

MIT Open Access Articles

Bayesian Inverse Problems with L_1 Priors: A Randomize-Then-Optimize Approach

The MIT Faculty has made this article openly available. **Please share** how this access benefits you. Your story matters.

Citation: Wang, Zheng, Johnathan M. Bardsley, Antti Solonen, Tiangang Cui, and Youssef M. Marzouk. "Bayesian Inverse Problems with L_1 Priors: A Randomize-Then-Optimize Approach." SIAM Journal on Scientific Computing 39, 5 (January 2017): S140–S166 © 2017 Society for Industrial and Applied Mathematics

As Published: <http://dx.doi.org/10.1137/16M1080938>

Publisher: Society for Industrial & Applied Mathematics (SIAM)

Persistent URL: <http://hdl.handle.net/1721.1/114625>

Version: Author's final manuscript: final author's manuscript post peer review, without publisher's formatting or copy editing

Terms of use: Creative Commons Attribution-Noncommercial-Share Alike



BAYESIAN INVERSE PROBLEMS WITH L_1 PRIORS: A RANDOMIZE-THEN-OPTIMIZE APPROACH

ZHENG WANG*, JOHNATHAN M. BARDSLEY†, ANTTI SOLONEN‡, TIANGANG CUI§, AND YOUSSEF M. MARZOUK*

Abstract. Prior distributions for Bayesian inference that rely on the l_1 -norm of the parameters are of considerable interest, in part because they promote parameter fields with less regularity than Gaussian priors (e.g., discontinuities and blockiness). These l_1 -type priors include the total variation (TV) prior and the Besov space $B_{1,1}^s$ prior, and in general yield non-Gaussian posterior distributions. Sampling from these posteriors is challenging, particularly in the inverse problem setting where the parameter space is high-dimensional and the forward problem may be nonlinear. This paper extends the randomize-then-optimize (RTO) method, an optimization-based sampling algorithm developed for Bayesian inverse problems with Gaussian priors, to inverse problems with l_1 -type priors. We use a variable transformation to convert an l_1 -type prior to a standard Gaussian prior, such that the posterior distribution of the transformed parameters is amenable to Metropolized sampling via RTO. We demonstrate this approach on several deconvolution problems and an elliptic PDE inverse problem, using TV or Besov space $B_{1,1}^s$ priors. Our results show that the transformed RTO algorithm characterizes the correct posterior distribution and can be more efficient than other sampling algorithms. The variable transformation can also be extended to other non-Gaussian priors.

Key words. Inverse problems, Bayesian inference, Monte Carlo methods

AMS subject classifications. 65J22, 62F15, 65C05

1. Introduction. Inverse problems are encountered in many fields of science and engineering—whenever unknown parameters in mathematical models of physical phenomena must be estimated from noisy, incomplete, and indirect measurements. While inverse problems can be solved using a variety of approaches [48], the Bayesian statistical approach [23, 47] is particularly attractive as it offers a coherent framework for quantifying parameter uncertainty, while naturally accommodating different types of data and rich models of prior information.

We begin our discussion of the Bayesian approach to inverse problems by considering a parametric statistical model of the form

$$(1.1) \quad y = f(\theta) + \epsilon,$$

where $y \in \mathbb{R}^m$ is a vector of measurements, $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$ is the forward model (also known as the “parameter-to-observable map”) relating the unknown parameters $\theta \in \mathbb{R}^n$ to the measurements y , and $\epsilon \in \mathbb{R}^m$ is the measurement error. We assume that the error ϵ is a Gaussian random vector with mean zero and covariance matrix $\Gamma_{\text{obs}} \in \mathbb{R}^{m \times m}$, i.e., $\epsilon \sim \mathcal{N}(0, \Gamma_{\text{obs}})$. We will consider both linear and nonlinear forward models f .

Next, define a prior probability density,

$$p(\theta) \propto \exp(-\lambda J(\theta)),$$

that encapsulates all *a priori* information on the parameters θ . Here, $\lambda \in \mathbb{R}$ is a hyperparameter and $J : \mathbb{R}^n \rightarrow \mathbb{R}$ is a prescribed function. Through Bayes’ rule, the prior density and the likelihood function defined by (1.1) together yield the posterior probability density of the parameters θ :

$$(1.2) \quad p(\theta|y) \propto p(y|\theta)p(\theta) \propto \exp\left(-\frac{1}{2}\|f(\theta) - y\|_{\Gamma_{\text{obs}}^{-1}}^2 - \lambda J(\theta)\right).$$

*Department of Aeronautics and Astronautics, MIT, Cambridge, MA 02139 (zheng_w@mit.edu, ymarz@mit.edu.)

†Department of Mathematical Sciences, Montana, University of Montana, Missoula, MT 59812 (bardsleyj@mso.umt.edu.)

‡Department of Mathematics and Physics, Lappeenranta University of Technology, Lappeenranta, Finland (antti.solonen@gmail.com.)

§School of Mathematical Sciences, Monash University, Victoria 3800, Australia (tiangang.cui@monash.edu.)

Solving the inverse problem in the Bayesian setting amounts to characterizing the posterior distribution (1.2), e.g., computing posterior moments or other posterior expectations. A flexible way to do so is via sampling, which has been a topic of research in Bayesian inverse problems for decades (see, e.g., [4, 23, 39, 47]). A widely used class of algorithms for sampling from the posterior is Markov chain Monte Carlo (MCMC); see, e.g., [13, 43, 28, 15, 11] for a general introduction. Most MCMC algorithms build on the Gibbs [12] or general Metropolis-Hastings [34, 20, 17] constructions. For example, [24, 23] implement Gibbs samplers for use on large-scale nonlinear inverse problems, while [19, 18] introduce adaptive Metropolis algorithms that work well on parameter inference problems of small to medium dimension. The need for adaptive algorithms underscores the idea that efficient MCMC sampling requires proposal distributions that capture the local or global structure of the target (posterior) distribution. Accordingly, the Metropolis-adjusted Langevin algorithm (MALA) [45] uses gradients of the target density to guide samples towards regions of higher probability, while [30] approximates local Hessians of the log-target density to construct Gaussian proposals for large-scale problems. Riemannian manifold MCMC [16] may use even higher-order derivative information, along with Hamiltonian Monte Carlo (HMC) [38, 21] proposals. Another issue, particularly relevant to Bayesian inverse problems where θ represents the discretization of a distributed parameter, is that most MCMC algorithms have mixing rates that deteriorate as the discretization is refined [44, 33, 32]. Recent work [6] has introduced Metropolis algorithms with discretization-invariant mixing properties. Dimension-independent likelihood-informed (DILI) samplers then combine discretization invariance with proposals informed by Hessians and other descriptors of the posterior geometry [7]. With the exception of HMC, however, even these relatively sophisticated samplers produce Gaussian proposals at each step. From a computing perspective, we also note that most MCMC algorithms are sequential in nature and may not scale well to massively parallel settings (e.g., via multiple chains) [14].

This paper builds on recent work that explores the potential for *optimization methods* to improve sampling. Broadly, these methods facilitate simulation from non-Gaussian proposal distributions that capture important aspects of posterior structure. Notable examples include randomized maximum likelihood [41], implicit sampling [5, 35], and randomize-then-optimize (RTO) [3]. Our focus in this work is on the RTO approach. RTO uses repeated solutions of a randomly perturbed optimization problem to produce samples from a non-Gaussian distribution, which is used as a Metropolis independence proposal. Although it is more expensive to implement per sample than many simpler Gaussian proposals, it often yields better MCMC mixing. In addition, because the proposals can be generated independently and in parallel, RTO can easily take advantage of large-scale parallel computing environments. However, RTO is only defined for certain classes of problems; in the case of Bayesian inverse problems, it is defined for problems with Gaussian priors and Gaussian measurement error.

The main contribution of this paper is to extend RTO to *non-Gaussian priors*, and to understand the efficiency of the resulting posterior sampling algorithm. We will focus on the case of l_1 -type priors, but the approach can be used on other priors as well. In using l_1 -type priors, we assume that there is a deterministic invertible matrix $D \in \mathbb{R}^{n \times n}$, such that the elements of the vector $D\theta$ are *a priori* independent and endowed with identical Laplace distributions. Thus, the prior is of the form

$$(1.3) \quad p(\theta) \propto \exp \left(-\lambda \sum_{i=1}^n |(D\theta)_i| \right),$$

where $\lambda \in \mathbb{R}$ is a hyperparameter. This choice yields a posterior of the form

$$(1.4) \quad p(\theta|y) \propto \exp \left(-\frac{1}{2} \|f(\theta) - y\|_{\Gamma_{\text{obs}}^{-1}}^2 - \lambda \sum_{i=1}^n |(D\theta)_i| \right).$$

For what is perhaps the most common l_1 -type prior used in Bayesian inverse problems, D is the

discrete one-dimensional derivative (or difference) matrix. This choice yields the total variation (TV) prior, which is related to the well-known regularization functional that penalizes the variation of a signal in order to promote a blocky, discontinuous solution [46, 49]. The TV prior can be derived from the assumption that the increments (i.e., the differences between neighboring parameter node values) are i.i.d. Laplace random variables [2], and it has the form (1.3) only when θ is the discretization of a one-dimensional signal. Another common class of l_1 -type priors are the Besov space $B_{1,1}^s$ priors [26], where D is now a matrix representing a discrete wavelet transform [9]; for the use of Besov priors on large-scale imaging test cases, see [10, 37]. These priors have the advantage that even in two or more dimensions, they retain the form (1.3) and hence the techniques of this paper can be used. Besov priors (with suitable parameters) have been shown to be discretization invariant [26, 8], in that they yield posterior means that converge under mesh refinement.

We extend RTO to the problem of sampling from (1.4) by introducing a multivariate “prior transformation.” This transformation deterministically couples a random variable with an l_1 -type prior to one with a Gaussian prior, and thus enables the use of RTO. A similar transformation for a scalar parameter θ has been suggested in [40]. The present multivariate transformation is more general, however. To the best of our knowledge, it has not been previously proposed, nor has its impact on sampling been investigated. After modifying the RTO algorithm to incorporate the transformation, we conduct a simple comparison of the resulting method with other algorithms, and then focus on numerically exploring the factors that influence its efficiency.

More broadly, variable transformations have been used to improve sampling in [22, 42]. For instance, [42] learns a parameterized multivariate transformation, designed to approximately Gaussianize an arbitrary target distribution, adaptively during MCMC. [22] introduces fixed isotropic (i.e., $\|\theta\|$ -dependent) transformations to obtain target distributions with super-exponentially light tails, so that random-walk Metropolis sampling is geometrically ergodic. In a similar fashion, we use our prior transformation to obtain a posterior distribution to which we can apply RTO. We also describe extensions of our approach to more general priors: first, when any *exact* (e.g., closed-form) coupling between the prior and a standard Gaussian is available, and second, when the prior transformation is only *approximate*. In the latter case, we modify the Metropolis step of our RTO sampler to correct for error in the prior transformation.

The remainder of the paper is organized as follows. We begin in Section 2 with a description of the RTO algorithm [3]. Then, in Section 3, we describe prior transformations that turn (1.4) into a target density amenable to RTO sampling. Finally, in Section 4, we present several numerical examples and comparisons of our method with other MCMC algorithms.

2. Randomize-then-optimize. In the context of Bayesian inverse problems, the randomize-then-optimize (RTO) [3] algorithm can be used to sample from the posterior distribution if the prior distributions on the parameters θ and the measurement error ϵ are both Gaussian. It generates proposal samples through optimization, and then “corrects” these samples using either importance sampling or Metropolis-Hastings. Here, we briefly review the original RTO algorithm; for simplicity, we use notation slightly different from that of [3].

2.1. Form of the target distribution. RTO requires that the target distribution be of a specific form; in particular, it requires that the target density (which for the purposes of this paper is the posterior density of θ) be written as

$$(2.1) \quad p(\theta|y) \propto \exp\left(-\frac{1}{2}\|F(\theta)\|^2\right),$$

where $F(\theta)$ is a vector-valued function of the parameters θ .

Given a Gaussian prior and Gaussian measurement errors, we can, without loss of generality, use linear transformations to “whiten” the prior and the error model so that the inverse problem has the form

$$(2.2) \quad y = f(\theta) + \epsilon, \quad \epsilon \sim N(0, I_m), \quad \theta \sim N(\theta_0, I_n),$$

where $\theta_0 \in \mathbb{R}^n$ is the prior mean; and I_n and I_m are identity matrices of size n and m , respectively. The resulting posterior density is given by

$$p(\theta|y) \propto \exp\left(-\frac{1}{2}\left\|\begin{bmatrix} \theta \\ f(\theta) \end{bmatrix} - \begin{bmatrix} \theta_0 \\ y \end{bmatrix}\right\|^2\right).$$

This density is in the form (2.1), where $F(\theta) = \begin{bmatrix} \theta - \theta_0 \\ f(\theta) - y \end{bmatrix}$ and $F : \mathbb{R}^n \rightarrow \mathbb{R}^{n+m}$.

2.2. The RTO–Metropolis–Hastings algorithm. We now outline how to use RTO to sample from a posterior of the form (2.1). First, a linearization point $\bar{\theta}$ is found and fixed throughout the algorithm. In [3], $\bar{\theta}$ is set to be the posterior mode, though this is not the only possible or useful choice. To obtain the posterior mode, we solve

$$(2.3) \quad \bar{\theta} = \arg \min_{\theta} \frac{1}{2} \|F(\theta)\|^2.$$

Second, the Jacobian of F , which we denote as J_F , is evaluated at $\bar{\theta}$, and an orthonormal basis $\bar{Q} \in \mathbb{R}^{(m+n) \times n}$ for the column space of $J_F(\bar{\theta})$, which we denote as $\text{col}(J_F(\bar{\theta}))$, is computed through a thin-QR factorization of $J_F(\bar{\theta})$. Third, independent samples $\xi^{(i)}$ are drawn from an n -dimensional standard Gaussian, and proposal points $\theta_{\text{prop}}^{(i)}$ are found by solving the optimization problem

$$(2.4) \quad \theta_{\text{prop}}^{(i)} = \arg \min_{\theta} \frac{1}{2} \left\| \bar{Q}^\top F(\theta) - \xi^{(i)} \right\|^2$$

for each sample $\xi^{(i)}$. Under conditions described in [3] and listed in Assumption B.1, the points $\theta_{\text{prop}}^{(i)}$ are distributed according to the proposal density

$$(2.5) \quad q(\theta_{\text{prop}}) = (2\pi)^{-\frac{n}{2}} \left| \bar{Q}^\top J_F(\theta_{\text{prop}}) \right| \exp\left(-\frac{1}{2} \left\| \bar{Q}^\top F(\theta_{\text{prop}}) \right\|^2\right),$$

where $|\cdot|$ denotes the absolute value of the matrix determinant. We focus on using this distribution as an independence proposal in Metropolis–Hastings, though it can also be used in importance sampling. The Metropolis–Hastings acceptance ratio, for a move from a point $\theta^{(i-1)}$ to the proposed point $\theta_{\text{prop}}^{(i)}$, is

$$\frac{p(\theta_{\text{prop}}^{(i)}|y)q(\theta^{(i-1)})}{p(\theta^{(i-1)}|y)q(\theta_{\text{prop}}^{(i)})} = \frac{w(\theta_{\text{prop}}^{(i)})}{w(\theta^{(i-1)})},$$

where $w(\theta)$ are

$$(2.6) \quad w(\theta) := \left| \bar{Q}^\top J_F(\theta) \right|^{-1} \exp\left(-\frac{1}{2} \|F(\theta)\|^2 + \frac{1}{2} \left\| \bar{Q}^\top F(\theta) \right\|^2\right).$$

The resulting MCMC method, which we call RTO–Metropolis–Hastings (RTO-MH), is summarized in Algorithm 2.1.

Remark 2.1. Other choices for the matrix \bar{Q} used in (2.4) and (2.5) are possible, provided that Assumption B.1, which leads to the sampling density $q(\theta)$ in (2.5), is satisfied. Also, in the computation of the Metropolis acceptance ratio, one can use a factorization of $J_F(\theta)$ or $\bar{Q}^\top J_F(\theta)$ and take advantage of properties of the log function; e.g., if $Q_\theta R_\theta = \bar{Q}^\top J_F(\theta)$ is the QR factorization of $\bar{Q}^\top J_F(\theta)$, then

$$\log \left| \bar{Q}^\top J_F(\theta) \right| = \sum_{i=1}^n \log[R_\theta]_{ii}.$$

Algorithm 2.1 RTO-MH

```
1: Find  $\bar{\theta}$  (e.g., the posterior mode) using (2.3)
2: Determine  $J_F(\bar{\theta})$ , the Jacobian of  $F$  at  $\bar{\theta}$ 
3: Compute  $\bar{Q}$ , whose columns are an orthonormal basis for  $\text{col}(J_F(\bar{\theta}))$ 
4: for  $i = 1, \dots, n_{\text{samps}}$  do in parallel
5:   Sample  $\xi^{(i)}$  from a standard  $n$ -dimensional Gaussian
6:   Solve for a proposal sample  $\theta_{\text{prop}}^{(i)}$  using (2.4)
7:   Compute  $w(\theta_{\text{prop}}^{(i)})$  from (2.6)
8: end for
9: Set  $\theta^{(0)} = \bar{\theta}$ 
10: for  $i = 1, \dots, n_{\text{samps}}$  do in series
11:   Sample  $v$  from a uniform distribution on  $[0,1]$ 
12:   if  $v < w(\theta_{\text{prop}}^{(i)}) / w(\theta^{(i-1)})$  then
13:      $\theta^{(i)} = \theta_{\text{prop}}^{(i)}$ 
14:   else
15:      $\theta^{(i)} = \theta^{(i-1)}$ 
16:   end if
17: end for
```

3. RTO-MH with a prior transformation. The previous section showed how Bayesian inverse problems with Gaussian priors and Gaussian measurement errors yield posterior densities that can be written in the form (2.1), as required by RTO. Now we propose a technique that uses RTO to sample from a posterior resulting from a Gaussian measurement model and a *non-Gaussian* prior. This is accomplished via a change of variables that transforms the non-Gaussian prior defined on the physical parameter $\theta \in \mathbb{R}^n$ to a standard Gaussian prior defined on a reference parameter $u \in \mathbb{R}^n$. The caveat is that the transformed forward model, now viewed as a function of u , is the original forward model composed with the nonlinear mapping function, and hence the transformation adds complexity to f .

3.1. Transformations for l_1 -type priors. In the following subsections, we exemplify this approach for l_1 -type priors. First, we describe the transformation of single parameter endowed with a Laplace prior (Section 3.1.1). We then extend that example to construct a transformation of multiple parameters for any l_1 -type prior (Section 3.1.2). Finally, we discuss general prior transformations and summarize the algorithm for performing RTO with a prior transformation (Section 3.2).

3.1.1. Single parameter with a Laplace prior. In this subsection, we consider an inverse problem of the form (1.1) but with only a single parameter and a single observation, $n = m = 1$:

$$y = f(\theta) + \epsilon, \quad \epsilon \sim N(0, \sigma_{\text{obs}}^2),$$

where $\sigma_{\text{obs}} \in \mathbb{R}^+$ is the standard deviation of the error. Instead of a Gaussian prior on θ , we use a Laplace prior

$$p(\theta) \propto \exp(-\lambda|\theta|).$$

Then, the posterior has the form

$$(3.1) \quad p(\theta|y) \propto \exp\left(-\frac{1}{2}\left(\frac{f(\theta) - y}{\sigma_{\text{obs}}}\right)^2 - \lambda|\theta|\right).$$

Due to the Laplace prior, $p(\theta|y)$ cannot directly be written in the form (2.1).

Let us construct an invertible mapping function $g_{1D} : \mathbb{R} \rightarrow \mathbb{R}$ that relates a Gaussian reference random variable $u \in \mathbb{R}$ to the Laplace-distributed physical parameter $\theta \in \mathbb{R}$, such that $\theta = g_{1D}(u)$. A monotone transformation that achieves this goal is

$$(3.2) \quad g_{1D}(u) \equiv \mathcal{L}^{-1}(\varphi(u)) = -\frac{1}{\lambda} \text{sign}(u) \log \left(1 - 2 \left| \varphi(u) - \frac{1}{2} \right| \right),$$

where \mathcal{L} is the cumulative distribution function (cdf) of the Laplace distribution and φ is the cdf of the standard Gaussian distribution. To prove that the reference random variable is in fact standard Gaussian, we calculate its cdf as:

$$\begin{aligned} \mathbb{P}(u < u_0) &= \mathbb{P}(g_{1D}^{-1}(\theta) < u_0) = \mathbb{P}(\theta < g_{1D}(u_0)) \\ &= \mathcal{L}(g_{1D}(u_0)) = \mathcal{L} \circ \mathcal{L}^{-1} \circ \varphi(u_0) = \varphi(u_0). \end{aligned}$$

Hence, this mapping function indeed transforms a standard Gaussian reference random variable u to the Laplace-distributed parameter θ , and thus $p(u) \propto \exp(-\frac{1}{2}u^2)$.

The mapping function g_{1D} and its derivative are depicted in Figure 3.1. The function is monotone, bijective, and continuously differentiable. Its derivative is

$$g'_{1D}(u) = \frac{\varphi'(u)}{\lambda \varphi(-|u|)},$$

where $\varphi'(u) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}u^2)$ is the probability density function of the standard Gaussian distribution.

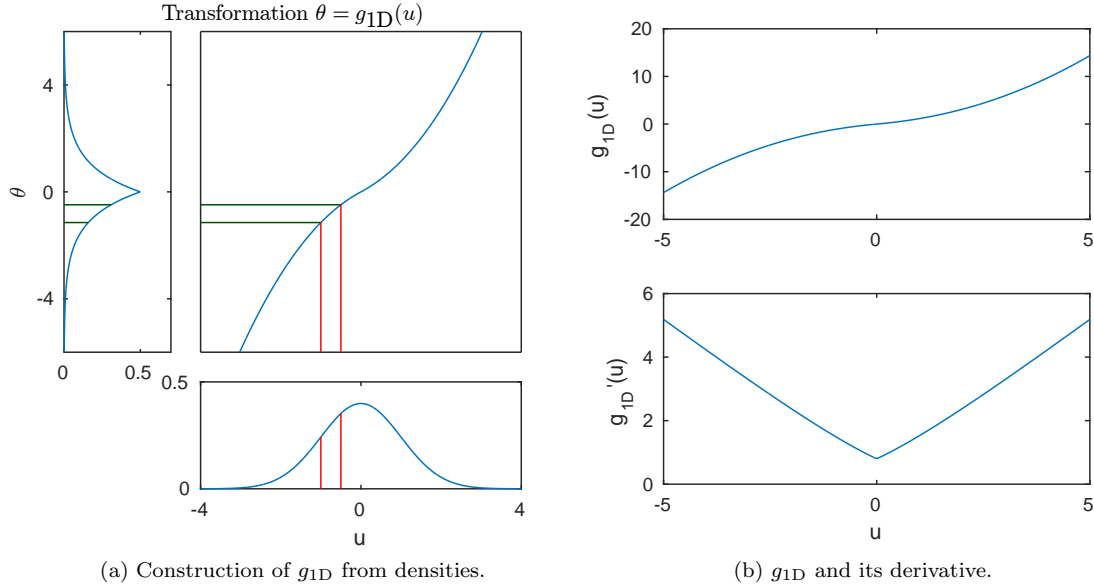


FIG. 3.1. Left: transformation from the standard Gaussian to a Laplace distribution (with $\lambda = 1$). The probability mass between the two vertical lines is equal to that between the horizontal lines. Right: Mapping function g_{1D} and its derivative. The mapping function is continuously differentiable.

Now we can solve Bayesian inverse problems on u and transform the posterior samples of u to posterior samples of θ using the mapping function. The form of the *transformed* posterior density, i.e., the posterior density of u , is given in the following lemma and proven in Appendix A.

LEMMA 3.1. Let (3.1) specify the posterior density of a Bayesian inference problem with parameter $\theta \in \mathbb{R}$. Given the variable transformation $\theta = g_{1D}(u)$ defined in (3.2), the posterior density of u has the form:

$$(3.3) \quad p(u|y) \propto \exp \left(-\frac{1}{2} \left(\frac{f \circ g_{1D}(u) - y}{\sigma_{obs}} \right)^2 - \frac{1}{2} u^2 \right).$$

After the transformation, the prior over the new variables simplifies to a standard Gaussian, and the forward model becomes more complex. In particular, the transformed forward model is the original forward model composed with the nonlinear mapping. The new posterior appears with a Gaussian prior and observational noise, and can be cast in the form (2.1). The resulting structure allows us to use RTO.

3.1.2. Multiple parameters with an l_1 prior. Now we build on the previous section in order to construct a prior transformation for a multivariate l_1 -type prior. Starting from an inverse problem of the form (1.1), we allow for multiple unknown parameters, $n \geq 1$, and multiple observations, $m \geq 1$. We impose the following l_1 -type prior on θ :

$$p(\theta) \propto \exp \left(-\lambda \|D\theta\|_1 \right) = \exp \left(-\lambda \sum_{i=1}^n |(D\theta)_i| \right),$$

where $D \in \mathbb{R}^{m \times n}$ is an invertible matrix and $(D\theta)_i$ denotes the i th element of vector $D\theta$. The posterior on θ is then

$$(3.4) \quad p(\theta|y) \propto \exp \left[-\frac{1}{2} (f(\theta) - y)^\top \Gamma_{obs}^{-1} (f(\theta) - y) \right] \exp \left(-\lambda \|D\theta\|_1 \right).$$

Reference random variables that are *a priori* i.i.d. Gaussian can be transformed to each Laplace-distributed element of $D\theta$ using the one-dimensional transformation g_{1D} defined in (3.2). Then, $D\theta = g(u)$, where $g : \mathbb{R}^n \rightarrow \mathbb{R}^m$ and

$$g(u) := [g_{1D}(u_1), \dots, g_{1D}(u_n)]^\top.$$

Thus, a prior transformation for the l_1 -type prior is

$$(3.5) \quad \theta = D^{-1}g(u),$$

resulting in the requirement that D be invertible. Then, the Jacobian of the transformation is $D^{-1}J_g$, where $J_g : \mathbb{R}^n \rightarrow \mathbb{R}^{m \times n}$ is the Jacobian of g given by

$$(3.6) \quad J_g(u) = \begin{bmatrix} g'_{1D}(u_1) & & & \\ & g'_{1D}(u_2) & & \\ & & \ddots & \\ & & & g'_{1D}(u_n) \end{bmatrix},$$

and g_{1D} is defined in (3.2).

Using this transformation, one can derive the posterior density over u by following the same steps as in the single variable case, with $D^{-1}g(u)$ in place of $g_{1D}(u)$, to obtain

$$p(u|y) \propto \exp \left(-\frac{1}{2} \left(f(D^{-1}g(u)) - y \right)^\top \Gamma_{obs}^{-1} \left(f(D^{-1}g(u)) - y \right) - \frac{1}{2} u^\top u \right).$$

The transformed posterior is in the form (2.1) and is amenable to RTO sampling. Figure 3.2 illustrates the effect of the transformation on an inverse problem with two unknown parameters, $D = I_2$, and a linear forward model; comparing the second and third columns, we note that the transformed prior becomes a standard Gaussian, while the transformed likelihood becomes non-Gaussian.

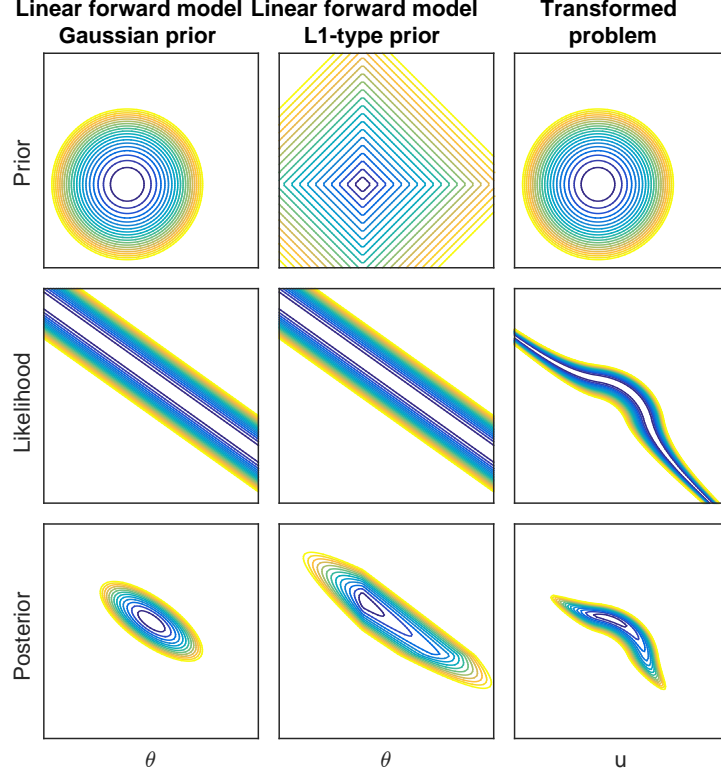


FIG. 3.2. Densities in illustrative inverse problems with two parameters. The plots depict the log-prior density, log-likelihood function, and log-posterior density. The three cases shown are a Gaussian prior with a linear forward model (left), l_1 -type prior with the same forward model (middle), and transformed l_1 -type prior with transformed likelihood (right). The transformation changes the prior to a Gaussian and makes the likelihood more complex. The rightmost posterior is smooth and can be written in the form (2.1).

3.2. RTO-MH with a general prior transformation. Given an inverse problem in the form (1.1) with a general non-Gaussian prior supported on \mathbb{R}^n , suppose that we can construct an invertible and continuously differentiable prior transformation $T: \mathbb{R}^n \rightarrow \mathbb{R}^n$ that couples a standard Gaussian random variable u to our non-Gaussian random variable θ . Both g_{1D} in (3.2) and $D^{-1}g$ in (3.5) are examples of such transformations T . Then the transformed posterior density is

$$\begin{aligned}
 (3.7) \quad p(u|y) &\propto \exp \left(-\frac{1}{2} \left(f \circ T(u) - y \right)^\top \Gamma_{\text{obs}}^{-1} \left(f \circ T(u) - y \right) - \frac{1}{2} u^\top u \right) \\
 &= \exp \left(-\frac{1}{2} \left\| \tilde{f}(u) - \tilde{y} \right\|^2 - \frac{1}{2} \|u\|^2 \right) \\
 &:= \exp \left(-\frac{1}{2} \left\| \tilde{F}(u) \right\|^2 \right),
 \end{aligned}$$

where $\tilde{f}(u) = \Gamma_{\text{obs}}^{-1/2} f \circ T(u)$ is the transformed forward model, $\tilde{y} = \Gamma_{\text{obs}}^{-1/2} y$ is the whitened data, and $\tilde{F}(u) = \begin{bmatrix} u \\ \tilde{f}(u) - \tilde{y} \end{bmatrix}$. We can use RTO to sample from the transformed posterior defined by (3.7).

To perform the optimization steps in RTO and to evaluate the proposal density of RTO, we

need the Jacobian of \tilde{F} , which has the form

$$(3.8) \quad J_{\tilde{F}}(u) = \begin{bmatrix} I \\ J_{\tilde{f}}(u) \end{bmatrix}.$$

Here, $J_{\tilde{f}}(u)$ is the Jacobian of the transformed forward model \tilde{f} and is given by

$$(3.9) \quad J_{\tilde{f}}(u) = \Gamma_{\text{obs}}^{-1/2} J_f(T(u)) J_T(u),$$

where $J_f : \mathbb{R}^n \rightarrow \mathbb{R}^{m \times n}$ is the Jacobian of the original forward model f and $J_T : \mathbb{R}^n \rightarrow \mathbb{R}^{n \times n}$ is the Jacobian of the prior transformation T . The final algorithm, incorporating a prior transformation in RTO-MH, is summarized in Algorithm 3.1.

Algorithm 3.1 RTO-MH with a Prior Transformation

- 1: Determine the prior mapping function $T : \mathbb{R}^n \rightarrow \mathbb{R}^n$ such that $u = T^{-1}(\theta)$ has a standard Gaussian distribution
 - 2: Find the mode $\bar{u} \in \mathbb{R}^n$ of transformed posterior density $p(u|y)$ defined by (3.7)
 - 3: Calculate $\bar{Q} \in \mathbb{R}^{n \times m+n}$, whose columns are an orthonormal basis for the column space of $J_{\tilde{F}}(\bar{u})$, as defined in (3.8)–(3.9)
 - 4: **for** $i = 1, \dots, n_{\text{samps}}$ **do** in parallel
 - 5: Draw a standard Gaussian sample $\xi^{(i)} \sim \mathcal{N}(0, I_n)$
 - 6: Compute RTO samples via $u_{\text{prop}}^{(i)} = \arg \min_u \|\bar{Q}^\top \tilde{F}(u) - \xi^{(i)}\|^2$ and weights
 $w(u_{\text{prop}}^{(i)}) = |\bar{Q}^\top J_{\tilde{F}}(u_{\text{prop}}^{(i)})|^{-1} \exp\left(-\frac{1}{2} \|\tilde{F}(u_{\text{prop}}^{(i)})\|^2 + \frac{1}{2} \|\bar{Q}^\top \tilde{F}(u_{\text{prop}}^{(i)})\|^2\right)$
 - 7: **end for**
 - 8: **for** $i = 1, \dots, n_{\text{samps}}$ **do** in series
 - 9: Sample v from a uniform distribution on $[0, 1]$
 - 10: **if** $v < w(u_{\text{prop}}^{(i)}) / w(u^{(i-1)})$ **then**
 - 11: $u^{(i)} = u_{\text{prop}}^{(i)}$
 - 12: **else**
 - 13: $u^{(i)} = u^{(i-1)}$
 - 14: **end if**
 - 15: **end for**
 - 16: **for** $i = 1, \dots, n_{\text{samps}}$ **do** in parallel
 - 17: Define $\theta^{(i)} = T(u^{(i)})$, the desired samples from $p(\theta|y)$
 - 18: **end for**
-

The computational cost of Algorithm 3.1 is dominated by that of Step 6, where repeated optimization problems are solved and the weights are calculated. Typically, within each optimization iteration, \tilde{f} is evaluated once and $J_{\tilde{f}}$ is applied to multiple vectors; after optimization, the weight $w(u_{\text{prop}}^{(i)})$ must be evaluated, which requires an evaluation of $J_{\tilde{f}}(u_{\text{prop}}^{(i)})$ and an $\mathcal{O}(n^3)$ computation of the log-determinant.

Under certain conditions on \tilde{F} , given in Assumption B.1 (substituting \tilde{F} for F), the samples u_{prop} generated by Steps 1–7 of Algorithm 3.1 are i.i.d. draws from the following probability density:

$$(3.10) \quad q(u_{\text{prop}}) = (2\pi)^{-\frac{n}{2}} \left| \bar{Q}^\top J_{\tilde{F}}(u_{\text{prop}}) \right| \exp\left(-\frac{1}{2} \|\bar{Q}^\top \tilde{F}(u_{\text{prop}})\|^2\right).$$

When the original forward model is linear, i.e., $f(\theta) = A\theta$, and the prior transformation in Section 3.1.2 is applied, the transformed problem automatically satisfies these conditions. This result is stated in the following theorem and proven in Appendix B.

THEOREM 3.2. *Let (3.4) specify the posterior density of a Bayesian inference problem with parameters $\theta \in \mathbb{R}^n$, and let the forward model in (3.4) be linear, $f(\theta) = A\theta$. After the prior transformation (3.5), the RTO algorithm described by Steps 1–7 of Algorithm 3.1 generates proposal samples with probability density given in (3.10).*

The proof of the theorem simply checks that the transformed problem satisfies the assumptions under which the RTO proposal density holds. For nonlinear forward models f , we leave these conditions as an assumption.

3.3. RTO-MH with an approximate prior transformation. The previous section addressed cases where an *exact* prior transformation T is known—i.e., where, if θ is distributed according to the prior, then $T^{-1}(\theta)$ has a standard Gaussian distribution. In some cases, determining such an exact transformation might not be feasible. Nonetheless we can still use approximate transformations—that is, transformations which only approximately “Gaussianize” the prior—to construct an RTO-MH algorithm.

Consider a transformation $\hat{T} : \mathbb{R}^n \rightarrow \mathbb{R}^n$ that couples a reference random variable u to our prior-distributed random variable θ . But now suppose that the distribution of the reference $u = \hat{T}^{-1}(\theta)$ is only *approximately* Gaussian. (To be clear, the expressions below will not require any Gaussian assumption on u ; the degree to which u departs from a standard Gaussian will affect the efficiency, not the correctness, of the following Metropolis-Hastings scheme.) These transformations can often be constructed numerically. For example, [36, 42, 31] describe how to construct parameterized maps from samples or unnormalized density evaluations of any atomless distribution. We can modify our method to work for approximate prior transformations such as these.

As with the exact map, let \hat{T} be invertible and continuously differentiable. We can apply the usual RTO procedure to obtain proposal samples $u_{\text{prop}}^{(i)}$ by solving

$$u_{\text{prop}}^{(i)} = \arg \min_u \left\| \bar{Q}^\top \begin{bmatrix} u \\ \Gamma_{\text{obs}}^{-\frac{1}{2}}(f \circ \hat{T}(u) - y) \end{bmatrix} - \xi^{(i)} \right\|^2$$

for Gaussian samples $\xi^{(i)}$. The proposed samples (in the reference space) will be distributed according to the density

$$\hat{q}(u_{\text{prop}}) = (2\pi)^{-\frac{n}{2}} \left| \bar{Q}^\top J_{\hat{F}}(u_{\text{prop}}) \right| \exp \left(-\frac{1}{2} \left\| \bar{Q}^\top \hat{F}(u_{\text{prop}}) \right\|^2 \right)$$

where

$$\hat{F}(u_{\text{prop}}) = \begin{bmatrix} u_{\text{prop}} \\ \Gamma_{\text{obs}}^{-\frac{1}{2}}(f \circ \hat{T}(u_{\text{prop}}) - y) \end{bmatrix}.$$

In order to obtain samples from the posterior, the RTO-MH algorithm must be modified to incorporate the density of the pullback of the true posterior under the map \hat{T} , which has the form

$$(3.11) \quad p(u|y) \propto \exp \left(-\frac{1}{2} \left\| \hat{f}(u) - \tilde{y} \right\|^2 \right) |J_{\hat{T}}(u)| p_\theta(\hat{T}(u)),$$

where $p_\theta(\cdot)$ is the prior density on θ , $\hat{f}(u) = \Gamma_{\text{obs}}^{-1} f \circ \hat{T}(u)$, and $|J_{\hat{T}}(\cdot)|$ is the Jacobian determinant of \hat{T} . Contrast (3.11) with (3.7); the key difference is that the standard Gaussian prior on u has been replaced with the pullback of p_θ under the map \hat{T} . If the prior transformation \hat{T} were exact, these two expressions would be equivalent. This process gives an altered formula for the weights in Step 6 of Algorithm 3.1:

$$w(u_{\text{prop}}^{(i)}) = \left| \bar{Q}^\top J_{\hat{F}}(u_{\text{prop}}^{(i)}) \right|^{-1} \exp \left(-\frac{1}{2} \left\| \hat{f}(u_{\text{prop}}^{(i)}) - \tilde{y} \right\|^2 + \frac{1}{2} \left\| \bar{Q}^\top \hat{F}(u_{\text{prop}}^{(i)}) \right\|^2 \right) |J_{\hat{T}}(u_{\text{prop}}^{(i)})| p_\theta(\hat{T}(u_{\text{prop}}^{(i)})).$$

The rest of the algorithm remains unchanged. In essence, the error in the approximate prior transformation is handled by appropriately altering the Metropolis-Hastings acceptance ratio.

4. Numerical examples. We apply RTO-MH with prior transformations to three numerical examples, labeled A, B, and C, all with l_1 -type priors. Examples A and B are (spatially) 1-D deconvolution problems with linear forward models, while Example C is a (spatially) 2-D inverse problem with a nonlinear forward model. In Example A, we use a TV prior and perform a simple comparison of the efficiency of our method with that of other MCMC samplers, including the Gibbs scheme proposed in [29] for linear inverse problems with l_1 -type priors. In Example B, we use a Besov $B_{1,1}^s$ space prior and examine the effects of parameter dimension n and hyperparameter λ on the performance of RTO. Finally, in Example C, we infer the coefficient field of a linear elliptic PDE; in this case, we use the 2-D Besov $B_{1,1}^s$ space prior. This example is meant to test RTO on a more difficult inverse problem, involving a nonlinear forward model and a parameter field in two spatial dimensions.

4.1. One-dimensional deconvolution problems. Examples A and B involve the deconvolution of a 1-D signal. We discretize a true signal, $\theta_{\text{true}}(x)$, defined on the domain $x \in [0, 1]$, using n grid points. The true signal is convolved with the function

$$(4.1) \quad k(x) = \begin{cases} 1 & \text{if } -\frac{1}{64} < x < \frac{1}{64} \\ 0 & \text{otherwise,} \end{cases}$$

and evaluated at $m = 30$ points to create measurements corresponding to integrals over interior segments of the domain. The data $y \in \mathbb{R}^m$ are generated by adding i.i.d. Gaussian noise with $\Gamma_{\text{obs}} = \sigma_{\text{obs}}^2 I$.

4.1.1. Example A: TV prior. In this example, the true signal is the square pulse,

$$\theta_{\text{true}}(x) = \begin{cases} 1 & \text{if } \frac{1}{3} < x < \frac{2}{3} \\ 0 & \text{otherwise} \end{cases},$$

which is also used in [27, 29]. Figure 4.1 depicts the true signal and the resulting data.

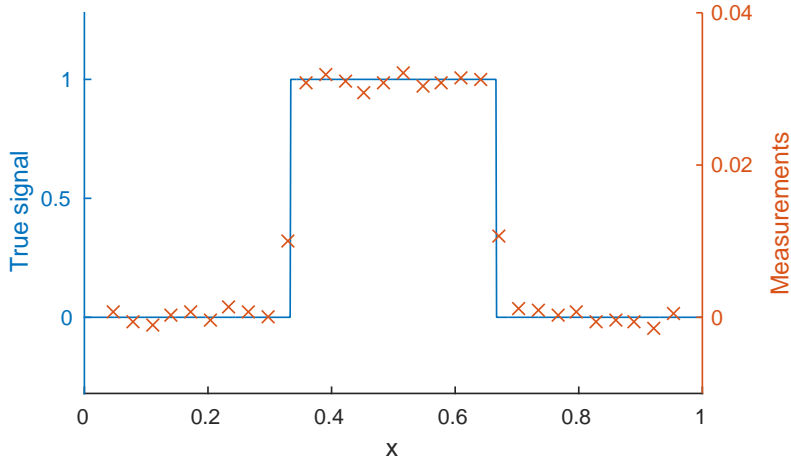


FIG. 4.1. The true signal and noisy measurements for example A.

We use a TV prior, i.e., $\pi(\theta) \propto \exp(-\lambda \|D\theta\|_1)$, with $\theta \in \mathbb{R}^n$, $n = 63$,

$$D = \begin{bmatrix} 1 & & & 1 \\ -1 & 1 & & \\ & \ddots & \ddots & \\ & & -1 & 1 \end{bmatrix}_{n \times n},$$

$\sigma_{\text{obs}} = 1 \cdot 10^{-3}$, and $\lambda = 8$. The first row of D imposes the condition that the sum of the boundary values is zero, making D invertible, which is required for the prior transformation to be well-defined.

We generate MCMC chains using three different algorithms: RTO-MH with a prior transformation, MALA, and the Gibbs scheme of [29]. To compare computational costs, we count the number of function evaluations used by each algorithm, with a Jacobian evaluation (used by MALA and RTO) counted as a single function evaluation for this linear problem. For each algorithm, we stopped the MCMC chain once the number of evaluations reached $4 \cdot 10^6$. The resulting MCMC chains are shown in Figure 4.2. For RTO, we used the default settings of the nonlinear least-squares solver `lsqnonlin` in MATLAB to perform all the optimizations. Our first attempt at MALA used the adaptive (AMALA) scheme of [1]. The resulting chain did not reach stationarity after $4 \cdot 10^6$ evaluations, as seen in Figure 4.2. Note that the vertical axis of the figure showing the AMALA chain is different from the others; the chain has not even located the region of high posterior probability. Instead, to obtain a convergent solution using MALA, we switched to a preconditioned MALA scheme, where the preconditioner was prescribed to be the posterior covariance matrix estimated from a converged MCMC chain generated by another algorithm (e.g., Gibbs sampling). Since finding this covariance requires a full exploration of the posterior, this scheme is not something that could be applied in practice; rather, it represents the “ideal” or endpoint of any AMALA scheme. But we show these MALA results here simply for comparative purposes. As seen in Figures 4.3 and 4.4, the posterior mean (also called the conditional mean (CM)) and posterior covariance from all three MCMC algorithms agree as we increase the maximum number of evaluations. This provides numerical evidence that RTO-MH with a prior transformation generates samples from the correct distribution.

Next, we assess effective sample size (ESS) per function/Jacobian evaluation and per CPU-second, as two measures of computational efficiency. ESS is the number of effectively independent samples in a Markov chain, i.e., the number of samples in a standard Monte Carlo estimator that has the same variance as an estimator computed from the correlated samples of the MCMC chain. It can be interpreted as a measure of the quality of the MCMC samples, where larger values of ESS indicate better chain mixing [15]. An accurate way to calculate the ESS of an MCMC chain of a single parameter is found in [50]; we do so for each component of our chains and report the minimum, median, and maximum (across components) ESS per evaluation and ESS per CPU-second in Table 4.1. The RTO method has a higher ESS per evaluation than the other benchmark algorithms. However, since the optimization and calculation of the weights in RTO involves additional computational overhead, MALA using the “ideal” preconditioner has a higher ESS per CPU-second than RTO-MH. As noted above, though, MALA with the “ideal” preconditioner is not a practically realizable algorithm. AMALA is a practical realization of preconditioned MALA, and the chain’s poor mixing is reflected in low ESS per CPU-second values. Overall, these results suggest that RTO-MH with a prior transformation is quite competitive for this test case, even without accounting for the fact that RTO can be run in parallel.

TABLE 4.1

Example A: ESS per evaluation or per CPU-second. Each Jacobian evaluation is considered to be equivalent in cost to one function evaluation. MALA (ideal) is preconditioned with the posterior covariance calculated from a converged chain of another method.

Method	ESS per evaluation			ESS per CPU-second		
	Minimum	Median	Maximum	Minimum	Median	Maximum
RTO with transf.	$2.48 \cdot 10^{-3}$	$7.43 \cdot 10^{-3}$	$8.72 \cdot 10^{-3}$	$4.77 \cdot 10^{-1}$	$1.43 \cdot 10^0$	$1.67 \cdot 10^0$
AMALA	$1.09 \cdot 10^{-6}$	$1.21 \cdot 10^{-6}$	$3.76 \cdot 10^{-6}$	$3.19 \cdot 10^{-4}$	$3.54 \cdot 10^{-4}$	$1.10 \cdot 10^{-3}$
MALA (ideal)	$1.08 \cdot 10^{-3}$	$1.24 \cdot 10^{-3}$	$1.48 \cdot 10^{-3}$	$1.39 \cdot 10^1$	$1.60 \cdot 10^1$	$1.90 \cdot 10^1$
Gibbs	$9.60 \cdot 10^{-6}$	$7.06 \cdot 10^{-5}$	$1.26 \cdot 10^{-4}$	$1.96 \cdot 10^{-2}$	$1.44 \cdot 10^{-1}$	$2.57 \cdot 10^{-1}$

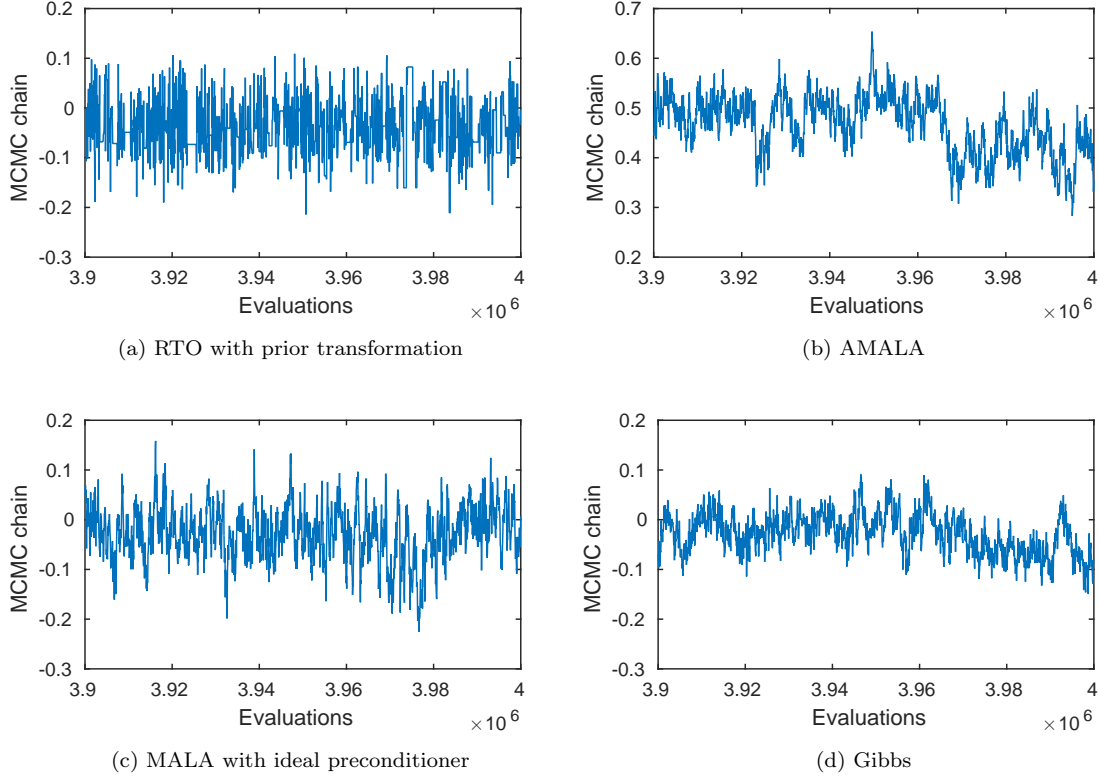


FIG. 4.2. *Example A: MCMC chains from various methods. Index 7 is plotted, which corresponds to the median ESS for RTO. The chain for AMALA is not yet stationary. The horizontal axis (number of function and Jacobian evaluations) reflects a common measure of computational cost for all methods.*

Remark 4.1. The CM estimate using a TV prior is in general not piecewise constant (i.e., blocky). In [27], it is proven that under refinement of parameter discretization, the CM estimate using a TV prior will become smooth.

4.1.2. Example B: Besov space prior. This second example is also a deconvolution of a 1-D signal. Here, the true signal is taken to be

$$\theta_{\text{true}}(x) = \begin{cases} 1 & \text{if } 2/15 < x < 7/15 \\ \frac{1}{2} & \text{if } 10/15 < x < 13/15 \\ 0 & \text{otherwise} \end{cases}.$$

Figure 4.5 shows the true signal and resulting data.

This time, we use the Besov $B_{1,1}^s$ prior with $s = 1$ and Haar wavelets, so that again $\pi(\theta) \propto \exp(-\lambda \|D\theta\|_1)$, where $\theta \in \mathbb{R}^n$, $D \in \mathbb{R}^{n \times n}$, and $\lambda \in \mathbb{R}$. In this case, the matrix D contains scaled wavelet basis functions (see details in Appendix C), and n must be a power of 2. We set the observational noise to be $\sigma_{\text{obs}} = 1 \cdot 10^{-3}$.

RTO with a prior transformation is used to sample from the posterior distributions. We perform two studies: first by fixing the hyperparameter to $\lambda = 32$ and scanning through parameter dimensions $n \in \{32, 64, 128, 258, 512\}$; and second, by fixing $n = 64$ and scanning through hyperparameter values $\lambda \in \{\frac{1}{2}, 1, 2, 4, 8, 16, 32, 64, 128\}$. We use chain lengths of $1 \cdot 10^4$, and tabulate the total ESS and the number of function and Jacobian evaluations. When we increase the dimension n , the posterior mean converges, as in Figure 4.6. This is expected due to the discretization-invariant nature of the

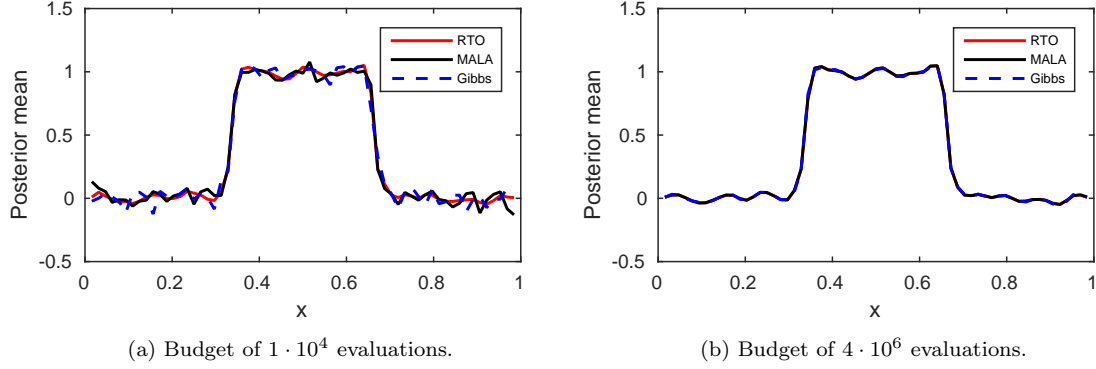


FIG. 4.3. *Example A: Sample estimates of the posterior mean, computed with transformed RTO (red), MALA (black), and Gibbs (blue).*

Besov $B_{1,1}^s$ prior [26, 8]. Next, as reported in Table 4.2, with each doubling of the dimension n , the ESS does not really decrease and the number of function evaluations increases only slightly. This is an important and encouraging result, as it is evidence of discretization invariance not only in the problem formulation, but in the performance of the transformed RTO-MH *sampling scheme*. Finally, as we increase the hyperparameter λ , the CM becomes smoother and the posterior standard deviation decreases, as in shown Figure 4.7. The sampling efficiency of our algorithm also deteriorates with increasing λ , as shown in Table 4.3. Overall, the results from these parameter studies indicate that RTO-MH with a prior transformation is effective even when the parameter dimension n is in the hundreds.

Remark 4.2. In Figure 4.6, the posterior standard deviation does not converge as the discretization is refined (i.e., as n increases). This behavior is not unexpected, as the prior standard deviation also does not converge under mesh refinement. In particular, the $B_{1,1}^s$ Besov space prior with Haar wavelets has finite pointwise variance only when $s > 1$, and not when $s = 1$. One can prove this property by summing the variance contributions from each level of wavelets in the Besov prior, as shown in Appendix D.

Remark 4.3. One possible reason for the decrease in sampling efficiency with higher λ is that the posterior samples lie further in the tails of the Laplace prior. As a result, the transformation is more nonlinear in the sense that the Hessian involving g''_{1D} is of higher magnitude.

TABLE 4.2

Example B: ESS and computational cost of RTO for various parameter dimensions, given chains of length $1 \cdot 10^4$.

n	Total ESS			Total evaluations	
	Minimum	Median	Maximum	Function	Jacobian
32	$2.68 \cdot 10^3$	$3.86 \cdot 10^3$	$4.61 \cdot 10^3$	$4.26 \cdot 10^5$	$4.26 \cdot 10^5$
64	$2.63 \cdot 10^3$	$3.65 \cdot 10^3$	$4.44 \cdot 10^3$	$4.55 \cdot 10^5$	$4.55 \cdot 10^5$
128	$2.10 \cdot 10^3$	$3.53 \cdot 10^3$	$5.07 \cdot 10^3$	$4.59 \cdot 10^5$	$4.59 \cdot 10^5$
256	$2.89 \cdot 10^3$	$3.69 \cdot 10^3$	$4.43 \cdot 10^3$	$4.61 \cdot 10^5$	$4.61 \cdot 10^5$
512	$2.06 \cdot 10^3$	$3.65 \cdot 10^3$	$4.41 \cdot 10^3$	$4.65 \cdot 10^5$	$4.65 \cdot 10^5$

4.2. Two-dimensional elliptic PDE inverse problem. Our next numerical example is an elliptic PDE coefficient inverse problem on a two-dimensional domain. The forward model maps the

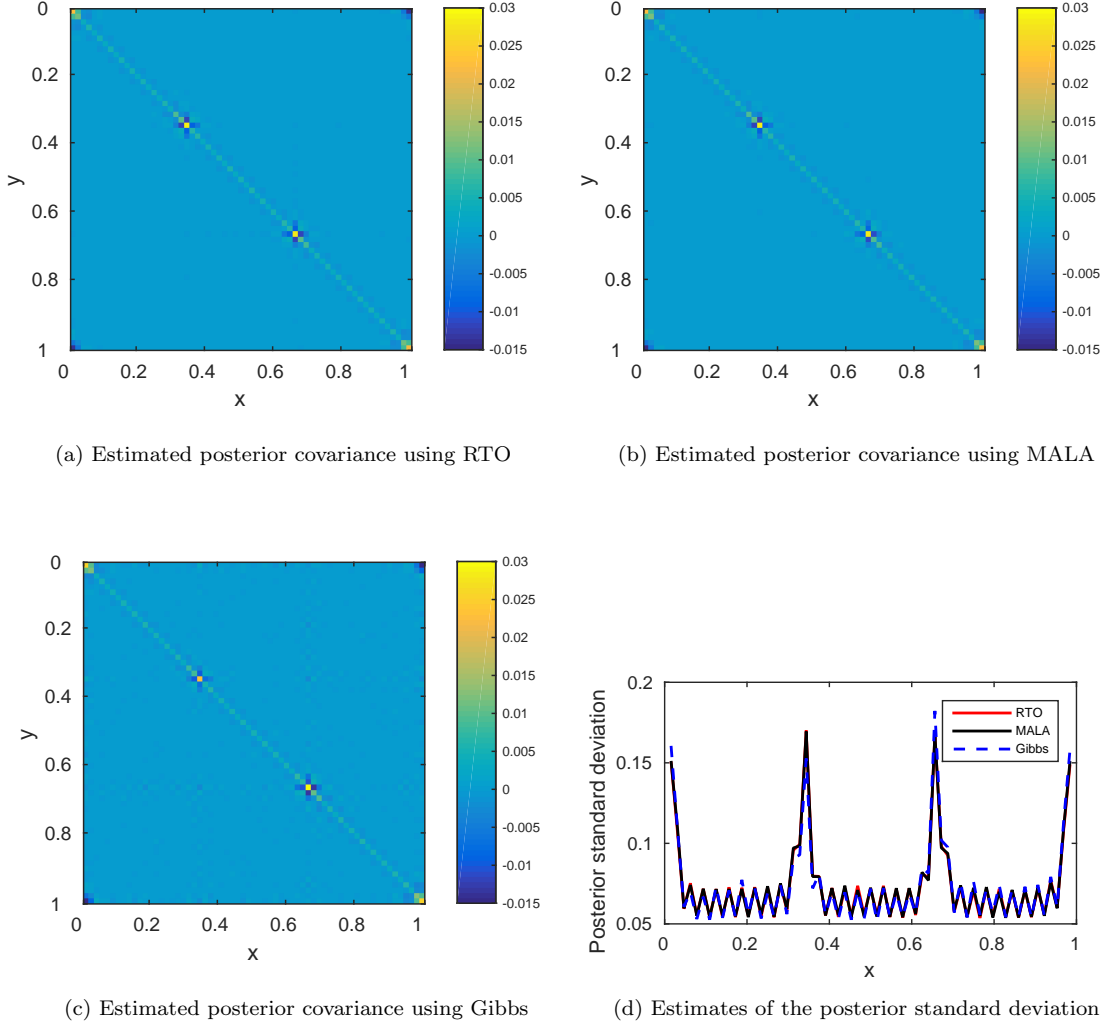


FIG. 4.4. *Example A: Sample estimates of the posterior covariance (top row and bottom left) and pointwise posterior standard deviation (bottom right), using a budget of $4 \cdot 10^6$ evaluations.*

log-conductivity field of the Poisson equation to observations of the potential field,

$$-\nabla \cdot \left(\exp(\theta(x)) \nabla s(x) \right) = h(x), \quad x \in [0, 1]^2,$$

where θ is the log-conductivity, s is the potential, and h is the forcing function. Neumann boundary conditions

$$\exp(\theta(x)) \nabla s(x) \cdot \vec{n}(x) = 0$$

are imposed, where $\vec{n}(x)$ is the normal vector at the boundary. To complete the system of equations, the average potential on the boundary is set to zero.

This PDE is solved using finite elements. The domain is partitioned into a $\sqrt{n} \times \sqrt{n}$ uniform grid of square elements, and we use linear shape functions in both directions. The parameters $\theta \in \mathbb{R}^n$ to be inferred are the nodal values of $\theta(x)$. Independent Gaussian noise with standard deviation $\sigma_{\text{obs}} = 2 \cdot 10^{-3}$ is added to the potential field s to give the observational data y .

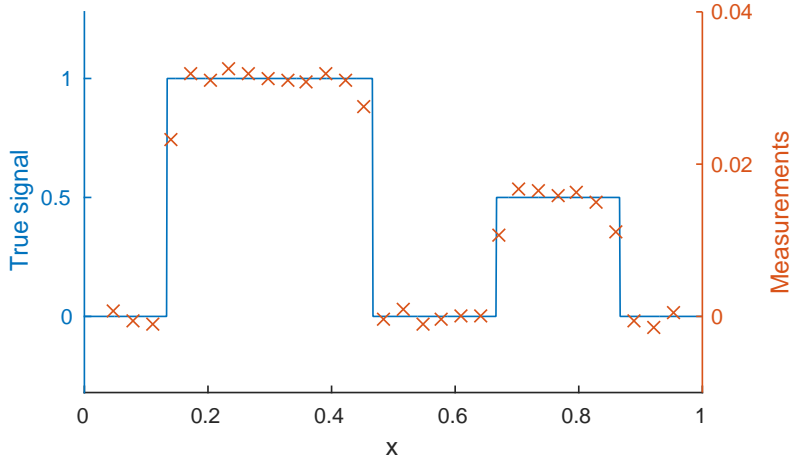


FIG. 4.5. *Example B: True signal and noisy measurements.*

TABLE 4.3

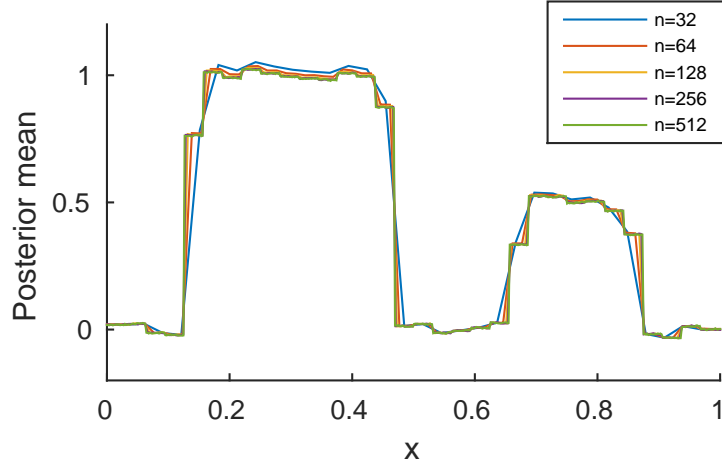
Example B: ESS and computational cost of RTO for varying prior hyperparameter values λ . Chains of length $1 \cdot 10^4$ are used.

λ	Total ESS			Total evaluations	
	Minimum	Median	Maximum	Function	Jacobian
0.5	$5.00 \cdot 10^3$	$5.83 \cdot 10^3$	$7.88 \cdot 10^3$	$5.47 \cdot 10^5$	$5.47 \cdot 10^5$
1	$5.66 \cdot 10^3$	$6.30 \cdot 10^3$	$8.11 \cdot 10^3$	$5.05 \cdot 10^5$	$5.05 \cdot 10^5$
2	$5.74 \cdot 10^3$	$6.71 \cdot 10^3$	$8.23 \cdot 10^3$	$4.73 \cdot 10^5$	$4.73 \cdot 10^5$
4	$5.82 \cdot 10^3$	$6.51 \cdot 10^3$	$8.01 \cdot 10^3$	$4.63 \cdot 10^5$	$4.63 \cdot 10^5$
8	$4.68 \cdot 10^3$	$5.69 \cdot 10^3$	$6.96 \cdot 10^3$	$4.69 \cdot 10^5$	$4.69 \cdot 10^5$
16	$3.20 \cdot 10^3$	$4.39 \cdot 10^3$	$5.29 \cdot 10^3$	$4.77 \cdot 10^5$	$4.77 \cdot 10^5$
32	$2.63 \cdot 10^3$	$3.65 \cdot 10^3$	$4.44 \cdot 10^3$	$4.55 \cdot 10^5$	$4.55 \cdot 10^5$
64	$2.32 \cdot 10^3$	$3.55 \cdot 10^3$	$4.34 \cdot 10^3$	$3.83 \cdot 10^5$	$3.83 \cdot 10^5$
128	$1.08 \cdot 10^3$	$2.19 \cdot 10^3$	$2.79 \cdot 10^3$	$3.02 \cdot 10^5$	$3.02 \cdot 10^5$

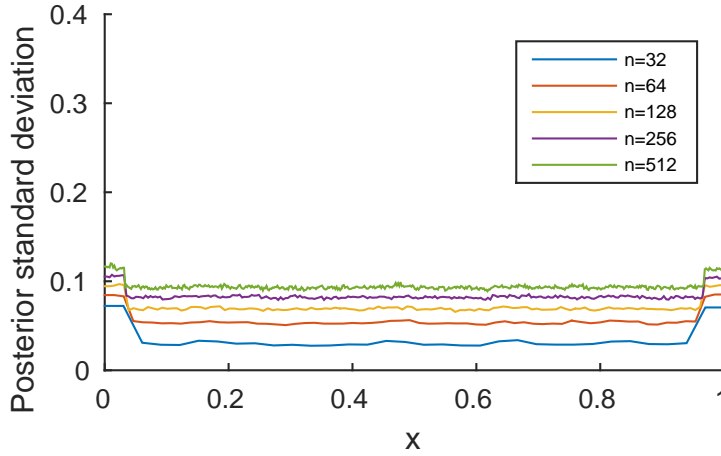
The forcing function h is a linear combination of thirteen Gaussian bumps: nine with weight 1 centered at the points (a, b) , $a, b \in \{0.05, 0.5, 0.95\}$, and four with weight $-9/4$ centered at the points (a, b) , $a, b \in \{0.25, 0.75\}$. The true parameter field θ_{true} , forcing function, and resulting noisy measurements are shown in Figure 4.8. A similar problem setup is found in [7].

4.2.1. Example C: Besov space prior in 2-D. To complete the setup of the Bayesian inverse problem, we impose a 2-D Besov $B_{1,1}^s$ prior, with a tensorized Haar wavelet basis, on θ . This l_1 -type prior is also written in the form (1.3). The columns of matrix D are Kronecker products of the columns of the matrix from the 1-D Besov $B_{1,1}^s$ space prior. The hyperparameter value is $\lambda = 32$ and the parameter dimension is set to $n = 256$, which gives rise to a 16×16 grid. The observational data are generated using a finer 128×128 grid.

We ran RTO-MH with a prior transformation and generated an MCMC chain of length $2 \cdot 10^5$. The computation used $9.3 \cdot 10^6$ function evaluations and $9.3 \cdot 10^6$ Jacobian evaluations to produce an ESS of $4.5 \cdot 10^2$. The posterior mean, estimated from the MCMC samples, appears similar to θ_{true} as shown in Figure 4.9. We also estimate the posterior standard deviation, shown in Figure 4.9; lower uncertainty regions seem to coincide with smaller log-conductivities. It is also



(a) Posterior mean



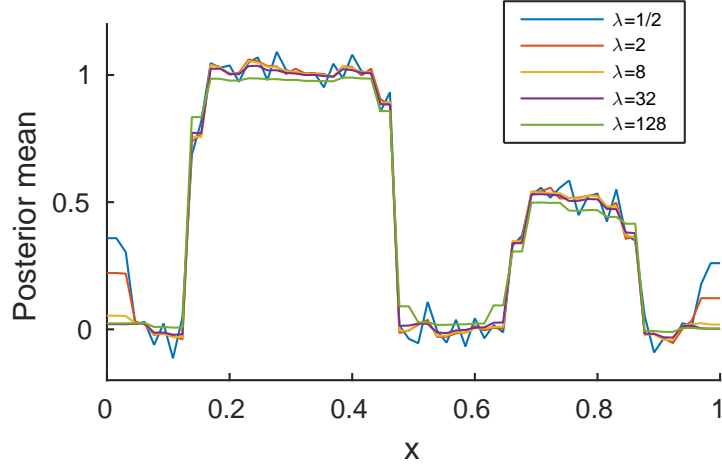
(b) Posterior standard deviation

FIG. 4.6. *Example B: Posterior mean and standard deviation for different values of the parameter dimension n . Hyperparameter λ is fixed to 32.*

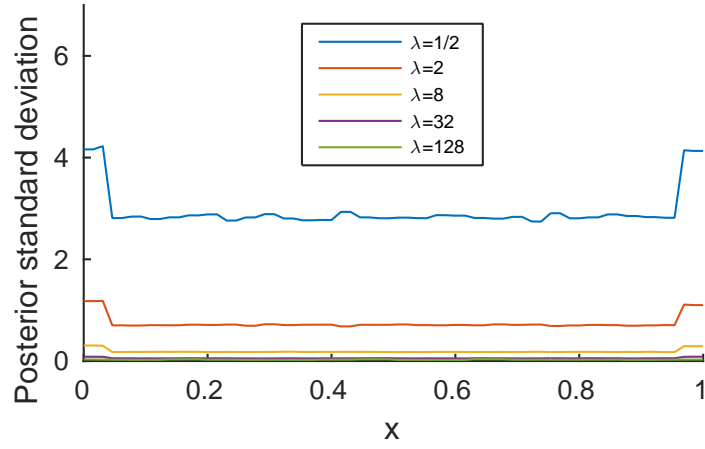
instructive to look at independent samples from the posterior, shown in Figure 4.10(a). They exhibit small-scale roughness (consistent with the Besov prior) and sample-to-sample variability; however, the corresponding samples from the posterior predictive distribution closely match the data, as in Figure 4.10(b). The fact that wider variability among the parameter samples corresponds to much narrower variability among the predictions reflects the smoothing properties of the forward operator and the ill-posedness of the inverse problem. Collectively the posterior samples $\{\theta^{(i)}\}$ characterize uncertainty in the solution of the inverse problem.

We note that the Gibbs sampler of [29] does not extend to nonlinear inverse problems such as this test case.

5. Concluding remarks. We have extended RTO, an optimization-based sampling algorithm, to posterior distributions arising in Bayesian inverse problems with non-Gaussian priors. As a concrete example, we consider l_1 -type priors such as TV and Besov $B_{1,1}^s$ priors. To transform the posterior into a form usable by RTO, we derive a deterministic map that transforms the prior to a



(a) Posterior mean



(b) Posterior standard deviation

FIG. 4.7. *Example B: Posterior mean and standard deviation for different values of the hyperparameter λ . Parameter dimension n is fixed to 64.*

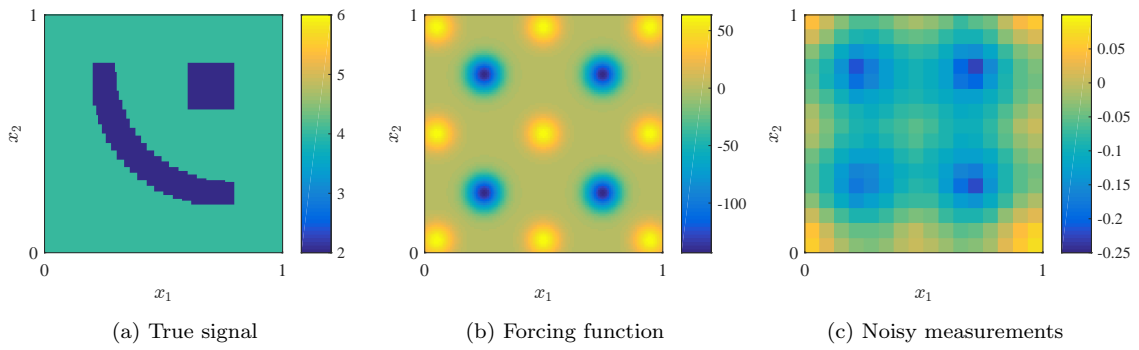


FIG. 4.8. *Example C: True signal, forcing function, and noisy measurements.*

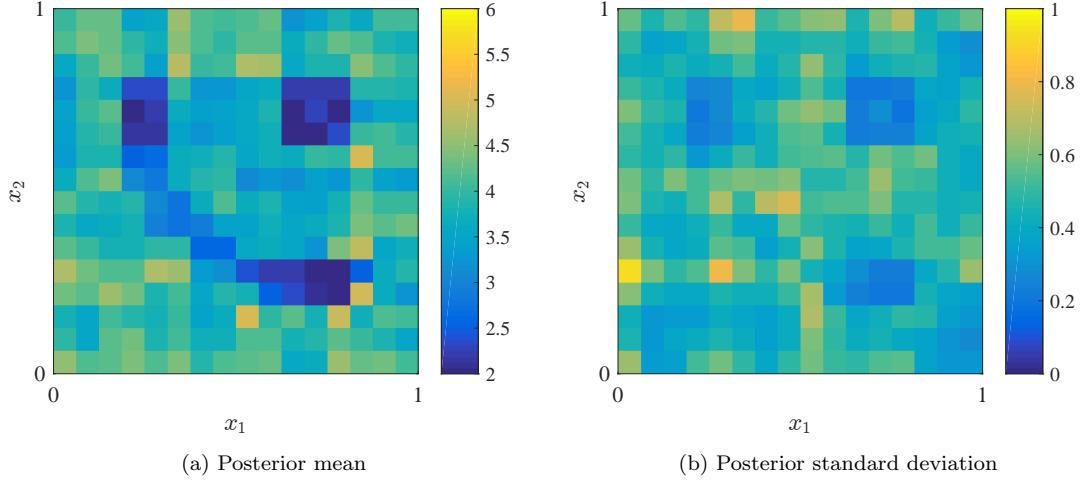


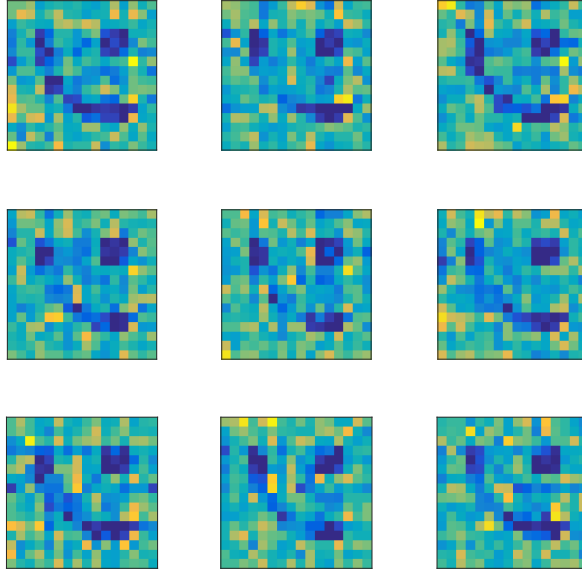
FIG. 4.9. *Example C: Posterior moments for the 2-D elliptic PDE problem.*

standard Gaussian. We embed the RTO proposal into a Metropolis-Hastings algorithm to generate asymptotically exact samples from the transformed posterior, and then apply the transformation to obtain samples from the original posterior. Some assumptions are required for the probability density of the RTO proposal samples to be known and computable. We prove that these assumptions are satisfied for linear forward models and our transformation of l_1 priors. Numerical studies suggest that our method can be more efficient than standard MCMC algorithms, and that its sampling performance does not deteriorate as the parameter discretization is refined. We also successfully employ the algorithm for posterior sampling in a nonlinear inverse problem with a Besov $B_{1,1}^s$ prior in two spatial dimensions, suggesting that it is a promising and versatile computational approach for challenging problems.

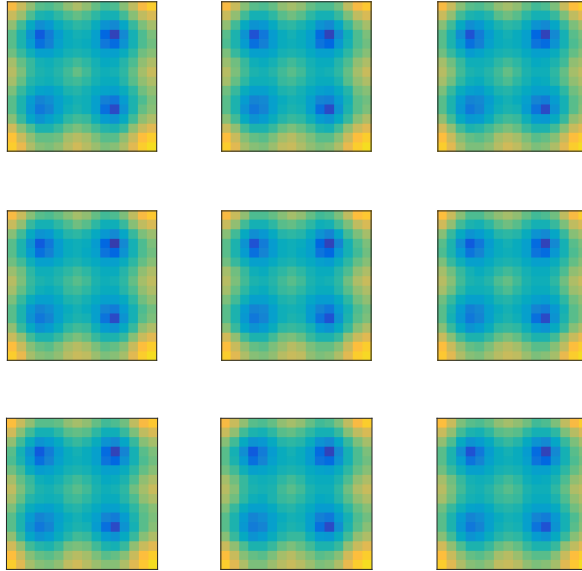
Acknowledgments. Z. Wang, A. Solonen, and Y. Marzouk acknowledge support from the eni-MIT Alliance research program. J. Bardsley was funded by the National Security Technologies, LLC, Site Directed Research and Development program.

Appendix A. Proof of Lemma 3.1. We now derive the posterior density on u . In this appendix, we use more precise notation for clarity. Let $\pi_\Theta(w)$ be the prior density on θ evaluated at $\theta = w$, $\pi_U(w)$ be the prior density on u evaluated at $u = w$, and so forth for the posterior densities. First, note that

$$\pi_\Theta(g_{1D}(u)) = \pi_U(u) \left| \frac{\partial}{\partial \theta} g_{1D}^{-1}(\theta) \right|,$$



(a) Posterior samples $\theta^{(i)}$



(b) Corresponding posterior predictive samples $f(\theta^{(i)})$

FIG. 4.10. *Example C: Posterior samples and corresponding posterior predictive samples. The former have small-scale roughness and sample-to-sample variability, while the latter closely match the potential field measurements.*

and thus

$$\begin{aligned}
\pi_{U|Y}(u|y) &= \pi_{\Theta|Y}(g_{1D}(u)|y) \overbrace{\left| \frac{\partial}{\partial u} g_{1D}(u) \right|}^{|J_{g_{1D}}|} \\
&\propto \exp \left[-\frac{1}{2} \left(\frac{f \circ g_{1D}(u) - y}{\sigma_{\text{obs}}} \right)^2 \right] \pi_{\Theta}(g_{1D}(u)) \left| \frac{\partial}{\partial u} g_{1D}(u) \right| \\
&\propto \exp \left[-\frac{1}{2} \left(\frac{f \circ g_{1D}(u) - y}{\sigma_{\text{obs}}} \right)^2 \right] \pi_U(u) \underbrace{\left| \frac{\partial}{\partial \theta} g_{1D}^{-1}(\theta) \right|}_{|J_{g_{1D}}^{-1}|} \underbrace{\left| \frac{\partial}{\partial u} g_{1D}(u) \right|}_{|J_{g_{1D}}|} \\
&\propto \exp \left[-\frac{1}{2} \left(\frac{f \circ g_{1D}(u) - y}{\sigma_{\text{obs}}} \right)^2 \right] \pi_U(u) \\
&\propto \exp \left[-\frac{1}{2} \left(\frac{f \circ g_{1D}(u) - y}{\sigma_{\text{obs}}} \right)^2 \right] \exp \left(-\frac{1}{2} u^2 \right).
\end{aligned}$$

We note that in the third line, by the inverse function theorem, $\left| \frac{\partial}{\partial \theta} g_{1D}^{-1}(\theta) \right|$ is the inverse of $\left| \frac{\partial}{\partial u} g_{1D}(u) \right|$ and the two terms cancel.

Appendix B. RTO proposal density and proof of Theorem 3.2. First, we recall the assumptions under which the RTO proposal density in (2.5) holds. Knowing the form of the proposal density is important because it allows us to “correct” the proposed samples and thus achieve exact sampling, for instance through the use of a Metropolis-Hastings scheme, or via importance sampling. The theorem that describes the required assumptions is found in [3] and restated below.

ASSUMPTION B.1 (Conditions for validity of the RTO proposal density).

- (i) $p(\theta|y) \propto \exp \left(-\frac{1}{2} \|F(\theta)\|^2 \right)$, where $\theta \in \mathbb{R}^n$.
- (ii) $F : \mathbb{R}^n \rightarrow \mathbb{R}^{n+m}$ is a continuously differentiable function with Jacobian J_F .
- (iii) $J_F(\theta) \in \mathbb{R}^{(n+m) \times n}$ has rank n for every θ in the domain of F .
- (iv) The matrix $\overline{Q}^\top J_F(\theta)$ is invertible for all θ in the domain of F , where

$$J_F(\bar{\theta}) = [\overline{Q}, \tilde{Q}] \begin{bmatrix} \overline{R} \\ 0 \end{bmatrix}$$

is the QR factorization of $J_F(\bar{\theta})$, with $\bar{\theta}$ fixed.

THEOREM B.2 (Proposal density for RTO [3]). *If Assumption B.1 holds, then the RTO algorithm described by Steps 1–7 of Algorithm 2.1 generates proposal samples distributed according to the probability density (2.5).*

We now prove Theorem 3.2 by checking Assumptions B.1(i) to B.1(iv) for the transformed forward model $\tilde{f}(u)$.

Proof of Theorem 3.2. If \tilde{f} is continuously differentiable, then \tilde{F} is continuously differentiable. Thus Assumptions B.1(i) and B.1(ii) are automatically satisfied. Assumption B.1(iii) is also satisfied since

$$J_{\tilde{F}}(\theta) = \begin{bmatrix} I \\ J_{\tilde{f}}(\theta) \end{bmatrix},$$

and regardless of $J_{\tilde{f}}(\theta)$, the columns of $J_{\tilde{F}}(\theta)$ are linearly independent due to the identity matrix in the first n rows of $J_{\tilde{F}}(\theta)$.

To show that Assumption B.1(iv) holds, we use the form of the transformed forward model. Let the original linear forward model be $f(\theta) = A\theta$. Then the transformed forward model is

$$\tilde{f}(u) = f(D^{-1}g(u)) = AD^{-1}g(u).$$

Following the computations used to obtain (3.7), the posterior on u takes the form

$$\begin{aligned} p(u|y) &\propto \exp \left[-\frac{1}{2}(\tilde{f}(u) - y)^\top \Gamma_{\text{obs}}^{-1}(\tilde{f}(u) - y) \right] \exp \left(-\frac{1}{2}u^\top u \right) \\ &= \exp \left(-\frac{1}{2} \left\| \tilde{F}(u) \right\|^2 \right), \end{aligned}$$

where

$$\tilde{F}(u) = \begin{bmatrix} u \\ \Gamma_{\text{obs}}^{-1/2}(\tilde{f}(u) - y) \end{bmatrix}, \quad J_{\tilde{F}}(u) = \begin{bmatrix} I \\ \Gamma_{\text{obs}}^{-1/2}AD^{-1}J_g(u) \end{bmatrix}.$$

Assumption B.1(iv) requires that the matrix $\bar{Q}^\top J_{\tilde{F}}(u)$ be invertible for all u in the domain of \tilde{F} . For any $u_1 \in \mathbb{R}^n$ and $u_2 \in \mathbb{R}^n$,

$$\begin{aligned} J_{\tilde{F}}(u_1)^\top J_{\tilde{F}}(u_2) &= I + J_g(u_1)D^{-\top}A^\top \Gamma_{\text{obs}}^{-1}AD^{-1}J_g(u_2) \\ &= J_g(u_1) \left(J_g(u_1)^{-1}J_g(u_2)^{-1} + D^{-\top}A^\top \Gamma_{\text{obs}}^{-1}AD^{-1} \right) J_g(u_2). \end{aligned}$$

$J_g(u)$ is a positive diagonal matrix for any u , and $D^{-\top}A^\top \Gamma_{\text{obs}}^{-1}AD^{-1}$ is symmetric positive semi-definite. Then, the middle matrix is symmetric positive definite. Thus, $J_{\tilde{F}}(u_1)^\top J_{\tilde{F}}(u_2)$ is the product of three invertible matrices and is therefore invertible.

\bar{Q} is obtained from the thin QR-decomposition of $J_{\tilde{F}}(\bar{u})$, where \bar{u} is the mode of the posterior defined on u . It follows that $J_{\tilde{F}}(u)^\top \bar{Q} = J_{\tilde{F}}(u)^\top J_{\tilde{F}}(\bar{u})\bar{R}^{-1}$ is invertible for any $u \in \mathbb{R}^n$. This shows that Assumption B.1(iv) holds. Hence, Assumptions B.1(i) to B.1(iv) hold for the transformed forward model $\tilde{f}(u)$ and Theorem B.2 yields Theorem 3.2. \square

Appendix C. Besov space priors as l_1 -type priors. Following [25], we start with a wavelet function $\psi \in \mathcal{L}_2([0, 1])$ defined such that the family of functions

$$\psi_{j,k}(x) = 2^{\frac{j}{2}}\psi(2^jx - k), \quad j, k \in \mathbb{Z}_+, \quad 0 \leq k \leq 2^j - 1,$$

is an orthonormal basis for $\mathcal{L}_2([0, 1])$. One example of such a function is the Haar wavelet,

$$\psi_{\text{Haar}} = \begin{cases} 1 & \text{when } 0 < x < \frac{1}{2} \\ -1 & \text{when } \frac{1}{2} < x < 1 \end{cases}.$$

With a wavelet and corresponding basis, we can represent functions by the expansion

$$f(x) = c_0 + \sum_{j=0}^{\infty} \sum_{k=0}^{2^j-1} w_{j,k} \psi_{j,k}(x), \quad c_0 := \int_0^1 f(x)dx, \quad w_{j,k} := \int_0^1 f(x) \psi_{j,k}(x)dx.$$

The Besov space $B_{p,q}^s([0, 1])$ contains functions over the interval $[0, 1]$ with a finite Besov $B_{p,q}^s([0, 1])$ norm, defined as

$$\|f\|_{B_{p,q}^s([0,1])} := \left(|c_0|^q + \sum_{j=0}^{\infty} 2^{jq \left(s + \frac{1}{2} - \frac{1}{p} \right)} \left(\sum_{n=0}^{2^j-1} |w_{j,k}|^p \right)^{\frac{q}{p}} \right)^{\frac{1}{q}},$$

where $s \in \mathbb{R}$ and $p, q \geq 1$ are properties of the space, and $c_0, w_{j,k} \in \mathbb{R}$ are the coefficients of the expansion. The discrete Besov $B_{p,q}^s$ space norm, defined for a vector $\theta \in \mathbb{R}^n$ of size $n = 2^l$, is

$$\|\theta\|_{B_{p,q}^s} := \left(|\hat{c}_0|^q + \sum_{j=0}^l 2^{jq(s+\frac{1}{2}-\frac{1}{p})} \left(\sum_{n=0}^{2^j-1} |\hat{w}_{j,k}|^p \right)^{\frac{q}{p}} \right)^{\frac{1}{q}},$$

where $\hat{c}_0, \hat{w}_{j,k} \in \mathbb{R}$ are the coefficients

$$\begin{aligned} \hat{c}_0 &= \frac{1}{n} \theta^\top \hat{\phi}_{0,0}, & \hat{w}_{j,k} &= \frac{1}{n} \theta^\top \hat{\psi}_{j,k}, \\ \hat{\phi}_{0,0} &= [1, \dots, 1]^\top, & \hat{\psi}_{j,k} &= \left[\psi_{j,k} \left(\frac{1}{2n} \right), \psi_{j,k} \left(\frac{3}{2n} \right), \dots, \psi_{j,k} \left(\frac{2n-1}{2n} \right) \right]^\top. \end{aligned}$$

Note that when $\theta \in \mathbb{R}^n$ is a discretization of the continuous function $f : [0, 1] \rightarrow \mathbb{R}$,

$$\theta = \left[f \left(\frac{1}{2n} \right), f \left(\frac{3}{2n} \right), \dots, f \left(\frac{2n-1}{2n} \right) \right]^\top.$$

Then, the discrete norm $\|\theta\|_{B_{p,q}^s}$ is an approximation to the continuous norm $\|f\|_{B_{p,q}^s}$. When $p = q = 1$, the discrete Besov $B_{p,q}^s$ space norm becomes

$$\begin{aligned} \|\theta\|_{B_{1,1}^s} &= |\hat{c}_0| + \sum_{j=0}^l \sum_{h=0}^{2^j-1} 2^{j(s-\frac{1}{2})} |\hat{w}_{j,k}| \\ &= \|WB\theta\|_1, \end{aligned}$$

where the matrix $W \in \mathbb{R}^{n \times n}$ is diagonal with

$$W_{1,1} = \frac{1}{\sqrt{n}} \quad \text{and} \quad W_{i,i} = \frac{1}{\sqrt{n}} 2^{j(s-\frac{1}{2})} \text{ when } 2^j + 1 \leq i \leq 2^{j+1},$$

and the matrix $B \in \mathbb{R}^{n \times n}$ is unitary with

$$B = \frac{1}{\sqrt{n}} \begin{bmatrix} \hat{\phi}_{0,0} & \hat{\psi}_{0,0} & \hat{\psi}_{1,0} & \hat{\psi}_{1,1} & \hat{\psi}_{2,0} & \dots \end{bmatrix}^\top.$$

Thus, we can write the Besov $B_{1,1}^s$ space prior in the form of (1.3) by

$$p(\theta) := \exp \left(-\lambda \|\theta\|_{B_{1,1}^s} \right) = \exp \left(-\lambda \|D\theta\|_1 \right),$$

where $D = WB$, with W and B defined as above.

Appendix D. Pointwise variance of Besov priors with Haar wavelets. Let f be a random function distributed according to the Besov $B_{1,1}^s$ prior, using Haar wavelets. f can be represented by the expansion

$$f(x) = c_0 + \sum_{j=0}^{\infty} \sum_{k=0}^{2^j-1} w_{j,h} \phi_{j,h}(x).$$

Fix any point x^* and consider the random variable $f(x^*)$. For each level j , there is only one basis, ϕ_{j,h^*} , that has a support containing x^* , where h^* depends on both x^* and j . Also, the magnitude of ϕ_{j,h^*} evaluated at x^* is $2^{\frac{j}{2}}$. Thus,

$$f(x^*) = c_0 + \sum_{j=0}^{\infty} \pm 2^{\frac{j}{2}} w_{j,h^*}$$

where, due to the Besov $B_{1,1}^s$ space prior,

$$c_0 \sim \text{Laplace}(0, 1), \quad w_{j,h^*} \sim \text{Laplace}(0, 2^{-j(s-\frac{1}{2})}).$$

We sum the variance contribution from each coefficient.

$$\text{Var}[f(x^*)] = 2 + \sum_{j=0}^{\infty} 2^j (2 \cdot 2^{-2j(s-\frac{1}{2})}) = 2 \left(1 + \sum_{j=0}^{\infty} 2^{-2j(s-1)} \right)$$

Hence, the pointwise variance is finite when $s > 1$ and does not converge when $s = 1$.

REFERENCES

- [1] Y. F. ATCHADÉ, *An adaptive version for the Metropolis adjusted Langevin algorithm with a truncated drift*, Methodology and Computing in applied Probability, 8 (2006), pp. 235–254.
- [2] J. M. BARDSLEY, *Laplace-distributed increments, the Laplace prior, and edge-preserving regularization*, Journal of Inverse and Ill-Posed Problems, 20 (2012), pp. 271–285.
- [3] J. M. BARDSLEY, A. SOLONEN, H. HAARIO, AND M. LAINE, *Randomize-then-optimize: A method for sampling from posterior distributions in nonlinear inverse problems*, SIAM Journal on Scientific Computing, 36 (2014), pp. A1895–A1910.
- [4] D. CALVETTI AND E. SOMERSALO, *An Introduction to Bayesian Scientific Computing: Ten Lectures on Subjective Computing*, vol. 2, Springer Science & Business Media, 2007.
- [5] A. CHORIN, M. MORZFELD, AND X. TU, *Implicit particle filters for data assimilation*, Communications in Applied Mathematics and Computational Science, 5 (2010), pp. 221–240.
- [6] S. L. COTTER, G. O. ROBERTS, A. M. STUART, AND D. WHITE, *MCMC methods for functions: modifying old algorithms to make them faster*, Statistical Science, 28 (2013), pp. 424–446.
- [7] T. CUI, K. J. LAW, AND Y. M. MARZOUK, *Dimension-independent likelihood-informed MCMC*, Journal of Computational Physics, 304 (2016), pp. 109–137.
- [8] M. DASHTI, S. HARRIS, AND A. STUART, *Besov priors for Bayesian inverse problems*, Inverse Problems and Imaging, 6 (2012), pp. 183–200.
- [9] I. DAUBECHIES, *Ten Lectures on Wavelets*, vol. 61, SIAM, 1992.
- [10] I. DAUBECHIES, M. DEFRISE, AND C. DE MOL, *An iterative thresholding algorithm for linear inverse problems with a sparsity constraint*, Communications on Pure and Applied Mathematics, 57 (2004), pp. 1413–1457.
- [11] D. GAMERMAN AND H. F. LOPES, *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian inference*, Chapman and Hall/CRC Press, 2006.
- [12] A. E. GELFAND AND A. F. SMITH, *Sampling-based approaches to calculating marginal densities*, Journal of the American Statistical Association, 85 (1990), pp. 398–409.
- [13] A. GELMAN, J. B. CARLIN, H. S. STERN, AND D. B. RUBIN, *Bayesian Data Analysis*, vol. 2, Chapman & Hall/CRC Boca Raton, FL, USA, 2014.
- [14] C. J. GEYER, *Practical Markov chain Monte Carlo*, Statistical Science, (1992), pp. 473–483.
- [15] W. R. GILKS, S. RICHARDSON, AND D. J. SPIEGELHALTER, *Markov Chain Monte Carlo in Practice*, Chapman & Hall/CRC Interdisciplinary Statistics, 1996.
- [16] M. GIROLAMI AND B. CALDERHEAD, *Riemann manifold Langevin and Hamiltonian Monte Carlo methods*, Journal of the Royal Statistical Society: Series B (Statistical Methodology), 73 (2011), pp. 123–214.
- [17] P. J. GREEN, *Reversible jump Markov chain Monte Carlo computation and Bayesian model determination*, Biometrika, 82 (1995), pp. 711–732.
- [18] H. HAARIO, M. LAINE, A. MIRA, AND E. SAKSMAN, *DRAM: efficient adaptive MCMC*, Statistics and Computing, 16 (2006), pp. 339–354.
- [19] H. HAARIO, E. SAKSMAN, AND J. TAMMINEN, *An adaptive Metropolis algorithm*, Bernoulli, (2001), pp. 223–242.
- [20] W. K. HASTINGS, *Monte Carlo sampling methods using Markov chains and their applications*, Biometrika, 57 (1970), pp. 97–109.
- [21] M. D. HOFFMAN AND A. GELMAN, *The No-U-turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo*, Journal of Machine Learning Research, 15 (2014), pp. 1593–1623.
- [22] L. T. JOHNSON AND C. J. GEYER, *Variable transformation to obtain geometric ergodicity in the random-walk Metropolis algorithm*, The Annals of Statistics, 40 (2012), pp. 3050–3076.
- [23] J. KAIPIO AND E. SOMERSALO, *Statistical and Computational Inverse Problems*, vol. 160, Springer Science & Business Media, 2006.
- [24] J. P. KAIPIO, V. KOLEHMAINEN, E. SOMERSALO, AND M. VAUHKONEN, *Statistical inversion and Monte Carlo sampling methods in electrical impedance tomography*, Inverse Problems, 16 (2000), p. 1487.
- [25] V. KOLEHMAINEN, M. LASSAS, K. NIINIMÄKI, AND S. SILTANEN, *Sparsity-promoting Bayesian inversion*, Inverse Problems, 28 (2012), p. 025005.

- [26] M. LASSAS, E. SAKSMAN, AND S. SILTANEN, *Discretization-invariant Bayesian inversion and Besov space priors*, Inverse Problems and Imaging, 3 (2009), pp. 87–122.
- [27] M. LASSAS AND S. SILTANEN, *Can one use total variation prior for edge-preserving Bayesian inversion?*, Inverse Problems, 20 (2004), p. 1537.
- [28] J. S. LIU, *Monte Carlo Strategies in Scientific Computing*, Springer Science & Business Media, 2008.
- [29] F. LUCKA, *Fast Markov chain Monte Carlo sampling for sparse Bayesian inference in high-dimensional inverse problems using $L1$ -type priors*, Inverse Problems, 28 (2012), p. 125012.
- [30] J. MARTIN, L. C. WILCOX, C. BURSTEDDE, AND O. GHATTAS, *A stochastic Newton MCMC method for large-scale statistical inverse problems with application to seismic inversion*, SIAM Journal on Scientific Computing, 34 (2012), pp. A1460–A1487.
- [31] Y. MARZOUK, T. MOSELHY, M. PARNO, AND A. SPANTINI, *Sampling via measure transport: An introduction*, in Handbook of Uncertainty Quantification, R. Ghanem, D. Higdon, and H. Owhadi, eds., Springer, 2016.
- [32] J. C. MATTINGLY, N. S. PILLAI, AND A. M. STUART, *Diffusion limits of the random walk Metropolis algorithm in high dimensions*, The Annals of Applied Probability, 22 (2012), pp. 881–930.
- [33] K. L. MENGENSEN AND R. L. TWEEDIE, *Rates of convergence of the Hastings and Metropolis algorithms*, The Annals of Statistics, 24 (1996), pp. 101–121.
- [34] N. METROPOLIS, A. W. ROSENBLUTH, M. N. ROSENBLUTH, A. H. TELLER, AND E. TELLER, *Equation of state calculations by fast computing machines*, The Journal of Chemical Physics, 21 (1953), pp. 1087–1092.
- [35] M. MORZFELD, X. TU, E. ATKINS, AND A. J. CHORIN, *A random map implementation of implicit filters*, Journal of Computational Physics, 231 (2012), pp. 2049–2066.
- [36] T. A. MOSELHY AND Y. M. MARZOUK, *Bayesian inference with optimal maps*, Journal of Computational Physics, 231 (2012), pp. 7815–7850.
- [37] J. L. MUELLER AND S. SILTANEN, *Linear and Nonlinear Inverse Problems with Practical Applications*, vol. 10, SIAM, 2012.
- [38] R. M. NEAL, *MCMC using Hamiltonian dynamics*, Handbook of Markov Chain Monte Carlo, 2 (2011), pp. 113–162.
- [39] G. K. NICHOLLS AND C. FOX, *Prior modeling and posterior sampling in impedance imaging*, in SPIE’s International Symposium on Optical Science, Engineering, and Instrumentation, International Society for Optics and Photonics, 1998, pp. 116–127.
- [40] D. S. OLIVER, *Metropolized Randomized Maximum Likelihood for sampling from multimodal distributions*, arXiv preprint arXiv:1507.08563, (2015).
- [41] D. S. OLIVER, N. HE, AND A. C. REYNOLDS, *Conditioning permeability fields to pressure data*, in ECMOR V-5th European Conference on the Mathematics of Oil Recovery, EAGE, September 1996.
- [42] M. PARNO AND Y. MARZOUK, *Transport map accelerated Markov chain Monte Carlo*, arXiv preprint arXiv:1412.5492, (2014).
- [43] C. ROBERT AND G. CASELLA, *Monte Carlo Statistical Methods*, Springer Science & Business Media, 2013.
- [44] G. O. ROBERTS, A. GELMAN, AND W. R. GILKS, *Weak convergence and optimal scaling of random walk Metropolis algorithms*, The Annals of Applied Probability, 7 (1997), pp. 110–120.
- [45] G. O. ROBERTS AND R. L. TWEEDIE, *Exponential convergence of Langevin distributions and their discrete approximations*, Bernoulli, (1996), pp. 341–363.
- [46] L. I. RUDIN, S. OSHER, AND E. FATEMI, *Nonlinear total variation based noise removal algorithms*, Physica D: Nonlinear Phenomena, 60 (1992), pp. 259–268.
- [47] A. M. STUART, *Inverse problems: a Bayesian perspective*, Acta Numerica, 19 (2010), pp. 451–559.
- [48] A. TARANTOLA, *Inverse Problem Theory and Methods for Model Parameter Estimation*, Other Titles in Applied Mathematics, SIAM, 2005.
- [49] C. R. VOGEL, *Computational Methods for Inverse Problems*, vol. 23, SIAM, 2002.
- [50] U. WOLFF AND ALPHA COLLABORATION, *Monte Carlo errors with less errors*, Computer Physics Communications, 156 (2004), pp. 143–153.