SIAM J. SCI. COMPUT.
Vol. 34, No. 3, pp. A1460–A1487

A STOCHASTIC NEWTON MCMC METHOD FOR LARGE-SCALE STATISTICAL INVERSE PROBLEMS WITH APPLICATION TO SEISMIC INVERSION*

JAMES MARTIN†, LUCAS C. WILCOX‡, CARSTEN BURSTEDDE§, AND OMAR GHATTAS¶

Abstract. We address the solution of large-scale statistical inverse problems in the framework of Bayesian inference. The Markov chain Monte Carlo (MCMC) method is the most popular approach for sampling the posterior probability distribution that describes the solution of the statistical inverse problem. MCMC methods face two central difficulties when applied to large-scale inverse problems: first, the forward models (typically in the form of partial differential equations) that map uncertain parameters to observable quantities make the evaluation of the probability density at any point in parameter space very expensive; and second, the high-dimensional parameter spaces that arise upon discretization of infinite-dimensional parameter fields make the exploration of the probability density function prohibitive. The challenge for MCMC methods is to construct proposal functions that simultaneously provide a good approximation of the target density while being inexpensive to manipulate. Here we present a so-called Stochastic Newton method in which MCMC is accelerated by constructing and sampling from a proposal density that builds a local Gaussian approximation based on local gradient and Hessian (of the log posterior) information. Thus, the method exploits tools (adjoint-based gradients and Hessians) that have been instrumental for fast (often mesh-independent) solution of deterministic inverse problems. Hessian manipulations (inverse, square root) are made tractable by a low-rank approximation that exploits the compact nature of the data misfit operator. This is analogous to a reduced model of the parameter-to-observable map. The method is applied to the Bayesian solution of an inverse medium problem governed by 1D seismic wave propagation. We compare the Stochastic Newton method with a reference black box MCMC method as well as a gradient-based Langevin MCMC method, and observe at least two orders of magnitude improvement in convergence for problems with up to 65 parameters. Numerical evidence suggests that a 1025 parameter problem converges at the same rate as the 65 parameter problem.

Key words. MCMC, Stochastic Newton, inverse problems, uncertainty quantification, Langevin dynamics, low-rank Hessian

AMS subject classifications. 35Q62, 35Q93, 35Q86, 49M15, 74J20, 74J25, 65C40, 65C60, 65M32

DOI. 10.1137/110845598

1. Introduction and background.

Uncertainty in reconstructing parameter fields from data is a fundamental feature of ill-posed inverse problems. Our lack of

*Submitted to the journal's Methods and Algorithms for Scientific Computing section August 24, 2011; accepted for publication (in revised form) February 13, 2012; published electronically June 5, 2012. This research was supported by AFOSR grant FA9550-09-1-0608, NSF grants DMS-0724746, ARC-0941678, and CMMI-1028889, DOE grants DE-FG02-08ER25860, DE-SC0002710, and DE-FC52-08NA28615, and NSF TeraGrid award TG-MCA04N026.

http://www.siam.org/journals/sisc/34-3/84559.html

†Corresponding author. Institute for Computational Engineering & Sciences, and Computational Science, Engineering, & Mathematics graduate program, The University of Texas at Austin, Austin, TX 78712 (jmartin@ices.utexas.edu). This author's research was supported by the DOE CSGF under grant DE-FG02-97ER25308.

‡Institute for Computational Engineering & Sciences, The University of Texas at Austin, Austin, TX 78712. Address beginning September 2012: Naval Postgraduate School, Department of Applied Mathematics, Monterey, CA 93943 (me@lucaswilcox.com).

§Institute for Computational Engineering & Sciences, The University of Texas at Austin, Austin, TX 78712. Current address: Institut für Numerische Simulation (INS), Rheinische Friedrich-Wilhelms-Universität Bonn, 53115 Bonn, Germany (carsten@ices.utexas.edu).

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

knowledge results from noisy measurements, sparse observations, uncertain forward models, and uncertain prior parameter information. The deterministic output least squares approach to inverse problems, which amounts to minimizing a regularized data misfit function, is incapable of accounting for uncertainties in the solution of the inverse problem. *Bayesian inference* provides a systematic framework for incorporating uncertainties in observations, forward models, and prior knowledge to quantify uncertainties in the model parameters. However, Bayesian solution of large-scale statistical inverse problems, i.e., those described by expensive forward models such as partial differential equations (PDEs), and for large numbers of model parameters that result from discretized parameter fields, is essentially intractable using conventional statistical techniques that view the forward model (i.e., the parameter-to-observable map) as a black box.

We address methods for sampling probability density functions (pdfs) that describe uncertain parameter fields in Bayesian solutions to statistical inverse problems governed by PDEs. Such problems have two properties that present significant challenges for standard Markov chain Monte Carlo (MCMC) sampling methods. First, each sample point requires solution of the forward problem, which can be exceedingly expensive. Second, discretization of the parameter space can result in very high dimensional pdfs. Here, we present a method that exploits the structure of the inverse operator to greatly speed up MCMC. The method, which we refer to as *Stochastic Newton*, can be derived by analogy with the classical Newton method for the associated deterministic inverse problem. Stochastic Newton employs a local Gaussian approximation to the target pdf—informed by local Hessian information—as a proposal density for MCMC. A low-rank approximation of the Hessian is invoked—reflecting the ill-posed nature of many PDE-based inverse problems—rendering the computation tractable. Alternatively, Stochastic Newton can be interpreted as a Hessian-preconditioned Langevin MCMC method. In the remainder of this section, we provide background on the Bayesian formulation of statistical inverse problems and on MCMC methods, and discuss alternative approaches.

1.1. Bayesian formulation of the statistical inverse problem. The great challenge in solving inverse problems lies in the fact that they are usually ill-posed: many different choices of model parameters may be consistent with the data. Nonuniqueness stems from sparsity of the observations and uncertainty in both the measurements and the model itself. A popular approach to obtaining a unique "solution" to the inverse problem is to formulate it as a least squares optimization problem: minimize the misfit between observed and predicted outputs in an appropriate norm while also minimizing a *regularization* term that penalizes unwanted features of the parameters. This is often called *Occam's approach*: find the "simplest" set of parameters that is consistent with the measured data. The inverse problem thus leads to a nonlinear optimization problem that is constrained by the forward model. Estimation of parameters using this regularization approach to inverse problems will yield an estimate of the "best" parameter values that simultaneously fit the data and honor the regularization penalty term. However, we are interested in not just point estimates of the best-fit parameters but a complete statistical description of the parameter values. The *Bayesian* approach does this by reformulating the inverse problem as a problem in *statistical inference*, incorporating uncertainties in the measurements, the forward model, and prior information on the parameters [27, 43]. The solution of this inverse problem is the *posterior* joint probability density of the parameters, which encodes the degree of confidence in their estimate. Thus we are able to quantify the resulting

uncertainty in the parameters, taking into account uncertainties in the data, model, and prior information.

Suppose the relationship between output observables $\boldsymbol{d}$ (the predicted outputs at the measurement locations and time instants) and uncertain model parameters $\boldsymbol{m}$ is denoted by $\boldsymbol{d} = f(\boldsymbol{m}, \boldsymbol{e})$, where $\boldsymbol{e}$ represents noise due to measurement and/or modeling errors. In other words, given the model parameters $\boldsymbol{m}$ and noise $\boldsymbol{e}$, the function $f(\boldsymbol{m}, \boldsymbol{e})$ solves the forward (PDE) problem to yield $\boldsymbol{d}$. Suppose also that we have the prior probability density $\pi_{\text{prior}}(\boldsymbol{m})$, which encodes the confidence we have in prior information on the unknown model parameters (i.e., independent of present observations), and the likelihood function $\pi_{\text{like}}(\boldsymbol{d}_{\text{obs}}|\boldsymbol{m})$, which describes the conditional probability that the model parameters $\boldsymbol{m}$ give rise to the actual measurements $\boldsymbol{d}_{\text{obs}}$. Then Bayes's theorem of inverse problems expresses the posterior probability density of the model parameters, π_{post}, given the data $\boldsymbol{d}_{\text{obs}}$, as the conditional probability

$$\pi_{\text{post}}(\boldsymbol{m}) := \pi(\boldsymbol{m}|\boldsymbol{d}_{\text{obs}}) \propto \pi_{\text{prior}}(\boldsymbol{m})\, \pi(\boldsymbol{d}_{\text{obs}}|\boldsymbol{m}). \tag{1.1}$$

Expression (1.1) provides the statistical solution of the inverse problem as a probability density for the model parameters $\boldsymbol{m}$. Often, particularly in high dimensions, we are interested not in a complete characterization of $\pi_{\text{post}}(\boldsymbol{m})$ (which may be intractable to compute and impossible to interpret) but in its moments (mean, covariance, etc.) or other functionals (e.g., event probabilities).

As a specific example, suppose the noise is additive and is modeled as Gaussian with zero mean and a covariance matrix $\boldsymbol{\Gamma}_{\text{noise}}$, and suppose the prior density of the model parameters is represented as Gaussian with $\bar{\boldsymbol{m}}_{\text{prior}}$ as the mean and $\boldsymbol{\Gamma}_{\text{prior}}$ as the covariance matrix; then the posterior probability density of the model parameters is given explicitly (within a normalizing constant) by

$$\pi_{\text{post}}(\boldsymbol{m}) \propto \exp\left[-\tfrac{1}{2}\|f(\boldsymbol{m}) - \boldsymbol{d}_{\text{obs}}\|^2_{\boldsymbol{\Gamma}^{-1}_{\text{noise}}} - \tfrac{1}{2}\|\boldsymbol{m} - \bar{\boldsymbol{m}}_{\text{prior}}\|^2_{\boldsymbol{\Gamma}^{-1}_{\text{prior}}}\right]. \tag{1.2}$$

This latter expression shows that even when the prior, measurement, and modeling uncertainties are Gaussian, the posterior density of the model parameters is generally not Gaussian, due to the nonlinearity of the parameter-to-observable map, $f(\boldsymbol{m})$. However, this expression exposes a significant connection between statistical and deterministic inversion. Suppose we wish to find the value of the most likely model parameters by maximizing the posterior density (1.2). This is equivalent to minimizing the negative argument of the exponential function—which is precisely the misfit function that is minimized by deterministic inverse methods, provided we interpret the prior as a regularization and weigh the data misfit by the inverse noise covariance. Moreover, it is straightforward to show that the inverse of the Hessian matrix of the deterministic regularized misfit function approximates the covariance matrix of the posterior density (the equivalence is exact when $f(\boldsymbol{m})$ is linear). This connection between the Hessian operator of the deterministic inverse problem and the inverse covariance matrix of the statistical inverse problem is crucial to the computational efficiency of the Stochastic Newton method.

While it is easy to write expressions for the posterior pdf such as (1.1) or (1.2), making use of these expressions poses a challenge, because the posterior pdf is a surface in high dimensions (equal to the number of model parameters $\boldsymbol{m}$), and because the solution of the forward problem (i.e., computing $f(\boldsymbol{m})$ given $\boldsymbol{m}$) is required to evaluate the probability of any point in parameter space (as can be seen in (1.2)). Straightforward grid-based sampling is limited to problems with a few parameters and cheap

ALGORITHM 1. Metropolis–Hastings Algorithm to sample pdf π.

Choose initial parameters $\boldsymbol{m}_0$
Compute $\pi(\boldsymbol{m}_0)$
for $k = 0, \ldots, N-1$ **do**
  Draw sample $\boldsymbol{y}$ from the proposal density $q(\boldsymbol{m}_k, \cdot)$
  Compute $\pi(\boldsymbol{y})$
  Compute $\alpha(\boldsymbol{m}_k, \boldsymbol{y}) = \min\left\{1, \frac{\pi(\boldsymbol{y})q(\boldsymbol{y},\boldsymbol{m}_k)}{\pi(\boldsymbol{m}_k)q(\boldsymbol{m}_k,\boldsymbol{y})}\right\}$
  Draw $u \sim \mathcal{U}([0,1))$
  if $u < \alpha(\boldsymbol{m}_k, \boldsymbol{y})$ **then**
    Accept: Set $\boldsymbol{m}_{k+1} = \boldsymbol{y}$
  else
    Reject: Set $\boldsymbol{m}_{k+1} = \boldsymbol{m}_k$
  end if
end for

forward simulations. Special sampling techniques, such as MCMC methods, have been developed to generate sample ensembles that typically require many fewer points than grid-based sampling; see, e.g., [27, 43, 44]. In particular, Metropolis–Hastings (M-H) methods employ a given *proposal* probability density $q(\boldsymbol{m}_k, \boldsymbol{y})$ at each sample point in parameter space $\boldsymbol{m}_k$ to generate a proposed sample point $\boldsymbol{y}$. Once generated, the M-H criterion chooses to either accept or reject the proposed sample point, and repeats from the new point, thereby generating a chain of samples from the posterior density $\pi_{\text{post}}(\boldsymbol{m})$. Algorithm 1.1 [27, section 3.6.2] presents pseudocode for the M-H method. For example, a popular choice for the proposal density is the isotropic Gaussian $q(\boldsymbol{m}_k, \boldsymbol{y}) = \frac{1}{(2\pi)^{n/2}} \exp[-\frac{1}{2}(\|\boldsymbol{m}_k - \boldsymbol{y}\|)^2]$; the resulting method is known as random walk metropolis. This proposal density is easy to sample but can lead to poor MCMC performance due to the mismatch between the proposal and posterior densities. This problem is greatly compounded when the parameter dimension is large, and in these cases it is critical that this mismatch be minimized to obtain acceptable MCMC performance. The challenge is to devise a proposal density $q(\boldsymbol{m}_k, \boldsymbol{y})$ that is both easy to sample and a good representation of the underlying posterior probability density.

A traditional approach is to utilize a single site updating scheme [25, 33]. This approach is more forgiving of naive proposal densities but requires as many forward simulations as parameters to perform a single parameter sweep. When the forward simulation is expensive and the parameter dimension is large, this approach is computationally intractable. In this paper, we therefore restrict our attention to multivariate proposal densities that update the entire parameter vector at once.

Given the connection between the inverse covariance matrix of the posterior pdf and the Hessian of the deterministic regularized misfit mentioned above, our goal is to capitalize on advances in algorithms for deterministic inverse problems to construct proposal densities for M-H MCMC that exploit the structure of the posterior pdf. In particular, we construct local Gaussian approximations of the posterior pdf from gradient and Hessian information of the negative log posterior. Drawing samples from this proposal density then requires solving systems that are identical to the Newton step for a deterministic inverse problem, thereby exploiting advances in fast Newton methods for deterministic inverse problems. Using modern adjoint techniques, gradients can be computed at a cost of a single linearized forward solve, as can actions of Hessians on vectors. These tools, combined with specialized solvers that exploit

the fact that many ill-posed inverse problems have compact data misfit operators, often permit solution of deterministic inverse problems in a dimension-independent (and typically small) number of iterations (see, e.g., [1]).

We study the efficiency of the Stochastic Newton method introduced here on a model seismic inverse problem, that of recovering the distribution of stiffness of an elastic medium from noisy observations of seismically induced ground motion at the surface. Stochastic Newton is compared with a freely available implementation of another popular method that attempts to exploit posterior covariance information, the delayed rejection adaptive Metropolis (DRAM) method, and with unpreconditioned Langevin MCMC. The results demonstrate large speedups over the other methods and suggest mesh independence of Stochastic Newton for problems with up to 1025 parameters. We demonstrate experimentally that Stochastic Newton is able to take large steps without compromising acceptance rates, and that convergence diagnostics and integrated autocorrelation functions show substantial improvement in the Stochastic Newton sample chains over traditional MCMC sample chains.

1.2. Approaches for sampling posterior pdfs. In this subsection, we review existing approaches to the solution of the statistical inverse problem and conclude by describing the relationship between the proposed Stochastic Newton method and existing methods. We restrict this review to methods for sampling pdfs that arise specifically from large-scale statistical inverse problems characterized by "expensive" forward models (e.g., those governed by PDEs) and high-dimensional parameter spaces (e.g., those that arise by discretization of heterogeneous PDE coefficients). For such problems, nearly every existing method ultimately gives the solution to the statistical inverse problem as a set of samples drawn from the posterior pdf. To make this sampling tractable, some form of reduction is often advocated. Below we review several different forms of reduction of the forward model and parameter space that have been proposed. We proceed from these reduced modeling approaches to increasingly "intrusive" sampling methods, eventually making use of first and second derivative information to characterize the posterior pdf.

1.2.1. Reduced modeling. A popular approach to working with a large number of parameters is to reduce the dimension of the problem in some way during the computation of the (expensive) parameter-to-observable map and later generate samples by interrogating this reduced representation at a correspondingly reduced cost. Projection-type reduced order models are one possible realization of this idea. Here, the state space is projected onto a limited number of basis functions to obtain an inexpensive reduced forward model. This is then used for posterior evaluation or sampling [3, 10, 16, 35, 46]. In addition, the parameter space may also be reduced to facilitate MCMC methods that work well in low dimensions [29]. The challenge has been to develop reduced models that are faithful over the full high-dimensional parameter space.

Other approaches use a truncated polynomial chaos (PC) expansion to represent the uncertain parameters and construct an approximate stochastic forward problem by Galerkin projection onto this PC basis [19]. This stochastic problem is more expensive than the original forward problem, but once obtained, the solution can be used to construct a surrogate for the posterior distribution, which can be evaluated repeatedly at negligible cost, making it ideal for MCMC sampling [30, 32]. Because the total number of terms in the PC expansion is combinatorial in the parameter dimension, a truncated Karhunen–Loève (KL) expansion (based on the prior distribution) may be employed to prevent the cost of the stochastic forward problem from becoming prohibitive [31]. However, it is necessary to ensure that enough KL modes are retained

so that the solution of the statistical inverse problem is not significantly biased toward the prior distribution.

Alternatively, after reduction using a PC basis, one can formulate a functional optimization problem over the stochastic space to characterize the solution to the inverse problem [4]. This idea can be combined with Smolyak sparse grids and stochastic collocation to tackle higher-dimensional problems as well [47]. One may also approximate the parameter-to-observable map with a Gaussian process model that is constructed via Bayesian model calibration over a limited set of training data (limited in both the number of experimental observations available as well as the number of forward model evaluations) [23, 28]. Additionally, the Gaussian process model may incorporate local Hessian information to estimate covariance matrices needed in the Gaussian process representation [9].

1.2.2. Adaptive sampling. As an alternative, we may instead "sample then reduce," wherein the full parameter space is sampled by an MCMC method that is able to cope with the high dimensionality and strong correlation structure inherent in ill-posed inverse problems. This can be of particular importance when modes of the parameter space that are important to the inverse problem do not align well with a coordinate basis or strong modes of the prior in the KL expansion, and any reduced basis generated by these approaches would require a prohibitive number of basis vectors to solve the problem with sufficient accuracy.

Delayed rejection adaptive metropolis (DRAM) MCMC adaptively constructs an approximation to the posterior covariance matrix to guide the sampling process and cope with the correlation structure [21]. DRAM requires only the ability to evaluate the posterior density at an arbitrary point and can thus be considered a black-box (or "nonintrusive") method. Similarly, the so-called t-walk requires only pointwise evaluations, but is specifically designed to be invariant to scale and correlation structure, allowing it to perform well on problems that have different scales or correlations in different regions of parameter space [13].

Many MCMC methods also employ derivative information to help guide sampling, which is more demanding of the types of information that need to be computed from the forward map. Langevin MCMC employs a stochastic differential equation (SDE) that has the desired posterior distribution as a stationary solution. Trajectories (realizations) of this SDE can thus be used to construct sample chains for the posterior distribution. When discretized, a finite timestep must be selected, and the discrete trajectories may no longer be faithful to the original SDE. Langevin MCMC restores convergence of the sample chain to the desired posterior distribution by considering each timestep as a proposal distribution for the M-H algorithm (see, e.g., [2, 39]). This also permits the use of inexpensive approximate gradient information (e.g., computed based on a coarse scale model) [14].

Another class of methods uses a two stage proposal process, where the proposal is first subjected to an accept/reject step based on an inexpensive approximate model (e.g., based on a coarse scale model), and the expensive true solution is computed only when the proposal is likely to be accepted [12, 15, 24].

Finally, Hamiltonian Monte Carlo (HMC) extends the parameter space at each MCMC sample to include a momentum variable, chooses a random sample from momentum space, and integrates a Hamiltonian system to generate proposal points. Derivative information of the posterior density is also used for this approach in the construction and solution of this system. A review of HMC methods can be found in [34].

1.2.3. Hessian-based sampling. Last but not least, we consider methods that make use of Hessian information (i.e., second derivatives) of the forward map. This information is generally more expensive to obtain but can prove highly beneficial to speeding up convergence of the sampling process. MCMC methods that utilize Hessian information have been considered previously [17, 18, 36] but are practicable only for a small number of parameters or for problems where an analytical expression for the Hessian is available. In [22], a BFGS-type approximation of the Hessian is considered for this purpose to avoid explicit computation of second derivatives and demonstrated on a 16 parameter Gaussian posterior distribution.

Another interesting approach makes use of the Fisher information as a natural metric for a Riemannian manifold [20]. Langevin MCMC and HMC can both be derived in this particular metric and show significant gains over the traditional varieties of MCMC by respecting the local structure of the parameter space. This method employs what amounts to the Gauss–Newton approximation of the Hessian of the negative log posterior, as well as additional third derivative terms. Computing the exact Gauss–Newton Hessian is generally intractable for large-scale inverse problems since it requires the solution of as many forward problems as the number of parameters.

Finally, the Stochastic Newton method we introduce in this paper can be understood as a relative of a preconditioned Langevin MCMC method, where the preconditioning is performed with the local Hessian of the negative log posterior. It is noteworthy that we obtain a similar preconditioning term to the one that appears in the Riemannian-manifold derivation of Langevin MCMC; in this paper, however, we construct an accurate low-rank representation of the Hessian and show that all necessary computations can be performed without constructing the full Hessian operator. This permits scalability to large parameter dimensions.

1.3. Outline of the paper. In section 2, we demonstrate the natural connections between deterministic optimization and the statistical inverse problem, use these connections to derive Stochastic Newton MCMC, and derive the low-rank approximations required to make the method tractable for large-scale inverse problems. Section 3 presents a motivating Bayesian statistical inverse problem based on seismic wave scattering. Finally, in section 4 we compare the performance of Stochastic Newton with DRAM MCMC and Langevin MCMC in various convergence metrics, demonstrate that Stochastic Newton offers a favorable tradeoff between increased complexity of the computations and improved MCMC convergence, and show examples that support good scalability with increasing dimensionality of parameter space.

2. Stochastic Newton MCMC. Large-scale optimization provides many tools and insights—in particular, Newton's method and its matrix-free variants—that accelerate the solution of deterministic inverse problems. In this section, we develop the Stochastic Newton method, which exploits natural connections between the deterministic inverse problem and the Bayesian statistical inverse problem to accelerate statistical sampling methods. Moreover, motivated by the spectral structure of underlying infinite-dimensional Hessian operators that appear in many ill-posed inverse problems, we introduce low-rank approximations that make the Stochastic Newton method tractable in high dimensions.

2.1. Connection with optimization. Consider a finite-dimensional parameter-to-observable map $\boldsymbol{d} = \boldsymbol{f}(\boldsymbol{m})$ that maps parameters $\boldsymbol{m} \in \mathbb{R}^n$ to observables $\boldsymbol{d} \in \mathbb{R}^m$. The deterministic inverse problem seeks to minimize $\frac{1}{2}\|\boldsymbol{f}(\boldsymbol{m}) - \boldsymbol{d}_{\text{obs}}\|^2_{\boldsymbol{W}}$, the misfit between the model predictions and the observed data $\boldsymbol{d}_{\text{obs}} \in \mathbb{R}^m$ in the $\boldsymbol{W}$-norm,

with $\boldsymbol{W} \in \mathbb{R}^m \times \mathbb{R}^m$. A quadratic regularization term $\frac{1}{2}\|\boldsymbol{m}-\bar{\boldsymbol{m}}\|^2_{\boldsymbol{R}}$ penalizes distance from a baseline vector of parameters $\bar{\boldsymbol{m}} \in \mathbb{R}^n$ in the $\boldsymbol{R}$-norm, with $\boldsymbol{R} \in \mathbb{R}^n \times \mathbb{R}^n$. Appropriate regularization of this form addresses ill-posedness of the inverse problem and guarantees uniqueness of the solution $\boldsymbol{m}^*$ to the following deterministic inverse problem:

$$\boldsymbol{m}^* = \arg\min_{\boldsymbol{m}} \left(\tfrac{1}{2}\|\boldsymbol{f}(\boldsymbol{m}) - \boldsymbol{d}_{\text{obs}}\|^2_{\boldsymbol{W}} + \tfrac{1}{2}\|\boldsymbol{m}-\bar{\boldsymbol{m}}\|^2_{\boldsymbol{R}} \right). \tag{2.1}$$

In the statistical inverse setting, we observe that Bayes's theorem (1.1) can be understood directly in the deterministic context if we inspect the negative log-posterior:

$$-\log \pi_{\text{post}} = -\log \pi_{\text{like}} - \log \pi_{\text{prior}} + \text{const.} \tag{2.2}$$

The constant of proportionality from Bayes's theorem is included above but affects neither the deterministic optimization nor the statistical inverse problem. In the statistical setting of the inverse problem, the misfit $\boldsymbol{f}(\boldsymbol{m}) - \boldsymbol{d}_{\text{obs}}$ is interpreted as a vector-valued random variable. When the measurement error and model error are unbiased, additive, and Gaussian, we have $(\boldsymbol{f}(\boldsymbol{m}) - \boldsymbol{d}_{\text{obs}}) \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Gamma}_{\text{noise}})$. (See, e.g., section 3.2.1 of [27].) The log-likelihood function in this context plays the role of the misfit term in the optimization formulation (2.1):

$$-\log \pi_{\text{like}}(\boldsymbol{d}_{\text{obs}}|\boldsymbol{m}) = \tfrac{1}{2}(\boldsymbol{f}(\boldsymbol{m}) - \boldsymbol{d}_{\text{obs}})^T \boldsymbol{\Gamma}^{-1}_{\text{noise}} (\boldsymbol{f}(\boldsymbol{m}) - \boldsymbol{d}_{\text{obs}}). \tag{2.3}$$

More general considerations of measurement and model error are possible and do not restrict the applicability of our method.

Similarly, if the prior density is Gaussian with mean $\bar{\boldsymbol{m}}_{\text{prior}}$ and covariance matrix $\boldsymbol{\Gamma}_{\text{prior}}$, then the log-prior term in (2.2) plays the role of the regularization from deterministic optimization:

$$-\log \pi_{\text{prior}}(\boldsymbol{m}) = \tfrac{1}{2}(\boldsymbol{m} - \bar{\boldsymbol{m}}_{\text{prior}})^T \boldsymbol{\Gamma}^{-1}_{\text{prior}} (\boldsymbol{m} - \bar{\boldsymbol{m}}_{\text{prior}}). \tag{2.4}$$

As before the assumption of a Gaussian prior can also be relaxed.

The negative log-posterior (2.2) is now understood directly as the cost function $V(\boldsymbol{m})$ from deterministic optimization, and therefore we can write the posterior density as

$$\pi_{\text{post}}(\boldsymbol{m}|\boldsymbol{d}_{\text{obs}}) \propto \exp(-V(\boldsymbol{m})), \tag{2.5}$$

where the cost function $V(\boldsymbol{m})$ is given by

$$V(\boldsymbol{m}) := \tfrac{1}{2}\|f(\boldsymbol{m}) - \boldsymbol{d}_{\text{obs}}\|^2_{\boldsymbol{\Gamma}^{-1}_{\text{noise}}} + \tfrac{1}{2}\|\boldsymbol{m} - \bar{\boldsymbol{m}}_{\text{prior}}\|^2_{\boldsymbol{\Gamma}^{-1}_{\text{prior}}}. \tag{2.6}$$

The connection between the cost functional from deterministic optimization and the posterior probability density in the statistical setting is made explicit when we seek the maximum a posteriori (MAP) estimate $\boldsymbol{m}_{\text{MAP}}$, which is given by maximizing the posterior or, equivalently, by minimizing the cost function. Thus, $\boldsymbol{m}_{\text{MAP}} = \boldsymbol{m}^*$ when the appropriate definitions of $\boldsymbol{W}$ and $\boldsymbol{R}$ are taken in (2.1). Next, we consider how to further exploit this connection between deterministic and statistical inversion.

2.2. The Gaussian linear case. When the parameter-to-observable map is linear, we write $\boldsymbol{f}(\boldsymbol{m}) = \boldsymbol{G}\boldsymbol{m}$ with $\boldsymbol{G} \in \mathbb{R}^{m \times n}$. In this case we observe that the negative log-posterior (or deterministic cost function)

$$V(\boldsymbol{m}) = \tfrac{1}{2}(\boldsymbol{G}\boldsymbol{m}-\boldsymbol{d}_{\text{obs}})^T \boldsymbol{\Gamma}^{-1}_{\text{noise}} (\boldsymbol{G}\boldsymbol{m}-\boldsymbol{d}_{\text{obs}}) + \tfrac{1}{2}(\boldsymbol{m}-\bar{\boldsymbol{m}}_{\text{prior}})^T \boldsymbol{\Gamma}^{-1}_{\text{prior}} (\boldsymbol{m}-\bar{\boldsymbol{m}}_{\text{prior}}) \tag{2.7}$$

is quadratic in the parameters $\boldsymbol{m}$. Making use of the expressions for the gradient $\boldsymbol{g}$ and Hessian $\boldsymbol{H}$ of $V(\boldsymbol{m})$,

$$\boldsymbol{g} := \boldsymbol{g}(\boldsymbol{m}) = \nabla V(\boldsymbol{m}) = \boldsymbol{G}^T \boldsymbol{\Gamma}_{\text{noise}}^{-1}(\boldsymbol{G}\boldsymbol{m} - \boldsymbol{d}_{\text{obs}}) + \boldsymbol{\Gamma}_{\text{prior}}^{-1}(\boldsymbol{m} - \bar{\boldsymbol{m}}_{\text{prior}}), \tag{2.8}$$

$$\boldsymbol{H} := \nabla^2 V(\boldsymbol{m}) = \boldsymbol{G}^T \boldsymbol{\Gamma}_{\text{noise}}^{-1} \boldsymbol{G} + \boldsymbol{\Gamma}_{\text{prior}}^{-1}, \tag{2.9}$$

we can rewrite the cost function in the form

$$V(\boldsymbol{m}) = \tfrac{1}{2}(\boldsymbol{m} - \boldsymbol{m}^*)^T \boldsymbol{H}(\boldsymbol{m} - \boldsymbol{m}^*) + \text{const}, \tag{2.10}$$

which makes clear, since both $\boldsymbol{\Gamma}_{\text{noise}}$ and $\boldsymbol{\Gamma}_{\text{prior}}$ and therefore $\boldsymbol{H}$ are positive definite, that a unique minimum of $V(\boldsymbol{m})$ exists and is given by requiring $\boldsymbol{g}(\boldsymbol{m}^*) = \boldsymbol{0}$:

$$\boldsymbol{m}^* = \boldsymbol{H}^{-1} \left(\boldsymbol{G}^T \boldsymbol{\Gamma}_{\text{noise}}^{-1} \boldsymbol{d}_{\text{obs}} + \boldsymbol{\Gamma}_{\text{prior}}^{-1} \bar{\boldsymbol{m}}_{\text{prior}} \right). \tag{2.11}$$

Moreover, the posterior pdf $\exp(-V(\boldsymbol{m}))$ can be seen to be Gaussian with mean given by the minimizer of $V(\boldsymbol{m})$, i.e., the solution of the deterministic inverse problem (2.1), and covariance given by the inverse of the Hessian, $\boldsymbol{H}^{-1}$; i.e., π_{post} is distributed as $\mathcal{N}(\boldsymbol{m}^*, \boldsymbol{H}^{-1})$. Hence, we see an explicit connection between the deterministic solution and its statistical counterparts, at least in the case of a linear parameter-to-observable map.

2.3. The nonlinear case and Stochastic Newton's method. When the parameter-to-observable map $\boldsymbol{f}(\boldsymbol{m})$ is nonlinear, the posterior is no longer Gaussian, and in general the minimum of the cost function no longer coincides with the mean of the posterior, nor does the inverse of the Hessian coincide with the covariance matrix of the posterior. However, we can still exploit connections between deterministic optimization methods for minimizing $V(\boldsymbol{m})$ and statistical methods for sampling the posterior π_{post}.

The gold standard for optimization is Newton's method, which begins with a local quadratic approximation $\tilde{V}(\boldsymbol{m})$ of the cost function about a given point $\boldsymbol{m}_k$, which can be written as

$$V(\boldsymbol{m}) \approx \tilde{V}(\boldsymbol{m}) = \frac{1}{2}(\boldsymbol{m} - \boldsymbol{m}_k)^T \boldsymbol{H}(\boldsymbol{m} - \boldsymbol{m}_k) + \boldsymbol{g}^T(\boldsymbol{m} - \boldsymbol{m}_k) + V(\boldsymbol{m}_k), \tag{2.12}$$

with gradient $\boldsymbol{g}(\boldsymbol{m}_k) = \nabla V(\boldsymbol{m}_k)$ and Hessian $\boldsymbol{H}(\boldsymbol{m}_k) = \nabla^2 V(\boldsymbol{m}_k)$.

In the vicinity of a local minimum, $\boldsymbol{H}$ is positive definite. However, at an arbitrary point $\boldsymbol{m}$, $\boldsymbol{H}$ is not guaranteed to be positive definite, and in such cases it is necessary to replace $\boldsymbol{H}$ with a suitably modified positive definite Hessian $\tilde{\boldsymbol{H}}$ in $\tilde{V}(\boldsymbol{m})$ in order to guarantee convergence. A simple choice for $\tilde{\boldsymbol{H}}$ is an eigenvalue decomposition of $\boldsymbol{H}$, with small or negative eigenvalues replaced with a minimum threshold value. Finally, we rearrange (2.12) as we did in the Gaussian linear case (2.10) to write

$$\tilde{V}(\boldsymbol{m}) = \frac{1}{2}(\boldsymbol{m} - \boldsymbol{m}_k + \tilde{\boldsymbol{H}}^{-1}\boldsymbol{g})^T \tilde{\boldsymbol{H}}(\boldsymbol{m} - \boldsymbol{m}_k + \tilde{\boldsymbol{H}}^{-1}\boldsymbol{g}) + \text{const}, \tag{2.13}$$

which shows that the minimizer of $\tilde{V}(\boldsymbol{m})$ is given by $\boldsymbol{m}_{k+1} = \boldsymbol{m}_k - \tilde{\boldsymbol{H}}^{-1}\boldsymbol{g}$. Note that $-\tilde{\boldsymbol{H}}^{-1}\boldsymbol{g}$ is the Newton step, and iterating this process leads to the classical Newton method.

In the statistical setting, inserting $\tilde{V}(\boldsymbol{m})$ into (2.5) leads to an approximation of π_{post} given by

$$\pi_{\text{post}}(\boldsymbol{m}) \approx \tilde{\pi}(\boldsymbol{m}) = \exp(-\tilde{V}(\boldsymbol{m})), \tag{2.14}$$

which is in fact a Gaussian, centered on the point $\boldsymbol{m}_{k+1}$ (the result of the deterministic Newton step) with covariance $\tilde{\boldsymbol{H}}^{-1}$.

Having constructed a local Gaussian approximation of the posterior pdf, we are now in a position to define the Stochastic Newton method, which is an MCMC method that uses the normalized proposal density

$$\tilde{\pi}(\boldsymbol{y}) = \frac{\det \tilde{\boldsymbol{H}}^{1/2}}{(2\pi)^{n/2}} \exp\left(-\frac{1}{2}\left(\boldsymbol{y} - \boldsymbol{m}_k + \tilde{\boldsymbol{H}}^{-1}\boldsymbol{g}\right)^T \tilde{\boldsymbol{H}}\left(\boldsymbol{y} - \boldsymbol{m}_k + \tilde{\boldsymbol{H}}^{-1}\boldsymbol{g}\right)\right). \tag{2.15}$$

Recall that the quadratic approximation is constructed using gradient and Hessian information at $\boldsymbol{m}_k$, namely $\tilde{\boldsymbol{H}}(\boldsymbol{m}_k)$ and $\boldsymbol{g}(\boldsymbol{m}_k)$. Thus, we have "tailored" the proposal density $q(\boldsymbol{m}_k, \boldsymbol{y}) = \tilde{\pi}(\boldsymbol{y})$ to the underlying posterior pdf using derivative information of $V(\boldsymbol{m})$. The Stochastic Newton step at each MCMC iteration proposes a sample $\boldsymbol{y}$ from the density $\tilde{\pi}(\boldsymbol{y})$, which is then subjected to the accept/reject framework of the M-H algorithm. Pseudocode for Stochastic Newton MCMC for this problem is given in Algorithm 2.

If in fact the posterior density π_{post} is Gaussian (see, e.g., section 2.2) and the Hessian $\tilde{\boldsymbol{H}}$ is exact, then $q(\boldsymbol{m}_k, \boldsymbol{y}) = \tilde{\pi}(\boldsymbol{y}) = \pi_{\text{post}}(\boldsymbol{y})$, and the M-H acceptance probability in Algorithm 2 reduces to

$$\alpha(\boldsymbol{m}_k, \boldsymbol{y}) = \min\left\{1, \frac{\pi(\boldsymbol{y})q(\boldsymbol{y}, \boldsymbol{m}_k)}{\pi(\boldsymbol{m}_k)q(\boldsymbol{m}_k, \boldsymbol{y})}\right\} = \min\left\{1, \frac{\pi(\boldsymbol{y})\pi(\boldsymbol{m}_k)}{\pi(\boldsymbol{m}_k)\pi(\boldsymbol{y})}\right\} = 1. \tag{2.16}$$

Thus in this case we achieve "perfect sampling," in which all samples are independent draws from the true posterior density $\pi_{\text{post}}(\boldsymbol{m})$ and are accepted with probability 1.

Before concluding this section, we make one final remark about the threshold value used to define $\tilde{\boldsymbol{H}}$. Because $\tilde{\boldsymbol{H}}^{-1}$ is used as the covariance matrix for the proposal distribution, this minimum threshold value for $\tilde{\boldsymbol{H}}$ guarantees a maximum covariance value for the proposal density. This threshold value can therefore be used as a tunable parameter in MCMC to restrict the maximum desired step length to improve performance if the sample acceptance rate is too low.

2.4. Low-rank Hessian approximation. The MCMC method we are proposing here has been contemplated before [18, 36, 37], but they are applied only to low-dimensional sampling problems and not computationally intensive inverse problems as we consider here. Attempting to apply the method as described above to such problems will quickly lead to failure, since constructing just one Hessian requires n forward solves (see, e.g., [5, 26]), that is, equal to the number of parameters. Thus MCMC becomes intractable for expensive forward problems (e.g., governed by PDEs) and in high dimensions (e.g., when the parameters describe a discretization of a field such as a PDE coefficient, initial condition, boundary condition, etc.).

However, experience with large-scale deterministic inverse problems has shown in many cases that the Hessian of the data misfit term in (2.1) is a compact operator whose range space is independent of mesh resolution (see, e.g., [45]). The intuition behind the compactness of the Hessian of the data misfit term,

$$\boldsymbol{H}_{\text{misfit}} = -\nabla^2 \log \pi_{\text{like}}, \tag{2.17}$$

ALGORITHM 2. Stochastic Newton MCMC Algorithm to sample π_{post}.

Choose initial $\boldsymbol{m}_0$
Compute $\pi_{\text{post}}(\boldsymbol{m}_0), \boldsymbol{g}(\boldsymbol{m}_0), \boldsymbol{H}(\boldsymbol{m}_0)$
for $k = 0, \ldots, N-1$ **do**
  Define $q(\boldsymbol{m}_k, \boldsymbol{y}) = \tilde{\pi}(\boldsymbol{y})$ as in (2.15)
  Draw sample $\boldsymbol{y}$ from the proposal density $q(\boldsymbol{m}_k, \cdot)$
  Compute $\pi_{\text{post}}(\boldsymbol{y}), \boldsymbol{g}(\boldsymbol{y}), \boldsymbol{H}(\boldsymbol{y})$
  Compute $\alpha(\boldsymbol{m}_k, \boldsymbol{y}) = \min\left(1, \frac{\pi_{\text{post}}(\boldsymbol{y}) q(\boldsymbol{y}, \boldsymbol{m}_k)}{\pi_{\text{post}}(\boldsymbol{m}_k) q(\boldsymbol{m}_k, \boldsymbol{y})}\right)$
  Draw $u \sim \mathcal{U}([0,1])$
  if $u < \alpha(\boldsymbol{m}_k, \boldsymbol{y})$ **then**
    Accept: Set $\boldsymbol{m}_{k+1} = \boldsymbol{y}$
  else
    Reject: Set $\boldsymbol{m}_{k+1} = \boldsymbol{m}_k$
  end if
end for

is that for many ill-posed inverse problems, the observations are sparse and typically inform only a limited number of modes of the parameter field; thus, the Jacobian matrix of observables $\boldsymbol{f}(\boldsymbol{m})$ with respect to parameters $\boldsymbol{m}$ is well-approximated by a low-rank matrix. In particular, it can be shown that the Hessian of the data misfit operator for the inverse medium scattering problem we consider in section 3 is a compact operator with exponentially decaying spectrum (when the medium is analytic) [8]. This property suggests a low-rank approximation of the data misfit Hessian, which permits us to avoid prohibitive computation of the full Hessian. Below, we exploit the compactness of the data misfit Hessian to make the Stochastic Newton MCMC method presented here tractable for large-scale problems.

In the Bayesian setting, the Hessian $\boldsymbol{H}$ can be written as a sum of data misfit and prior Hessians, i.e.,

$$\boldsymbol{H} = \boldsymbol{H}_{\text{misfit}} + \boldsymbol{\Gamma}_{\text{prior}}^{-1}. \tag{2.18}$$

Consider a decomposition of the prior such that $\boldsymbol{\Gamma}_{\text{prior}} = \boldsymbol{L}\boldsymbol{L}^T$, computed either as the symmetric square root $\boldsymbol{L} = \boldsymbol{\Gamma}_{\text{prior}}^{1/2}$ or as the Cholesky factorization.[1] Rewriting $\boldsymbol{H}$ as

$$\boldsymbol{H} = \boldsymbol{L}^{-T}\left(\boldsymbol{L}^T \boldsymbol{H}_{\text{misfit}} \boldsymbol{L} + \boldsymbol{I}\right)\boldsymbol{L}^{-1}, \tag{2.19}$$

we see that the expression $\boldsymbol{L}^T \boldsymbol{H}_{\text{misfit}} \boldsymbol{L}$ emerges as a natural candidate for a low-rank spectral approximation, since comparison with the identity provides a quantitative criterion for truncating the spectrum, and since $\boldsymbol{\Gamma}_{\text{prior}}$ is often a smoothing operator, and thus the collapse of the spectrum of $\boldsymbol{H}_{\text{misfit}}$ is then enhanced by preconditioning with $\boldsymbol{L}$. The low-rank approximation of $\boldsymbol{L}^T \boldsymbol{H}_{\text{misfit}} \boldsymbol{L}$ represents the parameter subspace in which the data are most informative about the parameters and least constrained by the prior.

[1] For problems with very large parameter dimension, this factorization may become prohibitively costly to perform. In this case, one prefers to exploit the structure of the prior to specify $\boldsymbol{L}$ or $\boldsymbol{\Gamma}_{\text{prior}}^{1/2}$ directly—or their action on a vector—via an $O(n)$ method. For example, for a smoothing prior, one can view $\boldsymbol{\Gamma}_{\text{prior}}^{1/2}$ as an elliptic solve via a multigrid method [7, 41].

Using Lanczos (or any of its siblings [40]), an r-dimensional low-rank approximation can be represented as $\boldsymbol{L}^T\boldsymbol{H}_{\text{misfit}}\boldsymbol{L} \approx \boldsymbol{V}_r\boldsymbol{D}_r\boldsymbol{V}_r^T$, where $\boldsymbol{V}_r \in \mathbb{R}^{n\times r}$ contains the dominant eigenvectors, and the diagonal matrix $\boldsymbol{D}_r \in \mathbb{R}^{r\times r}$ contains the dominant eigenvalues. The number of desired eigenvectors r is determined by truncating the decomposition once the eigenvalues fall below some threshold value $\alpha \ll 1$, below which it is assumed that the prior dominates the character of the Hessian. This truncation also ensures positive definiteness of the low-rank approximation of $\boldsymbol{H}$, which we identify as $\tilde{\boldsymbol{H}}$.

Tractability of the low-rank approximation and its use in the Stochastic Newton Method can be established as follows. First, Lanczos requires only matrix-vector products ("matvecs"), and therefore there is no need to explicitly form the (dense) Hessian. Second, Lanczos tends to perform only as many matvecs as there are extreme (dominant) eigenvalues, so that compactness of the Hessian bounds the number of required Lanczos iterations. Third, each matvec requires only a pair of forward and adjoint PDE solves (see, e.g., [5], [26, section 1.6.5]). Therefore, the approximation can be constructed in a number of PDE solves comparable to the number of dominant eigenvalues, r. For many ill-posed inverse problems in which the parameters are a discretization of an unknown field, the dominant eigenvalues are associated with smooth eigenvectors (physically, this is a consequence of the data being uninformative about small length scales); as such, the dominant eigenvalues are unaffected by subsequent refinement, once a suitable discretization level is achieved. Thus, r is often independent of n (see, e.g., [7]). Finally, we observe that all necessary MCMC computations involving the Hessian can be performed without ever explicitly constructing the dense operator as follows:

$$\tilde{\boldsymbol{H}} = \boldsymbol{L}^{-T}\big[\boldsymbol{V}_r\boldsymbol{D}_r\boldsymbol{V}_r^T + \boldsymbol{I}\big]\boldsymbol{L}^{-1}, \tag{2.20}$$

$$\tilde{\boldsymbol{H}}^{-1}\boldsymbol{g} = \boldsymbol{L}\big\{\boldsymbol{V}_r\big[(\boldsymbol{D}_r + \boldsymbol{I}_r)^{-1} - \boldsymbol{I}_r\big]\boldsymbol{V}_r^T + \boldsymbol{I}\big\}\boldsymbol{L}^T\boldsymbol{g}, \tag{2.21}$$

$$\tilde{\boldsymbol{H}}^{-1/2}\boldsymbol{x} = \boldsymbol{L}\big\{\boldsymbol{V}_r\big[(\boldsymbol{D}_r + \boldsymbol{I}_r)^{-1/2} - \boldsymbol{I}_r\big]\boldsymbol{V}_r^T + \boldsymbol{I}\big\}\boldsymbol{x}, \tag{2.22}$$

$$\det(\tilde{\boldsymbol{H}}^{1/2}) = (\det\boldsymbol{L})^{-1}\prod_{i=1}^{r}(d_i+1)^{1/2}. \tag{2.23}$$

Expression (2.21) computes the Newton step, (2.22) allows us to sample from a Gaussian distribution with covariance $\tilde{\boldsymbol{H}}^{-1}$, and finally (2.23) is necessary in the computation of the accept/reject criterion of M-H. With the exception of operations with the square root of the prior, $\boldsymbol{L}$, the complexity of operations in (2.21)–(2.22) is $O(rn)$, where as noted above r is often independent of n. The determinant (2.23) requires only $O(r)$ operations in practice, since $\det(\boldsymbol{L})$ can be precomputed once. Finally, the complexity of carrying out operations with $\boldsymbol{L}$ in (2.21)–(2.22) appears naively to be $O(n^2)$; however, as mentioned in footnote 1, for very large scale problems (particularly on parallel computers), one would avoid a naive factorization and instead interpret the action of $\boldsymbol{L}$ on a vector using a fast solver [7].

In summary, the low-rank representation $\tilde{\boldsymbol{H}}$ can in many cases be computed efficiently (i.e., in a small number of PDE solves, independent of mesh, and therefore of problem, size) and applied in $O(n)$ computational work.

2.5. Comparison with Langevin MCMC methods. Stochastic Newton also has a natural interpretation as a form of a Langevin MCMC method. In Langevin MCMC, we begin with the negative log-posterior $V(\boldsymbol{m})$ and construct trajectories of

the stochastic process from Langevin dynamics,

$$d\boldsymbol{X}_t = -\boldsymbol{A}\nabla V \mathrm{d}t + \sqrt{2}\boldsymbol{A}^{1/2} d\boldsymbol{W}_t, \tag{2.24}$$

which sample the desired probability density as $t \to \infty$ [42]. Here, $\boldsymbol{A}$ is a positive definite preconditioning matrix, and stochastic variables are denoted by $\boldsymbol{X}_t$ and $\boldsymbol{W}_t$, where $\boldsymbol{W}_t$ is the vector of standard independent Brownian motions. When $\boldsymbol{A}$ is the identity, we recover traditional Langevin dynamics.

To solve (2.24), time is discretized by the Euler–Maruyama method, with time step Δt, to yield the update

$$\boldsymbol{x}_{k+1} - \boldsymbol{x}_k = -\boldsymbol{A}\nabla V(\boldsymbol{x}_k)\Delta t + \mathcal{N}(\boldsymbol{0}, 2\Delta t \boldsymbol{A}). \tag{2.25}$$

Preconditioning with the local inverse Hessian $\boldsymbol{H}^{-1}$, choosing $\Delta t = 1$, and discarding the factor of 2, we can formally recover the Stochastic Newton method derived previously:

$$\boldsymbol{x}_{k+1} - \boldsymbol{x}_k = -\boldsymbol{H}^{-1}\nabla V(\boldsymbol{x}_k) + \mathcal{N}(\boldsymbol{0}, \boldsymbol{H}^{-1}). \tag{2.26}$$

Of course the Hessian is both nonconstant and not everywhere positive definite, and so Stochastic Newton is not rigorously understood as a Langevin MCMC, but there do exist definite parallels. Note that without preconditioning, i.e.,

$$\boldsymbol{x}_{k+1} - \boldsymbol{x}_k = -\nabla V(\boldsymbol{x}_k)\Delta t + \mathcal{N}(\boldsymbol{0}, 2\Delta t \boldsymbol{I}), \tag{2.27}$$

Langevin MCMC resembles a steepest descent method in the deterministic setting.

2.6. Comparison to other Gaussian MCMC proposal types. Stochastic Newton's use of a Hessian-based local Gaussian approximation as a proposal function can be contrasted with other types of Gaussian proposal functions. Figure 1 shows proposal density contours for several different proposal functions, using the Rosenbrock function as an example target density. All contours in the image are normalized so that they contain 5%, 50%, and 95% of the density, respectively. In this way, the best acceptance rates and sample chain convergence will be achieved for the proposal that matches the contours of the target density most closely.

3. Application to statistical seismic inverse problem. We demonstrate the Stochastic Newton method by solving a particular statistical inverse problem. Consider a theoretical seismic exploration experiment in which a surface explosion causes seismic waves to travel downward into the subsurface medium. If there are obstacles in the medium, or if the medium properties vary with depth, then a fraction of the seismic wave energy will scatter off of these boundaries and return to the surface to be observed at later times. The statistical inverse problem processes these observations to reconstruct a statistical description of the subsurface medium properties. Using this description we are able to estimate properties of the subsurface, including locations of buried objects or oil/mineral deposits.

The remainder of this section describes in detail the ingredients required for this statistical reconstruction. The first ingredient is the mathematical forward model, which maps input medium parameters to predicted observations. Second, the likelihood function uses these predicted observations to determine the probability that the given input medium parameters could have produced the observed experimental data. The third ingredient in any Bayesian analysis is the prior distribution, which encapsulates all of the assumptions about the subsurface medium before any experimental data is considered. Bayes's theorem combines these ingredients into the posterior

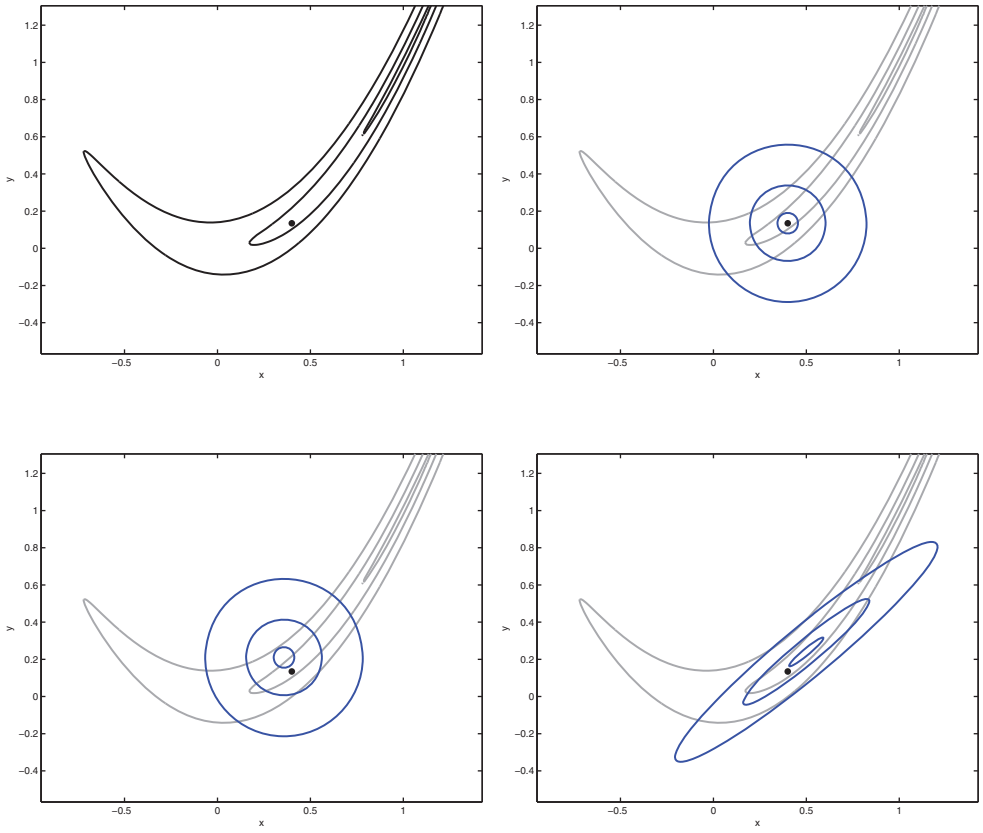

FIG. 1. *Visualizations of differing types of proposal distributions for MCMC. Top left shows contours of the classical Rosenbrock function from deterministic optimization, with effective (unnormalized) density* $\exp(-(1-x)^2-100(y-x^2)^2)$. *Top right shows the contours of the random walk proposal function overlaid on the Rosenbrock contours. Bottom left shows overlays of Langevin contours without preconditioning. Bottom right shows contours of the Stochastic Newton method–type proposal function.*

probability distribution over the set of input medium parameters, which is the statistical description of the subsurface medium which we seek. Finally, we describe the efficient computation of adjoint, gradient, and Hessian-vector product information which is required for use of our method.

3.1. The forward model. We model our exploration experiment using the 1D wave equation. The problem is solved on the spatial domain $\Omega = [0, L]$, where $z \in \Omega$ represents the depth beneath the surface at $z = 0$. At the maximum depth $z = L$, we use an absorbing boundary condition which allows plane waves to pass through the boundary without reflection.

The surface explosion is modeled with a right-hand side forcing input to the wave equation using a Ricker wavelet $F^{\text{ricker}}(t)$ in time with a mean spectrum energy density at 0.5 Hertz and a spatial delta function at the surface $\delta(z-0)$.

Finally, our model has two physical parameters, which are the density ρ and a stiffness parameter μ. In principle both parameters may vary freely with depth, but we will consider only variations in the stiffness $\mu(z; \boldsymbol{m})$ and assume a constant density $\rho = 1$. Note that we have included an explicit dependence on the model parameters $\boldsymbol{m}$.

The governing equations for the forward model are

$$\text{(PDE)} \qquad \rho\, u_{tt}(z,t) - \Big(\mu(z;\boldsymbol{m})\, u_z(z,t)\Big)_z = F^{\text{ricker}}(t)\delta(z-0),$$

$$\text{(Absorbing BC)} \qquad \mu(L;\boldsymbol{m})\, u_z(L,t) = -\sqrt{\rho\,\mu(L;\boldsymbol{m})}\, u_t(L,t),$$

$$\text{(Free Surface BC)} \qquad \mu(0;\boldsymbol{m})\, u_z(0,t) = 0,$$

$$\text{(IC)} \qquad u(z,0) = 0,$$

$$\text{(IC)} \qquad u_t(z,0) = 0.$$

These equations are solved numerically using finite elements on piecewise linear meshes in space using an explicit scheme in time, as in [11]. Most of the examples here are solved on 64 element (65 DOF) meshes, and a few are solved on 1024 element (1025 DOF) meshes. The physical parameter $\mu(z;\boldsymbol{m})$ is discretized as a linear combination of the same 65 or 1025 DOFs as the numerical PDE solution.

Finally, we observe the system by measuring the surface displacement at 120 equally spaced points in time. These measurements are assumed to contain errors at each observation time which are Gaussian, additive, and independent. The noise level is selected such that the resulting RMS signal to noise ratio is approximately 2:1.

It should be emphasized again that the role of the forward model $\boldsymbol{f}(\boldsymbol{m})$ is to map (stiffness) parameters $\boldsymbol{m}$ to surface displacement observations $\boldsymbol{d}$. In terms of the forward solution $u(z,t)$, the forward model can be expressed as a vector with components

$$f_i(\boldsymbol{m}) = u(0,t_i), \qquad i = 1,\ldots,120, \tag{3.1}$$

where $t_1,\ldots,t_{120}$ are the observation times. Although the underlying PDE is linear, this forward model map from parameters to observables is not.

3.2. The likelihood function. The likelihood function governs the probability that a candidate set of stiffness parameters $\mu(z;\boldsymbol{m})$ would reproduce the observation data $\boldsymbol{d}_{\text{obs}}$ that was measured in the exploration experiment. In our case, this observation data is synthetically generated according to the noise model assumed in the previous section.

We generate the experimental observation data on a mesh different from that used for statistical inversion (256 elements), and we additionally corrupt the observation data with additive Gaussian noise as discussed previously: $\boldsymbol{y}_{\text{obs}} = g(\boldsymbol{\mu}) + \varepsilon_{\text{noise}}$, where $\varepsilon_{\text{noise}} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Gamma}_{\text{noise}})$, and $\boldsymbol{\Gamma}_{\text{noise}} = \sigma^2_{\text{noise}}\boldsymbol{I}$. This is done to avoid "inverse crimes" [27], in which it might be artificially easy to invert for the desired parameters if the same mesh is used for the inversion as was used to generate the synthetic observation data.

Using the additive Gaussian noise model, our likelihood function is given as

$$\pi_{\text{like}}(\boldsymbol{d}_{\text{obs}}|\boldsymbol{m}) \propto \exp\Big[-\tfrac{1}{2}(\boldsymbol{f}(\boldsymbol{m}) - \boldsymbol{d}_{\text{obs}})^T \boldsymbol{\Gamma}^{-1}_{\text{noise}}(\boldsymbol{f}(\boldsymbol{m}) - \boldsymbol{d}_{\text{obs}})\Big]. \tag{3.2}$$

For the 2D problem (described in the next subsection), we use $\sigma_{\text{noise}} = 8\times 10^{-5}$, and for the 16D, 65D, and 1025D problems, we use $\sigma_{\text{noise}} = 2\times 10^{-5}$.

3.3. Parametrizations and priors. In this application, we set up four distinct statistical inverse problems which differ in the parametrization used to describe the medium and the prior imposed on each parametrization. The choice of prior in a statistical inverse problem can have a significant impact on both the computational

effort required to solve the problem as well as the posterior density itself. We select priors in this section that are intended to be typical of the priors that might be used in this case of a heterogeneous medium in which the fine scale variability of the medium is assumed negligible. Toward this end, we use Gaussian smoothness priors, which provide a flexible way to describe random fields with a desired degree of smoothness and are commonly employed in Bayesian inference of parameter fields. Except in the 1025D case described below, these parametrizations are considered as independent problems, and each have synthetic observation data which are unique to that parametrization.

In the simplest 2D case (i.e., with two independent parameters m_1, m_2) the medium is parameterized with four equal length layers, where we constrain the parameter values of the topmost and bottommost layers to be $\mu = 1$, leaving only two DOFs in the parametrization for the second and third layers. In this 2D case, we take the prior to be uniform over $[0.5, 10]$ (i.e., we specify as little a priori knowledge as possible except for the range of possible values):

$$\pi_{\text{pr}}(\boldsymbol{m}) \propto \begin{cases} 1 & \text{if} \quad 0.5 \le m_i \le 10 \quad \forall i, \\ 0 & \text{otherwise.} \end{cases} \tag{3.3}$$

In the intermediate 16D case, the medium is parameterized with 16 equal length layers, each containing four elements. In this case we do not further constrain any of the layer parameter values, but we use a (truncated) Gaussian smoothness prior to specify a priori knowledge that there should not be large jumps between parameter values in neighboring layers. The form of the prior is given explicitly for the layer parameter values m_i for $i = 1, \ldots, 16$:

$$\pi_{\text{prior}}(\boldsymbol{m}) \propto \begin{cases} \exp\Big(-\frac{1}{2}(\boldsymbol{m} - \bar{\boldsymbol{m}}_{\text{prior}})^T \boldsymbol{\Gamma}_{\text{prior}}^{-1} (\boldsymbol{m} - \bar{\boldsymbol{m}}_{\text{prior}}) & \text{if} \quad 0.5 \le m_i \le 10 \quad \forall i, \\ 0 & \text{otherwise,} \end{cases} \tag{3.4}$$

$$\bar{\boldsymbol{m}}^i_{\text{prior}} = 5, \tag{3.5}$$

$$\Gamma^{ij}_{\text{prior}} = \theta_1 \exp\left(\frac{-(z_i - z_j)^2}{2\theta_2^2} \right) + \varepsilon \delta_{ij}. \tag{3.6}$$

The values θ_1 and θ_2 specify the magnitude of the correlation and the correlation length, respectively. The layer depths z_i indicate the midpoint of each layer. In this example, the correlation length is chosen to be $\theta_2 = 0.125$ (a width of 2 layers). We add a small diagonal term $\varepsilon \delta^{ij}$ to ensure that the prior covariance remains numerically well conditioned. Here we choose $\varepsilon = 10^{-5}$.

Finally, the 65D and 1025D cases allow every discretization point in the mesh to be a separate parameter. (As such, the 1025D problem must be computed on a finer mesh.) We impose the same form of the prior as in the 16D case (again specifying a priori information about smoothness of the medium), where the z_i represent the depth of each mesh grid point, and we set the correlation length to $\theta_2 = 0.125$, which is intended to correspond to the same correlation length of 2 layers in the 16D case. As before, we add a small diagonal term $\varepsilon \delta^{ij}$ to ensure numerical well conditioning. We choose $\varepsilon = 10^{-5}$ in the 65D case and $\varepsilon = 10^{-12}$ in the 1025D case.

3.4. The statistical inverse problem. We are now prepared to describe the statistical inverse problem we seek to solve. Sections 3.1–3.3 describe in detail the

ingredients (forward model, likelihood, and prior) required to construct the posterior density using Bayes's theorem:

$$\pi_{\text{post}}(\boldsymbol{m}) \propto \pi_{\text{prior}}(\boldsymbol{m})\pi_{\text{like}}(\boldsymbol{d}_{\text{obs}}|\boldsymbol{m}). \tag{3.7}$$

Complete specification of a particular statistical inverse problem requires a set of observation data $\boldsymbol{d}_{\text{obs}}$ as defined in section 3.2 and a choice of medium parametrization and prior as defined in section 3.3.

In this paper, we consider three distinct inverse problems, corresponding to different choices of the 2D, 16D, and 65D medium parametrizations and associated prior distributions as described in section 3.3. In each problem, a sample from the prior distribution is selected to be the "ground truth" medium, which is then used to generate synthetic observation data as in section 3.2. As an experiment in "weak scaling" of our method, the same observation data are used in the 65D and 1025D experiments. In this sense, the 1025D problem is a refinement of the 65D problem, in which we desire to infer over a larger parameter space for the same fundamental underlying problem.

Finally, "solving" a statistical inverse problem reduces to the ability to interrogate $\pi_{\text{post}}(\boldsymbol{m})$. In high dimensions, this is a nontrivial problem even when the posterior density is known. Typically we are interested in the mean and covariance of the posterior distribution, and higher moments or other functionals of the distribution (e.g., event probabilities) may be desirable as well. Finally, probability distributions for specific quantities of interest (e.g., the softest type of rock in the medium $\min_z \mu(z;\boldsymbol{m})$) are often also essential for decision making purposes once the statistical inverse problem is characterized.

3.5. Efficient computation with adjoint methods. For this problem, efficient computation of gradient and Hessian information is crucial. In this section, we give the expressions derived from deterministic PDE constrained optimization, which we use to perform all derivative computations used in the numerical results. For practical reasons, we make little attempt in this paper to justify the expressions given here but refer the reader to standard references in PDE constrained optimization [6, 26].

Recall that the cost function from deterministic optimization is analogous to the negative log-posterior distribution for this problem. We assume the case of a Gaussian prior on the parameters $\boldsymbol{m}$. The negative log-posterior is written as

$$\begin{aligned} -\log \pi_{\text{post}} = {} & \tfrac{1}{2}(\boldsymbol{f}(\boldsymbol{m}) - \boldsymbol{d}_{\text{obs}})^T \boldsymbol{\Gamma}_{\text{noise}}^{-1}(\boldsymbol{f}(\boldsymbol{m}) - \boldsymbol{d}_{\text{obs}}) \\ & + \tfrac{1}{2}(\boldsymbol{m} - \bar{\boldsymbol{m}}_{\text{prior}})^T \boldsymbol{\Gamma}_{\text{prior}}^{-1}(\boldsymbol{m} - \bar{\boldsymbol{m}}_{\text{prior}}). \end{aligned} \tag{3.8}$$

In the deterministic context, the quantity (3.8) would be minimized as a function of the parameters $\boldsymbol{m}$, subject to the constraint that $u(z,t)$ satisfy the forward model given in the previous section.

The Lagrangian $L(u,p,\boldsymbol{m})$ serves as a tool for solving this constrained minimization problem, where the adjoint solution $p(z,t)$ is introduced as a Lagrange multiplier to enforce the given constraints.

To write the Lagrangian for the constrained optimization problem, we introduce the adjoint solution $p(z,t)$ which plays the role of the Lagrange multiplier. The Lagrangian can now be expressed in weak form in terms of the forward solution u,

the adjoint solution p, and the parameters $\boldsymbol{m}$:

$$
\begin{aligned}
&L(u,p,\boldsymbol{m}) \\
&\quad= \int_{t=0}^{t=T}\int_{z\in\Omega} \frac{1}{2\sigma_{\text{noise}}^2}\sum_{i=1}^{120}(u(z,t)-\boldsymbol{d}_{\text{obs}}^i)^2\delta(z-0)\delta(t-t_i) \\
&\qquad+\frac{1}{2}(\boldsymbol{m}-\bar{\boldsymbol{m}}_{\text{prior}})^T\boldsymbol{\Gamma}_{\text{prior}}^{-1}(\boldsymbol{m}-\bar{\boldsymbol{m}}_{\text{prior}})\mathrm{d}z\mathrm{d}t \\
&\qquad+\int_{t=0}^{t=T}\int_{z\in\Omega} p(z,t)\left[\rho u_{tt}(z,t)-\Big(\mu(z;\boldsymbol{m})\,u_z(z,t)\Big)_z-\delta(z-0)F^{\text{ricker}}(t)\right]\mathrm{d}z\mathrm{d}t \\
&\qquad+\int_{t=0}^{t=T}\Big[p(L,t)\left(\sqrt{\rho\,\mu(L;\boldsymbol{m})}\,u_t(L,t)+\mu(L;\boldsymbol{m})\,u_z(L,t)\right) \\
&\qquad\qquad-p(0,t)\mu(0;\boldsymbol{m})u_z(0,t)\Big]\mathrm{d}t \\
&\qquad+\int_{z\in\Omega}\rho\left[p(z,0)u_t(z,0)-p_t(z,0)u(z,0)\right]\mathrm{d}z \\
&\qquad+\sqrt{\rho\mu(L;\boldsymbol{m})}\,p(L,0)u(L,0).
\end{aligned}
$$

Setting $\delta_p L(u,p,\boldsymbol{m})=0$, we recover the original forward PDE with the proper boundary conditions, given in section 3.1.

Setting $\delta_u L(u,p,\boldsymbol{m})=0$ and carefully integrating by parts, we derive the adjoint PDE and boundary conditions which determine the adjoint solution $p(z,t)$:

(Adj. PDE)
$$\rho\,p_{tt}(z,t)-\Big(\mu(z;\boldsymbol{m})\,p_z(z,t)\Big)_z=-\frac{1}{\sigma_{\text{noise}}^2}\sum_{i=1}^{120}(u(z,t)-\boldsymbol{d}_{\text{obs}}^i)\delta(z-0)\delta(t-t_i),$$

(Absorbing BC) $\quad\mu(L;\boldsymbol{m})\,p_z(L,t)=\sqrt{\rho\,\mu(L;\boldsymbol{m})}\,p_t(L,t),$

(Free Surface BC) $\quad\mu(0;\boldsymbol{m})\,p_z(0,t)=0,$

(FC) $\quad p(z,T)=0,$

(FC) $\quad p_t(z,T)=0.$

The gradient $\boldsymbol{g}=\nabla_{\boldsymbol{m}}L(u,p,\boldsymbol{m})$ is then computed efficiently using the forward $u(z,t)$ and adjoint $p(z,t)$ functions satisfying the forward and adjoint equations, respectively:

$$
\begin{aligned}
\boldsymbol{g}&=\boldsymbol{\Gamma}_{\text{prior}}^{-1}(\boldsymbol{m}-\bar{\boldsymbol{m}}_{\text{prior}}) \\
&\quad+\int_0^T\int_\Omega\big[\nabla_{\boldsymbol{m}}\mu(z;\boldsymbol{m})\big]p_z(z,t)u_z(z,t)\,\mathrm{d}z\mathrm{d}t \\
&\quad+\int_0^T\frac{1}{2}\sqrt{\frac{\rho}{\mu(L;\boldsymbol{m})}}\,\big[\nabla_{\boldsymbol{m}}\mu(L;\boldsymbol{m})\big]p(L,t)u_t(L,t)\mathrm{d}t.
\end{aligned}
\tag{3.9}
$$

We next consider the block form of the full Hessian operator $\boldsymbol{H}=\nabla^2L(u,p,\boldsymbol{m})$, which acts on the *incremental variables* $(\tilde{p},\tilde{u},\tilde{\boldsymbol{m}})$. Since only the parameters $\tilde{\boldsymbol{m}}$ are of

interest, we consider the Schur complement of $\boldsymbol{H}$ with respect to the $\tilde{\boldsymbol{m}}$ block, which amounts to block elimination of the system

$$\begin{pmatrix} H_{pp} & H_{pu} & H_{p\boldsymbol{m}} \\ H_{up} & H_{uu} & H_{u\boldsymbol{m}} \\ H_{\boldsymbol{m}p} & H_{\boldsymbol{m}u} & H_{\boldsymbol{mm}} \end{pmatrix} \begin{pmatrix} \tilde{p} \\ \tilde{u} \\ \tilde{\boldsymbol{m}} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \nabla_{\boldsymbol{m}} \nabla_{\boldsymbol{m}} L(u, p, \boldsymbol{m}) \tilde{\boldsymbol{m}} \end{pmatrix}, \tag{3.10}$$

which implicitly defines the action of the Hessian on $\tilde{\boldsymbol{m}}$. Again, the details for the justification of the above expression are omitted, and we refer the reader to standard references in PDE constrained optimization.

To compute the action of the reduced Hessian operator on a given parameter function $\tilde{\boldsymbol{m}}(z)$, we first solve the *incremental forward equation*, given by row 1 of (3.10):

(Incremental PDE)
$$\rho\, \tilde{u}_{tt}(z,t) - \Big(\mu(z;\boldsymbol{m})\, \tilde{u}_z(z,t)\Big)_z = \Big(\big[\nabla_{\boldsymbol{m}} \mu(z;\boldsymbol{m}) \cdot \tilde{\boldsymbol{m}}\big]\, u_z(z,t)\Big)_z,$$

(Bottom BC)
$$\begin{aligned} &\mu(L;\boldsymbol{m}) \tilde{u}_z(L,t) + \big[\nabla_{\boldsymbol{m}} \mu(L;\boldsymbol{m}) \cdot \tilde{\boldsymbol{m}}\big]\, u_z(L,t) \\ &\quad = -\sqrt{\rho\, \mu(L;\boldsymbol{m})}\, \tilde{u}_t(L,t) - \frac{1}{2} \sqrt{\frac{\rho}{\mu(L;\boldsymbol{m})}}\, \big[\nabla_{\boldsymbol{m}} \mu(L;\boldsymbol{m}) \cdot \tilde{\boldsymbol{m}}\big] u_t(L,t), \end{aligned}$$

(Top BC)
$$\mu(0;\boldsymbol{m})\, \tilde{u}_z(0,t) + \big[\nabla_{\boldsymbol{m}} \mu(0;\boldsymbol{m}) \cdot \tilde{\boldsymbol{m}}\big] u_z(0,t) = 0,$$

(IC)
$$\tilde{u}(z,0) = 0,$$

(IC)
$$\tilde{u}_t(z,0) = 0.$$

Next we solve the *incremental adjoint equation*, given by row 2 of (3.10):

(Incremental Adj. PDE)
$$\begin{aligned} &\rho\, \tilde{p}_{tt}(z,t) - \Big(\mu(z;\boldsymbol{m})\, \tilde{p}_z(z,t)\Big)_z \\ &= \Big(\big[\nabla_{\boldsymbol{m}} \mu(z;\boldsymbol{m}) \cdot \tilde{\boldsymbol{m}}\big]\, p_z(z,t)\Big)_z - \frac{1}{\sigma^2_{\text{noise}}} \sum_{i=1}^{120} \tilde{u} \delta(z-0) \delta(t-t_i), \end{aligned}$$

(Bottom BC)
$$\begin{aligned} &\mu(L;\boldsymbol{m}) \tilde{p}_z(L,t) + \big[\nabla_{\boldsymbol{m}} \mu(L;\boldsymbol{m}) \cdot \tilde{\boldsymbol{m}}\big] p_z(L,t) \\ &\quad = \sqrt{\rho\, \mu(L;\boldsymbol{m})} \tilde{p}_t(L,t) + \frac{1}{2} \sqrt{\frac{\rho}{\mu(L;\boldsymbol{m})}}\, \big[\nabla_{\boldsymbol{m}} \mu(L;\boldsymbol{m}) \cdot \tilde{\boldsymbol{m}}\big] p_t(L,t), \end{aligned}$$

(Top BC)
$$\mu(0;\boldsymbol{m})\, \tilde{p}_z(0,t) + \big[\nabla_{\boldsymbol{m}} \mu(0;\boldsymbol{m}) \cdot \tilde{\boldsymbol{m}}\big] p_z(0,t) = 0,$$

(FC)
$$\tilde{p}(z,T) = 0,$$

(FC)
$$\tilde{p}_t(z,T) = 0.$$

Finally, the *Hessian-vector product* can be computed by row 3 of (3.10), using the solutions of the incremental forward and adjoint equations, respectively:

$$\begin{aligned}
\boldsymbol{H}\tilde{\boldsymbol{m}} &= \boldsymbol{\Gamma}_{\text{prior}}^{-1}\tilde{\boldsymbol{m}} \\
&\quad + \int_0^T \int_\Omega \Big\{ \big[\nabla_{\boldsymbol{m}}\mu(z;\boldsymbol{m})\big] \Big(\tilde{p}_z(z,t)u_z(z,t) + p_z(z,t)\tilde{u}_z(z,t)\Big) \\
&\qquad\qquad + \big[\nabla^2_{\boldsymbol{m}}\mu(z;\boldsymbol{m})\cdot\tilde{\boldsymbol{m}}\big]p_z(z,t)u_z(z,t)\Big\}\mathrm{d}z\mathrm{d}t \\
&\quad + \int_0^T \Big\{ \frac{1}{2}\sqrt{\frac{\rho}{\mu(L;\boldsymbol{m})}}\big[\nabla_{\boldsymbol{m}}\mu(L;\boldsymbol{m})\big]\Big(\tilde{p}(L,t)u_t(L,t) + p(L,t)\tilde{u}_t(L,t)\Big) \\
&\qquad\qquad - \frac{1}{4}\sqrt{\frac{\rho}{\mu(L;\boldsymbol{m})^3}}\,\big[\nabla^T_{\boldsymbol{m}}\mu(L;\boldsymbol{m})\big]\big[\nabla_{\boldsymbol{m}}\mu(L;\boldsymbol{m})\cdot\tilde{\boldsymbol{m}}\big]p(L,t)u_t(L,t) \\
&\qquad\qquad + \frac{1}{2}\sqrt{\frac{\rho}{\mu(L;\boldsymbol{m})}}\big[\nabla^2_{\boldsymbol{m}}\mu(L;\boldsymbol{m})\cdot\tilde{\boldsymbol{m}}\big]p(L,t)u_t(L,t)\Big\}\mathrm{d}t.
\end{aligned}$$

4. Numerical results. The primary goal of this section is to compare the performance of a variety of MCMC methods for the statistical inverse problems outlined in section 3.4. Four separate statistical inverse problems are considered, corresponding to the choices of of the medium parametrizations and associated priors outlined in section 3.3. We call these experiments 2D, 16D, 65D, and 1025D, respectively, according to the number of parameter dimensions contained in the problem.

The observation data $\boldsymbol{d}_{\text{obs}}$ for each of the experiments are synthetically generated using a "ground truth" medium in each case which is drawn from the prior distribution. Synthetic observation data is generated on a mesh different from that used for inversion.

In general the datasets used in each case are unique, with the specific exception that the 1025D problem is intended to be a precise refinement of the 65D problem: The same dataset is used in both experiments, and the initial starting points for the MCMC chains in the 1025D case are linear interpolations of the starting points for the 65D case. In this way, the same underlying *physical* problem is being solved in both cases, so that we may look at the scaling behavior of our method.

For each MCMC method to be compared, 64 MCMC chains are computed using a common set of 64 initial points. These points are selected from a long Stochastic Newton MCMC chain which is initialized at the MAP estimate. From this chain, several initial points are chosen which approximately maximize the minimum pairwise distances between points, so that the resulting set is distributed quasi-uniformly over the region of nonnegligible posterior probability density. In this way the initial points are overdispersed relative to the true posterior probability density (which is important for computation of the MPSRF, later), but they remove potential difficulties in comparing different "burn-in" times for different MCMC methods and make the results more comparable in general. In cases where the MCMC method requires tuning or choice of parameters, several parameter studies were performed to attempt to optimize the performance of the MCMC chain, wherein we choose the parameter(s) which provide the largest mean square jump distance while maintaining an acceptance rate of 30%–50% [38].

Secondary goals are to demonstrate features of this particular physical model which enable the use of Stochastic Newton MCMC and to examine quantities of interest which might be of scientific or engineering relevance.

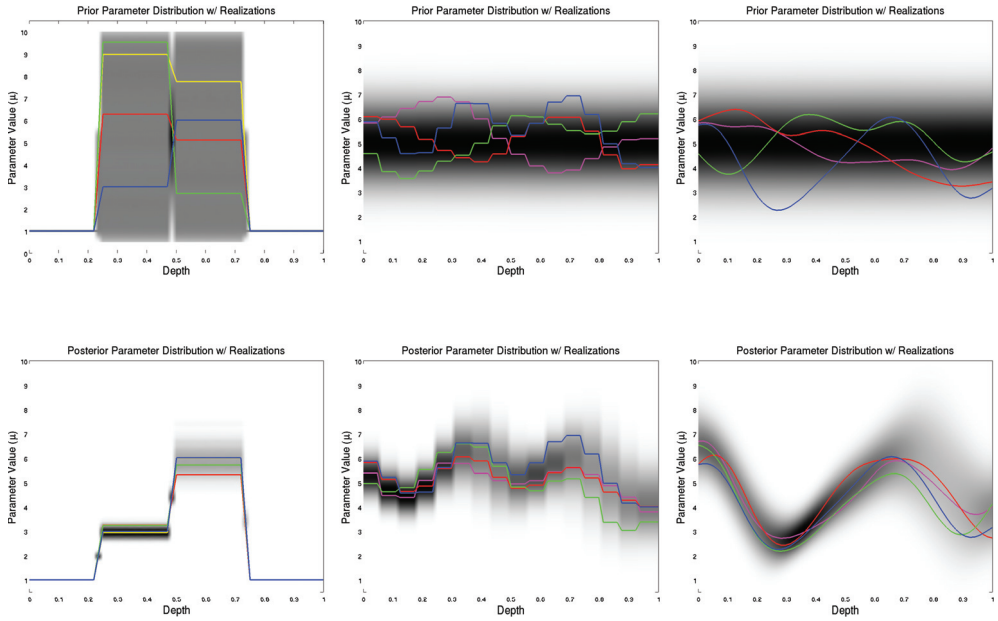

FIG. 2. *Visualizations of the* 1*D marginal prior probability distributions (top row) and posterior probability distributions (bottom row) are shown in gray scale above. Results are shown from left to right for each of the* 2*-,* 16*-, and* 65*D parametrizations of the medium, respectively. A few realizations from each distribution are overlaid to give an indication of the smoothness of the distributions. Parametrizations shown in blue on each of the plots represent the "true" underlying distribution from which the observation data were generated.*

4.1. Visualization of the posterior pdf. Attempting to construct a visualization which depicts the full correlation structure for a 65D object is an impossible task. In this section, we present the most generally informative visualization of the solution to the given statistical inverse problems that we are able to provide.

In Figure 2, we present the marginalized 1D probability distributions as a vertical gray scale stripe for each depth. Regions of darker gray indicate higher certainty that the true curve passes through a given value of the parameter at this depth. The images in the figure are constructed by placing these gray scale stripes side by side for every depth and as such present no indication of the correlation between parameter values at different depths.

To give a hint at the correlation structure, a few representative samples are shown, drawn from the prior PDF or drawn from the posterior PDF MCMC chain, respectively. In all cases, the blue curve represents the ground truth parameters, from which the synthetic observations were generated, and should be expected to pass through the regions of reasonable (or at least nonnegligible) probability.

4.2. MPSRF diagnostic. To compare the different MCMC methods, we employ the multivariate potential scale reduction factor (MPSRF) diagnostic [6]. This diagnostic compares averaged properties of the individual sample chains with properties of the pooled sample chain. When these properties are similar, we infer that each of the individual sample chains has converged.

This idea is made quantitative using the sample chain covariance. One estimate $\boldsymbol{W}$ uses the average of the individual sample chain covariances, which will tend to underestimate the true covariance of the distribution. Second, $\hat{\boldsymbol{V}}$ estimates the pooled

sample chain covariance between all of the chains, and will tend to overestimate the true sample covariance, due to the overdispersion of the initial points. The MPSRF statistic then computes the maximum linear projection of the ratio,

$$\sqrt{\hat{R}} = \max_{\|\boldsymbol{a}\|=1} \left(\frac{\boldsymbol{a}^T \hat{\boldsymbol{V}} \boldsymbol{a}}{\boldsymbol{a}^T \boldsymbol{W} \boldsymbol{a}} \right), \tag{4.1}$$

which overestimates and eventually approaches 1 as the two estimates become more and more similar.

Figure 3 displays the MPSRF curves for each MCMC method on each of the 2D, 16D, and 65D problems. We find in general that Stochastic Newton MCMC is always more efficient on a per sample basis, but the reference MCMC methods DRAM and Langevin are very competitive on the 2D and 16D problems in terms of total computation time, as Stochastic Newton is substantially more expensive. However, the reference methods fail to converge for the 65D problem under the MPSRF convergence diagnostic even in 10 hours of wallclock computation time and order 10^5 samples, while Stochastic Newton still appears to converge.

In Figure 4, we demonstrate scaling of the low-rank Stochastic Newton to large-scale problems by comparing the MPSRF convergence diagnostic for the 65D and 1025D problems plotted against the number of samples computed. If the 65D problem is well resolved, then we anticipate that the 1025D should display similar convergence diagnostics (as a function of number of samples), since it is in principle nothing more than a refinement of the same problem. Furthermore, we have claimed previously that each sample requires a dimension-independent number of PDE solves (depending only on the compact subspace of the Hessian), and therefore the full solution cost for the statistical inverse problem should be only a constant multiple of the cost of a single forward PDE solve, which is independent of the parameter dimension.

4.3. MCMC chain statistics. Table 1 shows the computational time per sample, mean square jump distance, and integrated autocorrelation times for six scalar quantities of interest.

Time per sample. Comparing the computational time required for each MCMC sample (TPS column of Table 1), we see that full Stochastic Newton easily has the highest per sample expense followed by reduced-rank Stochastic Newton. The reference MCMC methods are comparatively inexpensive.

Mean square jump distance. The mean square jump distance (MSJ column of Table 1) can also be used to give an indication of how well the MCMC chain is mixing within the desired posterior probability distribution. This metric is defined for a single MCMC chain with samples $\boldsymbol{m}_0, \ldots, \boldsymbol{m}_N$ as

$$\text{MSJ} := \frac{1}{N} \sum_{k=0}^{N-1} \|\boldsymbol{m}_{k+1} - \boldsymbol{m}_k\|^2. \tag{4.2}$$

The quantity reported in the table is averaged among all 64 parallel chains from a given method. In general, a larger mean square jump distance indicates faster mixing of the MCMC chain and tends to result in better chain convergence to the underlying posterior distribution.

Integrated autocorrelation time. In the 65D case, we also consider the integrated autocorrelation times for six scalar quantities of interest for each MCMC method considered. These quantities are computed for each sample as $\min_z \mu(z; \boldsymbol{m})$, $\max_z \mu(z; \boldsymbol{m})$, $\int_0^L \mu(z; \boldsymbol{m}) \mathrm{d}z$, $\mu(0; \boldsymbol{m})$, $\mu(L/2; \boldsymbol{m})$, and $\mu(L; \boldsymbol{m})$, respectively, and reported in Table 1.

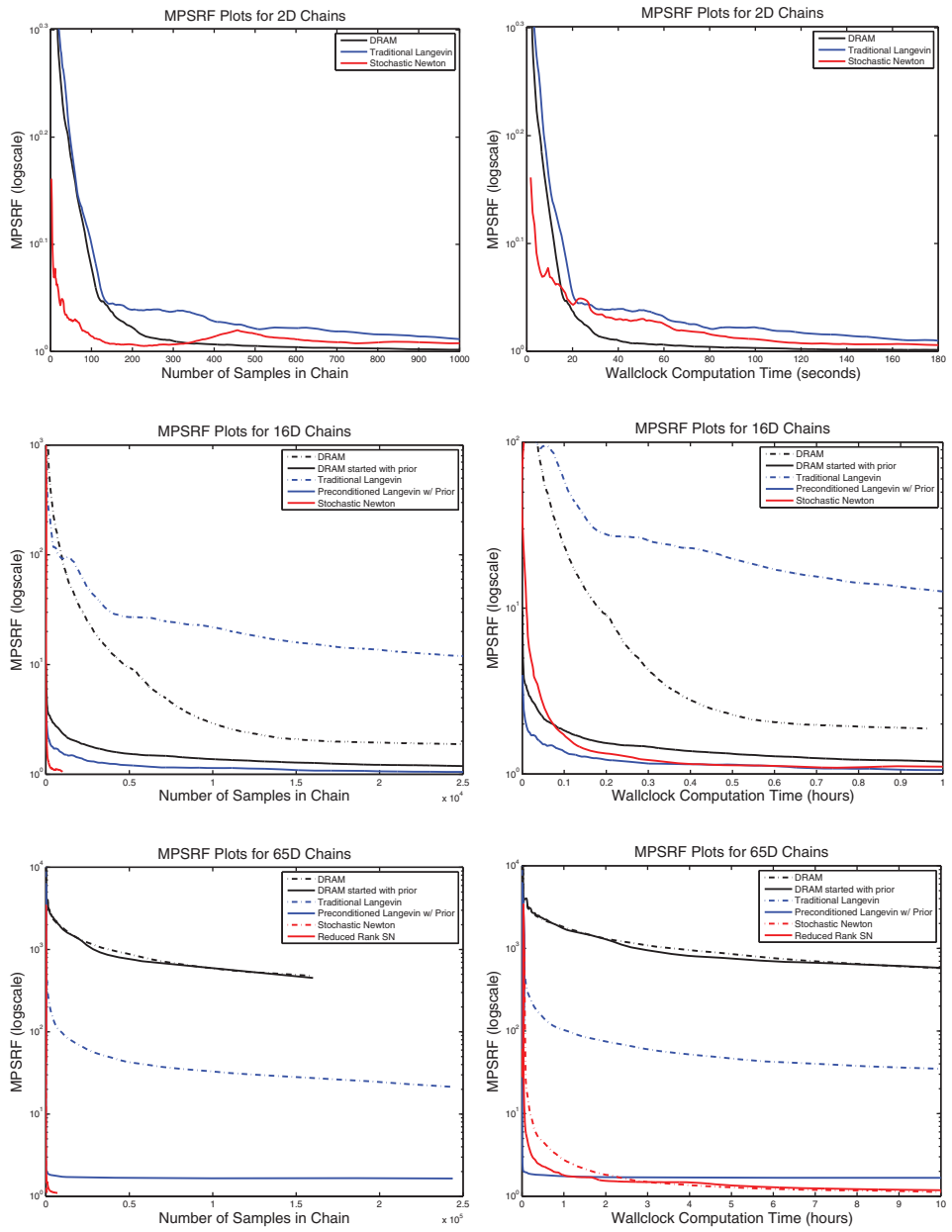

FIG. 3. *The MPSRF statistic is shown on a semilogarithmic plot for Stochastic Newton MCMC (in red) and two reference MCMC methods DRAM (in black) and Langevin (in blue). Color is available only in the online version. As convergence is reached, the MPSRF estimate is expected to decrease to* 1. *The MPSRF is plotted as a function of the number of samples in each of* 64 *parallel MCMC chains (left column) and as a function of the total per-chain wallclock computation time (right column). Stochastic Newton generally requires several PDE solves for each MCMC sample, while the reference methods DRAM and Langevin only require one and two PDE solves, respectively, which accounts precisely for the differences in the left and right columns. Results are shown from top to bottom for each of the* 2-, 16-, *and* 65*D parametrizations of the medium, respectively. In the smaller problems (top two rows), the reference MCMC methods are very competitive with Stochastic Newton. However, in the largest problem (bottom row), the reference MCMC methods fail to even converge in* 10 *hours of wallclock computation time and* $O(10^5)$ *samples under this metric.*

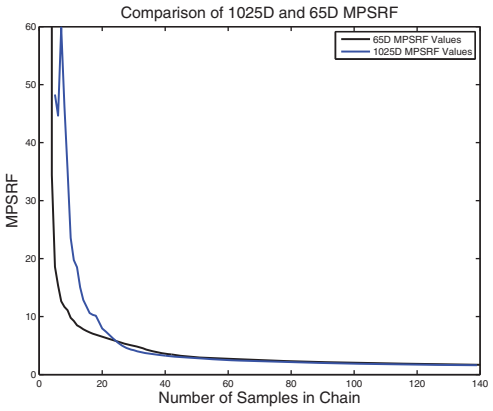

FIG. 4. *The MPSRF statistic for the* 1025*D and* 65*D Stochastic Newton MCMC chains is plotted as a function of the number of samples in each chain. We observe similar convergence rates for both problems in this metric, despite the factor of* 16 *difference in number of parameters. This demonstrates that Stochastic Newton is sensing only the intrinsic difficulty of the problem defined by the compact subspace of the Hessian operator, rather than the full* 65 *or* 1025 *parameter dimensions, which are otherwise fatal for the reference methods.*

TABLE 1

Time per sample in seconds (TPS), mean squared jump distance (MSJ), and integrated autocorrelation time comparison for a variety of MCMC methods. We compare the full-rank Stochastic Newton MCMC (SN), reduced-rank Stochastic Newton (rr SN), Langevin MCMC (L), prior-preconditioned Langevin MCMC (pp L), delayed rejection adaptive metropolis MCMC (DRAM), and prior-initialized DRAM (pi DRAM). Entries for which integrated autocorrelation is not listed are incomputable due to lack of chain convergence. The 65 *parameter experiment is considered for all statistics.*

			Integrated autocorrelation times					
MCMC type	TPS	MSJ	$\min \mu(z)$	$\max \mu(z)$	$\int_0^L \mu(z)\mathrm{d}z$	$\mu(0)$	$\mu(L/2)$	$\mu(L)$
SN	64	6.1	65	124	50	17	52	31
rr SN	16	6.8	85	95	46	37	56	32
L	0.42	3.0e-4	–	–	–	–	–	–
pp L	0.42	5.9	74	114	52	29	51	35
DRAM	0.35	1.2e-5	–	–	–	–	–	–
pi DRAM	0.35	1.2e-5	–	–	–	–	–	–

It is well known in Monte Carlo methods that averaging over N independent and identically distributed samples $\boldsymbol{m}_k \sim \pi_{\text{post}}$ will reduce the variance in the estimate by a factor of $\frac{1}{N}$. However, MCMC samples are most certainly not independent, and in general we observe that averaging over N samples from an MCMC chain will reduce the variance in the estimate by a factor of only $\frac{\tau}{N}$, where $\tau > 1$ is the integrated autocorrelation time [37]. This can be computed as

$$\tau = 1 + 2\sum_{s=1}^{\infty} \rho(s), \tag{4.3}$$

where $\rho(s)$ is the usual autocorrelation function for a lag s. In practice for finite length sample chains, $\rho(s)$ is a noisy function, and we report the maximum value of τ obtained by truncating the summation after any value of $s < 5000$. In some cases, the sum does not converge over the entire length of the sample chain. It is almost certain that these chains are not well converged, and these entries are not reported.

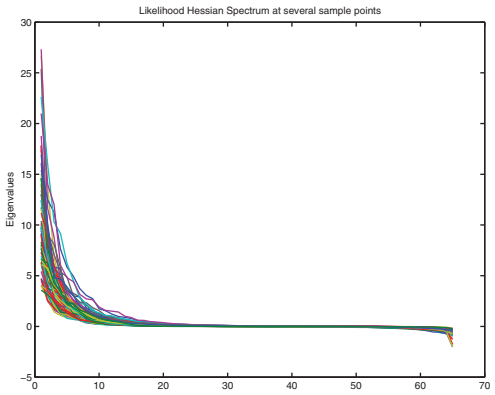

FIG. 5. *Likelihood Hessian spectrum curves, computed at each of* 64 *sample points distributed quasi-uniformly across the region of nonnegligible posterior pdf. The spectrum at every point collapses quickly to zero, as even a single noncompact spectrum would stand out among the rest in this view. We also observe small negative eigenvalues in some spectra, demonstrating nonlinearity and nonconvexity of our forward model.*

We observe similar integrated autocorrelation times for the full- and reduced-rank Stochastic Newton MCMC methods, as well as the prior preconditioned Langevin MCMC, indicating that these methods appear to be mixing well and at comparable rates. In this metric we actually do not observe significant difference between Stochastic Newton and prior preconditioned Langevin MCMC, which is likely due to the smoothing effects in our choice of prior.

4.4. Compactness of the likelihood Hessian. We demonstrate numerically that the Hessian matrix of the likelihood term for this problem is indeed compact, as this is a necessary condition for Stochastic Newton MCMC to be effective. Figure 5 contains spectra for the likelihood Hessian $\boldsymbol{H}_{\text{misfit}} = -\nabla^2 \log \pi_{\text{like}}$ computed at each of the 64 MCMC chain starting points.

The spectra shown do not contain the contribution to the Hessian from the prior term and are not modified (e.g., for positive definiteness) to demonstrate the low-rank character of the underlying physical model.

It is precisely this underlying compact nature of the forward model that enables similar convergence characteristics of the refined 1025D problem as those of the 65D problem, demonstrated in Figure 4.

5. Concluding remarks. We have presented a Stochastic Newton MCMC method that is aimed at ill-posed and large-scale statistical inverse problems. The key idea is to make use of gradient and Hessian information characterizing the posterior pdf. We apply concepts from deterministic optimization, making the connection to the classical Newton method, to efficiently construct a proposal density for MCMC sampling without ever building the full Hessian operators.

We apply the proposed method to a prototypical statistical inverse problem based on a seismological scattering experiment that is governed by a 1D wave equation. Stochastic Newton MCMC and two reference MCMC methods are applied to this problem for a variety of discretizations of the parameter space. When the number of parameters is small, all three methods are comparable in MCMC performance. However, when increasing the dimension of the parameter space, Stochastic Newton shows faster convergence and better mixing of the MCMC chain. Moreover, comparing its

performance for 65D and 1025D parametrizations of the same physical problem, we observe similar MCMC convergence characteristics. While this behavior is not yet provable theoretically, the numerical observations suggest an insensitivity of convergence of Stochastic Newton to the parameter dimension.

We hypothesize that the observed dimension independence of the proposed method (depicted in Figure 4) stems from its ability to detect the subspace of parameters for which the data are informative (and therefore the forward model is active), which is typically small for ill-posed inverse problems governed by PDEs. Once this data-informed subspace is sufficiently well resolved by a given parameter discretization, we anticipate that further parameter refinement does not affect the data misfit term and therefore does not affect the resulting posterior distribution or the low-rank character of the Hessian. We thus expect the number of PDE solves required for Stochastic Newton MCMC to be similarly unaffected as the parameter dimension is increased, enabling this method to be effective for PDE-based statistical inverse problems with high-dimensional parameter spaces.

We are currently investigating robustness of the method with respect to nonlinearity of the forward model and indefiniteness of the Hessian, which both can produce high sample rejection rates.

Acknowledgments. The authors would like to thank Youssef Marzouk, Tan Bui-Thanh, Georg Stadler, Ernesto Prudencio, Todd Oliver, Karl Schulz, and Colin Fox for their many engaging discussions and helpful advice over the course of this work. We thank the referees for their careful reading of the manuscript and their helpful suggestions. The most intensive calculations were performed on the Ranger supercomputer at the Texas Advanced Computing Center.

REFERENCES

[1] V. Akçelik, G. Biros, O. Ghattas, J. Hill, D. Keyes, and B. van Bloeman Waanders, *Parallel algorithms for PDE-constrained optimization*, in Parallel Processing for Scientific Computing, M. A. Heroux, P. Raghaven, and H. D. Simon, eds., SIAM, Philadelphia, 2006, pp. 291–322.

[2] A. Apte, M. Hairer, A. M. Stuart, and J. Voss, *Sampling the posterior: An approach to non-Gaussian data assimilation*, Phys. D, 230 (2007), pp. 50–64.

[3] S. R. Arridge, J. P. Kaipio, V. Kolehmainen, M. Schweiger, E. Somersalo, T. Tarvainen, and M. Vauhkonen, *Approximation errors and model reduction with an application in optical diffusion tomography*, Inverse Problems, 22 (2006), pp. 175–195.

[4] V. Asokan, B. Narayanan, and N. Zabaras, *Stochastic inverse heat conduction using a spectral approach*, Internat. J. Numer. Methods Engrg., 60 (2004), pp. 1569–1593.

[5] G. Biros and O. Ghattas, *Parallel Lagrange–Newton–Krylov–Schur methods for PDE–constrained optimization. Part* II: *The Lagrange–Newton solver and its application to optimal control of steady viscous flows*, SIAM J. Sci. Comput., 27 (2005), pp. 714–739.

[6] S. Brooks and A. Gelman, *General methods for monitoring convergence of iterative simulations*, J. Comput. Graph. Statist., 7 (1998), pp. 434–456.

[7] 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 Proceedings of SC12, Gordon Bell Prize, submitted.

[8] 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.

[9] T. Bui-Thanh, O. Ghattas, and D. Higdon, *Adaptive Hessian-based non-stationary Gaussian process response surface method for probability density approximation with application to Bayesian solution of large-scale inverse problems*, SIAM J. Sci. Comput., submitted.

[10] T. Bui-Thanh, K. Willcox, and O. Ghattas, *Parametric reduced-order models for probabilistic analysis of unsteady aerodynamic applications*, AIAA J., 46 (2008), pp. 2520–2529.

[11] C. Burstedde and O. Ghattas, *Algorithmic strategies for full waveform inversion: 1D experiments*, Geophys., 74 (2009), pp. WCC37–WCC46.

[12] J. A. Christen and C. Fox, *Markov chain Monte Carlo using an approximations*, J. Comput. Graph. Statist., 14 (2005), pp. 795–810.
[13] J. A. Christen and C. Fox, *A general purpose sampling algorithm for continuous distributions (the t-walk)*, Bayesian Anal., 5 (2010), pp. 263–283.
[14] P. Dostert, Y. Efendiev, T. Y. Hou, and W. Luo, *Coarse-gradient Langevin algorithms for dynamic data integration and uncertainty quantification*, J. Comput. Phys., 217 (2006), pp. 123–142.
[15] Y. Efendiev, T. Hou, and W. Luo, *Preconditioning Markov chain Monte Carlo simulations using coarse-scale models*, SIAM J. Sci. Comput., 28 (2006), pp. 776–803.
[16] D. Galbally, K. Fidkowski, K. Willcox, and O. Ghattas, *Nonlinear model reduction for uncertainty quantification in large-scale inverse problems*, Internat. J. Numer. Methods Engrg., 81 (2010), pp. 1581–1608.
[17] J. Geweke and H. Tanizaki, *On Markov chain Monte Carlo methods for nonlinear and non-Gaussian state-space models*, Comm. Statist. Simulation Comput., 28 (1999), pp. 867–894.
[18] J. Geweke and H. Tanizaki, *Note on the sampling distribution for the Metropolis-Hastings algorithm*, Comm. Statist. Theory Methods, 32 (2003), pp. 775–789.
[19] R. G. Ghanem and P. D. Spanos, *Stochastic Finite Elements: A Spectral Approach*, Springer-Verlag, New York, 1991.
[20] M. Girolami and B. Calderhead, *Riemann manifold Langevin and Hamiltonian Monte Carlo methods*, J. R. Stat. Soc. Ser. B Stat. Methodol., 73 (2011), pp. 123–214.
[21] H. Haario, M. Laine, A. Miravete, and E. Saksman, *DRAM: Efficient adaptive MCMC*, Stat. Comput., 16 (2006), pp. 339–354.
[22] K. M. Hanson and G. S. Cunningham, *Posterior sampling with improved efficiency*, in Medical Imaging: Image Processing, SPIE, Bellingham, WA, 1998, pp. 371–382.
[23] D. Higdon, J. Gattiker, B. Williams, and M. Rightley, *Computer model calibration using high-dimensional output*, J. Amer. Statist. Assoc., 103 (2008), pp. 570–583.
[24] D. Higdon, H. Lee, and C. Holloman, *Markov chain Monte Carlo-based approaches for inference in computationally intensive inverse problems*, in Bayesian Statistics 7, J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith, and M. West, eds. Oxford University Press, New York, 2003, pp. 181–197.
[25] D. Higdon, C. S. Reese, J. D. Moulton, J. A. Vrugt, and C. Fox, *Posterior exploration for computationally intensive forward models*, in Handbook of Markov Chain Monte Carlo, Chapman & Hall/CRC Press, Baca Raton, FL, 2010, pp. 401–418.
[26] M. Hinze, R. Pinnau, M. Ulbrich, and S. Ulbrich, *Optimization with PDE Constraints*, Springer-Verlag, New York, 2009.
[27] J. Kaipio and E. Somersalo, *Statistical and Computational Inverse Problems*, Appl. Math. Sci. 160, Springer-Verlag, New York, 2005.
[28] M. C. Kennedy and A. O'Hagan, *Bayesian calibration of computer models*, J. R. Stat. Soc. Ser. B Statist. Methodol., 63 (2001), pp. 425–464.
[29] C. Lieberman, K. Willcox, and O. Ghattas, *Parameter and state model reduction for large-scale statistical inverse problems*, SIAM J. Sci. Comput., 32 (2010), pp. 2523–2542.
[30] Y. Marzouk and D. Xiu, *A stochastic collocation approach to Bayesian inference in inverse problems*, Commun. Comput. Phys., 6 (2009), pp. 826–847.
[31] Y. M. Marzouk and H. N. Najm, *Dimensionality reduction and polynomial chaos acceleration of Bayesian inference in inverse problems*, J. Comput. Phys., 228 (2009), pp. 1862–1902.
[32] Y. M. Marzouk, H. N. Najm, and L. A. Rahn, *Stochastic spectral methods for efficient Bayesian solution of inverse problems*, J. Comput. Phys., 224 (2007), pp. 560–586.
[33] 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.
[34] R. M. Neal, *MCMC using Hamiltonian dynamics*, in Handbook of Markov Chain Monte Carlo, Chapman & Hall/CRC Press, Baca Raton, FL, 2010, pp. 113–162.
[35] N. C. Nguyen, G. Rozza, D. B. P. Huynh, and A. T. Patera, *Reduced basis approximation and a posteriori error estimation for parametrized parabolic PDEs; Application to real-time Bayesian parameter estimation*, in Large-Scale Inverse Problems and Quantification of Uncertainty, L. Biegler, G. Biros, O. Ghattas, M. Heinkenschloss, D. Keyes, B. Mallick, L. Tenorio, B. van Bloemen Waanders, and K. Willcox, eds., John Wiley & Sons, New York, 2011, pp. 151–178.
[36] Y. Qi and T. P. Minka, *Hessian-based Markov chain Monte-Carlo algorithms*, in Proceedings of the First Cape Cod Workshop on Monte Carlo Methods, Cape Cod, MA, 2002.
[37] C. P. Robert and G. Casella, *Monte Carlo Statistical Methods*, Springer Texts Statist., Springer-Verlag, Secaucus, NJ, 2004.

[38] G. O. Roberts and J. S. Rosenthal, *Optimal scaling for various Metropolis-Hastings algorithms*, Statist. Sci., 16 (2001), pp. 351–367.
[39] G. O. Roberts and R. L. Tweedie, *Exponential convergence of Langevin distributions and their discrete approximations*, Bernoulli, 2 (1996), pp. 341–363.
[40] Y. Saad, *Iterative Methods for Sparse Linear Systems*, 2nd ed., SIAM, Philadelphia, 2003.
[41] A. M. Stuart, *Inverse problems: A Bayesian perspective*, Acta Numer., 19 (2010), pp. 451–559.
[42] A. M. Stuart, P. Wiberg, and J. Voss, *Conditional path sampling of SDEs and the Langevin MCMC method*, Commun. Math. Sci., 2 (2004), pp. 685–697.
[43] A. Tarantola, *Inverse Problem Theory and Methods for Model Parameter Estimation*, SIAM, Philadelphia, 2005.
[44] L. Tierney, *Markov chains for exploring posterior distributions*, Ann. Statist., 22 (1994), pp. 1701–1762.
[45] C. R. Vogel, *Computational Methods for Inverse Problems*, Frontiers Appl. Math. 23, SIAM, Philadelphia, 2002.
[46] J. Wang and N. Zabaras, *Using Bayesian statistics in the estimation of heat source in radiation*, Int. J. Heat Mass Tran., 48 (2005), pp. 15–29.
[47] N. Zabaras and B. Ganapathysubramanian, *A scalable framework for the solution of stochastic inverse problems using a sparse grid collocation approach*, J. Comput. Phys., 227 (2008), pp. 4697–4735.