SIAM J. Sci. Comput.
Vol. 42, No. 2, pp. A1317–A1347

SCALABLE OPTIMIZATION-BASED SAMPLING ON FUNCTION SPACE[*]

JOHNATHAN M. BARDSLEY[†], TIANGANG CUI[‡], YOUSSEF M. MARZOUK[§], AND ZHENG WANG[‡]

Abstract. Optimization-based samplers such as randomize-then-optimize (RTO) [J. M. Bardsley et al., *SIAM J. Sci. Comput.*, 36 (2014), pp. A1895–A1910] provide an efficient and parallellizable approach to solving large-scale Bayesian inverse problems. These methods solve randomly perturbed optimization problems to draw samples from an approximate posterior distribution. "Correcting" these samples, either by Metropolization or importance sampling, enables characterization of the original posterior distribution. This paper focuses on the scalability of RTO to problems with high- or infinite-dimensional parameters. In particular, we introduce a new subspace strategy to reformulate RTO. For problems with intrinsic low-rank structures, this subspace acceleration makes the computational complexity of RTO scale linearly with the parameter dimension. Furthermore, this subspace perspective suggests a natural extension of RTO to a function space setting. We thus formalize a function space version of RTO and establish sufficient conditions for it to produce a valid Metropolis–Hastings proposal, yielding *dimension-independent* sampling performance. Numerical examples corroborate the dimension independence of RTO and demonstrate sampling performance that is also robust to small observational noise.

Key words. Markov chain Monte Carlo, Metropolis independence sampling, Bayesian inference, infinite-dimensional inverse problems, transport maps

AMS subject classifications. 15A29, 65C05, 65C60

DOI. 10.1137/19M1245220

1. Introduction.

The Bayesian framework is widely used for uncertainty quantification in inverse problems, i.e., inferring parameters of mathematical models given indirect and noisy data [27, 43]. In a Bayesian setting, the parameters are described as random variables and endowed with prior distributions. Conditioning on an observed data set yields the posterior distribution of these parameters, which characterizes uncertainty in possible parameter values. Solving the inverse problem amounts to computing posterior expectations, e.g., posterior means, variances, marginals, or other summary statistics.

Sampling methods, in particular, Markov chain Monte Carlo (MCMC) algorithms, provide a flexible yet provably convergent way of estimating posterior expectations [5]. The design of effective MCMC methods, however, rests on the careful construction of proposal distributions: efficiency demands proposal distributions that reflect

[*] Submitted to the journal's Methods and Algorithms for Scientific Computing section February 15, 2019; accepted for publication (in revised form) January 27, 2020; published electronically April 27, 2020.

https://doi.org/10.1137/19M1245220

Funding: The work of the first author was supported by the Gordon Preston Fellowship offered by the School of Mathematics at Monash University. The work of the second author was supported by the Australian Research Council, under grant CE140100049 (ACEMS). The work of the third and fourth authors was supported by the United States Department of Energy, Office of Advanced Scientific Computing Research, AEOLUS Mathematical Multifaceted Integrated Capability Center.

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

[‡] Corresponding author. School of Mathematics, Monash University, Clayton, VIC, 3800, Australia (tiangang.cui@monash.edu).

[§] Center for Computational Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139 (zheng_w@mit.edu, ymarz@mit.edu).

the geometry of the posterior [22], e.g., anisotropy, strong correlations, and even non-Gaussianity [35]. Another significant challenge in applying MCMC is parameter dimensionality. In many inverse problems governed by partial differential equations, the "parameter" is in fact a function of space and/or time that, for computational purposes, must be represented in a discretized form. Discretizations that sufficiently resolve the spatial or temporal heterogeneity of this function are often high dimensional. Yet, as analyzed in [30, 31, 39, 40], the performance of many common MCMC algorithms may degrade as the dimension of the discretized parameter increases, meaning that more MCMC iterations are required to obtain an effectively independent sample. One can design MCMC algorithms that do not degrade in this manner by formulating them in function space and ensuring that the proposal distribution satisfies a certain absolute continuity condition [14, 43]. These samplers are called *dimension independent* [14, 15]. Yet another core challenge is that MCMC algorithms are, in general, intrinsically serial: sampling amounts to simulating a discrete-time Markov process. The literature has seen many attempts at parallelizing the evaluation of proposed points [7] or sharing information across multiple chains [13, 26], but none of these is embarrassingly parallel.

A promising approach to many of these challenges is to *convert optimization methods into samplers* (i.e., Monte Carlo methods). This idea has been proposed in many forms: key examples include randomize-then-optimize (RTO) [2], Metropolized randomized-maximum-likelihood (RML) [34, 45], and implicit sampling [10, 32]. In its most basic form, RTO requires Bayesian inverse problems with Gaussian priors and noise models, although it can extend to problems with non-Gaussian priors via a change of variables [46]. Metropolized RML has problem requirements similar to those of RTO, but requires evaluating second-order derivatives of the forward model. Implicit sampling applies to target densities whose contours enclose star-convex regions; each proposal sample can then be generated cheaply by solving a line search.

In general, each of these algorithms solves randomly perturbed realizations of an optimization problem to generate samples from a probability distribution that is "close to" the posterior. The probability density function of this distribution is computable, and thus the distribution can be used as an independent proposal within MCMC or as a biasing distribution in importance sampling. For non-Gaussian targets, these proposal distributions are non-Gaussian. In general, they are *adapted* to the target distribution. The computational complexity and dimension scalability of the resulting sampler can be linked to the structure of the corresponding optimization problem. In addition, these sampling methods are embarrassingly parallel, and are easily implemented with existing optimization tools developed for solving deterministic inverse problems.

This paper considers optimization-based sampling in high dimensions. In particular, we focus on the scalable implementation and analysis of the RTO method. To begin, in section 2 we present interpretations of RTO that provide intuition for the method and its regime of applicability. Using these interpretations, we next motivate and construct a subspace-accelerated version of RTO whose computational complexity scales linearly with parameter dimension (section 3). This approach significantly accelerates RTO in high-dimensional settings. Subspace acceleration reveals that RTO's mapping acts differently on different subspaces of the parameter space. In section 4, we exploit this separation of the parameter subspaces to cast the transport map generated by RTO in an infinite-dimensional (i.e., function space) setting [43]. We also establish sufficient conditions for the probability distribution induced by RTO's mapping to be absolutely continuous with respect to the posterior. This result justifies RTO's observed *dimension-independent* sampling behavior: the acceptance rate and autocorrelation time of an MCMC chain using RTO as its proposal do not degrade as the parameter dimension increases. Similarly, the performance of importance sampling using RTO as a biasing distribution will stabilize in high dimensions. This result is analogous to the arguments in [3, 4, 15, 41, 43] showing that (generalized) preconditioned Crank–Nicolson (pCN), dimension-independent likelihood-informed MCMC, and other infinite-dimensional geometric MCMC methods are dimension independent. However, our MCMC construction relies on non-Gaussian proposals in a Metropolis independence setting, where the Markov chain can be run at essentially zero cost *after* the computationally costly step of drawing proposal samples and evaluating the proposal density. Because the latter step is embarrassingly parallel, the overall MCMC scheme is immediately parallelizable, unlike the above-mentioned MCMC samplers that rely on the iterative construction of Markov chains.

In section 5, we provide a numerical illustration of our algorithm on a one-dimensional elliptic PDE problem, exploring the factors that influence RTO's sampling efficiency. We observe that neither the parameter dimension nor the magnitude of the observational noise influence RTO's performance *per MCMC step*, though they both impact the computational cost of each step. Despite its more costly steps, RTO outperforms simple pCN in this example. In section 6, we further demonstrate the efficacy of our algorithm on a challenging two-dimensional parabolic PDE problem. Overall, our results show that RTO can tackle inverse problems with large parameter dimensions and arbitrarily small observational noise.

2. Background.

RTO generates samples from an approximation to the target (e.g., posterior) distribution in two steps. First, it repeatedly solves perturbed optimization problems to generate independent proposal samples. Second, it uses this collection of samples to describe an independent proposal for Metropolis–Hastings (MH) or a biasing distribution for self-normalized importance sampling. In this section, we first describe the target distributions to which RTO can be applied. We then provide interpretations of RTO from the geometric and transport perspectives, which lead to useful insights regarding both the sampling efficiency of RTO and sufficient conditions for the RTO procedure to be valid. For completeness, we conclude this section by summarizing RTO and other comparable optimization-based sampling methods using the transport map interpretation.

2.1. Target distribution.

RTO applies to target distributions on $\mathbb{R}^n$ whose densities can be written as

$$\pi_{\mathrm{tar}}(v) \propto \exp\left(-\frac{1}{2}\left\|H(v)\right\|^2\right), \tag{2.1}$$

where $H : \mathbb{R}^n \to \mathbb{R}^{m+n}$ is a vector-valued function of the parameters $v \in \mathbb{R}^n$ with an output dimension of $n+m$ for any $m \geq 1$. This structure is found in Bayesian inverse problems and other similar problems with n parameters, m observations, a Gaussian prior, and additive Gaussian observational noise. To illustrate, let

$$y = F(u) + \epsilon, \quad \epsilon \sim \mathrm{N}(0, \Gamma_{\mathrm{obs}}), \quad u \sim \mathrm{N}(m_{\mathrm{pr}}, \Gamma_{\mathrm{pr}}),$$

where $y \in \mathbb{R}^m$ are the data, $F : \mathbb{R}^n \to \mathbb{R}^m$ is the forward model, $u \in \mathbb{R}^n$ is the unknown parameter, and $\epsilon \in \mathbb{R}^m$ is the additive noise, assumed independent of u. Here, m_{pr} is the prior mean, and Γ_{obs} and Γ_{pr} are the covariance matrices of the observation noise and prior. We can simplify the problem via an affine change of variables that transforms the covariance matrices to identity matrices. Defining matrix factorizations of the covariances of prior and observation noise

$$S_{\mathrm{pr}}S_{\mathrm{pr}}^{\top} := \Gamma_{\mathrm{pr}}, \quad S_{\mathrm{obs}}S_{\mathrm{obs}}^{\top} := \Gamma_{\mathrm{obs}},$$

we have new whitened variables,

$$v := S_{\mathrm{pr}}^{-1}(u - m_{\mathrm{pr}}), \quad G(v) := S_{\mathrm{obs}}^{-1}\left[F\left(S_{\mathrm{pr}}v + m_{\mathrm{pr}}\right) - y\right], \quad e := S_{\mathrm{obs}}^{-1}\epsilon,$$

where $v \in \mathbb{R}^n$ is the whitened unknown parameter, $G : \mathbb{R}^n \to \mathbb{R}^m$ is the whitened forward model, and e is the whitened observational noise. The inverse problem becomes

$$0 = G(v) + e, \quad e \sim \mathrm{N}(0, \mathrm{I}), \quad v \sim \mathrm{N}(0, \mathrm{I}).$$

The data are shifted to the origin after whitening. The posterior density of the whitened variable v is then

$$p(v|y) = \pi_{\mathrm{tar}}(v) \propto \exp\left(-\frac{1}{2}\left\|\begin{bmatrix} v \\ G(v)\end{bmatrix}\right\|^2\right),$$

which is in the required form (2.1) with H defined as

$$H(v) := \begin{bmatrix} v \\ G(v)\end{bmatrix}. \tag{2.2}$$

Given a sample v from the target density $\pi_{\mathrm{tar}}(v)$, we can obtain a posterior sample of the original parameter u by applying the transformation

$$u = S_{\mathrm{pr}}v + m_{\mathrm{pr}}.$$

Notice that the form of $\pi_{\mathrm{tar}}(v)$ in (2.1) is identical to the probability density function of an $(n+m)$-dimensional standard normal distribution, $\pi(w) \propto \exp\left(-\frac{1}{2}\|w\|^2\right)$, evaluated at $w = H(v)$. This paints the geometric picture of the required target distribution: the target density $\pi_{\mathrm{tar}}(v)$, up to a normalizing constant, is the same as the density of the $(n+m)$-dimensional Gaussian distribution evaluated on the n-dimensional manifold $H(v) = (v, G(v)) \subset \mathbb{R}^{m+n}$ parameterized by $v \in \mathbb{R}^n$.

2.2. The RTO algorithm. The RTO algorithm requires an orthonormal basis for an n-dimensional subspace of $\mathbb{R}^{n+m}$. Let this basis be collected in a matrix $Q \in \mathbb{R}^{(m+n)\times n}$ with orthonormal columns. One common choice of Q follows from first finding a linearization point v_{ref}, which is often (but not necessarily) taken to be the maximum of the target density, i.e.,

$$v_{\mathrm{ref}} = \arg\min_{v} \frac{1}{2}\|H(v)\|^2. \tag{2.3}$$

Then one can compute Q from a thin QR factorization of $\nabla H(v_{\mathrm{ref}})$; this sets the basis to span the range of $\nabla H(v_{\mathrm{ref}})$.

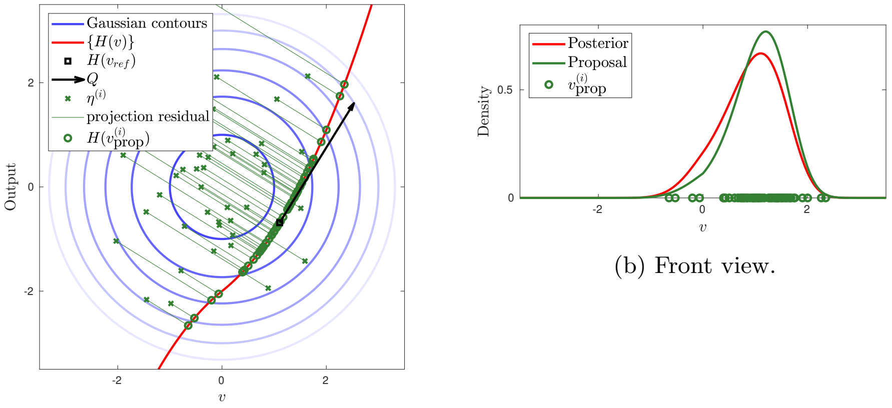
Using this matrix, RTO obtains proposal samples $v_{\mathrm{prop}}^{(i)}$ by repeatedly drawing independent $(n+m)$-dimensional standard normal vectors $\eta^{(i)}$ and solving the nonlinear system of equations

$$Q^{\top}H(v_{\mathrm{prop}}^{(i)}) = Q^{\top}\eta^{(i)}, \tag{2.4}$$

which is equivalent to solving the optimization problem

$$v_{\mathrm{prop}}^{(i)} = \arg\min_{v} \frac{1}{2}\left\|Q^{\top}\left(H(v) - \eta^{(i)}\right)\right\|^2 \tag{2.5}$$

if the minimum of the objective function in (2.5) is zero. To ensure that the system of



(a) Top view.

(b) Front view.

Fig. 1. *Geometric interpretation of RTO, in the case $n = m = 1$. RTO projects the $(n+m)$-dimensional Gaussian samples η (green crosses) onto the manifold $\{H(v)\}$ (red line) to determine the proposal samples v_{prop} (green circles). The projection residual $H(v^{(i)}_{\text{prop}}) - \eta$ is orthogonal to the range of Q. In the front view, the proposal samples v_{prop} (green circles) are shown to be distributed according to a proposal density (green line) that is close to the target density (red line).*

equations (2.4) has a unique solution and that the probability density of the resulting samples can be calculated explicitly, RTO requires the following conditions [2].

Assumption 2.1 (sufficient conditions for valid RTO).

1. *The function H is continuously differentiable with Jacobian ∇H.*
2. *The Jacobian $\nabla H(v)$ has full column rank for every v.*
3. *The map $v \mapsto Q^\top H(v)$ is invertible.*

Proposal samples generated via RTO can be interpreted as a projection of $(n+m)$-dimensional Gaussian samples onto the n-dimensional manifold $\{H(v) : v \in \mathbb{R}^n\}$. The samples are projected along the directions orthogonal to the range of Q such that the condition in (2.4) is satisfied. Figure 1 depicts the steps of RTO's proposal for the case $n = m = 1$. This geometric interpretation also illustrates the importance of the third condition for the RTO procedure to be valid: there will be a unique projected vector on the manifold $\{H(v)\}$ for any given $\eta \sim \mathrm{N}(0, \mathrm{I}_{n+m})$ provided the map $v \mapsto Q^\top H(v)$ is invertible.

The projection defined by RTO realizes the action of a particular transport map. Since the random vector $\eta \in \mathbb{R}^{n+m}$ is a standard Gaussian and the columns of Q are orthonormal, the projection of η, denoted by $\xi := Q^\top \eta \in \mathbb{R}^n$, is also a standard Gaussian. Writing the left-hand side of (2.4) compactly as

$$S(\cdot) = Q^\top H(\cdot),$$

the nonlinear system of equations in (2.4) can be expressed as

$$S(v) = \xi, \quad \text{where} \quad \xi \sim \mathrm{N}(0, \mathrm{I}_n). \tag{2.6}$$

This equation describes a deterministic coupling between the target random variable $v \in \mathbb{R}^n$ and the standard Gaussian "reference" random variable $\xi \in \mathbb{R}^n$. The coupling is defined by the forward model, the data, the observational noise, and the prior, through the function H and the matrix Q.

Under the conditions in Assumption 2.1, solving the nonlinear system (2.6) implicitly inverts the transport map S; that is, it evaluates S^{-1} on each ξ, to obtain a proposal $v = S^{-1}(\xi)$. The normalized probability density of v generated by RTO is given by the pushforward density of the n-dimensional standard Gaussian under the mapping S^{-1}:

$$\begin{aligned} \pi_{\text{RTO}}(v) &= \left|\det \nabla S(v)\right| \pi_{\text{ref}}\left(S(v)\right) \\ &= (2\pi)^{-\frac{n}{2}} \left|\det\left(Q^\top \nabla H(v)\right)\right| \exp\left(-\frac{1}{2}\left\|Q^\top H(v)\right\|^2\right). \end{aligned} \tag{2.7}$$

As shown in [2], RTO's proposal is exact (i.e., is the target) when the forward model is linear, and its proposal is expected to be close to the target when the forward model is close to linear. For weakly nonlinear problems, the proposal can be a good approximation to the posterior and hence can be used in MCMC and importance sampling. These proposal samples can be used either as an independent proposal in MH or as a biasing distribution in importance sampling. For the former case, the MH acceptance ratio can be written as

$$\frac{\pi_{\text{tar}}(v^{(i)}_{\text{prop}})\, \pi_{\text{RTO}}(v^{(i-1)})}{\pi_{\text{tar}}(v^{(i-1)})\, \pi_{\text{RTO}}(v^{(i)}_{\text{prop}})} = \frac{w(v^{(i)}_{\text{prop}})}{w(v^{(i-1)})},$$

where the weight $w(v)$ is defined as

$$w(v) = \left|\det\left(Q^\top \nabla H(v)\right)\right|^{-1} \exp\left(-\frac{1}{2}\left\|H(v)\right\|^2 + \frac{1}{2}\left\|Q^\top H(v)\right\|^2\right). \tag{2.8}$$

The resulting method, called RTO-MH, is summarized in Algorithm 2.1.

For importance sampling, since the normalizing constant of the target density is unknown, the weights must be normalized as

$$\tilde{w}(v^{(i)}) = w(v^{(i)}) / \sum_{j=1}^{N} w(v^{(j)}),$$

where N is the number of samples and the sum of weights $\tilde{w}(v^{(i)})$ is thus one. The proposal samples and weights can then be used to compute posterior expectations of some quantity of interest $g(v)$ using the self-normalizing importance sampling formula:

$$\int g(v)\pi_{\text{tar}}(v)\text{d}v = \sum_{i=1}^{N} \tilde{w}(v^{(i)}_{\text{prop}}) g(v^{(i)}_{\text{prop}}).$$

Remark 2.2. For Bayesian inverse problems, the RTO formulation presented here is limited to cases with Gaussian prior and Gaussian observation noise. By transforming non-Gaussian prior densities and/or observation noises into Gaussian densities, this limitation may be relaxed. See [46, 9] for examples.

Similarly to RTO, other optimization-based samplers such as random-map implicit sampling [32] and Metropolized RML [34] also use a standard normal as the reference distribution and push forward this distribution through some deterministic transformation. Each of these samplers specifies a different inverse transport S, as in (2.6), and then solves an optimization problem to evaluate S^{-1} on each reference

Algorithm 2.1 RTO-MH.

```
Find $v_{\mathrm{ref}}$ using (2.3)
Determine $\nabla H(v_{\mathrm{ref}})$
Compute $Q$, whose columns are an orthonormal basis for the range of $\nabla H(v_{\mathrm{ref}})$
for $i = 1, \ldots, n_{\mathrm{samps}}$ do in parallel
    Sample $\eta^{(i)}$ from an $(n+m)$-dimensional standard normal distribution
    Solve for a proposal sample $v_{\mathrm{prop}}^{(i)}$ using (2.5)
    Compute $w(v_{\mathrm{prop}}^{(i)})$ from (2.8)
Set $v^{(0)} = v_{\mathrm{ref}}$
for $i = 1, \ldots, n_{\mathrm{samps}}$ do in series
    Sample $t$ from a uniform distribution on [0,1]
    if $t < w(v_{\mathrm{prop}}^{(i)}) \big/ w(v^{(i-1)})$ then
        $v^{(i)} = v_{\mathrm{prop}}^{(i)}$
    else
        $v^{(i)} = v^{(i-1)}$
```

sample. For all three algorithms, the pushforward of the reference distribution can be used as a proposal distribution in MH or as a biasing distribution in importance sampling. A summary of each algorithm's mapping is given in Appendix A. The subspace acceleration strategies and infinite-dimensional formulation of RTO developed in this work may also benefit implicit sampling and RML. In addition, interpreting optimization-based samplers as transport maps and utilizing the importance sampling formula naturally open the door to constructing multilevel Monte Carlo estimators [21, 25] for Bayesian computation. This enables additional speedups and can bypass the complicated Markov chain construction processes used in the multilevel MCMC [18, 20]. Further research along this direction is in [8].

3. Scalable implementation of RTO.

In high-dimensional problems, the computation cost of operations involving the dense matrix Q in RTO poses a major computational challenge: the matrix-vector product with $Q^\top$ in each evaluation of the objective function in (2.5) costs $\mathcal{O}((n+m) \times n)$ floating point operations, where n is the number of parameters and m is the number of observed data. Assembling the matrix $Q^\top \nabla H(v)$ also requires $n+m$ matrix-vector products, and an additional $\mathcal{O}(n^3)$ floating point operations are needed to compute the determinant in the proposal density (2.7). For high-dimensional parameters, these operations are computationally prohibitive to apply. To overcome this challenge, we introduce a new subspace acceleration strategy to make these RTO operations scale linearly with the parameter dimension.

3.1. Subspace acceleration.

Our scalable implementation avoids computing and storing the QR factorization of the full-rank $(n+m) \times n$ matrix $\nabla H(v_{\mathrm{ref}})$. Instead, it opts to construct (and store) a singular value decomposition (SVD) of the smaller $m \times n$ linearized forward model $\nabla G(v_{\mathrm{ref}})$. To begin, we note from the definition (2.2) of H that

$$\nabla H(v) = \begin{bmatrix} \mathrm{I} \\ \nabla G(v) \end{bmatrix}.$$

Recall from RTO's mapping (2.6) that the RTO proposal samples are found by

$$Q^\top H(v) = \xi, \quad \text{where} \quad \xi \sim \mathrm{N}(0, \mathrm{I}_n),$$

where the columns of Q form an orthonormal basis for the range of $\nabla H(v_{\mathrm{ref}})$ and Q is computed from the thin QR decomposition of $\nabla H(v_{\mathrm{ref}})$. Since the 2-norm used in the objective function (2.5), the determinant in (2.7), and the standard Gaussian used in the RTO's mapping (2.6) are all invariant up to a rotation defined by an orthogonal matrix, any orthonormal basis for the range of $\nabla H(v_{\mathrm{ref}})$ plays the same role in RTO. This offers a viable way to avoiding computing the dense $(m+n) \times n$ matrix Q.

Instead of computing the QR factorization of ∇H, we consider the polar decomposition [23]:

$$\nabla H(v_{\mathrm{ref}}) = \widetilde{Q}\, \left(\nabla H(v_{\mathrm{ref}})^\top \nabla H(v_{\mathrm{ref}})\right)^{1/2},$$

where the matrix $\widetilde{Q} \in \mathbb{R}^{(m+n)\times n}$ has orthonormal columns and the matrix

$$\left(\nabla H(v_{\mathrm{ref}})^\top \nabla H(v_{\mathrm{ref}})\right)^{1/2} \in \mathbb{R}^{n\times n}$$

is positive definite by construction. This way, the matrix $\widetilde{Q}$ can be constructed as

$$\widetilde{Q} = \nabla H(v_{\mathrm{ref}}) \left(\nabla H(v_{\mathrm{ref}})^\top \nabla H(v_{\mathrm{ref}})\right)^{-\frac{1}{2}}.$$

In the above equation, the matrix $\nabla H(v_{\mathrm{ref}})^\top \nabla H(v_{\mathrm{ref}})$ is the Gauss–Newton approximation of the Hessian of the log-posterior density (referred to as Gauss–Newton Hessian hereafter) defined at the reference parameter v_{ref}.

PROPOSITION 3.1. *Let $\nabla G(v_{\mathrm{ref}})$ denote the forward model linearized at parameter v_{ref}, and consider its reduced SVD,*

$$\nabla G(v_{\mathrm{ref}}) = \Psi \Lambda \Phi^\top.$$

The nonlinear system $\widetilde{Q}^\top H(v) = \xi$ defining the RTO mapping can be rewritten as

$$\begin{cases} (\mathrm{I}_n - \Phi\Phi^\top)\, \xi = (\mathrm{I}_n - \Phi\Phi^\top)\, v, \\ \qquad \Phi\Phi^\top\, \xi = \Phi \left[(\Lambda^2 + \mathrm{I}_r)^{-\frac{1}{2}} \left(\Phi^\top v + \Lambda \Psi^\top G(v)\right)\right]. \end{cases} \tag{3.1}$$

The weighting function $w(v)$ in (2.8) can be expressed as

$$\begin{aligned} w(v) = {} & \left|\det\left(\widetilde{Q}^\top \nabla H(v)\right)\right|^{-1} \\ & \times \exp\left(-\frac{1}{2}\left\|G(v)\right\|^2 - \frac{1}{2}\left\|\Phi^\top v\right\|^2 + \frac{1}{2}\left\|(\Lambda^2 + \mathrm{I}_r)^{-\frac{1}{2}}\left(\Phi^\top v + \Lambda\Psi^\top G(v)\right)\right\|^2\right), \end{aligned} \tag{3.2}$$

where the determinant takes the simplified form

$$\left|\det\left(\widetilde{Q}^\top \nabla H(v)\right)\right| = \left|\det(\Lambda^2 + \mathrm{I}_r)^{-\frac{1}{2}}\right| \left|\det\left(\mathrm{I}_r + \Lambda\Psi^\top \nabla G(v) \Phi\right)\right|. \tag{3.3}$$

Proof. We will show that the matrices in the polar decomposition of $\nabla H(v_{\mathrm{ref}})$ can be obtained from the reduced SVD $\nabla G(v_{\mathrm{ref}}) = \Psi\Lambda\Phi^\top$. The eigendecomposition of the Gauss–Newton Hessian can be written in terms of the reduced SVD as

$$\nabla H(v_{\mathrm{ref}})^\top \nabla H(v_{\mathrm{ref}}) = \Phi(\Lambda^2 + \mathrm{I}_r)\Phi^\top + (\mathrm{I}_n - \Phi\Phi^\top), \tag{3.4}$$

where I_r and I_n are the identity matrices of size $r \times r$ and $n \times n$, respectively. Then, we have the identity

$$\left(\nabla H(v_{\mathrm{ref}})^\top \nabla H(v_{\mathrm{ref}})\right)^{-\frac{1}{2}} = \Phi(\Lambda^2 + \mathrm{I}_r)^{-\frac{1}{2}}\Phi^\top + (\mathrm{I}_n - \Phi\Phi^\top).$$

After some algebraic manipulation, this leads to the matrix

$$\widetilde{Q} = \begin{bmatrix} \Phi(\Lambda^2 + \mathrm{I}_r)^{-\frac{1}{2}}\Phi^\top + (\mathrm{I}_n - \Phi\Phi^\top) \\ \Psi\Lambda(\Lambda^2 + \mathrm{I}_r)^{-\frac{1}{2}}\Phi^\top \end{bmatrix}$$

and, hence, we have

$$\begin{aligned} \widetilde{Q}^\top H(v) &= \left[\Phi(\Lambda^2 + \mathrm{I}_r)^{-\frac{1}{2}}\Phi^\top + (\mathrm{I}_n - \Phi\Phi^\top)\right] v + \Phi\Lambda(\Lambda^2 + \mathrm{I}_r)^{-\frac{1}{2}}\Psi^\top G(v) \\ &= \Phi\left[(\Lambda^2 + \mathrm{I}_r)^{-\frac{1}{2}}\left(\Phi^\top v + \Lambda\Psi^\top G(v)\right)\right] + \left(\mathrm{I}_n - \Phi\Phi^\top\right) v. \end{aligned} \tag{3.5}$$

Thus the nonlinear system $\widetilde{Q}^\top H(v) = \xi$ can be rewritten in the form (3.1). Replacing Q with $\widetilde{Q}$ in the weighting function $w(v)$ (2.8), we have

$$w(v) = \left|\det\left(\widetilde{Q}^\top \nabla H(v)\right)\right|^{-1} \exp\left(-\frac{1}{2}\left\|H(v)\right\|^2 + \frac{1}{2}\left\|\widetilde{Q}^\top H(v)\right\|^2\right).$$

Since the matrix Φ has orthonormal columns, $\Phi\Phi^\top$ and $\mathrm{I}_n - \Phi\Phi^\top$ are orthogonal projectors. This leads to the identities

$$\begin{aligned} \left\|H(v)\right\|^2 = \|v\|^2 + \|G(v)\|^2 &= \left\|(\mathrm{I}_n - \Phi\Phi^\top)v\right\|^2 + \left\|\Phi^\top v\right\|^2 + \|G(v)\|^2, \\ \left\|\widetilde{Q}^\top H(v)\right\|^2 &= \left\|(\mathrm{I}_n - \Phi\Phi^\top)v\right\|^2 + \left\|(\Lambda^2 + \mathrm{I}_r)^{-\frac{1}{2}}\left(\Phi^\top v + \Lambda\Psi^\top G(v)\right)\right\|^2, \end{aligned}$$

by the definition of $H(v)$ in (2.2) and the definition of $\widetilde{Q}^\top H(v)$ in (3.5). Substituting the above identities into $w(v)$, we obtain the result in (3.2). Using (3.5), we obtain the linearization

$$\widetilde{Q}^\top \nabla H(v) = \mathrm{I} + \Phi\left[(\Lambda^2 + \mathrm{I}_r)^{-\frac{1}{2}}\Phi^\top - \Phi^\top + \Lambda(\Lambda^2 + \mathrm{I}_r)^{-\frac{1}{2}}\Psi^\top \nabla G(v)\right].$$

Hence, the determinant term is given by

$$\begin{aligned} \left|\det\left(\widetilde{Q}^\top \nabla H(v)\right)\right| &= \left|\det\left(\mathrm{I} + \Phi\left[(\Lambda^2 + \mathrm{I}_r)^{-\frac{1}{2}}\Phi^\top - \Phi^\top + \Lambda(\Lambda^2 + \mathrm{I}_r)^{-\frac{1}{2}}\Psi^\top \nabla G(v)\right]\right)\right| \\ &= \left|\det\left(\mathrm{I}_r + \left[(\Lambda^2 + \mathrm{I}_r)^{-\frac{1}{2}}\Phi^\top - \Phi^\top + \Lambda(\Lambda^2 + \mathrm{I}_r)^{-\frac{1}{2}}\Psi^\top \nabla G(v)\right]\Phi\right)\right| \\ &= \left|\det(\Lambda^2 + \mathrm{I}_r)^{-\frac{1}{2}}\right| \left|\det\left(\mathrm{I}_r + \Lambda\Psi^\top \nabla G(v)\Phi\right)\right|, \end{aligned}$$

where in the second line above we use Sylvester's determinant identity. This concludes the proof. □

Remark 3.2. For high-dimensional problems, it is not feasible to explicitly construct the linearized forward model $\nabla G(v)$. Instead, one should use matrix-free solvers such as Lanczos or randomized SVD (see [23, 24] and references therein) to compute the SVD of $\nabla G(v)$. This only involves evaluating matrix-vector products (MVPs) with $\nabla G(v)$ and its adjoint.

Equation (3.1) separates ξ into two parts: one in the column space of Φ and another in its orthogonal complement. Defining

$$v_r = \Phi^\top v \quad \text{and} \quad v = \Phi v_r + v_\perp,$$

where $v_\perp$ is an element in the orthogonal complement of range(Φ), we can solve the nonlinear system of equations (3.1) by first computing

$$v_\perp = \left(\mathrm{I}_n - \Phi\Phi^\top\right)\xi, \tag{3.6}$$

and then solving the r-dimensional optimization problem

$$v_r = \underset{v_r'}{\arg\min} \left\| \left(\Lambda^2 + \mathrm{I}_r\right)^{-\frac{1}{2}} \left(v_r' + \Lambda\Psi^\top G\left(v_\perp + \Phi v_r'\right) - \Phi^\top \xi\right) \right\|^2. \tag{3.7}$$

Equations (3.6) and (3.7) replace the n-dimensional optimization problem in (2.5). Note that at each given $v = v_\perp + \Phi v_r$, the vector-valued function within the 2-norm in (3.7) has the linearization

$$\left(\Lambda^2 + \mathrm{I}_r\right)^{-\frac{1}{2}} \left(\mathrm{I}_r + \Lambda\Psi^\top \nabla G(v)\Phi\right) \tag{3.8}$$

w.r.t. the reduced-dimensional parameter v_r. MVPs with the linearization (3.8) and its adjoint are needed by nonlinear optimization algorithms, e.g., quasi-Newton with line search or trust region with inexact Newton–CG [33], to solve (3.7). The scalable implementation of RTO is outlined in Algorithm 3.1.

Algorithm 3.1 Scalable implementation of RTO–MH.

```
Find $v_{\mathrm{ref}}$ using (2.3).
Determine the Jacobian matrix of the forward model, $\nabla G(v_{\mathrm{ref}})$.
Compute $\Psi$, $\Lambda$ and $\Phi$, which is the SVD of $\nabla G(v_{\mathrm{ref}})$.
for $i = 1, \ldots, n_{\mathrm{samps}}$ do in parallel
    Sample $\xi^{(i)}$ from an $n$-dimensional standard normal distribution.
    Solve for a proposal sample $v_{\mathrm{prop}}^{(i)} = v_\perp + \Phi v_r$ using (3.6) and (3.7).
    Compute $w(v_{\mathrm{prop}}^{(i)})$ from (3.2) using the determinant from (3.3).
Set $v^{(0)} = v_{\mathrm{ref}}$.
for $i = 1, \ldots, n_{\mathrm{samps}}$ do in series
    Sample $t$ from a uniform distribution on [0,1].
    if $t < w(v_{\mathrm{prop}}^{(i)}) \big/ w(v^{(i-1)})$ then
        $v^{(i)} = v_{\mathrm{prop}}^{(i)}$.
    else
        $v^{(i)} = v^{(i-1)}$.
```

3.2. Computational complexity and rank truncation.

The computational cost of the scalable RTO implementation derived above has two major sources. *First*, producing each RTO sample requires several optimization iterations. Each optimization iteration may evaluate the RTO objective function in (3.7), and MVPs with the linearization (3.8) and its adjoint, several times. These operations require evaluating the forward model, the actions of the linearized forward model and its adjoint, and the actions of the matrices Φ and Ψ several times. *Second*, for each RTO sample, we need to evaluate the determinant in (3.3) once to compute the weighting function. This in turn involves evaluating r MVPs with $\nabla G(v)$ and computing the determinant of an $r \times r$ matrix.

The following proposition summarizes the computational complexity of the operations involved in Algorithm 3.1.

Proposition 3.3. *We adopt the following assumptions on the computation of each RTO sample to establish the computational complexity of the scalable implementation of RTO.*

1. *On average, k_{opt} optimization iterations are needed to obtain each RTO sample. On average, k_{obj} objective function evaluations and k_{adj} MVPs with the linearization* (3.8) *and its adjoint are needed within each optimization iteration.*
2. *The number of floating point operations required to evaluate the forward model $G(v)$ is a function of the dimension of the discretized parameters, denoted by $C_1(n)$.*
3. *The number of floating point operations required to compute an MVP with the linearized forward model $\nabla G(v)$ and its adjoint is a function of the dimension of the discretized parameters, denoted by $C_2(n)$.*
4. *The data dimension is less than the parameter dimension, i.e., $m < n$.*

Then, counting the floating point operations needed to evaluate the objective function (3.7) *and the action of the linearization* (3.8)*, the number of floating point operations needed for each optimization iteration is*

$$\mathcal{O}\big((k_{\text{obj}} + k_{\text{adj}})(m\, r + n\, r)\big) + k_{\text{obj}}\, C_1(n) + k_{\text{adj}}\, C_2(n).$$

The number of floating point operations needed to evaluate the determinant in (3.3) *is $\mathcal{O}(m\, r^2 + r^3) + r\, C_2(n)$. Thus, a total of*

$$\mathcal{O}\big(k_{\text{opt}}\,(k_{\text{obj}} + k_{\text{adj}})(m\, r + n\, r) + m\, r^2 + r^3\big) + k_{\text{opt}}\, k_{\text{obj}}\, C_1(n) + (k_{\text{opt}}\, k_{\text{adj}} + r)\, C_2(n)$$

floating point operations are needed to compute one RTO sample, where the big–$\mathcal{O}$ term above refers to the total linear algebra cost, and the other terms refer to the total cost of evaluating $G(v)$ and the actions of $\nabla G(v)$.

Without loss of generality, the dimension n of the parameters is often proportional to the number of degrees of freedom of the discretized forward model, and thus the functions $C_1(n)$ and $C_2(n)$ are often *linear* or *quasi-linear* for scalable forward solvers, e.g., full multigrid solvers or preconditioned Krylov solvers. In this case, the computational complexity of each optimization iteration is dictated by the cost of solving the forward model and evaluating actions with its linearization. Similarly, the computational complexity of evaluating the determinant in (3.3) is dictated by the cost of the MVP with the linearized forward model. In contrast, without subspace acceleration, the complexity of computing the original objective function in (2.5) is quadratic in n, the complexity of computing the action of the linearization $Q^\top \nabla H(v)$ is also quadratic in n, and the complexity of computing the determinant in the weighting function (2.8) is cubic in n, since a dense matrix $Q \in \mathbb{R}^{(m+n)\times n}$ is involved. The cost of operating with the matrix Q will thus dominate the overall computational cost of standard RTO for high-dimensional problems. Subspace acceleration therefore significantly reduces the computational complexity of minimizing the RTO objective and calculating the determinant for each proposal. In addition, the size of the optimization problem in (3.7) is also reduced to the intrinsic rank r.

Rank truncation. For many inverse problems, the singular values of $\nabla G(v_{\text{ref}})$ decay quickly, as a consequence of a smoothing forward operator, noisy observations, and the correlation structure of the prior (where some smoothness is necessary to make the Bayesian inverse problem well-posed [43]). This fact is often used to reduce the parameter dimension of inverse problems by truncating the equivalent eigendecomposition (3.4) (cf. [6, 19, 42]), and hence to accelerate MCMC algorithms [15, 16, 29, 37] and to approximate posterior distributions [6, 19, 17, 47].

Using intuition derived from optimal posterior approximations in linear Bayesian inverse problems [42], we can derive heuristics for truncating the SVD in scalable RTO. This can be particular useful for cases where data are abundant, i.e., when $y \in \mathbb{R}^m$ is a large vector. Suppose we have a linear inverse problem, that is, $G(v) = Gv$ and $H(v) = Hv$. Computing the reduced SVD $\nabla G(v) \equiv G = \Psi\Lambda\Phi^\top$, the inverse of the posterior covariance is given by the Gauss–Newton Hessian, which has the eigendecomposition[1]

$$\nabla H(v_{\text{ref}})^\top \nabla H(v_{\text{ref}}) = \Phi(\Lambda^2 + \mathrm{I}_r)\Phi^\top + (\mathrm{I}_n - \Phi\Phi^\top).$$

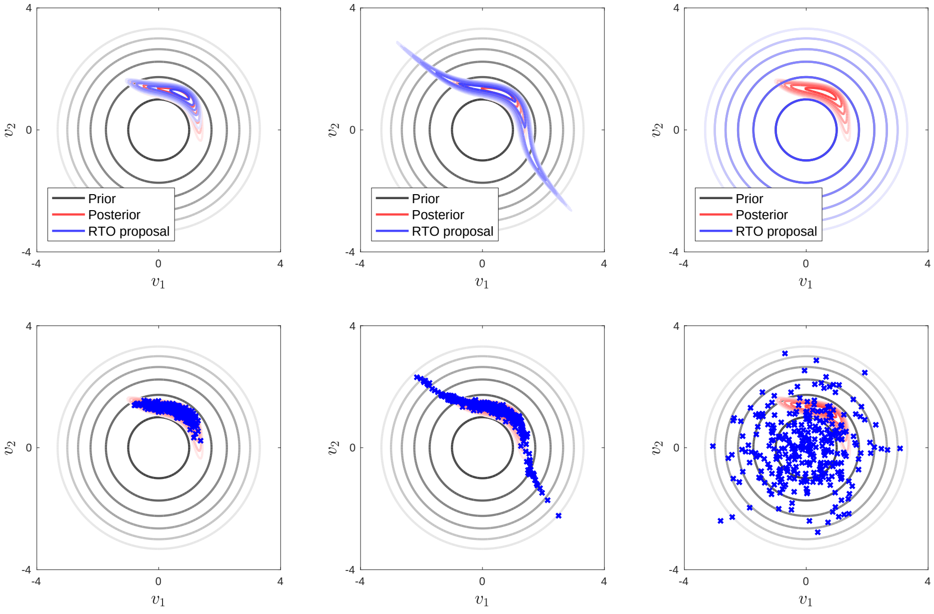
The subspace spanned by Φ contains the parameter directions where the posterior differs from the prior, since the prior (on the whitened variable v) has identity covariance matrix I_n. A small singular value λ_i implies that, along the corresponding right singular vector ϕ_i, the variance reduction from prior to posterior is small; in particular, the ratio of posterior to prior variance is nearly one [42]. We can thus neglect parameter directions corresponding to small singular values by truncating the SVD of $\nabla G(v_{\text{ref}})$. Suppose that the truncation rank is $t < r$; this leads to an approximate eigendecomposition in the form of

$$\nabla H(v_{\text{ref}})^\top \nabla H(v_{\text{ref}}) \approx \Phi_t(\Lambda_t^2 + \mathrm{I}_t)\Phi_t^\top + (\mathrm{I}_n - \Phi_t\,\Phi_t^\top), \tag{3.9}$$

where $\Phi_t \in \mathbb{R}^{n\times t}$ and $\Lambda_t \in \mathbb{R}^{t\times t}$ consist of the leading t right singular vectors and singular values, respectively. In the linear case, the RTO proposal is a Gaussian distribution with the covariance matrix given by the inverse of the truncated approximation in (3.9); this result directly follows from (3.5). As shown in [42], the inverse of the approximation in (3.9) is also an optimal approximation to the posterior covariance with respect to the natural (geodesic) distance on the manifold of symmetric positive definite matrices. In this situation, truncating the SVD of $\nabla G(v_{\text{ref}})$ for singular values that are smaller than one, e.g., 10^{-2} or 10^{-3}, yields negligible impact on the RTO proposal.

In nonlinear settings, we can adopt the same truncation strategy as a heuristic. The truncation will change RTO's map (3.1) and the resulting proposal distribution. Figure 2 shows the effect on RTO's proposal of truncating the SVD in a toy example with a nonlinear forward model and a standard normal prior. Truncation restricts the role of the data misfit term in the construction of the proposal distribution. As the rank r decreases, the proposal distribution becomes broader. In the extreme case, when r is truncated to zero, RTO's proposal reverts to the prior. Note, however, the non-Gaussianity of the RTO proposal for $r \geq 1$ in this nonlinear example. We will evaluate the impact of truncation on MCMC sampling efficiency in subsequent numerical examples.

[1] The linearization $\nabla H(v)$ does not depend on v for linear inverse problems. We use this notation for consistency with the nonlinear case.



(a) Rank = 2. (b) Rank = 1. (c) Rank = 0.

FIG. 2. *Truncating the SVD in a two-dimensional toy example with a nonlinear forward model and standard normal prior. Top: contours of the prior, posterior, and RTO's proposal density. Bottom: contours of the prior and posterior densities, and samples from RTO's proposal.*

4. RTO on function space. The scalable implementation presented in section 3 ensures that the computational cost of generating each RTO sample is dictated by the cost of evaluating the forward model and the adjoint model. When applying RTO as an independent proposal in the MH algorithm, or as the biasing distribution in self-normalized importance sampling, it is also critical to understand how its statistical performance (for instance, as measured by the acceptance rate of independence MH) depends on the dimension of the discretized parameters. To this end, we adopt the function space framework of [43] to analyze the RTO proposal. We will focus on the case of applying RTO within MCMC, though the analysis can easily be adapted to importance sampling. In this section, we will first provide background on MCMC in the function space setting, then interpret RTO's mapping in function space, and conclude by establishing sufficient conditions such that the statistical performance of RTO is invariant to the dimension of discretised parameters.

4.1. Function space MCMC. To be aligned with the framework of [43], we will consider the target distribution on the original parameter u (rather than the "whitened" parameter v), in a function space setting. To preserve interpretability, we will use the same notation in the function space setting as we do in the finite-dimensional setting to represent the parameters, prior mean, and prior covariance. One exception is that we will use $\Gamma_{\text{pr}}^{1/2}$ to denote the symmetric square root of the prior covariance operator, which is equivalent to any square root of the prior covariance up to a rotation.

We suppose that the parameter u is an element of a separable Hilbert space $\mathcal{H}$, endowed with a Gaussian prior measure μ_{pr} such that the prior covariance Γ_{pr} is a self-adjoint, positive definite, and trace-class operator on $\mathcal{H}$. The inner product on $\mathcal{H}$ is denoted by $\langle \cdot\,,\cdot\rangle_{\mathcal{H}}$, with the associated norm denoted by $\|\cdot\|_{\mathcal{H}}$. For brevity, where misinterpretation is not possible, we will drop the subscript $\mathcal{H}$. We assume that the data y remain finite dimensional, i.e., $y \in \mathbb{R}^m$, $\Gamma_{\text{obs}} \in \mathbb{R}^{m\times m}$, and $F : \mathcal{H} \to \mathbb{R}^m$ for $m < \infty$. This way, the target probability measure is expressed by the Radon–Nikodym derivative

$$\frac{\mathrm{d}\mu_{\text{tar}}}{\mathrm{d}\mu_{\text{pr}}}(u) \propto \exp\Big(-\frac{1}{2}(y - F(u))^\top \Gamma_{\text{obs}}^{-1}(y - F(u))\Big)$$

with respect to the the prior measure. The MH algorithm defines a Markov chain of random functions, asymptotically distributed according to the target measure, in the following way: given the current state of the Markov chain, $U^{(k)} = u$, a candidate state u' is drawn from a proposal $q(u,\cdot)$. Define the following pair of measures on $\mathcal{H}\times\mathcal{H}$:

$$\begin{aligned} \nu(du, du') &= q(u, du')\mu_{\text{tar}}(du),\\ \nu^{\perp}(du, du') &= q(u', du)\mu_{\text{tar}}(du'). \end{aligned} \tag{4.1}$$

Then, the next state of the Markov chain is set to $U^{(k+1)} = u'$ with probability

$$\alpha(u, u') = \min\Big\{1, \frac{d\nu^{\perp}}{d\nu}(u, u')\Big\}, \tag{4.2}$$

and to $U^{(k+1)} = u$ otherwise.

For a continuously differentiable (as in Assumption 2.1) and sufficiently bounded (as defined in Assumption 2.7 of [43]) forward model F, [43] shows that the target measure is dominated by the prior measure. As a result, refinements of the corresponding finite-dimensional target measure (induced by refinements of the parameter discretization) will converge to an infinite-dimensional limit. To make the acceptance probability of MH then invariant to parameter discretization, i.e., convergent to some positive infinite-dimensional limit and hence yielding a valid transition kernel [44], we require the absolute continuity condition $\nu^{\perp} \ll \nu$. We will refer to a MH algorithm as *well-defined* if this absolute continuity condition holds. Note that many MCMC methods designed for finite-dimensional problems may not be well-defined on $\mathcal{H}$—they may have vanishing acceptance probability and vanishing effective sample size with increasing parameter discretization dimension [39, 40]. For example, the acceptance probability of an MH algorithm using the standard random walk proposal scales as $\mathcal{O}(n^{-1})$ with parameter dimension, and thus it is not suitable for high-dimensional problems. We aim to show that MH with an RTO proposal is well-defined on $\mathcal{H}$.

4.2. RTO mapping in function space. Recall that the Cameron–Martin space associated with the prior measure μ_{pr}, $\mathcal{H}_{\text{CM}} = \Gamma_{\text{pr}}^{1/2}\,\mathcal{H} \subset \mathcal{H}$, is equipped with the inner product

$$\langle a, b\rangle_{\mathcal{H}_{\text{CM}}} := \langle a, b\rangle_{\Gamma_{\text{pr}}^{-1}} = \Big\langle \Gamma_{\text{pr}}^{-\frac{1}{2}}\, a, \Gamma_{\text{pr}}^{-\frac{1}{2}}\, b\Big\rangle$$

for any $a, b \in \mathcal{H}_{\text{CM}}$. Here we will show that a sample generated by the RTO mapping is a modification of a random function drawn from the prior measure along a finite-dimensional subspace of the Cameron–Martin space.

We first consider the properties of the rank-r reduced SVD of the linearized forward model $\nabla G(v_{\text{ref}}) = \Psi\Lambda\Phi^\top$ in the function space setting. The right singular vectors $\phi_1, \phi_2, \ldots, \phi_r$ are also eigenfunctions of the eigenvalue problem

$$\nabla G(v_{\text{ref}})^{\natural}\nabla G(v_{\text{ref}})\,\phi_i = \lambda_i\phi_i, \quad i = 1, \ldots, r,$$

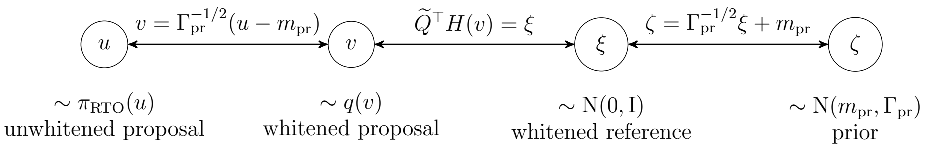

$v = \Gamma_{\mathrm{pr}}^{-1/2}(u - m_{\mathrm{pr}})$
$\widetilde{Q}^\top H(v) = \xi$
$\zeta = \Gamma_{\mathrm{pr}}^{-1/2}\xi + m_{\mathrm{pr}}$
u
v
ξ
ζ
$\sim \pi_{\mathrm{RTO}}(u)$
unwhitened proposal
$\sim q(v)$
whitened proposal
$\sim \mathrm{N}(0, \mathrm{I})$
whitened reference
$\sim \mathrm{N}(m_{\mathrm{pr}}, \Gamma_{\mathrm{pr}})$
prior

FIG. 3. *Relationship between four random variables, $u, \zeta \in \mathcal{H}$, $v, \xi \in \Gamma_{\mathrm{pr}}^{-\frac{1}{2}}\mathcal{H}$.*

where $\nabla G(v_{\mathrm{ref}})^\natural$ denotes the adjoint of the operator $\nabla G(v_{\mathrm{ref}})$. Recalling the whitening transform introduced in section 2.1, we have

$$\nabla G(v_{\mathrm{ref}})^\natural \nabla G(v_{\mathrm{ref}}) = \Gamma_{\mathrm{pr}}^{\frac{1}{2}} \nabla F\left(u_{\mathrm{ref}}\right)^\natural \Gamma_{\mathrm{obs}}^{-1} \nabla F\left(u_{\mathrm{ref}}\right) \Gamma_{\mathrm{pr}}^{\frac{1}{2}},$$

where $\nabla F : \mathcal{H} \to \mathbb{R}^m$ is the Fréchet derivative of the forward model and $\nabla F\left(u_{\mathrm{ref}}\right)^\natural$ is its adjoint. Defining a new set of functions

$$\chi_i = \Gamma_{\mathrm{pr}}^{1/2}\phi_i, \tag{4.3}$$

we also have an equivalent eigenvalue problem

$$\Gamma_{\mathrm{pr}} \nabla F\left(u_{\mathrm{ref}}\right)^\natural \Gamma_{\mathrm{obs}}^{-1} \nabla F\left(u_{\mathrm{ref}}\right) \chi_i = \lambda_i^2 \chi_i, \quad i = 1, \ldots, r.$$

Since the operator $\nabla F\left(u_{\mathrm{ref}}\right)^\natural \Gamma_{\mathrm{obs}}^{-1} \nabla F\left(u_{\mathrm{ref}}\right)$ is self-adjoint and has finite rank $r \leq m$ in the case of finite-dimensional data ($m < \infty$), we have that the eigenfunctions $\chi_i \in \Gamma_{\mathrm{pr}}\mathcal{H}$ and that the right singular vectors $\phi_i \in \Gamma_{\mathrm{pr}}^{1/2}\mathcal{H}$ for $i = 1, \ldots, r$.

Remark 4.1. Both $\{\chi_1, \chi_2, \ldots, \chi_r\}$ and $\{\phi_1, \phi_2, \ldots, \phi_r\}$ span finite-dimensional subspaces in the Cameron–Martin space. The basis functions $\{\phi_1, \phi_2, \ldots, \phi_r\}$ are orthogonal with respect to the inner product $\langle \cdot, \cdot \rangle$, whereas the basis functions $\{\chi_1, \chi_2, \ldots, \chi_r\}$ are orthogonal with respect to the Cameron–Martin inner product $\langle \cdot, \cdot \rangle_{\Gamma_{\mathrm{pr}}^{-1}}$.

In Figure 3, we specify the relationship between four random variables: $u, \zeta \in \mathcal{H}$, and $v, \xi \in \Gamma_{\mathrm{pr}}^{-\frac{1}{2}}\mathcal{H}$, where ζ is a newly defined random variable distributed according to the prior. Here we use u and v to denote random variables distributed according to the unwhitened and whitened RTO measures, respectively, rather than the corresponding target measures. This way, we have the identity

$$\langle v, \phi_i \rangle = \langle \Gamma_{\mathrm{pr}}^{-\frac{1}{2}}(u - m_{\mathrm{pr}}), \Gamma_{\mathrm{pr}}^{-\frac{1}{2}}\chi_i \rangle = \langle u - m_{\mathrm{pr}}, \chi_i \rangle_{\Gamma_{\mathrm{pr}}^{-1}}$$

for any $v \in \Gamma_{\mathrm{pr}}^{-\frac{1}{2}}\mathcal{H}$ and any right singular vector $\phi_i \in \Gamma_{\mathrm{pr}}^{1/2}\mathcal{H}$.

Given the basis functions $\{\chi_1, \chi_2, \ldots, \chi_r\}$, we introduce a linear map $R : \mathcal{H} \to \mathbb{R}^r$ whose components are

$$R_i(u) = \langle u, \chi_i \rangle_{\Gamma_{\mathrm{pr}}^{-1}}, \quad i = 1, \ldots, r,$$

and a projector $P : \mathcal{H} \to \mathrm{span}\{\chi_1, \chi_2, \ldots, \chi_r\}$ specified as

$$Pu = \sum_{i=1}^{r} \chi_i \, R_i(u).$$

For a random function $\zeta \in \mathcal{H}$ drawn from the prior measure, we can then express the RTO mapping (3.1) in the unwhitened coordinates

$$\begin{cases} (\mathrm{I} - P)(\zeta - m_{\mathrm{pr}}) &= (\mathrm{I} - P)(u - m_{\mathrm{pr}}), \\ P\,(\zeta - m_{\mathrm{pr}}) &= X\Big[(\Lambda^2 + \mathrm{I})^{-\frac{1}{2}}\big(R(u - m_{\mathrm{pr}}) + \Lambda\Psi^\top S_{\mathrm{obs}}^{-1}(F(u) - y)\big)\Big], \end{cases} \tag{4.4}$$

where $X = [\chi_1, \chi_2, \ldots, \chi_r]$. Analogously to the splitting of the RTO solution in the scalable implementation, we define

$$u_r = R(u - m_{\mathrm{pr}}) \quad \text{and} \quad u = Xu_r + u_\perp + m_{\mathrm{pr}}, \tag{4.5}$$

and projected random variables

$$\zeta_r = R(\zeta - m_{\mathrm{pr}}) \quad \text{and} \quad \zeta_\perp = (\mathrm{I} - P)(\zeta - m_{\mathrm{pr}}), \tag{4.6}$$

where $u_\perp$ and $\zeta_\perp$ are elements of the complement of range(X). Then, we can solve the nonlinear system of equations (4.4) by first letting $u_\perp = \zeta_\perp$ and then solving the r-dimensional system of equations

$$\Theta(u_r; u_\perp) = \zeta_r \tag{4.7}$$

for a given $u_\perp$, where the function

$$\Theta(u_r; u_\perp) = (\Lambda^2 + \mathrm{I})^{-\frac{1}{2}}\Big[u_r + \Lambda\Psi^\top S_{\mathrm{obs}}^{-1}(F(Xu_r + u_\perp + m_{\mathrm{pr}}) - y)\Big] \tag{4.8}$$

is an affine transformation of the nonlinear forward model F.

Remark 4.2. For problems with finite-dimensional data, the RTO mapping necessarily modifies a random function drawn from the prior measure only in the finite-dimensional subspace spanned by $\{\chi_1, \chi_2, \ldots, \chi_r\}$, which is a subspace of the Cameron–Martin space. This is well aligned with the nature of function space inverse problems, where the update from the prior to the posterior is expected to take place in the Cameron–Martin space. Other accelerations of function space MCMC, e.g., [15, 41], adopt similar approaches to modify their algorithms in some finite-dimensional subspace of the Cameron–Martin space. For problems with functional (infinite-dimensional) data, as long as the equivalence between the target measure and the prior measure can be established, one can truncate the SVD and apply RTO as in the finite data case. Such a truncation can also be the key to managing overall computational complexity.

4.3. Well-definedness of RTO on function space.

We will first prove that, under certain conditions, the RTO measure π_{RTO} is equivalent to the prior μ_{pr}. Under these conditions, we can then show that RTO is well-defined on $\mathcal{H}$.

THEOREM 4.3. *Let ζ be a random variable distributed according to the prior measure μ_{pr}, u be the random variable defined through the mapping in (4.4), and μ_{RTO} be the measure induced by u. Denote the subspace* $\mathrm{span}(\chi_1, \chi_2, \ldots, \chi_{\mathrm{r}})$ *and its complement by W and $W^\perp$, respectively. Without loss of generality, let the prior mean be zero. Suppose that for all $a \in \mathbb{R}^r$ and $b \in W^\perp$, the mapping*

$$a \mapsto a + \Lambda\Psi^\top S_{\mathrm{obs}}^{-1}F(Xa + b)$$

is Lipschitz continuous, injective, and its inverse is Lipschitz continuous. Then, the RTO measure μ_{RTO} is equivalent to the prior μ_{pr}.

Proof. In this proof only, we will employ the probability triplet $(\Omega, \mathcal{F}, \mathbb{P})$ and describe the random function formally as the map $u : \Omega \to \mathcal{H}$. We assume that the measurable space $(\Omega, \mathcal{F})$ is a Radon space. The four random variables in (4.5)–(4.7) can be defined as

$$\zeta_r : \Omega \to \mathbb{R}^r, \qquad u_r : \Omega \to \mathbb{R}^r, \qquad \zeta_\perp : \Omega \to W^\perp, \qquad u_\perp : \Omega \to W^\perp.$$

We use $\mathcal{B}(\cdot)$ to denote the Borel algebra. Let the notation $\mathbb{P}^u$ denote the pushforward measure of $\mathbb{P}$ through the mapping u:

$$\mathbb{P}^u := u_\sharp \mathbb{P} = \mathbb{P}\left(u^{-1}(\cdot)\right) = \mathbb{P}\left(u \in \,\cdot\,\right).$$

Using this notation, we have $\mu_{\text{pr}} = \mathbb{P}^\zeta$ and $\mu_{\text{RTO}} = \mathbb{P}^u$.

Since, under the mapping (4.4), the infinite-dimensional random variables $\zeta_\perp$ and $u_\perp$ take the same value, we use the regular conditional probability $\nu : W^\perp \times \mathcal{F} \to [0, 1]$ of the form

$$\nu : (b, A) \to \nu(b, A) = \mathbb{P}\left(u \in A \,|\, u_\perp = b\right) \quad \forall A \in \mathcal{F},\ \forall b \in W^\perp,$$

to analyze the RTO measure. For any $A \in \mathcal{B}(\mathcal{H})$ and $b \in W^\perp$, we define the set

$$A_r(b) := \{a \in \mathbb{R}^r \,|\, Xa + b \in A\}.$$

Then, for any $A \in \mathcal{B}(\mathcal{H})$, the RTO measure can be expressed in a conditional form

$$\mathbb{P}^u(A) = \mathbb{P}\left(\{u \in A\} \cap \{u_\perp \in W^\perp\}\right) = \int_{W^\perp} \mathbb{P}\left(u \in A \,|\, u_\perp = b\right) \mathbb{P}^{u_\perp}(\mathrm{d}b).$$

Given a fixed $b \in W^\perp$, RTO solves the equation $\Theta(u_r; b) = \zeta_r$, and thus we have

$$\begin{aligned} \mathbb{P}\left(u \in A \,|\, u_\perp = b\right) &= \mathbb{P}\left(u_r \in A_r(b) \,|\, u_\perp = b\right) \\ &= \mathbb{P}\left(\zeta_r \in \Theta(A_r(b); b) \,|\, \zeta_\perp = b\right) = \nu\left(b, \zeta_r^{-1} \circ \Theta(A_r(b); b)\right). \end{aligned}$$

This way, the RTO measure can be expressed in terms of the measure $\mathbb{P}^{u_\perp} = \mathbb{P}^{\zeta_\perp}$ and the conditional measure $\nu\left(b, \zeta_r^{-1}(\cdot)\right)$. This leads to

$$\mathbb{P}^u(A) = \int_{W^\perp} \nu\left(b, \zeta_r^{-1} \circ \Theta(A_r(b); b)\right) \mathbb{P}^{\zeta_\perp}(\mathrm{d}b). \tag{4.9}$$

The conditional measure $\nu\left(b, \zeta_r^{-1}(\cdot)\right) = \mathbb{P}(\zeta_r \in \cdot \,|\, \zeta_\perp = b)$ is the measure of a finite number of directions, ζ_r, of the prior conditioned on a particular value, $\zeta_\perp = b$. Since the basis functions $\{\chi_1, \chi_2, \ldots, \chi_r\}$ are orthogonal with respect to the Cameron–Martin inner product $\langle \cdot, \cdot \rangle_{\Gamma_{\text{pr}}^{-1}}$, the two random variables ζ_r and $\zeta_\perp$ are independent (see Proposition 1.26 of [38]). Hence, the conditional measure $\nu\left(b, \zeta_r^{-1}(\cdot)\right)$ is equivalent to the law of ζ_r. Let π_{ζ_r} denote its probability density function. Then, we have

$$\begin{aligned} \nu\left(b, \zeta_r^{-1} \circ \Theta(A_r(b); b)\right) &= \int_{\Theta(A_r(b); b)} \pi_{\zeta_r}(a) \mathrm{d}a \\ &= \int_{A_r(b)} \pi_{\zeta_r} \circ \Theta(a; b) \left|\det \nabla \Theta(a; b)\right| \mathrm{d}a \\ &= \int_{A_r(b)} \frac{\pi_{\zeta_r} \circ \Theta(a; b)}{\pi_{\zeta_r}(a)} \left|\det \nabla \Theta(a; b)\right| \, \pi_{\zeta_r}(a) \, \mathrm{d}a \\ &= \int_{A_r(b)} \frac{\pi_{\zeta_r} \circ \Theta(a; b)}{\pi_{\zeta_r}(a)} \left|\det \nabla \Theta(a; b)\right| \, \nu\left(b, \zeta_r^{-1}(\mathrm{d}a)\right). \end{aligned} \tag{4.10}$$

The change of variables in the above expression uses the fact that $\Theta(\,\cdot\,;b)$ is Lipschitz continuous and injective, and that its inverse is Lipschitz continuous. Note that for almost all $b \in W^\perp$ and $a \in \mathbb{R}^r$, the expression

$$\mathfrak{R}(a;b) := \frac{\pi_{\zeta_r} \circ \Theta(a;b)}{\pi_{\zeta_r}(a)} \, |\det \nabla\Theta(a;b)|$$

is positive. Substituting (4.10) into (4.9) and using the change of variables $b = (\mathrm{I}-P)z$ and $a = R(z)$ for any $z \in \mathcal{H}$, we obtain the RTO measure in the form

$$\mathbb{P}^u(A) = \int_{W^\perp}\int_{A_r(b)} \mathfrak{R}(a;b)\,\nu\left(b,\zeta_r^{-1}(\mathrm{d}a)\right)\mathbb{P}^{\zeta_\perp}(\mathrm{d}b) = \int_A \mathfrak{R}\Big(R(z);(\mathrm{I}-P)z\Big)\mathbb{P}^{\zeta}(\mathrm{d}z). \tag{4.11}$$

Therefore, the Radon–Nikodym derivative of the RTO measure with respect to the prior measure,

$$\frac{\mathrm{d}\mu_{\mathrm{RTO}}}{\mathrm{d}\mu_{\mathrm{pr}}}(u) = \mathfrak{R}\Big(R(z);(\mathrm{I}-P)z\Big),$$

is positive almost everywhere. This implies that μ_{RTO} is equivalent to μ_{pr}. □

Theorem 4.3 implies that the MH algorithm using RTO as its independence proposal yields dimension-independent performance in the function space setting. We formalize this notion in the following theorem.

THEOREM 4.4. *Suppose that the target measure μ_{tar} is equivalent to the prior measure μ_{pr}, i.e., $\mu_{\mathrm{tar}} \sim \mu_{\mathrm{pr}}$. Under the assumptions of Theorem 4.3, the acceptance probability of the MH algorithm using RTO as its independence proposal is positive almost surely with respect to $\mu_{\mathrm{pr}} \times \mu_{\mathrm{pr}}$.*

Proof. Following the result of Theorem 4.3 and the condition $\mu_{\mathrm{pr}} \sim \mu_{\mathrm{tar}}$, the Radon–Nikodym derivative of the target measure with respect to the RTO measure,

$$\omega(u) = \frac{\mathrm{d}\mu_{\mathrm{tar}}}{\mathrm{d}\mu_{\mathrm{RTO}}}(u),$$

is μ_{pr}-almost surely positive. The rest of the proof is a special case of Theorem 5.1 in [43]. Since RTO is an independent proposal, the resulting MH proposal measure becomes $q(u,\mathrm{d}u') = \mu_{\mathrm{RTO}}(\mathrm{d}u')$, and the pair of transition measures of MH become

$$\begin{aligned}\nu(\mathrm{d}u,\mathrm{d}u') &= \mu_{\mathrm{RTO}}(\mathrm{d}u')\,\mu_{\mathrm{tar}}(\mathrm{d}u),\\ \nu^\perp(\mathrm{d}u,\mathrm{d}u') &= \mu_{\mathrm{RTO}}(\mathrm{d}u)\,\mu_{\mathrm{tar}}(\mathrm{d}u').\end{aligned}$$

This way, the acceptance probability can be expressed as

$$\alpha(u,u') = \min\left(1, \frac{\mathrm{d}\mu_{\mathrm{tar}}}{\mathrm{d}\mu_{\mathrm{RTO}}}(u')\Big/\frac{\mathrm{d}\mu_{\mathrm{tar}}}{\mathrm{d}\mu_{\mathrm{RTO}}}(u)\right) = \min\left(1, \frac{\omega(u')}{\omega(u)}\right).$$

Because $\omega(u)$ is positive μ_{pr}-almost surely, the acceptance probability $\alpha(u,u')$ is positive $\mu_{\mathrm{pr}} \times \mu_{\mathrm{pr}}$-almost surely. □

RTO–MH is therefore well-defined in a function-space setting, under the conditions in Theorem 4.3. Thus refining the parameter discretization in a discrete setting should not diminish RTO–MH's sampling efficiency. Note that the Radon–Nikodym derivative $\omega(u)$ is also the importance ratio used in self-normalized importance sampling. Ensuring that $\omega(u)$ is positive (almost surely) can make the effective sample size of the self-normalized importance sampling estimator invariant to the discretized parameter dimension; see [1] and references therein for formal justifications.

5. Example 1: One-dimensional elliptic PDE.

The previous section provided a theoretical argument for RTO's dimension independence. This section numerically explores the factors that influence its sampling performance, using a simple one-dimensional elliptic PDE inverse problem. We describe the setup of the test case (section 5.1) and then explore the effects of parameter dimension (section 5.2) and observational noise (section 5.3). We conclude by comparing the performances of RTO and pCN (section 5.4).

5.1. Problem setup.

The diffusion equation is used to model the spatial distribution of many physical quantities, such as temperature, electrostatic potential, or pressure in porous media. We consider the following stationary diffusion equation,

$$-\frac{\mathrm{d}}{\mathrm{d}x}\left(\kappa(x)\frac{\mathrm{d}p}{\mathrm{d}x}(x)\right) = f(x), \quad 0 < x < 1,$$

with boundary conditions

$$\kappa(0)\frac{\mathrm{d}p}{\mathrm{d}x}(0) = -1, \qquad p(1) = 1,$$

and source term f. The diffusion coefficient κ is endowed with a log-normal prior distribution. In particular, $\log \kappa$ is a Gaussian process with a Laplace-like differential operator as its precision operator. After discretization on a uniform grid with n nodes, κ is thus specified as

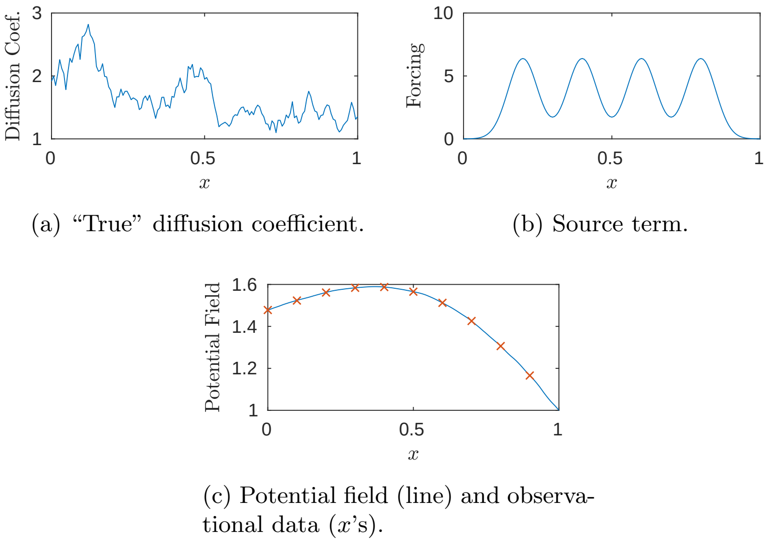
$$\kappa = 1.5 \exp\left(S_{\mathrm{pr}} v\right) + 0.1, \qquad S_{\mathrm{pr}}^{-1} = \sqrt{n}\begin{bmatrix} \sqrt{n} & & & & \sqrt{n} \\ -1 & 1 & & & \\ & -1 & 1 & & \\ & & & \ddots & \\ & & & -1 & 1 \end{bmatrix},$$

where $v \in \mathbb{R}^n$ is a vector of independent standard normals and we have abused notation so that $\kappa \in \mathbb{R}^n$ immediately above as well. For any realization of κ, the equation is solved numerically using finite differences with the three point central difference stencil. Derivatives of the potential field p with respect to κ are evaluated using the matrix-free adjoint model. In this setting, the dimension of discretized parameters is the same as the degrees of freedom in the forward model. Computing $S_{\mathrm{pr}}v$, solving the forward model, and solving the adjoint model (for one MVP with the Jacobian) all require $\mathcal{O}(n)$ floating point operations.

For the inverse problem, we suppose that the potential field is observed, with additive Gaussian noise, at nine equally spaced points along the domain. Our goal is to condition the field κ on these observations. We generate synthetic data using a mesh size of 151, which does not correspond to any mesh size used in solving the inverse problem, avoiding an inverse crime. The "true" diffusion coefficient, source term, potential field, and data are depicted in Figure 4.

5.2. Influence of parameter dimension.

In our first experiment, we solve the Bayesian inverse problem using RTO for a series of parameter dimensions ranging from $n = 41$ to $n = 10241$. We fix the observational noise standard deviation to 10^{-5} and, at each parameter dimension, run an MCMC chain of 5000 steps. The chains are started at the posterior mode. As shown in Figure 5, the posterior distributions obtained for the different discretizations match quite closely. As shown in Table 1, the

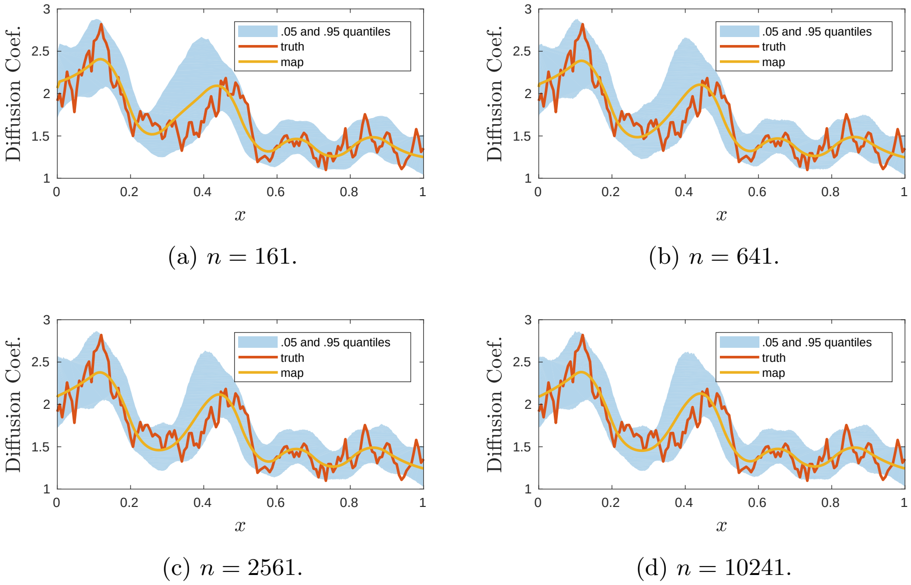


(a) "True" diffusion coefficient.

(b) Source term.

(c) Potential field (line) and observational data (x's).

Fig. 4. *Elliptic PDE problem setup.*



(a) $n = 161$.

(b) $n = 641$.

(c) $n = 2561$.

(d) $n = 10241$.

Fig. 5. *Summary statistics of posterior distributions computed via RTO-MH with varying parameter dimension n. 90% marginal credibility intervals (blue shaded region), true diffusivity coefficient (red line), and MAP estimate (yellow line).*

acceptance rate and effective sample size (ESS) are both high and essentially constant with respect to parameter dimension. (We report the median ESS over all components of the n-dimensional chain.) These results provide an empirical demonstration of RTO's dimension independence, meaning that the number of MCMC steps required to obtain a single effectively independent sample is independent of n.

The number of optimization iterations in each MCMC step is also roughly constant in n. To solve each optimization problem, we use the nonlinear least-squares

TABLE 1

ESS, average acceptance rate, and average number of optimization iterations per step of RTO, with varying parameter dimension. MCMC chain length is 5000 *steps.*

Parameter Dim.	41	81	161	321	641	1281	2561	5121	10241
ESS	4268.9	4206.7	4307.1	4343.5	4544.8	4464.5	4523.3	4484.9	4532.2
Acceptance rate	0.928	0.926	0.932	0.936	0.948	0.950	0.954	0.950	0.953
Opt. iterations	170.74	209.12	273.03	324.04	357.76	307.50	198.81	165.06	142.25

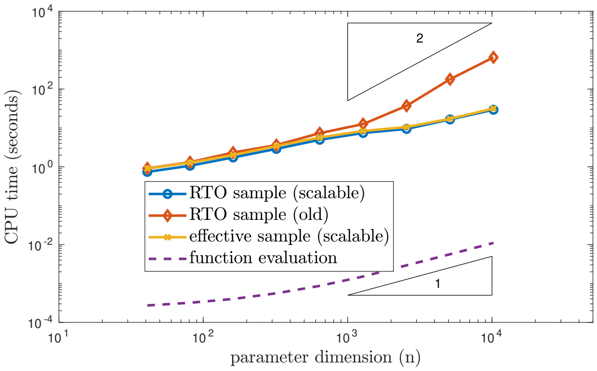
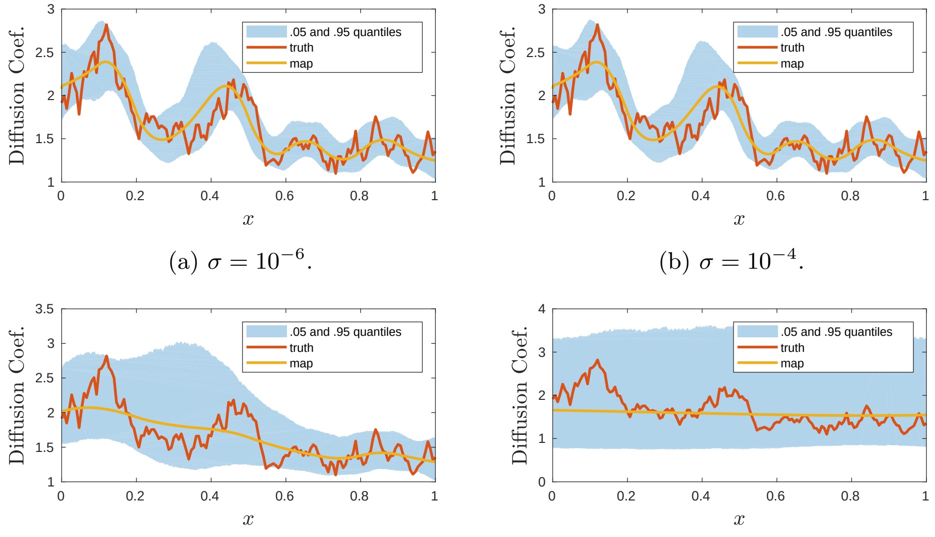


FIG. 6. *Computational cost for elements of RTO, varying parameter dimension.*

solver in MATLAB, provided with Jacobian-vector products. The solver uses a trust-region-reflective algorithm [11, 12] where each iteration approximately solves a large linear system using preconditioned CGs. We set the starting point for each sequence of optimization iterations to the posterior mode. The primary stopping criterion is a function tolerance (i.e., a lower bound on the change in the value of the objective) of 10^{-6}, which is below the level of discretization error.

Figure 6 shows the CPU times needed to generate one effectively independent sample, to generate one RTO sample, and to evaluate the forward model once. All three lines suggest that the CPU times increase linearly with the discretized parameter dimension. This confirms our analysis of the computational complexity of RTO in Proposition 3.3—in this example, the computational complexities of both the forward model and the RTO map are linear. It also implies that it takes the same number of MCMC steps to obtain a desired accuracy regardless of the discretized parameter dimension. This confirms our finding in section 4. We also report the CPU time for the standard RTO (with a dense matrix $Q \in \mathbb{R}^{(m+n)\times n}$) to generate one sample (red line in Figure 6). In this example, we observe that the computational complexity of the standard RTO is quadratic with the parameter dimension.

5.3. Influence of observational noise. In our second experiment, we examine the effect of observational noise magnitude on the sampling efficiency of RTO. We fix the parameter dimension to $n = 641$ and scan through observational noise standard deviations ranging from 10^{-7} to 10^0, which correspond to signal-to-noise ratios ranging from 1.5×10^7 to 1.5. Once again we run MCMC chains of length 5000. Changing the observational noise magnitude changes the posterior distribution, as shown in Figure 7. With extremely small observational noise, the probability mass of the posterior concentrates on the manifold where the parameter values yield outputs that exactly match the data. Generally, this collapse makes the posterior more difficult to simulate



(a) $\sigma = 10^{-6}$.

(b) $\sigma = 10^{-4}$.

(c) $\sigma = 10^{-2}$.

(d) $\sigma = 10^{0}$.

Fig. 7. *Summary statistics of posterior distributions computed via RTO-MH with varying observational noise σ. 90% marginal credibility intervals (blue shaded region), true diffusivity coefficient (red line), and MAP estimate (yellow line).*

Table 2

ESS, average acceptance rate, and average number of optimization iterations per step for RTO for varying observational noise magnitude. Chain length of 5000*.*

Noise std deviation	10^{-7}	10^{-6}	10^{-5}	10^{-4}	10^{-3}	10^{-2}	10^{-1}	10^{0}	10^{1}
Numerical ESS	4504.8	4427.4	4349.9	4423.0	4415.1	4187.2	4317.7	4476.9	5000.0
Acceptance rate	0.946	0.944	0.941	0.945	0.935	0.924	0.939	0.959	0.999
Opt. iterations	567.64	495.41	363.71	296.55	89.07	8.32	5.70	4.70	3.31

using most MCMC methods. In the case of RTO, it makes the optimization problems harder to solve. As shown in Table 2, even though the ESS and acceptance rate remain relatively constant with varying observational noise, the number of optimization iterations required to obtain each sample increases as the observational noise becomes very small. Thus, as the observational noise shrinks, more function evaluations are required for each MCMC step. This behavior is also illustrated in Figure 8, where the CPU time for a single function evaluation is constant, but the time for one MCMC step and for one independent sample increases.

Of course, the number of optimization iterations at each step depends on the choice of stopping tolerance. In these experiments, we fix the function tolerance (see section 5.2) to 10^{-6}. However, the forward model and the observed data enter the RTO objective function through the whitening transform (cf. section 2.1). This implicitly normalizes the observational noise by the standard deviation. This way, a fixed tolerance implicitly imposes an increasingly stringent condition for smaller observational noise, which explains the higher number of function evaluations required. Overall, though, these results suggest that RTO can be applied to inverse problems with extremely small observational noise provided that solving the optimization problems remains tractable.

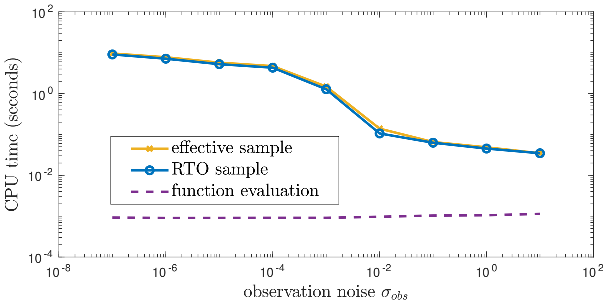
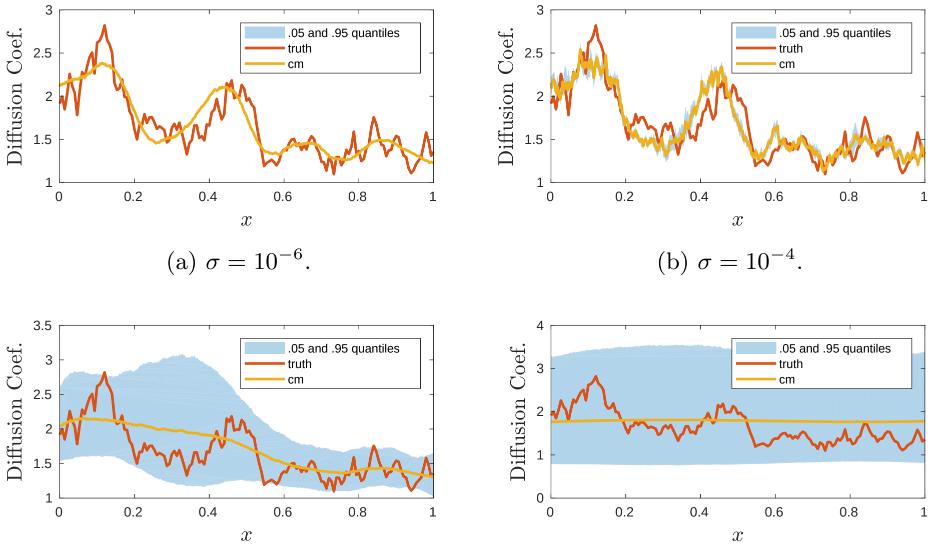


FIG. 8. *Computational cost for elements of RTO, varying observational noise.*



(a) $\sigma = 10^{-6}$.

(b) $\sigma = 10^{-4}$.

(c) $\sigma = 10^{-2}$.

(d) $\sigma = 10^{0}$.

FIG. 9. *Summary statistics of posterior distributions computed through pCN, varying observational noise σ. 90% credibility intervals (blue shaded region), true diffusivity coefficient (red line), and the conditional mean estimate (yellow line). The MCMC chain does not converge for $\sigma = 10^{-6}$ and $\sigma = 10^{-4}$.*

5.4. Comparing RTO with pCN. In our third experiment, we compare the computational efficiency of RTO and pCN [14]. The two algorithms are both dimension independent. We fix the parameter dimension to $n = 641$ and compare the algorithms' performance on inverse problems with different observational noise standard deviations, ranging from 10^{-6} to 10^{0}. For pCN, we use a chain length of 5×10^6 and remove the first 50% of the samples as burn-in. We manually tune the step size of pCN to obtain the largest empirical ESS. As shown in Figure 9, the posterior marginals from pCN match those obtained with RTO for the two larger observational noise values. For the two smaller observational noise values, however, pCN does not converge. In particular, examination of Figure 9 and of MCMC trace plots for the smaller noise cases shows that the pCN chain does not travel far from its starting point. Table 3 reveals that RTO requires less computational time per

TABLE 3
Comparing computational cost for RTO and pCN.

	CPU time (seconds) per ESS	
Observational noise σ	RTO	pCN
10^{-6}	7.772	$1.193 \cdot 10^{3*}$
10^{-4}	4.712	$1.103 \cdot 10^{3*}$
10^{-2}	0.139	7.739
10^{0}	0.049	0.250

*Estimated from a nonconverged MCMC chain. Actual values may be higher.

independent sample in *all* cases, even when the observational noise is larger. (Note that this performance metric, time per ESS, normalizes away the impact of different chain lengths.)

In this numerical example, RTO thus outperforms pCN by a large margin. Moreover, in the two cases with smaller observational noise, RTO is the only algorithm that produces meaningful estimates of the posterior. In summary, we find that RTO's sampling performance is robust to parameter dimension and observational noise, and can be more efficient than pCN.

6. Example 2: Two-dimensional parabolic PDE.

To further demonstrate the efficacy of RTO, we solve the inverse problem of identifying the coefficient of a two-dimensional parabolic PDE from point observations of its solution. Consider the problem domain $\Omega = [0, 3] \times [0, 1]$ with boundary $\partial\Omega$. We denote the spatial coordinate by $x = (x_1, x_2) \in \Omega$. We model the time-varying potential (solution) field $p(x, t)$ for a given conductivity (coefficient) field $\kappa(x)$ and forcing function $f(x, t)$ using the heat equation

$$\frac{\mathrm{d}p(x,t)}{\mathrm{d}t} = \nabla \cdot (\kappa(x)\nabla p(x,t)) + f(x,t), \quad x \in \Omega,\ t \in [0, T], \tag{6.1}$$

where $T = 2$. Parabolic PDEs of this type are widely used in modeling groundwater flow, optical diffusion tomography, the diffusion of thermal energy, and numerous other common scenarios for inverse problems. Let

$$\partial\Omega_{\mathrm{n}} = \{x \in \partial\Omega \,|\, x_2 = 0\} \cup \{x \in \partial\Omega \,|\, x_2 = 1\}$$

denote the top and bottom boundaries, and

$$\partial\Omega_{\mathrm{d}} = \{x \in \partial\Omega \,|\, x_1 = 0\} \cup \{x \in \partial\Omega \,|\, x_1 = 3\}$$

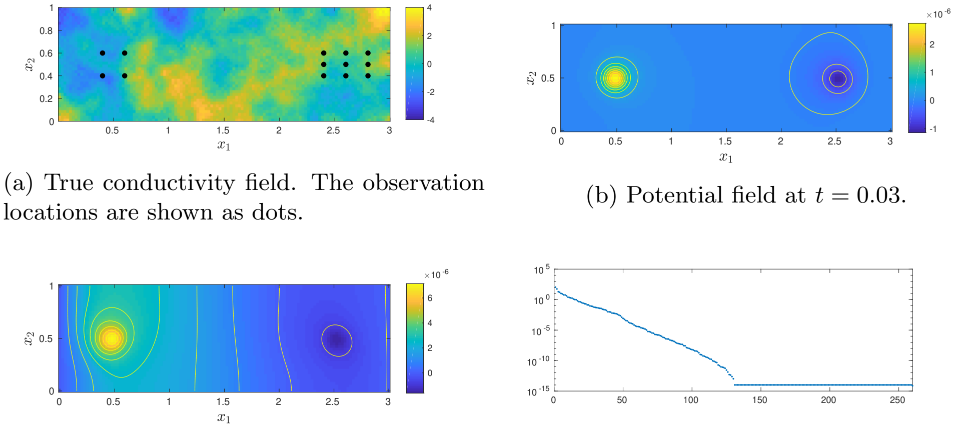
denote the left and right boundaries. For $t \geq 0$, we impose the mixed boundary condition

$$p(x,t) = 0 \ \forall x \in \partial\Omega_{\mathrm{d}} \quad \text{and} \quad (\kappa(x)\nabla p(x,t)) \cdot \vec{n}(x) = 0 \ \forall x \in \partial\Omega_{\mathrm{n}},$$

where $\vec{n}(x)$ is the outward normal vector on the boundary. We also impose a zero initial condition, i.e., $p(x, 0) = 0 \ \forall x \in \Omega$, and let the potential field be driven by a time-invariant forcing function

$$f(x,t) = c\Big(\exp\big(-\frac{1}{2r^2}\|x-a\|^2\big) - \exp\big(-\frac{1}{2r^2}\|x-b\|^2\big)\Big) \ \forall t \geq 0$$

with $r = 0.05$, which is the superposition of two Gaussian-shaped sink/source terms centered at $a = (0.5, 0.5)$ and $b = (2.5, 0.5)$, scaled by a constant $c = 6 \times 10^{-4}$.



(a) True conductivity field. The observation locations are shown as dots.

(b) Potential field at $t = 0.03$.

(c) Potential field at $t = 2.0$.

(d) Singular values of the linearized forward model at the MAP point.

FIG. 10. *Setup of the parabolic inversion example.*

The conductivity field $\kappa(x)$ is endowed with a log-normal prior. That is, letting $u(x) = \log \kappa(x)$, the prior for $u(x)$ takes the form $\mathrm{N}(m_{\mathrm{pr}}, \Gamma_{\mathrm{pr}})$. Here we prescribe zero prior mean, $m_{\mathrm{pr}} = 0$, and model the inverse of the prior covariance operator using the stochastic PDE approach (see [28, 43] and references therein):

$$-\triangle u(x) + \gamma u(x) = \mathcal{W}(x), \quad x \in \Omega, \tag{6.2}$$

where $\triangle$ is the Laplace operator and $\mathcal{W}(x)$ is the white noise process. We impose a no-flux boundary condition on the above stochastic PDE and set $\gamma = 5$.

Equations (6.1) and (6.2) are solved using the finite element method with bilinear basis functions. A mesh with 120 × 40 elements is used in this example. This leads to $n = 4800$ dimensional discretized parameters. The true conductivity field used for generating observed data is a realization from the prior distribution. The true conductivity field and the simulated potential field at different times are shown in Figures 10(a)–(c). The potential field is observed at 13 discrete locations (shown as dots in Figure 10(a)) at 20 discrete time points equally spaced between $t = 0.1$ and $t = 2$. We set the standard derivation of the observation noise to $\sigma = 3 \times 10^{-7}$, which corresponds to a signal-to-noise ratio of about 10. In the inverse problem, we use this $m = 260$ dimensional vector of data to estimate the conductivity field $\kappa(x)$.

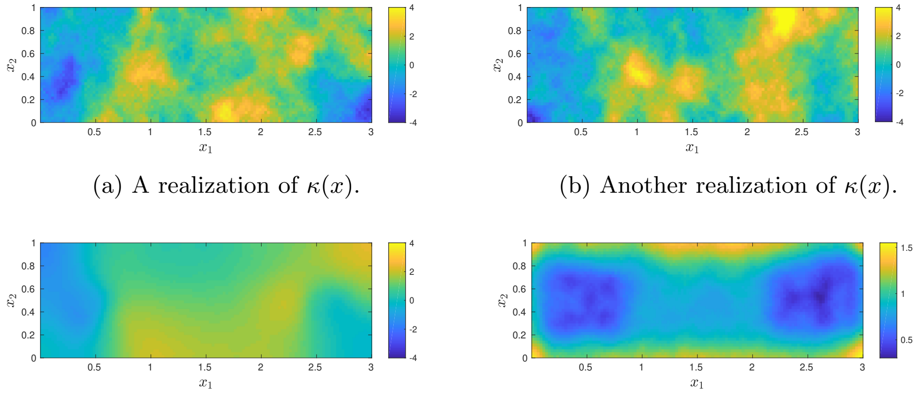
The forward model is linearized at the MAP point. As shown in Figure 10(d), we observe a sharp decay in the singular values of the linearized model, with these values dropping below machine precision after rank 130. We truncate the singular values at thresholds $\tau = 1$, 10^{-2}, and 10^{-4} to define three different RTO proposals. Then, using each RTO proposal, we generate 2500 samples to characterize the posterior using a Metropolis independence sampler (i.e., RTO-MH). The rank of the truncated SVD, statistics about the computation of each RTO sample, and ESS are reported in Table 4 for each truncation threshold. Note that all the truncated ranks are significantly smaller than the parameter dimension $n = 4800$.

Here, we observe that with a rather large truncation threshold ($\tau = 1$), we obtain a significantly lower ESS than with the other two truncation thresholds. This behavior agrees with the heuristics discussed in section 3.2: if one truncates the SVD more aggressively, the RTO proposal gets closer to the prior. Thus, it is expected

Table 4

Rank of the truncated SVD, average number of forward model evaluations, average number of MVPs with the linearized forward model and its adjoint, average number of optimization iterations per RTO sample, average CPU time per RTO sample, and ESS; all for varying SVD truncation thresholds. Chain length of 2500.

Truncation threshold	1	10^{-2}	10^{-4}
Rank	15	39	57
Number of evaluations of $G(v)$	18.8	12	12.4
Number of MVPs with $\nabla G(v)$	321.6	235.6	258.4
Optimization iterations per sample	17.8	11	11.4
CPU time (sec) per sample	354	257	283
Numerical ESS (out of 2500)	292	1140	1130



(a) A realization of $\kappa(x)$.

(b) Another realization of $\kappa(x)$.

(c) The posterior mean of $\kappa(x)$.

(d) The posterior standard deviation of $\kappa(x)$.

Fig. 11. *Sample realizations and summary statistics of the conductivity field $\kappa(x)$ distributed according to the posterior.*

that this RTO proposal will have lower statistical performance than an RTO proposal obtained with smaller τ (e.g., 10^{-2}). Once the truncation threshold is sufficiently small, however, we do not gain additional statistical performance by allowing more modes; compare the ESS at $\tau = 10^{-2}$ to that at $\tau = 10^{-4}$. This behavior is also in accordance with the truncation strategies and interpretation of the singular values discussed in section 3.2. Regarding the computational performance, we observe that more optimization iterations and longer CPU times are needed to obtain one RTO sample (on average) with the truncation threshold $\tau = 1$ than with the smaller truncation thresholds. We attribute this behavior to the fact that the truncated proposal does not constrain the parameter value in directions complementary to the range of Φ, and thus the optimization iterations may need to navigate through the tails of the posterior. For truncation thresholds 10^{-2} and 10^{-4}, the difference in the number of optimization iterations is insignificant. Overall, the truncation threshold of $\tau \approx 10^{-2}$ suggested in section 3.2 appears to be a reasonable choice in this example.

Two posterior samples and some summary statistics of the posterior, computed using RTO-MH with the truncation threshold $\tau = 10^{-2}$, are shown in Figure 11. We observe that the posterior samples and the posterior mean demonstrate a similar structure to the true conductivity field used to generate the synthetic data set. We also observe that the posterior standard deviation of the conductivity field is low in regions near the observation locations. In comparison, the posterior standard deviation is relatively high in regions near the boundary and between clusters of observation locations, where the observed data do not provide sufficient information to infer parameters.

We also attempted to compare RTO with pCN in this example. However, because of the rather informative data, pCN fails to produce an ergodic chain in a comparable amount of CPU time. An additional, but important, implementation note is that we generated RTO samples and evaluated the corresponding weighting functions (3.2) in parallel, and then quickly postprocessed the RTO samples using the Metropolis procedure to obtain posterior samples. Postprocessing is the only serial step of the calculation, and is very fast since all the costly calculations (sample generation, weight evaluation) are already completed. In this way, RTO can significantly reduce the wall clock time of Markov chain simulation compared to common MCMC methods that use state-dependent transition kernels, since posterior density evaluations and Markov chain simulation must be carried out sequentially in the latter case.

7. Discussion.

The main contribution of this work is a new scalable implementation of the RTO optimization-based sampling method. By using a polar decomposition rather than a QR factorization to build the RTO proposal, and deriving this polar decomposition from the SVD of a linearized forward model, we can reduce the computational cost of evaluating the RTO proposal (excepting perhaps the evaluation of the forward model itself) to linear complexity in the parameter dimension. This approach naturally splits the parameter space into two subspaces, and allows us to sample the RTO proposal and evaluate its density by solving smaller problems of size r, where r is an intrinsic dimension of the problem. This splitting also relates the RTO proposal to other parameter dimension reduction methods for Bayesian inverse problems. We formalize this RTO procedure in a function space setting, and show that the statistical performance of RTO is invariant to the discretized parameter dimension, under appropriate technical assumptions. Our results provide both practical algorithms and theoretical justification for applying RTO to high-dimensional inverse problems.

We then provide an empirical exploration of factors influencing the sampling efficiency of RTO, using various PDE-constrained Bayesian inverse problems. Our numerical results confirm that RTO has dimension-independent sampling efficiency, and also show that the observational noise magnitude affects the cost of solving each optimization problem but not the mixing of the RTO Metropolis independence sampler. Using a simple elliptic PDE example, we observe that RTO outperforms pCN for a wide range of problem settings. We also demonstrate the efficacy of RTO on a challenging two-dimensional parabolic PDE inverse problem, evaluating the impact of rank truncation on sampling efficiency and computational costs. These numerical results confirm our theoretical findings: RTO offers a viable way to tackle inverse problems with high-dimensional parameters and even very small observational noise.

There are many ways to extend the work described here. For example, one might use a mixture of several RTO proposals, defined by different linearizations, to better capture forward model nonlinearity in some extremely challenging inverse problems. Such mixtures might also help surmount the invertibility issues that arise when the assumptions of Theorem 4.3 are violated. For instance, one could employ a defensive mixture involving the prior distribution, along with localized proposals that are managed with trust-region strategies. The transport-map interpretation of RTO also suggests combining the RTO map with more elaborate local MCMC proposals on the Gaussian reference space, along the lines of [35]. In addition, since RTO's prior-to-proposal mapping has a well-defined continuous limit, one can naturally use RTO to generate coupled proposal samples at different discretization levels. These correlated samples can be used as control variates in the multilevel/multifidelity setting [21, 25, 36] to further accelerate the computation of posterior statistics.

Appendix A. Other optimization-based samplers. Similarly to RTO, other optimization-based sampling algorithms such as the random-map implementation of implicit sampling [32] and Metropolized RML [34] also yield deterministic couplings of two random variables. Here we briefly review the transport maps defined by the random-map implementation of implicit sampling and by Metropolized RML.

Implicit sampling requires that the target density have level sets that are "star-shaped," in that any ray starting from the mode passes through each level set exactly once. The target density is written as

$$\pi_{\mathrm{tar}}(v) \propto \exp\left(-\ell(v)\right), \tag{A.1}$$

where the negative log-target density ℓ has a minimum at the mode v_{MAP}. In order to draw proposal samples, we sample $\xi \in \mathbb{R}^n$ from a standard Gaussian and solve the following nonlinear system of equations to find a proposal $v_* \in \mathbb{R}^n$:

$$\begin{cases} \dfrac{L^{-1}(v_* - v_{\mathrm{MAP}})}{\|L^{-1}(v_* - v_{\mathrm{MAP}})\|} = \dfrac{\xi}{\|\xi\|}, \\ \ell(v_*) - \ell(v_{\mathrm{MAP}}) = \dfrac{1}{2}\|\xi\|^2. \end{cases} \tag{A.2}$$

The *direction* of the sample v_* (relative to the mode) is based on the direction of the sampled ξ. The *magnitude* of v_* is then found through a one-dimensional line search for the point where the negative log target ℓ satisfies

$$\ell(v_*) - \ell(v_{\mathrm{MAP}}) = \frac{1}{2}\|\xi\|^2.$$

In practice, L is chosen to be a square matrix such that $L^\top L := \left[\nabla^2 \ell(v_{\mathrm{MAP}})\right]^{-1}$, where $\nabla^2\ell(v_{\mathrm{MAP}})$ is the Hessian of ℓ evaluated at the MAP point.

Similarly to RTO, Metropolized RML requires that the target distribution have a Gaussian prior and additive Gaussian observational noise. Following the notation in section 2.1, we present a whitened version of Metropolized RML where the prior and observational noise covariances are transformed to the identity and the data are shifted to the origin. This way, the target density takes the form

$$\pi_{\mathrm{tar}}(v) \propto \exp\left(-\frac{1}{2}\|v\|^2 - \frac{1}{2}\|G(v)\|^2\right). \tag{A.3}$$

Defining a tuning parameter $\gamma \in (0,1)$, Metropolized RML adds the auxiliary variables $d \in \mathbb{R}^m$ and considers an *augmented* target distribution

$$\pi_{\mathrm{tar}}(v, d) \propto \exp\left(-\frac{1}{2}\|v\|^2 - \frac{1}{2\gamma}\|G(v) - d\|^2 - \frac{1}{2(1-\gamma)}\|d\|^2\right). \tag{A.4}$$

This defines a distribution on the joint space of parameters and data. Since the above joint distribution can also be written as

$$\begin{aligned} \pi_{\mathrm{tar}}(v, d) &\propto \exp\left(-\frac{1}{2}\|v\|^2 - \frac{1}{2}\|G(v)\|^2 - \frac{1}{2\gamma(1-\gamma)}\|d - (1-\gamma)G(v)\|^2\right) \\ &\propto \pi_{\mathrm{tar}}(v)\, \exp\left(-\frac{1}{2\gamma(1-\gamma)}\|d - (1-\gamma)G(v)\|^2\right), \end{aligned}$$

Table 5
Transport map interpretation of the three optimization-based samplers. In RTO, with default settings, the matrix Q comes from a thin QR factorization.

Algorithm	Target distribution	Transport map
RTO	$\pi_{\mathrm{tar}}(v) \propto \exp\left(-\frac{1}{2}\|H(v)\|^2\right)$	$Q^\top H(v) = \xi$, where $QR := \nabla H(v_{\mathrm{ref}})$
Implicit sampling	$\pi_{\mathrm{tar}}(v) \propto \exp\left(-\ell(v)\right)$	$\begin{cases} \dfrac{L^{-1}(v - v_{\mathrm{ref}})}{\|L^{-1}(v - v_{\mathrm{ref}})\|} = \dfrac{\xi}{\|\xi\|}, \\ \ell(v) - \ell(v_{\mathrm{ref}}) = \dfrac{1}{2}\|\xi\|^2, \end{cases}$ where $L^\top L := \left[\nabla^2 \ell(v_{\mathrm{ref}})\right]^{-1}$
RML	$\pi_{\mathrm{tar}}(v, d) \propto \exp\big(- \frac{1}{2}\|v\|^2$ $- \frac{1}{2\gamma}\|G(v) - d\|^2 - \frac{1}{2(1-\gamma)}\|d\|^2\big)$, where $\gamma \in (0, 1)$	$\begin{cases} v + \dfrac{1}{\rho}\nabla G(v)^\top \left(G(v) - d\right) = \xi_v, \\ \dfrac{1}{\rho}d - \left(\dfrac{1-\rho}{\rho}\right) G(v) = \xi_d, \end{cases}$ where $\rho \in (0, 1)$

it can be expressed as the product of the marginal distribution of v—which is the original target distribution—and the conditional distribution of d given v. Defining another tuning parameter $\rho \in (0, 1)$, Metropolized RML generates a pair of random variables $\xi_v \sim \mathrm{N}(0, \mathrm{I}_n)$ and $\xi_d \sim \mathrm{N}(0, \mathrm{I}_m)$ and solve the following randomly perturbed optimization problem,

$$(v_*, d_*) = \underset{(v,d)}{\arg\min} \left(\frac{1}{2}\|v - \xi_v\|^2 + \frac{1}{2\rho}\|G(v) - d\|^2 + \frac{1}{2(1-\rho)}\|d - \xi_d\|^2 \right),$$

to obtain a pair of proposal samples (v_*, d_*). Under the first order optimality condition, at the minima of the above objective function, the following system of nonlinear equations holds:

$$\text{(A.5)} \qquad \begin{cases} v_* + \dfrac{1}{\rho}\nabla G(v_*)^\top \left(G(v_*) - d_*\right) = \xi_v, \\ \dfrac{1}{\rho}d_* - \left(\dfrac{1-\rho}{\rho}\right) G(v_*) = \xi_d. \end{cases}$$

Thus, one can compute the joint density of (v_*, d_*) in the augmented parameter-and-data space using the mapping defined in (A.5). The samples are Metropolized in the augmented space to obtain correlated samples distributed according to the augmented target distribution $\pi_{\mathrm{tar}}(v, d)$. The components of v are then distributed according to the original target distribution. The parameters ρ and γ are tunable settings of the algorithm. In practice, they are set close to one and zero, respectively.

A summary of the mappings induced by RTO, implicit sampling, and RML is given in Table 5. Each algorithm describes a different map S, as in (2.6), to build the deterministic coupling. The actions of the inverse maps need to be computed using either nonlinear optimization algorithms or root finding methods (in one dimension).

Acknowledgments. The authors would like to thank the anonymous referees for their valuable comments on improving the manuscript. We also thank Benjamin Zhang for his thoughtful comments and suggestions.

REFERENCES

[1] S. Agapiou, O. Papaspiliopoulos, D. Sanz-Alonso, and A. M. Stuart, *Importance sampling: Intrinsic dimension and computational cost*, Statist. Sci., 32 (2017), pp. 405–431.

[2] J. M. Bardsley, A. Solonen, H. Haario, and M. Laine, *Randomize-then-optimize: A method for sampling from posterior distributions in nonlinear inverse problems*, SIAM J. Sci. Comput., 36 (2014), pp. A1895–A1910, https://doi.org/10.1137/140964023.

[3] A. Beskos, M. Girolami, S. Lan, P. E. Farrell, and A. M. Stuart, *Geometric MCMC for infinite-dimensional inverse problems*, J. Comput. Phys., 335 (2017), pp. 327–351.

[4] A. Beskos, G. O. Roberts, A. M. Stuart, and J. Voss, *MCMC methods for diffusion bridges*, Stoch. Dyn., 8 (2008), pp. 319–350.

[5] S. Brooks, A. Gelman, G. Jones, and X. L. Meng, eds., *Handbook of Markov Chain Monte Carlo*, Taylor & Francis, Boca Raton, FL, 2011.

[6] T. Bui-Thanh, O. Ghattas, J. Martin, and G. Stadler, *A computational framework for infinite-dimensional Bayesian inverse problems Part* I: *The linearized case, with application to global seismic inversion*, SIAM J. Sci. Comput., 35 (2013), pp. A2494–A2523.

[7] B. Calderhead, *A general construction for parallelizing Metropolis–Hastings algorithms*, Proc. Natl. Acad. Sci. USA, 111 (2014), pp. 17408–17413.

[8] C. Chen, T. Cui, Y. M. Marzouk, and Z. Wang, *Multilevel Optimisation-Based Importance Sampling Methods for Bayesian Inversion*, manuscript.

[9] V. Chen, M. M. Dunlop, O. Papaspiliopoulos, and A. M. Stuart, *Dimension–Robust MCMC in Bayesian Inverse Problems*, preprint, arXiv:1803.03344, 2019.

[10] A. Chorin, M. Morzfeld, and X. Tu, *Implicit particle filters for data assimilation*, Commun. Appl. Math. Comput. Sci., 5 (2010), pp. 221–240, https://doi.org/10.2140/camcos.2010.5.221.

[11] T. F. Coleman and Y. Li, *On the convergence of reflective Newton methods for large-scale nonlinear minimization subject to bounds*, Math. Program., 67 (1994), pp. 189–224.

[12] T. F. Coleman and Y. Li, *An interior trust region approach for nonlinear minimization subject to bounds*, SIAM J. Optim., 6 (1996), pp. 418–445.

[13] P. R. Conrad, A. D. Davis, Y. M. Marzouk, N. S. Pillai, and A. Smith, *Parallel local approximation MCMC for expensive models*, SIAM/ASA J. Uncertain. Quantif., 6 (2018), pp. 339–373.

[14] S. L. Cotter, G. O. Roberts, A. M. Stuart, and D. White, *MCMC methods for functions: Modifying old algorithms to make them faster*, Statist. Sci., 28 (2013), pp. 424–446, https://doi.org/10.1214/13-STS421.

[15] T. Cui, K. J. H. Law, and Y. M. Marzouk, *Dimension-independent likelihood-informed MCMC*, J. Comput. Phys., 304 (2016), pp. 109–137, https://doi.org/10.1016/j.jcp.2015.10.008.

[16] T. Cui, J. Martin, Y. M. Marzouk, A. Solonen, and A. Spantini, *Likelihood-informed dimension reduction for nonlinear inverse problems*, Inverse Problems, 29 (2014), 114015, https://doi.org/10.1088/0266-5611/30/11/114015.

[17] T. Cui, Y. M. Marzouk, and K. Willcox, *Scalable posterior approximations for large-scale Bayesian inverse problems via likelihood-informed parameter and state reduction*, J. Comput. Phys., 315 (2016), pp. 363–387.

[18] T. J. Dodwell, C. Ketelsen, R. Scheichl, and A. L. Teckentrup, *A hierarchical multilevel Markov chain Monte Carlo algorithm with applications to uncertainty quantification in subsurface flow*, SIAM/ASA J. Uncertain. Quantif., 3 (2015), pp. 1075–1108.

[19] H. P. Flath, L. C. Wilcox, V. Akcelik, J. Hill, B. van Bloemen Waanders, and O. Ghattas, *Fast algorithms for Bayesian uncertainty quantification in large-scale linear inverse problems based on low-rank partial Hessian approximations*, SIAM J. Sci. Comput., 33 (2011), pp. 407–432.

[20] V. H. Hoang, C. Schwab, and A. M. Stuart, *Complexity analysis of accelerated MCMC methods for Bayesian inversion*, Inverse Problems, 29 (2013), 085010.

[21] M. B. Giles, *Multilevel Monte Carlo path simulation*, Oper. Res., 56 (2008), pp. 607–617, https://doi.org/10.1287/opre.1070.0496.

[22] 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, https://doi.org/10.1111/j.1467-9868.2010.00765.x.

[23] G. H. Golub and C. F. Van Loan, *Matrix Computations*, Vol. 3, John Hopkins University Press, Baltimore, MD, 2012.

[24] N. Halko, P. G. Martinsson, and J. A. Tropp, *Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions*, SIAM Rev., 53 (2011), pp. 217–288.

[25] S. Heinrich, *Multilevel Monte Carlo methods*, in Large-Scale Scientific Computing, Springer, Berlin, 2001, pp. 58–67, https://doi.org/10.1007/3-540-45346-6.

[26] 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, Oxford, 2003, pp. 181–197.

[27] J. Kaipio and E. Somersalo, *Statistical and Computational Inverse Problems*, Appl. Math. Sci. 160, Springer, New York, 2006, https://doi.org/10.1007/b138659.

[28] F. Lindgren, H. Rue, and J. Lindstroм̈, *An explicit link between Gaussian fields and Gaussian Markov random fields: The stochastic partial differential equation approach*, J. R. Stat. Soc. Ser. B Stat. Methodol., 73 (2011), pp. 423–498.

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

[30] J. C. Mattingly, N. S. Pillai, and A. M. Stuart, *Diffusion limits of the random walk Metropolis algorithm in high dimensions*, Ann. Appl. Probab., 22 (2012), pp. 881–930.

[31] K. L. Mengersen and R. L. Tweedie, *Rates of convergence of the Hastings and Metropolis algorithms*, Ann. Statist., 24 (1996), pp. 101–121.

[32] M. Morzfeld, X. Tu, E. Atkins, and A. J. Chorin, *A random map implementation of implicit filters*, J. Comput. Phys., 231 (2012), pp. 2049–2066, https://doi.org/10.1016/j.jcp.2011.11.022.

[33] J. Nocedal and S. Wright, *Numerical Optimization*, Springer, New York, 2006.

[34] D. S. Oliver, *Metropolized randomized maximum likelihood for improved sampling from multimodal distributions*, SIAM/ASA J. Uncertain. Quantif., 5 (2017), pp. 259–277, https://doi.org/10.1137/15M1033320.

[35] M. D. Parno and Y. M. Marzouk, *Transport map accelerated Markov chain Monte Carlo*, SIAM/ASA J. Uncertain. Quantif., 6 (2018), pp. 645–682, https://doi.org/10.1137/17M1134640.

[36] B. Peherstorfer, K. Willcox, and M. Gunzburger, *Optimal model management for multifidelity Monte Carlo estimation*, SIAM J. Sci. Comput., 38 (2016), pp. A3163–A3194, https://doi.org/10.1137/15M1046472.

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

[38] G. D. Prato, *An Introduction to Infinite-Dimensional Analysis*, Springer, Berlin, 2006, https://doi.org/10.1007/3-540-29021-4.

[39] G. O. Roberts, A. Gelman, and W. R. Gilks, *Weak convergence and optimal scaling of random walk Metropolis algorithms*, Ann. Appl. Probab., 7 (1997), pp. 110–120.

[40] G. O. Roberts and R. L. Tweedie, *Exponential convergence of Langevin distributions and their discrete approximations*, Bernoulli, 2 (1996), pp. 341–363, https://doi.org/10.2307/3318418.

[41] D. Rudolf and B. Sprungk, *On a generalization of the preconditioned Crank–Nicolson Metropolis algorithm*, Found. Comput. Math., 18 (2018), pp. 309–343.

[42] A. Spantini, A. Solonen, T. Cui, J. Martin, L. Tenorio, and Y. Marzouk, *Optimal low-rank approximations of Bayesian linear inverse problems*, SIAM J. Sci. Comput., 37 (2015), pp. A2451–A2487.

[43] A. M. Stuart, *Inverse problems: A Bayesian perspective*, Acta Numer., 19 (2010), pp. 451–559, https://doi.org/10.1017/S0962492910000061.

[44] L. Tierney, *A note on Metropolis-Hastings kernels for general state spaces*, Ann. Appl. Probab., 8 (1998), pp. 1–9.

[45] K. Wang, T. Bui-Thanh, and O. Ghattas, *A randomized maximum a posteriori method for posterior sampling of high dimensional nonlinear Bayesian inverse problems*, SIAM J. Sci. Comput., 40 (2018), pp. A142–A171.

[46] Z. Wang, J. M. Bardsley, A. Solonen, T. Cui, and Y. M. Marzouk, *Bayesian inverse problems with L1 priors: A randomize-then-optimize approach*, SIAM J. Sci. Comput., 39 (2017), pp. S140–S166, https://doi.org/10.1137/16M1080938.

[47] O. Zahm, T. Cui, K. Law, A. Spantini, Y. Marzouk, *Certified Dimension Reduction in Nonlinear Bayesian Inverse Problems*, preprint, arXiv:1807.03712, 2018.