Journal of Computational Physics 229 (2010) 7317–7343

Contents lists available at ScienceDirect

Journal of Computational Physics

journal homepage: www.elsevier.com/locate/jcp

# Hierarchical Bayesian inference for Ill-posed problems via variational method

Bangti Jin [a,*], Jun Zou [b]

[a] Center for Industrial Mathematics, University of Bremen, D-28334 Bremen, Germany
[b] Department of Mathematics, The Chinese University of Hong Kong, Shatin, N.T., Hong Kong, PR China

ARTICLE INFO

*Article history:*
Received 6 June 2009
Received in revised form 30 May 2010
Accepted 9 June 2010
Available online 17 June 2010

*Keywords:*
Variational method
Hierarchical Bayesian inference
Inverse problem
Uncertainty quantification
Cauchy problem

ABSTRACT

This paper investigates a novel approximate Bayesian inference procedure for numerically solving inverse problems. A hierarchical formulation which determines automatically the regularization parameter and the noise level together with the inverse solution is adopted. The framework is of variational type, and it can deliver the inverse solution and regularization parameter together with their uncertainties calibrated. It approximates the posteriori probability distribution by separable distributions based on Kullback–Leibler divergence. Two approximations are derived within the framework, and some theoretical properties, e.g. variance estimate and consistency, are also provided. Algorithms for their efficient numerical realization are described, and their convergence properties are also discussed. Extensions to nonquadratic regularization/nonlinear forward models are also briefly studied. Numerical results for linear and nonlinear Cauchy-type problems arising in heat conduction with both smooth and nonsmooth solutions are presented for the proposed method, and compared with that by Markov chain Monte Carlo. The results illustrate that the variational method can faithfully capture the posteriori distribution in a computationally efficient way.



## 1. Introduction

In this paper, we are interested in a novel numerical method of Bayesian type for solving inverse problems, especially those related to heat conduction. Inverse problems arise in many disciplines, such as heat conduction [1], mechanics and geophysics, and play an important role in revealing the underlying physical mechanisms. Typically, inverse problems are ill-posed in the sense that the solution lacks a stable dependence on the data. Therefore, their stable and accurate numerical solutions are very challenging. One of the most popular approaches is Tikhonov regularization, which solves a nearby well-posed problem and takes its solution as an approximation. Iterative type methods, such as Landweber method and conjugate gradient method, equipped with a suitable stopping criterion can also be applied.

Bayesian inference approach provides another principled and flexible framework for inverse problems, and has distinct features over classical deterministic regularization methods. Firstly, it yields an ensemble of inverse solutions consistent with the given data, and thus it enables uncertainty quantification of a specific solution. This contrasts sharply with above-mentioned deterministic inverse techniques that content with singling out one solution out of the ensemble. Secondly, it provides a flexible regularization in that the difficult problem of choosing a regularization parameter is resolved through hierarchical modeling. Therefore, it has attracted considerable attention in a wide variety of applied disciplines,

* Corresponding author. Tel.: +49 421 218 63807.
*E-mail addresses:* kimbtsing@yahoo.com.cn (B. Jin), zou@math.cuhk.edu.hk (J. Zou).

0021-9991/$ - see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.jcp.2010.06.016

e.g. geophysics [2,3], image processing [4] and transient heat conduction [5–7]. For a comprehensive overview of methodological developments, we refer to monographs [2,4].

Hierarchical Bayesian inference has been applied to inverse heat conduction problems [5–8]. The numerical results presented in these studies are very encouraging in that the regularization parameter, noise level and inverse solution can be simultaneously estimated with their uncertainties calibrated. Despite the popularity of hierarchical Bayesian formulations in practical applications and demonstrated performances, the choice of the prior parameter pairs for the hyper-parameters was carried out in a rather ad hoc manner in existing studies. It remains unclear why these formulations work in practice, and no guidelines for their choice were available. Also the Bayesian solution, i.e., posterior probability density function (PPDF) is often numerically sampled, e.g. by Markov chain Monte Carlo (MCMC). However, the MCMC can be computationally expensive, and its convergence might be not easy to diagnose. To circumvent the computational problem, the authors [8] proposed considering the joint *maximum a posteriori* (MAP), and derived an augmented Tikhonov (a-Tikhonov for short) functional that determines the regularization parameter and the noise level along with the solution. Recently some mathematical underpinnings were also provided [9]. However, it yields only one solution like above-mentioned deterministic inverse techniques and does not calibrate the associated uncertainties, and thus it is not completely satisfactory from the point of view of Bayesian analysis.

This paper investigates an alternative framework based on the variational method. The new approach can quantify the uncertainties of the computed solution, thereby overcoming the drawback of the a-Tikhonov method. The approach was first developed in machine learning community [10–12], however, its application to inverse problems seems largely unexplored. This paper will offer some new theoretical results, e.g. its properties in the context of classical inverse theory and convergence properties of the algorithms, to shed some lights on the practical performance. Analyzing the properties of these approximations also provide one means to interrogate the properties of hierarchical formulations. Some heuristic guidelines for the choice of prior parameter pairs in hierarchical Bayesian formulations will be derived, and thus the study sheds new insights on hierarchical Bayesian formulations. The approach is generally applicable to both linear and nonlinear inverse problems with suitable extensions. We shall examine its applicability on severely ill-posed linear and nonlinear Cauchy problems, and carry out a detailed comparison of the new method with the true PPDF explored by the MCMC.

The rest of the paper is structured as follows. Fundamentals of Bayesian inference, hierarchical modeling and associated computational challenge are recalled in Section 2. The variational method for linear inverse problems is described in Section 3, two approximations of the PPDF are derived, and their theoretical properties are analyzed. Algorithms for computing the approximations together with their convergence properties are also discussed. Two generalizations, i.e., $\ell^r$ prior and nonlinear forward models, are briefly discussed in Section 4. Numerical results for the Cauchy-type problems with smooth and nonsmooth solutions to illustrate their features are presented in Section 5, and compared with that by the MCMC. Finally, we conclude the paper with Section 6.

## 2. Bayesian inference approach

This section describes the Bayesian framework for a finite-dimensional linear inverse problem

$$\mathbf{Hm} = \mathbf{d}, \tag{1}$$

where $\mathbf{H} \in \mathbb{R}^{n\times m}$, $\mathbf{m} \in \mathbb{R}^m$ and $\mathbf{d} \in \mathbb{R}^n$ represent system matrix, sought-for solution and given data, respectively. We shall denote $\mathbf{d}^\dagger$ the noise-free data, and assume that $\mathbf{d} = \mathbf{d}^\dagger + \boldsymbol{\omega}$ with $\boldsymbol{\omega}$ being a random vector with mean zero and variance $\sigma_0^2\mathbf{I}$. We shall focus on hyper-parameter treatment within hierarchical models and the associated computational challenge of exploring the posterior state space.

The primary goal of Bayesian inference is to deduce the distribution of the unknown parameters $\mathbf{m}$ conditioned on the data $\mathbf{d}$, i.e., the PPDF $p(\mathbf{m}|\mathbf{d})$. According to Bayes' rule, it is related to $\mathbf{d}$ by

$$p(\mathbf{m}|\mathbf{d}) = \frac{p(\mathbf{d}|\mathbf{m})p(\mathbf{m})}{\int p(\mathbf{d}|\mathbf{m})p(\mathbf{m})d\mathbf{m}}.$$

The functions $p(\mathbf{d}|\mathbf{m})$ and $p(\mathbf{m})$ are known as likelihood function and prior probability density, respectively, and they are two basic building blocks of Bayesian inference. Intuitively, it provides a mechanism to integrate the prior knowledge $p(\mathbf{m})$ with the information contained in the data $p(\mathbf{d}|\mathbf{m})$ to achieve the current state of knowledge, the PPDF $p(\mathbf{m}|\mathbf{d})$. The normalizing constant $\int p(\mathbf{d} \mid \mathbf{m})p(\mathbf{m})d\mathbf{m}$ is needed for estimating the credible interval [13], however its computation can be highly nontrivial, especially in high-dimensions. Fortunately, it is often unnecessary to compute the normalizing constant, e.g. the MCMC and optimization, and the PPDF $p(\mathbf{m}|\mathbf{d})$ may be simply evaluated as

$$p(\mathbf{m}|\mathbf{d}) \propto p(\mathbf{d}|\mathbf{m})p(\mathbf{m}). \tag{2}$$

The PPDF $p(\mathbf{m}|\mathbf{d})$ constitutes a complete description of the inverse problem, and it contains all the information available about $\mathbf{m}$. However, it is not directly informative, and various summarizing statistics, e.g. point estimates and credible intervals, have to be computed. Typical point estimates include posterior mean $\hat{\mathbf{m}}_{\mathrm{pm}}$ and MAP $\hat{\mathbf{m}}_{\mathrm{map}}$. However, we caution that point estimates may not be representative of the PPDF [5,6].

Regarding the two building blocks of a generic Bayesian model, the likelihood is usually straightforward to obtain. A simple model assumes that i.i.d. additive Gaussian random errors with mean zero and variance $\sigma^2 = \tau^{-1}$ account for the measurement noise. Then the likelihood $p(\mathbf{d}|\mathbf{m})$ is given by

$$p(\mathbf{d}|\mathbf{m}) \propto \tau^{\frac{n}{2}} e^{-\frac{\tau}{2}\|\mathbf{H}\mathbf{m}-\mathbf{d}\|_2^2}. \tag{3}$$

Throughout this paper, we shall reserve the notation $\sigma^2$ for noise variance, and denote the variance of any other quantities by var ().

The prior $p(\mathbf{m})$ encodes the knowledge of the unknown $\mathbf{m}$ before collecting the data, and thus it stays at the heart of any modeling tasks. A versatile tool for prior modeling is Markov random field (MRF). In the present study, the following simple MRF is adopted

$$p(\mathbf{m}) \propto \lambda^{\frac{m}{2}} e^{-\frac{\lambda}{2}\|\mathbf{L}\mathbf{m}\|_2^2}, \tag{4}$$

where the matrix $\mathbf{L}$ encapsulates the structure of the interactions between neighboring sites, and $\lambda$ is a scaling parameter dictating the strength of interaction. We shall assume $\ker \mathbf{H} \cap \ker \mathbf{L} = \{\mathbf{0}\}$ and the rank of $\mathbf{L}$ is $m$ for simplicity.

If the parameters $\tau$ and $\lambda$ are known, with likelihood (3) and prior model (4), the PPDF $p(\mathbf{m}|\mathbf{d})$ (2) can be evaluated as

$$p(\mathbf{m}|\mathbf{d}) \propto e^{-\frac{\tau}{2}\|\mathbf{H}\mathbf{m}-\mathbf{d}\|_2^2} e^{-\frac{\lambda}{2}\|\mathbf{L}\mathbf{m}\|_2^2}. \tag{5}$$

In Bayesian analysis, it is customary to select the MAP $\hat{\mathbf{m}}_{\text{map}} := \arg\max_{\mathbf{m}} p(\mathbf{m} \mid \mathbf{d})$ as the inverse solution. The MAP $\hat{\mathbf{m}}_{\text{map}}$ is easily derived as

$$\hat{\mathbf{m}}_{\text{map}} = \arg\min_{\mathbf{m}} \left\{ \mathcal{J}_\eta(\mathbf{m}) := \|\mathbf{H}\mathbf{m} - \mathbf{d}\|_2^2 + \eta\|\mathbf{L}\mathbf{m}\|_2^2 \right\},$$

which is exactly Tikhonov regularization with a regularization parameter $\eta = \lambda\tau^{-1}$, and the unique minimizer will be denoted by $\mathbf{m}_\eta$. Therefore, the prior $p(\mathbf{m})$ regularizes the inverse problem and $\lambda\tau^{-1}$ assumes the crucial role of a regularization parameter. However, the selection of $\eta$ is notoriously nontrivial in almost all inverse techniques. The parameters $\lambda$ and $\tau$ are still important in Bayesian inference due to their substantial effects on the PPDF and thus posterior point estimates.

The Bayesian paradigm proposes to resolve the issue flexibly through hierarchical modeling. The idea is to let the data $\mathbf{d}$ determine these parameters in the hope of effectively diminishing the effect of the initial (prior) assumptions of their values on the inverse solution. More precisely, the hyper-parameters, i.e., $\lambda$ and $\tau$, are also modeled as random variables and have their own priors. Then the PPDF (2) is augmented to determine the scaling parameter $\lambda$ and to detect the noise level $\tau$ as follows:

$$p(\mathbf{m}, \lambda, \tau|\mathbf{d}) \propto p(\mathbf{d}|\mathbf{m}, \tau)p(\mathbf{m}|\lambda)p(\lambda)p(\tau). \tag{6}$$

A standard practice to select priors for hyper-parameters, also known as hyper-priors, is to use conjugate priors, which enables combining neatly with the likelihood to facilitate subsequent mathematical manipulations of the PPDF while remains sufficiently flexible. For both $\lambda$ and $\tau$, the conjugate prior is a Gamma distribution $G(t;\alpha,\beta)$, which is defined by

$$G(t;\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} t^{\alpha-1} e^{-\beta t}, \tag{7}$$

where $\Gamma(\cdot)$ is the standard Gamma function, and $\alpha$ and $\beta$ are nonnegative constants. The mean, standard deviation and mode of $G(t;\alpha,\beta)$ are given by $\frac{\alpha}{\beta}$, $\frac{\sqrt{\alpha}}{\beta}$ and $\frac{\alpha-1}{\beta}$, respectively. In some real-world applications, nonconjugate priors can also be very useful and may affect the subsequent derivations, however, we will restrict our attention to conjugate priors. Upon adopting conjugate priors for both $\lambda$ and $\tau$, the PPDF (6) reads

$$p(\mathbf{m}, \lambda, \tau|\mathbf{d}) \propto \tau^{\frac{n}{2}} e^{-\frac{\tau}{2}\|\mathbf{H}\mathbf{m}-\mathbf{d}\|_2^2} \cdot \lambda^{\frac{m}{2}} e^{-\frac{\lambda}{2}\|\mathbf{L}\mathbf{m}\|_2^2} \cdot \lambda^{\alpha_0-1} e^{-\beta_0\lambda} \cdot \tau^{\alpha_1-1} e^{-\beta_1\tau}, \tag{8}$$

where $(\alpha_0,\beta_0)$ and $(\alpha_1,\beta_1)$ are the parameter pairs of the Gamma distribution for $\lambda$ and $\tau$, respectively.

Due to the presence of the hyper-parameters $\lambda$ and $\tau$, the PPDF (8) is nonstandard and implicit as opposed to the PPDF (5). The posterior state space is often of high dimensionality, and thus it can only be numerically explored. Among various numerical sampling techniques, the MCMC remains the most popular one [14,15]. Unfortunately, a large number of samples, e.g. $10^5 - 10^6$, are often required for obtaining reliable estimates, especially for the variance of the inverse solution, and thus a faithful exploration of the PPDF is expensive, and moreover the convergence of the Markov chain maybe not easy to diagnose.

To the best of authors' knowledge, few attempts besides numerical exploration via the MCMC have been made to simultaneously estimate of the hyper-parameters and the solution together with their uncertainties, and often only point estimates have been considered. One typical approach of the latter category is the a-Tikhonov method [9]. It considers the MAP of the PPDF (8), which amounts to minimizing the functional $\mathcal{J}(\mathbf{m},\lambda,\tau)$ defined by

$$\mathcal{J}(\mathbf{m},\lambda,\tau) = \frac{\tau}{2}\|\mathbf{H}\mathbf{m} - \mathbf{d}\|_2^2 + \frac{\lambda}{2}\|\mathbf{L}\mathbf{m}\|_2^2 + \beta_0\lambda - \alpha_0' \ln\lambda + \beta_1\tau - \alpha_1' \ln\tau, \tag{9}$$

where $\alpha_0' = \frac{m}{2} + \alpha_0 - 1$ and $\alpha_1' = \frac{n}{2} + \alpha_1 - 1$. One salient feature of the functional is that it determines regularization parameter and noise level simultaneously with the inverse solution. However, it aims at a single solution and not at the PPDF that

can be explored to gain additional information about the point estimate. Instead of retaining all possible values of the parameters and the solution, it chooses one specific set of values, thereby neglecting many other interpretations of the data. Intuitively, this can be problematic: for sharply peaked PPDFs, other values will have much lower posterior probabilities, but for broad PPDFs, choosing one single value will neglect many other equally plausible interpretations.

## 3. Variational principle

In this section, we describe the variational method for computing an approximation to the PPDF (8). It can deliver point estimates together with uncertainties for both the solution and hyper-parameters, and thus overcomes some drawbacks of point estimate-oriented methods, e.g. the a-Tikhonov method. The derivations will also shed new lights on hierarchical Bayesian formulations.

The basic idea is to approximate the PPDF $p(\mathbf{m},\lambda,\tau|\mathbf{d})$ by a "simpler" distribution using Kullback–Leibler divergence while hopefully retaining its main features. It has attracted considerable interests during the last few years, and the spectrum of its applications includes graphical models [11], MEG inverse problem [16] and image processing [17]. For an account of its application to signal processing, we refer to the monograph [18], and references therein for further literature. This section focuses on two approximations given by the framework.

Methodologically speaking, the fundamental idea of the variational method consists of first transforming the problem into an equivalent optimization problem and obtaining then an approximation by solving the optimization problem inexactly, in a manner analogous to the classical finite-element method. To this end, we need to choose the distance metric to derive an equivalent optimization problem and the type of assumptions to deliver computationally tractable approximations. Hence there are two essential ingredients of the variational method: metric and simplifying assumption. The Bayesian solution is a probability distribution (density function), and thus the equivalent transformation calls for probability metric. In principle, any valid probability metric, e.g. Kullback–Leibler divergence, Hellinger distance and Wasserstein distance, see Ref. [19] for an overview of some most commonly used probability metrics, can be adopted to measure the distance between an approximate PPDF and the true PPDF. While most of these distances do not directly lead to computationally tractable/efficient numerical algorithms, the Kullback–Leibler divergence does, as we shall see in the remaining part of this section. We review briefly the Kullback–Leibler divergence below, and for a detailed account of its mathematical properties, we refer to Ref. [20]. The Kullback–Leibler divergence $D_{KL}(q(\mathbf{m},\lambda,\tau)|p(\mathbf{m},\lambda,\tau|\mathbf{d}))$ between two probability density functions $q(\mathbf{m},\lambda,\tau)$ and $p(\mathbf{m},\lambda,\tau|\mathbf{d})$ is defined by

$$\begin{aligned} D_{KL}(q(\mathbf{m},\lambda,\tau)|p(\mathbf{m},\lambda,\tau|\mathbf{d})) &= \int\int\int q(\mathbf{m},\lambda,\tau)\ln\frac{q(\mathbf{m},\lambda,\tau)}{p(\mathbf{m},\lambda,\tau|\mathbf{d})}d\mathbf{m}d\lambda d\tau \\ &= \int\int\int q(\mathbf{m},\lambda,\tau)\ln\frac{q(\mathbf{m},\lambda,\tau)}{p(\mathbf{m},\lambda,\tau,\mathbf{d})}d\mathbf{m}d\lambda d\tau + \log p(\mathbf{d}), \end{aligned}$$

where $p(\mathbf{d}) = \int\int\int p(\mathbf{m},\lambda,\tau,\mathbf{d})d\mathbf{m}d\lambda d\tau$ is a normalizing constant. Recall Jensen's inequality

$$E[\varphi(Y)] \geqslant \varphi(E[Y]), \tag{10}$$

where the notation $E$ denotes taking expectation with respect to a certain probability measure, $Y$ is a random variable and $\varphi$ is a convex function. Since $q(\mathbf{m},\lambda,\tau)$ is a probability density function, we can apply inequality (10) with $\varphi(x) = -\ln(x)$ and $Y = \frac{p(\mathbf{m},\lambda,\tau|\mathbf{d})}{q(\mathbf{m},\lambda,\tau)}$ to deduce

$$\begin{aligned} D_{KL}(q(\mathbf{m},\lambda,\tau)|p(\mathbf{m},\lambda,\tau|\mathbf{d})) &= E[\varphi(Y)] \geqslant \varphi(E[Y]) \\ &= -\ln\int\int\int q(\mathbf{m},\lambda,\tau)\cdot\frac{p(\mathbf{m},\lambda,\tau|\mathbf{d})}{q(\mathbf{m},\lambda,\tau)}d\mathbf{m}d\lambda d\tau \\ &= -\ln\int\int\int p(\mathbf{m},\lambda,\tau|\mathbf{d})d\mathbf{m}d\lambda d\tau = -\ln 1 = 0. \end{aligned}$$

That is, the divergence $D_{KL}$ is nonnegative and vanishes if and only if $q$ coincides with $p$, and measures the 'proximity' between two distributions. Therefore we have effectively transformed the problem into an equivalent optimization problem by considering minimizing the divergence $D_{KL}$. If there is no restriction on the trial distribution $q(\mathbf{m},\lambda,\tau)$, minimizing the divergence $D_{KL}$ would reproduce the PPDF $p(\mathbf{m},\lambda,\tau|\mathbf{d})$. However, the PPDF $p(\mathbf{m},\lambda,\tau|\mathbf{d})$ is not available in closed form for $p(\mathbf{d})$ cannot be calculated analytically. Nonetheless, it is a fixed constant for a given set of data $\mathbf{d}$, and thus irrelevant in the optimization procedure. We will suppress the conditional notation in our subsequent expressions, and choose to minimize the following functional, which is also denoted by $D_{KL}$

$$D_{KL}(q(\mathbf{m},\lambda,\tau)|p(\mathbf{m},\lambda,\tau)) = \int\int\int q(\mathbf{m},\lambda,\tau)\ln\frac{q(\mathbf{m},\lambda,\tau)}{p(\mathbf{m},\lambda,\tau)}d\mathbf{m}d\lambda d\tau. \tag{11}$$

Unfortunately, this optimization problem without further assumptions is also computationally intractable to solve exactly. Nonetheless, an approximation can be obtained by solving the optimization problem inexactly under certain simplifying assumptions. This can be achieved by restricting the space of admissible solutions, in the same spirit of using a finite-dimensional finite element space instead of a computationally intractable infinite-dimensional function space in the finite

element method. This is the other essential ingredient of the variational method. To this end, let us examine more closely the origin of difficulty. The difficulty mainly stems from the strong interactions/coupling between $\mathbf{m}$ and $(\lambda,\tau)$, i.e., conditional dependence in probability terms. Therefore, conditional independence emerges as the key ingredient in developing approximations in the probability world. Consequently, the variational method adopts a separable, i.e., conditionally independent given the data $\mathbf{d}$, approximation for the posterior distributions of $\mathbf{m}$ and $(\lambda,\tau)$. The idea is closely related to the mean field theory of statistical mechanics for treating many-body systems. Therefore, we seek an approximation $q(\mathbf{m},\lambda,\tau)$ of factorized form

$$q(\mathbf{m},\lambda,\tau) = q(\mathbf{m})q(\lambda,\tau). \tag{12}$$

This is often the only assumption invoked on the "simpler" approximation $q(\mathbf{m},\lambda,\tau)$ to render the optimization problem (11) analytically and computationally tractable, and the resultant is termed as `Nondegenerate Approximation` (`Approx I` for short). Extra assumptions, e.g. degeneracy, on the approximations may be imposed to further reduce its complexity. We will investigate the case of a degenerate $q(\mathbf{m})$, and term it as `Degenerate Approximation` (`Approx II` for short).

### 3.1. Nondegenerate approximation

`Approx I` approximates the PPDF with a factorized distribution $q(\mathbf{m},\lambda,\tau) = q(\mathbf{m})q(\lambda,\tau)$. We have the following existence result, which shows that the optimization problem is well-defined.

**Theorem 3.1.** *There exists at least one minimizer to the optimization problem (11).*

**Proof.** By Jensen's inequality, the functional $D_{KL}$ is bounded from below, and thus there exists a minimizing sequence $q^n(\mathbf{m},\lambda,\tau) \equiv q^n(\mathbf{m})q^n(\lambda,\tau)$. Note that the divergence $D_{KL}$ is weakly lower semi-continuous on $L^1$ and that for any $C > 0$, the sublevel sets $\{q^n(\mathbf{m},\lambda,\tau) \in L^1 : D_{KL}(q^n(\mathbf{m},\lambda,\tau)|p(\mathbf{m},\lambda,\tau)) \leqslant C\}$ are weakly compact in $L^1$, see Lemmas 2.2 and 2.3 of Ref. [21]. Therefore, there exists at least one minimizer $q^*(\mathbf{m},\lambda,\tau) \equiv q^*(\mathbf{m})q^*(\lambda,\tau)$ to the optimization problem (11). □

The rest of this section analyzes the properties of the minimizers by examining the optimality system and describes an algorithm for computing the minimizer.

#### 3.1.1. Derivation of optimality conditions

We shall attempt to analyze the properties of the minimizers to gain understanding of the variational method and hierarchical formulations. To this end, we consider its optimality system. To enforce the normalization condition of the densities $q(\mathbf{m})$ and $q(\lambda,\tau)$, i.e., $\int q(\mathbf{m})d\mathbf{m} = 1$ and $\int\int q(\lambda,\tau)d\lambda d\tau = 1$, we introduce the Lagrange function $\mathcal{L}(q(\mathbf{m}),q(\lambda,\tau),\varrho_1,\varrho_2)$ of the divergence $D_{KL}$ defined as follows:

$$\mathcal{L}(q(\mathbf{m}),q(\lambda,\tau),\varrho_1,\varrho_2) = D_{KL}(q(\mathbf{m})q(\lambda,\tau)|p(\mathbf{m},\lambda,\tau)) + \varrho_1\left(\int q(\mathbf{m})d\mathbf{m} - 1\right) + \varrho_2\left(\int\int q(\lambda,\tau)d\lambda d\tau - 1\right),$$

where $\varrho_i$ are Lagrange multipliers associated with the normalizing conditions. Taking the variational derivative of the Lagrange function $\mathcal{L}(q(\mathbf{m}),q(\lambda,\tau),\varrho_1,\varrho_2)$ with respect to $q(\mathbf{m})$ and equating it to zero yields

$$\begin{aligned}\frac{\partial}{\partial q(\mathbf{m})}\mathcal{L}(q(\mathbf{m}),q(\lambda,\tau),\varrho_1,\varrho_2) &= \frac{\partial}{\partial q(\mathbf{m})}D_{KL}(q(\mathbf{m})q(\lambda,\tau)|p(\mathbf{m},\lambda,\tau)) + \varrho_1\\ &= \int\int\left[\frac{\partial}{\partial q(\mathbf{m})}\int q(\mathbf{m})\ln\frac{q(\mathbf{m})q(\lambda,\tau)}{p(\mathbf{m},\lambda,\tau)}d\mathbf{m}\right]q(\lambda,\tau)d\lambda d\tau + \varrho_1\\ &= \int\int[-\ln p(\mathbf{m},\lambda,\tau) + \ln q(\lambda,\tau) - 1 + \ln q(\mathbf{m})]q(\lambda,\tau)d\lambda d\tau + \varrho_1 = 0,\end{aligned}$$

where $\varrho_1$ is determined according to the optimality condition $\frac{\partial}{\partial\varrho_1}\mathcal{L}(q(\mathbf{m}),q(\lambda,\tau),\varrho_1,\varrho_2) = 0$, i.e., the normalization condition $\int q(\mathbf{m})d\mathbf{m} = 1$. Rearranging the above equation shows that at a critical point $q^*(\mathbf{m})q^*(\lambda,\tau)$ of the divergence $D_{KL}$, there holds

$$\ln q^*(\mathbf{m}) = \int\int \ln p(\mathbf{m},\lambda,\tau)q^*(\lambda,\tau)d\lambda d\tau - \ln Z_{q^*(\mathbf{m})} = E_{q^*(\lambda,\tau)}[\ln p(\mathbf{m},\lambda,\tau)] - \ln Z_{q^*(\mathbf{m})},$$

where the constant $\ln Z_{q^*(\mathbf{m})} = \varrho_1 - 1 + \int\int q^*(\lambda,\tau)\ln q^*(\lambda,\tau)d\lambda d\tau$. Here the notation $E_q$ denotes taking expectation with respect to a probability density function $q$. By substituting the formula for $p(\mathbf{m},\lambda,\tau)$, c.f. (8), into this equation, we obtain

$$\begin{aligned}\ln q^*(\mathbf{m}) &= E_{q^*(\lambda,\tau)}[\ln p(\mathbf{m},\lambda,\tau)] - \ln Z_{q^*(\mathbf{m})}\\ &= E_{q^*(\lambda,\tau)}\left[-\frac{\tau}{2}\|\mathbf{Hm}-\mathbf{d}\|_2^2 - \frac{\lambda}{2}\|\mathbf{Lm}\|_2^2 + T(\lambda,\tau)\right] - \ln Z_{q^*(\mathbf{m})},\\ &= -\frac{E_{q^*(\lambda,\tau)}[\tau]}{2}\|\mathbf{Hm}-\mathbf{d}\|_2^2 - \frac{E_{q^*(\lambda,\tau)}[\lambda]}{2}\|\mathbf{Lm}\|_2^2 + E_{q^*(\lambda,\tau)}[T(\lambda,\tau)] - \ln Z_{q^*(\mathbf{m})},\end{aligned}$$

where $T(\lambda,\tau)$ contains terms in $\ln p(\mathbf{m},\lambda,\tau)$ not involving $\mathbf{m}$. Since the last two terms are independent of $\mathbf{m}$ and thus contribute only to the normalizing condition and the first two terms are quadratic in $\mathbf{m}$, we deduce that $q^*(\mathbf{m})$ follows a Gaussian distribution. To simplify the expression, let the scalars $\lambda^*$, $\tau^*$ and $\eta^*$ be defined as $\lambda^* = E_{q^*(\lambda,\tau)}[\lambda]$ and $\tau^* = E_{q^*(\lambda,\tau)[\tau]}$, and $\eta^* = \frac{\lambda^*}{\tau^*}$, respectively. Now observe the elementary identity which follows by expanding the quadratic function at $\mathbf{m}^* = (\mathbf{H}^T\mathbf{H} + \eta^*\mathbf{L}^T\mathbf{L})^{-1}\mathbf{H}^T\mathbf{d}$

$$-\frac{\tau^*}{2}\|\mathbf{Hm}-\mathbf{d}\|_2^2 - \frac{\lambda^*}{2}\|\mathbf{Lm}\|_2^2 = -\frac{1}{2}(\mathbf{m}-\mathbf{m}^*)^T\left[\tau^*\mathbf{H}^T\mathbf{H} + \lambda^*\mathbf{L}^T\mathbf{L}\right](\mathbf{m}-\mathbf{m}^*) - \frac{\tau^*}{2}\|\mathbf{Hm}^*-\mathbf{d}\|_2^2 - \frac{\lambda^*}{2}\|\mathbf{Lm}^*\|_2^2.$$

Observe that $\mathbf{m}^* = \mathbf{m}_{\eta^*}$, the minimizer of the Tikhonov functional $\mathcal{J}_{\eta^*}$, and also the last two terms contribute only to the normalizing constant. Therefore, we deduce

$$q^*(\mathbf{m}) = \mathcal{N}(\mathbf{m}^*, (\tau^*\mathbf{H}^T\mathbf{H} + \lambda^*\mathbf{L}^T\mathbf{L})^{-1}),$$

where $\mathcal{N}$ denotes a normal distribution.

Next we derive the remaining part of the optimality system. Analogously, taking the variational derivative of the Lagrange function $\mathcal{L}(q(\mathbf{m}), q(\lambda,\tau), \varrho_1, \varrho_2)$ with respect to $q(\lambda,\tau)$ and equating it to zero yields

$$\begin{aligned}\frac{\partial}{\partial q(\lambda,\tau)}\mathcal{L}(q(\mathbf{m}), q(\lambda,\tau), \varrho_1, \varrho_2) &= \frac{\partial}{\partial q(\lambda,\tau)} D_{KL}(q(\mathbf{m})q(\lambda,\tau)|p(\mathbf{m},\lambda,\tau)) + \varrho_2 \\ &= \int\left[\frac{\partial}{\partial q(\lambda,\tau)}\int\int q(\lambda,\tau)\ln\frac{q(\mathbf{m})q(\lambda,\tau)}{p(\mathbf{m},\lambda,\tau)}d\lambda d\tau\right]q(\mathbf{m})d\mathbf{m} + \varrho_2 \\ &= \int[-\ln p(\mathbf{m},\lambda,\tau) + \ln q(\mathbf{m}) - 1 + \ln q(\lambda,\tau)]q(\mathbf{m})d\mathbf{m} + \varrho_2 = 0,\end{aligned}$$

where the Lagrange multiplier $\varrho_2$ is again determined according to the optimality condition $\frac{\partial}{\partial \varrho_2}\mathcal{L}(q(\mathbf{m}), q(\lambda,\tau), \varrho_1, \varrho_2) = 0$, i.e., the normalization condition $\int\int q(\lambda,\tau)d\lambda d\tau = 1$. Consequently, we deduce that

$$\ln q^*(\lambda,\tau) = \int \ln p(\mathbf{m},\lambda,\tau)q^*(\mathbf{m})d\mathbf{m} - \ln Z_{q^*(\lambda,\tau)} = E_{q^*(\mathbf{m})}[\ln p(\mathbf{m},\lambda,\tau)] - \ln Z_{q^*(\lambda,\tau)},$$

where the constant $Z_{q^*(\lambda,\tau)} = \varrho_2 - 1 + \int q^*(\mathbf{m})\ln q^*(\mathbf{m})d\mathbf{m}$. Let the constants $\alpha_0''$ and $\alpha_1''$ be defined by $\alpha_0'' = \frac{m}{2} + \alpha_0$ and $\alpha_1'' = \frac{n}{2} + \alpha_1$. Now the expression of $p(\mathbf{m},\lambda,\tau)$ gives

$$\begin{aligned}\ln q^*(\lambda,\tau) &= E_{q^*(\mathbf{m})}[\ln p(\mathbf{m},\lambda,\tau)] - \ln Z_{q^*(\lambda,\tau)} \\ &= E_{q^*(\mathbf{m})}\left[(\alpha_1''-1)\ln\tau - \left(\frac{1}{2}\|\mathbf{Hm}-\mathbf{d}\|_2^2 + \beta_1\right)\tau\right. \\ &\quad \left. + (\alpha_0''-1)\ln\lambda - \left(\frac{1}{2}\|\mathbf{Lm}\|_2^2 + \beta_0\right)\lambda + T(\mathbf{m})\right] - \ln Z_{q^*(\lambda,\tau)} \\ &= (\alpha_1''-1)\ln\tau - \left(\frac{1}{2}E_{q^*(\mathbf{m})}\left[\|\mathbf{Hm}-\mathbf{d}\|_2^2\right] + \beta_1\right)\tau + (\alpha_0''-1)\ln\lambda - \left(\frac{1}{2}E_{q^*(\mathbf{m})}\left[\|\mathbf{Lm}\|_2^2\right] + \beta_0\right)\lambda \\ &\quad + E_{q^*(\mathbf{m})}[T(\mathbf{m})] - \ln Z_{q^*(\lambda,\tau)},\end{aligned}$$

where $T(\mathbf{m})$ contains terms in $\ln p(\mathbf{m},\lambda,\tau)$ not involving $(\lambda,\tau)$. Taking into account the normalization condition for $q^*(\lambda,\tau)$ and comparing the expression with the defining equation (7) of a Gamma distribution, we conclude that the density $q^*(\lambda,\tau)$ is separable, and can be written as $q^*(\lambda,\tau) = q^*(\lambda)q^*(\tau)$ with both $q^*(\lambda)$ and $q^*(\tau)$ following a Gamma distribution, i.e.,

$$\begin{aligned}q^*(\lambda) &= G\left(\lambda;\alpha_0'',\frac{1}{2}E_{q^*(\mathbf{m})}\left[\|\mathbf{Lm}\|_2^2\right] + \beta_0\right), \\ q^*(\tau) &= G\left(\tau;\alpha_1'',\frac{1}{2}E_{q^*(\mathbf{m})}\left[\|\mathbf{Hm}-\mathbf{d}\|_2^2\right] + \beta_1\right).\end{aligned}$$

In particular, we have $\lambda^* = E_{q^*(\lambda,\tau)}[\lambda] = E_{q^*(\lambda)}[\lambda]$ and $\tau^* = E_{q^*(\lambda,\tau)}[\tau] = E_{q^*(\tau)}[\tau]$. Note that the separability of the density function $q^*(\lambda,\tau) = q^*(\lambda)q^*(\tau)$ follows from the derivation, and is not *a priori* assumed.

In summary, we deduce that at a critical point $q^*(\mathbf{m},\lambda,\tau) = q^*(\mathbf{m})q^*(\lambda)q^*(\tau)$ there hold

$$\begin{aligned}q^*(\mathbf{m}) &= \mathcal{N}\left(\mathbf{m}^*, \left(\tau^*\mathbf{H}^T\mathbf{H} + \lambda^*\mathbf{L}^T\mathbf{L}\right)^{-1}\right), \\ q^*(\lambda) &= G\left(\lambda;\alpha_0'',\frac{1}{2}E_{q^*(\mathbf{m})}\left[\|\mathbf{Lm}\|_2^2\right] + \beta_0\right), \\ q^*(\tau) &= G\left(\tau;\alpha_1'',\frac{1}{2}E_{q^*(\mathbf{m})}\left[\|\mathbf{Hm}-\mathbf{d}\|_2^2\right] + \beta_1\right).\end{aligned} \tag{13}$$

Therefore, the approximate Bayesian solution $q(\mathbf{m})$ follows a Gaussian distribution with mean $\mathbf{m}^*$ and covariance $\mathrm{cov}_{q^*(\mathbf{m})} = (\tau^*\mathbf{H}^T\mathbf{H} + \lambda^*\mathbf{L}^T\mathbf{L})^{-1}$, and the parameters $\lambda$ and $\tau$ still follow Gamma distributions with their prior parameter pairs, i.e., $(\alpha_0,\beta_0)$ and $(\alpha_1,\beta_1)$, appropriately updated. The latter convenience is brought by the adoption of conjugate priors.

The optimality system (13) will play an important role in analyzing of the minimizers and in designing algorithms for minimizing the functional.

*3.1.2. Properties of minimizer*

To gain some understanding of and insight into hierarchical formulations, we focus on the means of the approximate PPDF $q^*(\mathbf{m})q^*(\lambda)q^*(\tau)$, i.e., $\lambda^* = E_{q^*(\lambda)}[\lambda]$, $\tau^* = E_{q^*(\tau)}[\tau]$ and $\mathbf{m}^* = E_{q^*(\mathbf{m})}[\mathbf{m}]$. The point estimates $\tau^*$ and $\lambda^*$ accept nice interpretations. Note that

$$(\tau^*)^{-1} = \frac{\frac{1}{2}E_{q^*(\mathbf{m})}\left[\|\mathbf{Hm}-\mathbf{d}\|_2^2\right]+\beta_1}{\frac{n}{2}+\alpha_1} = \varsigma_\tau\frac{\beta_1}{\alpha_1} + (1-\varsigma_\tau)\frac{E_{q^*(\mathbf{m})}\left[\|\mathbf{Hm}-\mathbf{d}\|_2^2\right]}{n},$$

where the weight $\varsigma_\tau = \frac{\alpha_1}{\alpha_1+\frac{n}{2}} \in (0,1)$. Note that the quantity $\frac{\alpha_1}{\beta_1}$ is the mean of the prior distribution for $\tau$ and constitutes the prior estimate of the inverse variance, and the term $\frac{1}{n}E_{q^*(\mathbf{m})}\left[\|\mathbf{Hm}-\mathbf{d}\|_2^2\right]$ stands for the contribution from the data $\mathbf{d}$. Therefore, the reciprocal of the mean $\tau^*$ is a convex combinations of the initial value determined by the prior parameter pair $(\alpha_1,\beta_1)$ and the estimate determined by the data $\mathbf{d}$, and the weight $\varsigma_\tau$ plays the role of a normalized confidence parameter. The weight $\varsigma_\tau$ in turn is dictated by the prior parameter $\alpha_1$: $\varsigma_\tau \approx 1$ for large $\alpha_1$, and corresponds to the case of informative prior, whereas $\varsigma_\tau \approx 0$ for $\alpha_1 \approx 1$. In the latter case, the estimate $\tau^*$ is fully determined by the data. An analogous interpretation applies to the estimate $\lambda^*$.

By computing the expectation explicitly, we derive that

$$\tau^* = \frac{\alpha_1''}{\frac{1}{2}\|\mathbf{Hm}_{\eta^*}-\mathbf{d}\|_2^2 + \frac{1}{2}\mathrm{tr}((\mathbf{H}^\mathrm{T}\mathbf{H}+\eta^*\mathbf{L}^\mathrm{T}\mathbf{L})^{-1}\mathbf{H}^\mathrm{T}\mathbf{H})\frac{1}{\tau^*}+\beta_1},$$

which after rearranging the terms and taking into account the defining equation $(\sigma^2)^* = \frac{1}{\tau^*}$ gives an estimate for the noise variance

$$(\sigma^2)^* = \frac{\frac{1}{2}\|\mathbf{Hm}_{\eta^*}-\mathbf{d}\|_2^2+\beta_1}{\alpha_1'' - \frac{1}{2}\mathrm{tr}((\mathbf{H}^\mathrm{T}\mathbf{H}+\eta^*\mathbf{L}^\mathrm{T}\mathbf{L})^{-1}\mathbf{H}^\mathrm{T}\mathbf{H})}.$$

Up to now, we have not mentioned how to the specify the prior parameter pairs $(\alpha_0,\beta_0)$ and $(\alpha_1,\beta_1)$. In practice, these parameter pairs are chosen in an *ad hoc* manner [5,6,8], and there exists no known guideline for their appropriate choices despite the popularity of hierarchical Bayesian formulations. As a consequence, we do not know how good the estimate $(\sigma^2)^*$ can be and how to choose $(\alpha_1,\beta_1)$ so that the estimate is reasonable. To explore these issues, we have to analyze the properties of the minimizer. First, we recall that one of the most popular statistical approaches for estimating noise variance, the generalized cross-validation (GCV) [22], seeks an estimate $\mathcal{V}(\eta) = \frac{\|\mathbf{Hm}_\eta-\mathbf{d}\|_2^2}{\mathcal{T}(\eta)}$, where $\eta$ minimizes the GCV function $\mathcal{V}(\eta)/\mathcal{T}(\eta)$ with $\mathcal{T}(\eta)$ defined as $\mathcal{T}(\eta) = \mathrm{tr}\left(\mathbf{I}_n - \mathbf{H}(\mathbf{H}^\mathrm{T}\mathbf{H}+\eta\mathbf{L}^\mathrm{T}\mathbf{L})^{-1}\mathbf{H}^\mathrm{T}\right)$. In the PPDF (8), if a noninformative prior is adopted for $\tau$, which is mimicked by $\alpha_1 \approx 1$ and $\beta_1 \approx 0$, then the estimate $(\sigma^2)^*$ almost coincides with the GCV estimate $\mathcal{V}(\eta)$

$$(\sigma^2)^* \approx \mathcal{V}(\eta^*)$$

by observing the identity $\mathcal{T}(\eta) = n - \mathrm{tr}\left(\left(\mathbf{H}^\mathrm{T}\mathbf{H}+\eta\mathbf{L}^\mathrm{T}\mathbf{L}\right)^{-1}\mathbf{H}^\mathrm{T}\mathbf{H}\right)$. The GCV estimate $\mathcal{V}(\eta)$ can often approximate accurately the variance $\sigma_0^2$ for $\eta$ varying over a broad scale. This was observed in many numerical experiments [22], and we refer to Ref. [9] for a preliminary justification. Under the premise that the GCV estimate $\mathcal{V}(\eta)$ is reasonable, the equality should hold reasonably. To this end, $\beta_1$ should be small relative to $\|\mathbf{Hm}_\eta-\mathbf{d}\|_2^2$, and $\alpha_1$ might take value 1.

According to classical inverse theory [23], there exists no deterministic inverse theory for parameter choice rules disrespecting the true noise level $\sigma_0$, to which the hierarchical formulation (8) belongs. In particular, the inverse solution does not necessarily converge as the true noise level $\sigma_0$ diminishes to zero. Numerically, we always observe that $(\sigma^2)^*$ approximates excellently $\sigma_0^2$. Therefore, we choose to fix the estimate $(\sigma^2)^*$ at $\sigma_0^2$ for gaining further insights. To this end, we recall the definition of a generalized minimum-norm solution $\mathbf{m}^\dagger$ in classical inverse theory to the linear inverse problem (1), i.e., it satisfies

$$\mathbf{Hm}^\dagger = \mathbf{d}^\dagger \quad \text{and } \|\mathbf{Lm}^\dagger\|_2 \leqslant \|\mathbf{Lm}\|_2 \quad \forall \mathbf{m} \in \mathbb{R}^m \text{ such that } \mathbf{Hm} = \mathbf{d}^\dagger.$$

Under the assumption that $\ker\mathbf{H} \cap \ker\mathbf{L} = \{\mathbf{0}\}$, the generalized minimum-norm solution is unique. The condition $\mathbf{Hm}^\dagger = \mathbf{d}^\dagger$ requires consistency between the model $\mathbf{H}$ and the exact data $\mathbf{d}^\dagger$, and it holds for sufficiently accurate forward models. Moreover, we have the next result on the mean $\mathbf{m}^* = \mathbf{m}_{\eta^*}$, and for a proof, we refer to Appendix A.1. The result is of deterministic worst-case scenario type as in the classical inverse theory [23]. That is, we have abused the notation $\boldsymbol{\omega}$ of a random variable for its realization. Note that for a Gaussian random variable $\boldsymbol{\omega}$, the event that the norm of its realization $\|\boldsymbol{\omega}\|_2$ is of order $\sigma_0$ occurs with a overwhelming probability.

**Theorem 3.2.** *Assume that $\tau$ is fixed at $\sigma_0^{-2}$, the realization of the random variable $\boldsymbol{\omega}$ satisfies $\|\boldsymbol{\omega}\|_2^2 \leqslant c\sigma_0^2$ for any $\sigma_0$ and ker $\mathbf{H} \cap$ ker $\mathbf{L} = \{\mathbf{0}\}$. Then for fixed $\beta_0$ and $\alpha_0'' \sim \mathcal{O}(\sigma_0^{-d})$ with $0 < d < 2$, the mean $\mathbf{m}^*$ converges to the generalized minimum-norm solution $\mathbf{m}^\dagger$ as $\sigma_0$ tends to zero.*

Theorem 3.2 indicates that under some conditions on $\alpha_0''$ the mean $\mathbf{m}^*$ converges to the desired exact solution as the noise level tends to zero. Conversely, it also implies that hierarchical formulations with fixed $\alpha_0$ might fail, if the noise level $\sigma_0$ varies arbitrarily, albeit the dependence might be rather weak. This observation can be numerically confirmed, see Section 5. The choice $\alpha_0'' \sim \sigma_0^{-d}$ with $0 < d < 2$ is in accordance with the classical inverse theory [23]. The theorem imposes no condition on $\beta_0$, but from the proof in Appendix A.1, the choice $\beta_0 \sim \mathcal{O}\left(\|\mathbf{L}\mathbf{m}^\dagger\|_2^2\right) \approx \mathcal{O}(1)$ seems plausible. Numerically, it has only a negligible effect so long as its value is sufficiently small. In brevity, we may choose the parameter pairs $(\alpha_0, \beta_0)$ and $(\alpha_1, \beta_1)$ according to the following heuristic guidelines: $\alpha_0 \sim \mathcal{O}(\sigma_0^{-1})$, $\beta_0 \sim \mathcal{O}(1)$, $\alpha_1 \sim \mathcal{O}(1)$ and $\beta_1 \ll \sigma_0^2$. As the exact noise level $\sigma_0$ may be unknown, the guideline $\alpha_0 \sim \mathcal{O}(\sigma_0^{-1})$ suggests adaptively updating the parameter $\alpha_0$ using the estimated noise level $(\sigma^2)^*$ in practical implementation. A similar strategy has been derived for the a-Tikhonov method [24]. We will neither further pursue this strategy in the present paper, and nor implement it in the numerical experiments.

### 3.1.3. Algorithm I

We are now in a position to describe an algorithm of alternating iterative type for computing `Approx I`. The Kullback–Leibler divergence is known to be strictly convex [21]. However, the separability assumption renders the optimization non-convex as the domain of definition is no longer convex. Nonetheless, it remains strictly biconvex, i.e., strictly convex with respect to $q(\mathbf{m})$ (respectively $q(\lambda,\tau)$) for fixed $q(\lambda,\tau)$ (respectively $q(\mathbf{m})$). The algorithm derives directly from this analytical observation, and is sketched in Algorithm 1 (`Alg I` for short).

**Algorithm 1.** Variational approximation with $q(\mathbf{m},\lambda,\tau) = q(\mathbf{m})q(\lambda,\tau)$

1: Give an initial guess $q^0(\lambda,\tau)$ and tolerance *tol*, set $k = 0$ and compute $\eta_1 = E_{q^0(\lambda,\tau)}[\lambda]/E_{q^0(\lambda,\tau)}[\tau]$
2: **repeat**
3: Set $k = k + 1$.
4: Find $q^k(\mathbf{m})$ by

$$q^k(\mathbf{m}) = \arg\min_{q(\mathbf{m})} D_{KL}(q(\mathbf{m})q^{k-1}(\lambda,\tau)|p(\mathbf{m},\lambda,\tau)).$$

5: Find $q^k(\lambda,\tau)$ by

$$q^k(\lambda,\tau) = \arg\min_{q(\lambda,\tau)} D_{KL}(q^k(\mathbf{m})q(\lambda,\tau)|p(\mathbf{m},\lambda,\tau)).$$

6: Calculate $\lambda_{k+1} = E_{q^k(\lambda,\tau)}[\lambda]$ and $\tau_{k+1} = E_{q^k(\lambda,\tau)}[\tau]$, and set $\eta_{k+1} = \lambda_{k+1}\tau_{k+1}^{-1}$.
7: **until** $|\eta_{k+1} - \eta_k|/\eta_k \leqslant tol$
8: Return $q^k(\mathbf{m})q^k(\lambda,\tau)$ as the solution

Before commenting on its convergence, let us first further develop each step of the algorithm. Setting the first variation of the divergence $D_{KL}$ with respect to $q(\mathbf{m})$ to zero gives

$$q^k(\mathbf{m}) \propto \exp\left(E_{q^k(\lambda,\tau)}[\ln p(\mathbf{m},\lambda,\tau)]\right),$$

see (13). By repeating the computations in Section 3.1.1, we deduce that $q^k(\mathbf{m})$ follows a Gaussian distribution with covariance $\mathrm{cov}_{q^k(\mathbf{m})}$ and mean $\mathbf{m}_k$ given by

$$\mathrm{cov}_{q^k(\mathbf{m})}[\mathbf{m}] = \left[\tau_k\mathbf{H}^{\mathrm{T}}\mathbf{H} + \lambda_k\mathbf{L}^{\mathrm{T}}\mathbf{L}\right]^{-1} \quad \text{and} \quad \mathbf{m}_k := E_{q^k(\mathbf{m})}[\mathbf{m}] = \mathrm{cov}_{q^k(\mathbf{m})}[\mathbf{m}]\tau_k\mathbf{H}^{\mathrm{T}}\mathbf{d} \equiv \mathbf{m}_{\eta_k},$$

respectively, where $\tau_k = E_{q^{k-1}(\lambda,\tau)}[\tau]$, $\lambda_k = E_{q^{k-1}(\lambda,\tau)}[\lambda]$ and $\eta_k = \lambda_k\tau_k^{-1}$. In other words, the solution in Step 4 is explicitly given by

$$q^k(\mathbf{m}) = \mathcal{N}\left(\mathbf{m}_{\eta_k}, \left[\tau_k\mathbf{H}^{\mathrm{T}}\mathbf{H} + \lambda_k\mathbf{L}^{\mathrm{T}}\mathbf{L}\right]^{-1}\right),$$

which in practical implementation involves computing the mean $\mathbf{m}_{\eta_k}$ and variance $[\tau_k\mathbf{H}^{\mathrm{T}}\mathbf{H} + \lambda_k\,\mathbf{L}^{\mathrm{T}}\mathbf{L}]^{-1}$. Analogously, we can derive

$$q^k(\lambda,\tau) \propto \exp\left(E_{q^k(\mathbf{m})}[\ln p(\mathbf{m},\lambda,\tau)]\right).$$

Direct computation shows that $q^k(\lambda,\tau)$ takes a factorized form, i.e., $q^k(\lambda,\tau) = q^k(\lambda)q^k(\tau)$, and $q^k(\lambda)$ and $q^k(\tau)$ are Gamma distributions given respectively by

$$q^k(\lambda) = G\left(\lambda;\alpha_0'',\frac{1}{2}E_{q^k(\mathbf{m})}\left[\|\mathbf{L}\mathbf{m}\|_2^2\right] + \beta_0\right),$$
$$q^k(\tau) = G\left(\tau;\alpha_1'',\frac{1}{2}E_{q^k(\mathbf{m})}\left[\|\mathbf{H}\mathbf{m} - \mathbf{d}\|_2^2\right] + \beta_1\right).$$

Therefore, the computation of $q^k(\lambda)$ and $q^k(\tau)$ involves simply updating their respective parameter pairs, which can be carried out analytically for $q^k(\mathbf{m})$ follows a multivariate Gaussian. Moreover, observing the separability, we have $\lambda_{k+1} = E_{q^{k(\lambda,\tau)}}[\lambda] = E_{q^{k(\lambda)}}[\lambda]$ and $\tau_{k+1} = E_{q^{k(\lambda,\tau)}}[\tau] = E_{q^{k(\tau)}}[\tau]$, and they can be calculated straightforwardly since both $q^k(\lambda)$ and $q^k(\tau)$ follow a Gamma distribution, see (7).

Since each step of the algorithm decreases the functional value and the functional is bounded from below by zero, the sequence of functional values $\{D_{KL}(q^k(\mathbf{m})q^k(\lambda,\tau)|p(\mathbf{m},\lambda,\tau))\}_k$ is nonincreasing and converges monotonically. Moreover, we have the next convergence theorem of the sequence $\{(q^k(\mathbf{m})q^k(\lambda)q^k(\tau))\}_k$, and for a proof, we refer to Appendix A.2.

**Theorem 3.3.** *The sequence $\{(q^k(\mathbf{m})q^k(\lambda)q^k(\tau))\}_k$ generated by* `Alg I` *has a subsequence converging to a stationary point of the functional $D_{KL}$.*

As a consequence of the proof of the theorem (c.f. Appendix A.2), the convergence is actually determined by the sequence of scalars $\{(\lambda_k,\tau_k,\eta_k)\}_k$. This suggests a natural and simple stopping criterion for `Alg I`: monitoring one or several of these scalars. Specifically, the regularization parameter $\eta_k$ determines the mean $\mathbf{m}_{\eta_k}$, which is the quantity of primary interest. Therefore, we propose the following heuristic stopping criterion for the algorithm: the algorithm stops if there holds

$$\frac{|\eta_{k+1}-\eta_k|}{\eta_k} < tol \tag{14}$$

for some given tolerance *tol*. Computationally speaking, it is simpler than directly monitoring the decrease of the divergence itself, which for high-dimensional problems is more expensive to evaluate.

Finally, we note that in practice the GCV estimate $\mathcal{V}(\eta)$ is almost unchanged for $\eta$ varying over a broad scale while remaining a good approximation to $\sigma_0^2$, so is the estimate $(\sigma^2)^*$. In other words, the convergence of the scalar $(\sigma^2)^*$ is relatively fast during the iteration procedure. This motivates us to consider a formulation by fixing the value of $\tau$ at $\sigma_0^{-2}$ to shed further lights on the practical convergence behavior of `Alg I`. Specifically, by (21) from Appendix A.1, the sequence $\{\eta_k = \lambda_k\sigma_0^2\}_k$ generated by `Alg I` can be succinctly written in terms of the following fixed point iteration:

$$\eta_{k+1} = \frac{2\alpha_0''\sigma_0^2}{\|\mathbf{L}\mathbf{m}_k\|_2^2 + \sigma_0^2\sum_{i=1}^p \frac{1}{\gamma_i^2+\eta_k} + 2\beta_0},$$

where $\mathbf{m}_k = \mathbf{m}_{\eta_k}$, and $\gamma_i$ are generalized singular values of the matrix pair $(\mathbf{H},\mathbf{L})$ [22], see also Appendix A.1. The next theorem shows that the sequence $\{\eta_k\}_k$ converges always monotonically. This casts new light on `Alg I`: The sequence $\{\eta_k\}_k$ should be monotonic after some initial iterations, i.e., locally monotonic, and thus the algorithm merits a steady convergence.

**Theorem 3.4.** *For any initial guess $\eta_0$, the sequence $\{\eta_k\}_k$ converges monotonically.*

**Proof.** By the definition of $\eta_k$, we have

$$\begin{aligned}\eta_{k+1}-\eta_k &= \frac{2\alpha_0''\sigma_0^2}{\|\mathbf{L}\mathbf{m}_k\|_2^2 + \sigma_0^2\sum_{i=1}^p \frac{1}{\gamma_i^2+\eta_k} + 2\beta_0} - \frac{2\alpha_0''\sigma_0^2}{\|\mathbf{L}\mathbf{m}_{k-1}\|_2^2 + \sigma_0^2\sum_{i=1}^p \frac{1}{\gamma_i^2+\eta_{k-1}} + 2\beta_0}\\ &= \frac{2\alpha_0''\sigma_0^2}{D_k}\left[\|\mathbf{L}\mathbf{m}_{k-1}\|_2^2 - \|\mathbf{L}\mathbf{m}_k\|_2^2 + \sigma_0^2\sum_{i=1}^p\left(\frac{1}{\gamma_i^2+\eta_{k-1}} - \frac{1}{\gamma_i^2+\eta_k}\right)\right] := \frac{2\alpha_0''\sigma_0^2}{D_k}[I+\sigma_0^2 II],\end{aligned}$$

where the denominator $D_k$ is defined as

$$D_k := \left(\|\mathbf{L}\mathbf{m}_{k-1}\|_2^2 + \sigma_0^2\sum_{i=1}^p\frac{1}{\gamma_i^2+\eta_{k-1}} + 2\beta_0\right)\left(\|\mathbf{L}\mathbf{m}_k\|_2^2 + \sigma_0^2\sum_{i=1}^p\frac{1}{\gamma_i^2+\eta_k} + 2\beta_0\right)$$

and the terms $I$ and $II$ in the square bracket are respectively given by

$$\begin{aligned} I &:= \|\mathbf{L}\mathbf{m}_{k-1}\|_2^2 - \|\mathbf{L}\mathbf{m}_k\|_2^2,\\ II &:= \sum_{i=1}^p\left(\frac{1}{\gamma_i^2+\eta_{k-1}} - \frac{1}{\gamma_i^2+\eta_k}\right) = \sum_{i=1}^p\frac{1}{(\gamma_i^2+\eta_{k-1})(\gamma_i^2+\eta_k)}(\eta_k-\eta_{k-1}).\end{aligned}$$

Since the term $\|\mathbf{L}\mathbf{m}_{k-1}\|_2^2 - \|\mathbf{L}\mathbf{m}_k\|_2^2$ has the same sign with $(\eta_k-\eta_{k-1})$ [9], the sequence $\{\eta_k\}_k$ is monotonic. A trivial lower bound of $\{\eta_k\}_k$ is zero. Moreover, we note that

$$\eta_k = \frac{2\alpha_0''\sigma_0^2}{\|\mathbf{L}\mathbf{m}_{k-1}\|_2^2 + \sigma_0^2\sum_{i=1}^p \frac{1}{\gamma_i^2+\eta_{k-1}} + 2\beta_0} \leqslant \frac{\alpha_0''\sigma_0^2}{\beta_0}.$$

The uninform boundedness and its monotonicity imply the desired convergence. □

### 3.2. Degenerate approximation

In `Approx I`, the only assumption on $q(\mathbf{m})$ and $q(\lambda,\tau)$ is their conditional independence given the observational data $\mathbf{d}$. Extra assumptions might be invoked to further reduce its complexity. In `Approx II`, we now consider the case of a degenerate $q(\mathbf{m})$, i.e., it takes one value with probability one and the rest with probability zero. Then the Kullback–Leibler divergence $D_{KL}$ reduces to

$$D_{KL}(q(\mathbf{m},\lambda,\tau)|p(\mathbf{m},\lambda,\tau)) = \int\int q(\lambda,\tau)\ln\frac{q(\lambda,\tau)}{p(\mathbf{m},\lambda,\tau)}d\lambda d\tau. \tag{15}$$

**Theorem 3.5.** *There exists at least one minimizer to the optimization problem (15).*

**Proof.** Note that for fixed $\mathbf{m}$, $p(\mathbf{m},\lambda,\tau)$ can be regarded as an unnormalized Gamma distribution in both $\lambda$ and $\tau$, and by (7), the normalizing constant $c(\mathbf{m})$ is given by

$$c(\mathbf{m}) = \Gamma(\alpha_0'')^{-1}\Gamma(\alpha_1'')^{-1}\left(\beta_0+\frac{1}{2}\|\mathbf{Lm}\|_2^2\right)^{\alpha_0''}\left(\beta_1+\frac{1}{2}\|\mathbf{Hm}-\mathbf{d}\|_2^2\right)^{\alpha_1''}.$$

Now by Jensen's inequality (10), for fixed $\mathbf{m}$ the divergence $D_{KL}$ defined in (15) attains its minimum at $q(\lambda,\tau) = c(\mathbf{m})p(\mathbf{m},\lambda,\tau)$ with $c(\mathbf{m})$ defined above. With this choice of $q(\lambda,\tau)$, we integrate out $\lambda$ and $\tau$ to obtain the minimum value of the divergence $D_{KL}$ in terms of $\mathbf{m}$

$$\inf_{q(\lambda,\tau)} D_{KL}(q(\mathbf{m},\lambda,\tau)|p(\mathbf{m},\lambda,\tau)) = \ln c(\mathbf{m}) = \alpha_1''\ln\left(\frac{1}{2}\|\mathbf{Hm}-\mathbf{d}\|_2^2+\beta_1\right)+\alpha_0''\ln\left(\frac{1}{2}\|\mathbf{Lm}\|_2^2+\beta_0\right)+\ln\Gamma(\alpha_0'')^{-1}\Gamma(\alpha_1'')^{-1}.$$

The constant $\ln\Gamma(\alpha_0'')^{-1}\Gamma(\alpha_1'')^{-1}$ is irrelevant in the minimization and thus can be ignored. Consequently, we arrive at a reduced functional $\mathcal{G}(\mathbf{m})$ in terms of the solution $\mathbf{m}$ only

$$\mathcal{G}(\mathbf{m}) = \alpha_1''\ln\left(\frac{1}{2}\|\mathbf{Hm}-\mathbf{d}\|_2^2+\beta_1\right)+\alpha_0''\ln\left(\frac{1}{2}\|\mathbf{Lm}\|_2^2+\beta_0\right).$$

The functional $\mathcal{G}(\mathbf{m})$ is bounded from below and continuous on $\mathbb{R}^m$, and the assumption $\ker\mathbf{H}\cap\ker\mathbf{L} = \{\mathbf{0}\}$ implies that it is also coercive. Therefore, the functional $\mathcal{G}$ has at least one minimizer. Consequently, the divergence $D_{KL}$ has at least one minimizer. □

By differentiating the reduced functional $\mathcal{G}(\mathbf{m})$ with respect to $\mathbf{m}$ and setting it to zero yields

$$\alpha_1''\frac{\mathbf{H}^{\mathrm{T}}(\mathbf{Hm}-\mathbf{d})}{\frac{1}{2}\|\mathbf{Hm}-\mathbf{d}\|_2^2+\beta_1}+\alpha_0''\frac{\mathbf{L}^{\mathrm{T}}\mathbf{Lm}}{\frac{1}{2}\|\mathbf{Lm}\|_2^2+\beta_0} = 0,$$

which after rearranging gives

$$(\mathbf{H}^{\mathrm{T}}\mathbf{H}+\eta\mathbf{L}^{\mathrm{T}}\mathbf{L})\mathbf{m} = \mathbf{H}^{\mathrm{T}}\mathbf{d}$$

with $\eta = \frac{\alpha_0''}{\alpha_1''}\frac{\frac{1}{2}\|\mathbf{Hm}-\mathbf{d}\|_2^2+\beta_1}{\frac{1}{2}\|\mathbf{Lm}\|_2^2+\beta_0}$. This indicates that the critical point $\mathbf{m}^*$ is a minimizer of the Tikhonov functional $\mathcal{J}_{\eta^*}$, i.e., $\mathbf{m}^* = \mathbf{m}_{\eta^*}$. Now by recalling the optimal choice of $q^*(\lambda,\tau)$ for achieving the minimum divergence in the proof of Theorem 3.5, we deduce that $q^*(\lambda,\tau)$ is proportional to $p(\mathbf{m}^*,\lambda,\tau)$ and thus both $\lambda$ and $\tau$ follow a Gamma distribution. Consequently, the optimality system at a critical point $q^*(\mathbf{m},\lambda,\tau) = \delta(\mathbf{m}-\mathbf{m}^*)q^*(\lambda)q^*(\tau))$ of the divergence reads

$$\begin{aligned}
&\mathbf{m}^* = (\mathbf{H}^{\mathrm{T}}\mathbf{H}+\eta^*\mathbf{L}^{\mathrm{T}}\mathbf{L})^{-1}\mathbf{H}^{\mathrm{T}}\mathbf{d},\\
&q^*(\lambda) = G\left(\lambda;\alpha_0'',\frac{1}{2}\|\mathbf{Lm}^*\|_2^2+\beta_0\right),\\
&q^*(\tau) = G\left(\tau;\alpha_1'',\frac{1}{2}\|\mathbf{Hm}^*-\mathbf{d}\|_2^2+\beta_1\right).
\end{aligned} \tag{16}$$

Let $\lambda^* = E_{q^*(\lambda)}[\lambda]$ and $\tau^* = E_{q^*(\tau)}[\tau]$, then we have $\eta^* = \frac{\lambda^*}{\tau^*}$. Recall the defining equation $(\sigma^2)^* = 1/\tau^*$. Thus `Approx II` estimates $\sigma_0^2$ by

$$(\sigma^2)^* := \frac{1}{\tau^*} = \frac{\frac{1}{2}\|\mathbf{Hm}_{\eta^*}-\mathbf{d}\|_2^2+\beta_1}{\frac{n}{2}+\alpha_1}.$$

Analogous to the variance estimate by `Approx I`, the estimate $(\sigma^2)^*$ accepts interesting interpretations as a convex combination of prior and data contributions. Upon adopting a noninformative prior for $\tau$, i.e., $\alpha_1\approx 1$ and $\beta_1\approx 0$, the estimate $(\sigma^2)^*$ is related to $\mathcal{V}(\eta)$ by $(\sigma^2)^*\approx\mathcal{V}(\eta)\frac{\mathcal{T}(\eta)}{n}$. For inverse problems with exponentially decaying spectra, which holds for integral equations of the first kind with smooth kernels, the proportional factor $\frac{\mathcal{T}(\eta)}{n}$ varies slowly with $\eta$ and remains of order one

for $\eta$ over a very broad scale [9]. Thus the estimate $(\sigma^2)^*$ also represents an accurate approximation to the exact noise variance $\sigma_0^2$, and it was numerically confirmed for some benchmark inverse problems, e.g. Fredholm integral equations of the first kind and ill-posed Cauchy problems [9].

`Approx II` is closely related to the a-Tikhonov method. To illustrate the point, first observe that the optimality condition of the functional $\mathcal{G}(\mathbf{m})$ can be rewritten as

$$\mathbf{H}^{\mathrm{T}}\mathbf{H}\mathbf{m}+\frac{\alpha_0''}{\alpha_1''}\frac{\frac{1}{2}\|\mathbf{H}\mathbf{m}-\mathbf{d}\|_2^2+\beta_1}{\frac{1}{2}\|\mathbf{L}\mathbf{m}\|_2^2+\beta_0}\mathbf{L}^{\mathrm{T}}\mathbf{L}\mathbf{m}=\mathbf{H}^{\mathrm{T}}\mathbf{d}.$$

Similarly, by considering the optimality system of the a-Tikhonov functional, the regularization parameter $\eta=\lambda\tau^{-1}$ verifies $\eta=\frac{\alpha_0'}{\alpha_1'}\frac{\frac{1}{2}\|\mathbf{H}\mathbf{m}_\eta-\mathbf{d}\|_2^2+\beta_1}{\frac{1}{2}\|\mathbf{L}\mathbf{m}_\eta\|_2^2+\beta_0}$. By substituting it into the optimality condition of the Tikhonov functional $\mathcal{J}_\eta$, we arrive at its alternative characterization of the functional $\mathcal{J}(\mathbf{m},\lambda,\tau)$ in terms of $\mathbf{m}$

$$\mathbf{H}^{\mathrm{T}}\mathbf{H}\mathbf{m}+\frac{\alpha_0'}{\alpha_1'}\frac{\frac{1}{2}\|\mathbf{H}\mathbf{m}-\mathbf{d}\|_2^2+\beta_1}{\frac{1}{2}\|\mathbf{L}\mathbf{m}\|_2^2+\beta_0}\mathbf{L}^{\mathrm{T}}\mathbf{L}\mathbf{m}=\mathbf{H}^{\mathrm{T}}\mathbf{d}.$$

Therefore, the optimality condition of `Approx II` coincides with that of the a-Tikhonov functional $\mathcal{J}(\mathbf{m},\lambda,\tau)$ with $\alpha_0''$ and $\alpha_1''$ in place of $\alpha_0'$ and $\alpha_1'$. The difference is due to the fact that `Approx II` takes the mean of the PPDF, whereas the a-Tikhonov method considers the mode. Usually we have inequalities $m,\ n\gg 1$, and thus there holds $\alpha_0'\approx\alpha_0''$ and $\alpha_1'\approx\alpha_1''$. Consequently `Approx II` and the a-Tikhonov method yield practically identical numerical results. This shows the unifying nature of the variational method in that we can derive several approximations within the framework. However, `Approx II` also attempts to quantify the uncertainties of the hyper-parameters. Because of this interesting connection, the theoretical results [9] developed for the a-Tikhonov method apply equally well to `Approx II`. In particular, an analogue of Theorem 3.2 holds also for `Approx II`, and the comments after Theorem 3.2 apply as well.

Theoretically, the point estimates $\mathbf{m}_{\eta^*}$ by both `Approx I` and `II` are convergent under identical assumptions. Moreover, both variance estimates $(\sigma^2)^*$ are in excellent agreement when adopting a noninformative prior for the parameter $\tau$. However, `Approx I` and `II` can be different if $E_{q^*(\mathbf{m})}\left[\|\mathbf{L}\mathbf{m}\|_2^2\right]$ differs significantly from $\|\mathbf{L}\mathbf{m}_{\eta^*}\|_2^2$. The term $E_{q^*(\mathbf{m})}\left[\|\mathbf{L}\mathbf{m}\|_2^2\right]$ consists of two components, i.e., the mean $\|\mathbf{L}\mathbf{m}_{\eta^*}\|_2^2$ and the variance tr $((\tau^*\mathbf{H}^{\mathrm{T}}\mathbf{H}+\lambda^*\mathbf{L}^{\mathrm{T}}\mathbf{L})^{-1}\mathbf{L}^{\mathrm{T}}\mathbf{L})$. Therefore, these two methods would yield identical results if and only if in the term $E_{q^*(\mathbf{m})}\left[\|\mathbf{L}\mathbf{m}\|_2^2\right]$, the variance component is small compared to the mean component, i.e., $\mathrm{tr}((\tau^*\mathbf{H}^{\mathrm{T}}\mathbf{H}+\lambda^*\mathbf{L}^{\mathrm{T}}\mathbf{L})^{-1}\mathbf{L}^{\mathrm{T}}\mathbf{L})\ll\|\mathbf{L}\mathbf{m}_{\eta^*}\|_2^2$. Otherwise, they might deliver quite different solutions. This marked difference between `Approx I` and `II` arises when, loosely speaking, the MAP is not representative of the PPDF.

We now consider an algorithm for numerically realizing `Approx II`. It is of alternating iterative type like `Alg I`, and derives again from the biconvexity of the functional (15). The complete algorithm is shown in Algorithm 2 (`Alg II` for short). Observe again the computation of $\lambda_k$ and $\tau_k$ are straightforward as both $\lambda$ and $\tau$ follow a Gamma distribution.

**Algorithm 2.** Variational approximation with $q(\mathbf{m},\lambda,\tau)=q(\mathbf{m})q(\lambda,\tau)$ and $q(\mathbf{m})$ degenerate

1: Give an initial guess $q^0(\lambda,\tau)$ and tolerance *tol*, set $k=0$ and $\eta_0=0$
2: **repeat**
3: Set $k=k+1$
4: Calculate $\lambda_k=E_{q^{k-1}(\lambda,\tau)}[\lambda]$, $\tau_k=E_{q^{k-1}(\lambda,\tau)}[\tau]$, and $\eta_k=\lambda_k\tau_k^{-1}$
5: Find $\mathbf{m}_k$ by

$$\mathbf{m}_k=\left(\mathbf{H}^{\mathrm{T}}\mathbf{H}+\eta_k\mathbf{L}^{\mathrm{T}}\mathbf{L}\right)^{-1}\mathbf{H}^{\mathrm{T}}\mathbf{d}.$$

6: Calculate $q^k(\lambda,\tau)$ by

$$q^k(\lambda,\tau)=q^k(\lambda)q^k(\tau),$$

where $q^k(\lambda)$ and $q^k(\tau)$ are Gamma distributions given, respectively, by

$$q^k(\lambda)=G\left(\lambda;\alpha_0'',\frac{1}{2}\|\mathbf{L}\mathbf{m}_k\|_2^2+\beta_0\right)\ \text{and}\ q^k(\tau)=G\left(\tau;\alpha_1'',\frac{1}{2}\|\mathbf{H}\mathbf{m}_k-\mathbf{d}\|_2^2+\beta_1\right).$$

7: **until** $|\eta_k-\eta_{k-1}|/\eta_{k-1}\leqslant tol$
8: Return $q^k(\mathbf{m})q^k(\lambda,\tau)$ as the solution

Due to the close connection between `Approx II` and the a-Tikhonov method revealed above, `Alg II` is guaranteed to converge to a critical point of the functional, see Ref. [9] for an analysis of the a-Tikhonov method. The stopping criterion can be taken the same as that for `Alg I`, i.e., monitoring the relative change of the regularization parameter $\eta_k$, see (14).

## 4. Extensions

In this section we briefly discuss two generalizations of the quadratic model in the preceding section to more complicated models: nonGaussian prior and nonlinear forward models, which arise very often in real-world applications. The goal is to indicate the potential broad applicability of the variational method, instead of rigorous mathematical justification or theoretical properties.

### 4.1. $\ell^r$ prior

This part considers a nonGaussian prior, which recently has received significant interest in several areas, e.g. compressive sensing in signal processing and image processing [25]. More precisely, the prior $p(\mathbf{m}|\lambda)$ is of the form

$$p(\mathbf{m}|\lambda) \propto \frac{1}{Z(\lambda)} e^{-\lambda\|\mathbf{m}\|_r^r},$$

where $\|\mathbf{m}\|_r^r = \sum_i |m_i|^r$ and $Z(\lambda)$ is the normalizing constant. By a change of variable $\lambda s^r = t$, $Z(\lambda)$ can be calculated by observing

$$\int_0^\infty e^{-\lambda s^r} ds = \lambda^{-\frac{1}{r}} \int_0^\infty e^{-t} t^{\frac{1-r}{r}} dt \propto \lambda^{-\frac{1}{r}},$$

which further gives $Z(\lambda) = c_r \lambda^{-\frac{m}{r}}$ with $c_r$ being a constant. Consequently, the prior is given by

$$p(\mathbf{m}|\lambda) \propto \lambda^{-\frac{m}{r}} e^{-\lambda\|\mathbf{m}\|_r^r}.$$

For the $\ell^r$ prior, the corresponding Kullback–Leibler divergence cannot be calculated directly. We resort to a majorization–minimization approach. The basic idea is to utilize an auxiliary variable [17,26] to derive a lower bound for the prior, which is then iteratively updated. Recall by Young's inequality, i.e., for $a, b \geqslant 0$ and $r \in [0,2]$, there holds $a^{\frac{r}{2}} b^{1-\frac{r}{2}} \leqslant \frac{r}{2} a + (1-\frac{r}{2}) b$, we get

$$a^{\frac{r}{2}} \leqslant \frac{r}{2} \frac{a + (\frac{2}{r} - 1) b}{b^{1-\frac{r}{2}}}. \tag{17}$$

Next we define an auxiliary function $M(\mathbf{m}, \lambda, \mathbf{v})$ as below

$$M(\mathbf{m}, \lambda, \mathbf{v}) = c_r \lambda^{\frac{m}{r}} \exp\left( -\frac{r}{2} \lambda \sum_i \frac{m_i^2 + (\frac{2}{r} - 1) v_i}{v_i^{1-\frac{r}{2}}} \right),$$

where $\mathbf{v} = (v_1, \ldots, v_m)^{\mathrm{T}} \in \mathbb{R}_+^m$, $c_r$ is the constant in the normalizing constant $Z(\lambda)$. By virtue of inequality (17), we have

$$p(\mathbf{m}|\lambda) \geqslant M(\mathbf{m}, \lambda, \mathbf{v}) \quad \forall \mathbf{v} \in \mathbb{R}_+^m.$$

Observe that for any fixed $\mathbf{v}$, $M(\mathbf{m}, \lambda, \mathbf{v})$ is quadratic in $\mathbf{m}$, for which Bayesian inference can be analytically carried out, and bounds the prior from below. Consequently, a lower bound of the joint distribution $p(\mathbf{m}, \lambda, \tau)$ is given by

$$p(\mathbf{m}, \lambda, \tau) \geqslant p(\mathbf{d}|\mathbf{m}, \tau) M(\mathbf{m}, \lambda, \mathbf{v}) p(\lambda) p(\tau) \equiv F(\mathbf{m}, \lambda, \tau, \mathbf{v})$$

and thus we get an upper bound for the divergence, i.e.,

$$D_{KL}(q(\mathbf{m}, \lambda, \tau) | p(\mathbf{m}, \lambda, \tau)) \leqslant D_{KL}(q(\mathbf{m}, \lambda, \tau) | F(\mathbf{m}, \lambda, \tau, \mathbf{v})).$$

The upper bound can be tightened by minimizing it iteratively with respect to both $q(\mathbf{m}, \lambda, \tau)$ and $\mathbf{v}$, which results in a decreasing sequence of upper bounds and also a better approximation of the true prior $p(\mathbf{m}|\lambda)$ by the functional $M(\mathbf{m}, \lambda, \mathbf{v})$.

In sum, we replace the minimization of the Kullback–Leibler divergence by its upper bound in the spirit of classical majorization–minimization method. The structure of the upper bound suggests an alternating iterative procedure as before. The complete algorithm is given in Algorithm 3 (`Alg III` for short). In particular, for fixed $\mathbf{v}$, the standard solution of variational Bayesian analysis can be employed.

Algorithm 3. Variational approximation with $q(\mathbf{m},\lambda,\tau) = q(\mathbf{m})q(\lambda,\tau)$ for $\ell^r$ prior

1: Give an initial guess $q^0(\lambda,\tau)$, $\mathbf{v}^0$ and tolerance *tol*, and set $k = 0$ and compute $\eta_1 = E_{q^0(\lambda,\tau)}[\lambda]/E_{q^0(\lambda,\tau)}[\tau]$.
2: **repeat**
3: Set $k = k + 1$.
4: Find $q^k(\mathbf{m})$ by

$$q^k(\mathbf{m}) = \arg\min_{q(\mathbf{m})} D_{KL}(q(\mathbf{m})q^{k-1}(\lambda,\tau)|F(\mathbf{m},\lambda,\tau,\mathbf{v}^{k-1})).$$

5: Find $\mathbf{v}^k$ by

$$\mathbf{v}^k = \arg\min_{\mathbf{v}\in\mathbb{R}_+^m} D_{KL}(q^k(\mathbf{m})q^{k-1}(\lambda,\tau)|F(\mathbf{m},\lambda,\tau,\mathbf{v})).$$

6: Find $q^k(\lambda,\tau)$ by

$$q^k(\lambda,\tau) = \arg\min_{q(\lambda,\tau)} D_{KL}(q^k(\mathbf{m})q(\lambda,\tau)|F(\mathbf{m},\lambda,\tau,\mathbf{v}^k)).$$

7: Calculate $\lambda_{k+1} = E_{q^k(\lambda,\tau)}[\lambda]$ and $\tau_{k+1} = E_{q^k(\lambda,\tau)}[\tau]$, and set $\eta_{k+1} = \lambda_{k+1}\tau_{k+1}^{-1}$.
8: **until** $|\eta_{k+1} - \eta_k|/\eta_k \leqslant tol$.
9: Return $q^k(\mathbf{m})q^k(\lambda,\tau)$ as the solution.

Next we give explicit formulas for each step of the algorithm. Steps 4 and 6 can proceed as before. The distribution $q^k(\mathbf{m})$ is a multivariate Gaussian distribution with covariance $\mathrm{cov}_{q^k(\mathbf{m})}$ and mean $\mathbf{m}^k$ given by

$$\mathrm{cov}_{q^k(\mathbf{m})}[\mathbf{m}] = (\tau_k\mathbf{H}^{\mathrm{T}}\mathbf{H} + \lambda_k\mathbf{W}_k)^{-1}, \quad \mathbf{m}^k = \mathrm{cov}_{q^k(\mathbf{m})}[\mathbf{m}]\tau_k\mathbf{H}^{\mathrm{T}}\mathbf{d}$$

with $\mathbf{W}_k = \mathrm{diag}(r(v_i^k)^{\frac{r}{2}-1})$ being a diagonal matrix. The minimization with respect to $\mathbf{v}$ in Step 5 can also be carried out explicitly componentwise as below

$$v_i^k = \arg\min_{v_i} \frac{E_{q^k(\mathbf{m})}[m_i^2] + (\frac{2}{r}-1)v_i}{v_i^{1-\frac{r}{2}}}$$

from which it follows that $v_i^k = E_{q^k(\mathbf{m})}[m_i^2]$. Similarly, we can deduce that $q^k(\lambda,\tau) = q^k(\lambda)q^k(\tau)$ is a Gamma distribution given by

$$q^k(\lambda) = G\left(\lambda;\ \alpha_0 + \frac{m}{r}, \beta_0 + \sum_i (v_i^k)^{\frac{r}{2}}\right),$$
$$q^k(\tau) = G\left(\tau;\ \alpha_1 + \frac{n}{2}, \beta_1 + \frac{1}{2}E_{q^k(\mathbf{m})}\|\mathbf{H}\mathbf{m} - \mathbf{d}\|_2^2\right).$$

### 4.2. Nonlinear model

Many practical inverse problems are described by nonlinear forward models, and a linearized model may be insufficient. Thus it is of great interest to extend the variational Bayesian approach to nonlinear models directly. There have been some investigations, see e.g. the recent work [27] and references therein. We shall follow the idea of recursive linearization as the classical Gauss–Newton method for nonlinear optimization problems, which was recently investigated in [27].

To this end, let the nonlinear forward model be $\mathbf{H}(\mathbf{m}) : \mathbb{R}^m \to \mathbb{R}^n$. Upon adopting the additive Gaussian noise assumption on the data $\mathbf{d}$, we may write down the likelihood $p(\mathbf{d}|\mathbf{m},\tau)$ as

$$p(\mathbf{d}|\mathbf{m},\tau) \propto \tau^{\frac{n}{2}} \exp\left(-\frac{\tau}{2}\|\mathbf{H}(\mathbf{m}) - \mathbf{d}\|^2\right),$$

where again $\tau$ denotes the inverse variance. Proceeding as in Section 2, we can easily derive the PPDF. However, calculating the Kullback–Leibler divergence for general nonlinear forward models is analytically intractable. To ensure tractability and generality, we approximate the nonlinear forward model $\mathbf{H}(\mathbf{m})$ by its first-order Taylor expansion $\tilde{\mathbf{H}}(\mathbf{m})$ around the mode $\tilde{\mathbf{m}}$ of the posterior distribution (which for a Gaussian distribution is also the mean), i.e.,

$$\tilde{\mathbf{H}}(\mathbf{m}) = \mathbf{H}(\tilde{\mathbf{m}}) + \mathbf{J}(\mathbf{m} - \tilde{\mathbf{m}}),$$

where $\mathbf{J} = \nabla_{\mathbf{m}}\mathbf{H}$ is the Jacobian of the forward model $\mathbf{H}$ with respect to $\mathbf{m}$ evaluated at $\tilde{\mathbf{m}}$. With this linearization at hand, we recover the computational tractability, and in particular, `Alg I` might be employed for an approximation. Now we can

describe a variational method for nonlinear models, see Algorithm 4 (`Alg IV` for short) for the complete algorithm. Observe that the variational approximation $q^k(\mathbf{m})$ still follows a Gaussian distribution, despite the fact that the true conditional distribution $p(\mathbf{m}|\lambda,\tau,\mathbf{d})$ is nonGaussian. This contrasts sharply with the linear case discussed in Section 3. Note that in the inner loop of `Alg IV`, the variational approximation needs not to be carried out very accurately. Ref. [27] suggested one step for the inner loop, and demonstrated its performance for some problems arising from signal processing. In `Alg IV`, several different stopping criteria for the outer loop might be adopted, e.g. based on relative change of the inverse solution $\mathbf{m}$ or the regularization parameter $\eta$.

Algorithm 4. Variational approximation with $q(\mathbf{m},\lambda,\tau) = q(\mathbf{m})q(\lambda,\tau)$ for a nonlinear model $\mathbf{H}(\mathbf{m})$

1: Give an initial guess $q^0(\lambda,\tau)$, $\tilde{\mathbf{m}}^0$ and tolerance *tol*, and set $k = 0$
2: **repeat**
3: Calculate the linearized model, i.e., $\tilde{\mathbf{H}}(\mathbf{m}) = \mathbf{H}(\tilde{\mathbf{m}}^k) + \mathbf{J}^k(\mathbf{m} - \tilde{\mathbf{m}}^k)$
4: Set $k = k + 1$
5: Find variational approximation $q^k(\mathbf{m})q^k(\lambda,\tau)$ using $\tilde{\mathbf{H}}(\mathbf{m})$
6: **until** A stopping criterion is satisfied
7: Return $q^k(\mathbf{m})q^k(\lambda,\tau)$ as the solution

## 5. Numerical experiments and discussions

In this section, we evaluate the developed techniques on linear and nonlinear Cauchy-type problems in heat conduction. The mathematical formulation and its finite element discretization is first described, and then numerical results are presented.

### 5.1. Cauchy-type problems and discretization

Cauchy-type problems of determining an unknown boundary condition (coefficient) on a part of the boundary from overdetermined boundary data on the remaining part arise frequently in heat conduction. For instance, in the study of re-entrant space shuttles [1,28], one needs to infer the temperature on the outer surface of the shuttle from measurements on the inner surface. In quenching process [29,30], a coefficient profiling convective heat transfer through the boundary needs to be estimated from measurements of temperature and flux on the accessible part of the boundary.

Mathematically, the problem can be formulated as follows. Let $\Omega$ be an open and bounded domain in $\mathbb{R}^d (d \geqslant 2)$ with a boundary $\Gamma$. The boundary $\Gamma$ is divided into two disjointed parts $\Gamma = \Gamma_i \cup \Gamma_c$, which refer to the part of boundary inaccessible and accessible to experimental measurement, respectively. The steady-state heat conduction problem could be described by

$$-\nabla \cdot (\alpha \nabla u) = f \quad \text{in } \Omega, \tag{18}$$

where $\alpha(x)$ is the conductivity and $f(x)$ is the source term. Now (18) is subjected to

$$\alpha \frac{\partial u}{\partial n} = q \quad \text{on } \Gamma_c \quad \text{and } u = g \quad \text{on } \Gamma_o \subset \Gamma_c,$$

where $n$ is the unit outward normal to $\Gamma_c$. The inverse problem seeks the unknown temperature $u(x)$ (or an unknown coefficient) on the boundary $\Gamma_i$.

This problem is known as the Cauchy problem since both Dirichlet and Neumann boundary conditions are prescribed simultaneously on a part of the boundary, and the set of data is known as the Cauchy data. It is severely ill-posed, and thus is mathematically and numerically much more challenging than the forward problem. The unique continuation principle for elliptic equations ensures the uniqueness of the solution. The stability has been the subject of intensive investigations [33]. However, the solution may not exist when noise is present in the data, and it lacks a continuous dependence on the data. We refer to [31,8,28,30], and references therein for some numerical methods.

Next we describe the finite element discretization. We describe only the case of unknown temperature for simplicity, and refer to [32] for identifying boundary coefficients. Formally, the inverse problem can be written as an operator equation $F(\theta) = g$, where $\theta$ denotes the unknown temperature on the boundary $\Gamma_i$, and the operator $F$ is defined by $F : \theta \to u(\theta)|_{\Gamma_o}$ with $u = u(\theta)$ being the solution to (18) together with the following boundary conditions:

$$\alpha \frac{\partial u}{\partial n} = q \quad \text{on } \Gamma_c \quad \text{and } u = \theta \quad \text{on } \Gamma_i. \tag{19}$$

To discretize the linear operator $F$ into its discrete counterpart $\mathbf{F}$, we employ the finite element method. Let $\mathcal{T}_h$ be a shape regular and quasi-uniform triangulation of the domain $\Omega$. Then the piecewise linear finite element space $V_{h,\theta}$ is defined by

$$V_{h,\theta} = \left\{ \phi_h \in C(\bar{\Omega}) : \phi_h|_{\mathcal{T}_j} \in P_1(\mathcal{T}_j),\ \phi_h|_{\Gamma_i} = \theta \quad \forall \mathcal{T}_j \in \mathcal{T}_h \right\},$$

where $P_1(\mathcal{T}_j)$ denotes the space of all linear polynomials on the finite element $\mathcal{T}_j$. The finite element solution $u_h \in V_{h,\theta}$ to (18) and (19) solves the variational equation

$$\int_\Omega \alpha \nabla u_h \cdot \nabla \phi_h dx = \int_\Omega f\phi_h dx + \int_{\Gamma_c} q\phi_h ds \quad \forall \phi_h \in V_{h,0}.$$

The unknown temperature $\theta$ is parameterized by

$$\theta(x) = \sum_{j=1}^{m} m_j w_j(x),$$

where $w_j(x)$ are finite element basis functions defined on $\Gamma_i$, $m$ is the number of basis functions and $m_j$ are the unknowns to be estimated. We denote by $\mathbf{m}$ the vector $(m_1, m_2, \ldots, m_m)^{\mathrm{T}}$. The linearity of the problem enables the following splitting:

$$\mathbf{F}(\mathbf{m}) = \mathbf{H}\mathbf{m} + \mathbf{u}_{\mathrm{H}},$$

where the function $\mathbf{u}_{\mathrm{H}}$ is the restriction of the finite element solution to (19) with $\theta = 0$ to the boundary $\Gamma_o$, and $\mathbf{H} \in \mathbb{R}^{n\times m}$ is a sensitivity matrix with $n$ being the number of the measurements on the boundary $\Gamma_o$, i.e., the $j$th column of $\mathbf{H}$ is the finite element solution $v_h \in V_{h,w_j}$ of the variational equation

$$\int_\Omega \alpha \nabla v_h \cdot \nabla \phi_h dx = 0 \quad \forall \phi_h \in V_{h,0},$$

restricted on the boundary $\Gamma_o$. Then the inverse problem can be written in the form of (1) with $\mathbf{d} = \mathbf{g} - \mathbf{u}_{\mathrm{H}}$ being the data, where the vector $\mathbf{g}$ stands for the discrete temperature measurements on $\Gamma_o$.

### 5.2. Numerical experiments

This part presents numerical results to illustrate the variational method. Consider the Laplace equation, i.e. $\alpha(x) = 1$ and $f(x) = 0$ in (18). The domain $\Omega$ under consideration is a unit square $(0,1)\times(0,1)$, and the boundaries $\Gamma_i$ and $\Gamma_c$ are $\Gamma_i = [0,1]\times\{1\}$ and $\Gamma_c = \Gamma\backslash\Gamma_i$, respectively. The problem is discretized using 3200 triangular finite elements. Unless otherwise specified, the boundary $\Gamma_o$ is set to $\Gamma_o = \{0,1\}\times(0,1)$, and the number $n$ of measurements on the boundary $\Gamma_o$ is 80. All the computations were performed on a personal computer with 1.00 GB RAM.

The first two examples are taken from [8]. Then only the matrix $\mathbf{W} = \mathbf{L}^{\mathrm{T}}\mathbf{L}$ of the regularizing matrix $\mathbf{L}$ is needed. In our experiments, the matrix $\mathbf{W}$ for Examples 1 and 2 is given by $\mathbf{W}_1 = \mathbf{L}_1^{\mathrm{T}}\mathbf{L}_1$, where $\mathbf{L}_1 \in \mathbb{R}^{(m-1)\times m}$ is the first-order finite difference operator, and $\mathbf{W}_2 = m\mathbf{W}_1 + \mathbf{I}$, where $\mathbf{I} \in \mathbb{R}^{m\times m}$ is the identity matrix, which corresponds to a weighted $H^1$ norm. The resulting PPDF (8) is sampled using the standard Gibbs sampler, and the length of the Markov chain is 200,000 with the first 20,000 realizations discarded as transient states. The mixing of the Markov chain is monitored by visually inspecting the trace plot and calculating the correlation coefficient as in Ref. [8]. We note that a smaller number of MCMC iterations, e.g. $5\times10^4$, can yield a reasonable approximation to the mean [8], but it is insufficient for the variance components as the latter is far more challenging to approximate. The initial guesses for $\mathbf{m}$, $\lambda$ and $\tau$ for the Gibbs sampler are 0, 10 and 100, respectively, and the results are insensitive to these initial guesses. The values of parameter pairs $(\alpha_0,\beta_0)$ and $(\alpha_1,\beta_1)$ are fixed at $(1, 1\times10^{-3})$ and $(1, 1\times10^{-10})$, respectively. These initial guesses are used as $q^0(\lambda)q^0(\tau)$ for `Alg I` and `II` as well. Unless otherwise specified, these parameters apply also to the other two examples.

The synthetic noisy data $\mathbf{d}$ are generated by $d_i = \bar{d}_i + \max_{1\leqslant i\leqslant n}\{|\bar{d}_i|\}\varepsilon\omega_i$, where $\varepsilon$ denotes the relative noise level, and $\omega_i$ are standard normal random variables realized with MATLAB function `randn`. As the noisy data $\mathbf{d}$ depends on the specific noise realization of the random variable $\boldsymbol{\omega}$, the inverse solution differs for different noise realizations. In particular, the regularization parameter $\eta^*$ and the mean value $\mathbf{m}^*$ are realization-dependent. To assess the robustness of the methods under consideration, i.e., the MCMC, `Approx I` and `II`, with respect to noise realization, and to give the reader a feeling of their overall performance in a statistical sense, we shall test 1000 sets of realizations for Examples 1 and 2. We reiterate that `Approx II` coincides with the a-Tikhonov method [9] in terms of point estimate, see Section 3.2, and thus numerical results for the latter are not presented. Quantities of interest, e.g. regularization parameter $\eta^*$, relative reconstruction error $e = \|\mathbf{m}^* - \mathbf{m}^\dagger\|_2/\|\mathbf{m}^\dagger\|_2$ of the mean $\mathbf{m}^*$ and estimated noise level, are first computed for each realization and are then summarized in the form of probability densities estimated via kernel density estimation [8]. All the estimates presented below are defined Section 3, e.g. $\lambda^*$, $(\sigma^2)^* = 1/\tau^*$ and $\eta^* = \lambda^*/\tau^*$, with their superscript $^*$ being frequently omitted for notational simplicity, and the method used for obtaining them will be indicated by subscript, i.e., the subscripts ai, aii and mc are the shorthand notation for `Approx I`, `Approx II` and the MCMC, respectively. The nonlinear extensions in Section 4 are inherently of the type `Approx I` but with some further approximations, e.g. linearization, and will be differentiated by the respective algorithm, i.e., subscripts aiii and aiv for `Alg III` and `Alg IV`, respectively.

To evaluate the accuracy of these methods, we also show the optimal regularization parameter $\eta_{\mathrm{opt}}$, i.e., the one which achieves minimal reconstruction error in Tikhonov regularization, and the corresponding error $e_{\mathrm{opt}}$, for the first three examples. Since there is no known numerically efficient method for computing $\eta_{\mathrm{opt}}$, we have opted for a sampling strategy. More precisely, the optimal $\eta_{\mathrm{opt}}$ is computed as follows: we first evaluate the errors for 200 uniformly distributed values for the regularization parameter in a logarithmic scale in the interval $[1.0\times10^{-16}, 1]$, and then take the one yielding the smallest error as the optimal $\eta_{\mathrm{opt}}$.

*5.2.1. Example 1: smooth solution*

As the first example, we consider the case of reconstructing a smooth solution. The exact solution to the Laplace equation is given by

$$u(x) = \sin(\pi x_1)e^{\pi x_2} + x_1 + x_2, \quad x = (x_1, x_2) \in \Omega.$$

The boundary conditions can be computed straightforwardly.

First we study the robustness of the methods with respect to noise realization. The densities reflect the sampling distribution of point estimates, e.g. $\eta^*$ and $(\sigma^2)^*$. The probability density $p(\eta)$ of the regularization parameter $\eta$ for Example 1 with $\varepsilon = 3\%$ ($\varepsilon$ will be omitted hereafter) noise is shown in Fig. 1(a). The regularization parameter $\eta_{\text{ai}}$ by `Approx I` graphically coincides with that by the MCMC, and is smaller than that by `Approx II`, i.e., $\eta_{\text{aii}}$, however, the discrepancies are within a factor of two. Moreover, the densities are narrowly peaked, and thus both `Approx I` and `II` are robust with respect to noise realization. The mean $\mathbf{m}^*$ of the Bayesian solution is also random, and the probability density $p(e)$ of the reconstruction error $e$ of the mean $\mathbf{m}^*$ is shown in Fig. 1(b). The errors $e_{\text{ai}}$ and $e_{\text{aii}}$ are very similar albeit the latter is slightly smaller. The estimates $\sigma^2_{\text{ai}}$ and $\sigma^2_{\text{aii}}$ have the same magnitude, see Fig. 1(c), and both agree well with the exact variance $\sigma_0^2$. Noting the defining identity $\eta = \lambda\tau^{-1}$, the difference between $\eta_{\text{ai}}$ and $\eta_{\text{aii}}$ is mostly attributed to that between $E_{q^*(\mathbf{m})}\left[\|\mathbf{Lm}\|_2^2\right]$ and $\|\mathbf{Lm}^*\|_2^2$. For all three quantities under consideration, the results by `Approx I` and the MCMC almost always coincide, and thus `Approx I` faithfully captures the true PPDF.

Next we examine the methods for one specific noise realization in detail. Typical numerical results for Example 1 are shown in Figs. 2 and 3 and Table 1, where for the ease of comparison the optimal regularization parameter $\eta_{\text{opt}}$ and the corresponding error $e_{\text{opt}}$ are also presented. The posterior mean $\mathbf{m}_{\text{mc}}$ agrees excellently with the exact solution, and practically identical with $\mathbf{m}_{\text{ai}}$ and $\mathbf{m}_{\text{aii}}$: All three numerical solutions are graphically indistinguishable. The standard deviation std ($\mathbf{m}$) of the numerical solution is shown in Fig. 2(b). std ($\mathbf{m}$) can quantify the uncertainties associated with the mean, and thus can be used to assess the plausibility of a specific solution [13]. The uncertainty bounds, roughly indicated by the estimated variance, given by the MCMC and `Approx I` agree well, however, the latter slightly underestimates. Note that the std ($\mathbf{m}$) curve by the MCMC is jaggy despite the large number of samples used for estimation. Interestingly, the entire covariance structure is very accurately captured, see Fig. 3: The covariance shape estimated by the MCMC and that by `Approx I` almost coincide, although that by the MCMC has some small oscillations along various ridges, which might be attributed to the insufficient number of MCMC samples. The magnitude of the covariance decays rapidly away from diagonal. Note that `Approx II` yields

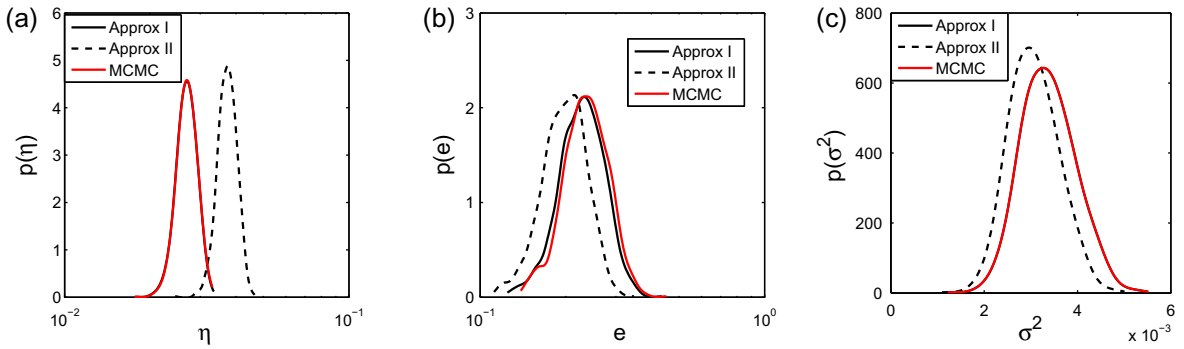


**Fig. 1.** The estimated probability density of (a) regularization parameter $\eta$, (b) accuracy error $e$, and (c) variance $\sigma^2$ for Example 1 with 3% noise.

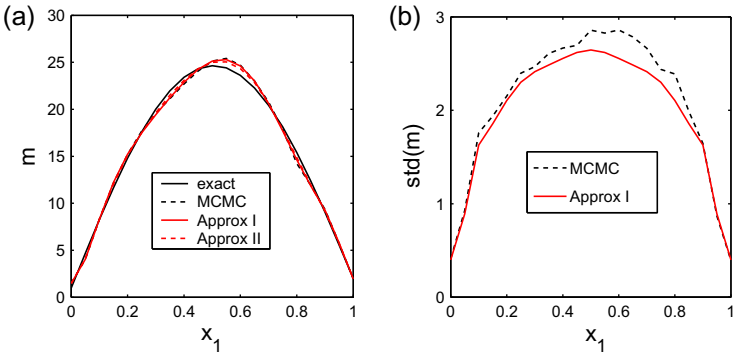


**Fig. 2.** (a) Mean and (b) standard deviation for Example 1 with 3% noise.

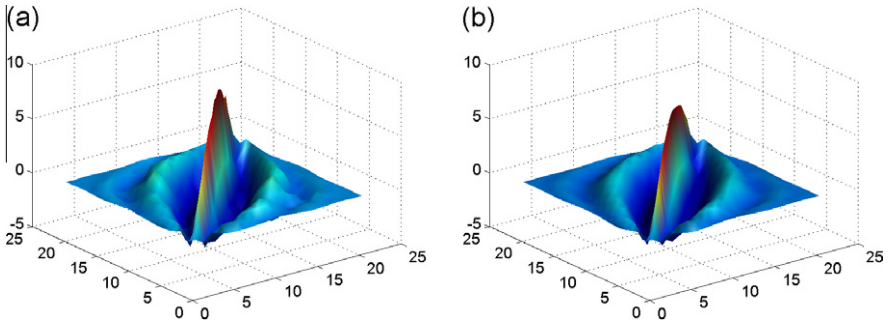


**Fig. 3.** The covariance for Example 1 with 3% noise by (a) MCMC and (b) `Approx I`.

**Table 1**
Numerical results for Example 1.

| $\varepsilon$ (%) | $\sigma_0$ | $\sigma_{mc}$ | $\sigma_{ai}$ | $\sigma_{aii}$ | $\eta_{mc}$ | $\eta_{ai}$ | $\eta_{aii}$ | $\eta_{opt}$ |
|---|---|---|---|---|---|---|---|---|
| 1 | 1.97e−2 | 2.10e−2 | 2.07e−2 | 1.96e−2 | 3.70e−5 | 3.58e−5 | 6.64e−5 | 1.43e−3 |
| 3 | 5.91e−2 | 6.28e−2 | 6.20e−2 | 5.90e−2 | 3.01e−4 | 2.87e−4 | 6.04e−4 | 3.83e−3 |
| 5 | 9.84e−2 | 1.05e−1 | 1.03e−1 | 9.86e−2 | 7.83e−4 | 7.52e−4 | 1.71e−3 | 5.67e−3 |
| | $e_{mc}$ | $e_{ai}$ | $e_{aii}$ | $e_{opt}$ | $\lambda_{mc}$ | $\lambda_{ai}$ | $\lambda_{aii}$ | $\lvert\rho_{\mathbf{m},\tau}\rvert_\infty$ | $\lvert\rho_{\mathbf{m},\lambda}\rvert_\infty$ |
| 1 | 2.54e−2 | 2.31e−2 | 1.85e−2 | 6.56e−3 | 8.37e−2 | 8.32e−2 | 1.73e−1 | 3.66e−2 | 1. 69e−1 |
| 3 | 3.44e−2 | 3.13e−2 | 2.27e−2 | 1.05e−2 | 7.63e−2 | 7.46e−2 | 1.74e−1 | 2.70e−2 | 1. 44e−1 |
| 5 | 3.29e−2 | 3.34e−2 | 2.11e−2 | 1.38e−2 | 7.15e−2 | 7.07e−2 | 1.76e−1 | 3.25e−2 | 1. 34e−1 |

no uncertainty estimate for the inverse solution $\mathbf{m}^*$. The variance estimate for the inverse solution $\mathbf{m}$, either by `Approx I` or MCMC, shrinks as the noise level $\varepsilon$ decreases, i.e., the probability bound sharpens accordingly. The computing times with MATLAB 7.0 required for `Approx I`, `Approx II` and the MCMC are 0.32 s, 0.15 s and 203 s, respectively. This clearly shows the computational efficiency of the variational method over the MCMC.

The estimates $\sigma_{mc}$, $\sigma_{ai}$ and $\sigma_{aii}$ agree well with the exact variance $\sigma_0$. The first two are identical within sampling error, and are less accurate than $\sigma_{aii}$, which concurs with previous numerical observations [8]. The probability bounds by the MCMC, `Approx I` and `II` are fairly close. For instance, in case of 3% noise, the estimated std ($\sigma$) by the three methods is $6.65 \times 10^{-4}$, $6.31 \times 10^{-4}$ and $5.71 \times 10^{-4}$, respectively. Both `Approx I` and `II` slightly under-estimate the uncertainties, however, the latter is more severe. The accuracy errors $e_{mc}$, $e_{ai}$ and $e_{aii}$ are still comparable albeit `Approx II` is marginally more accurate for all three noise levels, and all are comparable with optimal value $e_{opt}$. Also observe that the estimates $\lambda_{mc}$ and $\lambda_{ai}$ are very close, and their magnitude is about half of $\lambda_{aii}$. However, the estimated std ($\lambda$) differs markedly. For instance, in case of 3% noise, std ($\lambda$) by these three methods is $3.28 \times 10^{-2}$, $2.20 \times 10^{-2}$ and $5.12 \times 10^{-2}$, respectively, and thus the uncertainty bounds by `Approx I` and `II` are imprecise. The estimated value of $\lambda$ seems relatively independent of the noise level for all the methods, and thus the regularization parameter $\eta$ is of order $\sigma_0^2$. This can cause under-regularization in case of low noise levels in accordance with classical inverse theory and Theorem 3.2, although it is not conspicuous for severely ill-posed problems. Strategies that adapt automatically to the noise level might be necessary for arbitrarily varying noise levels, see Ref. [24] for such a strategy in the context of the a-Tikhonov method. The excellent agreement between results by the MCMC and that by `Approx I` indicates that `Approx I` captures faithfully the PPDF, and thus its guidelines and mathematical underpinnings may also apply to the hierarchical formulation.

The fundamental assumption of the variational method is conditional independence of $\mathbf{m}$ and $(\lambda, \tau)$ given the data $\mathbf{d}$. The invalidity of this assumption is expected to render the approximations inaccurate or even useless. To interrogate the assumption for this example, we calculate the correlation coefficient vectors $\rho_{\mathbf{m},\tau}$ and $\rho_{\mathbf{m},\lambda}$ between the vector $\mathbf{m}$ and $\lambda$ and $\tau$, respectively, from the MCMC samples. The correlation coefficient (vector) between the vector $\mathbf{m}$ and scalars are computed componentwise, i.e., with $m_i$. The results are shown in Fig. 4, where the abscissa $i$ denotes the $i$th component. Overall, the correlation coefficients between $\mathbf{m}$ and $\tau$ are small with a maximum norm $|\rho_{\mathbf{m},\tau}|_\infty$ smaller than 0.05 for all three noise levels, and thus the assumption seems valid. The correlation coefficients between $\mathbf{m}$ and $\lambda$ are slightly larger with the maximum norm $|\rho_{\mathbf{m},\lambda}|_\infty$ close to 0.18, see also Table 1. Therefore, the correlation between $\mathbf{m}$ and $(\lambda, \tau)$ is indeed relatively weak. This partially validates the conditional independence assumption and explain the excellent agreement between the covariances by the MCMC and that by `Approx I` observed in Fig. 3.

The convergence of `Alg I` and `II` is shown in Fig. 5(a) and (b), respectively. We show only the scalars $\lambda$, $\tau$ and $\eta$ because the convergence of the algorithms are fully determined by these quantities. To indicate quantitatively the convergence of these two algorithms, we calculate the asymptotic convergence rate $r^* := \lim_{k\to\infty} \frac{\eta_k - \eta^*}{\eta_{k-1} - \eta^*}$ for the regularization parameter, which can be empirically calculated from the sequence $\{\eta_k\}_k$. In case of 1%, 3% and 5% noise, the empirical convergence rate

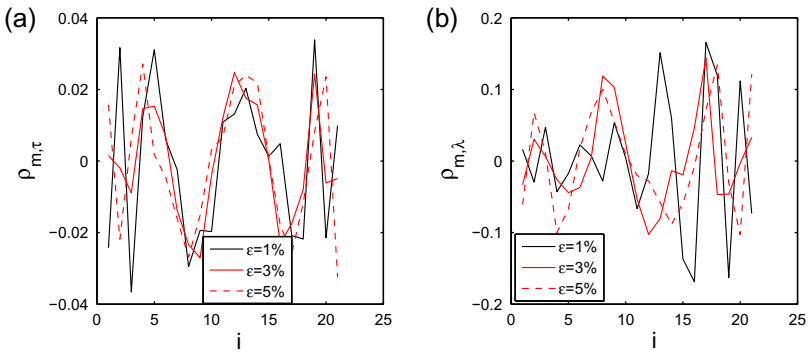


**Fig. 4.** The correlation coefficient between **m** and (a) $\tau$ and (b) $\lambda$ for Example 1.

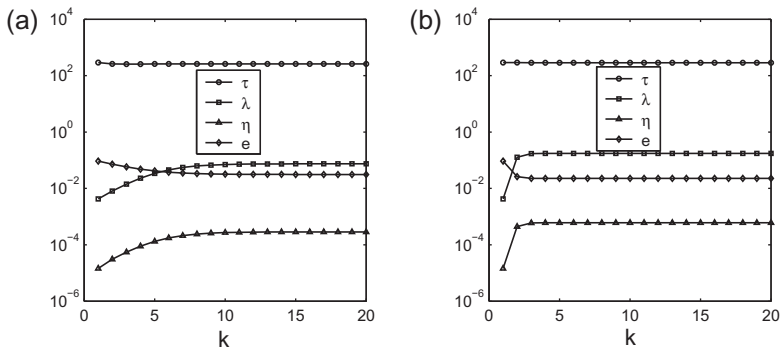


**Fig. 5.** Convergence of `Alg` (a) `I` and (b) `II` for Example 1 with 3% noise.

$r^*$ of `Alg I` is calculated to be $5.02 \times 10^{-1}$, $5.51 \times 10^{-1}$ and $5.79 \times 10^{-1}$, respectively, whereas that for `Alg II` is $1.22 \times 10^{-2}$, $2.50 \times 10^{-2}$ and $3.39 \times 10^{-2}$, respectively. Therefore, `Alg II` converges much faster than `I`, which is also clear from Fig. 5. The convergence of both algorithms deteriorates with the increase of the noise level, but the effect on `Alg I` is less substantial than that on `Alg II`. Interestingly, a striking convergence within one iteration is observed for the estimate $\sigma_{\text{aii}}$. The convergence for $\sigma_{\text{ai}}$ is slightly slower, i.e., convergence within three iterations. Nonetheless, it is still much faster than that for other quantities. This substantiates the working assumption of fixing $\tau$ at $\sigma_0^{-2}$ in Theorem 3.4. Theorem 3.4 predicts a local monotonic convergence of the sequence $\{\eta_k\}_k$, which is also corroborated by the numerical results.

Finally, we study the sensitivity of the numerical results with respect to the prior parameter pairs $(\alpha_0, \beta_0)$ and $(\alpha_1, \beta_1)$. Our analysis in Section 3 indicates that a parameter pair $(\alpha_1, \beta_1)$ with $\alpha_1 = 1$ and $\beta_1$ close to zero is sufficient, but it gives no explicit indication on $(\alpha_0, \beta_0)$. Recall the defining relation for $\eta$ of `Approx II`, i.e., $\eta = \frac{\alpha_0''}{\frac{1}{2}\|\mathbf{L}\mathbf{m}_\eta\|_2^2 + \beta_0}\sigma^2$. Now if the value of $\alpha_0$ not too large, it is dominated by $\frac{m}{2}$ in the numerator $\alpha_0'' = \frac{m}{2} + \alpha_0$, and thus can only have marginal impact on the inverse solution. Therefore, we conduct the sensitivity analysis only for $\beta_0$. The results are presented in Table 2. Large values of $\beta_0$, e.g. 1000, overwhelm the term $\frac{1}{2}\|\mathbf{L}\mathbf{m}_\eta\|_2^2$, and thus the latter would have little impact on the solution procedure. Then `Approx II` is tantamount to minimizing the residual, which renders the inverse solution under-regularized, see the last two rows of Table 2. Strikingly, the estimate $(\sigma^2)^*$ remains accurate. The term $\frac{1}{2}\|\mathbf{L}\mathbf{m}_\eta\|_2^2$ comes into play as the value of $\beta_0$ decreases, and the solution is insensitive to its actual value as long as it is small enough. This is confirmed by the observation that the results, e.g. error $e$, are practically identical for $\beta_0$ varying over the range $[1 \times 10^{-1}, 1 \times 10^{-7}]$. These heuristic analysis and

**Table 2**
Numerical results for Example 1 with 3% noise.

| $\beta_0$ | $\sigma_{\text{ai}}$ | $\sigma_{\text{aii}}$ | $\eta_{\text{ai}}$ | $\eta_{\text{aii}}$ | $e_{\text{ai}}$ | $e_{\text{aii}}$ | $\lambda_{\text{ai}}$ | $\lambda_{\text{aii}}$ | $\lvert\rho_{\mathbf{m},\tau}\rvert_\infty$ | $\lvert\rho_{\mathbf{m},\lambda}\rvert_\infty$ |
|---|---|---|---|---|---|---|---|---|---|---|
| 1e−7 | 6.20e−2 | 5.90e−2 | 2.87e−4 | 6.04e−4 | 3.13e−2 | 2.27e−2 | 7.46e−2 | 1.74e−1 | 3. 32e−2 | 1.86e−1 |
| 1e−5 | 6.20e−2 | 5.90e−2 | 2.87e−4 | 6.04e−4 | 3.13e−2 | 2.27e−2 | 7.46e−2 | 1.74e−1 | 2. 38e−2 | 1.29e−1 |
| 1e−3 | 6.20e−2 | 5.90e−2 | 2.87e−4 | 6.04e−4 | 3.13e−2 | 2.27e−2 | 7.46e−2 | 1.74e−1 | 2. 70e−2 | 1.44e−1 |
| 1e−1 | 6.20e−2 | 5.90e−2 | 2.86e−4 | 6.03e−4 | 3.13e−2 | 2.27e−2 | 7.45e−2 | 1.73e−1 | 3. 60e−2 | 8.74e−2 |
| 1e1 | 6.20e−2 | 5.89e−2 | 2.51e−4 | 5.23e−4 | 3.29e−2 | 2.42e−2 | 6.52e−2 | 1.50e−1 | 3. 25e−2 | 1.25e−1 |
| 1e3 | 6.24e−2 | 5.87e−2 | 2.05e−5 | 3.68e−5 | 8.40e−2 | 6.81e−2 | 5.27e−3 | 1.07e−2 | 5. 90e−2 | 1.88e−1 |

numerical observations hold also for Approx I, see Table 2. Finally, we remark that the conditional independence assumption seems relatively independent of the parameter $\beta_0$ for the magnitude of the correlation coefficients $|\rho_{\mathbf{m},\lambda}|_\infty$ and $|\rho_{\mathbf{m},\tau}|_\infty$ remains almost unchanged as the value of $\beta_0$ varies, see the last two columns of Table 2.

*5.2.2. Example 2: nonsmooth solution*

As the second example, we consider a more challenging case of reconstructing a nonsmooth solution. The boundary conditions are prescribed as follows:

$$u(x) = \begin{cases} 2x_1, & x \in [0,\frac{1}{2}] \times \{1\} \\ 2 - 2x_1, & x \in [\frac{1}{2},1] \times \{1\} \end{cases} \quad \text{and} \quad \frac{\partial u(x)}{\partial n} = 1, \quad x \in \Gamma_c.$$

The exact solution to the Laplace equation is unavailable, and thus the numerical solution to the forward problem on a finer mesh is taken as the exact data.

As before, we first investigate the robustness of the methods with respect to noise realization. The probability density $p(\eta)$ for Example 2 is shown in Fig. 6(a). The distributions of both $\eta_{\text{ai}}$ and $\eta_{\text{aii}}$ are narrowly peaked and thus insensitive to noise realizations, but the magnitude of former is smaller. In spite of the apparent discrepancies, the differences between the errors $e_{\text{ai}}$ and $e_{\text{aii}}$ are insignificant. The estimates $\sigma_{\text{ai}}$ and $\sigma_{\text{aii}}$ are almost identical. Therefore, the marked difference between $\eta_{\text{ai}}$ and $\eta_{\text{aii}}$ is attributed to that between $\|\mathbf{L}\mathbf{m}^*\|_2^2$ and $E_{q^*(\mathbf{m})}\left[\|\mathbf{L}\mathbf{m}\|_2^2\right]$, which in turn boils down to the dominance of the variance component over the mean component $\|\mathbf{L}\mathbf{m}^*\|_2^2$ in the term $E_{q^*(\mathbf{m})}\left[\|\mathbf{L}\mathbf{m}\|_2^2\right]$. An excellent agreement of the results by Approx I and the MCMC is again observed, and they graphically almost coincide with each other, as in Example 1.

Exemplary numerical results for Example 2 are shown in Figs. 7 and 8 and Table 3. The posterior mean $\mathbf{m}_{\text{mc}}$ agrees reasonably well with the exact solution, taking into account the nonsmoothness. However, the sharp corner is not accurately resolved because of the smoothing nature of the prior. The numerical reconstructions $\mathbf{m}_{\text{ai}}$ and $\mathbf{m}_{\text{aii}}$ are practically identical in terms of the error $e$, however, the profiles of $\mathbf{m}_{\text{ai}}$ and $\mathbf{m}_{\text{aii}}$ differ significantly since Approx II selects a larger regularization parameter. Interestingly, the errors $e_{\text{ai}}$ and $e_{\text{aii}}$ are close to each other, which is attributed to the fact that the nonsmooth part is poorly approximated and the corresponding error dominates. The probability bound is also adversely affected by the nonsmoothness and not so sharp as for Example 1. Moreover, Approx I under estimates std ($\mathbf{m}$) by about 20% compared with the MCMC result, nonetheless, the shape of the covariance structure is still resolved accurately, see Fig. 8. The mechanism of

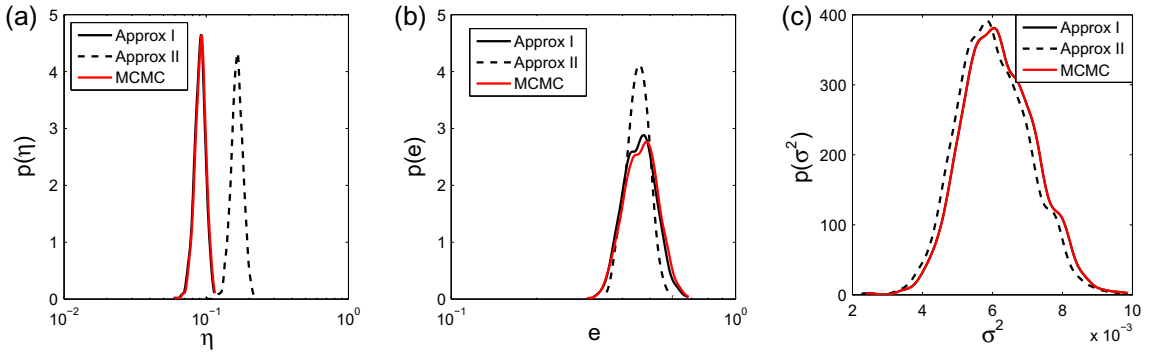


**Fig. 6.** The estimated probability density of (a) regularization parameter $\eta$, (b) accuracy error $e$, and (c) variance $\sigma^2$ for Example 2 with 3% noise.

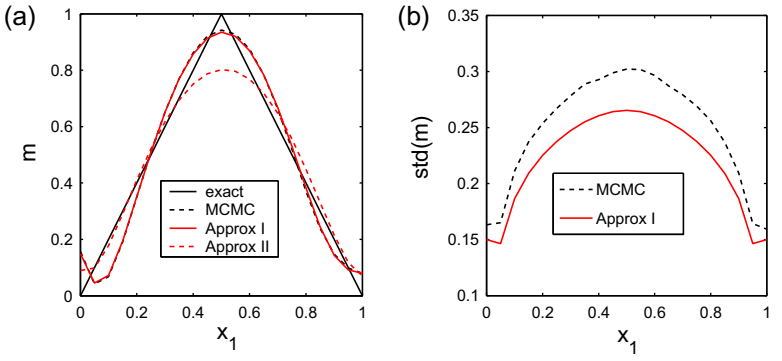


**Fig. 7.** (a) Mean and (b) standard deviation for Example 2 with 3% noise.

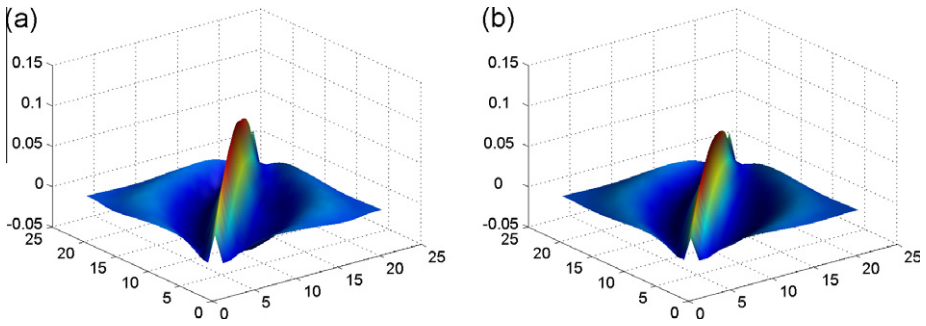


**Fig. 8.** The covariance for Example 2 with 3% noise by (a) MCMC and (b) `Approx I`.

**Table 3**
Numerical results for Example 2.

| $\varepsilon$ (%) | $\sigma_0$ | $\sigma_{mc}$ | $\sigma_{ai}$ | $\sigma_{aii}$ | $\eta_{mc}$ | $\eta_{ai}$ | $\eta_{aii}$ | $\eta_{opt}$ |
|---|---|---|---|---|---|---|---|---|
| 1 | 2.66e−2 | 2.82e−2 | 2.87e−2 | 2.70e−2 | 5.78e−4 | 5.54e−4 | 1.62e−3 | 2.12e−3 |
| 3 | 7.99e−2 | 8.46e−2 | 8.34e−2 | 8.29e−2 | 4.36e−3 | 3.98e−3 | 1.80e−2 | 8.41e−3 |
| 5 | 1.33e−1 | 1.41e−1 | 1.40e−1 | 1.41e−1 | 1.23e−2 | 1.06e−2 | 7.07e−2 | 1.52e−2 |
| | $e_{mc}$ | $e_{ai}$ | $e_{aii}$ | $e_{opt}$ | $\lambda_{mc}$ | $\lambda_{ai}$ | $\lambda_{aii}$ | $\lvert\rho_{\mathbf{m},\tau}\rvert_\infty$ | $\lvert\rho_{\mathbf{m},\lambda}\rvert_\infty$ |
| 1 | 9.36e−2 | 9.39e−2 | 8.63e−2 | 8.59e−2 | 7.27e−1 | 7.16e−1 | 2.23e0 | 3.65e−2 | 1.20e−1 |
| 3 | 1.26e−1 | 1.23e−1 | 1.16e−1 | 9.52e−2 | 6.07e−1 | 5.72e−1 | 2.62e0 | 6.71e−2 | 2.04e−1 |
| 5 | 1.21e−1 | 1.13e−1 | 2.26e−1 | 1.06e−1 | 6.11e−1 | 5.46e−1 | 3.53e0 | 7.65e−2 | 2.43e−1 |

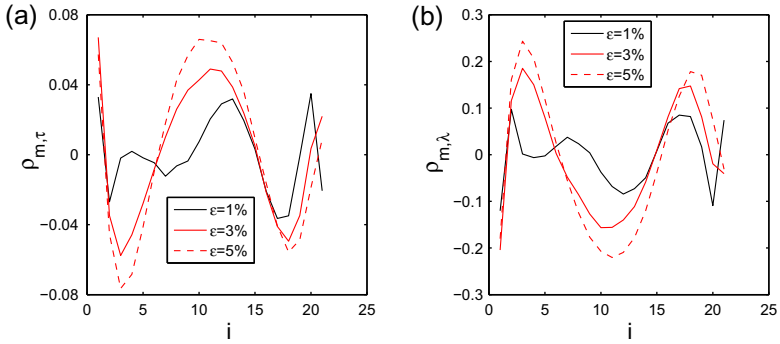


**Fig. 9.** The correlation coefficient between **m** and (a) $\tau$ and (b) $\lambda$ for Example 2.

under-estimation in the variance estimation remains unclear, and we refer to [34] for some discussions. The variance of **m** decreases as the noise level decreases albeit more slowly. The correlation coefficients between **m** and $\tau$ and $\lambda$ seem a bit larger, and tend to increase slightly as the noise level increases, see Fig. 9.

The estimates $\sigma_{mc}$, $\sigma_{ai}$ and $\sigma_{aii}$ agree well with the exact variance $\sigma_0$. The reconstruction errors $e_{mc}$ and $e_{ai}$ are very close to each other and both are still close to $e_{opt}$, the optimal one, and $e_{aii}$ suffers from over-regularization in case of 5% noise. Observe that the estimates $\lambda_{mc}$ and $\lambda_{ai}$ are close for all three noise levels and decrease slightly with the increase of the noise level, whereas $\lambda_{aii}$ increases mildly. Consequently, $\eta_{mc}$ and $\eta_{ai}$ seem of order $\sigma_0^2$, while $\eta_{aii}$ tends to decay at a rate faster than $\sigma_0^2$, which is not close to the optimal value in all three cases. This accounts for its severe over-regularization in case of 5% noise. Therefore, `Approx I` is less sensitive to the variation of noise level than `Approx II`. Interestingly, $\sigma_{aii}$ still represents an excellent estimate of $\sigma_0$. The numerical results are insensitive to the value of the parameter $\beta_0$ so long as its value is sufficiently small, see Table 4. `Approx I` yields comparable results with the MCMC, and thus it captures better the PPDF than `Approx II`. Also the correlation coefficients seem relatively independent of $\beta_0$ so long as its value is small.

The convergence of the algorithms is shown in Fig. 10. The empirical convergence rate $r^*$ of `Alg I` is $6.52 \times 10^{-1}$, $7.44 \times 10^{-1}$ and $7.84 \times 10^{-1}$, respectively, in case of 1%, 3% and 5% noise, whereas that of `Alg II` is $7.43 \times 10^{-2}$, $2.19 \times 10^{-1}$ and $3.89 \times 10^{-1}$, respectively. Therefore, the convergence rate for both algorithms is adversely affected by increasing the noise level. Also the nonsmoothness has a deleterious effect on the convergence, and its effect on `Alg II` is more significant than on `Alg I`. A striking convergence with one iteration is again observed for the scalar $\sigma_{aii}$.

**Table 4**
Numerical results for Example 2 with 3% noise.

| $\beta_0$ | $\sigma_{ai}$ | $\sigma_{aii}$ | $\eta_{ai}$ | $\eta_{aii}$ | $e_{ai}$ | $e_{aii}$ | $\lambda_{ai}$ | $\lambda_{aii}$ | $\lvert\rho_{\mathbf{m},\tau}\rvert_\infty$ | $\lvert\rho_{\mathbf{m},\lambda}\rvert_\infty$ |
|---|---|---|---|---|---|---|---|---|---|---|
| 1e−7 | 8.35e−2 | 8.29e−2 | 4.02e−3 | 1.80e−2 | 1.22e−1 | 1.16e−1 | 5.77e−1 | 2.62e0 | 6. 55e−2 | 2.07e−1 |
| 1e−5 | 8.35e−2 | 8.29e−2 | 4.02e−3 | 1.80e−2 | 1.22e−1 | 1.16e−1 | 5.77e−1 | 2.62e0 | 6. 51e−2 | 2.13e−1 |
| 1e−3 | 8.35e−2 | 8.29e−2 | 4.02e−3 | 1.80e−2 | 1.22e−1 | 1.16e−1 | 5.77e−1 | 2.62e0 | 6. 71e−2 | 2.04e−1 |
| 1e−1 | 8.35e−2 | 8.28e−2 | 3.94e−3 | 1.75e−2 | 1.23e−1 | 1.14e−1 | 5.66e−1 | 2.55e0 | 6. 88e−2 | 2.04e−1 |
| 1e1 | 8.33e−2 | 8.13e−2 | 1.47e−3 | 4.90e−3 | 4.94e−1 | 1.11e−1 | 2.12e−1 | 7.42e−1 | 4. 86e−2 | 1.64e−1 |
| 1e3 | 8.38e−2 | 8.00e−2 | 3.20e−5 | 7.18e−5 | 8.42e−1 | 5.46e−1 | 4.56e−3 | 1.12e−2 | 2. 61e−2 | 1.34e−1 |

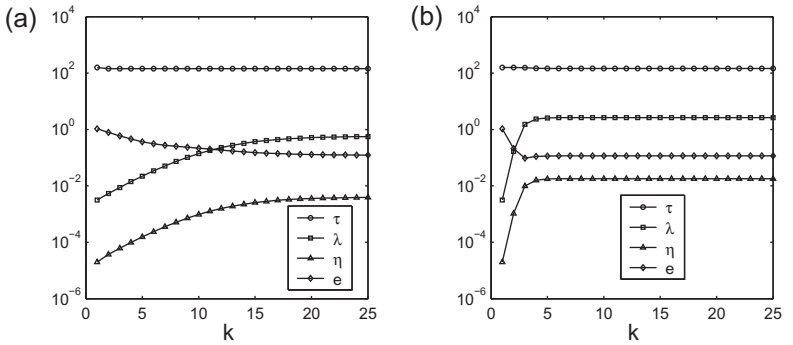


**Fig. 10.** Convergence of `Alg` (a) `I` and (b) `II` for Example 2 with 3% noise.

*5.2.3. Example 3: $\ell^r$ prior*

This example is adapted from Example 1, but with an $\ell^r$ prior, i.e. $p(\mathbf{m}|\lambda) \propto \lambda^{\frac{m}{r}} e^{-\lambda\|\mathbf{m}\|_r^r}$. We fix $r = 1.9$, which is nonquadratic, and thus the popular Gibbs sampler is not directly applicable. For this example, the parameter pair $(\alpha_0, \beta_0)$ is taken to be $(1 \times 10^3, 1 \times 10^{-3})$, and the auxiliary vector $\mathbf{v}$ is initialized with all entries equal to one.

The numerical results for Example 3 with various noise levels by `Alg III` and the MCMC are summarized in Table 5. The standard Metropolis–Hastings algorithm converges very slowly. To alleviate the problem, we employed a blocking strategy: update $\mathbf{m}$ only three sites each time with the Metropolis–Hastings algorithm and cycle over all the sites, and update $\lambda$ and $\tau$ with the Gibbs sampler. The proposal distribution in the algorithm is a symmetric Gaussian random walk sampler with standard deviation 0.1. The chain is run to a length of $5 \times 10^5$, and the last $4 \times 10^5$ samples are used for estimating relevant statistical quantities.

The means of the two approximations are fairly close to each other. Since the value of exponent $r$ is close to 2, we expect solutions similar to that of Example 1. This is numerically observed from Fig. 11(a). The standard deviation by MCMC is slightly larger than that by `Approx III`, see Fig. 11(b). However, there are many not-so-small oscillations throughout the covariance by the MCMC, see Fig. 12, despite the large number of samples used for estimation.

The results in Table 5 indicate that the automatically determined reconstruction by the hierarchical formulation is close to the optimal choice. For a graphical illustration, we refer to Fig. 11(a) for the mean, and 11(b) for the respective covariance. `Alg III` merits a steady and fast convergence as `Alg I` and `II` and in practice the convergence is achieved within five iterations, see Fig. 11(c). Ten iterations of the algorithm take 0.109 s, whereas the MCMC ($5 \times 10^5$ iterations) takes 385 s. This well illustrates the versatility of the variational method for approximate Bayesian inference.

*5.2.4. Example 4: nonlinear model*

The final example is a nonlinear inverse problem using Cauchy data, which arises in corrosion detection [30,32]. Now the goal is to estimate the Robin coefficient $\gamma(x)$ of a Robin boundary condition on $\Gamma_i$, i.e.,

$$\alpha \frac{\partial u}{\partial n} + \gamma u = \gamma u_a \quad x \in \Gamma_i.$$

**Table 5**
Numerical results for Example 3.

| $\varepsilon$ (%) | $\sigma_0$ | $\sigma_{aiii}$ | $\sigma_{mc}$ | $\eta_{aiii}$ | $\eta_{mc}$ | $\eta_{opt}$ | $e_{aiii}$ | $e_{mc}$ | $e_{opt}$ | $\lambda_{aiii}$ | $\lambda_{mc}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1.97e−2 | 2.09e−2 | 2.10e−2 | 9.45e−5 | 9.47e−5 | 7.49e−5 | 1.55e−2 | 1.80e−2 | 1. 53e−2 | 2.15e−1 | 2.15e−1 |
| 3 | 5.91e−2 | 6.57e−2 | 6.58e−2 | 9.35e−4 | 9.38e−4 | 2.43e−4 | 3.56e−2 | 3.71e−2 | 2. 65e−2 | 2.16e−1 | 2.16e−1 |
| 5 | 9.84e−2 | 1.20e−2 | 1.20e−1 | 3.15e−3 | 3.21e−3 | 5.35e−4 | 6.00e−2 | 6.11e−2 | 3. 42e−2 | 2.19e−1 | 2.19e−1 |

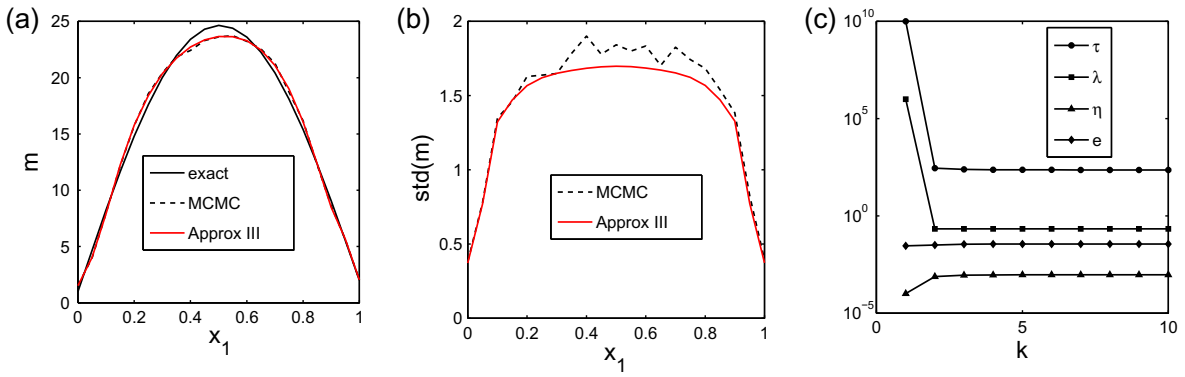


**Fig. 11.** Numerical results for Example 3 with 3% noise: (a) mean, (b) standard deviation and (c) convergence of `Alg III`.

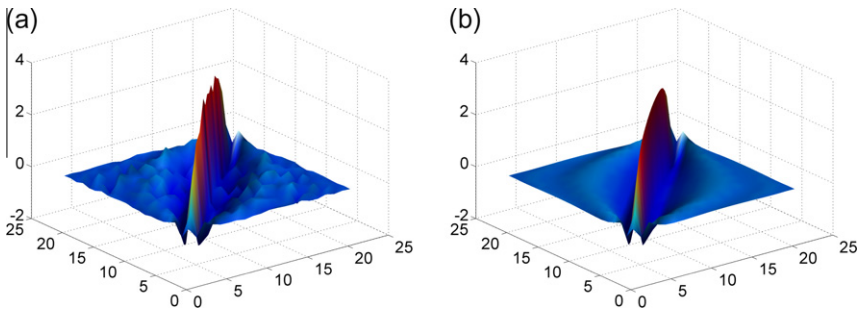


**Fig. 12.** The covariance for Example 3 with 3% noise by (a) MCMC and (b) `Approx III`.

Here $\frac{\partial u}{\partial n} = -1$ on $\Gamma_c$, $u_a = 0$ and $f = -4$. For this example, we set $\Gamma_o = \Gamma_c$. The Robin coefficient to be estimated is $\gamma = 1 + x_1$. As before, we discretize the coefficient $\gamma(x)$ with a mesh size $h = 0.05$, which gives a vector $\mathbf{m}$ of *21* scalar parameters. The regularization matrix $\mathbf{L}$ is the second-order finite difference operator. The parameter pair $(\alpha_0, \beta_0)$ is set to $(1 \times 10^2, 1 \times 10^{-3})$. The initial guess for `Alg IV` is set to *2*.

For this example, the standard Metropolis–Hastings algorithm together with a simple blocking strategy and a random walk proposal distribution suffers from very slow convergence. There are several other techniques that may improve the mixing the MCMC chain, e.g. prior preconditioning strategy in [35] and delayed rejection-adaptive Metropolis sampler (DRAM) [36]. The former combines a (checkerboard) blocking strategy with proposals from the prior (either directly or based on pivoted Chelosky decomposition), whereas the latter integrates the idea of multiple-stage proposal with utilizing accumulated knowledge of the PPDF from sample paths. Instead of employing these techniques, here we have opted for an intelligently designed MCMC algorithm to obtain a reference Bayesian solution. Firstly, we utilize the variational approximation as the proposal distribution (independent sampler) in the Metropolis–Hastings algorithm. Secondly, we accelerate the MCMC sampling by reduced-order modeling via proper orthogonal decomposition (POD) [32], which is one of the fastest algorithms available. The snapshots for generating the POD basis are taken at 5000 uniformly distributed points in the hypercube $[0.1, 3]^{21}$. For the inversion, 30 POD basis functions are employed, which gives an average relative error of $1.28 \times 10^{-6}$ at 1000 random test points in the hypercube and thus is deemed sufficient. In order to shorten the burning period for the MCMC chain, we took the mean of the variational approximation as the initial guess for $\mathbf{m}$. The chain is run to a length of $5 \times 10^4$, the acceptance ratio is 80%, and the last $4 \times 10^4$ samples are used for estimation. A much longer chain does not improve the result greatly. The computing time for the MCMC and the variational method is $1.61 \times 10^2$ s ($5 \times 10^4$ iterations) and 11 s (20 iterations), respectively, where for the MCMC the part of calculating the POD basis is excluded.

The numerical results for Example 4 with 3% noise in the data by `Alg IV` are shown in Figs. 13 and 14. The mean by the MCMC and the variational method almost coincides with each other and both are in good agreement with the exact one, with the respective reconstruction error $e_{\mathrm{mc}} = 1.93 \times 10^{-2}$ and $e_{\mathrm{aiv}} = 1.97 \times 10^{-2}$. The variational method estimates also excellently the standard deviation std($\mathbf{m}$) relative to the MCMC. Especially, the covariance shapes are almost identical, see Fig. 14. Note that the covariance by the MCMC is free from spurious oscillations, although the number of MCMC samples is not large. In particular, this shows that the variational method is a promising preconditioning technique to the MCMC for handling high-dimensional problems. Similar to other algorithms, `Alg IV` converges steadily and quickly, and the convergence is achieved within 10 iterations. In `Alg IV`, the inverse variance $\tau$ converges much faster than other quantities,

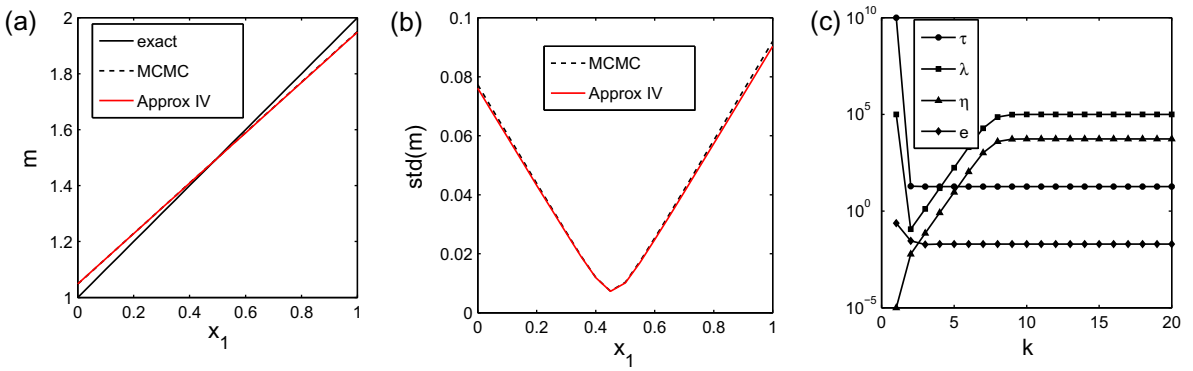


**Fig. 13.** Numerical results for Example 4 with 3% noise: (a) mean, (b) standard deviation, and (c) convergence of `Alg IV`.

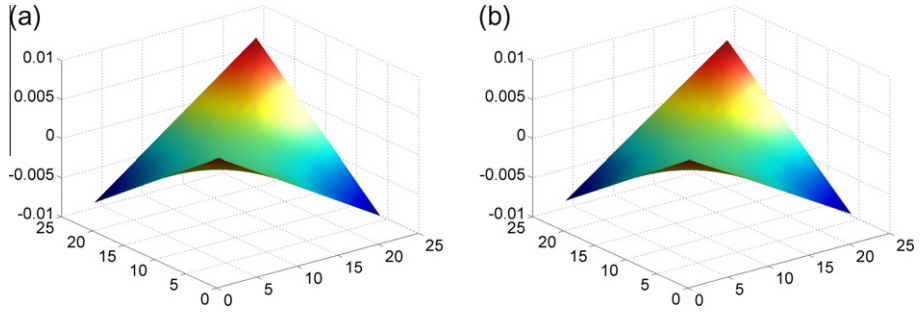


**Fig. 14.** The covariance for Example 4 with 3% noise in the data: (a) MCMC and (b) `Alg IV`.

and its convergence is practically achieved within three iterations. The estimated noise levels $\sigma_{\text{aiv}} = 2.32 \times 10^{-1}$ and $\sigma_{\text{mc}} = 2.32 \times 10^{-1}$ are both in good agreement with the exact one $\sigma_0 = 2.68 \times 10^{-1}$.

## 6. Concluding remarks

We have investigated the variational method for approximating hierarchical Bayesian formulations of ill-posed problems. The idea is to approximate the PPDF by conditionally independent distributions of the solution and parameters in the Kullback–Leibler divergence. Two approximate PPDF are derived within the framework. The existence of a minimizer to the functionals is established, properties of the minimizers are studied, and some heuristic guidelines for specifying the prior parameter pairs are provided. Alternating iterative algorithms are proposed for minimizing the functionals, and their convergence properties are discussed. Extensions of the framework to nonlinear problems, either nonGaussian priors or nonlinear forward models, have also been discussed. Numerical results for Cauchy type problems in heat transfer indicate that both `Approx I` and `II` can estimate accurately the inverse solution and the hyper-parameters with their uncertainties quantified, and the algorithms feature a fast and steady convergence. Moreover, `Approx I` and its variants can capture faithfully main features of the PPDF, and they provide valuable approximate solutions to nontrivial Bayesian inference problems. In addition, the approximation can serve as an effective promising preconditioner for the MCMC.

There are several avenues deserving further research. Firstly, the proximity of the variational approximations to the PPDF was assessed using the MCMC, which is often expensive. Therefore, it is of paramount importance to derive computable error bounds and to develop efficient numerical methods to estimate the errors. Secondly, the convergence rate of the algorithms remains to be established despite the steady and fast convergence numerically observed, which will shed new insights into their pros and cons and designing acceleration strategies. Thirdly, the extension of the framework to a functional analytic setting, e.g. implicit operator formulations involved in nonlinear inverse problems and nonstandard regularization terms, is of interest. In particular, although the generalizations of the variational method to general $\ell^r$ priors and nonlinear forward models have been briefly discussed, a refined theoretical investigation and more extensive numerical verifications would be very helpful. Finally, some engineering problems, e.g. super-resolution and hyperspectral imaging, call for complex regularization formulations involving multiple data fitting terms and/or regularization terms due to heteroscedastic nature of the data and multiscale feature of the solution. The development of the variational framework for these more general formulations is also impending.

**Acknowledgements**

The authors thank the editor guiding the paper, Professor George Em Karniadakis, and three anonymous referees for their many constructive comments which have led to a significant improvement of the quality and the presentation of the paper. This work was partially carried out during a visit of Bangti Jin at Department of Mathematics, Chinese University of Hong Kong and partially supported by a Direct Grant for Research 2009/2010 from CUHK. He would like to thank Professor Jun Zou for the hospitality. The work of the first author is supported by the Alexander von Humboldt foundation through a post-doctoral researcher fellowship, and the work of the second author was substantially supported by Hong Kong RGC Grants (Project 404407) and partially supported by CUHK Focused Investment Scheme 2008/2010.

## Appendix A. Proofs of Theorems 3.2 and 3.3

### A.1. Proof of Theorem 3.2

For fixed $\tau = \sigma_0^{-2}$, it follows from (13) that $\eta^* = \lambda^* \sigma_0^2$ satisfies

$$\eta^* \left[ \|\mathbf{L}\mathbf{m}_{\eta^*}\|_2^2 + \mathrm{tr}((\mathbf{H}^{\mathrm{T}}\mathbf{H} + \eta^* \mathbf{L}^{\mathrm{T}}\mathbf{L})^{-1}\mathbf{L}^{\mathrm{T}}\mathbf{L})\sigma_0^2 + 2\beta_0 \right] = 2\alpha_0'' \sigma_0^2. \tag{20}$$

We also recall the generalized singular value decomposition [22]. For any pair of matrices $\mathbf{H} \in \mathbb{R}^{n\times m}$ and $\mathbf{L} \in \mathbb{R}^{p\times m}$ with $n \geqslant m \geqslant p$ and rank $(\mathbf{L}) = p$, there exists

$$\mathbf{H} = \mathbf{U}\begin{pmatrix} \Sigma & 0 \\ 0 & \mathbf{I}_{m-p} \end{pmatrix}\mathbf{X}^{-1}, \quad \mathbf{L} = \mathbf{V}(\mathbf{M}\ \mathbf{0}_{p\times(m-p)})\mathbf{X}^{-1},$$

where the matrices $\mathbf{U} = [\mathbf{u}_1\ \mathbf{u}_2, \ldots, \mathbf{u}_m] \in \mathbb{R}^{n\times m}$ and $\mathbf{V} = [\mathbf{v}_1\ \mathbf{v}_2, \ldots, \mathbf{v}_p] \in \mathbb{R}^{p\times p}$ are column orthonormal, the matrix $\mathbf{X} = [\mathbf{x}_1\ \mathbf{x}_2, \ldots, \mathbf{x}_m] \in \mathbb{R}^{m\times m}$ is nonsingular, and $\Sigma = \mathrm{diag}\,(\sigma_1, \ldots, \sigma_p)$ and $\mathbf{M} = \mathrm{diag}(\mu_1, \ldots, \mu_p)$ are diagonal matrices. Moreover, the nonnegative diagonal entries of $\Sigma$ and $\mathbf{M}$ are ordered and normalized such that $0 \leqslant \sigma_1 \leqslant \cdots \leqslant \sigma_p \leqslant 1$, $1 \geqslant \mu_1 \geqslant \cdots \geqslant \mu_p > 0$, $\sigma_i^2 + \mu_i^2 = 1$ for $i = 1, 2, \ldots, p$. The ratios $\gamma_i = \frac{\sigma_i}{\mu_i}$ are known as the generalized singular values of the matrix pair $(\mathbf{H}, \mathbf{L})$.

**Lemma A.1.** *There exists at least one and at most $2p + 1$ positive solutions to (20).*

**Proof.** By applying GSVD, the regularization parameter $\eta^*$ solves the nonlinear equation

$$\eta^* \left[ \|\mathbf{L}\mathbf{m}_{\eta^*}\|_2^2 + \sigma_0^2 \sum_{i=1}^{p} \frac{1}{\gamma_i^2 + \eta^*} + 2\beta_0 \right] = 2\alpha_0'' \sigma_0^2.$$

Let $f(\eta) = \eta[\|\mathbf{L}\mathbf{m}_\eta\|_2^2 + \sigma_0^2 \sum_{i=1}^{p} \frac{1}{\gamma_i^2+\eta} + 2\beta_0]$, then $f(\eta)$ is continuous and

$$\lim_{\eta\to 0} f(\eta) = 0 \quad \text{and} \quad \lim_{\eta\to+\infty} f(\eta) = +\infty.$$

Therefore, by the continuity, there exists at least one positive solution to $f(\eta) = 2\alpha_0'' \sigma_0^2$. Let $y_i = \mathbf{u}_i^{\mathrm{T}}\mathbf{d}$ be the Fourier coefficients of the data $\mathbf{d}$. Then we have $\|\mathbf{L}\mathbf{m}_\eta\|_2^2 = \sum_{i=1}^{p} \frac{\gamma_i^2 y_i^2}{(\eta+\gamma_i^2)^2}$ [9]. Hence $f(\eta)$ can be expanded as a rational $f(\eta) = \frac{P(\eta)}{Q(\eta)}$ with $P(\eta)$ and $Q(\eta)$ being polynomials in $\eta$ of order $2p + 1$ and $2p$, respectively. Now the second assertion follows from the fundamental theorem of algebra. □

**Lemma A.2.** *Assume that the realization of the random variable $\boldsymbol{\omega}$ satisfies $\|\boldsymbol{\omega}\|_2^2 \leqslant c\sigma_0^2$. Then there exist two constants $c_{r,0}$ and $c_{r,1}$ dependent on $\alpha_0''$ such that $c_{r,0}\sigma_0^2 \leqslant \eta^* \leqslant c_{r,1}\sigma_0^2$.*

**Proof.** From (20), we note that

$$\eta^* = \frac{2\alpha_0''\sigma_0^2}{\|\mathbf{L}\mathbf{m}_{\eta^*}\|_2^2 + \sigma_0^2 \sum_{i=1}^{p} \frac{1}{\gamma_i^2+\eta^*} + 2\beta_0} \leqslant \frac{\alpha_0''\sigma_0^2}{\beta_0}. \tag{21}$$

The second inequality follows by setting $c_{r,1} = \frac{\alpha_0''}{\beta_0}$. From the variational characterization of the mean $\mathbf{m}_{\eta^*}$, it follows directly that:

$$\|\mathbf{H}\mathbf{m}_{\eta^*} - \mathbf{d}\|_2^2 + \eta^*\|\mathbf{L}\mathbf{m}_{\eta^*}\|_2^2 \leqslant \|\mathbf{H}\mathbf{m}^\dagger - \mathbf{d}\|_2^2 + \eta^*\|\mathbf{L}\mathbf{m}^\dagger\|_2^2.$$

Therefore, we deduce that $\|\mathbf{L}\mathbf{m}_{\eta^*}\|_2^2 \leqslant \frac{c\sigma_0^2}{\eta^*} + \|\mathbf{L}\mathbf{m}^\dagger\|_2^2$. Combined with (20), we deduce that

$$\eta^* \left( \frac{c\sigma_0^2}{\eta^*} + \|\mathbf{L}\mathbf{m}^\dagger\|_2^2 + \sigma_0^2 \sum_{i=1}^{p} \frac{1}{\gamma_i^2+\eta^*} + 2\beta_0 \right) \geqslant 2\alpha_0''\sigma_0^2.$$

Rearranging the terms and letting $c_{r,0} = \frac{2\alpha_0'' - p - c}{\|\mathbf{L}\mathbf{m}^\dagger\|_2^2 + 2\beta_0}$, we arrive at the desired lower bound. □

Note that the lower bound $c_{r,0}$ can be negative for fixed $\alpha_0''$. However, it can be made positive and meaningful if $\alpha_0''$ scales as $\sigma_0^{-d}$ with $0 < d < 2$. In particular, the lower bound can be set to $c_{\eta,0} = \frac{\alpha_0''}{2(\beta_0+c_m)}$ for sufficiently small $\sigma_0$.

**Proof of** Theorem 3.2

**Proof.** By the minimizing property of $\mathbf{m}_{\eta^*}$, we have

$$\|\mathbf{H}\mathbf{m}_{\eta^*} - \mathbf{d}\|_2^2 + \eta^*\|\mathbf{L}\mathbf{m}_{\eta^*}\|_2^2 \leqslant \|\mathbf{H}\mathbf{m}^\dagger - \mathbf{d}\|_2^2 + \eta^*\|\mathbf{L}\mathbf{m}^\dagger\|_2^2.$$

From this and Lemma A.2, we deduce that

$$\|\mathbf{H}\mathbf{m}_{\eta^*} - \mathbf{d}\|_2^2 \leqslant c\sigma_0^2 + \eta^*\|\mathbf{L}\mathbf{m}^\dagger\|_2^2 \leqslant c\sigma_0^2 + c_{r,1}\sigma_0^{2-d}\|\mathbf{L}\mathbf{m}^\dagger\|_2^2 \to 0,$$

as the noise level $\sigma_0 \to 0$. Appealing to Lemma A.2 and the minimizing property again, we obtain $\|\mathbf{L}\mathbf{m}_{\eta^*}\|_2^2 \leqslant \frac{c}{c_{r,0}}\sigma_0^d + \|\mathbf{L}\mathbf{m}^\dagger\|_2^2$. Now the assumption $\ker\mathbf{H} \cap \ker\mathbf{L} = \{\mathbf{0}\}$ implies that the sequence $\{\mathbf{m}_{\eta^*}\}$ is uniformly bounded. Therefore, there exists a subsequence, also denoted as $\{\mathbf{m}_{\eta^*}\}$, that converges to some $\bar{\mathbf{m}} \in \mathbb{R}^m$. Note that

$$\lim_{\sigma_0\to 0}\|\mathbf{H}\mathbf{m}_{\eta^*} - \mathbf{d}^\dagger\|_2^2 \leqslant \lim_{\sigma_0\to 0} 2\left[\|\mathbf{H}\mathbf{m}_{\eta^*} - \mathbf{d}\|_2^2 + \|\mathbf{d} - \mathbf{d}^\dagger\|_2^2\right] = 0 \quad \text{and} \quad \lim_{\sigma_0\to 0}\|\mathbf{L}\mathbf{m}_{\eta^*}\|_2^2 \leqslant \|\mathbf{L}\mathbf{m}^\dagger\|_2^2$$

i.e., $\|\mathbf{H}\bar{\mathbf{m}} - \mathbf{d}^\dagger\|_2^2 = 0$ and $\|\mathbf{L}\bar{\mathbf{m}}\|_2^2 \leqslant \|\mathbf{L}\mathbf{m}^\dagger\|_2^2$. The uniqueness of the generalized minimum-norm solution $\mathbf{m}^\dagger$ implies $\bar{\mathbf{m}} = \mathbf{m}^\dagger$. Now a subsequence convergence argument concludes the theorem. □

### A.2. Proof of Theorem 3.3

We shall need the following lemma.

**Lemma A.3.** *The sequence $\{\eta_k\}_k$ generated by* `Alg I` *is uniformly bounded.*

**Proof.** By bias-variance decomposition, $\tau_k$ and $\lambda_k$ can be explicitly written as

$$\lambda_k = \frac{2\alpha_0''}{\|\mathbf{L}\mathbf{m}_{k-1}\|_2^2 + \mathrm{tr}\left((\tau_{k-1}\mathbf{H}^{\mathrm{T}}\mathbf{H} + \lambda_{k-1}\mathbf{L}^{\mathrm{T}}\mathbf{L})^{-1}\mathbf{L}^{\mathrm{T}}\mathbf{L}\right) + 2\beta_0},$$

$$\tau_k = \frac{2\alpha_1''}{\|\mathbf{H}\mathbf{m}_{k-1} - \mathbf{d}\|_2^2 + \mathrm{tr}\left((\tau_{k-1}\mathbf{H}^{\mathrm{T}}\mathbf{H} + \lambda_{k-1}\mathbf{L}^{\mathrm{T}}\mathbf{L})^{-1}\mathbf{H}^{\mathrm{T}}\mathbf{H}\right) + 2\beta_1}.$$

Therefore, we have $0 \leqslant \lambda_k \leqslant \frac{\alpha_0''}{\beta_0}$ and $0 \leqslant \tau_k \leqslant \frac{\alpha_1''}{\beta_1}$, i.e., the sequences $\{\lambda_k\}_k$ and $\{\tau_k\}_k$ are uniformly bounded. The minimizing property of $\mathbf{m}_k = \mathbf{m}_{\eta_k}$ yields

$$\|\mathbf{H}\mathbf{m}_k - \mathbf{d}\|_2^2 + \eta_k\|\mathbf{L}\mathbf{m}_k\|_2^2 \leqslant \|\mathbf{H}\mathbf{0} - \mathbf{d}\|_2^2 + \eta_k\|\mathbf{L}\mathbf{0}\|_2^2 = \|\mathbf{d}\|_2^2, \quad \text{i.e.,} \ \|\mathbf{H}\mathbf{m}_k - \mathbf{d}\|_2^2 \leqslant \|\mathbf{d}\|_2^2.$$

By the definition of $\tau_k$ and the GSVD, we obtain

$$\tau_{k+1} = \frac{2\alpha_1''}{\|\mathbf{H}\mathbf{m}_k - \mathbf{d}\|_2^2 + \sum_{i=1}^p \frac{\gamma_i^2}{\tau_k\gamma_i^2+\lambda_k} + \frac{n-m}{\tau_k} + 2\beta_1} \geqslant \frac{2\alpha_1''}{\|\mathbf{d}\|_2^2 + \frac{n}{\tau_k} + 2\beta_1}.$$

To derive a positive lower bound for the sequence $\{\tau_k\}_k$, we define an auxiliary sequence $\{\tilde{\tau}_k\}_k$ by letting $\tilde{\tau}_0 = \tau_0$, and for $k = 0, 1, 2, \ldots$

$$\tilde{\tau}_{k+1} = \frac{2\alpha_1''}{\|\mathbf{d}\|_2^2 + \frac{n}{\tilde{\tau}_k} + 2\beta_1}.$$

We claim that $\tau_k \geqslant \tilde{\tau}_k$ for all $k$. It holds for $k = 0$ automatically, and for $k = 0, 1, \ldots$, we have

$$\tau_{k+1} - \tilde{\tau}_{k+1} \geqslant \frac{2\alpha_1''}{\|\mathbf{d}\|_2^2 + \frac{n}{\tau_k} + 2\beta_1} - \frac{2\alpha_1''}{\|\mathbf{d}\|_2^2 + \frac{n}{\tilde{\tau}_k} + 2\beta_1} = \frac{2n\alpha_1''}{\tau_k\tilde{\tau}_k\left(\|\mathbf{d}\|_2^2 + \frac{n}{\tau_k} + 2\beta_1\right)\left(\|\mathbf{d}\|_2^2 + \frac{n}{\tilde{\tau}_k} + 2\beta_1\right)}(\tau_k - \tilde{\tau}_k).$$

The assertion now follows by induction on $k$. Next we show that the sequence $\{\tilde{\tau}_k\}_k$ has a positive lower bound. Note that its definition gives

$$\tilde{\tau}_{k+1} - \tilde{\tau}_k = \frac{2n\alpha_1''}{\tilde{\tau}_{k-1}\tilde{\tau}_k\left(\|\mathbf{d}\|_2^2 + \frac{n}{\tilde{\tau}_k} + 2\beta_1\right)\left(\|\mathbf{d}\|_2^2 + \frac{n}{\tilde{\tau}_{k-1}} + 2\beta_1\right)}(\tilde{\tau}_k - \tilde{\tau}_{k-1}),$$

i.e., the sequence $\{\tilde{\tau}_k\}_k$ is monotonic. Moreover, it is bounded below and above by 0 and $\frac{\alpha_1''}{\beta_1}$, respectively. Therefore, the sequence $\{\tilde{\tau}_k\}_k$ converges monotonically. Upon convergence, the limit $\tilde{\tau}^*$ satisfies

$$\tilde{\tau}^* = \frac{2\alpha_1''}{\|\mathbf{d}\|_2^2 + \frac{n}{\tau^*} + 2\beta_1}, \quad \text{i.e.,} \ \tilde{\tau}^* = \frac{2\alpha_1'' - n}{\|\mathbf{d}\|_2^2 + 2\beta_1} = \frac{2\alpha_1}{\|\mathbf{d}\|_2^2 + 2\beta_1} > 0.$$

Therefore, the sequence $\{\tilde{\tau}_k\}_k$ is positively away from zero, and this in conjunction with the claim implies that there exists some positive constant $c_\tau$ such that $\tau_k \geqslant c_\tau$. Now the uniform boundedness of the regularization parameter $\eta_k$ follows directly from its definition $\eta_k = \lambda_k \tau_k^{-1}$ and the uniform boundedness of the sequences $\{\lambda_k\}_k$ and $\{\tau_k\}_k$. □

**Proof of** Theorem 3.3

**Proof.** By Lemma A.3 and its proof, the sequence $\{(\eta_k, \lambda_k, \tau_k)\}_k$ is uniformly bounded, and therefore, there exists a subsequence, denoted by $\{(\eta_{k_l}, \lambda_{k_l}, \tau_{k_l})\}_{k_l \in \mathbb{K}}$ with the index set $\mathbb{K} \subset \mathbb{N}$, and some $(\eta^*, \lambda^*, \tau^*) \in (\mathbb{R}^+)^3$ such that

$$\lim_{k_l \to \infty} \eta_{k_l} = \eta^*, \quad \lim_{k_l \to \infty} \lambda_{k_l} = \lambda^* \quad \text{and} \ \lim_{k_l \to \infty} \tau_{k_l} = \tau^*.$$

The convergence of the sequence $\{(\lambda_{k_l}, \tau_{k_l})\}_{k_l \in \mathbb{K}}$ yields

$$\lim_{k_l \to \infty} \left(\tau_{k_l} \mathbf{H}^{\mathrm{T}}\mathbf{H} + \lambda_{k_l} \mathbf{L}^{\mathrm{T}}\mathbf{L}\right)^{-1} = \left(\tau^* \mathbf{H}^{\mathrm{T}}\mathbf{H} + \lambda^* \mathbf{L}^{\mathrm{T}}\mathbf{L}\right)^{-1}, \quad \text{i.e.,} \ \operatorname{cov}_{q^{k_l}(\mathbf{m})}[\mathbf{m}] = \operatorname{cov}_{q^*(\mathbf{m})}[\mathbf{m}].$$

Appealing to the continuity of the Tikhonov solution $\mathbf{m}_\eta$ with respect to $\eta$, we derive from the convergence of the subsequence $\{\eta_{k_l}\}_{k_l \in \mathbb{K}}$ that $\lim_{k_l \to \infty} \mathbf{m}_{k_l} = \mathbf{m}_{\eta^*} := \mathbf{m}^*$. Noting the fact that $q^{k_l}(\mathbf{m})$ is solely determined by $\mathbf{m}_{k_l}$ and $\operatorname{cov}_{q^{k_l}(\mathbf{m})}[\mathbf{m}]$, we deduce $\lim_{k_l \to \infty} q^{k_l}(\mathbf{m}) = q^*(\mathbf{m})$. By the convergence of $\{\mathbf{m}_{k_l}\}_{k_l \in \mathbb{K}}$, we have

$$\lim_{k_l \to \infty} \|\mathbf{H}\mathbf{m}_{k_l} - \mathbf{d}\|_2^2 = \|\mathbf{H}\mathbf{m}^* - \mathbf{d}\|_2^2 \quad \text{and} \ \lim_{k_l \to \infty} \|\mathbf{L}\mathbf{m}_{k_l}\|_2^2 = \|\mathbf{L}\mathbf{m}^*\|_2^2.$$

This in conjunction with the convergence of $\{\operatorname{cov}_{q^{k_l}(\mathbf{m})}[\mathbf{m}]\}_{k_l \in \mathbb{K}}$ implies

$$\lim_{k_l \to \infty} \frac{2\alpha_0''}{\|\mathbf{L}\mathbf{m}_{k_l}\|_2^2 + \operatorname{tr}\left(\operatorname{cov}_{q^{k_l}(\mathbf{m})}[\mathbf{m}]\mathbf{L}^{\mathrm{T}}\mathbf{L}\right) + 2\beta_0} = \frac{2\alpha_0''}{\|\mathbf{L}\mathbf{m}^*\|_2^2 + \operatorname{tr}(\operatorname{cov}_{q^*(\mathbf{m})}[\mathbf{m}]\mathbf{L}^{\mathrm{T}}\mathbf{L}) + 2\beta_0} := \lambda^{**},$$

$$\lim_{k_l \to \infty} \frac{2\alpha_1''}{\|\mathbf{H}\mathbf{m}_{k_l} - \mathbf{d}\|_2^2 + \operatorname{tr}(\operatorname{cov}_{q^{k_l}(\mathbf{m})}[\mathbf{m}]\mathbf{H}^{\mathrm{T}}\mathbf{H}) + 2\beta_1} = \frac{2\alpha_1''}{\|\mathbf{H}\mathbf{m}^* - \mathbf{d}\|_2^2 + \operatorname{tr}(\operatorname{cov}_{q^*(\mathbf{m})}[\mathbf{m}]\mathbf{H}^{\mathrm{T}}\mathbf{H}) + 2\beta_1} := \tau^{**}.$$

Combined with the definitions of $q^{k_l}(\lambda)$ and $q^{k_l}(\tau)$, these two identities imply

$$\lim_{k_l \to \infty} q^{k_l}(\lambda) = q^*(\lambda) \quad \text{and} \ \lim_{k_l \to \infty} q^{k_l}(\tau) = q^*(\tau)$$

for some $q^*(\lambda)$ and $q^*(\tau)$ uniquely determined by $\lambda^{**}$ and $\tau^{**}$, respectively. Therefore, the subsequence $\{q^{k_l}(\mathbf{m})q^{k_l}(\lambda)q^{k_l}(\tau)\}_{k_l \in \mathbb{K}}$ converges to some $q^*(\mathbf{m}, \lambda, \tau) \equiv q^*(\mathbf{m})q^*(\lambda)q^*(\tau)$. By the definitions of $\lambda_{k_l+1}$ and $\tau_{k_l+1}$, we see that the subsequence $\{(\lambda_{k_l+1}, \tau_{k_l+1})\}_{k_l \in \mathbb{K}}$ also converges, and by repeating the preceding argumentation, we deduce that the subsequence $\{q^{k_l+1}(\mathbf{m})q^{k_l+1}(\lambda)q^{k_l+1}(\tau)\}_{k_l \in \mathbb{K}}$ converges to some $q^{**}(\mathbf{m}, \lambda, \tau) \equiv q^{**}(\mathbf{m})q^{**}(\lambda)\, q^{**}(\tau)$.

Next we show that the limit $q^*(\mathbf{m}, \lambda, \tau) \equiv q^*(\mathbf{m})q^*(\lambda)q^*(\tau)$ is a stationary point of the divergence $D_{KL}$. To this end, let $\mathcal{S}$ be the algorithmic map [37] defined by `Alg I`, i.e., the solution operator that maps $q^k(\mathbf{m}, \lambda, \tau)$ into $q^{k+1}(\mathbf{m}, \lambda, \tau)$. Then the continuity of the functional $D_{KL}$ implies that the map $\mathcal{S}$ is closed. Therefore, we deduce that $q^{**}(\mathbf{m}, \lambda, \tau) = \mathcal{S}q^*(\mathbf{m}, \lambda, \tau)$. Obviously, it holds true that

$$D_{KL}(q^{**}(\mathbf{m})q^{**}(\lambda, \tau)|p(\mathbf{m}, \lambda, \tau)) \leqslant D_{KL}(q^*(\mathbf{m})q^{**}(\lambda, \tau)|p(\mathbf{m}, \lambda, \tau)) \leqslant D_{KL}(q^*(\mathbf{m})q^*(\lambda, \tau)|p(\mathbf{m}, \lambda, \tau)).$$

This and the monotone convergence of the functional value imply

$$D_{KL}(q^{**}(\mathbf{m})q^{**}(\lambda, \tau)|p(\mathbf{m}, \lambda, \tau)) = D_{KL}(q^*(\mathbf{m})q^{**}(\lambda, \tau)|p(\mathbf{m}, \lambda, \tau)) = D_{KL}(q^*(\mathbf{m})q^*(\lambda, \tau)|p(\mathbf{m}, \lambda, \tau)). \tag{22}$$

From the identity $q^{**}(\mathbf{m}, \lambda, \tau) = \mathcal{S}q^*(\mathbf{m}, \lambda, \tau)$, i.e.,

$$D_{KL}(q^*(\mathbf{m})q^{**}(\lambda, \tau)|p) \leqslant D_{KL}(q^*(\mathbf{m})q(\lambda, \tau)|p) \quad \forall q(\lambda, \tau),$$
$$D_{KL}(q^{**}(\mathbf{m})q^{**}(\lambda, \tau)|p) \leqslant D_{KL}(q(\mathbf{m})q^{**}(\lambda, \tau)|p) \quad \forall q(\mathbf{m}).$$

Now the strict biconvexity of the functional $D_{KL}$ and (22) implies that $q^{**}(\lambda, \tau) = q^*(\lambda, \tau)$ and $q^{**}(\mathbf{m}) = q^*(\mathbf{m})$, i.e., $q^{**}(\mathbf{m}, \lambda, \tau) = q^*(\mathbf{m}, \lambda, \tau)$, and thus $q^*(\mathbf{m}, \lambda, \tau)$ is a stationary point. □

## References

[1] J.V. Beck, B. Blackwell, C.R. St Clair, Inverse Heat Conduction: Ill-Posed Problems, Wiley, New York, 1985.
[2] A. Tarantola, Inverse Problem Theory and Methods for Model Parameter Estimation, SIAM, Philadelphia, 2005.

[3] A. Malinverno, V.A. Briggs, Expanded uncertainty quantification in inverse problems: hierarchical Bayes and empirical Bayes, Geophysics 69 (2004) 1005–1016.
[4] J. Kaipio, E. Somersalo, Statistical and Computational Inverse Problems, Springer, New York, 2005.
[5] J. Wang, N. Zabaras, A Bayesian inference approach to the inverse heat conduction problem, Int. J. Heat Mass Transfer 47 (2004) 3927–3941.
[6] J. Wang, N. Zabaras, Hierarchical Bayesian models for inverse problems in heat conduction, Inverse Probl. 21 (2005) 183–206.
[7] A.F. Emery, E. Valenti, D. Bardot, Using Bayesian inference for parameter estimation when the system response and experimental conditions are measured with error and some variables are considered as nuisance variables, Meas. Sci. Technol. 18 (2007) 19–29.
[8] B. Jin, J. Zou, A Bayesian inference approach to the ill-posed Cauchy problem of steady-state heat conduction, Int. J. Numer. Methods Eng. 76 (2008) 521–544.
[9] B. Jin, J. Zou, Augmented Tikhonov regularization, Inverse Probl. 25 (2009) 025001.
[10] M.I. Jordan, Z. Ghahramani, T.S. Jaakkola, L.K. Saul, An introduction to variational methods for graphical models, Mach. Learn. 37 (1999) 183–233.
[11] H. Attias, A variational Bayesian framework for graphical models, in: S.A. Solla, T.K. Leen, K.R. Maller (Eds.), Advances in Neural Information Processing Systems, vol. 12, MIT Press, Cambridge, 2000, pp. 209–215.
[12] M.J. Beal, Variational Algorithms for Approximate Bayesian Inference, Ph.D. Dissertation, University College London, London, UK, 2003.
[13] L.E. Eberly, G. Casella, Estimating Bayesian credible intervals, J. Stat. Plan. Infer. 112 (2003) 115–132.
[14] W.R. Gilks, S. Richardson, D.J. Spiegelhalter, Markov Chain Monte Carlo in Practice, Chapman & Hall, 1996.
[15] J.S. Liu, Monte Carlo Strategies in Scientific Computing, Springer, 2008.
[16] M. Sato, T. Yoshioka, S. Kajihara, K. Toyama, N. Goda, K. Doya, M. Kawato, Hierarchical Bayesian estimation for MEG inverse problem, NeuroImage 23 (2004) 806–826.
[17] S.D. Babacan, R. Molina, A.K. Katsaggelos, Parameter estimation in TV image restoration using variational distribution approximation, IEEE Trans. Image Proc. 17 (2008) 326–339.
[18] A. Quinn, V. Šmídl V, The Variational Bayes Method in Signal Processing, Springer, Berlin, 2005.
[19] A.L. Gibbs, F.E. Su, On choosing and bounding probability metrics, Int. Stat. Rev. 70 (2002) 419–435.
[20] E. Resmerita, R.S. Anderssen, Joint additive Kullback–Leibler residual minimization and regularization for linear inverse problems, Math. Methods Appl. Sci. 30 (2007) 1527–1544.
[21] P.P.B. Eggermont, Maximum entropy regularization for Fredholm integral equations of the first kind, SIAM J. Math. Anal. 24 (1993) 1557–1576.
[22] P.C. Hansen, Rank-Deficient and Discrete Ill-Posed Problems, SIAM, Philadelphia, 1998.
[23] H.W. Engl, M. Hanke, A. Neubauer, Regularization of Inverse Problems, Kluwer, Dordrecht, 1996.
[24] K. Ito, B. Jin, J. Zou, A new choice rule for regularization parameters in Tikhonov regularization, Tech. Report 2008–2007(362), Department of Mathematics, Chinese University of Hong Kong, 2008.
[25] I. Daubechies, M. Dfriese, C. De Mol, An iterative thresholding algorithm for linear inverse problems with a sparsity constraint, Commun. Pure Appl. Math. 57 (2004) 1413–1457.
[26] S.D. Babacan, L. Mancera, R. Molina, A.K. Katsaggelos, Nonconvex priors in Bayesian compressed sensing, in: 17th European Signal Processing Conference (EUSIPCO 2009), Glasgow, Scotland, August 24–28, 2009.
[27] M.A. Chappell, A.R. Groves, B. Whitcher, M.W. Woolrich, Variational Bayesian inference for a nonlinear forward model, IEEE Trans. Signal Proc. 57 (2009) 223–236.
[28] B. Jin, L. Marin, The plane wave method for some inverse problems associated with Helmholtz-type equations, Eng. Anal. Bound. Elem. 32 (2008) 223–240.
[29] A.M. Osman, J.V. Beck, Nonlinear inverse problem for the estimation of time-and-space dependent heat transfer coefficients, J. Thermophys. 3 (1988) 146–152.
[30] B. Jin, J. Zou, Numerical estimation of the Robin coefficient in a stationary diffusion equation, IMA J. Numer. Anal., in press, doi:10.1093/imanum/drn066.
[31] Y.C. Hon, T. Wei, Backus–Gilbert algorithm for the Cauchy problem of the Laplace equation, Inverse Probl. 17 (2001) 261–271.
[32] B. Jin, Fast Bayesian approach for parameter estimation, Int. J. Numer. Methods Eng. 76 (2008) 230–252.
[33] V. Isakov, Inverse Problems for Partial Differential Equations, second ed., Springer, New York, 2006.
[34] B. Wang, M. Titterington, Inadequacy of interval estimates corresponding to variational Bayesian approximations, in: R. Cowell, Z. Ghahramani (Eds.), Proceedings of the Tenth International workshop on Artificial Intelligence and Statistics (6–8, January 2005, Barbados), pp. 373–380.
[35] H.K.H. Lee, D.M. Higdon, Z. Bi, M.A.R. Ferreira, M. West, Markov random field models for high-dimensional parameters in simulation of fluid flow in porous media, Technometrics 44 (2002) 230–241.
[36] H. Haario, M. Laine, A. Mira, E. Saksman, DRAM: efficient adaptive MCMC, Stat. Comput. 16 (2006) 339–354.
[37] M. Bazaraa, H. Sherali, C. Shetty, Nonlinear Programming: Theory and Algorithms, Wiley, New York, 1993.