that $(Vx)_k \ge 0$ iff $x_k \ge 0$ for each $k \in S$. Then, we need to show that

$$V_{S,S}x_S + V_{S,S^c}x_{S^c} \in [0,\infty)^{|S|} \quad \text{iff} \quad x_S \in [0,\infty)^{|S|}$$

which holds iff $V_{S,S^c} = 0$ and $V_{S,S}$ is an invertible matrix with nonnegative entries. Denoting $U = V^{-1}$, these conditions imply that $U_{S,S} = [V_{S,S}]^{-1}$ and $U_{S,S^c} = 0$.

Now, matrix V^{-1} must satisfy the same conditions, i.e. $U_{S,S^c} = 0$ (which is satisfied) and $[V^{-1}]_{S,S} = [V_{S,S}]^{-1}$ is an invertible matrix with nonnegative entries.

Thus, we must have that both $V_{S,S}$ and $[V_{S,S}]^{-1}$ have nonnegative entries. We can prove, for instance, by induction, that this implies that both $V_{S,S}$ and $[V_{S,S}]^{-1}$ have to be diagonal matrices with positive eigenvalues, up to a permutation of the coordinates in S. For $|S| = 2$ it is a necessary and sufficient condition (base of the induction). Suppose the statement is true for $|S| = m$; then, for $|S| = m + 1$, nonnegativity of the matrices' elements implies that both matrices can be written in the block form of sizes m and 1. For each matrix, the block of size one must be a positive number, since the matrices are invertible, and the block of size m must be a diagonal matrix with positive eigenvalues, up to a permutation of the coordinates. This implies the statement of the lemma.

□

REFERENCES

Bertsekas, D. P. (2006). *Convex Analysis and Optimization.* Athena Scientific and Tsinghua University Press.

Besag, J. (1986). On the statistical analysis of dirty pictures (with discussion). *J. Roy. Statist. Soc. B*, **48**, 259–302.

Bochkina, N. (2012). Consistency of the posterior distribution in generalised linear inverse problems. Submitted.

Chernozhukov, V. and Hong, H. (2004). Likelihood estimation and inference in a class of nonregular econometric models. *Econometrica*, **72**, (5), 1445–80.

Davis, D., Chen, Z., Hwang, J.-N., Tsang, L., and Njoku, E. (1995). Solving inverse problems by Bayesian iterative inversion of a forward model with applications to parameter mapping using SMMR remote sensing data. *IEEE Transactions on Geoscience and Remote Sensing*, **33**, 1182–93.

Dupé, F.-X., Fadili, M.-J., and Starch, J.-L. (2011). Inverse problems with Poisson noise: Primal and primal-dual splitting. In *ICIP*, (ed. B. Macq and P. Schelkens), pp. 1901–4. IEEE.

Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **6**, (425), 721–41.

Ghosal, S., Ghosh, J. K., and Samanta, T. (1995). On convergence of posterior distributions. *Annals of Statistics*, **23**, 2145–52.

Ghosal, S. and Samanta, T. (1995). Asymptotic behaviour of Bayes estimates and posterior distributions in multiparameter nonregular cases. *Math. Methods Statist.*, **4**, 361–88.

Ghosh, J. K., Ghosal, S., and Samanta, T. (1994). Stability and convergence of posterior in non-regular problems. In *Statistical Decision Theory and Related Topics*, (ed. S. S. Gupta and J. O. Berger), pp. 183–99. Springer.

Green, P. J. (1990). Bayesian reconstructions from emission tomography data using a modified EM algorithm. *IEEE Transactions on Medical Imaging*, **9**, 84–93.

Hirano, K. and Porter, J. R. (2003). Asymptotic efficiency in parametric structural models with parameter-dependent support. *Econometrica*, **71**, (5), 1307–38.

Hofinger, A. and Pikkarainen, H. K. (2007). Convergence rate for the Bayesian approach to linear inverse problems. *Inverse Problems*, **23**, (6), 2469–84.

Ibragimov, I. A. and Has'minskij, R. Z. (1981). *Statistical estimation: asymptotic theory*. Springer, New York.

Kuhn, H. W. and Tucker, A. W. (1951). Nonlinear programming. In *Proceedings of 2nd Berkeley Symposium*, pp. 481–92. Berkeley: University of California Press.

Le Cam, L. (1953). On some asymptotic properties of maximum likelihood estimates and related Bayes estimates. *University of California Publications in Statistics*, **1**, 277–330.

Le Cam, L. and Yang, G. (1990). *Asymptotics in statistics: some basic concepts*. Springer, New York.

Nelder, J. A. and Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society. Series A (General)*, **135**, (3), 370–84.

Rover, C., Guidi, R. M. G., Viceré, A., and Christensen, N. (2007). Coherent Bayesian analysis of inspiral signals. *Classical and Quantum Gravity*, **24**, 607–615.

Searle, S. R. (1982). *Matrix algebra useful for statistics*. Wiley, New York.

Tarantola, A. (2006). Popper, Bayes and the inverse problem. *Nature Physics*, **2**, 492–4.

van der Vaart, A. (1998). *Asymptotic Statistics*. Cambridge University Press.

Weir, I. S. (1997). Fully Bayesian reconstructions from single-photon emission computed tomography data. *Journal of the American Statistical Association*, **92**, (437), 49–60.

E-mail: N.Bochkina@ed.ac.uk E-mail: P.J.Green@bristol.ac.uk

School of Mathematics, University of Edinburgh, Edinburgh EH9 3JZ, UK. E-mail: N.Bochkina@ed.ac.uk

School of Mathematics, University of Bristol, Bristol BS8 1TW, UK, and
School of Mathematical Sciences, University of Technology, Sydney, Australia. E-mail: P.J.Green@bristol.ac.uk