SMAI-JCM
SMAI Journal of Computational Mathematics

Hyperparameter Estimation in Bayesian MAP Estimation: Parameterizations and Consistency

Matthew M. Dunlop, Tapio Helin & Andrew M. Stuart

Volume 6 (2020), p. 69-100.

<http://smai-jcm.centre-mersenne.org/item?id=SMAI-JCM_2020__6__69_0>

Publication membre du
Centre Mersenne pour l'édition scientifique ouverte
http://www.centre-mersenne.org/

Sousmission sur https://smai-jcm.math.cnrs.fr/index.php/SMAI-JCM

SMAI Journal of Computational Mathematics
Vol. 6, 69-100 (2020)

Hyperparameter Estimation in Bayesian MAP Estimation: Parameterizations and Consistency

Matthew M. Dunlop [1]
Tapio Helin [2]
Andrew M. Stuart [3]

[1] Courant Institute of Mathematical Sciences, New York University, New York, New York, 10012, USA
E-mail address: matt.dunlop@nyu.edu
[2] School of Engineering Science, Lappeenranta–Lahti University of Technology, Lappeenranta, 53850, Finland
E-mail address: tapio.helin@lut.fi
[3] Computing & Mathematical Sciences, California Institute of Technology, Pasadena, California, 91125, USA
E-mail address: astuart@caltech.edu.

Abstract. The Bayesian formulation of inverse problems is attractive for three primary reasons: it provides a clear modelling framework; it allows for principled learning of hyperparameters; and it can provide uncertainty quantification. The posterior distribution may in principle be sampled by means of MCMC or SMC methods, but for many problems it is computationally infeasible to do so. In this situation maximum a posteriori (MAP) estimators are often sought. Whilst these are relatively cheap to compute, and have an attractive variational formulation, a key drawback is their lack of invariance under change of parameterization; it is important to study MAP estimators, however, because they provide a link with classical optimization approaches to inverse problems and the Bayesian link may be used to improve upon classical optimization approaches. The lack of invariance of MAP estimators under change of parameterization is a particularly significant issue when hierarchical priors are employed to learn hyperparameters. In this paper we study the effect of the choice of parameterization on MAP estimators when a conditionally Gaussian hierarchical prior distribution is employed. Specifically we consider the centred parameterization, the natural parameterization in which the unknown state is solved for directly, and the noncentred parameterization, which works with a whitened Gaussian as the unknown state variable, and arises naturally when considering dimension-robust MCMC algorithms; MAP estimation is well-defined in the nonparametric setting only for the noncentred parameterization. However, we show that MAP estimates based on the noncentred parameterization are not consistent as estimators of hyperparameters; conversely, we show that limits of finite-dimensional centred MAP estimators are consistent as the dimension tends to infinity. We also consider empirical Bayesian hyperparameter estimation, show consistency of these estimates, and demonstrate that they are more robust with respect to noise than centred MAP estimates. An underpinning concept throughout is that hyperparameters may only be recovered up to measure equivalence, a well-known phenomenon in the context of the Ornstein–Uhlenbeck process. The applicability of the results is demonstrated concretely with the study of hierarchical Whittle–Matérn and ARD priors.

2010 Mathematics Subject Classification. 62G05, 62C10, 62G20, 45Q05.

Keywords. Bayesian inverse problems, hierarchical Bayesian, MAP estimation, optimization, nonparametric inference, hyperparameter inference, consistency of estimators.

1. Introduction

Let X, Y be separable Hilbert spaces, and let $A : X \to Y$ be a linear map. We consider the problem of recovering a state $u \in X$ from observations $y \in Y$ given by

$$y = Au + \eta, \quad \eta \sim N(0, \Gamma) \tag{1.1}$$

where η is random noise corrupting the observations. This is an example of a linear inverse problem, with the mapping $u \mapsto Au$ being the corresponding forward problem. In the applications we consider,

X is typically an infinite-dimensional space of functions, and Y a finite-dimensional Euclidean space $\mathbb{R}^J$.

Our focus in this paper is on the Bayesian approach to this inverse problem. We view y, u, η as random variables, assume that u and η are *a priori* independent with known distributions $\mathbb{P}(\mathrm{d}u)$ and $\mathbb{P}(\mathrm{d}\eta)$, and seek the posterior distribution $\mathbb{P}(\mathrm{d}u|y)$. Bayes' theorem then states that

$$\mathbb{P}(\mathrm{d}u|y) \propto \mathbb{P}(y|u)\mathbb{P}(\mathrm{d}u).$$

In the hierarchical Bayesian approach the prior depends on hyperparameters θ which are appended to the state u to form the unknown. The prior on (u, θ) is factored as $\mathbb{P}(\mathrm{d}u, \mathrm{d}\theta) = \mathbb{P}(\mathrm{d}u|\theta)\mathbb{P}(\mathrm{d}\theta)$ and Bayes' theorem states that

$$\mathbb{P}(\mathrm{d}u, \mathrm{d}\theta|y) \propto \mathbb{P}(y|u)\mathbb{P}(\mathrm{d}u|\theta)\mathbb{P}(\mathrm{d}\theta).$$

In this paper we study conditionally Gaussian priors in which $\mathbb{P}(\mathrm{d}u|\theta)$ is a Gaussian measure for every fixed θ.

Centred methods work directly with (u, θ) as unknowns, whilst noncentred methods work with (ξ, θ) where $u = \sqrt{C(\theta)}\xi$ and ξ is, *a priori*, a Gaussian white noise; thus $C(\theta)$ is the covariance of $u|\theta$. In the context of MCMC methods the use of noncentred variables has been demonstrated to confer considerable advantages. However the key message of this paper is that, when studying maximum a posteriori (MAP) estimation, and in particular consistency of learning hyperparameters θ in the data-rich limit, centred parameterization is preferable to noncentred parameterization.

1.1. **Literature Review**

The Bayesian approach is a fundamental and underpinning framework for statistical inference [7]. In the last decade it has started to become a practical computational tool for large scale inverse problems [24], realizing an approach to ill-posed inverse problems introduced in the 1970 paper [18]. The subject has developing mathematical foundations and attendant stability and approximation theories [16, 29, 30, 45, 40]. Furthermore, the subject of Bayesian posterior consistency is being systematically developed [4, 6, 36, 38, 28, 27, 42, 21, 20]. Furthermore, the paper [26] was the first to establish consistency in the context of hyperparameter learning, as we do here, and in doing so demonstrates that Bayesian methods have comparable capabilities to frequentist methods, regarding adaptation to smoothness, whilst also quantifying uncertainty. We comment further on the relationship of our work to [26] in more detail later in the paper, once the needed framework has been established.

For some problems it is still beyond reach to perform posterior sampling via MCMC or SMC methods. For this reason maximum a posterior (MAP) estimation, which provides a point estimator of the unknown and has a variational formulation, remains an important practical computational tool [24]. Furthermore MAP estimation links Bayesian inference with optimization approaches to inversion, and allows for the possibility of new optimization methods informed by the Bayesian perspective. As a consequence there is also a developing mathematical theory around MAP estimation for Bayesian ill-posed inverse problems, relating to both how to define a MAP estimator in the infinite dimensional setting [2, 12, 15, 22, 23], and to the subject of posterior consistency of MAP estimators [5, 3, 15, 37, 39].

The focus of this paper is hierarchical Bayesian inversion with Gaussian priors such as the Whittle–Matérn and ARD priors. See [44, 34] for references to the literature in this area. In this context the question of centred versus noncentred parameterization is important in defining the problem [41]. This choice also has significant ramifications for algorithms: there are many examples of settings in which the noncentred approach is preferable to the centred approach within the context of Gibbs-based MCMC sampling [1, 17, 43, 48] and even non-Bayesian methods such as ensemble Kalman inversion [10]. Nonetheless, in the context of MAP estimation, we demonstrate in this paper that the message is rather different: centred methods are preferable.

1.2. **Our Contribution**

The primary contributions of this paper are as follows:

- We demonstrate that, for MAP estimation, centred parameterizations are preferable to noncentred parameterizations when a goal of the inference is recovery of the hyperparameters θ. We provide conditions on the data model and prior distribution that lead to theorems describing the recovery, or lack of recovery, of the true hyperparameters in the simultaneous large data/small noise limit.
- We extend the theory to empirical Bayesian estimation of hyperparameters; we also demonstrate additional robustness that this method has over the centred parameterization.
- We demonstrate the precise sense in which hyperparameter recovery holds only up to measure equivalence.

In Section 2 we introduce the Bayesian setting in which we work, emphasizing hierarchical Gaussian priors and describing the centred and noncentred formulations. In Section 3 we review the concept of MAP estimation in a general setting, describing issues associated with working directly in infinite dimensions, and discussing different choices of parameterization. Section 4 contains the theoretical results concerning consistency of hyperparemeter estimation, setting up the data-rich scenario, and studying the properties of hyperparameter estimators for the centred, noncentred and empirical Bayes settings; we show in particular the applicability of the theory to the case of hierarchical Whittle–Matérn priors and Automatic Relevance Determination (ARD) priors. In Section 5 numerical results are given which illustrate the foregoing theory. In Section 6 we conclude. Some lemmas required in the analysis are given in an appendix.

2. **Bayesian Inverse Problems**

In this section we introduce the Bayesian hierarchical approach to the solution of inverse problems of the form considered in the introduction. In Section 2.1 we describe examples of Gaussian distributions motivating the hierarchical modelling in subsequent sections. Subsection 2.2 is devoted to a brief discussion of Bayesian inversion, hierarchical priors, centred versus noncentred parameterization and sampling methods associated with the different choice of hierarchical parameterization; this sets the context for our results comparing MAP estimation with centred and noncentred parameterizations.

2.1. **Gaussian Random Process Priors**

In this paper we focus on the case where the prior is (in the hierarchical case, conditionally,) Gaussian. Recall that probability measure μ_0 on X is a Gaussian measure if[1] $\ell^\sharp\mu_0$ is a Gaussian measure on $\mathbb{R}$ for any bounded linear functional $\ell : X \to \mathbb{R}$; equivalently, μ_0 is a Gaussian measure if $u \sim \mu_0$ implies that $\ell(u)$ is a Gaussian random variable on $\mathbb{R}$ for any such ℓ. If X is a space of functions on a domain $D \subseteq \mathbb{R}^d$, a random variable u on X with law μ_0 is referred to as a Gaussian process on D.[2] Such a Gaussian process is characterized completely by its mean function $m : D \to \mathbb{R}$ and covariance function $c : D \times D \to \mathbb{R}$:

$$m(x) = \mathbb{E}^{\mu_0}(u(x)) \quad \text{for all } x \in D,$$
$$c(x, x') = \mathbb{E}^{\mu_0}(u(x) - m(x))(u(x') - m(x')) \quad \text{for all } x, x' \in D.$$

[1] Given a measure μ on X and a measurable map $T : X \to X'$, $T^\sharp\mu_0$ denotes the *pushforward measure* on X', defined by $(T^\sharp\mu)(A) = \mu(T^{-1}(A))$ for all measurable $A \subseteq X'$.

[2] Also sometimes termed a Gaussian random field.

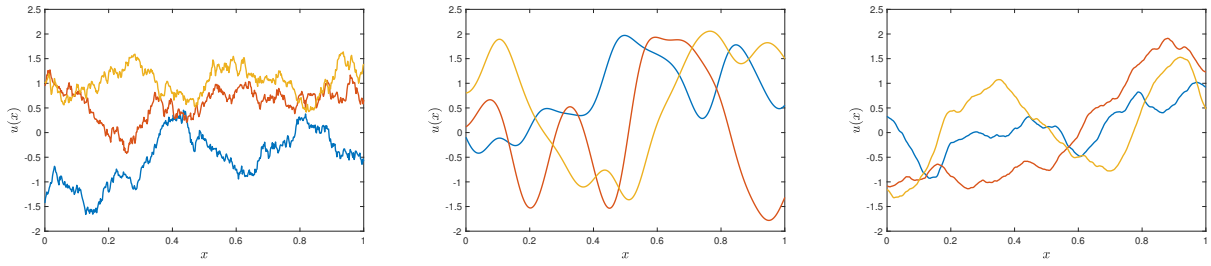

FIGURE 2.1. Three sample paths each from Gaussian processes on $[0,1]$ with (left) Ornstein–Uhlenbeck, (middle) squared exponential, and (right) Matern ($\nu = 3/2$) covariance functions.

Equivalently, it is characterized by its mean $m \in X$ and covariance operator $C : X \to X$,

$$m = \mathbb{E}^{\mu_0}(u), \quad C = \mathbb{E}^{\mu_0}(u-m) \otimes (u-m).$$

When $X = L^2(D)$, the covariance function is related to the covariance operator by

$$(C\varphi)(x) = \int_D c(x,x')\varphi(x')\,\mathrm{d}x' \quad \text{for all } \varphi \in X, x \in D,$$

that is, C is the integral operator with kernel c. In particular, if C is the inverse of a differential operator, c is the Green's function for that operator. We now detail a number of Gaussian processes that arise as examples throughout the paper.

Example 2.1 (Ornstein–Uhlenbeck)**.** Let $D = [0,1]$. Given $\sigma, \ell > 0$, define the covariance function

$$c_{OU}(t,t';\sigma,\ell) = \sigma^2 \exp\left(-\frac{|t-t'|}{\ell}\right).$$

This is the covariance associated with the stationary Ornstein–Uhlenbeck process on $[0,1]$ defined by

$$\mathrm{d}u_t = -u_t/\ell\,\mathrm{d}t + \sqrt{2\sigma^2/\ell}\,\mathrm{d}W_t, \quad u_0 \sim N(0,\sigma^2),$$

where σ^2 is the variance and ℓ the length scale. The sample paths of this process are almost surely Hölder with any exponent less than one half, everywhere in D.

Given observation of u_t over any interval $I \subseteq D$, the diffusion coefficient σ^2/ℓ may be found exactly by, for example, looking at quadratic variation. To see this, we rewrite in terms of $(\sigma,\beta) = (\sigma, \sigma^2/\ell)$, instead of treating (σ,ℓ) as the hyperparameters. With this parameterization we obtain

$$\mathrm{d}u_t = -\frac{\beta}{\sigma^2}u_t\,\mathrm{d}t + \sqrt{2\beta}\mathrm{d}W_t, \quad u_0 \sim N(0,\sigma^2).$$

By Girsanov's theorem, the law of u is equivalent to that for $\sqrt{2\beta}W_t$ for any choice of σ^2. Almost sure properties are shared between equivalent measures and for this reason it is possible to recover β from observation of u_t over any interval $I \subseteq D$, as it is from observation of $\sqrt{2\beta}W_t$ [43]. However joint recovery of β and σ^2 requires more data, such as observation of a sample path on $[0,\infty)$; see the discussion in [47].

Note also that the covariance function underlying this construction can be generalized to more general $D \subseteq \mathbb{R}^d$, using $|\cdot|$ to denote the Euclidean norm on $\mathbb{R}^d$ – it is then typically referred to as the exponential covariance function.

Example 2.2 (Squared Exponential)**.** Let $D \subseteq \mathbb{R}^d$. Given $\sigma, \ell > 0$, define the covariance function

$$c_{SE}(x,x';\sigma,\ell) = \sigma^2 \exp\left(-\frac{|x-x'|^2}{2\ell^2}\right).$$

Then the corresponding Gaussian process has samples which are almost surely infinitely differentiable everywhere; the parameters σ^2, ℓ represent variance and length-scale as for the Ornstein–Uhlenbeck covariance.

Example 2.3 (Whittle–Matérn)**.** Let $D \subseteq \mathbb{R}^d$. The Matérn (or Whittle–Matérn) covariance function provides an interpolation between the previous two examples in terms of sample regularity. The parameters $\sigma^2, \ell > 0$ have the same meaning as in the previous two examples and, additionally, we introduce the regularity parameter $\nu > 0$. Define the covariance function[3]

$$c_{WM}(x, x'; \sigma, \ell, \nu) = \sigma^2 \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{|x - x'|}{\ell} \right)^{\nu} K_\nu \left(\frac{|x - x'|}{\ell} \right)$$

where K_ν is the modified Bessel function of the second kind of order ν. Then the corresponding Gaussian process has samples which possess up to ν (fractional) weak derivatives almost surely; if the domain D is suitably regular they also possess up to ν Hölder derivatives almost surely. Note that we have

$$c_{WM}(x, x'; \sigma, \ell/\sqrt{2\nu}, \nu) \to \begin{cases} c_{OU}(x, x'; \sigma, \ell) & \text{as } \nu \to 1/2 \\ c_{SE}(x, x'; \sigma, \ell) & \text{as } \nu \to \infty. \end{cases}$$

If $D = \mathbb{R}^d$, the covariance function $c_{WM}(x, x'; \sigma, \ell, \nu)$ is the Green's function for the fractional differential operator given by

$$L(\sigma, \ell, \nu) = \frac{\Gamma(\nu)}{\sigma^2 \ell^d \Gamma(\nu + d/2)(4\pi)^{d/2}} (I - \ell^2 \Delta)^{\nu + d/2}. \tag{2.1}$$

This is the precision operator for the Gaussian measure. The corresponding covariance operator is given by $C(\sigma, \ell, \nu) = L(\sigma, \ell, \nu)^{-1}$. On more general domains $D \subseteq \mathbb{R}^d$, boundary conditions must be imposed on the Laplacian in order to ensure the invertibility of $L(\sigma, \ell, \nu)$; this generally affects the stationarity of samples, however conditions may be chosen such that stationarity of samples is (approximately) preserved [14, 25].

Finally, observe that if $-\Delta$ on a bounded domain D, subject to appropriate boundary conditions, diagonalizes with eigenbasis $\{\varphi_j\}$ and corresponding eigenvalues $\{\lambda_j\}$, then $C(\sigma, \ell, \nu)$ diagonalizes in the same basis with eigenvalues $\{\mu_j(\sigma, \ell, \nu)\}$,

$$\mu_j(\sigma, \ell, \nu) = \frac{\sigma^2 \ell^d \Gamma(\nu + d/2)(4\pi)^{d/2}}{\Gamma(\nu)} (1 + \ell^2 \lambda_j)^{-\nu - d/2}. \tag{2.2}$$

This is used later when considering consistency of point estimates.

2.2. **Hierarchical Inversion**

In this section we describe algorithmic issues arising from how we choose to parametrize the resulting Bayesian inverse problem. To facilitate this we first introduce the likelihood, and resulting posterior, arising from application of Bayes' theorem.

2.2.1. *Likelihood*

Using the model eq. (1.1), assuming $\eta \perp u$, we have $y|u \sim N(Au, \Gamma)$ and so

$$\mathbb{P}(y|u) \propto \exp(-\Phi(u; y)), \tag{2.3a}$$

$$\Phi(u; y) = \frac{1}{2} \|Au - y\|_\Gamma^2, \tag{2.3b}$$

[3]Some authors may include a factor $\sqrt{2\nu}$ before the distances $|x - x'|$. We omit it here for consistency with works such as [32, 44], which are key to the application of results in this paper.

where we have introduced the notation

$$\|z\|_{\mathsf{T}}^2 := \langle z, \mathsf{T}^{-1} z \rangle$$

for strictly positive-definite matrix or operator T on Hilbert space with inner-product $\langle \cdot, \cdot \rangle$; here we use the Euclidean inner-product on $Y = \mathbb{R}^J$. Other data models, such as those involving multiplicative or non-Gaussian noise, may lead to more complicated likelihood functions – we focus on Gaussian additive noise in this article for both clarity of presentation and analytical tractability.

2.2.2. *Natural Parametrization of the Posterior*

The posterior distribution is the law of the unknown state u given the data y, that is, the law $\mathbb{P}(u|y)$. Bayes' theorem shows how to construct the posterior in terms of the prior and likelihood. If the prior $\mu_0 = N(0, \mathsf{C}_0)$ is Gaussian, the posterior, given the likelihood described in subsection 2.2.1, is a Gaussian distribution in this conjugate setting: the posterior probability measure $\mu^y = N(m, \mathsf{C})$ satisfies

$$m = \mathsf{C}_0 A^* (\Gamma + A\mathsf{C}_0 A^*)^{-1} y, \quad \mathsf{C} = \mathsf{C}_0 - \mathsf{C}_0 A^* (\Gamma + A\mathsf{C}_0 A^*)^{-1} A\mathsf{C}_0. \tag{2.4}$$

Notice that in infinite dimensions justification of these formulae requires careful specification of the functional analytic setting [31].

In more general cases, such as when the forward map is non-linear or the prior is only conditionally Gaussian, sampling typically cannot be performed directly, and methods such as MCMC or SMC must be used instead to sample the posterior. We note here that when the prior is Gaussian, MCMC and SMC methods are available for sampling the posterior that are well-defined on function space and possess dimension-independent convergence properties [9, 13, 8].

In any setting where a Gaussian prior is consistent with prior knowledge, it is often the case that choice of a *particular* Gaussian with fixed parameters may be too restrictive in practice. For example, if a Whittle–Matérn Gaussian distribution is chosen, good prior estimates of the regularity parameter ν or length-scale ℓ may not be known, and differing choices of these parameters can lead to very different estimates under the posterior [35]. In the Bayesian paradigm we may treat these parameters as unknown random variables and place a prior distribution upon them.

We denote the hyperparameters by $\theta \in \Theta$, and assume Θ is finite-dimensional. Denoting ρ_0 the Lebesgue density of the prior on θ, we define the conditionally Gaussian prior distribution on $(u, \theta) \in X \times \Theta$ by

$$\mu_0(\mathrm{d}u, \mathrm{d}\theta) = \nu_0(\mathrm{d}u; \theta)\rho_0(\theta)\,\mathrm{d}\theta \tag{2.5}$$

where $\nu_0(\mathrm{d}u; \theta) = N(0, \mathsf{C}(\theta))$. Bayes' theorem is applied as above, and the posterior is now a measure on the product space $Z = X \times \Theta$:

$$\mu^y(\mathrm{d}u, \mathrm{d}\theta) \propto \exp(-\Phi(u; y))\,\mu_0(\mathrm{d}u, \mathrm{d}\theta). \tag{2.6}$$

As in the non-hierarchical setting, it is desirable to produce samples from the posterior in order to perform inference. The posterior is no longer Gaussian even when the forward map is linear, and so we cannot sample it directly. We can however take advantage of the conditional Gaussianity of the prior and the existence of dimension-robust MCMC sampling algorithms, as outlined in algorithm 1.

However, even though the update $u^{(k)} \mapsto u^{(k+1)}$ uses a dimension-robust algorithm, the update $\theta^{(k)} \mapsto \theta^{(k+1)}$ can be problematic even though it is only targeting a finite-dimensional distribution. The acceptance probability for a proposed update $\theta^{(k)} \mapsto \theta'$ involves the Radon–Nikodym derivative between the Gaussian distributions $\nu_0(\cdot; \theta^{(k)})$ and $\nu_0(\cdot; \theta')$. Such a derivative does not exist in general – by the Feldman–Hajèk theorem Gaussian measures in infinite dimensions are either equivalent or singular, and the restrictive conditions required for equivalence mean that in many naturally occurring situations, two Gaussian measures corresponding to different values of θ are singular. In practice this

Algorithm 1 Centred Metropolis-within-Gibbs

Choose $u^{(1)} \in X$, $\theta^{(1)} \in \Theta$.
for $k = 1 : K$ **do**
  Generate $u^{(k)} \mapsto u^{(k+1)}$ with a dimension-robust MCMC algorithm sampling $u|y, \theta^{(k)}$.
  Generate $\theta^{(k)} \mapsto \theta^{(k+1)}$ with an MCMC algorithm targeting $\theta|y, u^{(k+1)}$.
end for
return $\{(u^{(k)}, \theta^{(k)})\}_{k=1}^{K}$.

means that, for algorithm 1, any updates to θ have vanishingly small acceptance probability with respect to increasingly fine discretization of X; see [43] for a seminal analysis of this phenomenon. In the next subsubsection we discuss how this problem can be circumvented by means of a reparameterization.

2.2.3. *Reparameterization*

In the natural or *centred parameterization* [41], we treat the input u to the forward map as an unknown in the problem. However, the conditional nature of the prior on the pair (u, θ) leads to sampling problems related to measure singularity as described above. We therefore look for a way of parameterizing the prior that avoids this. We first make the observation that if $\xi \sim N(0, \mathsf{I})$, then for any fixed $\theta \in \Theta$ we have

$$\mathsf{C}(\theta)^{1/2}\xi \sim N(0, \mathsf{C}(\theta)) = \nu_0(\mathrm{d}u; \theta).$$

Therefore, if we choose $\xi \sim N(0, \mathsf{I})$ and $\theta \sim \rho_0$ *independently*, we have

$$(\mathsf{C}(\theta)^{1/2}\xi, \theta) \sim \nu_0(\mathrm{d}u; \theta)\rho_0(\mathrm{d}\theta) = \mu_0(\mathrm{d}u, \mathrm{d}\theta).$$

We can hence write a sample from μ_0 as a deterministic transform of a sample from the product measure $N(0, \mathsf{I}) \times \rho_0$ – this reparameterization is referred to as *noncentering* in the literature [41]. It has the advantage that we may pass it to the posterior distribution by sampling an appropriate surrogate distribution instead of directly targeting the posterior.

We now make the preceding statement precise. Let $\bar{X}$ be a space of distributions that white noise samples $\xi \sim N(0, \mathsf{I})$ belong to almost surely, and define the product spaces $Z = X \times \Theta$, $\bar{Z} = \bar{X} \times \Theta$. Define the mapping $T : \bar{Z} \to Z$ by $T(\xi, \theta) = (\mathsf{C}(\theta)^{1/2}\xi, \theta)$. Then we have the following.

Proposition 2.4 (Noncentering)**.** *Let μ^y denote the hierarchical posterior eq.* (2.6) *on Z with prior μ_0. Define the measures $\bar{\mu}_0, \bar{\mu}^y$ on $\bar{Z}$ by $\bar{\mu}_0 = N(0, \mathsf{I}) \times \rho_0$ and*[4]

$$\bar{\mu}^y(\mathrm{d}\xi, \mathrm{d}\theta) \propto \exp\big(-\Phi(T(\xi, \theta))\big)\, \bar{\mu}_0(\mathrm{d}\xi, \mathrm{d}\theta).$$

Then $\mu_0 = T^\sharp \bar{\mu}_0$ and $\mu^y = T^\sharp \bar{\mu}^y$.

Proof. The first equality follows from the preceding discussion, and the second from a standard property of pushforward measures:

$$\int f(x)\, (T^\sharp \mu)(\mathrm{d}x) = \int f(T(y))\, \mu(\mathrm{d}y).$$

■

The key consequence of this proposition is that if we sample $(\xi, \theta) \sim \bar{\mu}^y$, we have $T(\xi, \theta) \sim \mu^y$. We therefore use MCMC to target $\bar{\mu}^y$ instead of μ^y – since the field ξ and hyperparameter θ are independent under the prior, the previous measure singularity issues disappear. This leads us to algorithm 2.

Making the choice of noncentred variables over centred variables leads, in the context of Gibbs-based MCMC, to significant improvement in algorithmic performance, as detailed in a number of

[4]Here we have implicitly extended $\Phi : X \to \mathbb{R}$ to $\Phi : Z \to \mathbb{R}$ via projection: $\Phi(u, \theta) \equiv \Phi(u)$.

Algorithm 2 Noncentred Metropolis-within-Gibbs

Choose $\xi^{(1)} \in \bar{X}$, $\theta^{(1)} \in \Theta$.
for $k = 1 : K$ **do**
  Generate $\xi^{(k)} \mapsto \xi^{(k+1)}$ with a dimension-robust MCMC algorithm targeting $\xi|y, \theta^{(k)}$.
  Generate $\theta^{(k)} \mapsto \theta^{(k+1)}$ with an MCMC algorithm targeting $\theta|y, \xi^{(k+1)}$.
end for
return $\{T(\xi^{(k)}, \theta^{(k)})\}_{k=1}^{K}$.

papers [43, 41, 48, 1, 11]. However, as we demonstrate in the remainder of this paper, for MAP estimation different considerations come in to play, and centred methods are preferable.

3. Point Estimation

Sampling of the posterior distribution, for example using MCMC methods as mentioned in the previous section, or SMC methods as in [8], may be prohibitively expensive computationally if a large number of samples are required. It is then desirable to find a point estimate for the solution to the problem, as opposed to the full posterior distribution. The conditional mean is one such point estimate, but this typically requires samples in order to be computed. Two alternative point estimates that we study in this paper, and define in this section, are the MAP estimate and the empirical Bayes (EB) estimate, both of which can be computed through optimization procedures. The former can be interpreted as the mode of the posterior distribution, and the latter as a compromise between the mean and the mode. In Section 3.1 we introduce the basic MAP estimator and discuss its properties under change of variables. In Section 3.2 we generalize to centred and noncentred hierarchical formulations; mapping from one formulation to the other may be viewed as a hyperparameter dependent change of variables. In Section 3.3 we define the empirical Bayes estimator.

3.1. MAP Estimation and its dependence on parameterization

Suppose first that $X = \mathbb{R}^n$ and the posterior admits a Lebesgue density:

$$\mu^y(\mathrm{d}u) \propto \pi^y(u)\,\mathrm{d}u$$

A MAP estimate, or mode of the posterior distribution, is then any point $u \in X$ that maximizes π^y. Equivalently, it is any point that minimizes $-\log \pi^y$, which is usually more stable to deal with numerically. The existence of a Lebesgue density is central to this definition of MAP estimate. We, however, are primarily interested in the case that X is infinite-dimensional and a more general definition is therefore required. Dashti et al. [15] introduced such a generalization as follows.

Definition 3.1. Let μ be a Borel probability measure on a Banach space $\mathcal{X}$, and denote by $B_\delta(u)$ the ball of radius δ centred at $u \in \mathcal{X}$. A point $u_* \in \mathcal{X}$ is said to be a MAP estimator for the measure μ if

$$\lim_{\delta \to 0} \left(\frac{\mu(B_\delta(u_*))}{\max_{u \in \mathcal{X}} \mu(B_\delta(u))} \right) = 1.$$

More general definitions have subsequently been introduced [22, 12], but for the measures considered in this article they are equivalent to the definition above. If a Gaussian prior $\nu_0 = N(0, \mathsf{C}_0)$ is chosen and the data model eq. (1.1), eq. (2.3) is used, then it is known [15] that a point u is a MAP estimator if and only if it minimizes the Onsager-Machlup functional given by

$$\mathsf{I}(u) = \Phi(u; y) + \frac{1}{2}\|u\|_{\mathsf{C}_0}^2. \tag{3.1}$$

The quadratic penalty term is the Cameron-Martin norm associated to the Gaussian measure on Hilbert space X. Note that, as distinct from the finite-dimensional case, the quadratic term in $\mathsf{I}(u)$ is infinite at almost every point of the space X: $\mu_0(\{u \in X \mid \|u\|_{\mathsf{C}_0}^2 < \infty\}) = 0$. Although we have framed this discussion for the linear inverse problem eq. (1.1) subject to additive Gaussian noise, it applies to the nonlinear setting, with Gaussian priors, and Φ is simply the negative log-likelihood; however for this paper we consider only linear inverse problems with additive Gaussian noise and Φ is given by eq. (2.3b).

Let us now point out that MAP estimation makes a deep connection to classical applied mathematics approaches to inversion via optimization and for this reason it has an important place in the theory of Bayesian inversion. However an often-cited criticism of MAP estimation within the statistics community is that the methodology depends on the choice of parameterization of the model. To see this, assume again that $X = \mathbb{R}^n$ and that the posterior admits a Lebesgue density $\pi^y(u)$, so that the MAP estimator maximizes π^y. Suppose that we have a (smooth) bijective map $T : X \to X$, and instead write the unknown as $u = T(\xi)$ for some new coordinates ξ. Then the posterior in the coordinates ξ is given by

$$\bar{\pi}^y(\xi) = \pi^y(T(\xi)) \times |\det(\nabla T(\xi))|,$$

that is, for any bounded measurable $f : X \to \mathbb{R}$ we have

$$\int_X f(u)\pi^y(u)\,\mathrm{d}u = \int_X f(T(\xi))\bar{\pi}^y(\xi)\,\mathrm{d}\xi.$$

Due to the presence of this Jacobian determinant, the MAP estimators using the two coordinates generally differ. If there was no determinant term, we would have equivalence of the MAP estimators in the following sense, which is straightforward to verify.

Proposition 3.2. *Assume* $|\det(\nabla T(\xi))| \equiv 1$. *It holds that* $\xi_* \in \arg\max \bar{\pi}^y(\cdot)$ *if and only if* $T(\xi_*) \in \arg\max \pi^y(\cdot)$.

It is natural to study how this issue of reparameterization affects MAP estimators for hierarchical problems. In the previous section we chose a reparameterization motivated by the need to enable robust sampling of the posterior distribution. We show, however, that this reparameterization has undesirable effects on MAP estimation for hyperparameters.

3.2. **Hierarchical MAP Estimation**

In this subsection we extend the definition of a MAP estimator to the centred and noncentred hierarchical parameterization introduced in the previous section.

3.2.1. *Centred Hierarchical MAP Estimation*

We are interested in the case where μ_0 on $Z = X \times \Theta$ is given by eq. (2.5). The dependence of the covariance operator on the hyperparameter θ means that we cannot directly apply the above result for Gaussian measures to write down the Onsager-Machlup functional, as the normalization factor for the measure $\nu_0(\mathrm{d}u;\theta)$ depends on θ. If $X = \mathbb{R}^n$ is finite dimensional, we may write down the Onsager-Machlup functional as

$$\mathsf{I}_{\mathsf{C}}(u,\theta) = \Phi(u;y) + \frac{1}{2}\|u\|_{\mathsf{C}(\theta)}^2 + \frac{1}{2}\log\det\mathsf{C}(\theta) - \log\rho_0(\theta).$$

Now consider the case where $n \to \infty$ and $\mathbb{R}^n$ represents approximation of an infinite dimensional space X. Since the limiting operator $\mathsf{C}(\theta)$ is symmetric and compact, then the determinant of finite dimensional approximations tends to zero as $n \to \infty$. Additionally, the set of points for which the quadratic term is finite may depend on the hyperparameter θ – in particular such sets for different

values of θ may intersect only at the origin. As an example of this latter phenomenon, consider the Whittle–Matérn process with precision operator L given by eq. (2.1). The quadratic penalty term is $\langle u, Lu\rangle_X$ and, for different values of ν these correspond to different Sobolev space penalizations. In summary, both the definition and optimization of the functional I_C may be problematic in infinite dimensions; we show in what follows that this is also true for sequences of finite dimensional problems which approach the infinite dimensional limit.

Assuming now $X = \mathbb{R}^n$, if we fix $\theta \in \Theta$, then we can optimize $\mathsf{I}_\mathsf{C}(\cdot, \theta)$ to find $u(\theta) \in X$ such that

$$\mathsf{J}_\mathsf{C}(\theta) := \mathsf{I}_\mathsf{C}(u(\theta), \theta) \leq \mathsf{I}_\mathsf{C}(u, \theta) \quad \text{for all } u \in X.$$

In the linear setting eq. (1.1) that is our focus, using eq. (2.4), we have

$$u(\theta) = \mathsf{C}(\theta)A^*(\Gamma + A\mathsf{C}(\theta)A^*)^{-1}y. \tag{3.2}$$

We may then optimize $\mathsf{J}_\mathsf{C}(\cdot)$ to find $\theta_* \in \Theta$ such that

$$\mathsf{J}_\mathsf{C}(\theta_*) = \mathsf{I}_\mathsf{C}(u(\theta_*), \theta_*) \leq \mathsf{I}_\mathsf{C}(u(\theta), \theta) \leq \mathsf{I}_\mathsf{C}(u, \theta) \quad \text{for all } u \in X, \theta \in \Theta.$$

The task of optimizing I_C is hence reduced to that of optimizing J_C. In the next section we study the behaviour of minimizers of J_C as the quality of the data increases.

3.2.2. *Noncentred Hierarchical MAP Estimation*

If we work with the noncentred coordinates introduced in the previous section, the joint prior measure is the independent product of a Gaussian measure on $\bar{X}$ and with the hyperprior on Θ. MAP estimators can hence be seen to be well-defined on the infinite-dimensional space $\bar{Z} = \bar{X} \times \Theta$, and to be equivalent to minimizers of the Onsager-Machlup functional

$$\mathsf{I}_\mathsf{NC}(\xi, \theta) = \Phi(\mathsf{C}(\theta)^{1/2}\xi; y) + \frac{1}{2}\|\xi\|_I^2 - \log \rho_0(\theta).$$

Note that if we reverse the transformation and write $(\xi, \theta) = T^{-1}(u, \theta) = (\mathsf{C}(\theta)^{-1/2}u, \theta)$, we could equivalently define I_NC on $Z = X \times \Theta$ by

$$\mathsf{I}_\mathsf{NC}(u, \theta) = \Phi(u; y) + \frac{1}{2}\|u\|_{\mathsf{C}(\theta)}^2 - \log \rho_0(\theta),$$

in view of Proposition 3.2. This is I_C, with the problematic log-determinant term subtracted.

As in the centred case, we can now fix θ and optimize $\mathsf{I}_\mathsf{NC}(\cdot, \theta)$ over to find $\xi(\theta) \in \bar{X}$ such that

$$\mathsf{J}_\mathsf{NC}(\theta) := \mathsf{I}_\mathsf{NC}(\xi(\theta), \theta) \leq \mathsf{I}_\mathsf{NC}(\xi, \theta) \quad \text{for all } \xi \in \bar{X}.$$

Again, in the linear setting eq. (1.1), using eq. (2.4), we have that $\xi(\theta)$ is given by

$$\xi(\theta) = \mathsf{C}(\theta)^{1/2}A^*(\Gamma + A\mathsf{C}(\theta)A^*)^{-1}y.$$

Note that $u(\theta) = \mathsf{C}(\theta)^{1/2}\xi(\theta)$, which is consistent with Proposition 3.2. However, note that $\mathsf{J}_\mathsf{C} \neq \mathsf{J}_\mathsf{NC}$: only the former has the log-determinant term, and so the MAP estimate for the hyperparameters typically differs between the two parameterizations.

Remark 3.3. To understand that the difference between J_C and J_NC is related to the volume term arising from change of parameterization, consider the case $X = \mathbb{R}^n$. We start with the measure

$$\begin{aligned}\mu(\mathrm{d}u, \mathrm{d}\theta) &\propto \exp(-\mathsf{I}_\mathsf{C}(u, \theta))\mathrm{d}u\,\mathrm{d}\theta \\ &= \exp\left(-\Phi(u; y) - \frac{1}{2}\|u\|_{\mathsf{C}(\theta)}^2 - \frac{1}{2}\log\det \mathsf{C}(\theta) + \log \rho_0(\theta)\right)\mathrm{d}u\,\mathrm{d}\theta \\ &=: f(u, \theta)\,\mathrm{d}u\,\mathrm{d}\theta.\end{aligned}$$

We make the transformation $(u, \theta) = T(\xi, \theta) = (\mathsf{C}(\theta)^{1/2}\xi, \theta)$. The density in these new coordinates is now given by

$$h(\xi, \theta) = f(T(\xi, \theta)) \times |\det(\nabla T(\xi, \theta))|.$$

The Jacobian determinant may be calculated as

$$\det(\nabla T(\xi, \theta)) = \det(\nabla_\xi T_1(\xi, \theta)) \det(\nabla_\theta T_2(\xi, \theta)) = \det(\mathsf{C}(\theta)^{1/2}) \det(\mathsf{I}) = \det(\mathsf{C}(\theta))^{1/2}.$$

The log determinant terms hence cancel, giving

$$h(\xi, \theta) \propto \exp\left(-\Phi(\mathsf{C}(\theta)^{1/2}\xi; y) - \frac{1}{2}\|\xi\|_I^2 + \log \rho_0(\theta)\right) = \exp(-\mathsf{I}_{\mathsf{NC}}(\xi, \theta)).$$

3.3. **Empirical Bayesian Estimation**

Instead of jointly optimizing over the state u and hyperparameters θ, we may integrate out the state to obtain a measure just on θ. In this case, one considers finding the mode of the marginal measure

$$\mathbb{P}(\mathrm{d}\theta|y) = \int_X \mu^y(\mathrm{d}u, \mathrm{d}\theta) = \left(\frac{1}{\mathbb{P}(y)} \int_X \exp(-\Phi(u; y))\, \nu_0(\mathrm{d}u; \theta)\right) \rho_0(\theta)\, \mathrm{d}\theta.$$

The corresponding functional we wish to optimize to find θ is hence given by

$$\mathsf{J}_{\mathsf{E}}(\theta) = -\log\left(\int_X \exp(-\Phi(u; y))\, \nu_0(\mathrm{d}u; \theta)\right) - \log \rho_0(\theta). \tag{3.3}$$

In general the above functional cannot be written down more explicitly due to the intractability of the integral. When $X = \mathbb{R}^n$ is finite-dimensional, the integral may be approximated using a Monte Carlo average over samples $\{u_j\}_{j=1}^M \sim \exp(-\Phi(u; y))\, \nu_0(\mathrm{d}u; \theta')$ for any fixed $\theta' \in \Theta$:

$$\begin{aligned}
\mathsf{J}_{\mathsf{E}}(\theta) &= -\log\left(\int_X \frac{\nu_0(u; \theta)}{\nu_0(u; \theta')} \exp(-\Phi(u; y))\, \nu_0(\mathrm{d}u; \theta')\right) - \log \rho_0(\theta) \\
&\approx -\log \sum_{j=1}^M \exp\left(\frac{1}{2}\|u_j\|_{\mathsf{C}(\theta')}^2 - \frac{1}{2}\|u_j\|_{\mathsf{C}(\theta)}^2 + \frac{1}{2}\log\det \mathsf{C}(\theta')\mathsf{C}(\theta)^{-1}\right) - \log \rho_0(\theta) \\
&=: \mathsf{J}_{\mathsf{E}}(\theta; \theta', \{u_j\}),
\end{aligned}$$

where the log-sum-exp trick may be used numerically to avoid underflow [33, §3.5.3]. One may then aim to approximately optimize J_{E} via algorithm 3, which alternates approximating the integral above via samples from the conditional posterior given the current hyperparameter values, and optimizing over the hyperparameters given these samples, a form of expectation-maximization (EM) algorithm. The sampling in each step is typically performed using a dimension-robust MCMC algorithm, such as the pCN algorithm; the resulting random sequence $\{\theta^{(k)}\}$ can then be averaged, for example, to produce a single hyperparameter estimate.

Algorithm 3 EM Algorithm

Choose initial estimate $\theta^{(1)}$ for the hyperparameter.
for $k = 1 : K$ **do**
  Sample $\{u_j^{(k)}\}_{j=1}^M \sim \exp(-\Phi(u; y))\, \nu_0(\mathrm{d}u; \theta^{(k)})$.
  $\theta^{(k+1)} \leftarrow \operatorname*{argmin}_{\theta \in \Theta} \mathsf{J}_{\mathsf{E}}(\theta; \theta^{(k)}, \{u_j^{(k)}\})$.
end for

In the linear setting eq. (1.1), the integral in eq. (3.3) can be computed analytically using Gaussian structure. Rather than calculate the integral above directly, we note that we may rewrite the data in noncentred coordinates as

$$y = A\mathsf{C}(\theta)^{1/2}\xi + \eta, \quad \eta \sim N(0, \Gamma)$$

where $\xi \sim N(0, \mathsf{I})$; from this it can be seen that $\mathbb{P}(y|\theta) = N(0, \Gamma + A\mathsf{C}(\theta)A^*)$. Thus, by Bayes' theorem,

$$\mathbb{P}(\mathrm{d}\theta|y) \propto \frac{1}{\sqrt{\det(\Gamma + A\mathsf{C}(\theta)A^*)}} \exp\left(-\frac{1}{2}\|y\|^2_{\Gamma+A\mathsf{C}(\theta)A^*}\right)\rho_0(\theta)\,\mathrm{d}\theta.$$

Modes of this marginal measure are then given by minimizers of the functional

$$\mathsf{J}_\mathsf{E}(\theta) = \frac{1}{2}\|y\|^2_{\Gamma+A\mathsf{C}(\theta)A^*} + \frac{1}{2}\log\det(\Gamma + A\mathsf{C}(\theta)A^*) - \log\rho_0(\theta).$$

Despite involving norms and determinants on the data space rather than the state space, the form of J_E is actually very similar to that of J_C, as is shown in the following section.

Remark 3.4. In the spirit of this paper, we later consider the mode of $\mathbb{P}(\mathrm{d}\theta|y)$ as the empirical estimator for θ. Such a choice can also be considered as a regularized maximum likelihood estimator, where the hyperparameter density acts as a regularizer.

4. Consistency of Point Estimators

In the previous section we derived three different functionals, $\mathsf{J}_\mathsf{C}, \mathsf{J}_\mathsf{NC}$ and J_E. Optimizing each of these functionals leads to different estimates of the hyperparameters of the same underlying statistical model. In this section we study the behaviour of these estimates in a data-rich scenario. In Section 4.1 we spell out the precise data model that we use; it corresponds to a finite dimension N truncation of the linear inverse problem eq. (1.1), and since subsequent limit theorems focus on the situation in which the observational noise standard deviation γ is small, we write the resulting functionals to be optimized as $\mathsf{J}^{N,\gamma}_\mathsf{C}, \mathsf{J}^{N,\gamma}_\mathsf{NC}$ and $\mathsf{J}^{N,\gamma}_\mathsf{E}$. Proposition 4.3 gives the exact form for the resulting functionals and demonstrates the similar form taken by $\mathsf{J}^{N,\gamma}_\mathsf{C}$ and $\mathsf{J}^{N,\gamma}_\mathsf{E}$, whilst also showing that $\mathsf{J}^{N,\gamma}_\mathsf{NC}$ is substantially different. Subsection 4.2 contains the limit theorems which characterize the three different estimators in the data-rich limit. Theorem 4.6 shows that the centred and empirical Bayes approaches recover the true parameter value whilst the noncentred approach does not. In Section 4.3 we discuss examples.

4.1. The Data Model

In order to analyse the behaviour of these minimizers, we work in the simplified setup where the forward map A is linear, and A^*A is simultaneously diagonalizable with the family of covariance operators. Specifically, we make the following assumptions.

Assumptions 1. *We assume in what follows that:*

(i) The map A^*A and family of prior covariance operators $\{\mathsf{C}(\theta)\}_{\theta\in\Theta}$ are strictly positive and simultaneously diagonalizable with orthonormal eigenbasis $\{\varphi_j\}$, and we have

$$A^*A\varphi_j = a_j^2\varphi_j, \quad C(\theta)\varphi_j = \mu_j(\theta)\varphi_j \quad \textit{for all } j \in \mathbb{N}, \theta \in \Theta.$$

(ii) The noise covariance $\Gamma = \gamma^2\mathsf{I}$ is white.

Remark 4.1. The second assumption is essentially equivalent to assuming that the noise covariance Γ is non-degenerate: we may work with the transformed data $\Gamma^{-1/2}y$ and redefine A as $\Gamma^{-1/2}A$. We could hence equivalently replace A^*A with $A^*\Gamma^{-1}A$ in the first assumption.

We choose the basis $\{\psi_j\}$ for Y given by $\psi_j = A\varphi_j/\|A\varphi_j\| = A\varphi_j/a_j$; it can readily be checked that this is an orthonormal basis. Assume that the true state $u^\dagger$ that generates the data is drawn from the distribution $N(0, \mathsf{C}(\theta^\dagger))$ for some $\theta^\dagger \in \Theta$. We define the data $y^\gamma \in Y$ by

$$y^\gamma = Au^\dagger + \gamma\eta, \quad \eta \sim N(0, \mathsf{I}),$$

where we have made the dependence of the data on γ explicit. We define individual observations $y_j^\gamma \in \mathbb{R}$ of the data $y^\gamma \in Y$ as

$$\begin{aligned} y_j^\gamma &:= \langle y^\gamma, \psi_j\rangle \\ &= \frac{1}{a_j}\langle Au^\dagger, A\varphi_j\rangle + \gamma\langle \eta, \psi_j\rangle \\ &= \frac{1}{a_j}\langle u^\dagger, A^*A\varphi_j\rangle + \gamma\langle \eta, \psi_j\rangle \\ &= a_j u_j^\dagger + \gamma\eta_j, \quad \eta_j \overset{\text{iid}}{\sim} N(0,1), \quad j \in \mathbb{N}, \end{aligned} \tag{4.1}$$

where $u_j^\dagger := \langle u^\dagger, \varphi_j\rangle$. It is convenient to note that we have the equality in distribution with the noncentred-type representation

$$y_j^\gamma \overset{\text{d}}{=} \sqrt{a_j^2\mu_j(\theta^\dagger) + \gamma^2}\xi_j^\dagger, \quad \xi_j^\dagger \overset{\text{iid}}{\sim} N(0,1), \quad j \in \mathbb{N}. \tag{4.2}$$

As we establish results regarding convergence of minimizers in probability, there is no loss in generality in assuming that the data is given by eq. (4.2) instead of eq. (4.1).

The infinite collection of scalar problems eq. (4.1) is equivalent to the full infinite-dimensional problem. We consider a sequence of finite-dimensional problems arising from taking the first N of these observations, so that data provided for the N^{th} problem is given by

$$y_j^\gamma = a_j u_j^\dagger + \gamma\eta_j, \quad \eta_j \overset{\text{iid}}{\sim} N(0,1), \quad j = 1, \dots, N. \tag{4.3}$$

We take the prior distribution for these problems to be the projection of the full prior onto the span of the first N eigenfunctions $\{\varphi_j\}_{j=1}^N$, so that both the state and the data are finite-dimensional. To motivate why we use this projection of the prior distribution, we look at the structure of the likelihood. Writing $y_{1:N}^\gamma$ for the vector of observations $(y_1^\gamma, \dots, y_N^\gamma) \in \mathbb{R}^N$, the negative log-likelihood of $y_{1:N}^\gamma$ given u takes the form

$$\begin{aligned} \Phi_\gamma(u; y_{1:N}^\gamma) &= \frac{1}{2\gamma^2}\sum_{j=1}^N |\langle Au - y^\gamma, \psi_j\rangle|^2 \\ &= \frac{1}{2\gamma^2}\sum_{j=1}^N \left|\frac{1}{a_j}\langle A^*Au, \varphi_j\rangle - \langle y^\gamma, \psi_j\rangle\right|^2 \\ &= \frac{1}{2\gamma^2}\sum_{j=1}^N |a_j u_j - y_j^\gamma|^2 \end{aligned}$$

where $u_j := \langle u, \varphi_j\rangle$. The posterior on u_j for $j > N$ is hence uninformed by the observations and remains the same as the prior. To be more explicit, for the N^{th} problem we choose the conditional prior distribution $\nu_0^N(\cdot;\theta) = P_N^\sharp\nu_0(\cdot;\theta)$, where $P_N : X \to \mathbb{R}^N$ is given by $(P_N u)_j = u_j$ for $j = 1, \dots, N$. Since $\nu_0(\cdot;\theta) = N(0, \mathsf{C}(\theta))$ is Gaussian on X, this is equivalent to saying $\nu_0^N(\cdot;\theta) = N(0, P_N\mathsf{C}(\theta)P_N^*)$ is Gaussian on $\mathbb{R}^N$.

We denote by $\mathsf{J}_\mathsf{C}^{N,\gamma}$, $\mathsf{J}_\mathsf{NC}^{N,\gamma}$ and $\mathsf{J}_\mathsf{E}^{N,\gamma}$ the functionals J_C, J_NC and J_E respectively constructed for these finite dimensional problems. We study the convergence of estimates of the hyperparameter θ to its

true value $\theta^\dagger$ in the simultaneous limit of the number of observations $y_1^\gamma, \dots, y_N^\gamma$ going to infinity and the noise level γ going to zero.

Remark 4.2. The above truncation has no effect on the forms of the functionals $\mathsf{J}_{\mathsf{NC}}^{N,\gamma}$ and $\mathsf{J}_{\mathsf{E}}^{N,\gamma}$; $\mathsf{J}_{\mathsf{C}}^{N,\gamma}$ however does change. Nonetheless, if the non-truncated prior is used to write down $\mathsf{J}_{\mathsf{C}}^{N,\gamma}$, poor estimates for hyperparameters are obtained as the prior then dominates over the observations, see Section 5.2 for an illustration.

For brevity, in what follows we use the notation $f(\theta) \propto g(\theta)$ to mean that $f(\theta) = \alpha g(\theta) + \beta$ for some constants α, β – note that f and g then have the same minimizers.

Proposition 4.3. *Define $s_j^\gamma(\theta) = a_j^2 \mu_j(\theta) + \gamma^2$. Then we have*

$$\mathsf{J}_{\mathsf{C}}^{N,\gamma}(\theta) \propto \frac{1}{2N} \sum_{j=1}^{N} \left[\frac{(y_j^\gamma)^2}{s_j^\gamma(\theta)} - \log \frac{\mu_j(\theta^\dagger)}{\mu_j(\theta)} \right] - \frac{1}{N} \log \rho_0(\theta), \tag{4.4}$$

$$\mathsf{J}_{\mathsf{NC}}^{N,\gamma}(\theta) \propto \frac{1}{2N} \sum_{j=1}^{N} \frac{(y_j^\gamma)^2}{s_j^\gamma(\theta)} - \frac{1}{N} \log \rho_0(\theta), \tag{4.5}$$

$$\mathsf{J}_{\mathsf{E}}^{N,\gamma}(\theta) \propto \frac{1}{2N} \sum_{j=1}^{N} \left[\frac{(y_j^\gamma)^2}{s_j^\gamma(\theta)} - \log \frac{s_j^\gamma(\theta^\dagger)}{s_j^\gamma(\theta)} \right] - \frac{1}{N} \log \rho_0(\theta). \tag{4.6}$$

Remark 4.4. We have made the shifts

$$\mathsf{J}_{\mathsf{C}}^{N,\gamma}(\theta) \mapsto \mathsf{J}_{\mathsf{C}}^{N,\gamma}(\theta) - \frac{1}{2} \sum_{j=1}^{N} \log \mu_j(\theta^\dagger), \quad \mathsf{J}_{\mathsf{E}}^{N,\gamma}(\theta) \mapsto \mathsf{J}_{\mathsf{E}}^{N,\gamma}(\theta) - \frac{1}{2} \sum_{j=1}^{N} \log s_j^\gamma(\theta^\dagger).$$

These do not affect minimizers, as the shifts are constant in θ. These transformations are useful in the next section in the derivation of a limiting functional as $N \to \infty$ and $\gamma \to 0$.

Proof. Here, let us write $\mathsf{C}_N(\theta) = P_N \mathsf{C}(\theta) P_N^* \in \mathbb{R}^{N \times N}$ for the projected prior covariance. Instead of the expression for $u(\theta)$ given by eq. (3.2), we use the alternative expression

$$u(\theta) = (A^* \Gamma^{-1} A + \mathsf{C}_N(\theta)^{-1})^{-1} A^* \Gamma^{-1} y = \frac{1}{\gamma^2} \left(\frac{1}{\gamma^2} A^* A + \mathsf{C}_N(\theta)^{-1} \right)^{-1} A^* y$$

which follows from the Sherman–Morrison–Woodbury formula. Using the simultaneous diagonalizability, we then have that

$$u_j(\theta) := \langle u(\theta), \varphi_j \rangle = \frac{1}{\gamma^2} \left(\frac{a_j^2}{\gamma^2} + \frac{1}{\mu_j(\theta)} \right)^{-1} \langle A^* y, \varphi_j \rangle = \frac{a_j \mu_j(\theta)}{s_j^\gamma(\theta)} y_j^\gamma.$$

Now consider the functional

$$\mathsf{J}_0^{N,\gamma}(\theta) := \Phi_\gamma(u; \theta) + \frac{1}{2} \|u\|^2_{\mathsf{C}_N(\theta)}.$$

We may calculate

$$\begin{aligned} \mathsf{J}_0^{N,\gamma}(\theta) &= \frac{1}{2\gamma^2} \sum_{j=1}^{N} (a_j u_j(\theta) - y_j^\gamma)^2 + \frac{1}{2} \sum_{j=1}^{N} \frac{u_j(\theta)^2}{\mu_j(\theta)} \\ &= \frac{1}{2} \sum_{j=1}^{N} (y_j^\gamma)^2 \left[\frac{1}{\gamma^2} \left(\frac{a_j^2 \mu_j(\theta)}{s_j^\gamma(\theta)} - 1 \right)^2 + \frac{a_j^2 \mu_j(\theta)}{s_j^\gamma(\theta)^2} \right] \end{aligned}$$

$$= \frac{1}{2}\sum_{j=1}^{N} \frac{(y_j^\gamma)^2}{s_j^\gamma(\theta)^2}\left[\frac{1}{\gamma^2}\left(a_j^2\mu_j(\theta) - s_j^\gamma(\theta)\right)^2 + a_j^2\mu_j(\theta)\right]$$
$$= \frac{1}{2}\sum_{j=1}^{N} \frac{(y_j^\gamma)^2}{s_j^\gamma(\theta)}.$$

The expression for $\mathsf{J}_{\mathsf{NC}}^{N,\gamma}$ then follows. For $\mathsf{J}_{\mathsf{C}}^{N,\gamma}$, we note that

$$\frac{1}{2}\log\det \mathsf{C}_N(\theta) = \frac{1}{2}\sum_{j=1}^{N}\log\mu_j(\theta) \propto -\frac{1}{2}\sum_{j=1}^{N}\log\frac{\mu_j(\theta^\dagger)}{\mu_j(\theta)}$$

from which the result follows. Finally we deal with the empirical Bayes case $\mathsf{J}_{\mathsf{E}}^{N,\gamma}$. Observe that

$$\frac{1}{2}\|y^\gamma\|^2_{\Gamma+A\mathsf{C}_N(\theta)A^*} = \frac{1}{2}\sum_{i,j=1}^{N} y_i^\gamma y_j^\gamma \langle\psi_i, (\Gamma + A\mathsf{C}_N(\theta)A^*)^{-1}\psi_j\rangle$$
$$= \frac{1}{2}\sum_{i,j=1}^{N} y_i^\gamma y_j^\gamma \cdot \frac{1}{a_i a_j}\langle\varphi_i, A^*(\Gamma + A\mathsf{C}_N(\theta)A^*)^{-1}A\varphi_j\rangle.$$

the Sherman–Morrison–Woodbury identity again, we may write

$$A^*(\Gamma + A\mathsf{C}_N(\theta)A^*)^{-1}A = A^*\Gamma^{-1}A - A^*\Gamma^{-1}A(A^*\Gamma^{-1}A + \mathsf{C}_N(\theta)^{-1})^{-1}A^*\Gamma^{-1}A$$
$$= \frac{1}{\gamma^2}A^*A - \frac{1}{\gamma^2}A^*A\left(\frac{1}{\gamma^2}A^*A + \mathsf{C}_N(\theta)^{-1}\right)^{-1}\frac{1}{\gamma^2}A^*A,$$

and so by the simultaneous diagonalizability, and orthonormality of $\{\varphi_j\}$,

$$\frac{1}{2}\|y\|^2_{\Gamma+A\mathsf{C}_N(\theta)A^*} = \frac{1}{2}\sum_{j=1}^{N}\frac{(y_i^\gamma)^2}{a_j^2}\left[\frac{a_j^2}{\gamma^2} - \frac{a_j^2}{\gamma^2}\left(\frac{a_j^2}{\gamma^2} - \frac{1}{\mu_j(\theta)}\right)^{-1}\frac{a_j^2}{\gamma^2}\right]$$
$$= \frac{1}{2}\sum_{j=1}^{N}\frac{(y_i^\gamma)^2}{\gamma^2}\left[1 - \frac{a_j^2\mu_j(\theta)}{s_j^\gamma(\theta)}\right]$$
$$= \frac{1}{2}\sum_{j=1}^{N}\frac{(y_i^\gamma)^2}{s_j^\gamma(\theta)}.$$

To deal with the log-determinant term, we use Lemma A.1 to see that

$$\frac{1}{2}\log\det(\Gamma + A\mathsf{C}_N(\theta)A^*) = \frac{1}{2}\log\det(A^*\Gamma A + A^*A\mathsf{C}_N(\theta)A^*A) - \frac{1}{2}\log\det(AA^*).$$

Since $\{\varphi_j\}$ is an orthonormal basis for X, the first determinant may be calculated as

$$\frac{1}{2}\log\det(A^*\Gamma A + A^*A\mathsf{C}_N(\theta)A^*A) = \frac{1}{2}\sum_{j=1}^{N}\log(a_j^2\gamma^2 + a_j^4\mu_j(\theta)) = \frac{1}{2}\sum_{j=1}^{N}\log(a_j^2 s_j^\gamma(\theta))$$

and so

$$\frac{1}{2}\log\det(\Gamma + A\mathsf{C}_N(\theta)A^*) \propto -\frac{1}{2}\sum_{j=1}^{N}\log\frac{s_j^\gamma(\theta^\dagger)}{s_j^\gamma(\theta)}$$

from which the result follows. ∎

4.2. **Convergence of Minimizers**

We study convergence of the minimizers of the random functionals $\mathsf{J}_{\mathsf{C}}^{N,\gamma}, \mathsf{J}_{\mathsf{NC}}^{N,\gamma}$ and $\mathsf{J}_{\mathsf{E}}^{N,\gamma}$ in the simultaneous limit $N \to \infty$ and $\gamma \to 0$. We establish that, if the noise level decays sufficiently fast relative to the smallest value of the product of the singular values and the prior covariance, for the truncated problem, then the true hyperparameter is recovered in the cases of the centred MAP and empirical Bayes estimates. We also establish that it is not recovered in the case of the noncentred MAP estimate.

Let $\gamma_N > 0$ denote the noise level when N observations are taken. We define $s_j^{\gamma_N}(\theta) = a_j^2\mu_j(\theta) + {\gamma_N}^2$ as in Proposition 4.3, and define

$$b_j^N(\theta) = \frac{s_j^{\gamma_N}(\theta^\dagger)}{s_j^{\gamma_N}(\theta)}.$$

In order to establish the convergence, we make the following assumptions.

Assumptions 2. *We assume in what follows that:*

(i) $\Theta \subseteq \mathbb{R}^k$ *is compact.*

(ii) $\min_{j=1,\dots,N} a_j^2\mu_j(\theta)/\gamma_N^2 \to \infty$ *as* $N \to \infty$ *for all* $\theta \in \Theta$.

(iii) $g(\theta,\theta^\dagger) := \lim_{j\to\infty} \frac{\mu_j(\theta^\dagger)}{\mu_j(\theta)}$ *exists for all* $\theta \in \Theta$, *and the map* $\theta \mapsto g(\theta,\theta^\dagger) - \log g(\theta,\theta^\dagger)$ *is lower semicontinuous.*

(iv) If $g(\theta,\theta^\dagger) = 1$, *then* $\theta = \theta^\dagger$.

(v) The maps $\theta \mapsto \log \mu_j(\theta)$ *are Lipschitz on* Θ *for each* $j \in \mathbb{N}$, *with Lipschitz constants uniformly bounded in* j.

(vi) The maps $\theta \mapsto b_j^N(\theta)$ *are Lipschitz on* Θ *for each* $j = 1,\dots,N$, $N \in \mathbb{N}$, *with Lipschitz constants uniformly bounded in* j, N.

(vii) The map $\theta \mapsto \log \rho_0(\theta)$ *is Lipschitz on* Θ.

Assumption (i) is made to avoid complications with hyperparameter estimates potentially diverging. Assumption (ii) gives the rate at which the noise must decay relative to the decay of the singular values of the (whitened) forward map – the more ill-posed the problem is, and the weaker the prior is, the faster the noise must vanish. Assumption (iii) allows a limiting functional to be identified, and (iv) is an identifiability assumption which allows us to identify the true hyperparameter. Assumptions (v)-(vii) are made to ensure the functionals $\mathsf{J}_{\mathsf{C}}^{N,\gamma_N}, \mathsf{J}_{\mathsf{NC}}^{N,\gamma_N}, \mathsf{J}_{\mathsf{E}}^{N,\gamma_N}$ are also Lipschitz with Lipschitz constants (almost surely) uniformly bounded in N; note that when combined with the assumed compactness of Θ, we thus obtain existence of minimizers of these functionals over Θ.

Remark 4.5. Instead of having the noise level γ_N a function of the number of observations, we could also consider having the number of observations N_γ as a function of the noise level – this may be more appropriate in practice as one may not have control over the noise level. In this case, one would need to replace Assumption (ii) with

$$\min_{j=1,\dots,N_\gamma} a_j^2\mu_j(\theta)/\gamma \to \infty \text{ as } \gamma \to 0 \tag{4.7}$$

in order to obtain analogous results. We work with γ_N to make the arguments clearer: our sequences of functionals are indexed by a discrete rather than continuous parameter.

Theorem 4.6. *Let Assumptions 2 hold, and let* $\{\theta_{\mathsf{C}}^N\}, \{\theta_{\mathsf{E}}^N\}, \{\theta_{\mathsf{NC}}^N\}$ *denote sequences of minimizers over* Θ *of* $\{\mathsf{J}_{\mathsf{C}}^{N,\gamma_N}\}$, $\{\mathsf{J}_{\mathsf{E}}^{N,\gamma_N}\}$, $\{\mathsf{J}_{\mathsf{NC}}^{N,\gamma_N}\}$ *respectively.*

(i) $\theta_{\mathsf{C}}^N, \theta_{\mathsf{E}}^N \to \theta^\dagger$ *in probability as* $N \to \infty$.

(ii) Assume further that $g(\cdot, \theta^\dagger)$ *has a unique minimizer* θ_*. *Then* $\theta_{\mathsf{NC}}^N \to \theta_*$ *in probability as* $N \to \infty$.

Remark 4.7. The identification of hierarchical parameters such as the smoothness of a Gaussian distribution has been widely studied in regression problems. See e.g. example 2.1 and the work by Gloter and Hoffmann [19] on estimation of Hurst parameter and their discussion therein. In terms of inverse problems, Knapik et al. [26] recently studied consistency of empirical maximum likelihood estimators for linear diagonalizable problem. However, the key difference in these previous studies is the data generating distribution: we assume the data to be generated according to $\mathbb{P}(y|\theta^\dagger)$, whereas Knapik et al. consider $\mathbb{P}(y|u^\dagger)$ to be the data generating distribution, and the true regularity $\theta^\dagger$ is defined implicitly from the function $u^\dagger$.

The difference of the two observational models can be highlighted by considering repeated samples. In our model the variable u will vary in each data sample whereas for the other approach $u^\dagger$ remains fixed. In this work, we consider consistency properties in terms of vanishing noise which makes the difference less transparent and it requires further work to study how identifiability of $\theta^\dagger$ from $u^\dagger$ in previous studies and our assumptions are related.

In the setting employed in [26] the authors are able to take their analysis further: their main results in [26, Thm. 1 and 2] show convergence *rates* of the empirical estimator, like us in probability, and they use this to deduce that the empirical posterior on u contracts around the ground truth at an optimal rate.

Remark 4.8. In general it is the case that $\theta_* \neq \theta^\dagger$, and so the result concerning the convergence of $\{\theta_{\mathsf{NC}}^N\}$ is a negative result: the true hyperparameter is not recovered.

Proof. [Proof of Theorem 4.6] We establish the result in full for $\mathsf{J}_{\mathsf{C}}^{N,\gamma_N}$, and note the small modifications required to establish the results for $\mathsf{J}_{\mathsf{E}}^{N,\gamma_N}$ and $\mathsf{J}_{\mathsf{NC}}^{N,\gamma_N}$. We start by proving item (i). We have

$$\mathsf{J}_{\mathsf{C}}^{N,\gamma_N}(\theta) = \frac{1}{2N}\sum_{j=1}^{N}\left[\frac{(y_j^{\gamma_N})^2}{s_j^{\gamma_N}(\theta)} - \log\frac{\mu_j(\theta^\dagger)}{\mu_j(\theta)}\right] - \frac{1}{N}\log\rho_0(\theta).$$

We rewrite $y_j^{\gamma_N}$ using the representation eq. (4.2):

$$y_j^{\gamma_N} \overset{\mathrm{d}}{=} \sqrt{s_j^{\gamma_N}(\theta^\dagger)}\zeta_j, \quad \zeta_j \overset{\mathrm{iid}}{\sim} N(0,1),$$

and so

$$\mathsf{J}_{\mathsf{C}}^{N,\gamma_N}(\theta) \overset{\mathrm{d}}{=} \frac{1}{2N}\sum_{j=1}^{N}\left[b_j^N(\theta)\zeta_j^2 - \log\frac{\mu_j(\theta^\dagger)}{\mu_j(\theta)}\right] - \frac{1}{N}\log\rho_0(\theta).$$

We can see formally from the assumptions that, for each $\theta \in \Theta$, $b_j^N(\theta) \to g(\theta, \theta^\dagger)$ as $j, N \to \infty$, and so the strong law of large numbers suggests that

$$\mathsf{J}_{\mathsf{C}}^{N,\gamma_N}(\theta) \to \mathsf{J}_{\mathsf{C}}(\theta) := \frac{1}{2}g(\theta,\theta^\dagger) - \frac{1}{2}\log g(\theta,\theta^\dagger) \tag{4.8}$$

almost surely. Observe that J_{C} is minimized if and only if $g(\theta, \theta^\dagger) = 1$, which by Assumptions 2(iv) occurs if and only if $\theta = \theta^\dagger$. We hence wish to establish convergence of the minimizers of $\mathsf{J}_{\mathsf{C}}^{N,\gamma_N}$ to that of J_{C}. In order to show this convergence, we use the approach of [46]. Specifically we use the result of Exercise 3.2.3, which follows from Corollary 3.2.3(ii) and the Arzelà-Ascoli theorem. We must establish that:

(a) $\mathsf{J}_{\mathsf{C}}^{N,\gamma_N}$ converges pointwise in probability to J_{C};

(b) the maps $\theta \mapsto \mathsf{J}_{\mathsf{C}}^{N,\gamma_N}(\theta)$ are Lipschitz on Θ for each N, with (random) Lipschitz coefficients uniformly bounded in N almost surely;

(c) J_{C} is lower semicontinuous with a unique minimum at $\theta^\dagger$; and

(d) $\theta_{\mathsf{C}}^N = \mathcal{O}_P(1)$.

The point (c) is true by assumption, and (d) follows since Θ is compact. To establish that point (a) holds, we note that it suffices to show that, in probability, for each $\theta \in \Theta$,

$$\left|\frac{1}{2N}\sum_{j=1}^{N} b_j^N(\theta)(\zeta_j^2 - 1)\right| \to 0 \tag{4.9}$$

$$\left|\frac{1}{2N}\sum_{j=1}^{N}\left(b_j^N(\theta) - \log\frac{\mu_j(\theta^\dagger)}{\mu_j(\theta)}\right) - \frac{1}{N}\log\rho(\theta) - \mathsf{J}_{\mathsf{C}}(\theta)\right| \to 0. \tag{4.10}$$

Note that the expression eq. (4.10) is deterministic. Define the map

$$G_N(\theta) = \frac{1}{2N}\sum_{j=1}^{N} b_j^N(\theta)(\zeta_j^2 - 1).$$

We show that $G_N(\theta) \to 0$ weakly for all $\theta \in \Theta$; since the limit is constant, the convergence then also occurs in probability. Combining Lemma A.2 with Assumptions 2(ii),(iii), we see that

$$\frac{1}{N}\sum_{j=1}^{N} b_j^N(\theta) \to g(\theta, \theta^\dagger) \tag{4.11}$$

for each $\theta \in \Theta$. The proof of Lemma A.2 implies, in particular, that the sequence $\{b_j^N(\theta)\}_{j,N}$ is uniformly bounded for each θ. Since $\zeta_j^2 \overset{\text{iid}}{\sim} \chi_1^2$, we have that the characteristic function of $G_N(\theta)$ satisfies[5]

$$\begin{aligned}
\mathbb{E}\Big(\exp\left(itG_N(\theta)\right)\Big) &= \prod_{j=1}^{N}\mathbb{E}\left(\exp\left(it\cdot\frac{1}{2N}b_j^N(\theta)(\zeta_j^2-1)\right)\right)\\
&= \prod_{j=1}^{N}\left(1-\frac{b_j^N(\theta)it}{N}\right)^{-\frac{1}{2}}\exp\left(-it\cdot\frac{1}{2N}b_j^N(\theta)\right)\\
&= \exp\left(-\frac{1}{2}\sum_{j=1}^{N}\left[\log\left(1-\frac{b_j^N(\theta)it}{N}\right)+\frac{b_j^N(\theta)it}{N}\right]\right)\\
&= \exp\left(-\frac{1}{2}\sum_{j=1}^{N}\left[-\frac{b_j^N(\theta)it}{N}-\frac{1}{2}\left(\frac{b_j^N(\theta)it}{N}\right)^2-\mathcal{O}(N^{-3})+\frac{b_j^N(\theta)it}{N}\right]\right)\\
&= \exp\left(-\frac{1}{4}\sum_{j=1}^{N}\left[\frac{b_j^N(\theta)^2t^2}{N^2}-\mathcal{O}(N^{-3})\right]\right).
\end{aligned}$$

From the boundedness of $\{b_j^N(\theta)\}_{j,N}$, we deduce that the sum in the exponent tends to zero as $N \to \infty$. It follows that

$$\mathbb{E}\left(\exp\left(itG_N(\theta)\right)\right) \to \exp(0) = \mathbb{E}^{Z\sim\delta_0}\left(\exp(itZ)\right)$$

[5]Here log refers to the principal branch of the complex logarithm – note that we are bounded away from the branch cut since the argument always has real part 1.

and so $G_N(\theta) \to 0$ weakly; the convergence eq. (4.9) follows. We now rewrite the expression in eq. (4.10) as

$$\frac{1}{2N}\sum_{j=1}^{N}\left(b_j^N(\theta) - \log\frac{\mu_j(\theta^\dagger)}{\mu_j(\theta)}\right) - \frac{1}{N}\log\rho_0(\theta) - \mathsf{J}_\mathsf{C}(\theta)$$
$$= \frac{1}{2N}\sum_{j=1}^{N}\left(b_j^N(\theta) - g(\theta,\theta^\dagger)\right) - \frac{1}{2N}\sum_{j=1}^{N}\left(\log\frac{\mu_j(\theta^\dagger)}{\mu_j(\theta)} - \log g(\theta,\theta^\dagger)\right) - \frac{1}{N}\log\rho_0(\theta).$$

The first sum vanishes as $N \to \infty$ due to the convergence eq. (4.11), the second vanishes due to Assumptions 2(ii), and third clearly vanishes. The convergence eq. (4.10) follows, and hence so does the pointwise convergence in probability $\mathsf{J}_\mathsf{C}^{N,\gamma_N} \to \mathsf{J}_\mathsf{C}$. It remains to show the Lipschitz condition (b). We have, for any $\theta_1, \theta_2 \in \Theta$,

$$|\mathsf{J}_\mathsf{C}^{N,\gamma_N}(\theta_1) - \mathsf{J}_\mathsf{C}^{N,\gamma_N}(\theta_2)| \le \frac{1}{2N}\sum_{j=1}^{N}|b_j^N(\theta_1) - b_j^N(\theta_2)|\zeta_j^2$$
$$+ \frac{1}{2N}\sum_{j=1}^{N}|\log\mu_j(\theta_1) - \log\mu_j(\theta_2)| + \frac{1}{2}|\log\rho_0(\theta_1) - \log\rho_0(\theta_2)|.$$

By Assumptions 2(v)-(vii) the Lipschitz property follows. The almost sure boundedness of the Lipschitz constants follows from the strong law of large numbers, since the i.i.d. random variables ζ_j^2 have finite second moments.

In the case of $\mathsf{J}_\mathsf{E}^{N,\gamma_N}$, the limiting functional is the same: $\mathsf{J}_\mathsf{E} = \mathsf{J}_\mathsf{C}$. The proof for convergence of minimizers differs only in the expression eq. (4.10), wherein the logarithmic term in the sum is replaced by $\log b_j^N(\theta)$; this does not affect the convergence of the expression.

We now study (ii). The functional $\mathsf{J}_\mathsf{NC}^{N,\gamma_N}$ differs from $\mathsf{J}_\mathsf{C}^{N,\gamma_N}$ only in the absence of the logarithmic term – it is easy to see that the limiting functional is then given by

$$\mathsf{J}_\mathsf{NC}^{N,\gamma}(\theta) := \frac{1}{2}g(\theta,\theta^\dagger),$$

and that the required conditions (a)–(d) above are satisfied, since existence of a unique minimizer θ_* of $g(\cdot,\theta^\dagger)$ is assumed. The same result from [46] may then be used to obtain the stated result. ∎

Remark 4.9. An important implication of this result is that the hyperparameters can only be determined up to measure equivalence. By the Feldman–Hájek theorem, the measures $N(0,\mathsf{C}(\theta^\dagger))$ and $N(0,\mathsf{C}(\theta))$ are equivalent if and only if

$$\sum_{j=1}^{\infty}\left(\frac{\mu_j(\theta^\dagger)}{\mu_j(\theta)} - 1\right)^2 < \infty$$

which in particular implies that the limit $g(\theta,\theta^\dagger)$ is identically 1. The limiting functional J_C in equation (4.8) is hence minimized by any θ that gives rise to an equivalent measure.

Remark 4.10. In some situations the limiting functional $g(\theta,\theta^\dagger)$ is infinite whenever $\theta \neq \theta^\dagger$. Even though this limit clearly identifies the true hyperparameters, Theorem 4.6 does not directly apply, since, for example, Assumptions 2(iii),(vii) cannot hold. One approach to avoid this is to replace the objective functional $\mathsf{J}_\mathsf{C}^{N,\gamma_N}(\theta)$ by $\mathsf{J}_\mathsf{C}^{N,\gamma_N}(\theta)^{\varepsilon_N}$ for some positive sequence $\varepsilon_N \to 0$ – note that this does not affect the sequence of minimizers since $t \mapsto t^{\varepsilon_N}$ is strictly increasing for all N. Such a sequence $\{\varepsilon_N\}$ may be chosen in practice to be such that $(\mu_N(\theta^\dagger)/\mu_N(\theta))^{\varepsilon_N}$ converges to a finite value for each θ as $N \to \infty$. Examples of situations where these infinite limits occur, and appropriate choices of sequences $\{\varepsilon_N\}$ to obtain finite limits, are discussed in what follows.

4.3. **Examples**

We now provide examples which elucidate Theorem 4.6.

Example 4.11 (Whittle–Matérn)**.** Consider the case where the conditional Gaussian priors are Whittle–Matérn distributions on a bounded domain $D \subseteq \mathbb{R}^d$. As mentioned in example 2.3, the covariance operators diagonalize in the eigenbasis $\{\varphi_j\}$ of the Laplacian on D. Since D is bounded we simply *define* the Whittle–Matern process to have covariance given by the inverse of (2.1), and where we equip the Laplacian with Dirichlet, Neumann or periodic boundary conditions; we note that for all three such sets of boundary conditions, the eigenvalues λ_j of the negative Laplacian tend to infinity. We first consider the case where we are hierarchical about the standard deviation σ and the length-scale ℓ, and denote $\theta = (\sigma, \ell) \in \Theta$. Fixing the regularity parameter $\nu > 0$, the eigenvalues are given by

$$\begin{aligned}\mu_j(\theta) &= \kappa(\nu)\sigma^2\ell^d(1+\ell^2\lambda_j)^{-\nu-d/2}\\ &= \kappa(\nu)\sigma^2\ell^{-2\nu}(\ell^{-2}+\lambda_j)^{-\nu-d/2}\end{aligned}$$

for some constant $\kappa(\nu)$. We may then calculate

$$g(\theta,\theta^\dagger) = \lim_{j\to\infty}\left(\frac{\sigma^\dagger}{\sigma}\right)^2\left(\frac{\ell}{\ell^\dagger}\right)^{2\nu}\left(1+\frac{\ell^{-2}-(\ell^\dagger)^{-2}}{(\ell^\dagger)^{-2}+\lambda_j}\right)^{\nu} = \left(\frac{\sigma^\dagger}{\sigma}\right)^2\left(\frac{\ell}{\ell^\dagger}\right)^{2\nu}.$$

We then see that $g(\theta,\theta^\dagger) = 1$ if and only if[6] $\sigma\ell^{-\nu} = \sigma^\dagger(\ell^\dagger)^{-\nu}$. This equality is satisfied by infinitely many pairs (σ,ℓ). In order to apply Theorem 4.6 we require that the equality is only satisfied by the true hyperparameters. Therefore, instead of attempting to infer the pair (σ,ℓ), we attempt to infer the pair $(\sigma,\beta) := (\sigma,\sigma\ell^{-\nu})$; this is closely related to the discussion around the Ornstein–Uhlenbeck process in example 2.1. We then have

$$\mu_j(\theta) = \kappa(\nu)\beta^2\left(\left(\frac{\beta}{\sigma}\right)^{2/\nu}+\lambda_j\right)^{-\nu-d/2}$$

which leads to

$$g(\theta,\theta^\dagger) = \left(\frac{\beta^\dagger}{\beta}\right)^2.$$

When σ is fixed, by applying Theorem 4.6, we can deduce that the parameter β is identifiable using the centred MAP and empirical Bayesian methods; the proof that the requisite assumptions are satisfied under appropriate conditions is provided in Lemma A.3. In particular, assuming the algebraic decay $a_j \asymp j^{-a}$ and $\gamma_N \asymp N^{-w}$, Assumptions 2(ii) is equivalent to

$$w > a + \frac{\nu}{d} + \frac{1}{2}. \tag{4.12}$$

We also see that the parameter β is not identifiable via the noncentred MAP method, since $g(\cdot;\theta)$ is minimized by taking β as large as possible.

In the case where we are hierarchical about the regularity parameter ν, the assumptions of Theorem 4.6 do not hold. Nonetheless, the limiting functional can still be formally calculated as

$$\mathsf{J}_\mathsf{C}(\nu) = \begin{cases}\infty & \nu \neq \nu^\dagger\\ 1 & \nu = \nu^\dagger\end{cases}$$

which is clearly minimized if and only if $\nu = \nu^\dagger$. As discussed in Remark 4.10, we can rescale to obtain a finite limiting functional; in this case making the choice $\varepsilon_N = 1/\log(1+\lambda_N)$ achieves this.

[6]This condition is slightly weaker than that required for measure equivalence – for the measures to be equivalent we require in addition that $d \leq 3$, see for example Theorem 1 in [17].

Example 4.12 (Automatic Relevance Determination)**.** The Automatic Relevance Determination (ARD) kernel is typically defined by

$$c(x, x'; \theta) = \sigma^2 \exp\left(-\frac{1}{2}\sum_{k=1}^{d}\left(\frac{x_k - x'_k}{\theta_k}\right)^2\right).$$

This is the Green's function for the anisotropic heat equation at time $t = 1$:

$$\frac{\partial u}{\partial t}(t, x) = \sum_{k=1}^{d} \theta_k^2 \frac{\partial u^2}{\partial x_k^2}(t, x), \quad u(0, x) = \sigma^2 \xi(x).$$

The corresponding covariance operator is hence given by

$$\mathsf{C}(\theta) = \sigma^2 \exp(\Delta_\theta) := \sigma^2 \exp\left(-\sum_{k=1}^{d} \theta_k^2 \frac{\partial^2}{\partial x_k^2}\right).$$

On rectangular domains this family of operators is simultaneously diagonalizable under the Laplacian eigenbasis. For example, if $D = (0, 1)^d$ and we impose Dirichlet boundary conditions on the Laplacian, then the eigenvalues are given by

$$\mu_{i_1,\dots,i_d}(\theta) = \sigma^2 \exp\left(-\pi^2 \sum_{k=1}^{d} \theta_k^2 i_k^2\right).$$

The results we have concerning consistency are given in terms of eigenvalues indexed by a single index j rather than a multi-index $(i_1, \dots, i_d)$. Rather than consider a particular enumeration of the multi-indices, we instead aim to infer each hyperparameter θ_k individually by only sending $i_k \to \infty$ – this amounts to taking a subset of the observations. The problem of inferring each θ_k is then essentially equivalent to inference of the length-scale parameter of squared exponential prior with $d = 1$.

Note that Theorem 4.6 does not apply in this case – the limiting functional J_C is infinite everywhere except for the true hyperparameter, as was the case when inferring the parameter ν in the previous example. Again, following Remark 4.10, we can rescale to obtain a finite objective function; in this case making the choice $\varepsilon_N = 1/N^2$ suffices. ARD versions of general Whittle–Matérn covariances can also be obtained by replacing the negative Laplacian $-\Delta$ with its anistropic analogue $-\Delta_\theta$ within the precision operator. It can be verified that the requisite assumptions for Theorem 4.6 are satisfied in this case when $\nu < \infty$; the proof is almost identical to that of Lemma A.3 and is hence omitted for brevity.

5. Numerical Experiments

In this section we present a number of numerical experiments in order to both validate the theory presented, and illustrate how the theory may extend beyond what has been proven. Subsection 5.1 introduces a diagonalizable deblurring problem which is considered in the subsequent subsections. Subsection 5.2 looks at the behaviour of minimizers of $\mathsf{J}_\mathsf{C}^{N,\gamma}$ with and without the prior truncation, as discussed in Remark 4.2. Subsection 5.3 looks at the traces of the errors between the hyperparameter estimates, comparing the convergence rates between the different functionals. Subsection 5.4 considers the setup of example 4.11, wherein the variance and length-scale parameters are to be jointly inferred; the minimizers are confirmed numerically to lie on the curve of hyperparameters which give rise to equivalent measures. Finally, Section 5.5 considers settings that enable us to test whether the assumptions of the theory are sharp – in particular we see that they appear sharp only for the centred MAP approach, with the empirical Bayes estimates appearing to be more robust with respect to noise.

5.1. **Deblurring Problem**

In this subsection we consider the case that the forward map is given by a linear blurring operator. Let $\{\varphi_j\}_{j=0}^{\infty}$ denote the cosine Fourier basis on $D = (0,1)$,

$$\varphi_j(x) = \sqrt{2}\cos(\pi j x),$$

and define $A : L^2(D) \to L^2(D)$ by

$$\langle Au, \varphi_j \rangle = \begin{cases} j^{-2}\langle u, \varphi_j \rangle & j \geq 1 \\ 0 & j = 0. \end{cases}$$

Then the map A may be viewed as the solution operator $f \mapsto u$ for the problem

$$-\Delta u(x) = \pi^2 f(x) \quad \text{for } x \in D, \quad u'(0) = u'(1) = 0, \quad \int_0^1 u(x)\,\mathrm{d}x = 0. \tag{5.1}$$

It could equivalently viewed as a convolution operator, writing

$$(Au)(x) = \int_0^1 G(x,x')u(x')\,\mathrm{d}x'$$

where $G(x,x')$ is the Green's function for the system eq. (5.1). This choice of forward operator is convenient as it diaganalizes in the same basis as the Whittle–Matérn covariance operators on D, which are what we use throughout this subsection. In fig. 5.1 we show the true state $u^\dagger$ that we fix throughout this subsection, and its image $Au^\dagger$ under A. It is drawn from a Whittle–Matérn distribution with parameters $\sigma^\dagger = \ell^\dagger = 1$, $\nu^\dagger = 3/2$. To be explicit, in the notation of Section 4, we have

$$a_j = 1/j^2, \quad \mu_j(\theta) = \sigma^2 \ell^{-2\nu}(\pi^2 j^2 + \ell^{-2})^{-\nu}.$$

5.2. **Prior Truncation**

We first provide some numerical justification for the truncation of the prior at the same level as the observations when using the centred parameterization, as discussed in Remark 4.2. We fix a maximum discretization level $N_{\max} = 10^5$, and look at the behaviour of minimizers of the two functionals

$$\mathsf{J}_{\mathsf{C}}^{N,\gamma_N}(\theta) \propto \frac{1}{2}\sum_{j=1}^{N}\frac{(y_j^\gamma)^2}{s_j^\gamma(\theta)} - \frac{1}{2}\sum_{j=1}^{N}\log\frac{\mu_j(\theta^\dagger)}{\mu_j(\theta)} - \log\rho_0(\theta),$$

$$\widetilde{\mathsf{J}_{\mathsf{C}}^{N,\gamma_N}}(\theta) \propto \frac{1}{2}\sum_{j=1}^{N}\frac{(y_j^\gamma)^2}{s_j^\gamma(\theta)} - \frac{1}{2}\sum_{j=1}^{N_{\max}}\log\frac{\mu_j(\theta^\dagger)}{\mu_j(\theta)} - \log\rho_0(\theta),$$

as N is increased. We consider a conditional Whittle–Matérn prior, treating the inverse length-scale $\theta = \ell^{-1}$ as a hyperparameter, and set $\gamma_N = 1/N^5$ so that eq. (4.12) is satisfied. In fig. 5.2 we show how the errors between the estimated inverse length-scales and the truth compare between the two functionals as N increases. It can be seen that the error for the truncated prior is bounded above by that for the full prior, as expected.

5.3. **Centred, Noncentred and Empirical Bayes**

We now compare numerically the behaviour of optimizers of the three functionals $\mathsf{J}_{\mathsf{C}}^{N,\gamma_N}$, $\mathsf{J}_{\mathsf{NC}}^{N,\gamma_N}$ and $\mathsf{J}_{\mathsf{E}}^{N,\gamma_N}$, and verify that the conclusions of Theorem 4.6 hold. As above, we consider a conditional Whittle–Matérn prior with the inverse length-scale $\theta = \ell^{-1}$ as a hyperparameter, and set $\gamma_N = 1/N^5$. In fig. 5.3 we show how the errors between the three sequences of minimizers and the truth compare

as N increases. We see that the noncentred MAP error diverges, as expected: the limiting functional is given by

$$\mathsf{J}_{\mathsf{NC}}(\theta) = \frac{1}{2}\left(\frac{\ell}{\ell^\dagger}\right)^3,$$

which is minimized as $\ell^{-1} \to \infty$. The empirical Bayes and centred MAP errors both generally decrease as N is increased, again as expected, with the empirical Bayes estimate slightly outperforming the centred MAP estimate for moderate N; the noncentred MAP estimator fails to converge.

Also in fig. 5.3 we show the same errors averaged over 1000 independent realizations of the truth $u^\dagger \sim N(0, \mathsf{C}(\theta^\dagger))$ and noise $\eta \sim N(0, \mathsf{I})$, and the same behaviour is observed. A reason for the empirical Bayes estimate outperforming the noncentred MAP estimate for moderate N may be that the terms in the summation in the functional eq. (4.6) taking the form $x_j - \log x_j$, rather than $x_j - \log x_j^\varepsilon$ for some $x_j^\varepsilon \approx x_j$ as in eq. (4.4), which is minimized by $x_j = 1$. For larger N there is very little difference between the two functionals, since $\gamma_N \to 0$.

5.4. **Equivalent Families of Measures**

We now consider the same setup as the previous subsubsection, but treat both the inverse length-scale ℓ^{-1} and the standard deviation σ as hyperparameters: $\theta = (\sigma, \ell^{-1})$. The resulting family of conditional prior measures are then equivalent along any curve $\{(\sigma, \ell^{-1}) \,|\, \sigma\ell^{-\nu} = \text{constant}\}$, as discussed in example 4.11, and so the hyperparameters cannot be identified beyond this curve. This is illustrated in fig. 5.4. In the top row we plot the functional $\mathsf{J}_{\mathsf{C}}^{N,\gamma_N}(\sigma, \ell)$ for $(\sigma, \ell^{-1}) \in (0,5)^2$, with N increasing from left to right. In the bottom row we plot the sets

$$\left\{(\sigma, \ell^{-1}) \,\middle|\, \ell \in \operatorname*{arg\,min}_{\ell^{-1}\in(0,5)} \mathsf{J}_{\mathsf{C}}^{N,\gamma_N}(\sigma, \ell)\right\}, \quad \left\{(\sigma, \ell^{-1}) \,\middle|\, \sigma \in \operatorname*{arg\,min}_{\sigma\in(0,5)} \mathsf{J}_{\mathsf{C}}^{N,\gamma_N}(\sigma, \ell)\right\},$$

along with the curve $\sigma\ell^{-\nu} = \sigma^\dagger(\ell^\dagger)^{-\nu}$, i.e. $g(\cdot, \theta^\dagger)^{-1}(1)$; the global minimizer (i.e. the intersection of these sets) is also shown as a green dot. We see that the sets of minimizers concentrate on the limiting curve $g(\cdot, \theta^\dagger)^{-1}(1)$ as N is increased.

For reference, we also consider the same experiments, but working with the reparameterization $\theta = (\sigma, \beta)$ introduced in example 4.11 so that β should be identifiable. In fig. 5.5 we see that this is indeed the case, with the curves now concentrating on the line $\beta = \sigma^\dagger(\ell^\dagger)^{-\nu} = 1$ as N is increased.

5.5. **Noise Decay Rate**

We choose here now to be hierarchical about just the inverse length-scale $\theta = \ell^{-1}$. In the theory we made the assumption Assumptions 2(ii) concerning the decay rate of the forward map and covariance singular values versus the decay of the noise level. For Whittle–Matérn priors, assuming the algebraic decay $a_j \asymp j^{-a}$ and $\gamma_N \asymp N^{-w}$, the required condition on w for Assumptions 2(ii) to hold is given by eq. (4.12). In the setup considered here, this translates to $w > 4$. We now investigate numerically whether this condition is sharp, making the three choices $\gamma_N = N^{-w}$ for $w = 3.5, 4, 4.5$. The resulting error traces are shown in fig. 5.6 for the centred MAP and empirical Bayesian methods. It appears that the condition is likely to be sharp for the centred optimization, given that convergence fails at the borderline case. However. For the empirical Bayesian optimization the condition does not appear to be necessary, with convergence occurring in all cases, suggesting it is a more stable estimator than the MAP.

In light of Remark 4.5, we also consider the same setup, but with a fixed noise level and increasing N. Making the choice $N_\gamma = \gamma^{-1/w}$, the condition on w, equivalent to eq. (4.7), is the same as before. In fig. 5.7 we show the errors for the choices $w = 3.5.4, 4.5$, and the same trends are observed as for fig. 5.6.

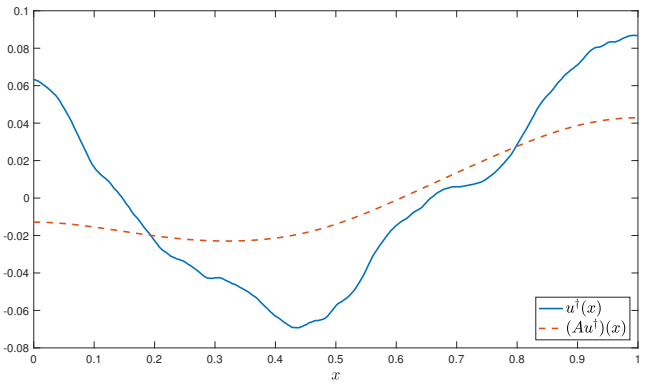

FIGURE 5.1. The true state $u^\dagger$ used throughout Section 5.1, and its image $Au^\dagger$ under the blurring operator A.

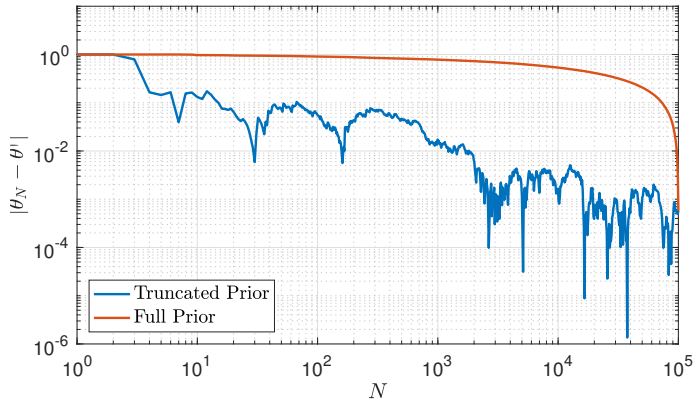

FIGURE 5.2. The trace of the errors between the minimizers of $\mathsf{J}_{\mathsf{NC}}^{N,\gamma_N}$ and $\widetilde{\mathsf{J}_{\mathsf{NC}}^{N,\gamma_N}}$ and the true hyperparameter, as defined in Section 5.2, as N is increased.

6. Conclusions

Learning hyperparameters in Bayesian hierarchical inference is important in two main contexts: when the hyperparameters themselves are the primary object of inference, and the underlying quantity which depends on them *a priori* is viewed as a nuisance parameter; when the hyperparameters themselves are not of direct interest, but choosing them carefully aids in inferring the underlying quantity which depends on them *a priori*. In both settings it is of interest to understand when hyperparameters can be accurately inferred from data. In this paper we have studied this question within the context of MAP estimation. Our work suggests the benefits of using the centred parameterization over the noncentred one, and also supports the use of empirical Bayes procedures. This is interesting because the relative merits of centering and noncentering in this context differ from what is found for sampling methods such as MCMC.

The theorem is confined to a straightforward situation, concerning linear inverse problems, in which the relevant operators are simultaneously diagonalizable. It also imposes conditions on the parameters defining the problem; numerical experiments indicate that these are sharp for the centred MAP estimator, but not for the empirical Bayes estimator, demonstrating that the latter is preferable. It

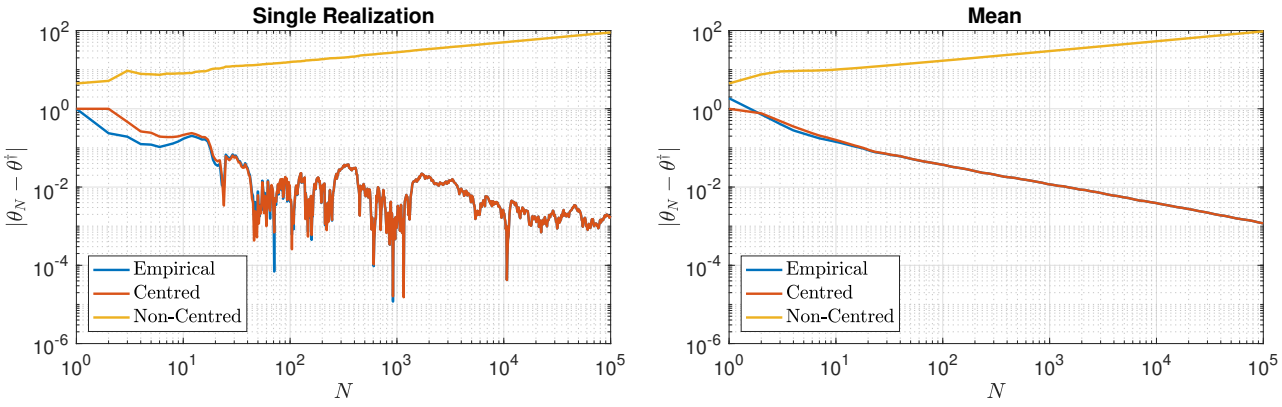

FIGURE 5.3. Comparison of the errors between the minimizers of the three functionals $\mathsf{J}_{\mathsf{E}}^{N,\gamma_N}, \mathsf{J}_{\mathsf{C}}^{N,\gamma_N}, \mathsf{J}_{\mathsf{NC}}^{N,\gamma_N}$ and the true hyperparameter, as N is increased. The left figure shows the error traces for a single realization of the truth and the noise, and the right figure shows the errors averaged over 1000 such realizations.

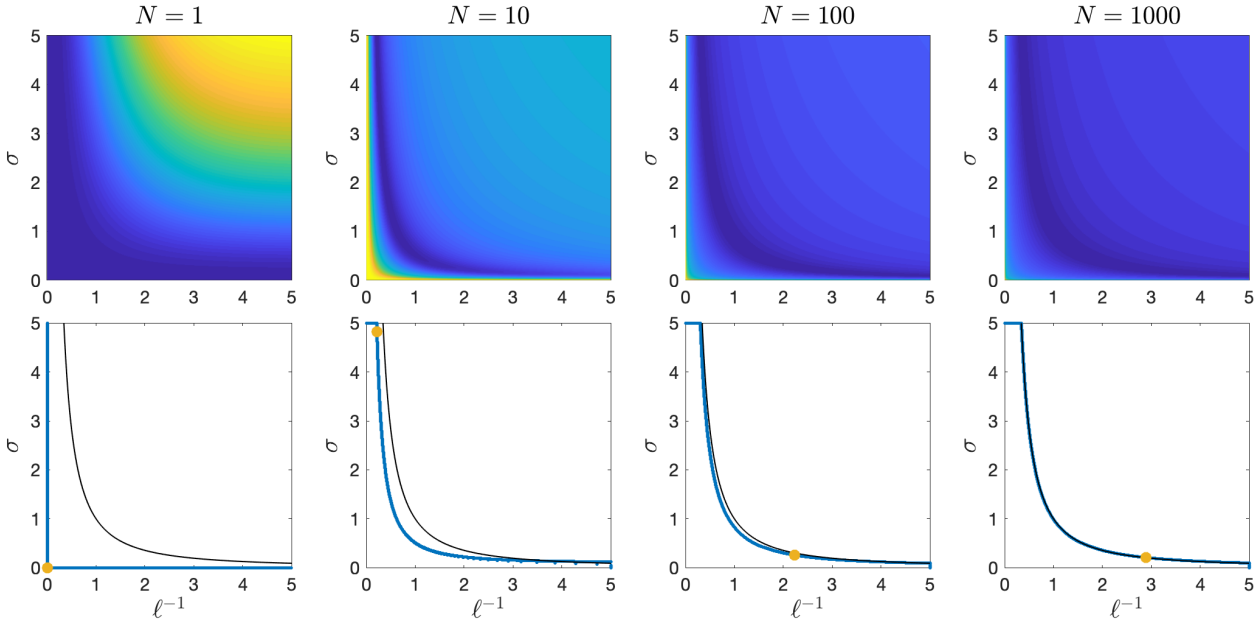

FIGURE 5.4. (Top) The objective function $\mathsf{J}_{\mathsf{C}}^{N,\gamma_N}(\sigma, \ell)$ for $N = 1, 10, 100, 1000$. (Bottom) The locations of the minimizers of each $\mathsf{J}_{\mathsf{C}}^{N,\gamma_N}(\sigma, \ell)$ across each row and column of the computed approximations (blue), the curve of parameters that produce equivalent measures to the true parameter (black), and the global optimizer (yellow).

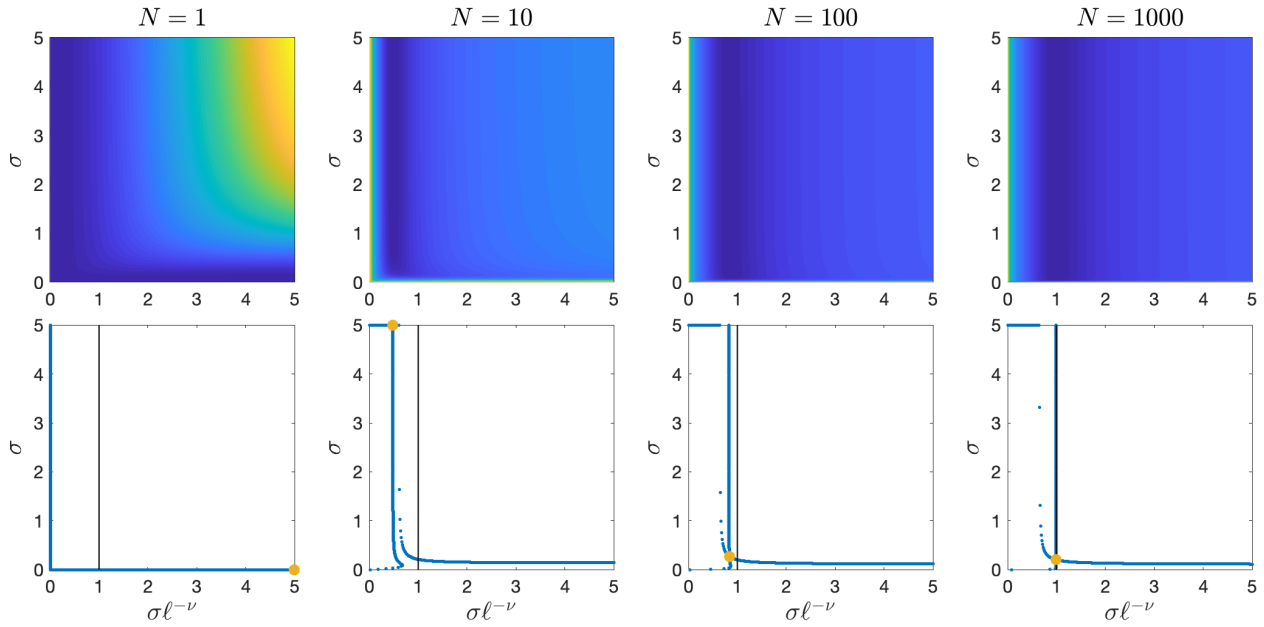

FIGURE 5.5. (Top) The objective function $\mathsf{J}_{\mathsf{C}}^{N,\gamma_N}(\sigma,\beta)$ for $N = 1, 10, 100, 1000$. (Bottom) The locations of the minimizers of each $\mathsf{J}_{\mathsf{C}}^{N,\gamma_N}(\sigma,\beta)$ across each row and column of the computed approximations (blue), the curve of parameters that produce equivalent measures to the true parameter (black), and the global optimizer (yellow).

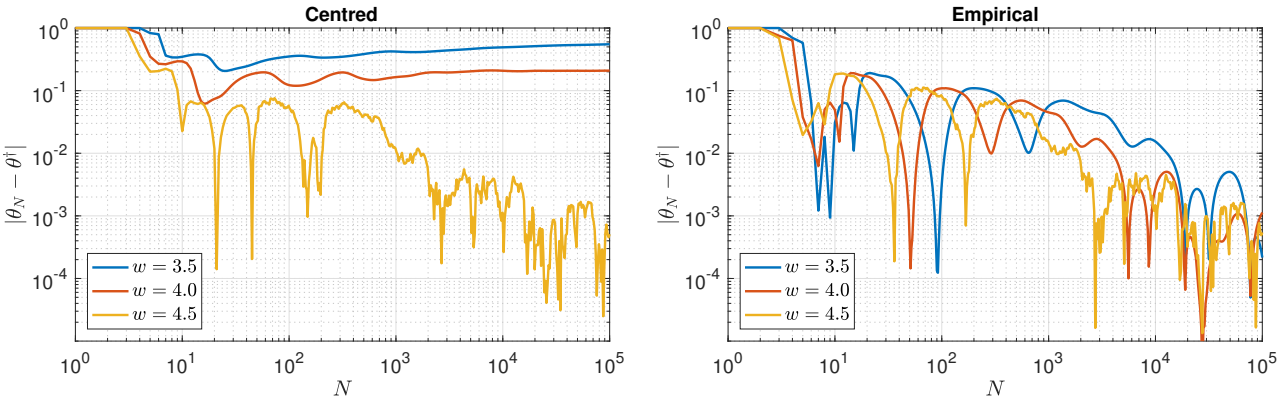

FIGURE 5.6. Traces of the errors between optimizers of $\mathsf{J}_{\mathsf{C}}^{N,\gamma_N}(\theta)$ (left), $\mathsf{J}_{\mathsf{E}}^{N,\gamma_N}(\theta)$ (right) and the true hyperparameter, as N is increased. Here the noise level γ_N is taken as $\gamma_N = N^{-w}$ for $w = 3.5, 4, 4.5$.

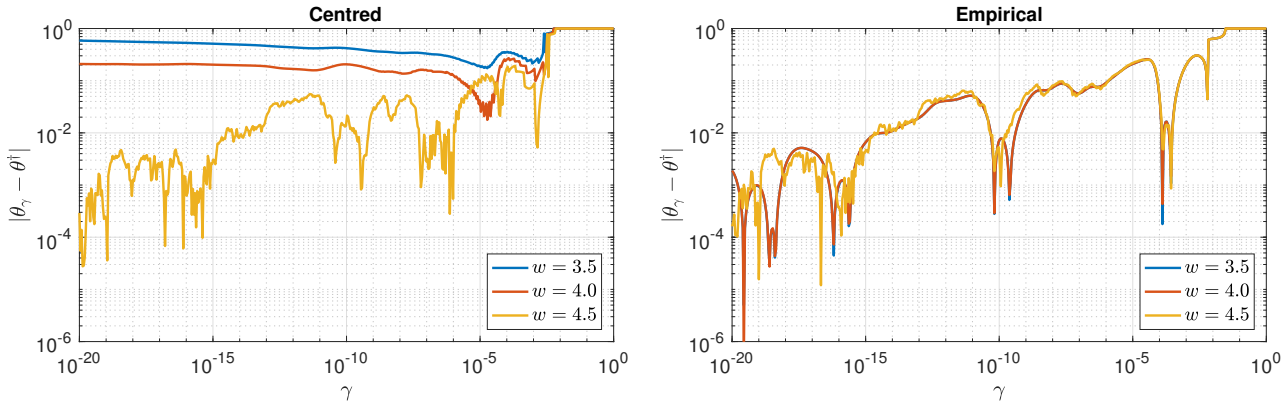

FIGURE 5.7. Traces of the errors between optimizers of $\mathsf{J}_{\mathsf{C}}^{N_\gamma,\gamma}(\theta)$ (left), $\mathsf{J}_{\mathsf{E}}^{N_\gamma,\gamma}(\theta)$ (right) and the true hyperparameter, as γ is decreased. Here the number of observations N_γ is taken as $N_\gamma = \gamma^{-1/w}$ for $w = 3.5, 4, 4.5$.

would also be of interest to push the boundaries of the theory outside this regime to the non-diagonal setting and even into nonlinear inverse problems. It would also be of interest to study fully Bayesian posterior inference for the hyperparameters, and Bernstein-von Mises theorems; this may be related the re-scalings needed at the end of examples 4.11 and 4.12.

Acknowledgements

The work of AMS and MMD is funded by US National Science Foundation (NSF) grant DMS 1818977 and AFOSR Grant FA9550-17-1-0185. The work of TH is funded by the Academy of Finland grant 326961.

Appendix A. Supporting Lemmas

In this appendix we provide a number of lemmas that are used during proofs and examples in the main text.

Lemma A.1. *Let $m \geq n$, $A \in \mathbb{R}^{m\times n}$ and $Q \in \mathbb{R}^{m\times m}$. Then*

$$\det(A^*QA) = \det(Q)\det(AA^*).$$

Proof. Let $A = U\Sigma V^*$ be the singular value decomposition of A, with $U \in \mathbb{R}^{m\times m}$, $V \in \mathbb{R}^{n\times n}$ unitary, and $\Sigma \in \mathbb{R}^{m\times n}$. Then we have

$$\det(A^*QA) = \det(V\Sigma^*U^*QU\Sigma V^*) = \det(\Sigma^*U^*QU\Sigma)\det(V^*V) = \det(\Sigma^*U^*QU\Sigma).$$

We have that

$$\begin{aligned}(\Sigma^*U^*QU\Sigma)_{ij} &= \begin{cases}\Sigma_{ii}(U^*QU)_{ij}\Sigma_{jj} & i,j \leq m\\ 0 & i > m \text{ or } j > m\end{cases}\\ &= (\hat{\Sigma}U^*QU\hat{\Sigma})_{ij}\end{aligned}$$

where $\hat{\Sigma} \in \mathbb{R}^{m\times m}$ is given by $\hat{\Sigma}_{ij} = \Sigma_{ij}$. Since all matrices are now square, we see that

$$\begin{aligned}\det(\Sigma^* U^* Q U \Sigma) &= \det(\hat{\Sigma} U^* Q U \hat{\Sigma}) \\ &= \det(Q)\det(U^*U)\det(\hat{\Sigma}^2) \\ &= \det(Q)\det(\Sigma\Sigma^*) \\ &= \det(Q)\det(AA^*).\end{aligned}$$

■

Lemma A.2. *Let $\{a_j\}, \{\mu_j\}, \{\bar{\mu}_j\}$ and $\{\gamma_j\}$ be positive sequences with $\bar{\mu}_j/\mu_j \to g \geq 0$. Then if*

$$\min_{j=1,\dots,N} \frac{a_j^2 \mu_j}{\gamma_N^2} \to \infty \quad \text{as } N \to \infty$$

we have

$$\frac{1}{N}\sum_{j=1}^{N} \frac{a_j^2\bar{\mu}_j + \gamma_N^2}{a_j^2\mu_j + \gamma_N^2} \to g \quad \text{as } N \to \infty.$$

Proof. We write

$$\begin{aligned}\frac{a_j^2\bar{\mu}_j + \gamma_N^2}{a_j^2\mu_j + \gamma_N^2} &= \frac{\bar{\mu}_j + \gamma_N^2/a_j^2}{\mu_j + \gamma_N^2/a_j^2} \\ &= \frac{\bar{\mu}_j}{\mu_j} + \left(\frac{\bar{\mu}_j + \gamma_N^2/a_j^2}{\mu_j + \gamma_N^2/a_j^2} - \frac{\bar{\mu}_j}{\mu_j}\right) \\ &= \frac{\bar{\mu}_j}{\mu_j} + \frac{\gamma_N^2/a_j^2(\mu_j - \bar{\mu}_j)}{\mu_j^2 + \gamma_N^2/a_j^2\mu_j} \\ &= \frac{\bar{\mu}_j}{\mu_j} + \frac{1}{a_j^2\mu_j/\gamma_N^2 + 1}\left(1 - \frac{\bar{\mu}_j}{\mu_j}\right).\end{aligned}$$

Now observe that

$$\left|\frac{1}{N}\sum_{j=1}^{N} \frac{a_j^2\bar{\mu}_j + \gamma_N^2}{a_j^2\mu_j + \gamma_N^2} - g\right| \leq \left|\frac{1}{N}\sum_{j=1}^{N}\left(\frac{a_j^2\bar{\mu}_j + \gamma_N^2}{a_j^2\mu_j + \gamma_N^2} - \frac{\bar{\mu}_j}{\mu_j}\right)\right| + \left|\frac{1}{N}\sum_{j=1}^{N}\frac{\bar{\mu}_j}{\mu_j} - g\right|.$$

The second term on the right hand side tends to zero by the assumed convergence. From the above, the first term is equal to

$$\begin{aligned}\left|\frac{1}{N}\sum_{j=1}^{N}\frac{1}{a_j^2\mu_j/\gamma_N^2 + 1}\left(1 - \frac{\bar{\mu}_j}{\mu_j}\right)\right| &\leq \max_{j=1,\dots,N}\frac{1}{a_j^2\mu_j/\gamma_N^2 + 1}\left|1 - \frac{\bar{\mu}_j}{\mu_j}\right| \\ &\leq C\left(\min_{j=1,\dots,N} a_j^2\mu_j/\gamma_N^2 + 1\right)^{-1}\end{aligned}$$

again using the assumed convergence of the ratio $\bar{\mu}_j/\mu_j$; the result follows. ■

In the following, given two sequence $\{a_j\}, \{b_j\}$, we write $a_j \asymp b_j$ if there exist constants $c_1, c_2 > 0$ such that $c_1 a_j \leq b_j \leq c_2 a_j$ for all j.

Lemma A.3. *Let $\Theta = [\beta_-, \beta_+] \subseteq (0,\infty)$. Given $\nu, \sigma > 0$, $d \in \mathbb{N}$ and a positive sequence $\lambda_j \asymp j^{2/d}$ define*

$$\mu_j(\beta) = \beta^2\left(\left(\frac{\beta}{\sigma}\right)^{2/\nu} + \lambda_j\right)^{-\nu - d/2}.$$

Assume that $a_j \asymp j^{-a}$ *and* $\gamma_N \asymp N^{-w}$*, where* $w, a > 0$ *are such that*

$$w > a + \frac{\nu}{d} + \frac{1}{2}.$$

Then Assumptions 2(i)-(vi) hold.

Proof.

(i) This is true by assumption.

(ii) We assume without loss of generality that a_j, λ_j are monotonically decreasing. Then

$$\min_{j=1,\dots,N} \frac{a_j^2 \mu_j(\beta)}{\gamma_N^2} = \frac{a_N^2 \mu_N(\beta)}{\gamma_N^2}.$$

We may bound the right hand side as

$$\frac{a_N^2 \mu_N(\beta)}{\gamma_N^2} \asymp N^{2(w-a)} \beta^2 \left(\left(\frac{\beta}{\sigma} \right)^{2/\nu} + \lambda_N \right)^{-\nu - d/2} \asymp N^{2(w-a-\nu/d-1/2)},$$

which diverges given the assumption on the parameters.

(iii) In example 4.11 it is demonstrated that

$$g(\beta, \beta^\dagger) = \lim_{j \to \infty} \frac{\mu_j(\beta^\dagger)}{\mu_j(\beta)} = \left(\frac{\beta^\dagger}{\beta} \right)^2$$

for all $\beta \in \Theta$. The map $g(\beta, \beta^\dagger) - \log g(\beta, \beta^\dagger)$ is clearly continuous on Θ, and so in particular lower semicontinuous.

(iv) This is clearly true.

(v) We have that

$$\log \mu_j(\beta) = 2 \log \beta - \left(\nu + \frac{d}{2} \right) \log \left(\left(\frac{\beta}{\sigma} \right)^{2/\nu} + \lambda_j \right)$$

which is smooth on Θ, and so

$$\begin{aligned} \left| \frac{\mathrm{d}}{\mathrm{d}\beta} \log \mu_j(\beta) \right| &= \left| \frac{2}{\beta} - \left(\nu + \frac{d}{2} \right) \frac{2}{\nu} \left(\frac{\beta}{\sigma} \right)^{2/\nu - 1} \left(\left(\frac{\beta}{\sigma} \right)^{2/\nu} + \lambda_j \right)^{-1} \right| \\ &\leq \frac{2}{\beta_-} + \left(\nu + \frac{d}{2} \right) \frac{2}{\nu} \frac{\sigma}{\beta_-}. \end{aligned}$$

It follows that $\log \mu_j(\beta)$ is Lipschitz with Lipschitz constants bounded in j.

(vi) The map $b_j^N(\beta)$ is smooth on Θ, and we have that

$$|(b_j^N)'(\beta)| = \left| b_j^N(\beta) \frac{\mu_j'(\beta)}{\mu_j(\beta) + \gamma_N^2 / a_j^2} \right| \leq \left| b_j^N(\beta) \right| \left| \frac{\mu_j'(\beta)}{\mu_j(\beta)} \right| = \left| b_j^N(\beta) \right| \left| \frac{\mathrm{d}}{\mathrm{d}\beta} \log \mu_j(\beta) \right|,$$

and the final term on the right hand side is uniformly bounded by part (v). Finally observe that

$$|b_j^N(\theta)| \leq \frac{\mu_j(\beta^\dagger)}{\mu_j(\beta)} + \frac{\gamma_N^2}{a_j^2 \mu_j(\beta)}.$$

The first term can be seen to be uniformly bounded by noting that $\beta_- > 0$, and the second term by using part (ii). The map $\beta_j^N(\beta)$ is hence Lipschitz with Lipschitz constants bounded in j, N.

■

References

[1] Sergios Agapiou, Johnathan M. Bardsley, Omiros Papaspiliopoulos, and Andrew M. Stuart. Analysis of the Gibbs sampler for hierarchical inverse problems. *SIAM/ASA J. Uncertain. Quantif.*, 2(1):511–544, 2014.

[2] Sergios Agapiou, Martin Burger, Masoumeh Dashti, and Tapio Helin. Sparsity-promoting and edge-preserving maximum a posteriori estimators in non-parametric Bayesian inverse problems. *Inverse Probl.*, 34(4), 2018.

[3] Sergios Agapiou, Masoumeh Dashti, and Tapio Helin. Rates of contraction of posterior distributions based on p-exponential priors. `https://arxiv.org/abs/1811.12244`, 2018.

[4] Sergios Agapiou, Stig Larsson, and Andrew M. Stuart. Posterior contraction rates for the Bayesian approach to linear ill-posed inverse problems. *Stochastic Processes Appl.*, 123(10):3828–3860, 2013.

[5] Sergios Agapiou and Peter Mathé. Posterior Contraction in Bayesian Inverse Problems Under Gaussian Priors. In *New Trends in Parameter Identification for Mathematical Models*, pages 1–29. Springer, 2018.

[6] Sergios Agapiou, Andrew M. Stuart, and Yuan-Xiang Zhang. Bayesian posterior contraction rates for linear severely ill-posed inverse problems. *J. Inverse Ill-Posed Probl.*, 22(3):297–321, 2014.

[7] James O. Berger. *Statistical Decision Theory and Bayesian Analysis*. Springer, 2013.

[8] Alexandros Beskos, Ajay Jasra, Ege A. Muzaffer, and Andrew M. Stuart. Sequential Monte Carlo methods for Bayesian elliptic inverse problems. *Stat. Comput.*, 25(4):727–737, 2015.

[9] Alexandros Beskos, Gareth Roberts, Andrew M. Stuart, and Jochen Voss. MCMC methods for diffusion bridges. *Stoch. Dyn.*, 8(03):319–350, 2008.

[10] Neil K. Chada, Marco A. Iglesias, Lassi Roininen, and Andrew M. Stuart. Parameterizations for ensemble Kalman inversion. *Inverse Probl.*, 34(5), 2018.

[11] Victor Chen, Matthew M. Dunlop, Omiros Papaspiliopoulos, and Andrew M. Stuart. Dimension-Robust MCMC in Bayesian Inverse Problems. `https://arxiv.org/abs/1806.00519`, 2018.

[12] Christian Clason, Tapio Helin, Remo Kretschmann, and Petteri Piiroinen. Generalized modes in Bayesian inverse problems. *SIAM/ASA J. Uncertain. Quantif.*, 7(2):652–684, 2019.

[13] Simon L. Cotter, Gareth Roberts, Andrew M. Stuart, and David White. MCMC methods for functions: modifying old algorithms to make them faster. *Stat. Sci.*, 28(3):424–446, 2013.

[14] Yair Daon and Georg Stadler. Mitigating the Influence of the Boundary on PDE-based Covariance Operators. *Inverse Probl. Imaging*, 12(5):1083–1102, 2018.

[15] Masoumeh Dashti, Kody JH Law, Andrew M. Stuart, and Jochen Voss. MAP estimators and their consistency in Bayesian nonparametric inverse problems. *Inverse Probl.*, 29(9), 2013.

[16] Masoumeh Dashti and Andrew M. Stuart. *The Bayesian approach to inverse problems*, pages 311–428. Springer, 2017.

[17] Matthew M. Dunlop, Marco A. Iglesias, and Andrew M. Stuart. Hierarchical Bayesian level set inversion. *Stat. Comput.*, pages 1–30, 2016.

[18] Joel N. Franklin. Well-posed stochastic extensions of ill-posed linear problems. *J. Math. Anal. Appl.*, 31(3):682–716, 1970.

[19] Arnaud Gloter, Marc Hoffmann, et al. Estimation of the Hurst parameter from discrete noisy data. *Ann. Stat.*, 35(5):1947–1974, 2007.

[20] Shota Gugushvili, Aad W. van der Vaart, and Dong Yan. Bayesian inverse problems with partial observations. *Trans. A. Razmadze Math. Inst.*, 172(3):388–403, 2018.

[21] Shota Gugushvili, Aad W. van der Vaart, and Dong Yan. Bayesian linear inverse problems in regularity scales. `https://arxiv.org/abs/1802.08992`, 2018.

[22] Tapio Helin and Martin Burger. Maximum a posteriori probability estimates in infinite-dimensional Bayesian inverse problems. *Inverse Probl.*, 31(8), 2015.

[23] Tapio Helin and Matti Lassas. Hierarchical models in statistical inverse problems and the Mumford–Shah functional. *Inverse Probl.*, 27(1), 2010.

[24] Jari Kaipio and Erkki Somersalo. *Statistical and Computational Inverse Problems*, volume 160. Springer, 2006.

[25] Ustim Khristenko, Laura Scarabosio, Piotr Swierczynski, Elisabeth Ullmann, and Barbara Wohlmuth. Analysis of boundary effects on PDE-based sampling of Whittle-Matérn random fields. *SIAM/ASA J. Uncertain. Quantif.*, 7(3):948–974, 2019.

[26] Bartek T Knapik, Botond T. Szabó, Aad W. van der Vaart, and J. Harry van Zanten. Bayes procedures for adaptive inference in inverse problems for the white noise model. *Probab. Theory Relat. Fields*, 164(3-4):771–813, 2016.

[27] Bartek T Knapik, Aad W. van der Vaart, and J. Harry van Zanten. Bayesian recovery of the initial condition for the heat equation. *Commun. Stat., Theory Methods*, 42(7):1294–1313, 2013.

[28] Bartek T Knapik, Aad W. van der Vaart, J. Harry van Zanten, et al. Bayesian inverse problems with Gaussian priors. *Ann. Stat.*, 39(5):2626–2657, 2011.

[29] Sari Lasanen. Non-Gaussian statistical inverse problems. Part I: Posterior distributions. *Inverse Probl. Imaging*, 6(2):215–266, 2012.

[30] Sari Lasanen. Non-Gaussian statistical inverse problems. Part II: Posterior convergence for approximated unknowns. *Inverse Probl. Imaging*, 6(2):267–287, 2012.

[31] Markku S. Lehtinen, Lassi Paivarinta, and Erkki Somersalo. Linear inverse problems for generalised random variables. *Inverse Probl.*, 5(4):599–612, 1989.

[32] Finn Lindgren, Håvard Rue, and Johan Lindström. An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. *J. R. Stat. Soc., Ser. B, Stat. Methodol.*, 73(4):423–498, 2011.

[33] Kevin P Murphy. *Machine Learning: A Probabilistic Perspective*. MIT Press, 2012.

[34] Radford M. Neal. *Bayesian Learning for Neural Networks*. PhD thesis, University of Toronto, 1995.

[35] Radford M. Neal. Monte Carlo implementation of Gaussian process models for Bayesian regression and classification. `https://arxiv.org/abs/physics/9701026`, 1997.

[36] Richard Nickl. Bernstein-von Mises theorems for statistical inverse problems I: Schrödinger equation. `https://arxiv.org/abs/1707.01764`, 2017.

[37] Richard Nickl and Kolyan Ray. Nonparametric statistical inference for drift vector fields of multi-dimensional diffusions. `https://arxiv.org/abs/1810.01702`, 2018.

[38] Richard Nickl and Jakob Söhl. Bernstein-von Mises theorems for statistical inverse problems II: compound Poisson processes. *Electron. J. Stat.*, 13(2):3513–3571, 2019.

[39] Richard Nickl, Sara van de Geer, and Sven Wang. Convergence rates for Penalised Least Squares Estimators in PDE-constrained regression problems. `https://arxiv.org/abs/1809.08818`, 2018.

[40] Houman Owhadi, Clint Scovel, and Tim Sullivan. On the brittleness of Bayesian inference. *SIAM Rev.*, 57(4):566–582, 2015.

[41] Omiros Papaspiliopoulos, Gareth Roberts, and Martin Sköld. A general framework for the parametrization of hierarchical models. *Stat. Sci.*, pages 59–73, 2007.

[42] Kolyan Ray. Bayesian inverse problems with non-conjugate priors. *Electron. J. Stat.*, 7:2516–2549, 2013.

[43] Gareth Roberts and Osnat Stramer. On inference for partially observed nonlinear diffusion models using the Metropolis–Hastings algorithm. *Biometrika*, 88(3):603–621, 2001.

[44] Lassi Roininen, Janne M. J. Huttunen, and Sari Lasanen. Whittle-Matérn priors for Bayesian statistical inversion with applications in electrical impedance tomography. *Inverse Probl. Imaging*, 8(2):561–586, 2014.

[45] Andrew M. Stuart. Inverse problems: a Bayesian perspective. In *Acta Numerica*, volume 19, pages 451–559. Cambridge University Press, 2010.

[46] Aad W. van der Vaart and Jon A. Wellner. Weak convergence. In *Weak convergence and empirical processes*, pages 16–28. Springer, 1996.

[47] J. Harry van Zanten. A Note on Consistent Estimation of Multivariate Parameters in Ergodic Diffusion Models. *Scand. J. Stat.*, 28(4):617–623, 2001.

[48] Yaming Yu and Xiao-Li Meng. To center or not to center: that is not the question – an Ancillarity–Sufficiency Interweaving Strategy (ASIS) for boosting MCMC efficiency. *J. Comput. Graph. Stat.*, 20(3):531–570, 2011.