A RANDOMIZED MAXIMUM A POSTERIORI METHOD FOR POSTERIOR SAMPLING OF HIGH DIMENSIONAL NONLINEAR BAYESIAN INVERSE PROBLEMS*

KAINAN WANG†, TAN BUI-THANH‡, AND OMAR GHATTAS§

Abstract. We present a randomized maximum a posteriori (rMAP) method for generating approximate samples of posteriors in high dimensional Bayesian inverse problems governed by large-scale forward problems. We derive the rMAP approach by (1) casting the problem of computing the MAP point as a stochastic optimization problem; (2) interchanging optimization and expectation; and (3) approximating the expectation with a Monte Carlo method. For a specific randomized data and prior mean, rMAP reduces to the randomized maximum likelihood (RML) approach. It can also be viewed as an iterative stochastic Newton method. An analysis of the convergence of the rMAP samples is carried out for both linear and nonlinear inverse problems. Each rMAP sample requires solution of a PDE-constrained optimization problem; to solve these problems, we employ a state-of-the-art trust region inexact Newton conjugate gradient method with sensitivity-based warm starts. An approximate Metropolization approach is presented to reduce the bias in rMAP samples. Various numerical methods will be presented to demonstrate the potential of the rMAP approach in posterior sampling of nonlinear Bayesian inverse problems in high dimensions.

Key words. randomized maximum a posteriori, inverse problems, uncertainty quantification, Markov chain Monte Carlo, trust region inexact Newton conjugate gradient

AMS subject classifications. 35Q62, 62F15, 35R30, 35Q93, 65C60

DOI. 10.1137/16M1060625

1. Introduction. We consider a class of inverse problems that seek to determine a distributed parameter in a partial differential equation (PDE) model, from indirect observations of outputs of the model. We adopt the framework of Bayesian inference, which accounts for uncertainties in observations, the map from parameters to observables via solution of the forward model, and prior information on the parameters. In particular, we seek a statistical description of all possible (sets of) parameters that conform to the available prior knowledge and at the same time are consistent with the observations via the parameter-to-observable map. The solution of a Bayesian inverse problem is the posterior measure, which encodes the degree of confidence on each set of parameters as the solution to the inverse problem under consideration.

Mathematically, the posterior is a surface in high dimensional parameter space. Even when the prior and noise probability distributions are Gaussian, the posterior need not be due to the nonlinearity of the parameter-to-observable map. For large-scale inverse problems, exploring non-Gaussian posteriors in high dimensions (to compute statistics such as the mean, covariance, and/or higher moments) is ex-

†Institute for Computational Engineering and Sciences, The University of Texas at Austin, Austin, TX 78712. Current address: Sanchez Oil & Gas Corp., Houston, TX 77002 (kennan.wong@gmail.com).

‡Department of Aerospace Engineering and Engineering Mechanics, and Institute for Computational Engineering and Sciences, The University of Texas at Austin, Austin, TX 78712 (tanbui@ices.utexas.edu).

§Institute for Computational Engineering and Sciences, Jackson School of Geosciences, and Department of Mechanical Engineering, The University of Texas at Austin, Austin, TX 78712 (omar@ices.utexas.edu).

tremely challenging. The usual method of choice for computing statistics is Markov chain Monte Carlo (MCMC) [20, 32, 37, 43, 53, 54, 55], which judiciously samples the posterior distribution, so that sample statistics can be used to approximate the exact distributions. The problem, however, is that standard MCMC methods often require millions of samples for convergence; since each sample requires an evaluation of the parameter-to-observable map, this entails millions of expensive forward PDE simulations—a prohibitive proposition. On one hand, with the rapid development of parallel computing, parallel MCMC methods [6, 16, 58, 61, 62] are studied to accelerate the computation. While parallelization allows MCMC algorithms to produce more samples in less time with multiple processors, such accelerations typically do not improve the mixing and convergence of MCMC algorithms. More sophisticated MCMC methods that exploit the gradient and higher derivatives of the log posterior (and hence the parameter-to-observable map) [2, 13, 15, 21, 22, 25, 27, 42, 44, 50] can, on the other hand, improve the mixing, acceptance rate, and convergence of MCMC. Several of these methods exploit local curvature in parameter space as captured by the Hessian operator of the negative logarithm of the posterior. This requires manipulating the Hessian of the data misfit functional (i.e., the negative log likelihood). The stochastic Newton method (SN) [13, 42, 50] makes these Hessian manipulations tractable by invoking a low rank approximation, motivated by the theoretically established or experimentally observed compactness of this operator for many large-scale ill-posed inverse problems.

However, despite its successful application to million-parameter problems governed by expensive-to-solve PDEs [9, 36], two barriers exist that prevent further scaling of SN to challenging problems. First, even computing low rank Hessian information for every sample in parameter space can be prohibitive. Second, when the curvature of the negative log posterior changes rapidly, SN's local Gaussian approximation may not provide a good enough model for the posterior, and hence the MCMC proposal may not be effective. This may result in low acceptance rates and excessive numbers of forward PDE solves.

In this paper, we consider an optimization boosted sampling framework, the randomized maximum a posteriori (rMAP) method that is inspired by the randomized maximum likelihood (RML) [39, 48] and the randomize-then-optimize (RTO) approaches [1]. Through computing each sample by PDE-constrained optimization [3, 4, 24, 33], it can explore the parameter space more efficiently. It can also be viewed as a nonlinear SN method that executes multiple Newton iterations in every MCMC step to generate a better proposal and to allow an improved acceptance rate. On the other hand, solving optimization problems is expensive, and hence we discuss several improvements and extensions to make the rMAP method more applicable towards solving real problems.

We present our discussions in the following order. Section 2 introduces a statistical inversion setting based on the Bayesian framework in infinite dimensions. The core of the paper is section 3. In this section, we first convert the maximum a posteriori (MAP) problem into a stochastic programming problem, which is then solved using sample average approximation. This rMAP method rediscovers the RML method as a special case. Results for convergence of the rMAP ensemble using stochastic programming theory are presented, and the extension of the rMAP to infinite dimensional problems is discussed at length. We also show that rMAP is a generalization of SN—for linear inverse problems, they become identical. It is worth noting that for nonlinear inverse problems, rMAP samples are not the actual but rather the approximate samples of the underlying posterior distribution. Therefore, in section 4 we also

discuss an approximate Metropolization technique to reduce the bias between sample approximation and the true posterior distribution. We discuss in section 5 a finite element discretization of the infinite dimensional Bayes inverse problem. We also describe how to solve the optimization problem efficiently at each sampling step. In particular, we present a sensitivity approach to obtain "good" initial guesses for further accelerating the optimization procedure. In section 6, various numerical results showing the efficiency of proposed strategies compared to state-of-the-art alternatives are presented for 1D analytical problems as well as 2D inverse problems governed by the Helmholtz equation. Finally, we conclude the paper in section 7.

2. Infinite dimensional Bayesian inverse problem setting. We consider the following generic forward model:

$$\mathcal{B}(u, w) = 0 \quad \text{in } \Omega,$$

which, for example, can be PDEs modeling the physical problem under consideration. The forward problem involves solving for the forward state w given a modeling of the distributed parameter u. In the inverse problem, the task is to reconstruct u given some available observations of w on parts of the domain Ω. One widely accepted model for the relationship between model parameters and observations is the additive noise model,

$$\mathbf{d} = \mathcal{G}(u) + \boldsymbol{\eta}, \tag{1}$$

with $\mathbf{d} = [d_1, \ldots, d_K]^T$ denoting all observed data, $\mathcal{G} := [w(\mathbf{x}_1), \ldots, w(\mathbf{x}_K)]^T$ denoting the parameter-to-observable (or forward) map, i.e., the map from the distributed parameter u to the observables $w(\mathbf{x}_i)$ at locations $\{\mathbf{x}_j\}, j = 1, 2, \ldots, K$, and noise being represented by $\boldsymbol{\eta}$, a random vector normally distributed by $\mathcal{N}(\mathbf{0}, \mathbf{L})$ with bounded covariance matrix $\mathbf{L}$. For simplicity, we take $\mathbf{L} = \sigma^2 \mathbf{I}$, where $\mathbf{I}$ is the identity matrix of appropriate dimension. For notational convenience, throughout the paper we use boldface italic letters for vectors and matrices and Roman letters for infinite dimensional counterparts. For example, u denotes a function in $L^2(\Omega)$, while $\boldsymbol{u}$ represents its discrete counterpart.

The inverse problem can be formulated as choosing model parameters that minimize the discrepancy between model prediction and observations:

$$\min_u \Phi(u, \mathbf{d}) := \frac{1}{2} |\mathbf{d} - \mathcal{G}(u)|_{\mathbf{L}}^2 \tag{2}$$

subject to the forward problem

$$\mathcal{B}(u, w) = 0, \tag{3}$$

where $|\cdot|_{\mathbf{L}} := \left| \mathbf{L}^{-\frac{1}{2}} \cdot \right|$ denotes the weighted Euclidean norm induced by the inner product in $\mathbb{R}^K$. This optimization problem, however, is ill-posed. An intuitive reason is that the dimension of vector of observations $\mathbf{d}$ is often much smaller than that of the parameter u (typically infinite before discretization), and hence $\mathbf{d}$ provides limited information about the distributed parameter u. As a result, the null space of the Jacobian of the parameter-to-observable map $\mathbf{F}$ is nonempty. In particular, for a class of inverse problems, we have shown that the Gauss–Newton approximation of the Hessian (which is the product of the Jacobian transpose and the Jacobian, and is also equal to the full Hessian of the misfit Φ evaluated at the optimal parameter in the

zero residual case, i.e., when the data is noise free) is a compact operator [10, 11, 12], and hence its range space is effectively finite dimensional.

In this paper, we choose to tackle the ill-posedness using a Bayesian framework [17, 26, 37, 40, 41, 51, 59]. We seek a statistical description of all possible parameter fields u that conform to some prior knowledge and at the same time are consistent with the observations. The Bayesian approach accomplishes this through a statistical inference framework that incorporates uncertainties in the observations, the forward map $\mathcal{G}$, and the prior information. To begin, we postulate the prior as a Gaussian measure $\mu := \mathcal{N}(u_0, \mathcal{C})$ with mean function u_0 and covariance operator $\mathcal{C}$ on u in $L^2(\Omega)$, where

$$\mathcal{C} := \alpha^{-1}(I - \Delta)^{-s} =: \alpha^{-1}\mathcal{A}^{-s}, \quad \alpha > 0,$$

with the domain of definition of $\mathcal{A}$ defined as

$$D(\mathcal{A}) := \left\{ u \in H^2(\Omega) : \frac{\partial u}{\partial \mathbf{n}} = 0 \text{ on } \partial\Omega \right\}.$$

Here, $H^2(\Omega)$ is the usual Sobolev space. Assume that the mean function u_0 resides in the Cameron–Martin space of μ; then one can show (see, e.g., [59]) that the prior measure μ is well-defined when $s > d/2$ (d is the spatial dimension), and in this case, any realization from the prior distribution μ almost surely resides in the Hölder space $X := C^{0,\beta}(\Omega)$ with $0 < \beta < s/2$. That is, $\mu(X) = 1$, and the Bayesian posterior measure ν satisfies the Radon–Nikodym derivative

$$(4) \qquad \frac{\partial \nu}{\partial \mu}(u|\mathbf{d}) \propto \exp\left(-\Phi(u, \mathbf{d})\right)$$

if $\mathcal{G}$ is a continuous map from X to $\mathbb{R}^K$.

The MAP point (see, e.g., [23, 59] for the definition of the MAP point in infinite dimensional settings) is given by

$$(5) \qquad u^{\mathrm{MAP}} := \arg\min_u \mathcal{J}(u; u_0, \mathbf{d}) := \frac{1}{2}|\mathbf{d} - \mathcal{G}(u)|_{\mathbf{L}}^2 + \frac{1}{2}\|u - u_0\|_{\mathcal{C}}^2,$$

where $\|\cdot\|_{\mathcal{C}} := \left\|\mathcal{C}^{-\frac{1}{2}}\cdot\right\|$ denotes the weighted $L^2(\Omega)$ norm induced by the $L^2(\Omega)$ inner product $\langle\cdot,\cdot\rangle$. We shall also use $\langle\cdot,\cdot\rangle$ to denote the duality pairing on $L^2(\Omega)$.

It should be pointed out that the last term in (5) can be considered as a prior-inspired regularization; the MAP point is thus a solution to the corresponding deterministic inverse problem. However, the Bayesian approach goes well beyond the deterministic solution to provide a complete statistical description of the inverse solution: the posterior encodes the degree of confidence (probability) in the estimate of all possible parameter fields.

In addition to the MAP point, we also wish to interrogate the posterior distribution for statistics such as conditional mean and interval estimates. This requires sampling of the distribution, where empirical statistics from produced samples can effectively approximate those of the posterior. Popular sampling methods usually suffer from problems such as the curse of dimensionality. On the other hand, successful computational methods for MAP estimation have been studied extensively. These facts motivate us to explore sampling methods that are facilitated by MAP estimates, which we discuss in detail below.

3. A randomized maximum a posteriori approach. In this section we present an approach, which we call the randomized maximum a posteriori (rMAP) method, to compute approximate samples for the posterior distribution. The idea is to first randomize the cost function to cast the MAP statement (5) into a stochastic programming problem, which is then solved using the Monte Carlo method (also known as the sample average approximation [57]). The resulting rMAP method resembles the RML developed in [39, 48] as a special case. We therefore rediscover the RML method from a completely new, i.e., stochastic programing, viewpoint. It is this view that allows us to provide new theoretical results on the RML approach for nonlinear inverse problems that were previously not available. Indeed, the fact that RML samples are exact samples of the posterior for linear inverse problems currently seems to be the only available result on the RML method [1, 39, 48]. We shall also show that the rMAP method (which from now on will be used interchangeably with the RML method) can be considered as a means to incorporate uncertainty into the solution of deterministic inverse approaches.

To begin, let us consider finite dimensional parameter space[1] for simplicity of the exposition, i.e., $\boldsymbol{u}, \boldsymbol{u}_0 \in \mathbb{R}^N$. The posterior measure ν in this case has the density π_{post} with respect to the Lebesgue measure:

$$\pi_{\text{post}} \propto \pi_{\text{like}} \times \pi_{\text{prior}},$$

where the likelihood is given by $\pi_{\text{like}} \propto \exp\left(-\Phi\left(u, \mathbf{d}\right)\right) = \exp\left(-\frac{1}{2}\left|\mathbf{d} - \mathcal{G}\left(u\right)\right|_{\mathbf{L}}^2\right)$ and the prior by $\pi_{\text{prior}} \propto \exp\left(-\frac{1}{2}\left|\boldsymbol{u} - \boldsymbol{u}_0\right|_{\mathcal{C}}^2\right)$. The MAP problem (5) becomes

(6) $$\boldsymbol{u}^{\text{MAP}} := \arg\min_{\boldsymbol{u}} \mathcal{J}\left(\boldsymbol{u}; \boldsymbol{u}_0, \mathbf{d}\right) := \frac{1}{2}\left|\mathbf{d} - \mathcal{G}\left(\boldsymbol{u}\right)\right|_{\mathbf{L}}^2 + \frac{1}{2}\left|\boldsymbol{u} - \boldsymbol{u}_0\right|_{\mathcal{C}}^2,$$

where $\mathcal{C} \in \mathbb{R}^{N \times N}$ is the covariance matrix in this case. Throughout this paper, we denote by $\mathbb{E}$ the expectation. We now randomize the cost function and hence the MAP problem (6).

LEMMA 3.1. *Let $\boldsymbol{\theta} \in \mathbb{R}^K$ and $\boldsymbol{\varepsilon} \in \mathbb{R}^N$ be two independent random vectors distributed by $\pi_{\boldsymbol{\theta}}$ and $\pi_{\boldsymbol{\varepsilon}}$ with zero mean, i.e., $\mathbb{E}_{\boldsymbol{\theta}}\left[\boldsymbol{\theta}\right] = \mathbf{0}$ and $\mathbb{E}_{\boldsymbol{\varepsilon}}\left[\boldsymbol{\varepsilon}\right] = \mathbf{0}$. The following result holds:*

$$\mathcal{J}\left(\boldsymbol{u}; \boldsymbol{u}_0, \mathbf{d}\right) = \mathbb{E}_{\boldsymbol{\theta} \times \boldsymbol{\varepsilon}}\left[\mathcal{J}^r\left(\boldsymbol{u}; \boldsymbol{u}_0, \mathbf{d}, \boldsymbol{\theta}, \boldsymbol{\varepsilon}\right)\right] - \mathbb{E}_{\boldsymbol{\theta}}\left[\boldsymbol{\theta}^T \boldsymbol{\theta}\right] - \mathbb{E}_{\boldsymbol{\varepsilon}}\left[\boldsymbol{\varepsilon}^T \boldsymbol{\varepsilon}\right],$$

where

$$\mathcal{J}^r\left(\boldsymbol{u}; \boldsymbol{u}_0, \mathbf{d}, \boldsymbol{\theta}, \boldsymbol{\varepsilon}\right) = \frac{1}{2}\left|\mathbf{d} + \boldsymbol{\theta} - \mathcal{G}\left(\boldsymbol{u}\right)\right|_{\mathbf{L}}^2 + \frac{1}{2}\left|\boldsymbol{u} - \boldsymbol{u}_0 - \boldsymbol{\varepsilon}\right|_{\mathcal{C}}^2,$$

with $\mathbb{E}_{\boldsymbol{\theta} \times \boldsymbol{\varepsilon}}$ denoting the expectation with respect to the product measure $\pi_{\boldsymbol{\theta}} \times \pi_{\boldsymbol{\varepsilon}}$ induced by $\left(\boldsymbol{\theta}, \boldsymbol{\varepsilon}\right)$. Consequently,

(7) $$\boldsymbol{u}^{MAP} := \arg\min_{\boldsymbol{u}} \mathcal{J}\left(\boldsymbol{u}; \boldsymbol{u}_0, \mathbf{d}\right) = \arg\min_{\boldsymbol{u}} \mathbb{E}_{\boldsymbol{\theta} \times \boldsymbol{\varepsilon}}\left[\mathcal{J}^r\left(\boldsymbol{u}; \boldsymbol{u}_0, \mathbf{d}, \boldsymbol{\theta}, \boldsymbol{\varepsilon}\right)\right].$$

Proof. Since $\boldsymbol{\theta}$ and $\boldsymbol{\varepsilon}$ are independent, we have

$$\mathbb{E}_{\boldsymbol{\theta} \times \boldsymbol{\varepsilon}}\left[\mathcal{J}^r\left(\boldsymbol{u}; \boldsymbol{u}_0, \mathbf{d}, \boldsymbol{\theta}, \boldsymbol{\varepsilon}\right)\right] = \frac{1}{2}\mathbb{E}_{\boldsymbol{\theta}}\left[\left|\mathbf{d} + \boldsymbol{\theta} - \mathcal{G}\left(\boldsymbol{u}\right)\right|_{\mathbf{L}}^2\right] + \frac{1}{2}\mathbb{E}_{\boldsymbol{\varepsilon}}\left[\left|\boldsymbol{u} - \boldsymbol{u}_0 - \boldsymbol{\varepsilon}\right|_{\mathcal{C}}^2\right]$$

$$= \mathcal{J}\left(\boldsymbol{u}; \boldsymbol{u}_0, \mathbf{d}\right) + \mathbb{E}_{\boldsymbol{\theta}}\left[\boldsymbol{\theta}^T \mathbf{L}^{-1}\left(\mathbf{d} - \mathcal{G}\left(\boldsymbol{u}\right)\right)\right] - \mathbb{E}_{\boldsymbol{\varepsilon}}\left[\boldsymbol{\varepsilon}^T \mathcal{C}^{-1}\left(\boldsymbol{u} - \boldsymbol{u}_0\right)\right] + \mathbb{E}_{\boldsymbol{\theta}}\left[\boldsymbol{\theta}^T \boldsymbol{\theta}\right] + \mathbb{E}_{\boldsymbol{\varepsilon}}\left[\boldsymbol{\varepsilon}^T \boldsymbol{\varepsilon}\right],$$

---

[1]Finite dimensionality could result from a discretization of distributed parameters (see, e.g., [8] for a constructive finite element discretization).

which proves the first assertion since $\mathbb{E}_{\boldsymbol{\theta}}\left[\boldsymbol{\theta}\right] = \mathbf{0}$ and $\mathbb{E}_{\boldsymbol{\varepsilon}}\left[\boldsymbol{\varepsilon}\right] = \mathbf{0}$. The second assertion is obvious since $\mathbb{E}_{\boldsymbol{\theta}}\left[\boldsymbol{\theta}^T\boldsymbol{\theta}\right]$ and $\mathbb{E}_{\boldsymbol{\varepsilon}}\left[\boldsymbol{\varepsilon}^T\boldsymbol{\varepsilon}\right]$ are constants independent of $\boldsymbol{u}$. $\qquad\square$

Lemma 3.1, particularly identity (7), shows that the MAP point can be considered as the solution of the following stochastic programming problem:

$$(8) \qquad \min_{\boldsymbol{u}} \mathbb{E}_{\boldsymbol{\theta}\times\boldsymbol{\varepsilon}}\left[\mathcal{J}^r\left(\boldsymbol{u};\boldsymbol{u}_0,\mathbf{d},\boldsymbol{\theta},\boldsymbol{\varepsilon}\right)\right] = \mathbb{E}_{\boldsymbol{\theta}\times\boldsymbol{\varepsilon}}\left[\min_{\boldsymbol{u}} \mathcal{J}^r\left(\boldsymbol{u};\boldsymbol{u}_0,\mathbf{d},\boldsymbol{\theta},\boldsymbol{\varepsilon}\right)\right],$$

where we have interchanged the order of minimization and expectation.[2] At this moment, (8) holds for finite dimensional cases, and whether it is also true for infinite dimensional settings is unknown. Our next step is to approximate the expectation on the right-hand side of (8) using the Monte Carlo approach (also known as the sample average approximation [57]). In particular, with n independent and identically distributed (i.i.d.) samples $(\boldsymbol{\theta}_j, \boldsymbol{\varepsilon}_j)$ from the product measure $\pi_{\boldsymbol{\theta}} \times \pi_{\boldsymbol{\varepsilon}}$, we have

$$(9) \qquad \min_{\boldsymbol{u}} \mathbb{E}_{\boldsymbol{\theta}\times\boldsymbol{\varepsilon}}\left[\mathcal{J}^r\left(\boldsymbol{u};\boldsymbol{u}_0,\mathbf{d},\boldsymbol{\theta},\boldsymbol{\varepsilon}\right)\right] \approx \frac{1}{n}\sum_{j=1}^{n}\min_{\boldsymbol{u}} \mathcal{J}^r\left(\boldsymbol{u};\boldsymbol{u}_0,\mathbf{d},\boldsymbol{\theta}_j,\boldsymbol{\varepsilon}_j\right).$$

Let us define

$$(10) \quad \boldsymbol{u}_j := \arg\min_{\boldsymbol{u}} \mathcal{J}^r\left(\boldsymbol{u};\boldsymbol{u}_0,\mathbf{d},\boldsymbol{\theta}_j,\boldsymbol{\varepsilon}_j\right) = \frac{1}{2}\left|\mathbf{d}+\boldsymbol{\theta}_j-\mathcal{G}\left(\boldsymbol{u}\right)\right|^2_{\mathbf{L}} + \frac{1}{2}\left|\boldsymbol{u}-\boldsymbol{u}_0-\boldsymbol{\varepsilon}_j\right|^2_{\mathcal{C}},$$

and we are in a position to define the rMAP method in Algorithm 1. As can be seen, the observation vector $\mathbf{d}$ and the prior mean $\boldsymbol{u}_0$ are randomized in the first two steps, which are then followed by solving a randomized MAP problem in the third step. Finally, we take each perturbed MAP point $\boldsymbol{u}_j$ as an approximate sample of the posterior π_{post}.

Algorithm 1 The rMAP algorithm.

Input: Choose the sample size n
1: **for** $j = 1,\ldots,n$ **do**
2:     Draw $\boldsymbol{\varepsilon}_j \sim \pi_{\boldsymbol{\varepsilon}}$
3:     Draw $\boldsymbol{\theta}_j \sim \pi_{\boldsymbol{\theta}}$
4:     Compute rMAP sample $\boldsymbol{u}_j$ via (10)
5: **end for**

Throughout the paper, we choose the product measure to be $\pi_{\boldsymbol{\theta}} \times \pi_{\boldsymbol{\varepsilon}} = \mathcal{N}\left(\mathbf{0},\mathbf{L}\right) \times \mathcal{N}\left(\mathbf{0},\mathcal{C}\right)$, and in this case the rMAP approach becomes the RML method [1, 38, 49]. That is, the RML method is a special case of our framework. In other words, by first casting the MAP computation into a stochastic programming problem and then solving it using the sample average approximation, we have arrived at a constructive derivation of the RML method. One can show that the RML samples are exactly those of the posterior when the forward map $\mathcal{G}\left(\boldsymbol{u}\right)$ is linear [1, 38, 49]. This seems to be the only theoretical result currently available for RML. Our stochastic programming viewpoint shows that the RML method is nothing more than a sample average approximation to the stochastic optimization problem (8), whose solution is the MAP point. However, the sample average does not converge to the MAP point, as we now show. Let us define

$$(11) \qquad\qquad S\left(\boldsymbol{u}_0,\mathbf{d},\boldsymbol{\theta},\boldsymbol{\varepsilon}\right) := \arg\min_{\boldsymbol{u}} \mathcal{J}^r\left(\boldsymbol{u};\boldsymbol{u}_0,\mathbf{d},\boldsymbol{\theta},\boldsymbol{\varepsilon}\right);$$

---

[2]The conditions under which the interchange is valid can be consulted in [56, Theorem 14.60].

that is, $S(\boldsymbol{u}_0, \mathbf{d}, \boldsymbol{\theta}, \boldsymbol{\varepsilon})$ is the "optimizer operator." Clearly, this operator maps a pair $(\boldsymbol{\theta}_j, \boldsymbol{\varepsilon}_j)$ to an RML sample

$$\boldsymbol{u}_j := \arg\min_{\boldsymbol{u}} \mathcal{J}^r\left(\boldsymbol{u}; \boldsymbol{u}_0, \mathbf{d}, \boldsymbol{\theta}_j, \boldsymbol{\varepsilon}_j\right) = S\left(\boldsymbol{u}_0, \mathbf{d}, \boldsymbol{\theta}_j, \boldsymbol{\varepsilon}_j\right).$$

PROPOSITION 3.2. *Assume $S(\boldsymbol{u}_0, \mathbf{d}, \boldsymbol{\theta}, \boldsymbol{\varepsilon})$ is measurable with respect to the product measure $\pi_{\boldsymbol{\theta}} \times \pi_{\boldsymbol{\varepsilon}}$; then*

$$\frac{1}{n}\sum_{j=1}^{n}\boldsymbol{u}_j \overset{a.s.}{\to} \mathbb{E}_{\boldsymbol{\theta}\times\boldsymbol{\varepsilon}}\left[S\left(\boldsymbol{u}_0, \mathbf{d}, \boldsymbol{\theta}, \boldsymbol{\varepsilon}\right)\right].$$

Proof. The result is a simple consequence of the law of large numbers. □

Note that setting $\boldsymbol{\theta} = \mathbf{0}$ and $\boldsymbol{\varepsilon} = \mathbf{0}$ in (11) reveals that $S(\boldsymbol{u}_0, \mathbf{d}, \mathbf{0}, \mathbf{0})$ is the solution of a deterministic inverse problem with prior-inspired regularization. If we view $\boldsymbol{\theta}$ and $\boldsymbol{\varepsilon}$ as the uncertainty in data $\mathbf{d}$ and the baseline (the prior mean) parameter $\boldsymbol{u}_0$, respectively, the rMAP method can be considered as a Monte Carlo approach to propagate the uncertainty from $\mathbf{d}$ and $\boldsymbol{u}_0$ to that of the inverse solution.

COROLLARY 3.3. *When the forward map $\mathcal{G}(\boldsymbol{u})$ is linear, the following holds:*

$$\frac{1}{n}\sum_{j=1}^{n}\boldsymbol{u}_j \overset{a.s.}{\to} \boldsymbol{u}^{MAP},$$

and each rMAP sample $\boldsymbol{u}_j$ is in fact the actual sample of the posterior.

We now extend the rMAP method to posterior distribution in function spaces. In this case, $\mathcal{C}$ is a covariance operator from $L^2(\Omega)$ to $L^2(\Omega)$, $\mathbb{R}^K \ni \boldsymbol{\theta} \sim \mathcal{N}(\mathbf{0}, \mathbf{L})$, and $L^2(\Omega) \ni \varepsilon \sim \mathcal{N}(0, \mathcal{C})$. For notational convenience, let us define

$$\hat{\mathbf{d}} := \mathbf{d} + \boldsymbol{\theta} \quad \text{and} \quad \hat{u} := u_0 + \varepsilon.$$

The randomized MAP problem is now defined as

(12)
$$\begin{aligned}
\hat{u}^{\mathrm{MAP}} &:= \arg\min_{u} \mathcal{J}^r\left(u; \hat{u}, \hat{\mathbf{d}}\right) \\
&:= \arg\min_{u} \frac{1}{2}\left|\hat{\mathbf{d}} - \mathcal{G}(u)\right|_{\mathbf{L}}^2 + \frac{1}{2}\|u\|_{\mathcal{C}}^2 + \langle u, \hat{u}\rangle_{\mathcal{C}}.
\end{aligned}$$

Note that the last two terms in (12) are not the same as the last term in (5). The reason is that the Cameron–Martin space of μ has zero measure [30, 52], and hence $\hat{u}$ almost surely does not belong to this space. As a result, the term $\frac{1}{2}\|\hat{u}\|_{\mathcal{C}}^2$ is almost surely infinite, which should be removed as done in (12). On the other hand, a solution to (5) or (12) is necessary in the Cameron–Martin space since, otherwise, the term $\|u\|_{\mathcal{C}}^2$ is infinite. The existence of such a solution has been shown in [59], and hence (12) is meaningful. Furthermore, the last term $\langle u, \hat{u}\rangle_{\mathcal{C}}$ should be understood in the limit sense (see [52, 59]) since $\hat{u} \in L^2(\Omega)$ and the Cameron–Martin space is dense in $L^2(\Omega)$. Now we are in a position to analyze the rMAP samples in function spaces.

LEMMA 3.4. *If the forward map $\mathcal{G}(u)$ is linear in u, then $\hat{u}^{MAP}$ is distributed by the posterior measure (4).*

Proof. To begin, assume $\mathcal{G}(u) = \mathbf{B}u$. Taking the first variation of $\mathcal{J}^r\left(u; \hat{u}, \hat{\mathbf{d}}\right)$ with respect to u in the direction $\tilde{u}$ (in the Cameron–Martin space) gives

$$\left\langle \nabla \mathcal{J}\left(u; \hat{u}, \hat{\mathbf{d}}\right), \tilde{u}\right\rangle = \left\langle \mathcal{L}u - \mathbf{B}^*\mathbf{L}^{-1}\hat{\mathbf{d}} - \mathcal{C}^{-1}\hat{u}, \tilde{u}\right\rangle,$$

where $\mathbf{B}^* : \mathbb{R}^K \to L^2(\Omega)$ is the adjoint of $\mathbf{B}$, and we have defined

$$\mathcal{L} := \mathbf{B}^*\mathbf{L}^{-1}\mathbf{B} + \mathcal{C}^{-1}.$$

Again, the term $\left\langle \mathcal{C}^{-1}\hat{u}, \tilde{u}\right\rangle$ should be understood in the limit sense (see [52, 59]). By definition, $\hat{u}^{\mathrm{MAP}}$ is a solution of $\left\langle \nabla \mathcal{J}\left(u; \hat{u}, \hat{\mathbf{d}}\right), \tilde{u}\right\rangle = 0 \ \forall \tilde{u}$. Consequently, we have

$$(13) \qquad \hat{u}^{\mathrm{MAP}} = \mathcal{L}^{-1}\left(\mathbf{B}^*\mathbf{L}^{-1}\hat{\mathbf{d}} + \mathcal{C}^{-1}\hat{u}\right).$$

Since both $\hat{u}$ and $\hat{\mathbf{d}}$ are Gaussian, $\hat{u}^{\mathrm{MAP}}$ is also a Gaussian random function. Assume that $\hat{\mathbf{d}}$ and $\hat{u}$ are independent; after some simple algebra and manipulation the mean of $\hat{u}^{\mathrm{MAP}}$ can be written as

$$(14) \qquad \mathbb{E}\left[\hat{u}^{\mathrm{MAP}}\right] = \mathcal{L}^{-1}\left(\mathbf{B}^*\mathbf{L}^{-1}\mathbf{d} + \mathcal{C}^{-1}u_0\right),$$

which is exactly the MAP point in (5). Furthermore, the covariance operator of $\hat{u}^{\mathrm{MAP}}$ reads

$$(15) \qquad \mathbb{E}\left[\left(\hat{u}^{\mathrm{MAP}} - u^{\mathrm{MAP}}\right) \otimes \left(\hat{u}^{\mathrm{MAP}} - u^{\mathrm{MAP}}\right)\right] = \mathcal{L}^{-1}.$$

On the other hand, using conditional Gaussian measures [59], one can show that the posterior measure ν is a Gaussian with mean function

$$(16) \qquad \bar{u} = u_0 + \mathcal{C}\mathbf{B}^*\left(\mathbf{L} + \mathbf{B}\mathcal{C}\mathbf{B}^*\right)^{-1}\left(\mathbf{d} - \mathbf{B}u_0\right)$$

and covariance operator

$$(17) \qquad \mathcal{C}_{\mathrm{post}} = \mathcal{C} - \mathcal{C}\mathbf{B}^*\left(\mathbf{L} + \mathbf{B}\mathcal{C}\mathbf{B}^*\right)\mathbf{B}\mathcal{C}.$$

The fact that (14) and (15) are identical to (16) and (17), respectively, follows directly from the "matrix" inversion lemma [28]. $\qquad\square$

3.1. rMAP as the SN method for linear inverse problems. We begin by extending the finite dimensional SN method in [42] to infinite dimensions. To that end, we define the SN proposal in function space as

$$(18) \qquad v_{SN} = u - \left[\nabla^2 \mathcal{J}(u; u_0, \mathbf{d})\right]^{-1} \nabla \mathcal{J}(u; u_0, \mathbf{d}) + \mathcal{N}\left(0, \left[\nabla^2 \mathcal{J}(u; u_0, \mathbf{d})\right]^{-1}\right),$$

where, from the definition of $\mathcal{J}$ in (5), we define

$$(19\mathrm{a}) \qquad \nabla \mathcal{J}(u; u_0, \mathbf{d}) = \nabla \mathcal{G}^*(u)\mathbf{L}^{-1}\left[\mathcal{G}(u) - \mathbf{d}\right] + \mathcal{C}^{-1}(u - u_0),$$

$$(19\mathrm{b}) \qquad \nabla^2 \mathcal{J}(u; u_0, \mathbf{d}) = \nabla\left[\nabla \mathcal{G}^*(u)\right]\mathbf{L}^{-1}\left[\mathcal{G}(u) - \mathbf{d}\right] + \nabla \mathcal{G}^*(u)\mathbf{L}^{-1}\nabla \mathcal{G}(u) + \mathcal{C}^{-1}.$$

Clearly, the infinite dimensional SN proposal reduces to that proposed in [42] for finite dimensional problems. Here comes the relation between rMAP and SN methods.

LEMMA 3.5. *The rMAP approach is identical to the SN method for linear inverse problems.*

Proof. Since the forward map is linear, i.e., $\mathcal{G}(u) = \mathbf{B}u$, the posterior is a Gaussian measure as discussed above. A simple manipulation gives

$$\nabla \mathcal{J}(u; u_0, \mathbf{d}) = \mathcal{L}u - \mathbf{B}^* \mathbf{L}^{-1} \mathbf{d} - \mathcal{C}^{-1} u_0 \quad \text{and} \quad \nabla^2 \mathcal{J}(u; u_0, \mathbf{d}) = \mathbf{B}^* \mathbf{L}^{-1} \mathbf{B} + \mathcal{C}^{-1}.$$

Consequently,

$$(20) \qquad v_{SN} = u^{\mathrm{MAP}} + \mathcal{N}\left(0, \mathcal{L}^{-1}\right),$$

where $u^{\mathrm{MAP}} = \mathcal{L}^{-1}\left(\mathbf{B}^* \mathbf{L}^{-1} \mathbf{d} + \mathcal{C}^{-1} u_0\right)$ as in the proof of Lemma 3.4. Due to the linearity of $\mathcal{G}$, we need only use one Newton iteration to obtain $\hat{u}^{\mathrm{MAP}}$, and it is exactly given by (13).

In order to show the equivalence between rMAP and SN, we need to prove that v_{SN} and $\hat{u}^{\mathrm{MAP}}$ come from the same distribution. But this is obvious by inspection: the mean function and the covariance function of v_{SN} are exactly given by (14) and (15), i.e., the mean and the covariance of $\hat{u}^{\mathrm{MAP}}$. □

3.2. rMAP as an iterative SN method for nonlinear inverse problems.
For a nonlinear forward map, rMAP is no longer the same as the SN method. Instead, as we now show, it can be considered as an iterative SN method (iSN) when the full Hessian is approximated by the Gauss–Newton Hessian. To begin, we note that the rMAP sample $\hat{u}^{\mathrm{MAP}}$ is a solution of the equation

$$(21) \qquad \nabla \mathcal{J}\left(u; \hat{u}, \hat{\mathbf{d}}\right) = 0,$$

which can be solved using the Newton method. Each Newton iteration reads

$$u^{k+1} = u^k - \left[\nabla^2 \mathcal{J}\left(u^k; \hat{u}, \hat{\mathbf{d}}\right)\right]^{-1} \nabla \mathcal{J}\left(u^k; \hat{u}, \hat{\mathbf{d}}\right), \quad k = 1, \dots.$$

Now, the Gauss–Newton part of the full Hessian (19b) is given by

$$\nabla^2 \mathcal{J}_g(u) = \nabla \mathcal{G}^*(u) \mathbf{L}^{-1} \nabla \mathcal{G}(u) + \mathcal{C}^{-1},$$

which is independent of u_0 and $\mathbf{d}$. The SN proposal in this case can be written as

$$v_{SN} = u - \left[\nabla^2 \mathcal{J}_g(u)\right]^{-1} \nabla \mathcal{J}(u; u_0, \mathbf{d}) + \mathcal{N}\left(0, \left[\nabla^2 \mathcal{J}_g(u)\right]^{-1}\right),$$

with u denoting the current state of the SN Markov chain under consideration. On the other hand, the rMAP method with Gauss–Newton Hessian can be written as

$$u^{k+1} = u^k - \left[\nabla^2 \mathcal{J}_g\left(u^k\right)\right]^{-1} \nabla \mathcal{J}\left(u^k; \hat{u}, \hat{\mathbf{d}}\right), \quad k = 1, \dots.$$

In particular, as we define the initial guess to be $u^1 = u$, we have

$$(22) \qquad u^2 = u - \left[\nabla^2 \mathcal{J}_g(u)\right]^{-1} \nabla \mathcal{J}\left(u; \hat{u}, \hat{\mathbf{d}}\right).$$

Now, by definition of $\hat{u}$ and $\hat{\mathbf{d}}$, there exist $\tilde{u}$ and $\tilde{\mathbf{d}}$ such that

$$\hat{u} = u_0 + \tilde{u} \quad \text{and} \quad \hat{\mathbf{d}} = \mathbf{d} + \tilde{\mathbf{d}},$$

where

$$\tilde{u} \sim \mathcal{N}(0, \mathcal{C}) \quad \text{and} \quad \tilde{\mathbf{d}} \sim \mathcal{N}(\mathbf{0}, \mathbf{L}).$$

Consequently, by linearity of $\nabla \mathcal{J}(u; \cdot, \cdot)$ with respect to the last two arguments (see (19a)) we have

$$\nabla \mathcal{J}\left(u; \hat{u}, \hat{\mathbf{d}}\right) = \nabla \mathcal{J}(u; u_0, \mathbf{d}) - \nabla \mathcal{G}^*(u) \mathbf{L}^{-1} \tilde{\mathbf{d}} - \mathcal{C}^{-1} \tilde{u},$$

and (22) becomes

$$u^2 = u - \left[\nabla^2 \mathcal{J}_g(u)\right]^{-1} \nabla \mathcal{J}(u; u_0, \mathbf{d}) - \underbrace{\left[\nabla^2 \mathcal{J}_g(u)\right]^{-1} \left(\nabla \mathcal{G}^*(u) \mathbf{L}^{-1} \tilde{\mathbf{d}} + \mathcal{C}^{-1} \tilde{u}\right)}_{u^\dagger}.$$

Next, the proof of Lemma 3.4 shows that $u^\dagger$ is distributed by $\mathcal{N}\left(0, \left[\nabla^2 \mathcal{J}_g(u)\right]^{-1}\right)$. Therefore, u^2 and v_{SN} are identically distributed. The difference between the rMAP and SN methods is now clear: the SN method uses u^2 as the MCMC proposal, while the rMAP first continues to iterate until (21) is (approximately) satisfied and then takes the last u_k as the proposal. In this sense, rMAP can be viewed as an iterative SN method.

3.3. Relation between rMAP and the randomize-then-optimize approach. This section draws a connection between the rMAP method and the (RTO) approach [1]. We shall show that they are identical for a linear forward map (linear inverse problems) but different if the forward map is nonlinear. We also propose a modification for the RTO method.

The difference between RML and RTO is best demonstrated for finite dimensional parameter space. In this case, the jth rMAP can be computed as

$$(23) \qquad \boldsymbol{u}_j^{rMAP} := \arg\min_{\boldsymbol{u}} \frac{1}{2} \left| \mathbf{L}^{-\frac{1}{2}} \left(\mathbf{d} + \boldsymbol{\theta}_j - \mathcal{G}(\boldsymbol{u})\right) \right|^2 + \frac{1}{2} \left| \mathcal{C}^{-\frac{1}{2}} \left(\boldsymbol{u} - \boldsymbol{u}_0 - \boldsymbol{\varepsilon}_j\right) \right|^2,$$

while the jth RTO sample [1] can be written as

$$(24) \qquad \boldsymbol{u}_j^{RTO} := \arg\min_{\boldsymbol{u}} \frac{1}{2} \left\| \mathbf{Q}^T \left[\begin{array}{c} \mathbf{L}^{-\frac{1}{2}} \left(\mathcal{G}(\boldsymbol{u}) - \mathbf{d} - \boldsymbol{\theta}_j\right) \\ \mathcal{C}^{-\frac{1}{2}} \left(\boldsymbol{u} - \boldsymbol{u}_0 - \boldsymbol{\varepsilon}_j\right) \end{array} \right] \right\|^2,$$

where $\mathbf{Q}$ is the first factor in the "thin" QR factorization of

$$(25) \qquad \overline{\mathbf{G}} := \mathbf{G}\left(\boldsymbol{u}^{\mathrm{MAP}}\right) := \left[\mathbf{L}^{-\frac{1}{2}} \nabla \mathcal{G}\left(\boldsymbol{u}^{\mathrm{MAP}}\right), \mathcal{C}^{-\frac{1}{2}}\right]^T = \mathbf{Q}\mathbf{R}$$

evaluated at the MAP point. Due to the presence of $\mathcal{C}^{-1}$, $\mathbf{G}$ has full column rank, and hence $\mathbf{R}$ is invertible. *Clearly, rMAP samples $\boldsymbol{u}_j^{rMAP}$ are not the same as RTO samples $\boldsymbol{u}_j^{RTO}$ since they are extrema of different cost functions in general.* Now, let us assume that the forward map is linear, i.e., $\mathcal{G}(\boldsymbol{u}) = \mathbf{B}\boldsymbol{u}$. Setting the derivative, with respect to $\boldsymbol{u}$, of the cost function in (23) to zero yields the following equation for the jth rMAP sample $\boldsymbol{u}_j^{rMAP}$:

$$\overline{\mathbf{G}}^T \left[\begin{array}{c} \mathbf{L}^{-\frac{1}{2}} \left(\mathbf{B}\boldsymbol{u} - \mathbf{d} - \boldsymbol{\theta}_j\right) \\ \mathcal{C}^{-\frac{1}{2}} \left(\boldsymbol{u} - \boldsymbol{u}_0 - \boldsymbol{\varepsilon}_j\right) \end{array} \right] = 0.$$

Using (25) and the fact that $\mathbf{Q}$ is orthonormal, we arrive at

$$\overline{\mathbf{G}}^T \mathbf{Q} \mathbf{Q}^T \left[\begin{array}{c} \mathbf{L}^{-\frac{1}{2}} \left(\mathbf{B} \boldsymbol{u} - \mathbf{d} - \boldsymbol{\theta}_j \right) \\ \mathcal{C}^{-\frac{1}{2}} \left(\boldsymbol{u} - \boldsymbol{u}_0 - \boldsymbol{\varepsilon}_j \right) \end{array} \right] = 0,$$

which is exactly the equation for the jth RTO sample $\boldsymbol{u}_j^{RTO}$ if one sets the derivative, with respect to $\boldsymbol{u}$, of the cost function in (24) to zero. In other words, we have shown that *RTO is identical to rMAP for linear inverse problems.*

Up to this point we have observed that the RTO method requires a QR factorization of $\overline{\mathbf{G}}$ which could be computationally intractable for large-scale inverse problems in high dimensional parameter spaces. We propose to use $\overline{\mathbf{G}}$ in place of $\mathbf{Q}$. For a general forward map, the modified RTO problem reads (compared to (24))

$$(26) \qquad \boldsymbol{u}_j^{RTO} := \arg \min_{\boldsymbol{u}} \frac{1}{2} \left\| \overline{\mathbf{G}}^T \left[\begin{array}{c} \mathbf{L}^{-\frac{1}{2}} \left(\mathbf{B} \boldsymbol{u} - \mathbf{d} - \boldsymbol{\theta}_j \right) \\ \mathcal{C}^{-\frac{1}{2}} \left(\boldsymbol{u} - \boldsymbol{u}_0 - \boldsymbol{\varepsilon}_j \right) \end{array} \right] \right\|^2,$$

and hence RTO samples now satisfy the equation

$$(27) \qquad \mathbf{G}^T \left(\boldsymbol{u} \right) \overline{\mathbf{G}} \, \overline{\mathbf{G}}^T \left[\begin{array}{c} \mathbf{L}^{-\frac{1}{2}} \left(\mathbf{B} \boldsymbol{u} - \mathbf{d} - \boldsymbol{\theta}_j \right) \\ \mathcal{C}^{-\frac{1}{2}} \left(\boldsymbol{u} - \boldsymbol{u}_0 - \boldsymbol{\varepsilon}_j \right) \end{array} \right] = 0.$$

The modified approach has a couple of advantages: (1) QR-factorization of (possibly large-scale) $\overline{\mathbf{G}}$ is no longer needed, and (2) there is no need to construct $\overline{\mathbf{G}}$ since all we need is its action, which can be computed efficiently using an adjoint technique. The determinant of $\mathbf{G} \left(\boldsymbol{u}^{RTO} \right)^T \overline{\mathbf{G}}$ is necessary if the RTO density is needed, but this is readily available from the MAP calculation and the computation of $\boldsymbol{u}^{RTO}$.

4. Metropolis-adjusted rMAP method. Recall from Lemma 3.4 that for linear inverse problems, the rMAP sample is exactly distributed by the posterior measure ν. When the forward map is nonlinear, Proposition 3.2 shows that this is no longer true. In this case, rMAP samples have bias which should be removed via, for example, the standard Metropolization [53]. The work in [48] shows that for some nonlinear test problems, the acceptance rate is above 90%, and the authors proposed to accept all rMAP samples. This simple strategy has been shown to work well in many cases (see, e.g., [34, 49]), though the resulting Markov chain can over/underestimate the actual posterior. We shall show that this is the case for our inverse problem, and a debiasing procedure is necessary. An exact Metropolization has been proposed in [48], but it is intractable except for problems with (very) small parameter dimension. We therefore propose an approximate Metropolized step, and this is done using a finite dimensional framework. To that end, we replace $\hat{u}$ by finite dimensional vector $\boldsymbol{u}$, e.g., a vector of finite element nodal values.

Following [46], we begin by defining

$$(28) \qquad \delta = \mathcal{G} \left(\hat{\boldsymbol{u}}^{\mathrm{MAP}} \right) - \hat{\mathbf{d}}.$$

Note that $\hat{\boldsymbol{u}}^{\mathrm{MAP}}$ also satisfies (21), which for the finite dimensional setting becomes

$$(29) \qquad \nabla \mathcal{G}^* \left(\hat{\boldsymbol{u}}^{\mathrm{MAP}} \right) \mathbf{L}^{-1} \left[\mathcal{G} \left(\hat{\boldsymbol{u}}^{\mathrm{MAP}} \right) - \hat{\mathbf{d}} \right] + \mathcal{C}^{-1} \left(\hat{\boldsymbol{u}}^{\mathrm{MAP}} - \hat{\boldsymbol{u}} \right) = 0.$$

We can view (28) and (29) together as the definition of a map $\mathcal{T} : (\hat{\boldsymbol{u}}, \hat{\mathbf{d}}) \rightarrow (\hat{\boldsymbol{u}}^{\mathrm{MAP}}, \delta)$, and we assume that this map is locally invertible. This allows us to

explicitly write $\mathcal{T}^{-1}$ as

$$(30) \qquad \begin{bmatrix} \hat{\boldsymbol{u}} \\ \hat{\mathbf{d}} \end{bmatrix} = \mathcal{T}^{-1}(\hat{\boldsymbol{u}}^{\mathrm{MAP}}, \delta) = \begin{bmatrix} \hat{\boldsymbol{u}}^{\mathrm{MAP}} + \mathcal{C}\nabla\mathcal{G}^* \left(\hat{\boldsymbol{u}}^{\mathrm{MAP}}\right) \mathbf{L}^{-1}\delta \\ \mathcal{G}\left(\hat{\boldsymbol{u}}^{\mathrm{MAP}}\right) - \delta \end{bmatrix}.$$

That is, we know $\mathcal{T}$ implicitly through its inverse. Given the fact that the distribution of $(\hat{\boldsymbol{u}}, \hat{\mathbf{d}})$ is available, the distribution of $(\hat{\boldsymbol{u}}^{\mathrm{MAP}}, \delta)$ can be computed via the Jacobian of the transformation:

$$J := \frac{\partial \left(\hat{\boldsymbol{u}}, \hat{\mathbf{d}}\right)}{\partial \left(\hat{\boldsymbol{u}}^{\mathrm{MAP}}, \delta\right)} = \begin{bmatrix} \boldsymbol{I} + \mathcal{C}\nabla^2\mathcal{G}^*\left(\hat{\boldsymbol{u}}^{\mathrm{MAP}}\right) \otimes \mathbf{L}^{-1}\delta & \mathcal{C}\nabla\mathcal{G}^*\left(\hat{\boldsymbol{u}}^{\mathrm{MAP}}\right) \mathbf{L}^{-1} \\ \nabla\mathcal{G}\left(\hat{\boldsymbol{u}}^{\mathrm{MAP}}\right) & -\boldsymbol{I} \end{bmatrix}.$$

Simple algebra yields the determinant as

$$(31) \quad |J| = \left| \det\left(\boldsymbol{I} + \mathcal{C}\nabla\mathcal{G}^*\left(\hat{\boldsymbol{u}}^{\mathrm{MAP}}\right) \mathbf{L}^{-1}\nabla\mathcal{G}\left(\hat{\boldsymbol{u}}^{\mathrm{MAP}}\right) + \mathcal{C}\nabla^2\mathcal{G}^*\left(\hat{\boldsymbol{u}}^{\mathrm{MAP}}\right) \otimes \mathbf{L}^{-1}\delta \right) \right|,$$

which is nothing but the determinant of $\nabla^2 J$ scaled by a determinant of the prior covariance.

Let us denote by $h\left(\hat{\boldsymbol{u}}^{\mathrm{MAP}}, \delta\right)$ the density of proposing the pair $\left(\hat{\boldsymbol{u}}^{\mathrm{MAP}}, \delta\right)$ via the map $\mathcal{T}$ described above. Clearly, it is the push-forward of the probability of the pair $\left(\hat{\boldsymbol{u}}, \hat{\mathbf{d}}\right)$. By the measure preservation property and the change of variables formula, we have

$$h\left(\hat{\boldsymbol{u}}^{\mathrm{MAP}}, \delta\right) = f\left(\mathcal{T}^{-1}\left(\hat{\boldsymbol{u}}^{\mathrm{MAP}}, \delta\right)\right) |J|,$$

where f is defined as

$$(32) \qquad f\left(\hat{\boldsymbol{u}}, \hat{\mathbf{d}}\right) \sim \exp\left[-\frac{1}{2}\left(\hat{\boldsymbol{u}} - \boldsymbol{u}_0\right)^T \mathcal{C}^{-1}\left(\hat{\boldsymbol{u}} - \boldsymbol{u}_0\right) - \frac{1}{2}\left(\hat{\mathbf{d}} - \mathbf{d}\right)^T \mathbf{L}^{-1}\left(\hat{\mathbf{d}} - \mathbf{d}\right) \right].$$

While the marginal distribution of $\hat{\boldsymbol{u}}^{\mathrm{MAP}}$ is desirable, the marginalization process on $h\left(\hat{\boldsymbol{u}}^{\mathrm{MAP}}, \delta\right)$ is not trivial (as shall be shown). This suggests that we can conduct the sampling in the augmented space defined by $\left(\hat{\boldsymbol{u}}^{\mathrm{MAP}}, \delta\right)$. What remains is to construct a joint posterior density of $\left(\hat{\boldsymbol{u}}^{\mathrm{MAP}}, \delta\right)$ such that marginalizing out δ yields exactly the posterior π_{post}. Similar to [47, 60], we may choose the joint posterior as

$$\pi\left(\hat{\boldsymbol{u}}^{\mathrm{MAP}}, \delta\right) \propto \exp\left[-\frac{1}{2}\left(\hat{\boldsymbol{u}}^{\mathrm{MAP}} - \boldsymbol{u}_0\right)^T \mathcal{C}^{-1}\left(\hat{\boldsymbol{u}}^{\mathrm{MAP}} - \boldsymbol{u}_0\right) \right.$$
$$\left. -\frac{\eta_1}{2}\left(\mathcal{G}(\hat{\boldsymbol{u}}^{\mathrm{MAP}}) - \delta - \mathbf{d}_0\right)^T \mathcal{L}^{-1}\left(\mathcal{G}(\hat{\boldsymbol{u}}^{\mathrm{MAP}}) - \delta - \mathbf{d}_0\right) - \frac{\eta_2}{2}\delta^T \mathcal{L}^{-1}\delta \right].$$

It can be shown that if $\eta_1 = \eta_2/(\eta_2 - 1)$, the marginal distribution of $\hat{\boldsymbol{u}}^{\mathrm{MAP}}$ is exactly the posterior distribution. Therefore, the conventional Metropolis–Hastings algorithm can be applied to the joint distribution $\pi\left(\hat{\boldsymbol{u}}^{\mathrm{MAP}}, \delta\right)$, with $h\left(\hat{\boldsymbol{u}}^{\mathrm{MAP}}, \delta\right)$ as the proposal distribution. The acceptance rate in this case reads

$$\alpha\left(\hat{\boldsymbol{u}}^{\mathrm{MAP}}, \delta\right) = \min\left(1, \frac{\pi\left(\hat{\boldsymbol{u}}^{\mathrm{MAP}*}, \delta^*\right) h\left(\hat{\boldsymbol{u}}_k^{\mathrm{MAP}}, \delta_k\right)}{\pi\left(\hat{\boldsymbol{u}}_k^{\mathrm{MAP}}, \delta_k\right) h\left(\hat{\boldsymbol{u}}^{\mathrm{MAP}*}, \delta^*\right)} \right),$$

where $\left(\hat{\boldsymbol{u}}_k^{\mathrm{MAP}}, \delta_k\right)$ is the previous sample and $\left(\hat{\boldsymbol{u}}^{\mathrm{MAP}_*}, \delta^*\right)$ is the proposed sample for the next state.

While the above augmented space method guarantees that $\hat{\boldsymbol{u}}_k^{\mathrm{MAP}}$ is correctly distributed by the posterior distribution π_{post} in the limit, it can be prohibitively expensive, especially for large-scale inverse problems. The reason is that adding the data δ increases dimensionality, which can be significant if the data dimension is large. Though MCMC methods are independent of the dimension, the number of samples could be excessively large in order to obtain a reasonable result (the curse of dimensionality). Moreover, the evaluation of the full Hessian as required in the computation of the Jacobian can also be very expensive. To address these challenges, we now present an approximate marginalization of δ that allows us to carry out the MCMC method in the original space of $\hat{\boldsymbol{u}}^{\mathrm{MAP}}$. A direct consequence of this approximation is that $\hat{\boldsymbol{u}}_k^{\mathrm{MAP}}$ are no longer the truth samples of the posterior π_{post} even in the limit. Nevertheless, our experiences, including the numerical results presented in this paper, show that the results from the approximate Markov chain are very close to those from the genuine Markov chain. To begin, we observe that

$$f\left(\mathcal{T}^{-1}\left(\hat{\boldsymbol{u}}^{\mathrm{MAP}}, \delta\right)\right) = p\left(\hat{\boldsymbol{u}}^{\mathrm{MAP}}\right) \zeta\left(\delta\right) \eta\left(\hat{\boldsymbol{u}}^{\mathrm{MAP}}\right),$$

where

$$p\left(\hat{\boldsymbol{u}}^{\mathrm{MAP}}\right) = \exp\left(-\frac{1}{2}\left|\hat{\boldsymbol{u}}^{\mathrm{MAP}} - \boldsymbol{u}_0\right|_{\mathcal{C}}^2 - \frac{1}{2}\left|\mathcal{G}\left(\hat{\boldsymbol{u}}^{\mathrm{MAP}}\right) - \mathbf{d}_0\right|_{\mathbf{L}}^2\right)$$

is proportional to the posterior distribution,

$$\zeta\left(\delta\right) = \exp\left(-\frac{1}{2}\left(\delta - \mathcal{H}\mathcal{K}\right)^T \mathcal{H}^{-1}\left(\delta - \mathcal{H}\mathcal{K}\right)\right),$$

and

$$\eta\left(\hat{\boldsymbol{u}}^{\mathrm{MAP}}\right) = \exp\left(\frac{1}{2}\mathcal{K}^T \mathcal{H}\mathcal{K}\right),$$

where $\mathcal{H}$ and $\mathcal{K}$ are given by

$$\mathcal{H}^{-1} = \mathbf{L}^{-1} + \mathbf{L}^{-1}\nabla\mathcal{G}\left(\hat{\boldsymbol{u}}^{\mathrm{MAP}}\right)\mathcal{C}\nabla\mathcal{G}^*(\hat{\boldsymbol{u}}^{\mathrm{MAP}})\mathbf{L}^{-1}$$

and

$$\mathcal{K} = \mathbf{L}^{-1}\left(\left(\mathcal{G}\left(\hat{\boldsymbol{u}}^{\mathrm{MAP}}\right) - \mathbf{d}_0\right) + \nabla\mathcal{G}\left(\hat{\boldsymbol{u}}^{\mathrm{MAP}}\right)\left(\hat{\boldsymbol{u}}^{\mathrm{MAP}} - \boldsymbol{u}_0\right)\right).$$

Since the terms including δ constitute a Gaussian kernel, such a decomposition allows us to marginalize δ and obtain the probability of proposing $\hat{\boldsymbol{u}}^{\mathrm{MAP}}$:

$$q\left(\hat{\boldsymbol{u}}^{\mathrm{MAP}}\right) = \int h\left(\hat{\boldsymbol{u}}^{\mathrm{MAP}}, \delta\right) d\delta = p\left(\hat{\boldsymbol{u}}^{\mathrm{MAP}}\right) \eta\left(\hat{\boldsymbol{u}}^{\mathrm{MAP}}\right) \omega\left(\hat{\boldsymbol{u}}^{\mathrm{MAP}}\right) |J|,$$

where $\omega\left(\hat{\boldsymbol{u}}^{\mathrm{MAP}}\right)$ results from integrating with respect to δ, and it possesses the explicit form

$$\omega\left(\hat{\boldsymbol{u}}^{\mathrm{MAP}}\right) \propto |\mathcal{H}|^{\frac{1}{2}} = |\mathbf{L}|^{-\frac{1}{2}} |J|^{-\frac{1}{2}}.$$

Substituting these formulas into the decomposition of $q\left(\hat{\boldsymbol{u}}^{\mathrm{MAP}}\right)$, we obtain the following ratio of posterior distribution over proposal distribution:

$$\theta\left(\hat{\boldsymbol{u}}^{\mathrm{MAP}}\right) = \frac{p\left(\hat{\boldsymbol{u}}^{\mathrm{MAP}}\right)}{q\left(\hat{\boldsymbol{u}}^{\mathrm{MAP}}\right)} \propto \exp\left(-\frac{1}{2}\mathcal{K}^T\mathcal{H}\mathcal{K}\right)|\mathbf{L}|^{\frac{1}{2}}\,|J|^{-\frac{1}{2}}\,.$$

With this ratio, we are able to compute the acceptance ratio between a newly proposed state $\hat{\boldsymbol{u}}_*^{\mathrm{MAP}}$ and a current state $\hat{\boldsymbol{u}}_k^{\mathrm{MAP}}$. One computational consideration in practice would be that directly computing the gradient of the forward map, $\nabla\mathcal{G}\left(\hat{\boldsymbol{u}}^{\mathrm{MAP}}\right)$, can be expensive when the number of measurements is high. A further practical simplification would be approximating α with only the $|\mathbf{L}|^{\frac{1}{2}}\,|J|^{-\frac{1}{2}}$. Thus, the acceptance ratio we adopt has the form

$$(33) \qquad \tilde{\alpha}\left(\hat{\boldsymbol{u}}_*^{\mathrm{MAP}}, \hat{\boldsymbol{u}}_k^{\mathrm{MAP}}\right) = \frac{\theta\left(\hat{\boldsymbol{u}}_*^{\mathrm{MAP}}\right)}{\theta\left(\hat{\boldsymbol{u}}_k^{\mathrm{MAP}}\right)} \approx \frac{\left|J(\hat{\boldsymbol{u}}_k^{\mathrm{MAP}})\right|^{\frac{1}{2}}}{\left|J(\hat{\boldsymbol{u}}_*^{\mathrm{MAP}})\right|^{\frac{1}{2}}}.$$

In addition, we drop the higher order term in the Jacobian. In other words, we replace the full Hessian in (31) with a Gauss–Newton Hessian. These simplifications appear to be reasonable, as we will show in the numerical results.

It should be pointed out that we have recently shown that the misfit (Gauss–Newton) Hessian is a compact operator [10, 11]. Moreover, $\mathcal{C}^{\frac{1}{2}}$ is also a compact operator by the definition of the Gaussian measure. It follows that $\mathcal{C}^{\frac{1}{2}}\nabla^2\Phi_g\left(\hat{\boldsymbol{u}}^{\mathrm{MAP}}, \hat{\mathbf{d}}\right)\mathcal{C}^{\frac{1}{2}}$ is compact and admits low rank approximation. This is in fact one of the key points that is exploited to construct a scalable and mesh-independent method in our previous work on extreme scale Bayesian inversion [9, 14]. Thus, computing $|J|$ can be done in a scalable manner independent of the mesh size using the randomized SVD technique [31], for example.

5. Finite element discretization and optimization. For the practical problems we consider we assume the spatial dimension to be at least two; therefore we choose $s > 1$ so that the infinite dimensional framework is well-defined as discussed in section 2. As a result, evaluating the prior and/or generating a prior sample requires us to discretize and/or solve a fractional PDE. Similarly to [8] (and references therein) we combine the finite element method (FEM) [18] and the matrix transfer technique (see, e.g., [35]) to discretize the (truncated) Karhunen–Loève (KL) expansion of the prior. For the discretization of the forward equation and hence the likelihood, we also use the same FEM.

Using finite element approximation, the MAP problem (5) becomes a (possibly) high dimensional and nonlinear optimization problem. It is thus necessary to use the state-of-the-art scalable optimization solver to minimize the cost. Here we choose the trust region inexact Newton conjugate gradient (TRINCG) method, for which some of the main ideas can be found in, e.g., [5, 7, 19, 45]. The method combines the rapid locally quadratic convergence rate properties of the Newton method, the effectiveness of trust region globalization for treating ill-conditioned problems, and the Eisenstat–Walker idea of preventing oversolving. In the numerical results section, we demonstrate the efficiency of this trust region method over popular Levenberg–Marquardt (LM) techniques. As we shall see, in some difficult examples, choosing TRINCG becomes critical in controlling computation time for rMAP sampling.

5.1. Good initial guess for the rMAP algorithm. One of the most important aspects of numerical optimization, particularly with the Newton method, is how to choose a good initial guess. The closer the initial guess is to the basin of attraction of a local minimum, the faster the convergence. This is clearly important since we desire to minimize the cost of computing rMAP proposals. One way to achieve this is through using sensitivity analysis, which we now describe. To begin, we distinguish ∇, the derivative with respect to u, from derivatives with other variables: for example, $\nabla_{\hat{u}_i}$ and $\nabla_{\hat{\mathbf{d}}_i}$ denote derivatives with respect to $\hat{u}_i$ and $\hat{\mathbf{d}}_i$, respectively. Consider two consecutive rMAP samples $\hat{u}_i^{\mathrm{MAP}}$ and $\hat{u}_{i+1}^{\mathrm{MAP}}$ that satisfy

$$(34) \qquad F\left(\hat{u}_i^{\mathrm{MAP}}; \hat{u}_i, \hat{\mathbf{d}}_i\right) := \nabla \mathcal{J}\left(\hat{u}_i^{\mathrm{MAP}}; \hat{u}_i, \hat{\mathbf{d}}_i\right) = 0,$$

$$(35) \qquad F\left(\hat{u}_{i+1}^{\mathrm{MAP}}; \hat{u}_{i+1}, \hat{\mathbf{d}}_{i+1}\right) := \nabla \mathcal{J}\left(\hat{u}_{i+1}^{\mathrm{MAP}}; \hat{u}_{i+1}, \hat{\mathbf{d}}_{i+1}\right) = 0.$$

Now, let us define

$$\tilde{u} = \hat{u}_{i+1} - \hat{u}_i \quad \text{and} \quad \tilde{\mathbf{d}} = \hat{\mathbf{d}}_{i+1} - \hat{\mathbf{d}}_i.$$

Assuming that $\hat{u}_i^{\mathrm{MAP}}$ is already computed from (34), we now construct an initial guess for solving (35) using the Newton method:

$$(36) \qquad u^{init} = \hat{u}_i^{\mathrm{MAP}} + \underbrace{\left\langle \nabla_{\hat{\mathbf{d}}_i} \hat{u}_i^{\mathrm{MAP}}, \tilde{\mathbf{d}} \right\rangle + \left\langle \nabla_{\hat{u}_i} \hat{u}_i^{\mathrm{MAP}}, \tilde{u} \right\rangle}_{T},$$

which is simply the first order Taylor approximation of $\hat{u}_{i+1}^{\mathrm{MAP}}$ around $\left(\hat{u}_i, \hat{\mathbf{d}}_i\right)$.

What remains is to compute T in (36). To this end, we expand the gradient in (35) using the first order Taylor expansion to obtain the following equation for T:

$$(37) \qquad \nabla^2 \mathcal{J}\left(\hat{u}_i^{\mathrm{MAP}}; \hat{u}_i, \hat{\mathbf{d}}_i\right) T \approx \nabla \mathcal{G}^*\left(\hat{u}_i^{\mathrm{MAP}}\right) \mathbf{L}^{-1} \tilde{\mathbf{d}} + \mathcal{C}^{-1} \tilde{u}.$$

Solving (37) requires an adjoint solve to evaluate the right-hand side and the inverse of $\nabla^2 \mathcal{J}\left(\hat{u}_i^{\mathrm{MAP}}; \hat{u}_i, \hat{\mathbf{d}}_i\right)$ (the Hessian evaluated at the ith rMAP sample). If $\left|\hat{u}_i^{\mathrm{MAP}} - \hat{u}_{i+1}^{\mathrm{MAP}}\right|$ is small, u^{init} is a very good approximation of $\hat{u}_{i+1}^{\mathrm{MAP}}$. Thus, solving (35) with u^{init} as the initial guess helps substantially reduce the number of optimization iterations (and hence the number of forward PDE solves). In practice, we linearize around the MAP point (5), and this approach further reduces the number of PDE solves since $\nabla^2 \mathcal{J}\left(u^{\mathrm{MAP}}; u_0, \mathbf{d}\right)$ is fixed and can be well approximated using low rank approximation [9, 14].

6. Numerical results. In this section, we present sampling results using several test cases. In section 6.1, we first demonstrate the effectiveness of using the approximated Metropolization for efficient rMAP sampling. We then use two analytical functions to compare the sampling efficiency of the approximated rMAP to the RTO method and to the SN method described above. In section 6.2, we again use the approximated rMAP method to sample a Bayesian inverse problem on a 2D Helmholtz forward model. Therein, we compare the computational efficiency of the popular LM method (see, e.g., [49]) to that of the TRINCG method for each rMAP sample and the effectiveness of using a good initial guess as discussed in section 5. In order to examine statistical convergence of rMAP methods, we also compare rMAP samples with those from the delayed rejection adaptive Metropolis (DRAM) sampler [29].

6.1. Analytical function example. Let us start by numerically demonstrating how rMAP and RTO cost functions in (23) and (24), respectively, change the original cost function in (5). To this end, we consider two analytical cost functions (negative log posterior)

$$(38a) \qquad \mathcal{J}_1 := \frac{1}{2}\left(\boldsymbol{u} - 0.8\right)^2 + \frac{1}{2 \times 0.2^2}\left(\boldsymbol{u}^2 - 1\right)^2,$$

$$(38b) \qquad \mathcal{J}_2 := \frac{1}{2}\left(\boldsymbol{u} - 1\right)^2 + \frac{1}{2 \times 0.2^2}\left(\boldsymbol{u}^3 - 0.8\right)^2.$$

6.1.1. Comparison of augmented space Metropolization and its approximation. We first compare the accuracy and statistical efficiency of the augmented space Metropolization and its approximation presented in section 4. To this end we choose to work with the cost function (38a) but with three different data variances $\sigma = \{0.2, 0.5, 1.0\}$, i.e.,

$$\mathcal{J}_1 := \frac{1}{2}\left(\boldsymbol{u} - 0.8\right)^2 + \frac{1}{2\sigma^2}\left(\boldsymbol{u}^2 - 1.0\right)^2,$$

to construct three different posterior distributions. We sample these distributions with both augmented space Metropolized rMAP and the approximated Metropolized rMAP with 5000 samples. We plot the histograms at 200, 1000, and 5000 samples in Figure 1. The results seem to indicate that the approximated method (magenta histograms) converges more rapidly than the augmented space counterpart (cyan histograms). This is also reflected in a comparison of acceptance ratio in Figure 2, in which we observe that the acceptance rate for the approximate method is higher. One of the reasons for the low acceptance rate of the augmented space method is again due to the increase of dimensionality. For very skewed distribution (e.g., the fifth and the sixth row of Figure 1), however, the approximate method is less accurate than the augmented space counterpart. For the other cases, both methods are comparable. Recall that the approximate method is also much less expensive. For these reasons, we will use the approximate method throughout the rest of the paper.

6.1.2. Comparing rMAP and RTO methods. Figure 3 shows the original cost functionals $\mathcal{J}_1, \mathcal{J}_2$ and their randomization with rMAP and RTO methods. (Note that both the original RTO and our modified version give identical results for all analytical results, and hence we do not distinguish them.) Here, we use the same $\boldsymbol{\theta}$ and $\boldsymbol{\varepsilon}$ for both rMAP and RTO. As can be seen, both randomized costs preserve the characteristics, e.g., multimodality and skewness, of the original cost function. However, they differ from the original cost function as well as from one other, which agrees with our findings in section 3.3.

We next examine the sensitivity of both rMAP and RTO with multimodality and optimization solvers. To that end, we first use the MATLAB routine *fminunc*, the unconstrained optimization solver, and use the MAP point as the initial guess to compute rMAP and RTO samples for the $\mathcal{J}_1$ cost functional. As can be seen in Figures 4(a) and 4(d), both methods are stuck in a mode. Instead, if we use $\hat{\boldsymbol{u}}_j := \boldsymbol{u}_0 + \boldsymbol{\varepsilon}_j$ as an initial guess for computing the jth sample, we obtain the results in Figures 4(b) and 4(e), respectively. Clearly, both methods explore both modes well. Thus, *for rMAP and RTO to work with the local optimization solver, it is important that initial guesses are well distributed in the parameter space.* In fact, good initial guesses also help significantly reduce the number of forward solves, as we will show in the following subsection.
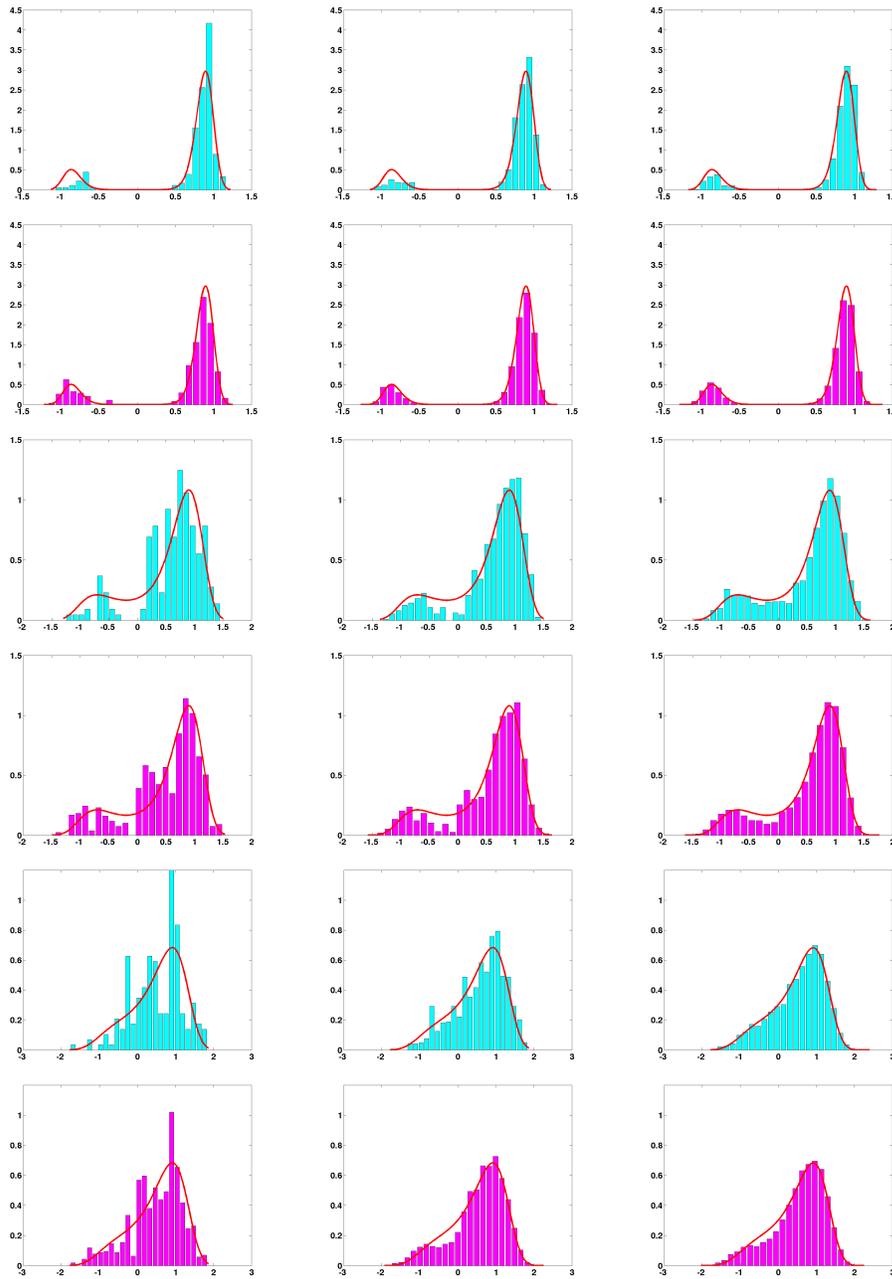
FIG. 1. *Comparison of sampling efficiency between the augmented space Metropolization and its approximation. From left to right, the number of samples in each column is* 200, 1000, *and* 5000. *From top to bottom, cyan plots are histograms produced by augmented space methods and magenta plots are produced by the approximated Metropolization. (See online version for color.)*

As a comparison, we employ the MATLAB constrained optimization solver *fminbnd,* with prescribed bound $-100 \leq \boldsymbol{u} \leq 100$ more than sufficient to cover the
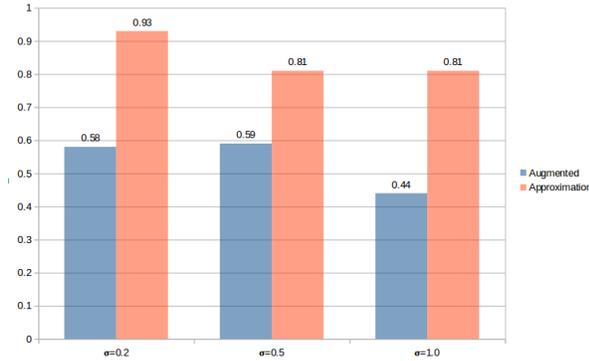
FIG. 2. *Comparison of acceptance ratios between the augmented space Metropolization and its approximation. In all cases, augmented space methods have a lower acceptance ratio due to the higher dimensionality.*

(a) $\mathcal{J}_1$, and its rMAP and RTO.   (b) $\mathcal{J}_2$, and its rMAP and RTO.
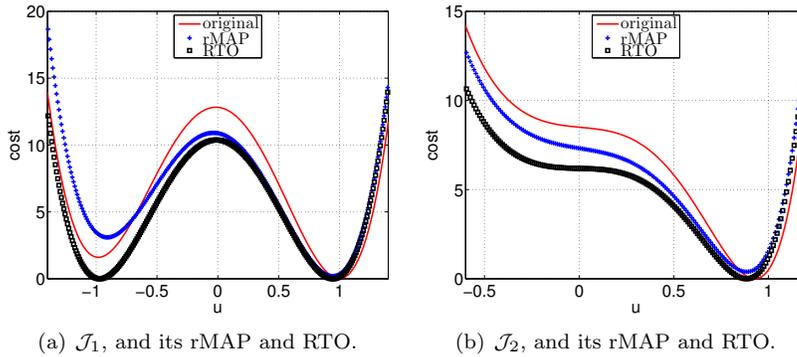
FIG. 3. *Randomization of the cost functionals in* (38) *with rMAP and RTO methods.*

modes. This optimization solver computes initial guesses using the *golden section rule*. The results for rMAP and RTO are shown in Figures 4(c) and 4(f): rMAP still works well in this case, while RTO is stuck in the left mode. Thus, rMAP seems to be more robust with optimization solvers. From numerical experiments we observe that rMAP tends to displace the original function more than RTO does, and this may partially explain the robustness of the former. However, in this experiment, rMAP seems to have introduced an artificial mode for the original function, as we now show in Figure 5, for cost function $\mathcal{J}_2$. Note that the original cost function $\mathcal{J}_2$ has only one mode, but it can become multimodal for a range of $\boldsymbol{\varepsilon}$ and $\boldsymbol{\theta}$. As can be observed in Figures 5(a) and 5(c), rMAP puts a lot of samples in an artificial mode that was not in the original function, while RTO does not seem to generate the artifact. With the square root Jacobian correction in section 4, we can, in Figure 5(b), both remove that artificial mode and improve the histogram for the actual mode. We can also improve the RTO samples by first taking the RTO density as the importance sampling density and then using the importance weights to correct for RTO samples. The result in Figure 5(d) shows that this strategy effectively improves RTO's estimation of the posterior distribution.

(a) rMAP: MAP initial guess. (b) rMAP: random initial guess. (c) rMAP: "Golden section" initial guess.

(d) RTO: MAP initial guess. (e) RTO: random initial guess. (f) RTO: "Golden section" initial guess.
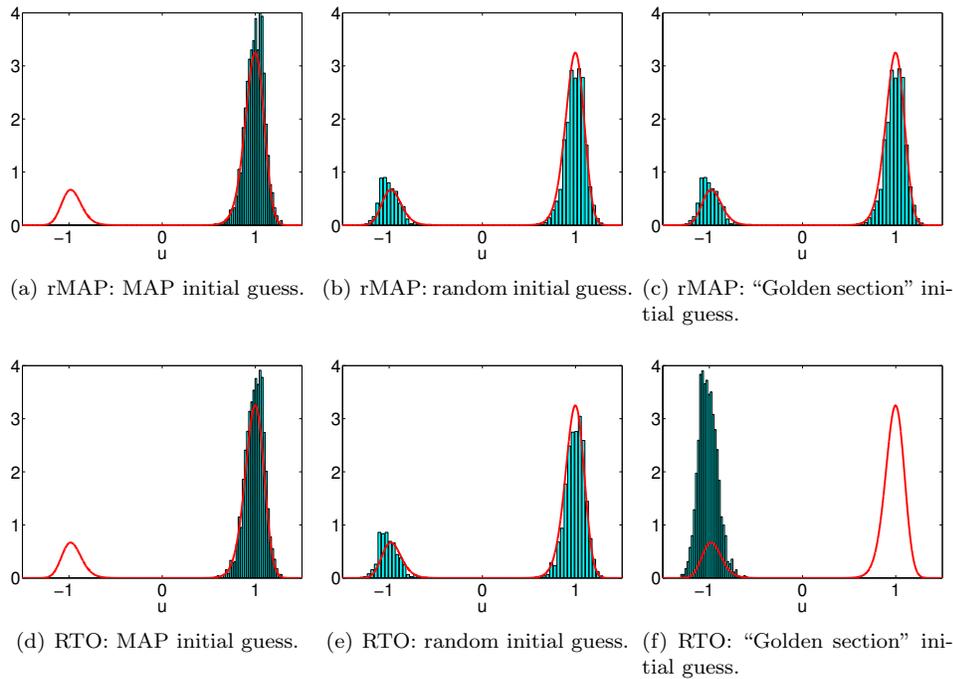
Fig. 4. *Sensitivity of rMAP and RTO with local optimization solvers and initial guesses. Figures* 4(a) *and* 4(d) *show the results of* fminunc *and MAP initial guess. Figures* 4(b) *and* 4(e) *show the results of* fminunc *and random prior means as initial guesses. Figures* 4(c) *and* 4(f) *show the results of* fminbnd *and the default golden section rule initial guess. The cost functional* $\mathcal{J}_1$ *is used to conduct these experiments.*

6.1.3. Comparing rMAP and SN methods. In this section, we will numerically confirm our discussion from section 3.2 on the improvement of rMAP over the SN method. For concreteness, we choose $\mathcal{J}_1$ in (38a), a multimodal function, for the comparison. We have shown in section 3.2 that rMAP can be viewed as an iterative SN method. It is this deterministic iteration that can help rMAP explore the sample space more rapidly. In particular, rMAP can be interpreted as a globalization strategy. It is in fact a move away from the inefficiencies of random-walk/diffusion processes toward powerful optimization methods that use derivative information to traverse the posterior.

For numerical comparison, we compute 1000 samples from the Metropolis-adjusted rMAP sampler, and in this case the total number of Newton iterations is approximately 20,000. Since the parameter dimension is one, the total number of (forward and adjoint) PDE solves is 40,000. For the SN Newton method, we take 100,000 samples. Three independent chains with three different initial states, namely the origin and the left and right modes of the posterior distribution, are computed for both samplers. Figure 6 shows the histogram of each chain together with the exact density. We observe that rMAP chains are capable of sampling both modes and the sampling results are independent of starting points. On the contrary, SN chains show dependency on the starting points, and they are stuck in local minima.

6.1.4. Statistical convergence of rMAP. We also numerically examine Proposition 3.2 using cost function $\mathcal{J}_1$. First, we compute the expectation $\mathbb{E}_{\boldsymbol{\theta} \times \boldsymbol{\varepsilon}}[S(\boldsymbol{u}_0, \mathbf{d}, \boldsymbol{\theta}, \boldsymbol{\varepsilon})]$

(a) rMAP.   (b) Corrected rMAP.

(c) RTO.   (d) Corrected RTO.
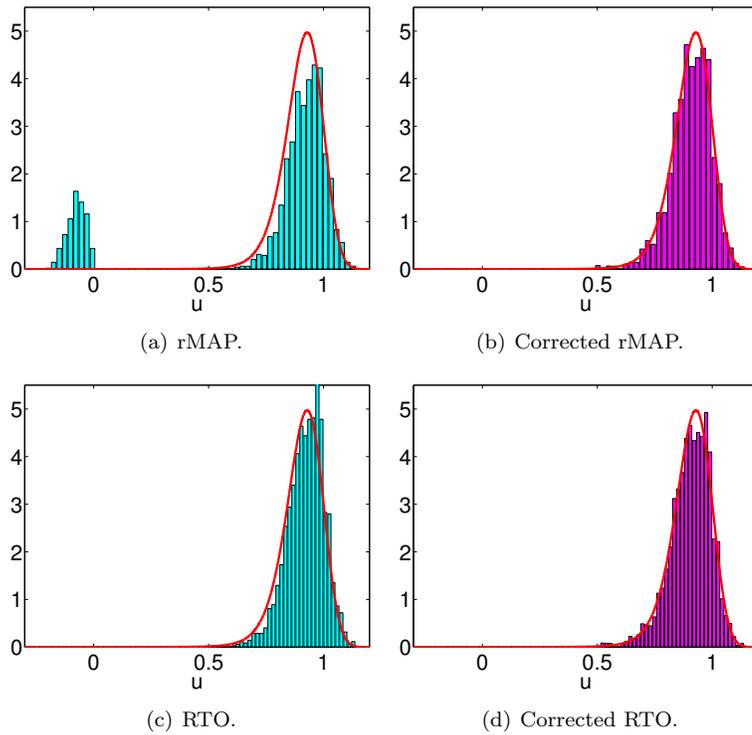
Fig. 5. *Illustration of artificial mode created by rMAP and correction strategies for both rMAP and RTO. The correction for rMAP was achieved by the square root of the Jacobian in section 4, and the correction for RTO was obtained via important sampling weights. The numerical experiments were done for the cost functional $\mathcal{J}_2$.*

using a tensor product Gauss–Hermite quadrature. Ten independent rMAP chains are computed, each of which has one million samples. We compute the averages $\{\frac{1}{n}\sum_{j=1}^{n}\boldsymbol{u}_j\}_{n=1}^{N}$, $N = 10^6$, over each chain, and the results are compared to the quadrature-based expectation. In Figure 7, it is shown that the approximate mean of rMAP samples aligns well with the limit $\mathbb{E}_{\boldsymbol{\theta}\times\boldsymbol{\varepsilon}}\left[S\left(\boldsymbol{u}_0,\mathbf{d},\boldsymbol{\theta},\boldsymbol{\varepsilon}\right)\right]$ and hence confirms our theoretical result in Proposition 3.2.

6.2. Helmholtz problems. Although our proposed framework is valid for Bayesian inverse problems governed by any system of forward PDEs, here we illustrate the use of the framework on a frequency domain acoustic wave equation in the form of the Helmholtz equation. Namely, the forward model $\mathcal{B}\left(u,w\right)$ is defined, in an open and bounded domain Ω, as

$$-\nabla^2 w - e^{2u}w = 0 \text{ in } \Omega,$$

$$\frac{\partial w}{\partial \mathbf{n}} = g \text{ on } \partial\Omega,$$

where w is the acoustic field, u is the logarithm of the distributed wave number field on Ω, $\mathbf{n}$ is the unit outward normal on $\partial\Omega$, and g is the prescribed Neumann source on the boundary.

In subsection 6.2.1, we first discuss the computation of the gradient and Hessian of the objective function using the adjoint method. The adjoint method enables tractable

(a) rMAP.

(b) SN (initial state at the left mode).

(c) SN (initial state at the origin).

(d) SN (initial state at the right mode).
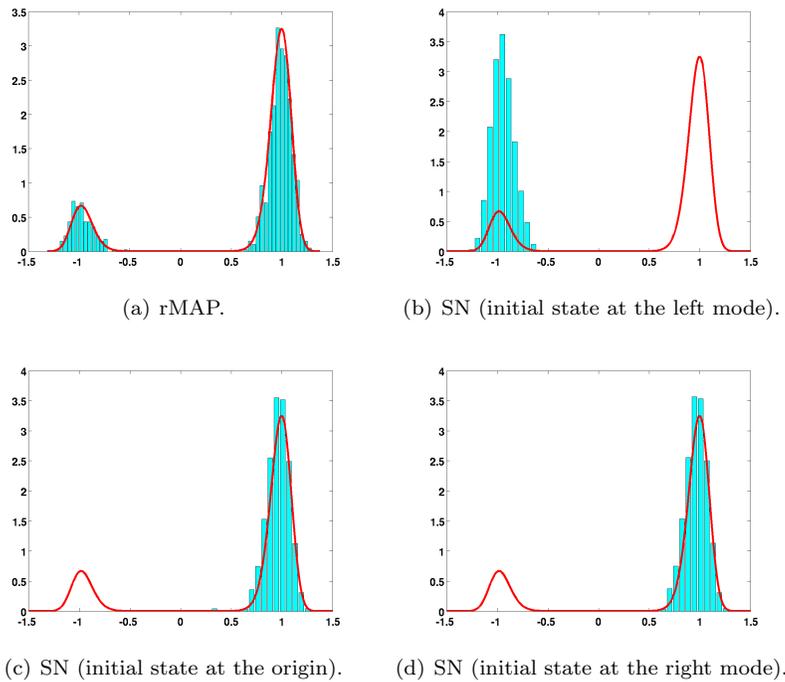
FIG. 6. *Comparison of Metropolis-adjusted rMAP and SN MCMC methods for sampling multimodal problems. Three starting points are chosen for these two samplers, namely the left mode, zero, and the right mode. The histograms are the same irrespective of the starting points for the rMAP method, and hence only one plot is shown here. While SN chains are trapped in local minima, rMAP counterparts traverse the posterior very well.*
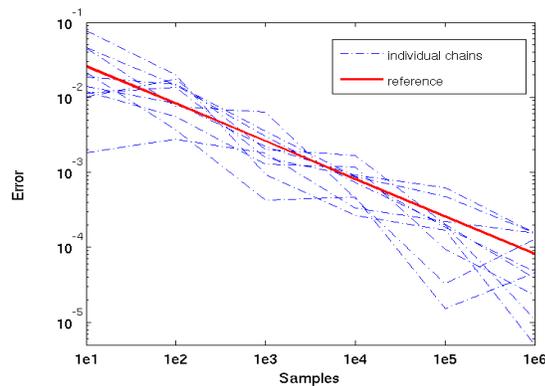
FIG. 7. *Convergence test of rMAP samples against a quadrature evaluated expectation value. Blue dashed lines show the errors of the 10 rMAP chains, each containing one million samples. They align well with the the solid red line, which represents the theoretical, $n^{-\frac{1}{2}}$ convergence rate from the central limit theorem. (See online version for color.)*

computation of the MAP estimator, which is crucial to the rMAP algorithm. In subsection 6.2.2, we analyze the sampling results using the rMAP algorithm. Through a comparison of the different optimization settings described above, we demonstrate

the efficiency achieved by using the TRINCG solver and a good initial guess. In addition, the rMAP samples are compared with DRAM samples, where we observe that Metropolis-adjusted rMAP samples provide statistical estimates with similar quality compared to those obtained from DRAM, while requiring much less computation.

6.2.1. Computation of the gradient and Hessian-vector product. In this section, we briefly discuss how to efficiently compute the gradient and Hessian-vector product. Using the standard reduced space approach (see, e.g., [11]), one can show that the (reduced) gradient $\nabla \mathcal{J}\left(u; \hat{u}, \hat{\mathbf{d}}\right)$ acting in any direction $\tilde{u}$ is given by

$$\left\langle \nabla \mathcal{J}\left(u; \hat{u}, \hat{\mathbf{d}}\right), \tilde{u}\right\rangle = -2 \int_{\Omega} \tilde{u} e^{2u} w \tau \, d\Omega,$$

where the adjoint state τ satisfies the adjoint equation

$$(39a) \qquad -\nabla^2 \tau - e^{2u} \tau = -\frac{1}{\sigma^2} \sum_{j=1}^{K} (w - d_j) \delta(\mathbf{x} - \mathbf{x}_j) \ \text{ in } \Omega,$$

$$(39b) \qquad \frac{\partial \tau}{\partial \mathbf{n}} = 0 \text{ on } \partial \Omega.$$

On the other hand, the Hessian acting in directions $\tilde{u}$ and $\widetilde{\tilde{u}}$ reads

$$\left\langle \left\langle \nabla^2 \mathcal{J}\left(u; \hat{u}, \hat{\mathbf{d}}\right), \tilde{u}\right\rangle, \widetilde{\tilde{u}}\right\rangle = -4 \int_{\Omega} \tilde{u} \, \widetilde{\tilde{u}} \, e^{2u} w \tau \, d\Omega - 2 \int_{\Omega} \tilde{u} e^{2u} \tilde{w} \tau \, d\Omega - 2 \int_{\Omega} \tilde{u} e^{2u} w \tilde{\tau} \, d\Omega,$$

where the incremental forward state $\tilde{w}$ obeys the incremental forward equation

$$(40a) \qquad -\nabla^2 \tilde{w} - e^{2u} \tilde{w} = 2 \, \widetilde{\tilde{u}} \, e^{2u} w \ \text{ in } \Omega,$$

$$(40b) \qquad \frac{\partial \tilde{w}}{\partial \mathbf{n}} = 0 \text{ on } \partial \Omega,$$

and the incremental adjoint state $\tilde{\tau}$ obeys the incremental adjoint equation

$$(41a) \qquad -\nabla^2 \tilde{\tau} - e^{2u} \tilde{\tau} = 2 \, \widetilde{\tilde{u}} \, e^{2u} \tau - \frac{1}{\sigma^2} \sum_{j=1}^{K} \tilde{w} \delta(\mathbf{x} - \mathbf{x}_j) \ \text{ in } \Omega,$$

$$(41b) \qquad \frac{\partial \tilde{\tau}}{\partial \mathbf{n}} = 0 \text{ on } \partial \Omega.$$

We shall compare our TRNCG optimization solver with the popular LM approach (see, e.g., [45, 49]). For that reason, we need to compute the Gauss–Newton Hessian-vector product. It can be shown that the Gauss–Newton Hessian acting in directions $\tilde{u}$ and $\widetilde{\tilde{u}}$ reads

$$\left\langle \left\langle \nabla^2 \mathcal{J}_{GN}\left(u; \hat{u}, \hat{\mathbf{d}}\right), \tilde{u}\right\rangle, \widetilde{\tilde{u}}\right\rangle = -2 \int_{\Omega} \tilde{u} e^{2u} w \tilde{\tau} \, d\Omega,$$

where the incremental forward state $\tilde{w}$ still satisfies (40), but the incremental adjoint state $\tilde{\tau}$ now obeys the following incremental adjoint equation:

$$(42a) \qquad -\nabla^2 \tilde{\tau} - e^{2u} \tilde{\tau} = -\frac{1}{\sigma^2} \sum_{j=1}^{K} \tilde{w} \delta(\mathbf{x} - \mathbf{x}_j) \ \text{ in } \Omega,$$

$$(42b) \qquad \frac{\partial \tilde{\tau}}{\partial \mathbf{n}} = 0 \text{ on } \partial \Omega.$$

6.2.2. Sampling results. Now we show the application of rMAP methods to quantify the uncertainty for the inverse problem governed by the above Helmholtz forward model. We create two experiments to compare and test the methods described above. Finite element discretization of the prior results in a parameter field with 94 parameters for both experiments. Since the experiments aim at testing algorithms rather than demonstrating Bayesian modeling, we conveniently fix the noise level for both experiments to be 1%. On the other hand, we use different prior coefficients α to control the "easiness" of sampling. We choose $\alpha = 8.0$ for the first experiment and $\alpha = 3.0$ for the second experiment—these numbers are chosen after trials to clearly represent two situations: a prior dominant case and a likelihood dominant case. When the model is prior dominant, the inverse problem resembles a linear inverse problem for which, following Lemma 3.4, rMAP should provide exact posterior samples. On the other hand, for the likelihood dominant case, due to the nonlinearity of the forward model, the rMAP samples are no longer exact posterior ones, and Metropolization becomes necessary.

For each of these two experiments, we draw a sample from the prior distribution and solve the forward equation (3) to generate a set of synthetic data as shown in Figure 8. Then we sample the Bayesian model with the following four variants of the rMAP method: TRINCG and LM, with/without good initial guesses. As a comparison, we also sample the model with a DRAM sampler of five million samples, which we consider large enough to be convergent.


(a) Synthetic u ($\alpha = 8.0$).   (b) Synthetic u ($\alpha = 3.0$).
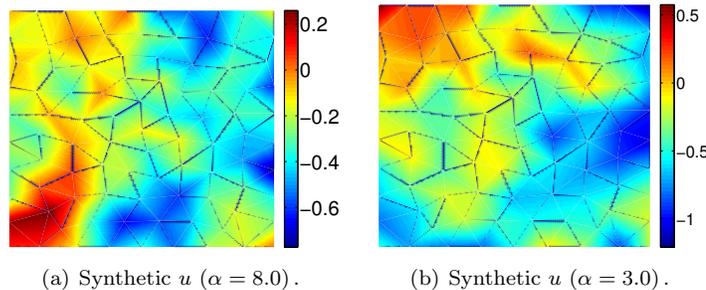
FIG. 8. *Synthetic parameter u for two numerical experiments.*

Since rMAP samples are not exact posterior samples for nonlinear problems, it is not necessary to demand high accuracy (and hence high cost) in each optimization solution. Yet, we still hope that for these loosely approximate rMAP samples, the proposed Metropolization can effectively correct them towards the posterior distribution. To that end, we set large tolerances: $\varepsilon_F = \varepsilon_X = \varepsilon_G = 10^{-4}$ for the first experiment and $\varepsilon_F = \varepsilon_X = \varepsilon_G = 10^{-6}$ for the second. For a similar reason, we limit the maximal allowed number of iterations to 150 and 200, respectively, to further control the computational costs.

For each experiment and each variant of the method, we compute 1000 rMAP samples. Within each experiment, we use the same randomly perturbed sequences $\{\hat{u}\}_{i=1}^{1000}$ and $\{\hat{\mathbf{d}}\}_{i=1}^{1000}$ for all four rMAP methods. Ideally, with this setting these methods should have produced exactly the same rMAP samples if each optimizer had converged. In practice, the acquired samples are different among these methods due to the tolerance and iteration control.

Figures 9 and 10 show the estimated conditional mean and variance for the high

prior and the low prior cases, respectively. In both cases, the plain rMAP samples have nonnegligible approximation errors. These errors are successfully corrected with a Metropolization using weights described in section 4. We point out that among the four variants of the rMAP methods, the one that uses TRINCG and good initial guesses shows optimal performance. Its statistical estimates are close to that of the DRAM sampler in both experiments. This indicates the fast convergence of the TRINCG method: even though we have relaxed the convergence criteria and limited the number of iterations, TRINCG has always been able to get close to the real optimizer rapidly. Our proposed method of computing initial guesses has further ensured its efficiency. For a closer look, in Figures 11 and 12 we show a comparison between sampling estimates of the DRAM and the rMAP using TRINCG with good initial guesses.

(a) mean (rMAP).

(b) mean (weighted-rMAP).

(c) Variance (rMAP).

(d) Variance (weighted-rMAP).
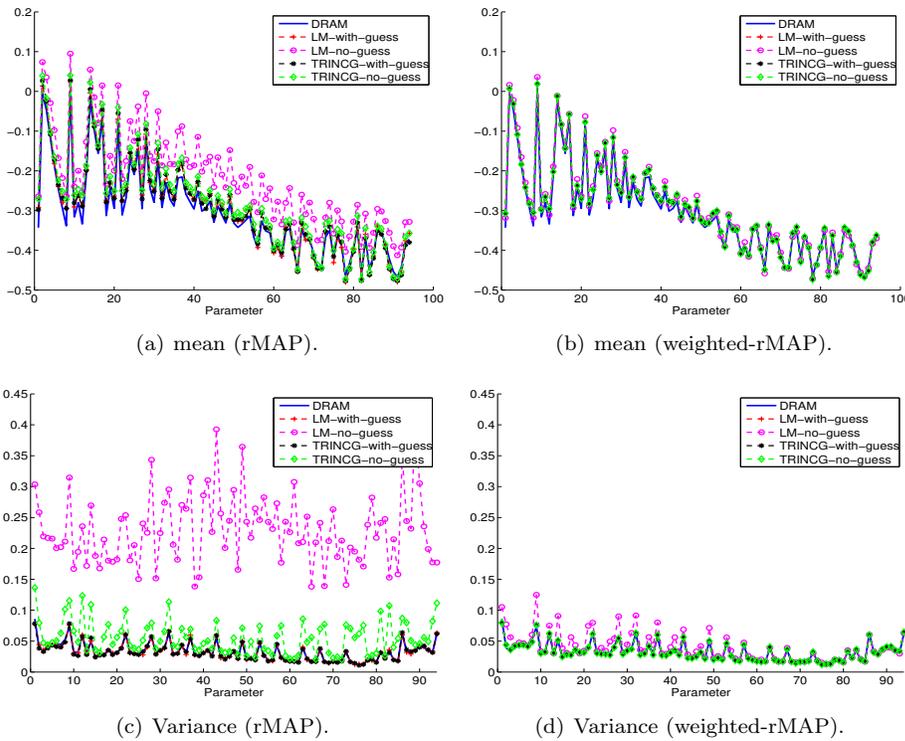
FIG. 9. *Case with* $\alpha = 8.0$*: Comparison of estimated statistics from all samplers. The top row shows the conditonal mean estimate between* (a) *DRAM and rMAP samples and* (b) *DRAM and Metropolized rMAP samples. The bottom row shows the corresponding comparison of variance estimates.*

Next, we analyze the computational efficiency of the rMAP samplers. Note that the DRAM samples are highly correlated due to the large dimensionality of parameter space; meanwhile, as we will show below, rMAP samples are almost statistically independent, even for nonlinear problems. In order to compare computational performance between rMAP and DRAM as well, we utilize a concept of effective sample size (ESS), which is defined, for a sampler with a total of n samples, as

$$\text{(43)} \qquad \text{ESS} = \frac{n}{\overline{\tau}},$$

(a) mean (rMAP).

(b) mean (weighted-rMAP).

(c) Variance (rMAP).

(d) Variance (weighted-rMAP).
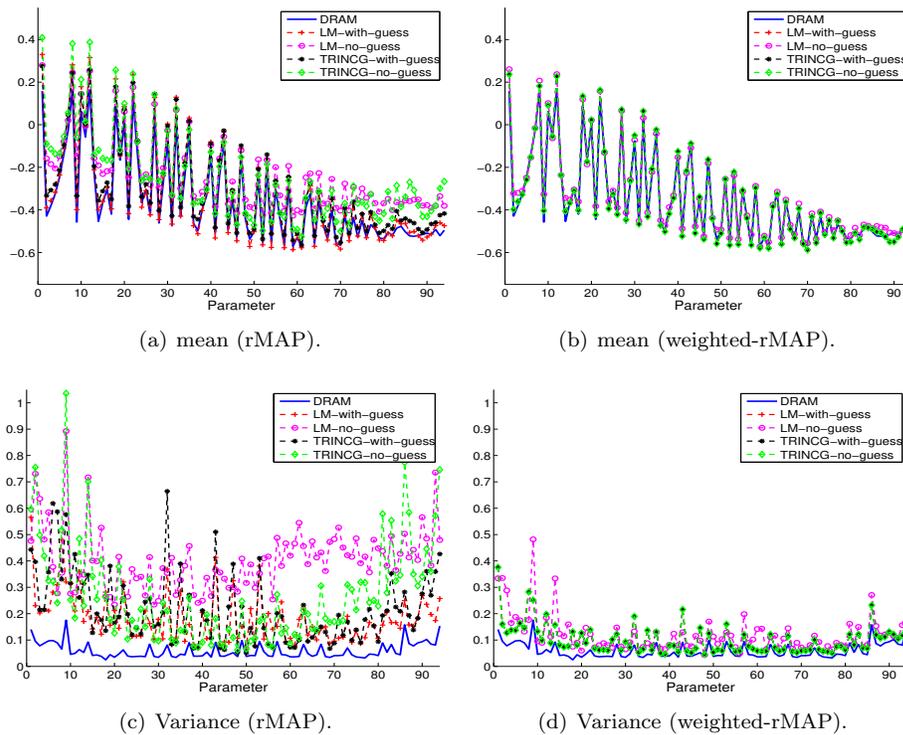
FIG. 10. *Case with $\alpha = 3.0$: Comparison of estimated statistics from all samplers. The top row shows the conditonal mean estimate between* (a) *DRAM and rMAP samples and* (b) *DRAM and Metropolized rMAP samples. The bottom row shows the corresponding comparison of variance estimates.*

and, for a model with L parameters in total, the averaged integrated auto-correlation time (IACT) $\overline{\tau}$ is computed by

$$\overline{\tau} = \frac{1}{L} \sum_{l=1}^{L} \left(1 + 2 \sum_{k=1}^{\infty} \tau(k) \right),$$

in which the auto-correlation function (ACF) $\tau(k)$ for a time series X_t with mean value μ and variance σ^2 is defined as

$$\tau(k) = \frac{E\left[(X_t - \mu)(X_{t+k} - \mu) \right]}{\sigma^2}.$$

Since PDE solve is the most time-consuming part, we take the total number of PDE solves (assuming the cost of solving forward, adjoint, incremental forward, and incremental adjoint equations is the same) as the measure of the computational cost.

Figure 13 shows the comparison of IACT for all parameters. For simplicity, we only show the IACT for rMAP samples obtained through the TRINCG together with good initial guesses. We then obtain the mean IACTs to be $\overline{\tau}_{\mathrm{DRAM}} = 461.90$, $\overline{\tau}_{\mathrm{rMAP}} = 1.00$, and $\overline{\tau}_{\mathrm{weighted\text{-}rMAP}} = 1.11$ for the first experiment, and $\overline{\tau}_{\mathrm{DRAM}} = 564.32$, $\overline{\tau}_{\mathrm{rMAP}} = 1.10$, and $\overline{\tau}_{\mathrm{weighted\text{-}rMAP}} = 1.2743$ for the second experiment. Therefore, 1000 rMAP samples correspond to about 415,000 DRAM samples when $\alpha = 8.0$ and correspond to about 443,000 DRAM samples when $\alpha = 3.0$. As a result, for

(a) DRAM.   (b) rMAP.   (c) Weighted-rMAP.

(d) DRAM.   (e) rMAP.   (f) Weighted-rMAP.
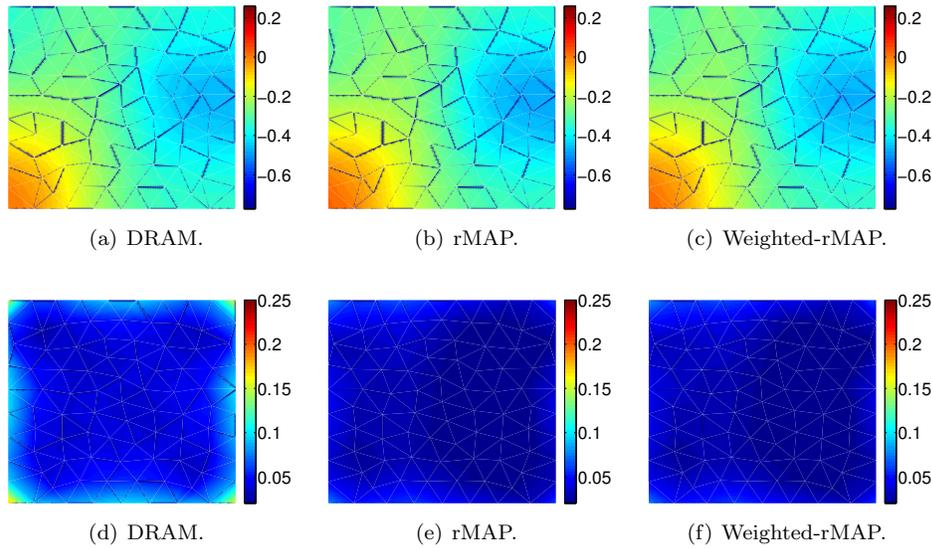
FIG. 11. *Comparison of estimates for the first experiment with $\alpha = 8.0$. The top row shows the conditional mean estimate. The bottom row shows comparison of variance estimation. All the rMAP samples are obtained from the TRINCG method with good initial guesses.*


(a) DRAM.   (b) rMAP.   (c) Weighted-rMAP.

(d) DRAM.   (e) rMAP.   (f) Weighted-rMAP.
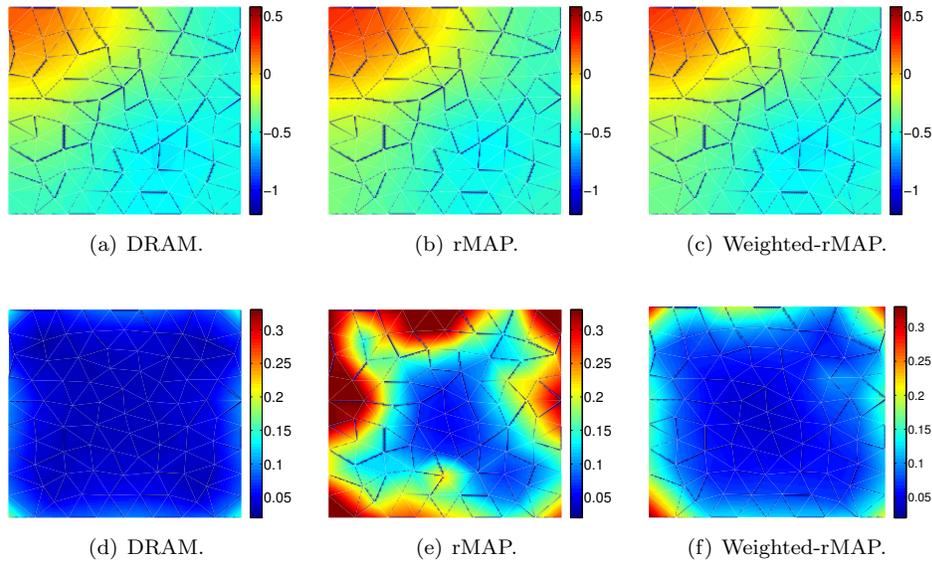
FIG. 12. *Comparison of estimates for the second experiment with $\alpha = 3.0$. The top row shows the conditional mean estimate. The bottom row shows comparison of variance estimation. All the rMAP samples are obtained from the TRINCG method with good initial guesses.*

comparing computational costs in both experiments, we take into consideration 1000 rMAP samples and 400,000 DRAM samples.

We compare costs of different sampling/optimization strategies in Tables 1 and

(a) $\alpha = 8.0$.   (b) $\alpha = 3.0$.
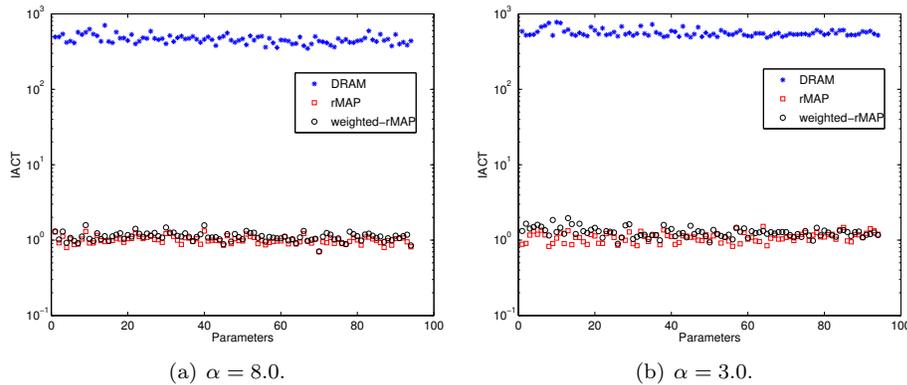
FIG. 13. *IACT for DRAM, rMAP, and weighted-rMAP for all parameters. Samples from TRINCG with good initial guesses are used to compute these IACTs.*

2. It is obvious that, compared with the LM method, TRINCG improves in efficiency both with and without a warm-start; for example, when good initial guesses are adopted, LM is about 60% and about 320% more expensive than TRINCG, respectively. The importance of the warm-start strategy is also evident from these tables. In particular, good initial guesses speed up the LM algorithm significantly (at least fivefold) in the prior-dominated case such that the rMAP sampler with LM performs better than DRAM with a statistically comparable number of samples. Nonetheless, the computational costs of the LM method in Table 2 are higher than those of the corresponding DRAM sampler even with good initial guesses, leaving TRINCG as the only tractable choice for rMAP sampling of this "difficult" likelihood-dominated problem.

TABLE 1

Cost for the case $\alpha = 8.0$. The cost is measured in the number of PDE solves in generating 1000 rMAP samples using four combinations: with either TRINCG or LM and with or without warm-start strategy. As a comparison, the cost for the DRAM sampler to get 400,000 samples is shown in the last row.

rMAP	Good initial guess	Random initial guess
TRINCG	208367	288254
LM	334618	1626279
DRAM	732917	

TABLE 2

Cost for the case $\alpha = 3.0$. The cost measured in the number of PDE solves in generating 1000 rMAP samples using four combinations: with either TRINCG or LM and with or without warm-start strategy. As a comparison, the cost for the DRAM sampler to get 400,000 samples is shown in the last row.

rMAP	Good initial guess	Random initial guess
TRINCG	511956	568671
LM	1639601	2705973
DRAM	706017	

7. Conclusions. In this paper we present a randomized maximum a posteriori (rMAP) approach to approximately sample posteriors of nonlinear Bayesian inverse problems in high dimensional parameter spaces. The idea is to cast the standard MAP computation as a stochastic optimization problem and use the sample average approach to approximate the expectation. We have shown that the randomized maximum likelihood method is a special case of the proposed rMAP method. The stochastic programming viewpoint allows us to provide additional theoretical results in both finite and infinite dimensions and for both linear and nonlinear inverse problems, leading to a better understanding of rMAP. The appeal of the proposed approach is that each rMAP sample requires solution of a PDE-constrained optimization problem which can be carried out efficiently using a trust region inexact Newton conjugate gradient (TRINCG) method. To further reduce the cost of each rMAP sample, we develop a warm-start strategy using sensitivity analysis via an efficient adjoint technique. Viewing rMAP as an iterative SN method reveals that rMAP is in fact a move away from the inefficiencies of random-walk/diffusion processes toward powerful optimization methods that use derivative information to traverse the posterior.

We have established a connection between the rMAP approach and a closely related randomize-then-optimize method. We show that they are identical for linear inverse problems but different for nonlinear ones. Since rMAP samples are approximate samples of the posterior, we present an approximate Metropolization to reduce the bias. We have also discussed finite element method (FEM) discretization of the infinite dimensional Bayesian inverse problem, solving optimization problems at each sampling step with the TRINCG method, and a sensitivity analysis–based warm-start strategy. Analytical and numerical experiments are presented to confirm various theoretical results and demonstrate the potential of the rMAP approach for difficult nonlinear Bayesian inverse problems.

REFERENCES

[1] J. M. BARDSLEY, A. SOLONEN, H. HAARIO, AND M. LAINE, *Randomize-then-optimize: A method for sampling from posterior distributions in nonlinear inverse problems*, SIAM J. Sci. Comput., 36 (2014), pp. A1895–A1910, https://doi.org/10.1137/140964023.

[2] A. BESKOS, F. J. PINSKI, J. M. SANZ-SERNA, AND A. M. STUART, *Hybrid Monte Carlo on Hilbert spaces*, Stochastic Process. Appl., 121 (2011), pp. 2201–2230.

[3] L. T. BIEGLER, O. GHATTAS, M. HEINKENSCHLOSS, AND B. VAN BLOEMEN WAANDERS, EDS., *Large-Scale PDE-Constrained Optimization*, Lecture Notes in Comput. Sci. Engrg. 30, Springer-Verlag, Heidelberg, 2003.

[4] A. BORZÌ AND V. SCHULZ, *Computational Optimization of Systems Governed by Partial Differential Equations*, Comput. Sci. Engrg. 8, SIAM, Philadelphia, 2012, https://doi.org/10.1137/1.9781611972054.

[5] M. A. BRANCH, T. F. COLEMAN, AND Y. LI, *A subspace, interior, and conjugate gradient method for large-scale bound-constrained minimization problems*, SIAM J. Sci. Comput., 21 (1999), pp. 1–23, https://doi.org/10.1137/S1064827595289108.

[6] A. BROCKWELL, *Parallel Markov chain Monte Carlo simulation by pre-fetching*, J. Comput. Graph. Statist., 15 (2006), pp. 246–261.

[7] T. BUI-THANH, *Model-Constrained Optimization Methods for Reduction of Parameterized Large-Scale Systems*, Ph.D. thesis, Department of Aeronautics and Astronautics, MIT, Cambridge, MA, 2007.

[8] T. BUI-THANH AND Q. P. NGUYEN, *FEM-based discretization-invariant MCMC methods for PDE-constrained Bayesian inverse problems*, Inverse Probl. Imaging, 10 (2016), pp. 943–975.

[9] T. BUI-THANH, C. BURSTEDDE, O. GHATTAS, J. MARTIN, G. STADLER, AND L. C. WILCOX, *Extreme-scale UQ for Bayesian inverse problems governed by PDEs*, in SC12: Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis, IEEE Computer Society, Washington, DC, 2012, pp. 1–11.

[10] T. BUI-THANH AND O. GHATTAS, *Analysis of the Hessian for inverse scattering problems. Part I: Inverse shape scattering of acoustic waves*, Inverse Problems, 28 (2012), 055001.

[11] T. BUI-THANH AND O. GHATTAS, *Analysis of the Hessian for inverse scattering problems. Part II: Inverse medium scattering of acoustic waves*, Inverse Problems, 28 (2012), 055002.

[12] T. BUI-THANH AND O. GHATTAS, *Analysis of the Hessian for inverse scattering problems. Part III: Inverse medium scattering of electromagnetic waves in three dimensions,* Inverse Problems Imaging, 7 (2013), pp. 1139–1155.

[13] T. BUI-THANH AND O. GHATTAS, *A Scaled Stochastic Newton Algorithm for Markov Chain Monte Carlo Simulations*, unpublished.

[14] T. BUI-THANH, O. GHATTAS, J. MARTIN, AND G. STADLER, *A computational framework for infinite-dimensional Bayesian inverse problems. Part I: The linearized case, with application to global seismic inversion*, SIAM J. Sci. Comput., 35 (2013), pp. A2494–A2523, https://doi.org/10.1137/12089586X.

[15] T. BUI-THANH AND M. A. GIROLAMI, *Solving large-scale PDE-constrained Bayesian inverse problems with Riemann manifold Hamiltonian Monte Carlo*, Inverse Problems, 30 (2014), 114014.

[16] J. BYRD, *Parallel Markov Chain Monte Carlo*, Ph.D. thesis, University of Warwick, Coventry, UK, 2010.

[17] D. CALVETTI AND E. SOMERSALO, *Introduction to Bayesian Scientific Computing: Ten Lectures on Subjective Computing*, Springer, New York, 2007.

[18] P. G. CIARLET, *The Finite Element Method for Elliptic Problems*, North–Holland, Amsterdam, New York, 1978. Reprinted as Classics Appl. Math. 40, SIAM, Philadelphia, 2002, https://doi.org/10.1137/1.9780898719208.

[19] T. F. COLEMAN AND Y. LI, *An interior trust region approach for nonlinear minimization subject to bounds*, SIAM J. Optim., 6 (1996), pp. 418–445, https://doi.org/10.1137/0806023.

[20] S. L. COTTER, G. O. ROBERTS, A. M. STUART, AND D. WHITE, *MCMC methods for functions: Modifying old algorithms to make them faster*, Statist. Sci., 28 (2013), pp. 424–446.

[21] T. CUI, K. J. LAW, AND Y. M. MARZOUK, *Dimension-independent likelihood-informed MCMC*, J. Comput. Phys., 304 (2016), pp. 109–137.

[22] T. CUI, J. MARTIN, Y. M. MARZOUK, A. SOLONEN, AND A. SPANTINI, *Likelihood-Informed Dimension Reduction for Nonlinear Inverse Problems*, preprint, https://arxiv.org/abs/1403.4680, 2014.

[23] M. DASHTI, K. LAW, A. STUART, AND J. VOSS, *MAP estimators and their consistency in Bayesian nonparametric inverse problems*, Inverse Problems, 29 (2013), 095017.

[24] J. C. DE LOS REYES, *Numerical PDE-Constrained Optimization*, Springer, New York, 2015.

[25] S. DUANE, A. D. KENNEDY, B. PENDLETON, AND D. ROWETH, *Hybrid Monte Carlo*, Phys. Lett. B, 195 (1987), pp. 216–222.

[26] J. N. FRANKLIN, *Well-posed stochastic extensions of ill–posed linear problems*, J. Math. Anal. Appl., 31 (1970), pp. 682–716.

[27] M. GIROLAMI AND B. CALDERHEAD, *Riemann manifold Langevin and Hamiltonian Monte Carlo methods*, J. Roy. Statist. Soc. Ser. B Statist. Methodol., 73 (2011), pp. 123–214.

[28] G. H. GOLUB AND C. F. VAN LOAN, *Matrix Computations*, 3rd ed., Johns Hopkins Stud. Math. Sci., Johns Hopkins University Press, Baltimore, MD, 1996.

[29] H. HAARIO, M. LAINE, A. MIRAVETE, AND E. SAKSMAN, *DRAM: Efficient adaptive MCMC*, Statist. Comput., 16 (2006), pp. 339–354.

[30] M. HAIRER, *An Introduction to Stochastic PDEs*, Lecture Notes, The University of Warwick/Courant Institute, 2009; available online at http://www.hairer.org/notes/SPDEs.pdf or https://arxiv.org/pdf/0907.4178.pdf.

[31] N. HALKO, P.-G. MARTINSSON, AND J. A. TROPP, *Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions*, SIAM Rev., 53 (2011), pp. 217–288, https://doi.org/10.1137/090771806.

[32] W. K. HASTINGS, *Monte Carlo sampling methods using Markov chains and their applications*, Biometrika, 57 (1970), pp. 97–109.

[33] M. HINZE, R. PINNAU, M. ULBRICH, AND S. ULBRICH, *Optimization with PDE Constraints*, Springer, New York, 2009.

[34] M. A. IGLESIAS, K. J. H. LAW, AND A. M. STUART, *Evaluation of Gaussian approximations for data assimilation in reservoir models*, Comput. Geosci., 17 (2013), pp. 851–885.

[35] M. ILIĆ, F. LIU, I. TURNER, AND V. ANH, *Numerical approximation of a fractional-in-space diffusion equation*, Frac. Calc. App. Anal., 8 (2005), pp. 323–341.

[36] T. ISAAC, N. PETRA, G. STADLER, AND O. GHATTAS, *Scalable and efficient algorithms for the propagation of uncertainty from data through inference to prediction for large-scale problems, with application to flow of the Antarctic ice sheet*, J. Comput. Phys., 296 (2015), pp. 348–368.

[37] J. Kaipio and E. Somersalo, *Statistical and Computational Inverse Problems*, Appl. Math. Sci. 120, Springer-Verlag, New York, 2005.

[38] P. K. Kitanidis, *Quasi-linear geostatistical theory for inversing*, Water Resour. Res., 31 (1995), pp. 2411–2419.

[39] P. K. Kitanidis, *On the geostatistical approach to the inverse problem*, Adv. Water Res., 19 (1996), pp. 333–342.

[40] S. Lasanen, *Discretizations of Generalized Random Variables with Applications to Inverse Problems*, Ph.D. thesis, University of Oulu, Oulu, Finland, 2002.

[41] M. S. Lehtinen, L. Päivärinta, and E. Somersalo, *Linear inverse problems for generalized random variables*, Inverse Problems, 5 (1989), pp. 599–612.

[42] J. Martin, L. C. Wilcox, C. Burstedde, and O. Ghattas, *A stochastic Newton MCMC method for large-scale statistical inverse problems with application to seismic inversion*, SIAM J. Sci. Comput., 34 (2012), pp. A1460–A1487, https://doi.org/10.1137/110845598.

[43] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, *Equation of state calculations by fast computing machines*, J. Chem. Phys., 21 (1953), pp. 1087–1092.

[44] R. M. Neal, *MCMC using Hamiltonian dynamics*, in Handbook of Markov Chain Monte Carlo, Chapman & Hall/CRC Press, Boca Raton, FL, 2011, pp. 113–162.

[45] J. Nocedal and S. J. Wright, *Numerical Optimization*, 2nd ed., Springer-Verlag, Berlin, 2006.

[46] D. S. Oliver, *Minimization for conditional simulation: Relationship to optimal transport*, J. Comput. Phys., 265 (2014), pp. 1–15.

[47] D. S. Oliver, *Metropolized Randomized Maximum Likelihood for Sampling from Multimodal Distributions*, preprint, https://arxiv.org/abs/1507.08563, 2016.

[48] D. S. Oliver, H. He, and A. C. Reynolds, *Conditioning permeability fields to pressure data*, in Proceedings of the European Conference for the Mathematics of Oil Recovery, 1996, pp. 1–11.

[49] D. S. Oliver, A. C. Reynolds, and N. Liu, *Inverse Theory for Petroleum Reservoir Characterization and History matching*, Cambridge University Press, Cambridge, UK, 2008.

[50] N. Petra, J. Martin, G. Stadler, and O. Ghattas, *A computational framework for infinite-dimensional Bayesian inverse problems, Part II: Stochastic Newton MCMC with application to ice sheet flow inverse problems*, SIAM J. Sci. Comput., 36 (2014), pp. A1525–A1555, https://doi.org/10.1137/130934805.

[51] P. Piiroinen, *Statistical Measurements, Experiments, and Applications*, Ph.D. thesis, Department of Mathematics and Statistics, University of Helsinki, Helsinki, Finland, 2005.

[52] G. D. Prato and J. Zabczyk, *Stochastic Equations in Infinite Dimensions*, Cambridge University Press, Cambridge, UK, 1992.

[53] C. P. Robert and G. Casella, *Monte Carlo Statistical Methods,* Springer Texts in Statist., Springer-Verlag, New York, 2005.

[54] G. O. Roberts and J. S. Rosenthal, *Optimal scaling of discrete approximations to Langevin diffusions*, J. R. Statist. Soc. B, 60 (1997), pp. 255–268.

[55] G. O. Roberts and R. L. Tweedie, *Exponential convergence of Langevin distributions and their discrete approximations*, Bernoulli, 2 (1996), pp. 341–363.

[56] R. T. Rockafellar and R. J.-B. Wets, *Variational Analysis*, Springer-Verlag, Berlin, 1998.

[57] A. Shapiro, D. Dentcheva, and A. Ruszczyński, *Lectures on Stochastic Programming: Modeling and Theory*, MOS-SIAM Ser. Optim. 9, SIAM, Philadelphia, 2009, https://doi.org/10.1137/1.9780898718751.

[58] I. Strid, *Efficient parallelisation of Metropolis–Hastings algorithms using a prefetching approach*, Comput. Statist. Data Anal., 54 (2010), pp. 2814–2835.

[59] A. M. Stuart, *Inverse problems: A Bayesian perspective*, Acta Numer., 19 (2010), pp. 451–559.

[60] D. A. Van Dyk and X.-L. Meng, *The art of data augmentation*, J. Comput. Graph. Statist., 10 (2001), pp. 1–50.

[61] K. Wang, *Parallel Markov Chain Monte Carlo Methods for Large Scale Statistical Inverse Problems*, Ph.D. thesis, Texas A&M University, College Station, TX, 2014.

[62] D. J. Wilkinson, *Handbook of Parallel Computing and Statistics*, Marcel Dekker/CRC Press, New York, 2005, pp. 481–512.