

Approximate Bayesian model inversion for PDEs with heterogeneous and state-dependent coefficients

D.A. Barajas-Solano*, A.M. Tartakovsky

Pacific Northwest National Laboratory, Richland, WA 99354, United States of America



ARTICLE INFO

Article history:

Received 30 November 2018

Received in revised form 3 June 2019

Accepted 6 June 2019

Available online 8 June 2019

Keywords:

Approximate Bayesian inference

Model inversion

Variational inference

Empirical Bayes

ABSTRACT

We present two approximate Bayesian inference methods for parameter estimation in partial differential equation (PDE) models with space-dependent and state-dependent parameters. We demonstrate that these methods provide accurate and cost-effective alternatives to Markov Chain Monte Carlo simulation. We assume a parameterized Gaussian prior on the unknown functions, and approximate the posterior density by a parameterized multivariate Gaussian density. The parameters of the prior and posterior are estimated from sparse observations of the PDE model's states and the unknown functions themselves by maximizing the evidence lower bound (ELBO), a lower bound on the log marginal likelihood of the observations. The first method, Laplace-EM, employs the expectation maximization algorithm to maximize the ELBO, with a Laplace approximation of the posterior on the E-step, and minimization of a Kullback-Leibler divergence on the M-step. The second method, DSVI-EB, employs the doubly stochastic variational inference (DSVI) algorithm, in which the ELBO is maximized via gradient-based stochastic optimization, with noisy gradients computed via simple Monte Carlo sampling and Gaussian backpropagation. We apply these methods to identifying diffusion coefficients in linear and nonlinear diffusion equations, and we find that both methods provide accurate estimates of posterior densities and the hyperparameters of Gaussian priors. While the Laplace-EM method is more accurate, it requires computing Hessians of the physics model. The DSVI-EB method is found to be less accurate but only requires gradients of the physics model.

© 2019 Elsevier Inc. All rights reserved.

1. Introduction

Partial differential equation (PDE) models of many physical systems involve space-dependent parameters and constitutive relationships that are usually only partially observed. Model inversion aims to estimate these unknown functions of space and the system's state from sparse measurements of the state, associated quantities of interest, and the unknown functions themselves. Bayesian inference provides a probabilistic framework for model inversion [1], in which data is assimilated by computing the posterior density of the parameters in terms of the likelihood of the observations given the PDE model and the prior density of the parameters codifying modeling assumptions. Unlike deterministic parameter estimation methods [2,3], the Bayesian framework provides a probabilistic characterization of the estimated parameter that can be employed for quantifying uncertainty and evaluating modeling assumptions. For linear problems with Gaussian likelihoods and

* Corresponding author.

E-mail addresses: David.Barajas-Solano@pnnl.gov (D.A. Barajas-Solano), Alexandre.Tartakovsky@pnnl.gov (A.M. Tartakovsky).

Gaussian priors, Bayesian inference can be done exactly (known in the context of state estimation for dynamical systems as the Kalman filter [4]). Unfortunately, physics models define nonlinear maps between the state and the model parameters, preventing carrying out exact inference even in the case of Gaussian likelihoods and Gaussian priors. The Markov Chain Monte Carlo (MCMC) method is robust for general nonlinear problems but is computationally expensive [5]. Despite recent advances in Hamiltonian Monte Carlo and ensemble and parallel MCMC [6–8], the number of forward simulations and likelihood evaluations required by MCMC sampling poses a challenge for model inversion of PDE models with high-dimensional parameters. Here, we propose two cost-effective alternatives to MCMC for estimating unknown parameters and constitutive relationships in PDE models.

Gaussian process (GP) regression, known as kriging in spatial geophysics, is commonly used to construct probabilistic models of heterogeneous parameters; therefore, GPs serve as a reasonable choice of prior for unknown parameters. In the context of Bayesian inference with GP priors, GP regression is equivalent to exact Bayesian inference for assimilating direct measurements of unknown parameters. Similarly, the marginal likelihood of parameter measurements can be computed in closed form, therefore allowing for model selection to be carried out by empirical Bayesian inference, also known as type-II maximum likelihood estimation [9].

Assimilating measurements of the state of PDE models is on the other hand less straightforward. Recently, a framework has been proposed to combine GP priors on the state and a discretization of the governing PDEs to assimilate state observations [10,11]. In this framework, state estimation and type-II maximum likelihood can be carried out in closed form when the governing equations are linear on the state; for the nonlinear case, inference is carried out approximately via linearization of the governing equations.

Parameter estimation for PDE models presents another layer of challenge as governing equations commonly induce nonlinear relations between parameters and states. A common example is the Laplace equation with space-dependent unknown diffusion coefficient, which is linear on the state, but induces a nonlinear relation between the state and the diffusion coefficient. For the general case of parameter estimation with nonlinearity introduced by the physics model, approximate Bayesian inference methods are necessary. The standard approximate inference tool is MCMC sampling of the Bayesian posterior. Given unbounded computational resources, MCMC will provide arbitrarily accurate results, but in practice MCMC often requires an intractable amount of forward simulations of the physics model. Algorithms such as Hamiltonian Monte Carlo (HMC) and the Metropolis-adjusted Langevin algorithm (MALA) employ first-order information in the form sensitivities of the physics model to improve the mixing and convergence of the Markov chain random walk, but nevertheless the total number of forward and sensitivity simulations remains a challenge. As an alternative, approaches such as the Laplace approximation and variational inference aim to approximate the exact posterior with an analytical, parameterized density.

In this manuscript we propose employing approximate Bayesian inference with GP priors to approximate the posterior of PDE parameters and to estimate the hyperparameters of their GP prior. We propose two optimization-based methods: The first, Laplace-EM, is based on the Laplace approximation [9,12,13] and the expectation maximization (EM) algorithm [14,12]. The second, doubly stochastic variational inference for empirical Bayes inference (DSVI-EB) is based on the DSVI algorithm [15,16]. The proposed methods employ first and second-order information, i.e., gradient and Hessian of physics models, evaluated via the discrete adjoint method. Both presented methods enjoy advantageous computational properties over MCMC and other approximate Bayesian inference algorithms such as expectation propagation [17] and the Laplace approximation-based method of [13] that renders each of them attractive for model inversion depending on the nature of the inversion problem. In particular, the Laplace-EM method is accurate for approximating the unimodal posteriors of the numerical examples of this manuscript, but requires computing Hessians. On the other hand, DSVI-EB is less accurate but only requires computing gradients, and can be trivially parallelized. We note that Gaussian mixtures can be employed in variational inference to approximate multimodal posteriors [18], but in the present work we limit our focus to unimodal posteriors. Furthermore, both methods are applicable to non-factorizing likelihoods, do not require computing moments of the likelihood, and do not require third- or higher order derivatives of the physics model. Finally, variational inference and the Laplace approximation have been employed for model inversion [19–22,13], but to the best of our knowledge, have not been used in the context of the empirical Bayesian framework to estimate GP prior hyperparameters, with the exception of the work of [13]. That work, based on the Laplace approximation, requires computing third-order derivatives of the physics model, which may be costly to compute, whereas the presented methods do not require third-order derivatives.

The manuscript is structured as follows: In Section 2 we formulate the empirical Bayesian inference problem for physics models and GP priors. In Section 3 we propose the approximate Bayesian inference and summarize the expectation maximization (EM) algorithm. The Laplace-EM algorithm is introduced in Section 4, and the DSVI-EB algorithm is described in Section 5. The computational complexity of the algorithms is discussed in Section 6. The application of the proposed methods is presented in Section 7. Finally, conclusions are given in Section 8.

2. Problem formulation

We consider physical systems modeled by stationary PDEs over the simulation domain $\Omega \subset \mathbb{R}^d$, $d \in [1, 3]$. We denote by $u: \Omega \rightarrow U \subset \mathbb{R}$ the system's state, and by $y: \Omega \times U \rightarrow \mathbb{R}$ the system's parameter, an unknown scalar function of space and the system's state. Our goal is to estimate the unknown function $y(\mathbf{x}, u)$ from sparse, noisy measurements of $u(\mathbf{x})$ and $y(\mathbf{x}, u)$. The PDE and boundary conditions are discretized for numerical computations, resulting in the set of M algebraic equations $\mathbf{L}(\mathbf{u}, \mathbf{y}) = 0$, where $\mathbf{u} \in \mathbb{R}^M$ denotes the vector of M state degrees of freedom, and $\mathbf{y} \in \mathbb{R}^N$ denotes the discretized

parameter vector, corresponding to the value of $y(\mathbf{x}, u)$ at the N discrete locations $\{\xi_i \in \Omega \times U\}_{i=1}^N$. Furthermore, we denote by Ξ the matrix of coordinates $\Xi \equiv (\xi_1, \dots, \xi_N)$.

We assume that the sparse observations of the discrete state and parameters, \mathbf{u}_s and \mathbf{y}_s , respectively, are collected with iid normal observation errors, that is,

$$\mathbf{u}_s = \mathbf{H}_u \mathbf{u} + \boldsymbol{\epsilon}_{us}, \quad \boldsymbol{\epsilon}_{us} \sim \mathcal{N}(0, \sigma_{us}^2 \mathbf{I}_{M_s}), \quad (1)$$

$$\mathbf{y}_s = \mathbf{H}_y \mathbf{y} + \boldsymbol{\epsilon}_{ys}, \quad \boldsymbol{\epsilon}_{ys} \sim \mathcal{N}(0, \sigma_{ys}^2 \mathbf{I}_{N_s}), \quad (2)$$

where $\mathbf{u}_s \in \mathbb{R}^{M_s}$, $M_s \ll M$ are the state observations, $\mathbf{y}_s \in \mathbb{R}^{N_s}$, $N_s \ll N$ are the parameter observations, $\mathbf{H}_u \in \mathbb{R}^{M_s \times M}$ is the state observation operator, and $\mathbf{H}_y \in \mathbb{R}^{N_s \times N}$ is the parameter observation operator. The vectors $\boldsymbol{\epsilon}_{us}$ and $\boldsymbol{\epsilon}_{ys}$ denote the iid normal observation errors with standard deviations σ_{us} and σ_{ys} , respectively. The observation errors satisfy $\mathbb{E}[\boldsymbol{\epsilon}_{us} \boldsymbol{\epsilon}_{ys}^\top] = 0$. In (1) and (2), \mathbf{I}_{M_s} and \mathbf{I}_{N_s} denote the $M_s \times M_s$ and $N_s \times N_s$ identity matrices, respectively. Then, the likelihood of the observations $\mathcal{D}_s \equiv \{\mathbf{u}_s, \mathbf{y}_s\}$ given \mathbf{y} is defined as

$$\log p(\mathcal{D}_s | \mathbf{y}) \equiv -\frac{1}{2\sigma_{us}^2} \|\mathbf{u}_s - \mathbf{H}_u \mathbf{u}\|_2^2 - \frac{1}{2\sigma_{ys}^2} \|\mathbf{y}_s - \mathbf{H}_y \mathbf{y}\|_2^2 + \text{const.}, \quad (3)$$

where \mathbf{u} satisfies the physics constraint $\mathbf{L}(\mathbf{u}, \mathbf{y}) = 0$ given \mathbf{y} , and the constant is independent of \mathbf{y} .

In probabilistic terms, our goal is to estimate the posterior density of \mathbf{y} given the data \mathcal{D}_s . By Bayes' theorem, this posterior is given by

$$p(\mathbf{y} | \mathcal{D}_s, \boldsymbol{\theta}) = \frac{p(\mathcal{D}_s | \mathbf{y}) p(\mathbf{y} | \boldsymbol{\theta})}{p(\mathcal{D}_s | \boldsymbol{\theta})}, \quad (4)$$

where $p(\mathbf{y} | \boldsymbol{\theta})$ is the prior density of \mathbf{y} parameterized by H hyperparameters organized into the vector $\boldsymbol{\theta}$. For simplicity we assume that $\boldsymbol{\theta} \in \mathbb{R}^H$. Constraints on the values of the hyperparameters can be enforced via changes of variables. E.g., for a certain hyperparameter $\eta_i \in (0, +\infty)$ we choose $\theta_i = \log \eta_i$ so that $\theta_i \in \mathbb{R}$. The density $p(\mathcal{D}_s | \boldsymbol{\theta})$ is the *marginal likelihood* or *evidence* of the data, given by

$$p(\mathcal{D}_s | \boldsymbol{\theta}) = \int p(\mathcal{D}_s | \mathbf{y}) p(\mathbf{y} | \boldsymbol{\theta}) d\mathbf{y}. \quad (5)$$

If one is not interested in the uncertainty in estimating \mathbf{y} given the data, one can compute in lieu of the full posterior (4) the *maximum a posteriori* (MAP) point estimate of \mathbf{y} , defined as the mode of the posterior, that is,

$$\hat{\mathbf{y}} \equiv \arg \max_{\mathbf{y}} \log p(\mathbf{y} | \mathcal{D}_s, \boldsymbol{\theta}) = \arg \max_{\mathbf{y}} \log p(\mathbf{y}, \mathcal{D}_s | \boldsymbol{\theta}), \quad (6)$$

where $p(\mathbf{y}, \mathcal{D}_s | \boldsymbol{\theta}) = p(\mathcal{D}_s | \mathbf{y}) p(\mathbf{y} | \boldsymbol{\theta})$ is the joint density of the data and the parameters given $\boldsymbol{\theta}$. Here we used the fact that the marginal likelihood $p(\mathcal{D}_s | \boldsymbol{\theta})$ is independent of \mathbf{y} .

We employ a zero-mean Gaussian process prior, that is,

$$p(\mathbf{y} | \boldsymbol{\theta}) = \mathcal{N}(\mathbf{y} | 0, \mathbf{C}_p(\boldsymbol{\theta}) \equiv C(\Xi, \Xi | \boldsymbol{\theta})), \quad (7)$$

where $\mathcal{N}(\cdot | \boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes the multivariate normal density with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, and $C(\cdot, \cdot | \boldsymbol{\theta})$ is a parameterized covariance kernel.

The posterior density depends on the prior hyperparameters, which can be chosen based on prior expert knowledge, or estimated from data. In the empirical Bayes approach, also known as *type-II maximum likelihood* or marginal likelihood estimation, point estimates of the hyperparameters are obtained by maximizing the marginal likelihood with respect to $\boldsymbol{\theta}$, i.e., $\boldsymbol{\theta} \equiv \arg \max_{\boldsymbol{\theta}} p(\mathcal{D}_s | \boldsymbol{\theta})$. In the fully Bayes approach, we instead pose a hyperprior on the hyperparameters, which is updated with data by the Bayes' theorem. In this work we will pursue the empirical Bayes approach.

Due to the nonlinear map from \mathbf{y} to \mathbf{u} defined by the physics constraint, the Bayesian inference problem of evaluating the posterior and marginal likelihood cannot be done in closed form. Exact inference therefore requires sampling the posterior via MCMC, which is intractable for sufficiently large N and M . As a consequence, estimating hyperparameters via marginal likelihood estimation is also intractable. As an alternative to exact inference, in this work we propose various approximate inference algorithms.

3. Approximate inference and expectation maximization

The goal is to approximate the exact posterior $p(\mathbf{y} | \mathcal{D}_s, \boldsymbol{\theta})$ by a density $q(\mathbf{y})$. More precisely, let $\mathbf{y} \in \mathbb{R}^N$, $\mathcal{B}(\mathbb{R}^N)$ denote the Borel σ -algebra on \mathbb{R}^N , and $\mu(\cdot)$ denote the Lebesgue measure on $\mathcal{B}(\mathbb{R}^N)$. We aim to approximate $p(\mathbf{y} | \mathcal{D}_s, \boldsymbol{\theta})$ by a density q defined as the Radon-Nikodym derivative $q = d\mathcal{Q}/d\mu(\mathbb{R}^N)$ for some $\mathcal{Q} \in (\mathbb{R}^N, \mathcal{B}(\mathbb{R}^N))$. The Kullback-Leibler (KL) divergence $D_{\text{KL}}(q(\mathbf{y}) \| p(\mathbf{y} | \mathcal{D}_s, \boldsymbol{\theta}))$ provides a means to rewriting the marginal likelihood, (5), in terms of $q(\mathbf{y})$. Namely, substituting (4) into the definition of the KL divergence gives

$$\begin{aligned}
D_{\text{KL}}(q(\mathbf{y}) \parallel p(\mathbf{y} \mid \mathcal{D}_s, \boldsymbol{\theta})) &= - \int q(\mathbf{y}) \log \frac{p(\mathbf{y} \mid \mathcal{D}_s, \boldsymbol{\theta})}{q(\mathbf{y})} d\mathbf{y} \\
&= - \int q(\mathbf{y}) \log \frac{p(\mathcal{D}_s \mid \mathbf{y}) p(\mathbf{y} \mid \boldsymbol{\theta})}{q(\mathbf{y}) p(\mathcal{D}_s \mid \boldsymbol{\theta})} d\mathbf{y} \\
&= -\mathcal{F}[q(\mathbf{y}), \boldsymbol{\theta}] + \log p(\mathcal{D}_s \mid \boldsymbol{\theta}),
\end{aligned} \tag{8}$$

where $\mathcal{F}[q(\mathbf{y}), \boldsymbol{\theta}]$ is given by

$$\mathcal{F}[q(\mathbf{y}), \boldsymbol{\theta}] = \mathbb{E}_{q(\mathbf{y})} [\log p(\mathcal{D}_s \mid \mathbf{y})] - D_{\text{KL}}(q(\mathbf{y}) \parallel p(\mathbf{y} \mid \boldsymbol{\theta})), \tag{9}$$

and $\mathbb{E}_{q(\mathbf{y})}[\cdot] \equiv \int (\cdot) q(\mathbf{y}) d\mathbf{y}$ denotes expectation with respect to the density $q(\mathbf{y})$. Reorganizing (8) we have the following alternative expression for (5):

$$\log p(\mathcal{D}_s \mid \boldsymbol{\theta}) = \mathcal{F}[q(\mathbf{y}), \boldsymbol{\theta}] + D_{\text{KL}}(q(\mathbf{y}) \parallel p(\mathbf{y} \mid \mathcal{D}_s, \boldsymbol{\theta})).$$

Given that the KL divergence is always non-negative, we have the inequality $\log p(\mathcal{D}_s \mid \boldsymbol{\theta}) \geq \mathcal{F}[q(\mathbf{y}), \boldsymbol{\theta}]$; therefore, the operator \mathcal{F} is often called the *evidence lower bound* (ELBO). The inequality becomes an equality when $q(\mathbf{y}) = p(\mathbf{y} \mid \mathcal{D}_s, \boldsymbol{\theta})$, that is, when the variational density is equal to the exact posterior. In the empirical Bayes setting, this suggests the strategy of selecting both q and $\boldsymbol{\theta}$ by maximizing the ELBO [14], i.e.,

$$(q(\mathbf{y}), \boldsymbol{\theta}) \equiv \arg \max_{(z(\mathbf{y}), \boldsymbol{\theta})} \mathcal{F}[z(\mathbf{y}), \boldsymbol{\theta}], \tag{10}$$

where the search for $q(\mathbf{y})$ is over probability densities corresponding to probability measures in $(\mathbb{R}^N, \mathcal{B}(\mathbb{R}^N))$, and the search for $\boldsymbol{\theta}$ is over \mathbb{R}^H , where H is the number of hyperparameters. Instead of maximizing over $q(\mathbf{y})$ and $\boldsymbol{\theta}$ simultaneously, we can do it iteratively by alternating two maximization steps, resulting in the iterative scheme

$$\begin{aligned}
\text{E-step: } q^{(j+1)}(\mathbf{y}) &\text{ set as } \arg \max_{z(\mathbf{y})} \mathcal{F}[z(\mathbf{y}), \boldsymbol{\theta}^{(j)}] \\
\text{M-step: } \boldsymbol{\theta}^{(j+1)} &\text{ set as } \arg \max_{\boldsymbol{\phi}} \mathcal{F}[q^{(j+1)}(\mathbf{y}), \boldsymbol{\phi}],
\end{aligned} \tag{11}$$

thus recovering the expectation maximization (EM) algorithm [14].

It remains to specify how the maximization problems (10) and (11) will be solved, particularly with respect to how to optimize over the space of possible densities $q(\mathbf{y})$ approximating the true posterior. The approximate inference algorithms presented in this manuscript are based on two families of approximations of the posterior. The Laplace-EM algorithm (Section 4) uses a *local* approximation around the MAP for a given $\boldsymbol{\theta}$, and optimizes the ELBO using the EM algorithm, (11). The DSVI-EB algorithm (Section 5) uses a parameterized density $q(\mathbf{y} \mid \boldsymbol{\phi})$ with *variational* parameters $\boldsymbol{\phi}$ to be selected jointly with $\boldsymbol{\theta}$ via stochastic optimization.

4. Laplace-EM algorithm

The Laplace approximation is an approach for approximating unimodal posteriors densities. It consists of fitting a multivariate Gaussian density around the MAP for a given choice of hyperparameters $\boldsymbol{\theta}$. The j th E-step of the EM algorithm, (11), consists of finding the posterior for a given set of hyperparameters, $\boldsymbol{\theta}^{(j)}$. This suggests we can replace the E-step by a Laplace approximation to the posterior, giving raise to the Laplace-EM algorithm.

We proceed to briefly describe the Laplace approximation. Expanding up to second order the log posterior (see (4)) around the MAP (6) yields

$$\log p(\mathbf{y} \mid \mathcal{D}_s, \boldsymbol{\theta}) = -\log p(\mathcal{D}_s \mid \boldsymbol{\theta}) + \log p(\hat{\mathbf{y}}, \mathcal{D}_s \mid \boldsymbol{\theta}) + \frac{1}{2}(\mathbf{y} - \hat{\mathbf{y}})^\top [\nabla \nabla \log p(\mathbf{y}, \mathcal{D}_s \mid \boldsymbol{\theta})|_{\mathbf{y}=\hat{\mathbf{y}}}] (\mathbf{y} - \hat{\mathbf{y}}) + \dots,$$

where $\nabla \nabla \log p(\mathbf{y}, \mathcal{D}_s \mid \boldsymbol{\theta})|_{\mathbf{y}=\hat{\mathbf{y}}}$ denotes the Hessian of the log joint density $\log p(\mathbf{y}, \mathcal{D}_s \mid \boldsymbol{\theta})$ around the MAP. This quadratic expression suggests approximating the posterior by the multivariate Gaussian density $q(\mathbf{y}) \equiv \mathcal{N}(\mathbf{y} \mid \boldsymbol{\mu}_q, \boldsymbol{\Sigma}_q)$ with mean $\boldsymbol{\mu}_q$ given by the MAP and covariance $\boldsymbol{\Sigma}_q$ given by the Hessian of the log joint density. In other words, we have

$$\begin{aligned}
\boldsymbol{\mu}_q &\equiv \arg \min_{\mathbf{y}} [-\log p(\mathbf{y}, \mathcal{D}_s \mid \boldsymbol{\theta})] \\
&= \arg \min_{\mathbf{y}} \left\{ -\log p(\mathcal{D}_s \mid \mathbf{y}) + \frac{1}{2} [\mathbf{y}^\top \mathbf{C}_p^{-1}(\boldsymbol{\theta}) \mathbf{y} + \log \det \mathbf{C}_p(\boldsymbol{\theta}) + N \log 2\pi] \right\},
\end{aligned} \tag{12}$$

and

$$\boldsymbol{\Sigma}_q \equiv -\nabla \nabla \log p(\mathbf{y}, \mathcal{D}_s \mid \boldsymbol{\theta})|_{\mathbf{y}=\boldsymbol{\mu}_q} = \mathbf{H} + \mathbf{C}_p^{-1}(\boldsymbol{\theta}), \tag{13}$$

where $\mathbf{H} \equiv -\nabla\nabla \log p(\mathcal{D}_s | \mathbf{y})|_{\mathbf{y}=\boldsymbol{\mu}_q}$ denotes the Hessian of the likelihood around $\boldsymbol{\mu}_q$. Note that $\boldsymbol{\Sigma}_q$ and $\boldsymbol{\mu}_q$ depend on $\boldsymbol{\theta}$ indirectly through the dependence of the MAP on $\boldsymbol{\theta}$.

In this work we solve the minimization problem (12) via gradient-based optimization. The necessary gradient of $\log p(\mathcal{D}_s | \mathbf{y})$ with respect to \mathbf{y} is computed via the discrete adjoint method described in Appendix C. The Hessian of $\log p(\mathcal{D}_s | \mathbf{y})$ with respect to \mathbf{y} , necessary to evaluate (13), is also computed via the discrete adjoint method.

We propose the Laplace-EM algorithm, where the Laplace approximation provides an approximation to the E-step. For the M-step, we keep the Laplace approximation fixed and maximize the ELBO with respect to the hyperparameters of the GP prior, $\boldsymbol{\theta}$. From (9) we see that for fixed $q(\mathbf{y})$, $\boldsymbol{\theta}$ appears only through the KL divergence $D_{\text{KL}}(q(\mathbf{y}) \| p(\mathbf{y} | \boldsymbol{\theta}))$; therefore, it suffices to minimize this KL divergence at the M-step. The Laplace-EM M-step reads

$$\boldsymbol{\theta}^{(j+1)} = \arg \min_{\boldsymbol{\theta}} D_{\text{KL}}(q(\mathbf{y}) \| p(\mathbf{y} | \boldsymbol{\theta})). \quad (14)$$

For the GP prior, this KL divergence is given in closed form by

$$D_{\text{KL}}(q(\mathbf{y}) \| p(\mathbf{y} | \boldsymbol{\theta})) = \frac{1}{2} \left[\text{tr}(\mathbf{C}_p^{-1} \boldsymbol{\Sigma}_q) + \boldsymbol{\mu}_q^\top \mathbf{C}_p^{-1} \boldsymbol{\mu}_q - N + \log \frac{\det \mathbf{C}_p}{\det \boldsymbol{\Sigma}_q} \right]. \quad (15)$$

If using a gradient method, the gradient of the KL divergence is given by

$$\frac{\partial}{\partial \theta_i} D_{\text{KL}}(q(\mathbf{y}) \| p(\mathbf{y} | \boldsymbol{\theta})) = -\frac{1}{2} \boldsymbol{\mu}_q^\top \mathbf{C}_p^{-1} \frac{\partial \mathbf{C}_p}{\partial \theta_i} \mathbf{C}_p^{-1} \boldsymbol{\mu}_q + \frac{1}{2} \text{tr} \left[\mathbf{C}_p^{-1} \frac{\partial \mathbf{C}_p}{\partial \theta_i} (\mathbf{I}_N - \mathbf{C}_p^{-1} \boldsymbol{\Sigma}_q) \right], \quad (16)$$

where \mathbf{I}_N denotes the $N \times N$ identity matrix.

The Laplace-EM algorithm is summarized in Algorithm 1. In practice, the EM iterations are halted once either a maximum number of iterations are completed, or once the relative change in hyperparameters is below a certain threshold, that is, when

$$\max \left\{ \left| \theta_i^{(j+1)} - \theta_i^{(j)} \right| / \left| \theta_i^s \right| \right\}_{i=1}^{N_\theta} \leq \text{rtol},$$

where N_θ is the number of prior hyperparameters, the θ_i^s , $i \in [1, N_\theta]$ are prescribed hyperparameter scales (that provide a sense of the magnitude of the hyperparameters), and rtol is the prescribed tolerance.

Algorithm 1 Laplace-EM.

Require: $\boldsymbol{\theta}^{(0)}$, $\mathbf{C}_p(\boldsymbol{\theta})$, $\log p(\mathcal{D}_s | \mathbf{y})$

$j \leftarrow 0$

repeat

 Compute $\boldsymbol{\mu}_q$ using (12)

 Compute $\boldsymbol{\Sigma}_q$ using (13)

 Solve (14) for $\boldsymbol{\theta}^{(j+1)}$

$j \leftarrow j + 1$

until Convergence

▷ E-step

▷ M-step

We note that the Laplace approximation is a commonly used tool for unimodal non-Gaussian inference [9,12,13]. Directly maximizing with respect to $\boldsymbol{\theta}$, the Laplace approximation to the marginal likelihood requires evaluating third-order derivatives of the log-likelihood function (3) with respect to \mathbf{y} . This is due to the implicit dependence of the MAP on $\boldsymbol{\theta}$, which requires evaluating third-order derivatives of the physics constraint. The use of the expectation maximization algorithm allows us to side-step the need of third-order derivatives. Other methods for non-Gaussian inference such as expectation propagation [17] require multiple evaluations of the moments of the likelihood function, and are therefore not considered in this work.

5. Doubly stochastic variational inference

In variational inference (VI) [23], we restrict our choice of $q(\mathbf{y})$ to a parameterized family $q(\mathbf{y} | \boldsymbol{\phi})$. In this context we refer to q as the *variational density* and $\boldsymbol{\phi}$ as the *variational parameters*. Following (10), we will estimate the variational parameters and the GP prior hyperparameters simultaneously by maximizing the corresponding ELBO, that is,

$$(\boldsymbol{\phi}, \boldsymbol{\theta}) = \arg \max_{(\boldsymbol{\phi}, \boldsymbol{\theta})} \mathcal{F}[q(\mathbf{y} | \boldsymbol{\phi}), \boldsymbol{\theta}]. \quad (17)$$

In this section we present our proposed implementation of variational inference for empirical Bayes. The main challenges of VI are (i) approximating the expectations on the expression for the ELBO, (9), and (ii) optimizing such approximations. To address these challenges we employ the *doubly stochastic variational inference* (DSVI) framework [15,16], in which a noisy

simple Monte Carlo estimate of the ELBO (1st source of stochasticity) is minimized via a gradient-based stochastic optimization algorithm (2nd source of stochasticity). In particular, we employ stochastic gradient ascent with the adaptive step-size sequence proposed by [24]. The gradients of the ELBO estimate with respect to variational parameters and prior hyperparameters are computed via Gaussian backpropagation [25,16,26,24].

5.1. Gaussian backpropagation

To maximize the ELBO via gradient-based stochastic optimization, we construct unbiased estimates of the ELBO and its gradients with respect to ϕ and θ . Computing the gradient $\nabla_{\phi} \mathcal{F}$ is not trivial as it involves expectations over q , which depends on ϕ .

We restrict ourselves to the multivariate Gaussian variational family $q(\mathbf{y} | \phi) = \mathcal{N}(\mathbf{y} | \mu_q, \Sigma_q)$, with variational mean $\mu_q \in \mathbb{R}^N$ and covariance $\Sigma_q = \mathbf{R}_q \mathbf{R}_q^T \in \mathbb{R}^{N \times N}$, where \mathbf{R}_q is a lower triangular factor matrix. For this choice we have $\phi = \{\mu_q, \mathbf{R}_q\}$. Similar to the Laplace approximation, this choice is justified for unimodal posteriors. We then introduce the change of variables $\mathbf{y} = \mu_q + \mathbf{R}_q \mathbf{z}$, with $\mathbf{z} \sim \mathcal{N}(0, \mathbf{I}_N)$. Substituting this change of variables into (9), we can rewrite the involved expectations in terms of expectations over $\mathcal{N}(\mathbf{z} | 0, \mathbf{I}_N)$, resulting in

$$\mathcal{F}[q(\mathbf{y} | \phi), \theta] = \mathbb{E}_{\mathcal{N}(\mathbf{z}|0, \mathbf{I}_N)}[\log p(\mathcal{D}_s | \mathbf{y})] + \mathbb{E}_{\mathcal{N}(\mathbf{z}|0, \mathbf{I}_N)}[\log p(\mathbf{y} | \theta)] + \log \det \mathbf{R}_q + \mathcal{H}[\mathcal{N}(\mathbf{z}|0, \mathbf{I}_N)], \quad (18)$$

where $\mathbf{y} = \mu_q + \mathbf{R}_q \mathbf{z}$, and $\mathcal{H}[\mathcal{N}(\mathbf{z} | 0, \mathbf{I}_N)] = N(1 + \log 2\pi)/2$ is the differential entropy of the standard multivariate normal. We then define the following unbiased estimate of the ELBO,

$$f(\mathbf{z}; \phi, \theta) = \log p(\mathcal{D}_s | \mathbf{y}) + \log \det \mathbf{R}_q - \frac{1}{2} \left[\mathbf{y}^T \mathbf{C}_p^{-1} \mathbf{y} + \log \det \mathbf{C}_p - N \right], \quad (19)$$

with gradients

$$\nabla_{\mu_q} f(\mathbf{z}; \phi, \theta) = \nabla_{\mathbf{y}} \log p(\mathcal{D}_s | \mathbf{y}) - \mathbf{C}_p^{-1} \mathbf{y}, \quad (20)$$

$$\nabla_{\mathbf{R}_q} f(\mathbf{z}; \phi, \theta) = [\nabla_{\mathbf{y}} \log p(\mathcal{D}_s | \mathbf{y})] \mathbf{z}^T - \mathbf{C}_p^{-1} \mathbf{y} \mathbf{z}^T + (\mathbf{R}_q^{-1})^T, \quad (21)$$

$$\frac{\partial}{\partial \theta_i} f(\mathbf{z}; \phi, \theta) = \frac{1}{2} \mathbf{y}^T \mathbf{C}_p^{-1} \frac{\partial \mathbf{C}_p}{\partial \theta_i} \mathbf{C}_p^{-1} \mathbf{y} - \frac{1}{2} \text{tr} \left(\mathbf{C}_p^{-1} \frac{\partial \mathbf{C}_p}{\partial \theta_i} \right), \quad (22)$$

where again $\mathbf{y} = \mu_q + \mathbf{R}_q \mathbf{z}$. The details of the derivations of (20)–(22) are presented in Appendix A. It can be verified that $\mathbb{E}_{\mathcal{N}(\mathbf{z}|0, \mathbf{I}_N)}[f(\mathbf{z}; \phi, \theta)] = \mathcal{F}[q(\mathbf{y} | \phi), \theta]$, so that the estimates are unbiased.

The variance of the estimate (19) and its gradients can be reduced by using the simple Monte Carlo (MC) or batch estimate and the corresponding gradients

$$f_n(\phi, \theta) = \frac{1}{n} \sum_{k=1}^n f(\mathbf{z}^{(k)}; \phi, \theta), \quad \mathbf{z}^{(k)} \sim \mathcal{N}(0, \mathbf{I}_N), \quad (23)$$

$$\nabla_{(\cdot)} f_n(\phi, \theta) = \frac{1}{n} \sum_{k=1}^n \nabla_{(\cdot)} f(\mathbf{z}^{(k)}; \phi, \theta), \quad (24)$$

where n is the size of the batch. The variance of the batch estimate (23) is lower by a factor of n , but requires computing the gradients of the log-likelihood n times.

Our DSVI algorithm is summarized in Algorithm 2. The stochastic gradient ascent algorithm with adaptive step-size sequence, proposed by [24], is reproduced in Appendix B for completeness.

Algorithm 2 Doubly stochastic variational inference.

Require: $\phi^{(0)}, \theta^{(0)}, \mathbf{C}_p(\theta), \log p(\mathcal{D}_s | \mathbf{y})$

$j \leftarrow 0$

repeat

 Sample n realizations of $\mathbf{z} \sim \mathcal{N}(0, \mathbf{I}_N)$

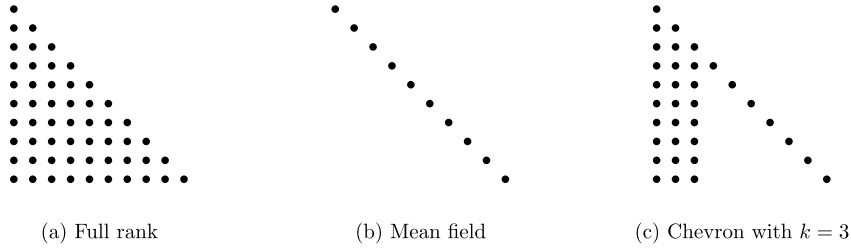
 Compute $\nabla_{\phi} f_n(\phi^{(j)}, \theta^{(j)})$ and $\nabla_{\theta} f_n(\phi^{(j)}, \theta^{(j)})$ using (20)–(22) and (24)

 Calculate step-size vectors $\rho_{\phi}^{(j)}$ and $\rho_{\theta}^{(j)}$ using (B.3) and (B.4)

$\phi^{(j+1)} \leftarrow \phi^{(j)} + \rho_{\phi}^{(j)} \circ \nabla_{\phi} f_n(\phi^{(j)}, \theta^{(j)})$ (B.1)

$\theta^{(j+1)} \leftarrow \theta^{(j)} + \rho_{\theta}^{(j)} \circ \nabla_{\theta} f_n(\phi^{(j)}, \theta^{(j)})$ (B.2)

until Convergence

Fig. 1. Parameterization of the factor matrix \mathbf{R}_q .

5.2. Parameterization of the factor matrix \mathbf{R}_q

It remains to discuss the parameterization of the factor \mathbf{R}_q . In this manuscript, we consider three alternatives: a full parameterization, the so-called *mean field* parameterization, and a constrained Chevron parameterization. The sparsity patterns of these parameterizations are shown in Fig. 1.

In the *full rank* parameterization [24], we take \mathbf{R}_q to be the non-unique Cholesky factor, that is, a $N \times N$ lower triangular matrix with unconstrained entries (Fig. 1a). In this case, we have $\phi \in \mathbb{R}^{N+N(N+1)/2}$. The number of variational parameters is therefore $O(N^2)$, which may render their optimization difficult. In order to address this challenge, we can employ the mean field and Chevron parameterizations, which result in a total number of variational parameters that is linear on N .

In the mean field parameterization, we take \mathbf{R}_q to be a strictly positive diagonal matrix, i.e., $\mathbf{R}_q = \text{diag}[\exp(\omega_q)]$, with $\omega_q \in \mathbb{R}^N$ (Fig. 1b) and $\exp(\cdot)$ understood as element-wise. This parameterization assumes that the variational density covariance is diagonal, and the exponential ensures that the non-zero entries of \mathbf{R}_q are strictly positive. In this case, we have $\phi \equiv \{\mu_q, \omega_q\} \in \mathbb{R}^{2N}$. The gradient of $f(\mathbf{z}; \phi, \theta)$ with respect to ω_q is given by

$$\nabla_{\omega_q} f(\mathbf{z}; \phi, \theta) = [-\nabla_{\mathbf{y}} \log p(\mathcal{D}_s | \mathbf{y}) + \mathbf{C}_p^{-1} \mathbf{y}] \circ \mathbf{z} \circ \exp(\omega_q) - \mathbf{I}_N, \quad (25)$$

where \circ denotes element-wise product, and exponentiation is taken as element-wise. The details of the derivation are presented in Appendix A. This parameterization assumes that the posterior components of \mathbf{y} are essentially uncorrelated and therefore cannot resolve the correlations of the true posterior, which are expected to be non-trivial for highly correlated prior covariance structures and for a small number of observations. As a consequence, the mean field parameterization tends to underestimate the posterior variance [23].

Finally, the constrained Chevron parameterization is similar to the full parameterization, but we set entries below the diagonal and for column number larger than the Chevron parameter $k < N$ to zero (Fig. 1c) [27]. In this case, we have $\phi \in \mathbb{R}^{N+(2N-k)(k+1)/2}$. The number of variational parameters for this parameterization is $O(N(k+1))$, a reduction with respect to the full parameterization, while maintaining some degree of expressivity for capturing correlations of the true posterior.

The DSVI-EB method follows the *automatic differentiation variational inference* (ADVI) algorithm [24], in which gradients of the joint probability $p(\mathcal{D}_s | \mathbf{y})$ with respect to variational parameters are computed using Gaussian backpropagation and reverse-mode automatic differentiation. ADVI is formulated for the full Bayes case and implements the full and mean-field parameterizations of \mathbf{R}_q . In comparison, our work is formulated for the empirical Bayes case, employs the adjoint method to compute gradients of physics solvers, and implements the constrained Chevron parameterization in addition to the full and mean-field parameterizations.

Two schemes are common in the literature for computing the gradients of the noisy ELBO estimate: the REINFORCE algorithm [28], also known as the likelihood ratio method or the log-derivative trick, and Gaussian backpropagation [26], also known as the reparameterization trick [25,16]. The REINFORCE algorithm employs gradients of the variational density with respect to its parameters, which is convenient as it only requires zero-order information of the physics model. Unfortunately the REINFORCE estimates of the ELBO gradients are well-known to be of high variance and must be paired with a variance reduction technique [15]. Gaussian backpropagation, on the other hand, results in lower-variance gradient estimates at the cost of requiring first-order information of the physics model.

An alternative formulation of VI is presented in [18], where the authors employ mixtures of diagonal multivariate Gaussian densities as the variational posterior, and approximate the ELBO using a second order Taylor expansion around the mean of each mixture component. The mean, diagonal covariance and mixture weights are estimated by maximizing the ELBO via coordinate ascent. This entirely deterministic approach is formulated for the inference problem and does not consider optimization over prior hyperparameters. A similar approach is also presented in [29] in the context of empirical Bayes.

6. Computational cost

In this section, we discuss the computational effort of the Laplace-EM and DSVI-EB algorithms. We compute the gradient and the Hessian of the likelihood via the discrete adjoint method (see Appendix C for details). Note that the Laplace-EM

method requires both gradients and Hessians, while the DSVI-EB method only requires gradients. For the physics constraint $\mathbf{L}(\mathbf{u}, \mathbf{y}) = 0$, the computation of the gradient requires the solution of one (linear) backward sensitivity problem of size $M \times M$. Similarly, the computation of the Hessian requires one backward sensitivity problem and N forward sensitivity problems, each of size $M \times M$. For the following discussion we assume that the cost of each forward and backward sensitivity problem is of order $O(M^\gamma)$, $\gamma > 1$.

For the Laplace-EM algorithm we discuss the cost per each EM cycle. Each E-step requires one Cholesky factorization of $\mathbf{C}_p(\theta^{(j)})$, of cost N^3 , the solution of (12) via gradient-based optimization, and the computation of the Hessian. Therefore, the cost of each E-step is $O(\max(M^\gamma, N^3))$. Each iteration of the M-step requires one Cholesky factorization of $\mathbf{C}_p(\theta)$. Therefore, the total cost of each EM cycle is again $O(\max(M^\gamma, N^3))$.

For the DSVI-EB algorithm, each iteration requires evaluating n gradients and one Cholesky factorization of \mathbf{C}_p . If n is chosen independent of M , we have that the total cost per iteration is also $O(\max(M^\gamma, N^3))$.

Finally, in general we expect the number of iterations for each E- and M-step, and the number of EM cycles and DSVI iterations, to increase with increasing N . The analysis of how said numbers scale with N is beyond the scope of this manuscript.

7. Numerical experiments

In this section, we present the application of the proposed approximate inference algorithms to the identification of the diffusion coefficient in diffusion equations. In particular, we are interested in identifying the diffusion coefficient $k(\mathbf{x}, u)$ of the homogeneous diffusion equation $\nabla \cdot (k(\mathbf{x}, u) \nabla u) = 0$ in $\Omega \subset \mathbb{R}^d$, from both measurements of the diffusion coefficient and of the state u . For the linear case ($k \equiv k(\mathbf{x})$), the diffusion equation models phenomena such as stationary heat transfer and Darcy flow. For the nonlinear case ($k \equiv k(u)$), one recovers the so-called Richards equation for horizontal flows in unsaturated porous media.

7.1. Linear diffusion problem

We consider the one-dimensional diffusion equation with Dirichlet boundary conditions

$$\frac{\partial}{\partial x} \left[k(x) \frac{\partial}{\partial x} u(x) \right] = 0, \quad x \in [0, 1], \quad (26)$$

$$u(0) = u_L, \quad u(1) = u_R, \quad (27)$$

where $u: [0, 1] \rightarrow \mathbb{R}$ is the state and $k: [0, 1] \rightarrow \mathbb{R}^+$ is the diffusion coefficient. The state is discretized into M degrees of freedom u_i organized into the vector $\mathbf{u} \in \mathbb{R}^M$. The diffusion coefficient is discretized into N degrees of freedom $k_i = \exp y_i$ corresponding to N spatial coordinates $\{x_i\}_{i=1}^N$, with the y_i organized into the vector $\mathbf{y} \in \mathbb{R}^N$. The discretized problem (26) and (27) is of algebraic form $\mathbf{L}(\mathbf{u}, \mathbf{y}) \equiv \mathbf{S}(\mathbf{y})\mathbf{u} - \mathbf{b}(\mathbf{y}) = 0$, where $\mathbf{S}: \mathbb{R}^N \rightarrow \mathbb{R}^{M \times M}$ and $\mathbf{b}: \mathbb{R}^N \rightarrow \mathbb{R}^M$. In (26), \mathbf{y} can only be identified from measurements of u up to an additive constant [1], and measurements of \mathbf{y} are required to estimate it uniquely.

We apply the presented model inversion algorithms to estimating two synthetic diffusion coefficients from measurements of the state u and the log-diffusion coefficient y . The synthetic \mathbf{y} fields are generated as realization of zero-mean GPs with two different covariance kernels. The first reference field is generated with the squared exponential (SE) covariance kernel

$$\mathbf{C}_{\text{SE}}(x, x' | \theta) = \sigma^2 \exp \left[- (x - x')^2 / 2\lambda^2 \right] + \sigma_n^2 1_{x=x'}, \quad (28)$$

and the second reference field is generated with the Matérn covariance kernel with shape coefficient $\nu = 3/2$ (M32) [9],

$$\mathbf{C}_{\text{M32}}(x, x' | \theta) = \sigma^2 \left(1 + \frac{\sqrt{3}|x - x'|}{\lambda} \right) \exp \left(- \frac{\sqrt{3}|x - x'|}{\lambda} \right) + \sigma_n^2 1_{x=x'}. \quad (29)$$

For the SE and M32 covariances, $\theta \equiv (\sigma, \lambda)$, and σ_n is set to 1×10^{-2} . We refer to σ and λ as the standard deviation and correlation length, respectively, of the prior covariances.

The \mathbf{y} and \mathbf{u} observations are taken at randomly selected degrees of freedom, with observation error standard deviations $\sigma_{us} = \sigma_{ys} = 1 \times 10^{-3}$. The reference values of \mathbf{y} and \mathbf{u} and the corresponding observations are shown in Fig. 2. It can be seen that due to the properties of the SE and M32 kernels, the SE field is smoother than the M32 field. For the SE reference field, 10 observations of \mathbf{y} and 1 observation of \mathbf{u} are taken. For the M32 reference field, which is less smooth, 15 observations of \mathbf{y} and 1 observation of \mathbf{u} are taken. The reference values of θ for the SE and M32 fields are presented in Tables 1 and 2, respectively. Finally, the boundary conditions are set to $u_L = 1.0$ and $u_R = 0.0$, and the numbers M and N of degrees of freedom are set to 50.

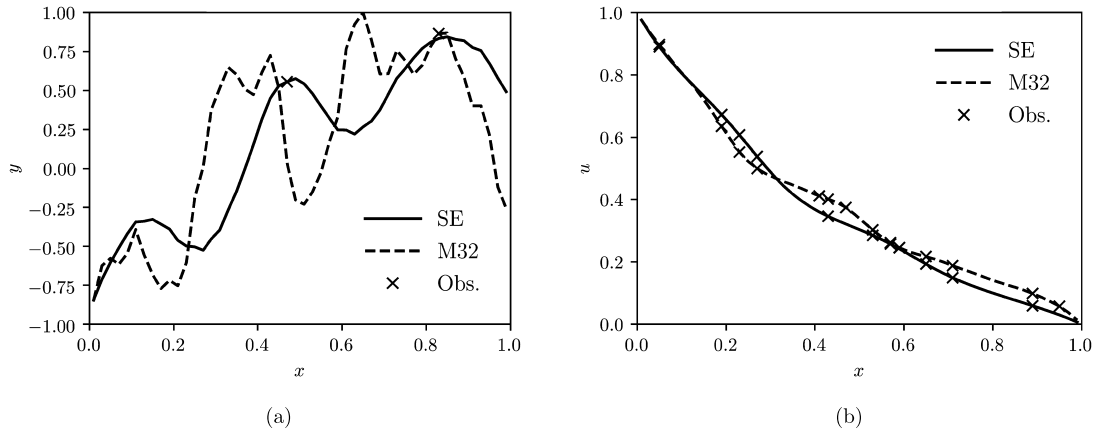


Fig. 2. Reference diffusion coefficients and state fields (lines), and observations (crosses), for the one-dimensional linear diffusion problem. Squared-exponential (SE) reference fields in continuous lines. Matérn 3/2 reference fields in dashed lines.

Table 1

Reference and estimated hyperparameters, and simple MC estimate of the ELBO, for the one-dimensional linear-diffusion problem and the SE reference field.

		Hyperparameters	
		$\hat{\mathcal{F}}$	λ
Reference			
Laplace-EM		−37.37(5)	0.608
DSVI	Full rank	−43.38(8)	0.551
	Chevron $k = 20$	−49.54(9)	0.653
	Chevron $k = 10$	−50.47(9)	0.672
	Chevron $k = 5$	−47.85(6)	0.681
	Mean field	−49.34(6)	0.687

Table 2

Reference and estimated hyperparameters, and simple MC estimate of the ELBO, for the one-dimensional linear-diffusion problem and the M32 reference field.

		Hyperparameters	
		$\hat{\mathcal{F}}$	λ
Reference			
Laplace-EM		−64.72(34)	0.629
DSVI	Full rank	−62.82(14)	0.576
	Chevron $k = 20$	−71.09(14)	0.637
	Chevron $k = 10$	−72.50(12)	0.642
	Chevron $k = 5$	−80.09(17)	0.598
	Mean field	−81.56(13)	0.677

7.1.1. Empirical Bayesian inference

Figs. 3 and 4 show the estimated diffusion coefficient for the SE and M32 fields, respectively, together with the 95% confidence intervals centered around the posterior mean. The estimates are computed using Laplace-EM and DSVI-EB with Chevron parameterization and $k = 20$. It can be seen that for both the SE and M32 fields, both methods provide accurate estimates of the reference field, with the reference field falling inside the confidence interval of the estimates (with localized exceptions for the SE field and DSVI-EB with Chevron parameterization in the vicinity of the $x = 1.0$ boundary, as shown in Fig. 3b). It can also be seen that, compared to Laplace-EM, DSVI-EB with Chevron parameterization and $k = 20$ leads to lower estimates of the posterior standard deviation. The accuracy of standard deviation estimates is discussed in more detail in Section 7.1.2.

Tables 1 and 2 show the estimated prior hyperparameters for the SE and M32 fields, respectively, together with simple MC estimates of the ELBO, $\hat{\mathcal{F}}$, computed using 1×10^4 realizations of the estimated posterior. It can be seen that hyperparameter estimates are similar for Laplace-EM and the DSVI-EB. In particular, estimates of the correlation length are close to reference values, while the standard deviation is underestimated across all methods and for both reference fields. Estimates are also similar for the different factor matrix parameterizations of DSVI-EB, with the exception of the full rank parameterization that resulted in more pronounced underestimation of both the standard deviation and correlation length.

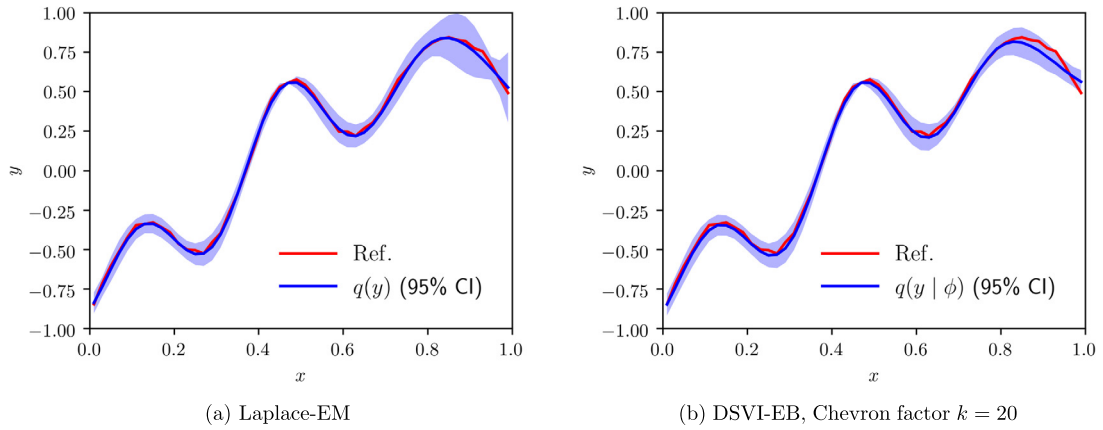


Fig. 3. Reference and estimated diffusion coefficient for the one-dimensional linear diffusion problem and the SE reference field. (For interpretation of the colors in the figure(s), the reader is referred to the web version of this article.)

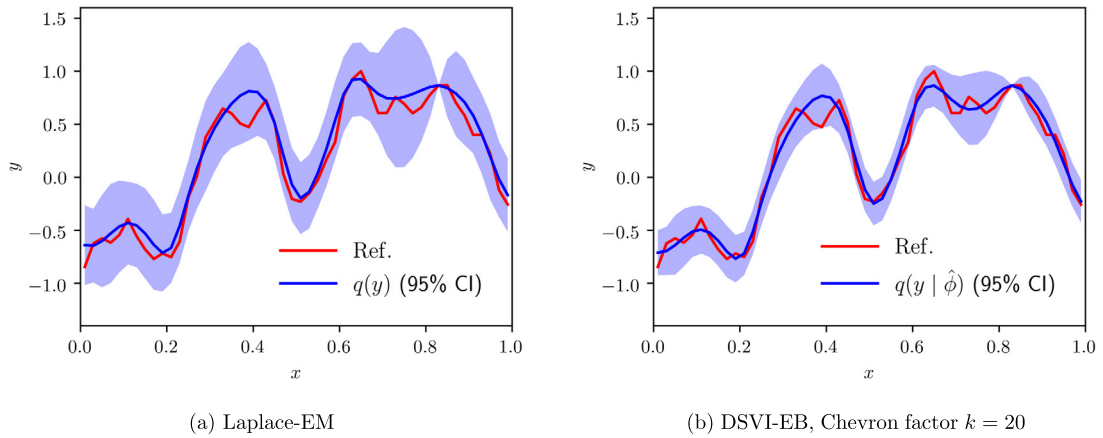


Fig. 4. Reference and estimated diffusion coefficient for the one-dimensional linear diffusion problem and the M32 reference field.

In terms of the ELBO, for the SE field, Laplace-EM results in the highest value, followed DSVI-EB with full rank factor parameterization. For the M32 field, DSVI-EB with full rank factor parameterization leads to the highest ELBO, followed by Laplace-EM. These results indicate that full rank representations of the covariance matrix of the estimated posterior density result in better estimates of the true posterior, whereas reduced representations such as Chevron and mean field are less accurate. In practice, it can be seen in Fig. 3b that the reduced representation is less capable of resolving the uncertainty of the y estimate in the vicinity of the $x = 1.0$ boundary. Nevertheless, reduced representations are not significantly worse than fuller representations for estimating prior hyperparameters.

7.1.2. Comparison against MCMC

We proceed to evaluate the accuracy of the proposed inference algorithms at approximating the posterior density $p(\mathbf{y} | \mathcal{D}_s, \theta)$. For this purpose, we employ MCMC simulation as the benchmark as it is known to converge to the exact posterior density. In order to restrict the focus to the approximation of the posterior density, we set the prior hyperparameters to fixed values equal to the reference values, θ_{ref} , and employ the Laplace-EM¹ and DSVI algorithms to estimate the posterior $p(\mathbf{y} | \mathcal{D}_s, \theta_{\text{ref}})$. We compare the estimated posterior mean and standard deviation against the sample mean and standard deviation computed from 1×10^4 MCMC realizations of the posterior generated using the No-U-Turn Sampler (NUTS) [8]. Due to the computational cost of MCMC sampling, we were unable to obtain MCMC samples of the posterior for the M32 field; therefore, the comparison between NUTS, Laplace-EM, and DSVI is performed for the SE field.

Fig. 5 presents a point-wise comparison against MCMC of the estimated mean and standard deviation computed using both Laplace-EM and DSVI with the full rank parameterization and the Chevron parameterization with $k = 20$ and 5. It can be seen that all estimates of the mean are very accurate, which indicates that the presented approximate inference algo-

¹ Note that in this context the Laplace-EM algorithm is reduced to the Laplace approximation for given θ (e.g. a single E-step in the EM algorithm), but we will refer to the associated results as Laplace-EM results for the sake of convenience.

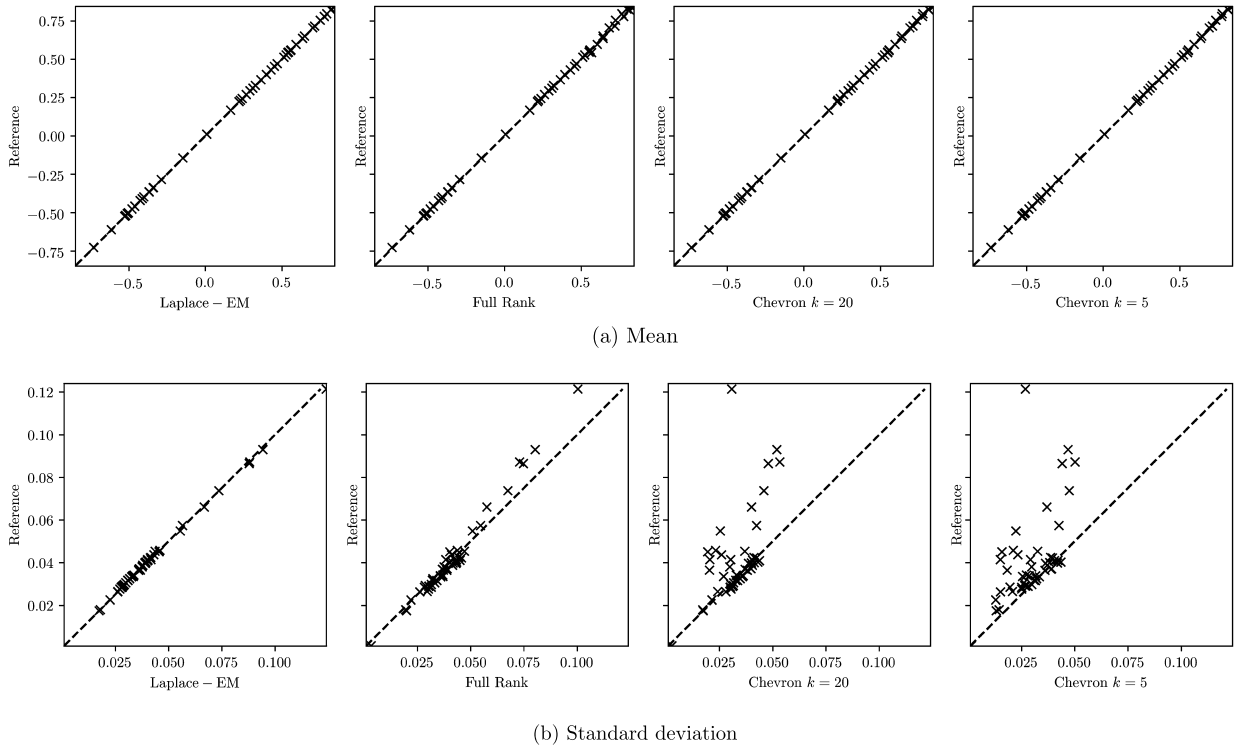


Fig. 5. Posterior mean and standard deviation estimated via approximate inference, compared against sample mean and standard deviation computed from MCMC realizations of the posterior (reference).

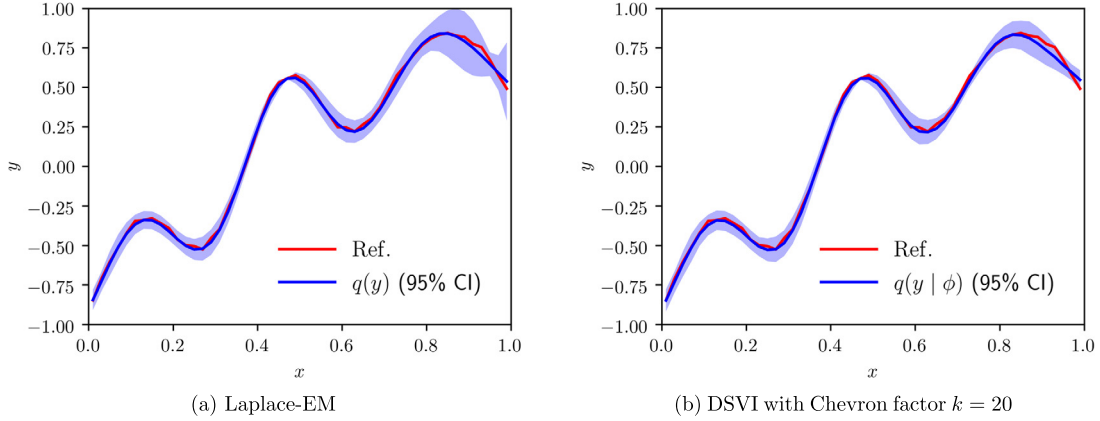


Fig. 6. Estimated posterior mean and 95% confidence interval computed from the estimated posterior variance, for the one-dimensional linear diffusion problem for fixed $\theta = \theta_{\text{ref}}$.

rihtms provide accurate estimates of the mean of unimodal posterior densities. This result is expected for the Laplace-EM algorithm where the estimated posterior mean is set to the MAP, but for the DSVI algorithm this is less of a given.

For the standard deviation, the Laplace-EM method provides the most accurate estimates, followed by DSVI with the full rank parameterization. This reinforces the conclusion drawn previously that full rank representations of the covariance lead to better estimates of the true posterior. Furthermore, it can be seen that the Chevron representation accurately resolves the bulk of point-wise standard deviation values (clustered at the bottom left of each plot in Fig. 5b) but leads to noticeable underestimation of the larger point-wise values (i.e. the top half of Fig. 5b). The underestimation of the standard deviation is more pronounced for decreasing k , and is the most pronounced for the mean field parameterization (not shown), which as remarked previously tends to underestimate the variance of the posterior [23].

Finally, in Fig. 6 we present the posterior mean and variance for Laplace-EM and DSVI with Chevron parameterization and $k=20$, obtained for fixed $\theta = \theta_{\text{ref}}$. Comparing Fig. 3 against Fig. 6 reveals that even though the empirical Bayes estimation procedure results in a standard deviation estimate lower than the reference value (see Table 1), the posterior density with

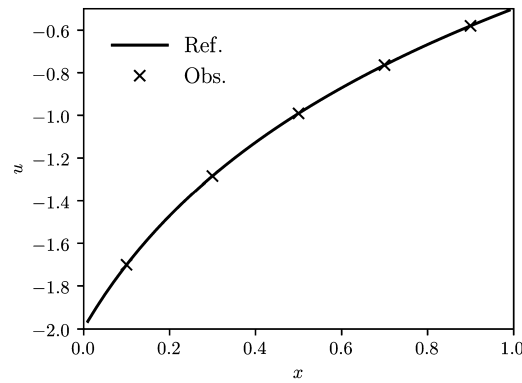


Fig. 7. Reference state field (continuous lines), and observations (crosses), for the one-dimensional nonlinear diffusion problem.

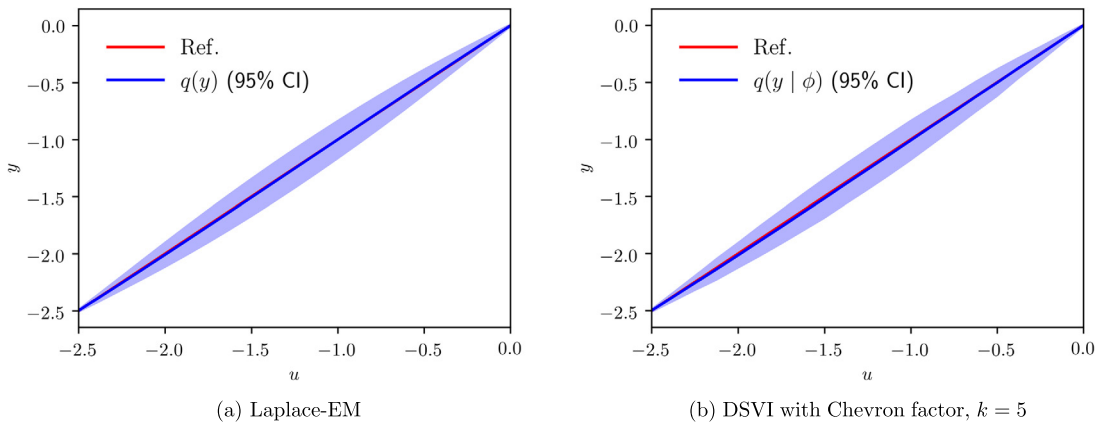


Fig. 8. Estimated diffusion coefficient for the one-dimensional nonlinear diffusion problem.

empirical Bayes hyperparameter estimates is a good approximate to the posterior density with reference hyperparameters, $p(\mathbf{y} | \mathcal{D}_s, \theta_{\text{ref}})$.

7.2. Nonlinear diffusion problem

We consider the one-dimensional nonlinear diffusion equation with Dirichlet boundary conditions

$$\frac{\partial}{\partial x} \left[k(u(x)) \frac{\partial}{\partial x} u(x) \right] = 0, \quad x \in [0, 1], \quad (30)$$

$$u(0) = u_L, \quad u(1) = u_R, \quad u_L < u_R \leq 0, \quad (31)$$

where $u: [0, 1] \rightarrow (-\infty, 0]$ is the state and $k: (\infty, 0] \rightarrow \mathbb{R}^+$ is the diffusion coefficient. Similarly to Section 7.1, the state is discretized into M degrees of freedom u_i organized into the vector $\mathbf{u} \in \mathbb{R}^M$. The diffusion coefficient function is discretized into N degrees of freedom $k_i = \exp y_i$ corresponding to N values of u over $[u_{\min}, 0]$ (where $u_{\min} < u_L$), organized into the vector $\mathbf{y} \in \mathbb{R}^N$. The discretized problem (30) and (31) is of the algebraic form $\mathbf{L}(\mathbf{u}, \mathbf{y}) = \mathbf{S}(\mathbf{u}, \mathbf{y})\mathbf{u} - \mathbf{b}(\mathbf{u}, \mathbf{y}) = 0$, where $\mathbf{S}: \mathbb{R}^M \times \mathbb{R}^N \rightarrow \mathbb{R}^{M \times M}$ and $\mathbf{b}: \mathbb{R}^M \times \mathbb{R}^N \rightarrow \mathbb{R}^M$. Inspection of (30) reveals that $y(u) \equiv \log k(u)$ can only be identified over the range $[u_L, u_R]$ and up to an additive constant.² To disambiguate the estimate, we provide measurements of $y(u)$ at $u = u_{\min}$ and $u = 0$.

We apply the presented model inversion algorithms to estimating a function $k(u)$ from 5 measurements of the state u and 2 measurements of $y(u) \equiv \log k(u)$ at $u = u_{\min}$ and $u = 0$. The reference diffusion coefficient is $k(u) = \exp u$ ($y(u) = u$). The \mathbf{u} observations are taken at randomly selected degrees of freedom, and are shown in Fig. 7. Observation error standard deviations σ_{us} and σ_{ys} are set to 1×10^{-2} . Boundary conditions are set to $u_L = -2.0$ and $u_R = -0.5$, and u_{\min} is set to -2.5 . Finally, the numbers M and N are set to 50 and 21, respectively.

Fig. 8 shows the estimated diffusion coefficient using the proposed model inversion methods, together with the 95% confidence intervals centered around the posterior mean. Presented are the results for the Laplace-EM method and the DSVI-EB

² This can be verified by introducing the Kirchhoff transformation $f(u) = \int_{u_{\min}}^u k(u) du$, with which (30) can be written as a linear equation on f .

Table 3

Reference and estimated hyperparameters, and simple MC estimate of the ELBO, for the one-dimensional nonlinear-diffusion problem.

		$\hat{\mathcal{F}}$	Hyperparameters	
			σ	λ
Laplace-EM		−12.68(3)	4.650	6.893
DSVI	Full rank	−13.97(4)	4.024	5.968
	Chevron $k = 10$	−14.56(4)	4.007	6.265
	Chevron $k = 5$	−15.02(4)	4.343	6.778
	Chevron $k = 2$	−16.23(4)	4.768	7.516
	Mean field	−16.66(4)	5.353	9.297

method with Chevron parameterization and $k = 5$. It can be seen that the estimated posterior mean and confidence intervals for both methods are nearly identical. As prior covariance $C(u, u | \theta)$, we employ the squared exponential model (28) with σ_n set to 1×10^{-2} . As in the linear case, both methods accurately estimate the reference function $y(u)$, and the reference function falls inside the 95% confidence intervals provided by the estimated posterior covariance.

Table 3 presents the estimated hyperparameters of the prior for the Laplace-EM and the DSVI-EB method, together with simple MC estimates of the ELBO computed using 1×10^4 realizations of the corresponding estimated posterior densities. It can be seen that estimated hyperparameters are different for the different methods (note that here we don't have reference values for the hyperparameters of the prior, as the reference $k(u)$ is not drawn from a GP model). Nevertheless, it can be seen that both methods result in similar estimates of y . In agreement with the linear case, the Laplace-EM and the DSVI-EB method with full rank parameterization result in the largest values of ELBO. Additionally, it can be seen that the ELBO decreases with increasing sparsity of the posterior covariance factor parameterization (i.e. with decreasing Chevron factor k), being the lowest for the mean field parameterization. This illustrates the compromise between the sparsity of the covariance factor and its expressive capacity for approximating the true posterior, that is, that less sparse covariance factors produce more accurate approximate posteriors.

8. Conclusions and discussion

We have presented two approximate empirical Bayesian methods, Laplace-EM and DSVI-EB, for estimating unknown parameters and constitutive relations in PDE models. Compared to other methods for approximate Bayesian inference, the proposed methods do not require third-order derivatives of the physics model, do not involve computing moments of non-Gaussian likelihoods, and are applicable to non-factorizing likelihoods. Furthermore, the calculation of the batch estimate of the ELBO and its gradients employed in the DSVI-EB method is trivially parallelizable, leading to savings in computational time. The numerical experiments presented show that both methods accurately approximate the posterior density and the hyperparameters of the GP prior. In particular, we find that the Laplace-EM method more accurately approximates the posterior density, at the cost of computing Hessians of the physics model, which increase the computational cost of each EM cycle. The DSVI-EB method, on the other hand, tends to underestimate the posterior standard deviation, but does not require Hessians. Consistent with the literature, we find that the accuracy of the DSVI-EB method at approximating the posterior decreases with increasing sparsity of the covariance factor parameterization employed.

We note that the presented formulation and numerical experiments were restricted to unimodal posterior densities. For multimodal posteriors, Gaussian mixtures can be employed as variational densities in variational inference [18]. The extension of DSVI-EB to Gaussian mixture variational families will be considered in future work. It remains an open question how to apply the Laplace approximation for multimodal posteriors [30]. A possible approach is to employ a mixture of Laplace approximations for each mode. We will pursue this avenue in future work.

For a very large number of degrees of freedom of the discretization of the unknown functions, the computational cost of the proposed methods is dominated by the associated cubic complexity. Future work will aim to address the challenge of cubic complexity by employing sparse GP inference.

Acknowledgements

This work was supported by the Applied Mathematics Program within the U.S. Department of Energy Office of Advanced Scientific Computing Research under Contract DE-AC02-06CH11347. Pacific Northwest National Laboratory is operated by Battelle for the DOE under Contract DE-AC05-76RL01830.

Appendix A. Gaussian backpropagation rules

In this section we present the derivation of the gradients (20)–(22).

For the gradient with respect to the variational mean, (20), we have by the chain rule, in index notation,

$$\frac{\partial}{\partial \mu_{q,i}} \log p(\mathcal{D}_s | \mathbf{y}) = \frac{\partial y_j}{\partial \mu_{q,i}} \frac{\partial}{\partial y_j} \log p(\mathcal{D}_s | \mathbf{y}) = \delta_{ji} \frac{\partial}{\partial y_j} \log p(\mathcal{D}_s | \mathbf{y}),$$

where we have used $\partial y_j / \partial \mu_{q,i} = \delta_{ji}$. It follows that $\nabla_{\mu_q} \log p(\mathcal{D}_s | \mathbf{y}) = \nabla_{\mathbf{y}} \log p(\mathcal{D}_s | \mathbf{y})$. Similarly, we have

$$\nabla_{\mu_q} \frac{1}{2} \mathbf{y}^\top \mathbf{C}_p^{-1} \mathbf{y} = \nabla_{\mathbf{y}} \frac{1}{2} \mathbf{y}^\top \mathbf{C}_p^{-1} \mathbf{y} = \mathbf{C}_p^{-1} \mathbf{y},$$

and thus we recover (20).

For the gradient with respect to the Cholesky factor \mathbf{R}_q , we note that $\partial y_k / \partial R_{q,ij} = \delta_{ki} z_j$. By the chain rule, we have

$$\frac{\partial}{\partial R_{q,ij}} \log p(\mathcal{D}_s | \mathbf{y}) = \frac{\partial y_k}{\partial R_{q,ij}} \frac{\partial}{\partial y_k} \log p(\mathcal{D}_s | \mathbf{y}) = \frac{\partial}{\partial y_i} \log p(\mathcal{D}_s | \mathbf{y}) z_j,$$

so that $\nabla_{\mathbf{R}_q} \log p(\mathcal{D}_s | \mathbf{y}) = [\nabla_{\mathbf{y}} \log p(\mathcal{D}_s | \mathbf{y})] \mathbf{z}^\top$. Similarly,

$$\nabla_{\mathbf{R}_q} \frac{1}{2} \mathbf{y}^\top \mathbf{C}_p^{-1} \mathbf{y} = \left[\nabla_{\mathbf{y}} \frac{1}{2} \mathbf{y}^\top \mathbf{C}_p^{-1} \mathbf{y} \right] \mathbf{z}^\top = \mathbf{C}_p^{-1} \mathbf{y} \mathbf{z}^\top.$$

Finally, we have $\nabla_{\mathbf{R}_q} \log \det \mathbf{R}_q = (\mathbf{R}_q^{-1})^\top$, from which we recover (21).

The gradients with respect to the prior hyperparameters, (22) can be derived from [9], Eqs. (A.14) and (A.15).

For the gradient with respect to ω_q of the mean-field parameterization $\mathbf{R}_q = \text{diag}[\exp \omega_q]$, we employ the relation

$$\frac{\partial R_{q,ij}}{\partial \omega_{q,k}} = \begin{cases} \exp \omega_{q,k} & \text{for } k = i = j, \\ 0 & \text{otherwise.} \end{cases}$$

By the chain rule, we have

$$\frac{\partial}{\partial \omega_{q,k}} \log p(\mathcal{D}_s | \mathbf{y}) = \frac{\partial R_{q,ij}}{\partial \omega_{q,k}} \frac{\partial}{\partial R_{q,ij}} \log p(\mathcal{D}_s | \mathbf{y}) = \frac{\partial}{\partial y_k} \log p(\mathcal{D}_s | \mathbf{y}) z_k \exp \omega_{q,k},$$

summation over k not implied. It follows that $\nabla_{\omega_q} \log p(\mathcal{D}_s | \mathbf{y}) = [\nabla_{\mathbf{y}} \log p(\mathcal{D}_s | \mathbf{y})] \circ \mathbf{z} \circ \exp \omega_q$. Similarly,

$$\frac{\partial}{\partial \omega_{q,k}} \frac{1}{2} \mathbf{y}^\top \mathbf{C}_p^{-1} \mathbf{y} = \frac{\partial R_{q,ij}}{\partial \omega_{q,k}} (\mathbf{C}_p^{-1})_{im} y_m z_j = (\mathbf{C}_p^{-1})_{km} y_m z_k \exp \omega_{q,k},$$

summation over k not implied. Finally,

$$\nabla_{\omega_q} \log \det \mathbf{R}_q = \nabla_{\omega_q} \log \prod_k \exp \omega_{q,k} = \nabla_{\omega_q} \sum_k \log \exp \omega_{q,k} = \mathbf{I}_N,$$

from which we recover (25).

Appendix B. Stochastic gradient ascent with adaptive step-size sequence

Here we reproduce for completeness the stochastic gradient ascent algorithm with adaptive step-size sequence proposed in [24]. The presentation is expanded to the empirical Bayes context for the update of prior hyperparameters. At each iteration, the variational parameters and prior hyperparameters are updated using the rules

$$\boldsymbol{\phi}^{(j+1)} = \boldsymbol{\phi}^{(j)} + \boldsymbol{\rho}_{\boldsymbol{\phi}}^{(j)} \circ \nabla_{\boldsymbol{\phi}} f_n^{(j)}, \quad (\text{B.1})$$

$$\boldsymbol{\theta}^{(j+1)} = \boldsymbol{\theta}^{(j)} + \boldsymbol{\rho}_{\boldsymbol{\theta}}^{(j)} \circ \nabla_{\boldsymbol{\theta}} f_n^{(j)}, \quad (\text{B.2})$$

where $f_n^{(j)} \equiv f_n(\boldsymbol{\phi}^{(j)}, \boldsymbol{\theta}^{(j)})$, and the vectors of step-sizes $\boldsymbol{\rho}_{\boldsymbol{\phi}}^{(j)}$ and $\boldsymbol{\rho}_{\boldsymbol{\theta}}^{(j)}$ are given by

$$\boldsymbol{\rho}_{\boldsymbol{\phi} \setminus \boldsymbol{\theta}}^{(j)} = \eta(j+1)^{-1/2+\epsilon} \left(\tau + \sqrt{\mathbf{s}_{\boldsymbol{\phi} \setminus \boldsymbol{\theta}}^{(j)}} \right), \quad (\text{B.3})$$

and the sequence

$$\mathbf{s}_{\boldsymbol{\phi} \setminus \boldsymbol{\theta}}^{(j)} = \alpha \left(\nabla_{\boldsymbol{\phi} \setminus \boldsymbol{\theta}} f_n^{(j)} \right)^2 + (1 - \alpha) \mathbf{s}_{\boldsymbol{\phi} \setminus \boldsymbol{\theta}}^{(j-1)} \text{ for } j > 0, \quad \mathbf{s}_{\boldsymbol{\phi} \setminus \boldsymbol{\theta}}^{(0)} = \left(\nabla_{\boldsymbol{\phi} \setminus \boldsymbol{\theta}} f_n^{(0)} \right)^2, \quad (\text{B.4})$$

where $\sqrt{\cdot}$ and $(\cdot)^2$ are understood as element-wise. The parameters τ , α , and ϵ are set to 1.0, 0.1, and 1×10^{-16} , respectively, while the parameter $\eta > 0$ is chosen on a case-by-case basis.

Appendix C. Discrete adjoint method for Darcy flow

In this section we describe the computation of the gradient and Hessian of the log-likelihood, $\nabla_{\mathbf{y}} \log p(\mathcal{D}_s | \mathbf{y})$ via the discrete adjoint method [31,32]. For this purpose we introduce the function

$$h(\mathbf{u}, \mathbf{y}) = -\frac{1}{2\sigma_{us}^2} \|\mathbf{u}_s - \mathbf{H}_u \mathbf{u}\|_2^2 - \frac{1}{2\sigma_{ys}^2} \|\mathbf{y}_s - \mathbf{H}_y \mathbf{y}\|_2^2, \quad (\text{C.1})$$

so that $\nabla_{\mathbf{y}} \log p(\mathcal{D}_s | \mathbf{y}) = \nabla h(\mathbf{u}(\mathbf{y}), \mathbf{y})$ by virtue of (3) (as the constant in (3) is independent of \mathbf{y}). In the following we will employ the following notation: Let a be a scalar function, \mathbf{b} and \mathbf{c} be vector functions, and γ be a scalar variable; then, $\partial a / \partial \mathbf{b}$ denotes the row vector with entries $\partial a / \partial b_i$, $\partial \mathbf{b} / \partial \gamma$ denotes the column vector with entries $\partial b_i / \partial \gamma$, and $\partial \mathbf{b} / \partial \mathbf{c}$ be the matrix with ij th entry $\partial b_i / \partial c_j$.

Differentiation $h(\mathbf{u}, \mathbf{y})$ with respect to y_i gives

$$\frac{dh}{dy_j} = \frac{\partial h}{\partial y_j} + \frac{\partial h}{\partial \mathbf{u}} \frac{\partial \mathbf{u}}{\partial y_j}. \quad (\text{C.2})$$

Similarly, differentiating the physics constraint $\mathbf{L}(\mathbf{u}, \mathbf{y}) = 0$ with respect to \mathbf{y} gives

$$\frac{\partial \mathbf{L}}{\partial y_j} + \frac{\partial \mathbf{L}}{\partial \mathbf{u}} \frac{\partial \mathbf{u}}{\partial y_j} = 0 \quad (\text{C.3})$$

which implies $\partial \mathbf{u} / \partial y_j = -(\partial \mathbf{L} / \partial \mathbf{u})^{-1} (\partial \mathbf{L} / \partial y_j)$. Substituting this relation into (C.2) gives the following expression for the j th component of the gradient:

$$\frac{dh}{dy_j} = \frac{\partial h}{\partial y_i} + \boldsymbol{\lambda}^\top \frac{\partial \mathbf{L}}{\partial y_j} \quad (\text{C.4})$$

where the adjoint variables $\boldsymbol{\lambda}$ satisfies the adjoint equation

$$\left(\frac{\partial \mathbf{L}}{\partial \mathbf{u}} \right)^\top \boldsymbol{\lambda} + \left(\frac{\partial h}{\partial \mathbf{u}} \right)^\top = 0. \quad (\text{C.5})$$

It can be seen that computing the gradient $\nabla h(\mathbf{u}(\mathbf{y}), \mathbf{y})$ requires a single linear backward sensitivity problem, (C.5), of size $M \times M$.

For the Hessian, we differentiate (C.2) with respect to y_i , obtaining

$$\frac{d^2 h}{dy_i dy_j} = \frac{\partial h}{\partial \mathbf{u}} \frac{\partial^2 \mathbf{u}}{\partial y_i \partial y_j} + D_{i,j}^2 h, \quad (\text{C.6})$$

where $D_{i,j}^2 h$ is given by

$$\frac{\partial^2 h}{\partial y_i \partial y_j} + \frac{\partial^2 h}{\partial y_i \partial \mathbf{u}} \frac{\partial \mathbf{u}}{\partial y_j} + \frac{\partial^2 h}{\partial y_j \partial \mathbf{u}} \frac{\partial \mathbf{u}}{\partial y_i} + \frac{\partial^2 h}{\partial \mathbf{u}^2} \left(\frac{\partial \mathbf{u}}{\partial y_i} \otimes \frac{\partial \mathbf{u}}{\partial y_j} \right), \quad (\text{C.7})$$

and $\partial^2 h / \partial \mathbf{u}^2$ denotes the Hessian of h with respect to \mathbf{u} , i.e. the matrix with ij th entry $\partial^2 h / \partial u_i \partial u_j$. Similarly, differentiating (C.3) with respect to y_i gives

$$\frac{\partial \mathbf{L}}{\partial \mathbf{u}} \frac{\partial^2 \mathbf{u}}{\partial y_i \partial y_j} + D_{i,j}^2 \mathbf{L} = 0, \quad (\text{C.8})$$

where $D_{i,j}^2 \mathbf{L}$ is given element-wise in a manner similar to (C.7). (C.8) implies $\partial^2 \mathbf{u} / \partial y_i \partial y_j = -(\partial \mathbf{L} / \partial \mathbf{u})^{-1} D_{i,j}^2 \mathbf{L}$. Substituting into (C.6) gives the following expression for the ij th component of the Hessian:

$$\frac{d^2 h}{dy_i dy_j} = \boldsymbol{\lambda}^\top D_{i,j}^2 \mathbf{L} + D_{i,j}^2 h. \quad (\text{C.9})$$

Computing the Hessian therefore requires the solution of N linear forward sensitivity problems, (C.3), for each $\partial \mathbf{u} / \partial y_i$, and one backward sensitivity solution for the adjoint variables, each problem of size $M \times M$.

References

- [1] A.M. Stuart, Inverse problems: a Bayesian perspective, *Acta Numer.* 19 (2010) 451–559, <https://doi.org/10.1017/S0962492910000061>.
- [2] M. Hanke, A regularizing Levenberg-Marquardt scheme, with applications to inverse groundwater filtration problems, *Inverse Probl.* 13 (1) (1997) 79, <http://stacks.iop.org/0266-5611/13/i=1/a=007>.
- [3] D.A. Barajas-Solano, B.E. Wohlberg, V.V. Vesselinov, D.M. Tartakovsky, Linear functional minimization for inverse modeling, *Water Resour. Res.* 51 (2014) 4516–4531, <https://doi.org/10.1002/2014WR016179>.
- [4] G. Evensen, *Data Assimilation: The Ensemble Kalman Filter*, Springer-Verlag, Berlin, Heidelberg, 2006.
- [5] T. Salimans, D. Kingma, M. Welling, Markov chain Monte Carlo and variational inference: bridging the gap, in: F. Bach, D. Blei (Eds.), *Proceedings of the 32nd International Conference on Machine Learning*, in: *Proceedings of Machine Learning Research*, vol. 37, PMLR, Lille, France, 2015, pp. 1218–1226, <http://proceedings.mlr.press/v37/salimans15.html>.
- [6] J. Goodman, J. Weare, Ensemble samplers with affine invariance, *Commun. Appl. Math. Comput. Sci.* 5 (1) (2010) 65–80, <https://doi.org/10.2140/camcos.2010.5.65>.
- [7] W. Neiswanger, C. Wang, E. Xing, Asymptotically exact, embarrassingly parallel MCMC, *ArXiv e-prints*, arXiv:1311.4780.
- [8] M.D. Hoffman, A. Gelman, The no-u-turn sampler: adaptively setting path lengths in hamiltonian Monte Carlo, *J. Mach. Learn. Res.* 15 (2014) 1593–1623, <http://jmlr.org/papers/v15/hoffman14a.html>.
- [9] C.E. Rasmussen, C.K.I. Williams, *Gaussian Processes for Machine Learning*, *Adaptive Computation and Machine Learning*, The MIT Press, 2005.
- [10] M. Raissi, P. Perdikaris, G.E. Karniadakis, Numerical gaussian processes for time-dependent and nonlinear partial differential equations, *SIAM J. Sci. Comput.* 40 (1) (2018) A172–A198, <https://doi.org/10.1137/17M1120762>.
- [11] M. Raissi, P. Perdikaris, G.E. Karniadakis, Machine learning of linear differential equations using gaussian processes, *J. Comput. Phys.* 348 (2017) 683–693, <https://doi.org/10.1016/j.jcp.2017.07.050>, <http://www.sciencedirect.com/science/article/pii/S0021999117305582>.
- [12] C.M. Bishop, *Pattern Recognition and Machine Learning*, *Information Science and Statistics*, Springer-Verlag, Berlin, Heidelberg, 2006.
- [13] N.D. Lawrence, G. Sanguinetti, M. Rattray, Modelling transcriptional regulation using gaussian processes, in: B. Schölkopf, J.C. Platt, T. Hoffman (Eds.), *Advances in Neural Information Processing Systems*, vol. 19, MIT Press, 2007, pp. 785–792, <http://papers.nips.cc/paper/3119-modelling-transcriptional-regulation-using-gaussian-processes.pdf>.
- [14] R.M. Neal, G.E. Hinton, *A View of the EM Algorithm that Justifies Incremental, Sparse, and Other Variants*, Springer, Netherlands, Dordrecht, 1998, pp. 355–368.
- [15] R. Ranganath, S. Gerrish, D. Blei, Black box variational inference, in: S. Kaski, J. Corander (Eds.), *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics*, in: *Proceedings of Machine Learning Research*, vol. 33, PMLR, Reykjavik, Iceland, 2014, pp. 814–822.
- [16] M. Titsias, M. Lázaro-Gredilla, Doubly stochastic variational Bayes for non-conjugate inference, in: E.P. Xing, T. Jebara (Eds.), *Proceedings of the 31st International Conference on Machine Learning*, in: *Proceedings of Machine Learning Research*, vol. 32, PMLR, Beijing, China, 2014, pp. 1971–1979.
- [17] T.P. Minka, Expectation propagation for approximate Bayesian inference, in: *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, UAI'01, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2001, pp. 362–369, <http://dl.acm.org/citation.cfm?id=2074022.2074067>.
- [18] P. Tsilifis, I. Bilonis, I. Katsounaros, N. Zabarar, Computationally efficient variational approximations for Bayesian inverse problems, *J. Verif. Valid. Uncertain. Quantificat.* 1 (3) (2016) 031004.
- [19] B. Jin, J. Zou, Hierarchical Bayesian inference for ill-posed problems via variational method, *J. Comput. Phys.* 229 (19) (2010) 7317–7343, <https://doi.org/10.1016/j.jcp.2010.06.016>, <http://www.sciencedirect.com/science/article/pii/S0021999110003311>.
- [20] I.M. Franck, P. Koutsourelakis, Sparse variational Bayesian approximations for nonlinear inverse problems: applications in nonlinear elastography, *Comput. Methods Appl. Mech. Eng.* 299 (2016) 215–244, <https://doi.org/10.1016/j.cma.2015.10.015>, <http://www.sciencedirect.com/science/article/pii/S0045782515003345>.
- [21] N. Guha, X. Wu, Y. Efendiev, B. Jin, B.K. Mallick, A variational Bayesian approach for inverse problems with skew-t error distributions, *J. Comput. Phys.* 301 (2015) 377–393, <https://doi.org/10.1016/j.jcp.2015.07.062>, <http://www.sciencedirect.com/science/article/pii/S002199911500515X>.
- [22] K. Yang, N. Guha, Y. Efendiev, B.K. Mallick, Bayesian and variational Bayesian approaches for flows in heterogeneous random media, *J. Comput. Phys.* 345 (2017) 275–293, <https://doi.org/10.1016/j.jcp.2017.04.034>, <http://www.sciencedirect.com/science/article/pii/S0021999117303054>.
- [23] D.M. Blei, A. Kucukelbir, J.D. McAuliffe, Variational inference: a review for statisticians, *J. Am. Stat. Assoc.* 112 (518) (2017) 859–877, <https://doi.org/10.1080/01621459.2017.1285773>.
- [24] A. Kucukelbir, D. Tran, R. Ranganath, A. Gelman, D.M. Blei, Automatic differentiation variational inference, *J. Mach. Learn. Res.* 18 (14) (2017) 1–45.
- [25] D.P. Kingma, M. Welling, Auto-encoding variational Bayes, *ArXiv e-prints*, arXiv:1312.6114.
- [26] D.J. Rezende, S. Mohamed, D. Wierstra, Stochastic backpropagation and approximate inference in deep generative models, *ArXiv e-prints*, arXiv:1401.4082.
- [27] E. Challis, D. Barber, Gaussian Kullback-Leibler approximate inference, *J. Mach. Learn. Res.* 14 (2013) 2239–2286, <http://jmlr.org/papers/v14/challis13a.html>.
- [28] R.J. Williams, *Simple Statistical Gradient-Following Algorithms for Connectionist Reinforcement Learning*, Springer, US, Boston, MA, 1992, pp. 5–32.
- [29] K. Friston, J. Mattout, N. Trujillo-Barreto, J. Ashburner, W. Penny, Variational free energy and the Laplace approximation, *NeuroImage* 34 (1) (2007) 220–234, <https://doi.org/10.1016/j.neuroimage.2006.08.035>, <http://www.sciencedirect.com/science/article/pii/S1053811906008822>.
- [30] E. Ruli, N. Sartori, L. Ventura, Improved Laplace approximation for marginal likelihoods, *Electron. J. Stat.* 10 (2) (2016) 3986–4009, <https://doi.org/10.1214/16-EJS1218>.
- [31] M.B. Giles, M.C. Duta, J.-D. Müller, N.A. Pierce, Algorithm developments for discrete adjoint methods, *AIAA J.* 41 (2) (2003) 198–205.
- [32] D. Ghate, M. Giles, Efficient Hessian calculation using automatic differentiation, in: *25th AIAA Applied Aerodynamics Conference*, 2007, p. 4059.