An Optimization Based Algorithm for Bayesian Inference

by

Zheng Wang

B.A.Sc., Engineering Science, University of Toronto (2013)

Submitted to the Department of Aeronautics and Astronautics
in partial fulfillment of the requirements for the degree of

Master of Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2015

Author ..... Signature redacted .....
Department of Aeronautics and Astronautics
May 7, 2015

Certified by..... Signature redacted .....
Youssef M. Marzouk
Associate Professor of Aeronautics and Astronautics
Thesis Supervisor

Accepted by ..... Signature redacted .....
Paulo C. Lozano
Associate Professor of Aeronautics and Astronautics
Chair, Graduate Program Committee

An Optimization Based Algorithm for Bayesian Inference

by

Zheng Wang

Submitted to the Department of Aeronautics and Astronautics
on May 19, 2015, in partial fulfillment of the
requirements for the degree of
Master of Science

Abstract

In the Bayesian statistical paradigm, uncertainty in the parameters of a physical system is characterized by a probability distribution. Information from observations is incorporated by updating this distribution from prior to posterior. Quantities of interest, such as credible regions, event probabilities, and other expectations can then be obtained from the posterior distribution. One major task in Bayesian inference is then to characterize the posterior distribution, for example, through sampling.

Markov chain Monte Carlo (MCMC) algorithms are often used to sample from posterior distributions using only unnormalized evaluations of the posterior density. However, high dimensional Bayesian inference problems are challenging for MCMC-type sampling algorithms, because accurate proposal distributions are needed in order for the sampling to be efficient. One method to obtain efficient proposal samples is an optimization-based algorithm titled 'Randomize-then-Optimize' (RTO).

We build upon RTO by developing a new geometric interpretation that describes the samples as projections of Gaussian-distributed points, in the joint data and parameter space, onto a nonlinear manifold defined by the forward model. This interpretation reveals generalizations of RTO that can be used. We use this interpretation to draw connections between RTO and two other sampling techniques, transport map based MCMC and implicit sampling. In addition, we motivate and propose an adaptive version of RTO designed to be more robust and efficient. Finally, we introduce a variable transformation to apply RTO to problems with non-Gaussian priors, such as Bayesian inverse problems with L1-type priors. We demonstrate several orders of magnitude in computational savings from this strategy on a high-dimensional inverse problem.

Thesis Supervisor: Youssef M. Marzouk
Title: Associate Professor of Aeronautics and Astronautics

Acknowledgments

The author would like to thank professor Youssef Marzouk and Helen Zhang for their indispensable advice and support. He would also like to thank his colleagues in the Aerospace Computational Design Laboratory for their helpful advice, intriguing conversations, and overall camaraderie.

This project would not be possible without funding from Eni. The author has also received financial support from NSERC. Thank you.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

1.1.1 Inverse problems

In the field of computational modeling, much work is done on the refinement of model parameters from observed data. This process is an example of solving an *inverse problem.* Inverse problems appear in many applications, including tomography, chemical kinetics, robot localization, reservoir geomechanics, and weather prediction.

Typically, scientists and engineers have access to a mathematical or computer model that simulates the response of the system, called the forward model. In particular, the forward model has parameter values as the input and calculates the desired measurements as the output. This forward model is used in conjunction with observed, "real" measurements to infer the input parameters. Thus, original data can be used to provide insight into system, to make informed predictions, or to guide decision-making.

To illustrate this process, we give here the example of reservoir geomechanics. The inverse problem, in this case, is to characterize the underground permeability and porosity of a reservoir using pressure measurements obtained at wells, represented in Fig. 1-1. The parameters of interest are the two rock properties, permeability and porosity, and the observations are the pressure measurements. A flow-geomechanical

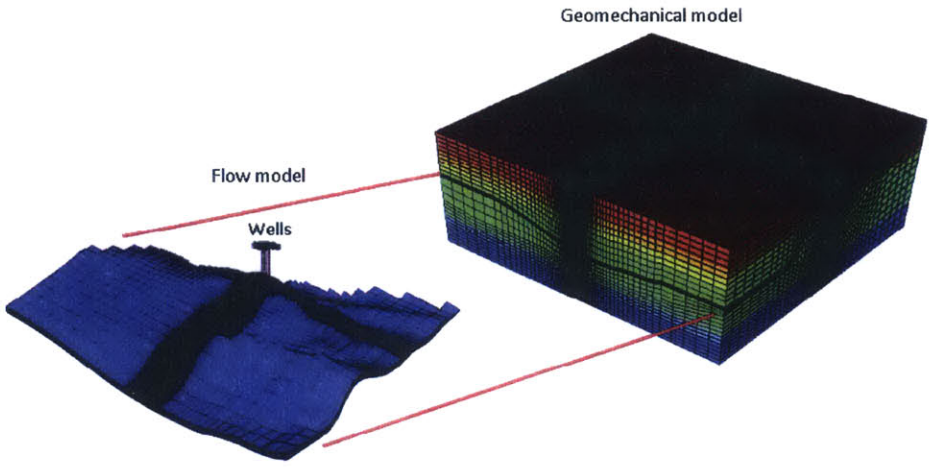

Figure 1-1: Reservoir geomechanics. Example application for inverse problems. Illustration is taken from [13].

simulation constitutes the forward model. Solving the inverse problem involves using well pressure measurements in a physical reservoir and finding, though successive computer simulations, likely values for permeability and porosity. These parameter values can then be used to predict future production from the wells and help make informed management decisions.

Inverse problems are also an integral part of filtering (also known as data assimilation) and experimental design. In filtering, inverse problems are solved sequentially to obtain state estimations of a dynamic system in time. A current state is then propagated using model dynamics to create forecasts, which are subsequently incorporated into the state estimation at the next time step. In experimental design, observations from previously-conducted experiments are used to characterize system parameters in order to determine the best experiments to perform in the future.

Unfortunately, many inverse problems are ill-posed, meaning that the solution to the inverse problem may not exist, may not be unique, or may be highly sensitive to small changes in the data. This can occur when widely different parameter values yield very close outputs in the model. In addition, for some applications, we are not only interested in the particular values of the system parameters, but also want to characterize the uncertainty surrounding those values. The is known as uncertainty quantification. The uncertainty can be described by a variance, a credible region, or

an entire distribution. Characterizing the uncertainty of the parameters is important for incorporating risk in analysis and decision making.

1.1.2 Bayesian inference

Bayesian inference is a framework for solving inverse problems that addresses both ill-posedness and uncertainty quantification. In Bayesian inference, we describe the uncertainty of a parameter using a distribution. An initial prior distribution describes the belief state of the parameter before any observations are taken into account. Then, the prior is updated to the posterior distribution using Bayes' rule.

$$\underbrace{p(\theta|y)}_{\text{posterior}} = \frac{\overbrace{p(y|\theta)}^{\text{likelihood}}\overbrace{p(\theta)}^{\text{prior}}}{\underbrace{\int_{\hat{\theta}} p(y|\hat{\theta})p(\hat{\theta})}_{\text{evidence}}}$$

where θ is the parameter and y is the observation. The evidence is a normalizing constant and is often costly to compute.

Solving a Bayesian inference problem typically reduces to characterizing the posterior distribution. Given the posterior, we can calculate the posterior mean, variance, higher moments, and event probabilities by taking expectations. We can also determine the posterior mode[1] and credible regions. There are many ways to characterize the posterior. [14] and [24] are two useful references for inverse problems and Bayesian inference. In this thesis, we focus on using samples to characterize the Bayesian posterior. With posterior samples, we can use Monte-Carlo integration to calculate any expectation of interest.

1.2 Markov-chain Monte Carlo

Markov-chain Monte Carlo (MCMC) algorithms are widely used to generate samples from a distribution for which the normalizing constant is difficult to compute, as is

[1]The posterior mode is sometimes referred to as the maximum a posteriori (MAP) point.

the case for a posterior distribution. MCMC generates dependent samples from a target distribution by simulating a Markov chain. These algorithms are commonplace and are described within several textbooks [9] [20] [4].

Random-walk Metropolis

A canonical MCMC algorithm is the random-walk Metropolis, described in Algorithm 1. In order to generate a chain of samples distributed according to the posterior, we propose a point from a Gaussian centered at the current point in the chain. A simple calculation is used to determine whether to accept and move to the proposed point, or to reject and remain at the current point. Each proposal requires a calculation of the unnormalized posterior density at the new point, which requires a forward model evaluation. When the Markov chain is continued ad infinitum, the distribution of its samples will approach the posterior[2].

Algorithm 1 Random-walk Metropolis

1: start with $\theta^{(0)}$

2: **for** $i = 1, \cdots, n_{\text{samp}}$ **do**

3: propose $\hat{\theta}^{(i)} \sim N(\theta^{(i-1)}, \sigma^2 I)$

4: $\theta^{(i)} = \begin{cases} \hat{\theta}^{(i)} & \text{with probability } \min\left(\frac{p(\hat{\theta}^{(i)}|y)}{p(\theta^{(i-1)}|y)}, 1\right) \\ \theta^{(i-1)} & \text{otherwise} \end{cases}$

Here, n_{samp} is the number of samples in the chain, $\hat{\theta}$ are the proposal points, and θ are posterior samples.

Independence Metropolis-Hastings

One useful MCMC algorithm to know for the rest of this thesis is independence Metropolis-Hastings. In contrast to random-walk Metropolis, the proposal sample does not depend on the current point in the chain; rather, it is drawn independently from a fixed proposal distribution. Algorithm 2 outlines the steps. The parts colored

[2]The distribution of the samples from such a Markov chain will approach the posterior under reasonably weak conditions.

in red highlight the difference between independence Metropolis-Hastings and random-walk Metropolis.

Algorithm 2 independence Metropolis-Hastings

1: start with $\theta^{(0)}$

2: **for** $i = 1, \cdots, n_{\text{samp}}$ **do**

3: propose $\hat{\theta}^{(i)} \sim q(\cdot)$

4: $$\theta^{(i)} = \begin{cases} \hat{\theta}^{(i)} & \text{with probability } \min\left(\frac{p(\hat{\theta}^{(i)}|y)q(\theta^{(i-1)})}{p(\theta^{(i-1)}|y)q(\hat{\theta}^{(i)})}, 1\right) \\ \theta^{(i-1)} & \text{otherwise} \end{cases}$$

Here, n_{samp} is the number of samples in the chain, $\hat{\theta}$ are proposal samples, θ are posterior samples, and $q(\cdot)$ is a fixed proposal distribution.

1.2.1 Optimization-based samplers

The efficiency of MCMC algorithms is tied to the correlation in the Markov chain. A chain of samples that are highly correlated will result in a higher variance (and error, on average) when used to calculate any expectation. One measure of the information contained in a chain of samples is the effective sample size (ESS). The ESS is the number of independent samples from the posterior that would give the same variance in calculating an expectation. Inefficient MCMC algorithms produce highly correlated chains and, hence, longer chains are required to obtain a desired ESS. The efficiency of MCMC samplers depends heavily on effective proposals.

Many MCMC techniques use adapted Gaussian proposals. The random-walk Metropolis algorithm, see Sec. 1.2, is one example. Two other examples are delayed-rejection adaptive Metropolis (DRAM) [11] and Metropolis-adjusted Langevin algorithm (MALA) [21]. DRAM proposes from a Gaussian adapted to the sample-estimated posterior covariance. When a sample is rejected, DRAM proposes from a second Gaussian with a reduced covariance. MALA uses derivative information to shift its proposal from being centered at the current sample towards the high-posterior region. All of these samplers propose from a Gaussian to update the Markov chain, and are reasonably efficient in low dimensions. However, for especially non-Gaussian posteriors

and high dimensional parameters, these and other simple MCMC samplers become inefficient.

One technique to improve sampling efficiency is to leverage optimization to draw from effective proposal distributions, proposal distributions that are close to the posterior. Optimization-based sampling techniques use optimization to propose from *non-Gaussian* distributions that depend on the posterior distribution or the inverse problem itself. Since their proposals are non-Gaussian, these algorithms have a higher possibility to be efficient for high dimensional, non-Gaussian posteriors. Implicit sampling [5] and Randomize-then-optimize (RTO) [2] are two such techniques, the latter of which will be the focus of this thesis.

1.3 Thesis Contribution

In this thesis, we focus on solving inverse problems using Bayesian inference by sampling from the posterior using RTO. The major contributions of this thesis are:

1. A new geometric interpretation of the RTO algorithm.
2. An adaptive version of RTO that is more robust and efficient.
3. A prior transformation technique to extend RTO to Bayesian inference problems with non-Gaussian priors.

The geometric interpretation provides intuition about the conditions under which RTO performs well. Adaptive RTO uses this interpretation to improve upon the original algorithm. Finally, a prior transformation technique allows us to use RTO on a broader range of inverse problems.

Each of these contributions is described in further detail in its own chapter. In Chapters 2 and 3, we use the assumption of a Gaussian prior and Gaussian observational noise. In Chapter 4 we relax this assumption to consider non-Gaussian priors, in particular, L1-type priors.

Chapter 2

Geometric Interpretation of RTO

This chapter describes and reinterprets the randomize-then-optimize (RTO) algorithm [2] for sampling from a Bayesian posterior. RTO is motivated through a geometric perspective. The posterior density is interpreted as a manifold intersecting a high-dimensional multivariate Gaussian. RTO is then reintroduced as a projection of samples from the high-dimensional multivariate Gaussian onto the manifold described by the forward model.

Following naturally from this geometric interpretation, we describe generalizations to RTO and discover a parameterized family of RTO-like proposal distributions to which the prior distribution belongs. This family of RTO-like proposals can be thought of as 'tuning knobs' of the sampling algorithm; the original RTO formulation provides heuristic values for the knobs, and practitioners may adjust the knobs to obtain desirable performance from RTO. The next chapter will demonstrate a few cases where the default, heuristic values of the knobs are inefficient or invalid. The family of RTO-like proposal distributions will then be used in an more robust adaptive version of RTO.

A second interpretation of RTO (and the RTO-like proposals) recasts it as an implicit approximate transport map. Using this formulation, we connect RTO to transport-map accelerated MCMC [19] and implicit sampling [17]. In particular, RTO and implicit sampling can both be thought of as using implicitly defined approximate

transport maps[1]. Transport maps are defined from the forward model for RTO, defined from the the target distribution for implicit sampling, and evaluated by solving optimization problems. This contrasts the transport-map accelerated MCMC approach, where an *explicit* approximate transport map is represented as a multivariate polynomial expansion and evaluated directly.

This chapter is organized as follows: Section 2.1 describes the geometric interpretation of the posterior in Bayesian inference; Section 2.2 reintroduces RTO as a projection, and provides intuitive reasoning as to why the algorithm generates proposal samples that are distributed close the posterior; Section 2.3 generalizes RTO to uncover a parameterized family of RTO-like proposals and describes the conditions under which the proposals are valid; Section 2.4 connects RTO to transport-map accelerated MCMC, and Section 2.5 connects RTO to implicit sampling.

2.1 Geometric Interpretation of the Posterior

Before describing the details of RTO, we begin by exploring the structure of the posterior distribution that RTO exploits. RTO requires that the posterior be in least-squares form. Let n be the number of parameters and m be the number of observations. The least-squares form is defined as

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

where $\theta \in \mathbb{R}^n$ is the parameter vector, $p(\cdot|y) : \mathbb{R}^n \to \mathbb{R}_+$ is the posterior density function, $F(\cdot) : \mathbb{R}^n \to \mathbb{R}^{m+n}$ is a parameter-to-response function which, in the context of Bayesian inference, contains the forward model, and $Y \in R^{m+n}$ is the response vector, which in the context of Bayesian inference contains the observation.

Since we can always transform an inverse problem with a non-standard Gaussian prior and a non-standard Gaussian observational noise to a problem with a standard Gaussian for both, we consider an inverse problem with standard Gaussian prior and

[1]RTO and implicit sampling both use approximate transport maps to provide samples from the exact posterior.

noise without loss of generality.

$$y = f(\theta) + \epsilon \qquad \theta \sim N(0, I) \qquad \epsilon \sim N(0, I)$$

where $y \in \mathbb{R}^m$ is the observation vector, $f(\cdot) : \mathbb{R}^n \to \mathbb{R}^m$ is the forward model, and $\epsilon \in \mathbb{R}^m$ is the observational noise. The posterior density function that arises from this inverse problem is

$$p(\theta|y) \propto p(y|\theta)p(\theta) \tag{2.1}$$

$$\propto \exp\left(-\frac{1}{2}\|f(\theta) - y\|^2\right) \exp\left(\frac{1}{2}\|\theta\|^2\right) \tag{2.2}$$

$$= \exp\left(-\frac{1}{2}\left\| \underbrace{\begin{bmatrix} \theta \\ f(\theta) \end{bmatrix}}_{F(\theta)} - \underbrace{\begin{bmatrix} 0 \\ y \end{bmatrix}}_{Y} \right\|^2\right). \tag{2.3}$$

As shown above, when the prior and observational noise are Gaussian, the posterior can be written in least-squares form. Here we incorporate prior information on each parameter as separate responses to obtain $m + n$ responses as the output of F.

An interesting thing to note in Eq. 2.1 is that the form of the posterior distribution resembles that of a multivariate Gaussian in $\mathbb{R}^{m+n}$. In particular, if we replace the parameter-to-response function $F(\theta)$ with a vector of independent random variables, we obtain a standard Gaussian centered at Y. We can interpret Eq. 2.1 as being constrained on a $\mathbb{R}^n$ dimension manifold in $\mathbb{R}^{m+n}$ that is parameterized by $F(\theta)$. In other words, the unnormalized posterior density evaluated at a particular value of θ is the same as the density of a standard Gaussian in $\mathbb{R}^{m+n}$ evaluated at $F(\theta) \equiv [\theta, f(\theta)]^T$.

When $m = n = 1$, we can visualize the higher-dimensional multivariate Gaussian. $F(\theta)$ becomes a line, a 1-D manifold embedded in 2D, as in Fig. 2-1. The local posterior density at a particular value of θ, shown on the horizontal axis, is determined by the height of the 2-D Gaussian, centered at $[0, y]^T$, evaluated at $[\theta f(\theta)]^T$ on the line.

One strategy to obtain samples from the posterior is to first sample from a proposal

distribution that is close to the posterior. Then, these samples are corrected using independence Metropolis-Hastings or importance sampling. RTO, map-accelerated MCMC, and implicit sampling all employ this strategy. The main challenge is then to obtain samples close enough to the posterior such that correction in independence Metropolis-Hastings is efficient. In the next section, we describe the procedure RTO uses to obtain proposal samples that are close to the posterior.

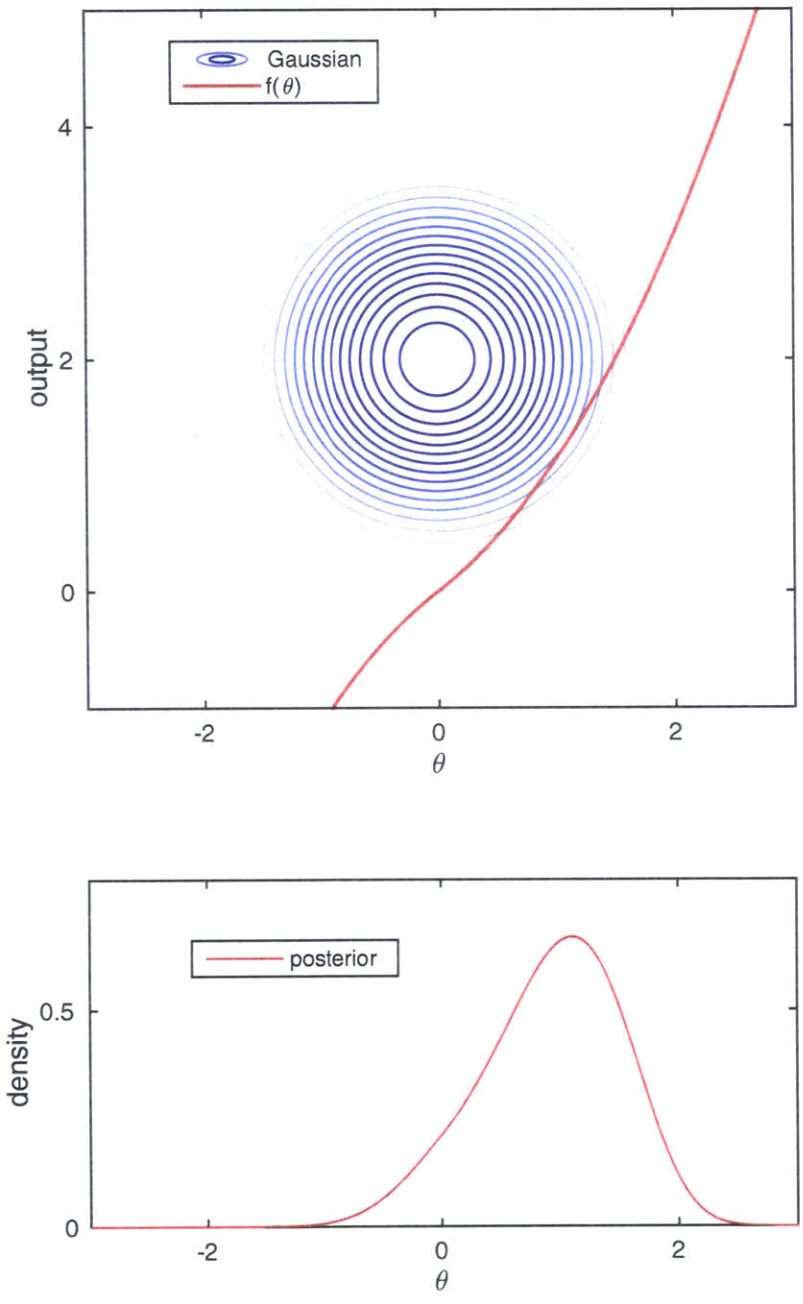

Figure 2-1: Posterior in least-squares form. (Top) 1-D manifold defined by the forward model intersecting a multivariate Gaussian. (Bottom) Posterior density. The posterior density at θ is equal to the density of the high dimensional Gaussian evaluated at $[\theta, f(\theta)]^T$.

2.2 RTO as a Projection on to a Manifold

RTO is motivated from the geometric posterior description described in the previous section. It generates proposal samples by sampling from the high-dimensional multivariate Gaussian, and then transporting, or projecting, these $m+n$-dimensional samples on to the n-dimensional manifold described by the forward model. The projection directions are orthogonal to the tangent space at the posterior mode. In other words, the samples are free to move in the directions orthogonal to the space spanned by the columns of the Jacobian of $F(\theta)$ at the posterior mode.

Fig. 2-2 is a 2D visualization where $n = m = 1$. The red dots describe the initial $m+n$-dimensional samples, and the purple dots are the projected samples that lie on the n-dimensional manifold. As seen in Fig. 2-2, the samples generated in such a manner are distributed according to a proposal distribution that is very close to the posterior. We expect this to be the case when the forward model is smooth and close to linear in the high posterior region.

It is very easy to verify that when the forward model is linear, as is the case for a straight line in 2D, the RTO proposal is exactly the posterior and they are Gaussian. When the forward model is nonlinear, but smooth, the posterior may be very non-Gaussian, meaning that a Gaussian proposal may be a poor choice for independence Metropolis-Hastings. However, RTO's proposal distribution may be very efficient.

A brief description of the RTO algorithm follows. RTO first solves an optimization problem to find the posterior mode. This is done through

$$\bar{\theta} = \arg\min_{\theta} \frac{1}{2}\|F(\theta) - Y\|^2 \tag{2.4}$$

where $\bar{\theta} \in \mathbb{R}^n$ is the posterior mode. RTO determines the tangent space at the posterior mode by storing an orthonormal basis $\bar{Q} \in \mathbb{R}^{(m+n)\times n}$, which is found through a QR factorization of $J(\bar{\theta})$. It then samples the $m+n$-dimensional standard Gaussian for a point $\Xi^{(i)}$ and projects this point onto the manifold by solving the optimization

problem

$$\theta^{(i)} = \arg\min_{\theta} \frac{1}{2} \left\| \bar{Q}^T \left(F(\theta) - (Y + \Xi^{(i)}) \right) \right\|^2. \tag{2.5}$$

Then, under certain conditions described in the next section, the proposal distribution can be evaluated as

$$q(\theta) \propto \left| \bar{Q}^T J(\theta) \right| \exp\left(-\frac{1}{2} \left\| \bar{Q}^T (F(\theta) - Y) \right\|^2 \right) \tag{2.6}$$

where $|\bar{Q}^T J(\theta)|$ indicates the absolute value of the determinant of the matrix $\bar{Q}^T J(\theta)$. The RTO algorithm is summarized in Alg. 3.

Algorithm 3 RTO

1: Find posterior mode $\bar{\theta}$ using optimization
2: Find $\bar{Q}$, the orthonormal basis that spans the columns of $J(\bar{\theta})$
3: **for** $i = 1, \cdots, n_{\text{samps}}$ **do** in parallel
4: Sample $\Xi^{(i)}$ from $m + n$-dimensional Gaussian
5: Find proposal samples $\hat{\theta}^{(i)}$ by minimizing Eq. 2.5.
6: Find $q(\hat{\theta}^{(i)})$ from Eq. 2.6
7: **for** $i = 1, \cdots, n_{\text{samps}}$ **do** in series
8: Use independence Metropolis-Hastings to obtain $\theta^{(i)}$

Note that $\bar{Q}$ can be found from a thin-QR decomposition of $J(\bar{\theta})$. In theory, the algorithm can be implemented directly using $J(\bar{\theta})$ in Eq. 2.5.

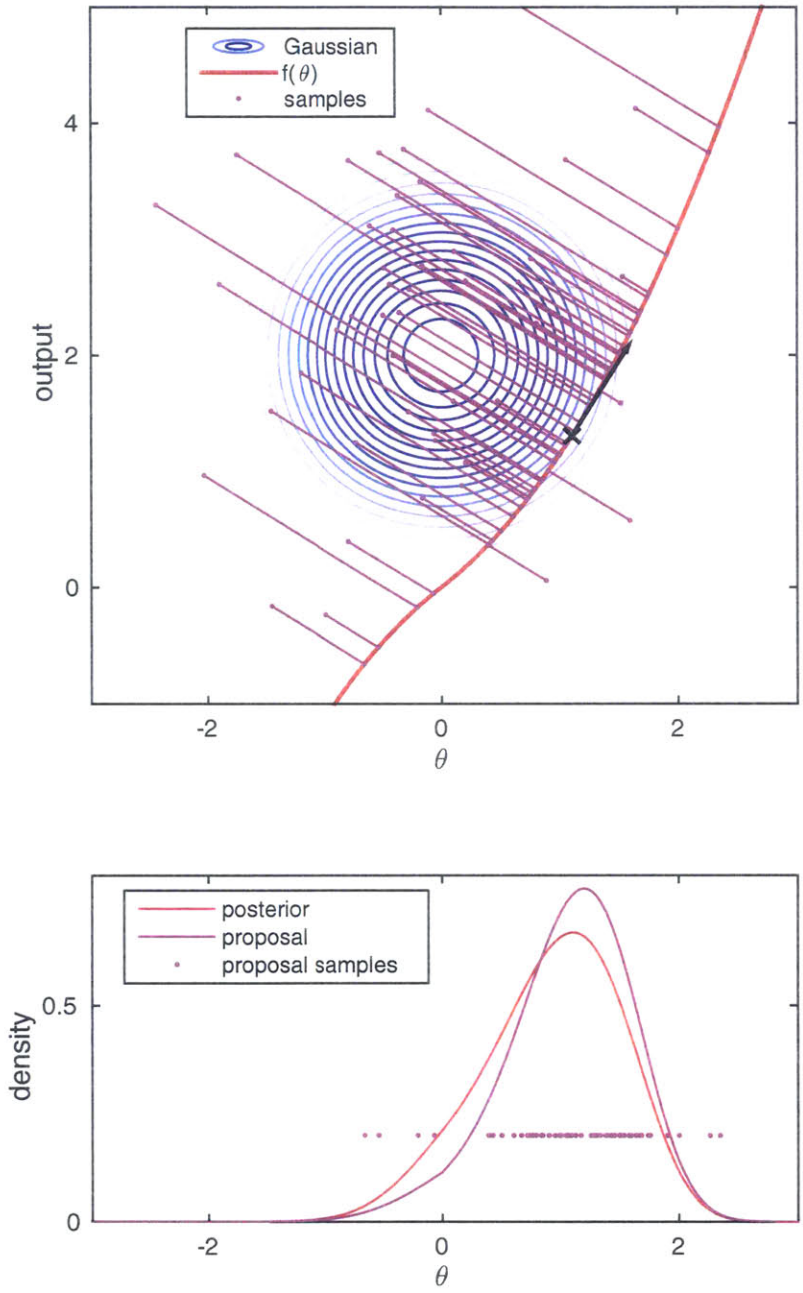

Figure 2-2: RTO's procedure to generate proposal samples. (Top) Gaussian samples shown, in purple, are projected on to the manifold. The black X is $F(\bar{\theta})$ and the arrow is $\bar{Q}$. (Bottom) The resulting samples are distributed according to the proposal density.

2.3 Validity and Generalizations of RTO

In this section, we explore when RTO is valid and generalize the algorithm by considering changes to $\bar{Q}$ and Y, interpreting the changes geometrically. This will lead to a more flexible description of an RTO-like proposal distribution, and reveal a family of such distributions for which we can evaluate the proposal density and from which we are able to sample. Being able to evaluate the proposal density is important because it allows us to use independence Metropolis-Hastings to obtain true posterior samples from the proposal samples.

The theorem that describes the circumstances under which the proposal distribution from RTO is valid is found in [2]. We paraphrase the main points of the theorem below. The conditions are:

(i) $p(\theta|y) \propto \exp\left(-\frac{1}{2}\|F(\theta) - Y\|^2\right)$, where $\theta \in \mathbb{R}^n$, $Y \in \mathbb{R}^m$.

(ii) $F : \mathbb{R}^n \to \mathbb{R}^m$ is a continuously differentiable function with Jacobian J.

(iii) $J(\theta) \in \mathbb{R}^{m \times n}$ has rank n for every θ in the domain of F.

(iv) the posterior mode $\bar{\theta} = \arg\min_x p(\theta|y)$ is unique.

(v) $J(\theta) = [\bar{Q}, \tilde{Q}][R, 0]^T$ is the QR-factorization of $J(\theta)$.

(vi) the matrix $\bar{Q}^T J(\theta)$ is invertible for all θ in the domain of F.

(vii) there is a θ such that $\bar{Q}^T(F(\theta) - (Y + \Xi)) = 0$ for any Ξ.

Theorem 2.3.1. *Assuming (i) to (vii) directly above are true, then lines 1-6 of Alg. 3 will sample from the distribution with the probability density described in Eq. 2.6.*

The proof of Theorem 2.3.1 is found in [2]. In that paper, the authors considered any target distribution that can be written in least-squares form. In Bayesian inference, we often have a posterior distribution with a Gaussian prior and observational noise. When we have a continuously differentiable function $f(\cdot)$, $F(\cdot)$ is also continuously differentiable. Thus, assumptions (i) and (ii) are automatically satisfied. Next, we show that (iii) is also true.

Theorem 2.3.2. *When we have a posterior distribution with a Gaussian prior and observational noise, $J(\theta) \in \mathbb{R}^{m \times n}$ has rank n for every θ in the domain of F.*

Proof.

$$J(\theta) \equiv \frac{\partial}{\partial \theta} F(\theta) = \frac{\partial}{\partial \theta} \begin{bmatrix} \theta \\ f(\theta) \end{bmatrix} = \begin{bmatrix} I \\ \nabla_f(\theta) \end{bmatrix}$$

Regardless of $\nabla_f(\theta)$, the columns of $J(\theta)$ will be linearly independent due to the identity matrix at the top. Hence, it will have rank n. □

Conditions (iv) and (v) are only used to ensure a unique $\bar{Q}$. The final two conditions (vi) and (vii), are not automatically satisfied and are left as assumptions. In the next chapter, we will show the breakdown of RTO when these assumptions do not hold.

The steps to proving Theorem 2.3.1 only use the facts that $\bar{Q}$ has orthonormal columns and that $\bar{Q}^T F(\theta)$ is invertible for all θ. In addition, the proof does not use any information on how Y is obtained. We can imagine using the same theory, and essentially the same algorithm, to sample from a proposal very similar to RTO by changing $\bar{Q}$ and Y. This leads to a family of proposal distributions parameterized by $\bar{Q}$ and Y.

An interesting observation is that the position of the original samples Ξ orthogonal to $\bar{Q}$ do not affect the resulting posterior samples in RTO. This leads to one minor modification to RTO, where we now sample $\xi \in \mathbb{R}^n$ instead of $\Xi \in \mathbb{R}^{m+n}$ and change Eq. 2.5 to

$$\theta^{(i)} = \arg\min_{\theta} \frac{1}{2} \left\| \bar{Q}^T (F(\theta) - Y) + \xi^{(i)} \right\|^2 \tag{2.7}$$

Here, $\xi \equiv \bar{Q}^T \Xi$.

Alg. 4 describes the procedure to sample from an RTO-like proposal. Fig. 2-3 and Fig. 2-4 depict two choices of $\bar{Q}$ and Y yielding different proposals.

Algorithm 4 Generate RTO-like proposal samples

1: Specify a particular $\bar{Q}$

2: Specify a particular Y

3: **for** $i = 1, \cdots, n_{\text{samps}}$ **do** in parallel

4: Sample $\xi^{(i)}$ from n-dimensional Gaussian

5: Find proposal samples $\hat{\theta}^{(i)}$ by minimizing Eq. 2.7.

6: Find $q(\hat{\theta}^{(i)})$ from Eq. 2.6

A particularly interesting choice of $\bar{Q}$ and Y is

$$\bar{Q} = \begin{bmatrix} I \\ 0 \end{bmatrix} \qquad Y = \begin{bmatrix} 0 \\ \lambda \end{bmatrix}$$

where λ is any fixed value. The resulting proposal distribution corresponds to a standard Gaussian centered at zero, which is exactly the prior. Hence, the prior is a member of the RTO-like proposal family. We will see in Chapter 3 that this prior can be used as a 'safe' starting point for an adaptive algorithm that changes $\bar{Q}$ and Y during sampling.

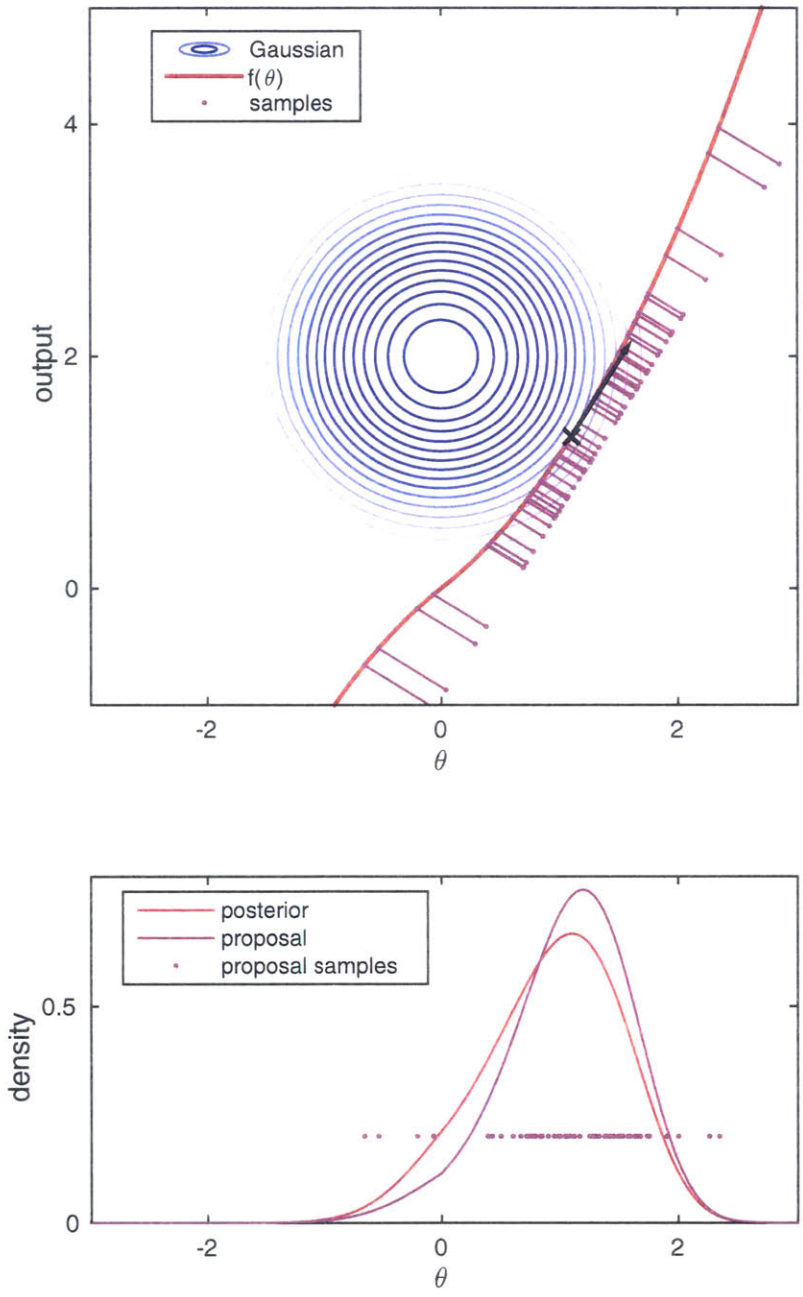

Figure 2-3: Modification to RTO's proposal. Instead of sampling from a joint Gaussian, we sample from the 1D Gaussian oriented along Q^T. The resulting samples are exactly the same.

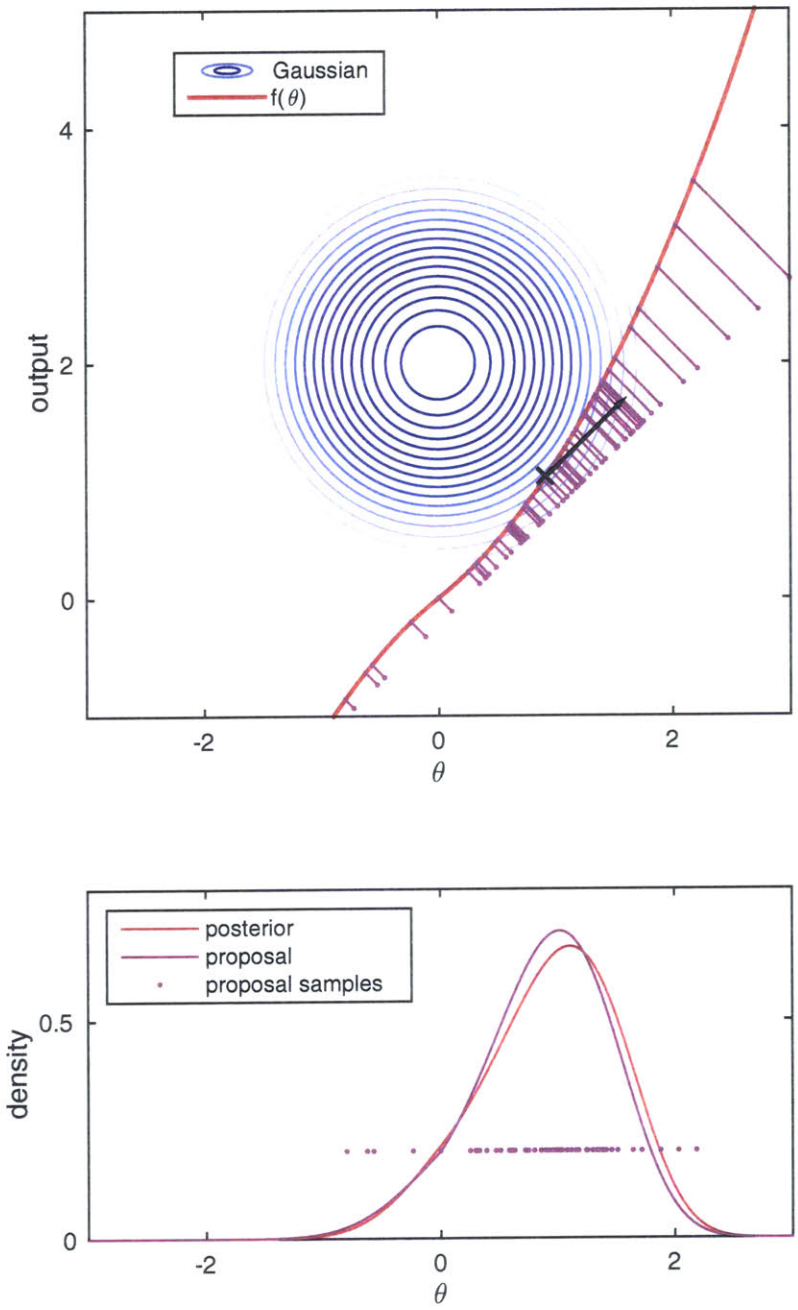

Figure 2-4: An RTO-like proposal obtained by modifying $\bar{Q}$ and Y. The proposal distribution is different from RTO's but its density can still be calculated.

2.4 Connections to Transport Maps

2.4.1 Transport-map accelerated MCMC

Transport-map accelerated MCMC defines a map $T(\cdot) : \mathbb{R}^n \to \mathbb{R}^n$ that acts on reference samples to obtain proposal samples. The transport map is constructed from posterior samples. It approximates a *true* map that pushes a reference density, such as a standard Gaussian, to the posterior exactly. The samples pushed forward using the approximate transport map are used as proposal samples. A lower-triangular structure is imposed on the map:

$$T(\theta_1, \theta_2, \cdots, \theta_n) = \begin{bmatrix} T_1(\theta_1) \\ T_2(\theta_1, \theta_2) \\ \vdots \\ T_n(\theta_1, \theta_2, \cdots, \theta_n) \end{bmatrix}$$

Here, θ_i is the ith component of θ, and $T_i(\cdot) : \mathbb{R}^i \to \mathbb{R}$ is the ith component of T. Each T_i is represented as a polynomial expansion. This triangular structure and polynomial representation allows the map to be evaluated and inverted easily. [19] derives the proposal density that results from pushing samples though the transport map, and also highlights additional constraints on the transport map needed for independence Metropolis-Hastings to be ergodic.

2.4.2 RTO as an approximate map

RTO can also be viewed as a transport map. In particular, the inverse map is defined from the relationship

$$\underbrace{\bar{Q}^T(F(\theta) - Y)}_{T^{-1}(\theta)} = \xi. \tag{2.8}$$

In transport-map accelerated MCMC, an explicit, invertible map is determined from approximate samples of the posterior. These approximate samples could be obtained from an unconverged MCMC chain, for example. In contrast, RTO defines

an implicit map before any sampling occurs. Using a transport map defined implicitly from the forward model, RTO avoids requiring an initial period of inefficient sampling from the posterior before obtaining a good proposal. However, since the RTO's map is not defined explicitly, we can not directly evaluate it. We can only evaluate its inverse using the forward model and hence, must employ optimization techniques to push forward one reference sample.

2.5 Connections to Implicit Sampling

2.5.1 Implicit sampling

Implicit sampling incorporates a very similar idea to RTO. It also uses a transport map from a Gaussian random variable to obtain samples close to the posterior. Two main *differences* between it and the RTO presented here, are that the proposal samples are used in self-normalized importance sampling rather than in independence Metropolis-Hastings, and that typically multiple inverse problems are solved sequentially in a filtering context[2] rather than a single posterior distribution being explored at a time.

First, we paraphrase the random map implementation of implicit sampling. In implicit sampling, we consider more general posterior densities where

$$p(\theta|y) \propto \exp\left(-R(\theta)\right). \tag{2.9}$$

Here $R(\cdot) : \mathbb{R}^n \to \mathbb{R}^n$ is the negative-log-posterior function. In our case, for posteriors in least-squares form, $R(\theta) = \frac{1}{2}\|F(\theta) - Y\|^2$.

Implicit sampling imposes the ansatz that the map $\theta = T(\xi)$ solves the following scalar equation

$$R(\theta) - R(\bar{\theta}) = \frac{1}{2}\xi^T\xi \tag{2.10}$$

where

$$\bar{\theta} = \arg\min_{\theta} R(\theta).$$

[2]Filtering can be thought of as sequentially propagating uncertainty through a time-dependent model and performing Bayesian inference when observations arrive.

Intuitively, Eq. 2.10 matches the negative-log-posterior with the negative-log of the density from a standard Gaussian. [5] list four additional desirable conditions for the map $T(\cdot)$ in implicit sampling:

(i) $T(\cdot)$ is one-to-one and onto with probability one.

(ii) $T(\cdot)$ is smooth.

(iii) $T(\xi = 0) = \bar{\theta}$.

(iv) $\left|\frac{\partial T(\xi)}{\partial \xi}\right|$ can be computed easily.

Eq. 2.10 and conditions (i) to (iv) above do not uniquely define $T(\cdot)$. The random map implementation of implicit sampling [17] chooses the following map:

$$\theta = T(\xi) = \bar{\theta} + \lambda(\xi) L^T \frac{\xi}{\|\xi\|} \tag{2.11}$$

where L is any invertible matrix we choose, and we solve for $\lambda(\xi)$ that satisfies Eq. 2.10 using optimization. This implementation can be thought of as sampling ξ from a standard Gaussian, taking its direction $\frac{\xi}{\|\xi\|}$ and magnitude $\|\xi\|$, and marching from the posterior mode $\bar{\theta}$ in the direction $L^T \frac{\xi}{\|\xi\|}$ until Eq. 2.10 is satisfied using the magnitude $\|\xi\|$. One suggested choice of L is $L^T L = H$, where H is the Hessian of R. The proposal density is evaluated as

$$q(\theta) \propto \left|\frac{\partial T(\xi)}{\partial \xi}\right|^{-1} \exp\left(-\frac{1}{2}\xi^T \xi\right) \tag{2.12}$$

where $\xi = T^{-1}(\theta)$ and

$$\left|\frac{\partial T(\xi)}{\partial \xi}\right| = 2|L|\left(\xi^T \xi\right)^{1-n/2} \left|\frac{\lambda^{n-1}}{2(\nabla_\theta R) L^T (\xi/\|\xi\|)}\right|$$

which is proven in [17]. Note that the expression given in [17] is

$$\frac{p(\theta|y)}{q(\theta)} \propto \left|\frac{\partial T(\xi)}{\partial \xi}\right| \exp\left(-R(\bar{\theta})\right)$$

which can be derived from Eq. 2.9, 2.10 and 2.12.

The implicit map is well defined when Eq. 2.11 is one-to-one. This occurs when the level sets of the posterior is 'star shaped'. In other words, for any level set $R(\theta) - R(\bar{\theta}) = c > 0$, a straight line passing through the posterior mode $\bar{\theta}$ must intersects the level set at exactly two points. This is exaplained in more detail in a section of [10]. When this condition is not satisfied, implicit sampling will not sample from the true posterior. This is analogous to the breakdown of RTO described in Sec. 3.1.

2.5.2 RTO's alternative ansatz to define an implicit map

However, there are striking similarities between the proposals used in implicit sampling and RTO. Both algorithms solve an optimization problem to transport Gaussian distributed samples to proposal samples close to the posterior; both proposal generating processes can be defined using a map; and both maps satisfy several desirable properties. This section explores these connections in more detail.

In RTO, we satisfy a different ansatz but the same conditions (i) through (iv). From Sec. 2.4, RTO's map $T(\cdot)$ is defined implicitly from the *vector* equation, see 2.8.

$$\bar{Q}^T(F(\theta) - Y) = \xi.$$

This equation replaces Eq. 2.10 and also defines a unique map.

We now show that the map defined by Eq. 2.8 satisfies the conditions (i) to (iv) above under the same assumptions as Theorem 2.3.1. Condition (i) asks for $T(\cdot)$ to be one-to-one and onto. In RTO, when $\bar{Q}^T J(\theta)$ is invertible, and there exists a θ such that $\bar{Q}^T(F(\theta) - (Y + \Xi)) = 0$ for any Ξ, then condition (i) holds. Condition (ii) requires that $T(\cdot)$ be smooth. If $f(\cdot)$ is differentiable, then $F(\cdot)$ is differentiable. Then, $T^{-1}(\cdot)$ is smooth, and $T(\cdot)$ is smooth. Condition (iii) is shown to be true by setting the derivative of the objective in Eq. 2.4 to zero and comparing it to Eq. 2.8 with the substitution $\xi = 0$. Condition (iv) holds due to Theorem 2.3.1.

To summarize, we can interpret RTO and implicit sampling as implicitly defining a particular map from a Gaussian random variable ξ to the proposal distribution. Both

these algorithms satisfy several desirable conditions, listed as (i) to (iv) above, which produce the exact posterior when the forward model is linear, produce a proposal close to the posterior when the forward model is nonlinear, and allow us to determine their proposal densities. The main difference between the algorithms is the ansatz. Implicit sampling matches the log-posterior to the log-density at ξ, and RTO matches the residual, $F(\theta) - Y$, in certain directions, to ξ.

Chapter 3

Adaptive RTO

This chapter discusses a few drawbacks of the original RTO algorithm and uses the geometric interpretation developed previously to construct an improved, adaptive version of the algorithm. Adaptive RTO is designed to be more robust and efficient than the original. Numerical examples demonstrate that when RTO fails in sampling from the posterior, adaptive RTO can succeed, and when RTO samples from the posterior correctly, adaptive RTO does so in a more efficient manner.

Previously, we identified two assumptions that RTO and the forward model needed to satisfy in order for RTO to work properly. When these assumptions do not hold, RTO does not sample from the posterior. In fact, a naïve implementation of RTO fails dramatically, and the obtained 'posterior' samples collapse in parameter space, as will be shown later. A change in $\bar{Q}$, one of the tunable parameters for RTO-like proposals, can fix this issue.

A second issue is when RTO's proposal distribution varies significantly from the posterior. This leads to inefficient sampling in independence Metropolis-Hastings. This can be easily demonstrated using a skewed posterior in 1D. The differing proposal and posterior distributions are caused by the constraint that the center of the Gaussian is projected to the mode of the posterior. This problem can be alleviated by changing the point Y, another tunable parameter for RTO-like proposals.

In light of these problems, we propose an adaptive version of RTO that, through changing $\bar{Q}$ and Y, searches within the RTO-like proposal family to find a distribution

that satisfies the assumptions and is as close to the posterior as possible. One metric for measuring the distance between two distributions is the K-L divergence. Using the K-L divergence, we obtain an optimization problem constrained on a manifold, which can be solved without any additional forward model evaluations.

The resulting method is an adaptive RTO algorithm that simultaneously samples from the posterior and determines the optimal RTO-like proposal. Experiments show that by using adaptive RTO, we sample from the true posterior on problems where the original assumptions for RTO do not hold, and that we are more efficient than RTO on problems where RTO is valid.

This chapter is organized as follows: Section 3.1 describes problem-specific challenges that RTO may face; Section 3.2 outlines the core idea of adaptive RTO; Section 3.2.1 describes the details of the optimization problem that is solved during adaption; Section 3.2.2 summarizes all the steps in the adaptive algorithm, and Section 3.3 shows two numerical examples comparing the original RTO and the adaptive version.

3.1 Drawbacks of RTO

In this section we highlight two drawbacks of the original RTO algorithm that motivate an adaptive alternative. These are two cases where the original RTO algorithm either does not work or is inefficient. They are: when the assumptions required for RTO do not hold, and when RTO's proposal is very different from the posterior.

3.1.1 Assumptions do not hold

From Section 2.3, the two main assumptions in order for RTO to work are:

- $\bar{Q}^T J(\theta)$ is invertible for any θ.
- There exists a θ such that $\bar{Q}^T(F(\theta) - Y) = \xi$ for any ξ.

When these assumptions do not hold, a naïve implementation of RTO fails dramatically, as shown in Fig. 3-1. This phenomenon results from the fact that not all samples

can be projected orthogonal to $\bar{Q}$ *and* onto the manifold. When samples cannot be projected onto the manifold, the optimization routine stops at a non-zero minimum, which causes these samples to cluster. The proposal samples are not distributed according to the analytical formula. In addition, the samples cluster around the point where $\bar{Q}^T J(\theta)$ is singular and do not cover the entire range of the posterior. Following through with the algorithm and blindly using its proposal density within independence Metropolis-Hastings will not result in sampling from the true posterior. In fact, the problem is further exacerbated through independence Metropolis-Hastings, since the proposal density formula evaluated around the cluster is very small and will cause independence M-H to reject any steps proposed outside of the cluster.

Another view of this breakdown is that the map from the Gaussian random samples ξ and the parameter values θ is no longer one-to-one. In this case, there may be zero or more than one $(\theta, f(\theta))$ points on the manifold that intersect a projected line tangent to $\bar{Q}$ and originating at $Y + \Xi$.

Fortunately, we are able to detect this issue while sampling the proposal. During the optimization step of RTO, the presence of a non-zero local minimum θ_* in the residual, i.e. when the residual is not zero after the optimization halts, indicates that $\bar{Q}^T J(\theta_*)$ is not invertible at that local minimum and that the proposal density formula does not hold. This can be seen by taking the derivative of the objective function with respect to θ and setting it to zero at the local minimum.

$$\frac{\partial}{\partial \theta}\left[\bar{Q}^T(F(\theta) - (Y + \epsilon))\right]\Big|_{\theta=\theta_*} = 0$$

$$2J(\theta_*)^T \bar{Q}\bar{Q}^T(F(\theta_*) - (Y + \epsilon)) = 0$$

Thus, at the local minimum, either the residual is reduced to zero, in other words $\bar{Q}^T(F(\theta_*) - (Y + \epsilon)) = 0$, or $J(\theta_*)^T\bar{Q} \in \mathbb{R}^{n\times n}$ is singular. Hence, when we find a local minimum with a non-zero residual, $\bar{Q}^T J(\theta_*)$ is not invertible at that local minimum and assumptions of RTO are violated.

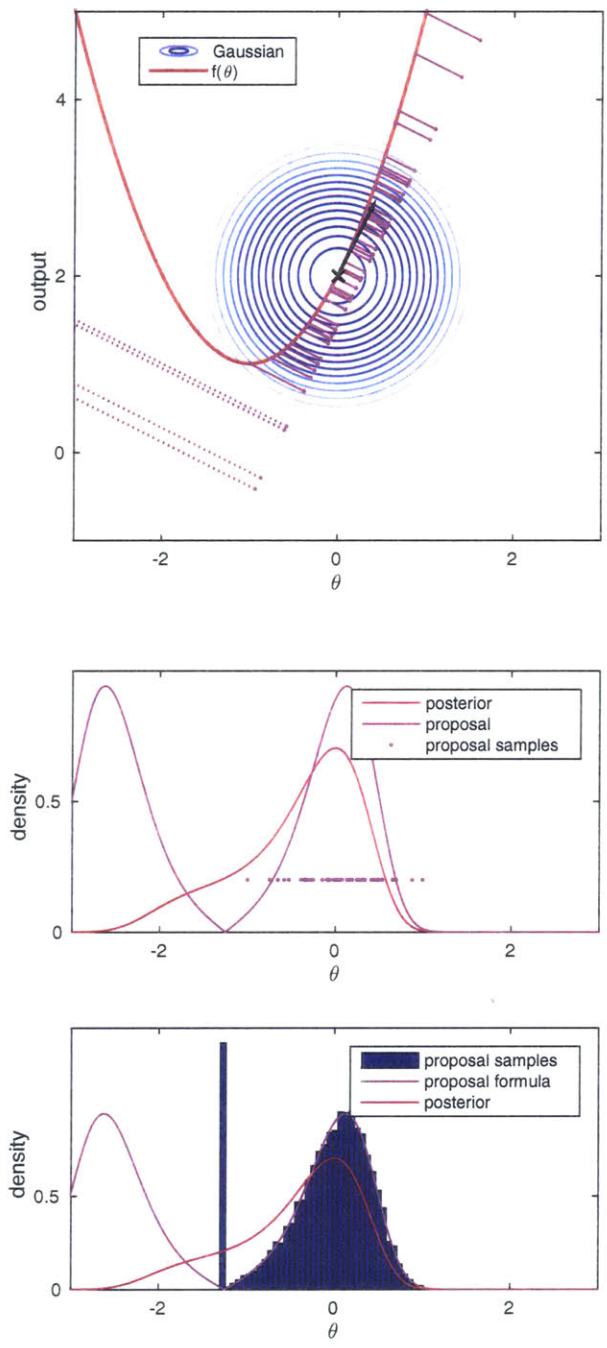

Figure 3-1: Case where RTO assumptions do not hold. (Top) Gaussian and manifold. (Middle) Posterior, proposal formula and projected samples. (Bottom) Histogram of actual samples drawn from RTO. There is a quadratic forward model with one parameter and one observation. The projected path from some random samples ξ do not pass through the manifold. The resulting analytical form of the proposal distribution is no longer valid.

3.1.2 Proposal differs from posterior

The second drawback of RTO occurs when the proposal is valid but is very different from the posterior. In this case, the sampling is inefficient. In other words, we obtain a highly-correlated chain from independence Metropolis-Hastings, and would require many steps of the Markov chain to obtain a large effective sample size. This is because the posterior over proposal density (sometimes called the importance weight) varies a lot from sample to sample. More steps in independence M-H more are likely to be rejected.

We illustrate this example (in 1D) using a posterior that is highly skewed, seen in Fig. 3-2. In the original RTO formulation, there is an equal probability of sampling ϵ to the left and the right of $\bar{\theta}$, and, thus, there must be equal probability mass in RTO's proposal on either side of $\bar{\theta}$. As such, the proposal median must be located at the posterior mode $\bar{\theta}$. This constraint guarantees that the proposal from RTO will not be able to approximate skewed posteriors well. Fig. 3-3 shows that this scenario can be avoided if we change Y and use an altered RTO-like proposal.

These two drawbacks indicate that the original RTO algorithm's performance is problem-dependent. The sampling algorithm may fail completely, such as when RTO's assumptions do not hold, or be very inefficient, such when the proposal is far from the posterior. They also suggest that changing $\bar{Q}$ and Y can be helpful in different scenarios. In the next section, we build upon this intuition and outline an adaptive RTO algorithm that changes $\bar{Q}$ and Y as we sample from the posterior.

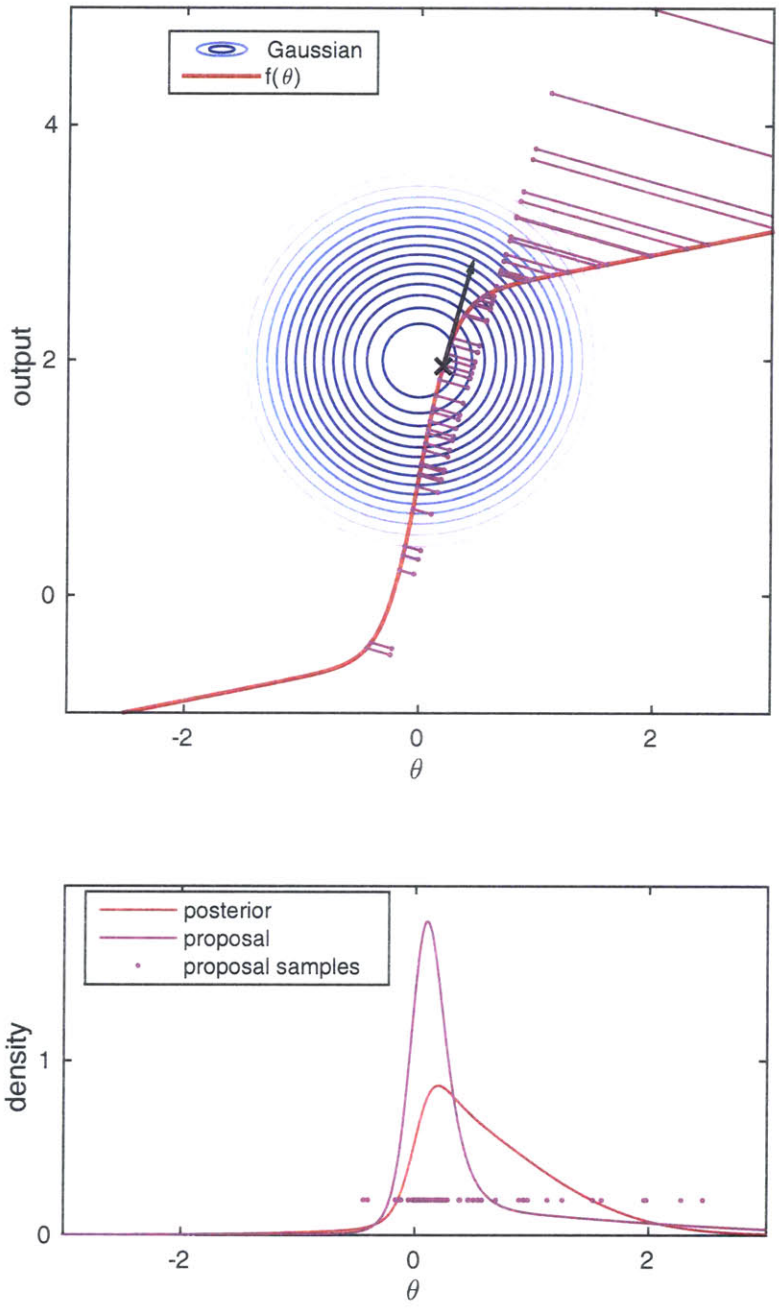

Figure 3-2: Case of RTO where proposal varies from posterior. (Top) Gaussian and manifold. (Bottom) Posterior, proposal formula and projected samples. Although the proposal formula is correct, the proposal distribution is very different from the posterior.

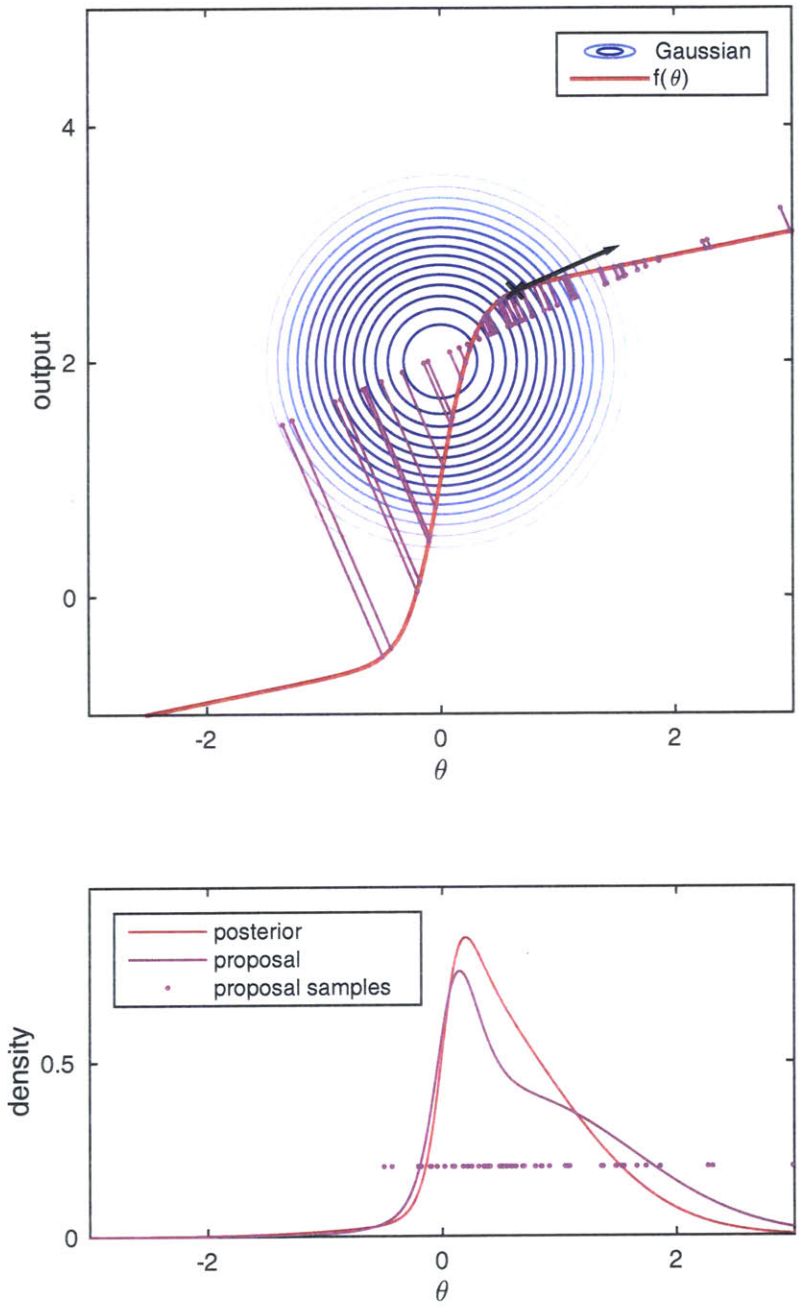

Figure 3-3: RTO-like proposal on previous problem. (Top) Gaussian and manifold. (Bottom) Posterior, proposal formula and projected samples. Choosing a different $\bar{Q}$ and Y creates a proposal closer to the posterior.

3.2 Adapting the Proposal

The drawbacks outlined in the previous section motivate a change in $\bar{Q}$ and Y. This involves choosing a different proposal within the RTO-like proposal family. We can freely choose any distribution from the RTO-like proposal family provided that the RTO's assumptions hold for the corresponding $\bar{Q}$. This is because we are able to sample from any of them and we can evaluate any of their densities.

We propose an adaptive version of RTO where we avoid the two undesirable situations outlined in Section 3.1 by changing $\bar{Q}$ and Y from their heuristic values in RTO. In particular, we find the 'best' values of $\bar{Q}$ and Y such that we obtain the most efficient proposal distribution in the family of RTO-like proposal distributions.

An efficient proposal is one that is close to the posterior, so named because it makes independence Metropolis-Hastings sampling more efficient. Following a similar adaptive approach to [19], we choose to quantitatively measure the distance between the two distributions using the Kullback-Leibler (KL) divergence *from* the posterior *to* the proposal. This idea is similar in spirit to variational techniques where the closest distribution in a parameterized family is used as an approximation.

3.2.1 Optimization of the proposal distribution

The details of optimizing the KL divergence are outlined here. The KL divergence is defined as

$$\begin{aligned} D_{KL}[p(\theta|y)||q(\theta)] &= \mathbb{E}_{p(\theta|y)}\left[\log\left(\frac{p(\theta|y)}{q(\theta)}\right)\right] \\ &= \mathbb{E}_{p(\theta|y)}\left[-\log|\bar{Q}^T J(\theta)| + \frac{1}{2}\left\|\bar{Q}^T\left(F(\theta)-\bar{Y}\right)\right\|^2\right] + \mathrm{C} \end{aligned}$$

Evaluating the KL divergence requires an expectation over the posterior. During sampling, we have approximate samples from the posterior using the Markov chain generated up to this point. Given posterior samples $\theta^{(i)}$, their function evaluations $F(\theta^{(i)})$, and their Jacobians $J(\theta^{(i)})$, we can estimate the KL divergence by

$$D_{KL}[p(\theta|y)||q(\theta)] \approx \sum_{i=1}^{n_{\text{samp}}} \left[-\log|\bar{Q}^T J(\theta^{(i)})| + \frac{1}{2}\left\|\bar{Q}^T\left(F(\theta^{(i)}) - \bar{Y}\right)\right\|^2 \right] + \text{C}$$

Then, the optimization problem is

$$\min_{\bar{Q},\bar{Y}} \sum_{i=1}^{n_{\text{samp}}} \left[-\log|\bar{Q}^T J(\theta^{(i)})| + \frac{1}{2}\left\|\bar{Q}^T\left(F(\theta^{(i)}) - \bar{Y}\right)\right\|^2 \right] \quad \text{subject to} \quad \bar{Q}^T\bar{Q} = I \quad (3.1)$$

This is an optimization problem over a manifold, in particular, a product of a Grassmann manifold and an Euclidean manifold. The directions $\bar{Q}$ is constrained to a Grassmann manifold, and the center Y is constrained to an Euclidean manifold. Manifold optimization problems are well-studied and can be solved using specialized algorithms, such as steepest descent and conjugate gradient over manifolds [8]. In this thesis, we use the MATLAB library *manopt* [3].

The following section outlines the new adaptive RTO algorithm.

3.2.2 Adaptive RTO: Algorithm Overview

Adaptive RTO is summarized as follows

1. Start with a valid RTO-like proposal density (such as the prior).
2. Obtain samples from the posterior using the current proposal and independence M-H.
3. Optimize for the best $\bar{Q}$ and Y given the samples of the posterior obtained thus far.
4. Repeat steps 2 and 3 for a fixed number of adaptation steps.
5. Use the final adapted proposal with independence M-H to obtain a large number of posterior samples efficiently.

A more detailed version is presented in Algorithm 5.

Algorithm 5 Adaptive RTO

Initialize an empty set of posterior points $\mathcal{L} = \emptyset$.
Start with a valid proposal defined by $\bar{Q}^{(0)}$ and $Y^{(0)}$.
for $k = 1, \cdots, n_{\text{adapt}}$ **do** in series
    for $i = 1, \cdots, n_{\text{small}}$ **do** in parallel
        Obtain proposal $\hat{\theta}^{(i)}$, $F(\hat{\theta}^{(i)})$, and $J(\hat{\theta}^{(i)})$ using Algorithm 4 with $\bar{Q} = \bar{Q}^{(k-1)}$ and $Y = Y^{(k-1)}$.
    for $i = 1, \cdots, n_{\text{small}}$ **do** in series
        Use independence Metropolis-Hastings to obtain posterior samples $\theta^{(i)}$, $f(\theta^{(i)})$, and $J(\theta^{(i)})$.
    Add the current posterior model-Jacobian pairs $\{[\theta^{(1)}, F(\theta^{(1)}), J(\theta^{(1)}], \cdots\}$ to the set of posterior points $\mathcal{L}$.
    Use that set $\mathcal{L}$ to optimize Eq. 3.1 for $\bar{Q}^{(k)}$ and $Y^{(k)}$.

for $i = 1, \cdots, n_{\text{large}}$ **do** in parallel
    Obtain proposal samples $\hat{\theta}^{(i)}$ using Algorithm 4 with $\bar{Q} = \bar{Q}^{(n_{\text{adapt}})}$ and $Y = Y^{(n_{\text{adapt}})}$.
for $i = 1, \cdots, n_{\text{large}}$ **do** in series
    Use independence Metropolis-Hastings to obtain posterior samples $\theta^{(i)}$.

Here, n_{adapt} is the number of adaptation intervals, n_{small} is the number of samples to take during each adaptation step, and n_{large} is the number of samples to take using the final adapted proposal distribution. For the initial choice of $\bar{Q}^{(0)}$ and $Y^{(0)}$, we can either use RTO's original heuristic choices, which is valid only when RTO's assumptions hold, or the prior distribution, which is always valid. The idea is to obtain more samples that accurately describe the posterior in conjunction with a RTO-like proposal distribution closer to the posterior.

Note that the manifold optimization problem does not require any additional forward model evaluations, which are typically the most computationally expensive

components. The optimization only uses the list of stored forward model evaluations and Jacobians to describe the posterior. When the algorithm starts off poorly, with an under-sampled or unrepresentative list of posterior points, adaptive RTO may select an invalid proposal during the optimization step. Experimentally, we observe that adaptive RTO is more robust when we use a larger number of samples in each adaptation step.

Typically, large numbers of stored forward model evaluations and Jacobians are desired. Specific implementation details concerning how many points to keep in the list and whether to thin out the points, such as, for example, by keeping every n^{th} point, merit further exploration, and are likely to be problem-dependent. Efficient manifold optimization in high dimensions and accelerated starting procedures to avoid under-sampling are potential directions for future development.

3.3 Numerical Examples

The following two numerical examples demonstrate two cases where adaptive RTO improves upon the original version. In the first test case, RTO's assumptions were not originally satisfied. We observe that RTO breaks down, but adaptive RTO does not. In the second test case, RTO's assumptions *are* satisfied, and we show that adaptive RTO is more efficient than the original.

3.3.1 Boomerang example

For this example, there are two parameters and one observation. The forward model is

$$f(\theta) = \begin{cases} 3(\theta_2 + 2\theta_1 - 1) & \text{when } \theta_1 \leq -1 \\ 3(\theta_2 - \theta_1^2) & \text{when } -1 < \theta_1 \leq 1 \\ 3(\theta_2 - 2\theta_1 + 1) & \text{when } \theta_1 > 1 \end{cases}$$

It is plotted in Fig. 3-4. We observe $y = 1$ and the prior is a Gaussian with mean $[1, 0]^T$ and identity covariance. The resulting posterior is also shown in Fig. 3-4.

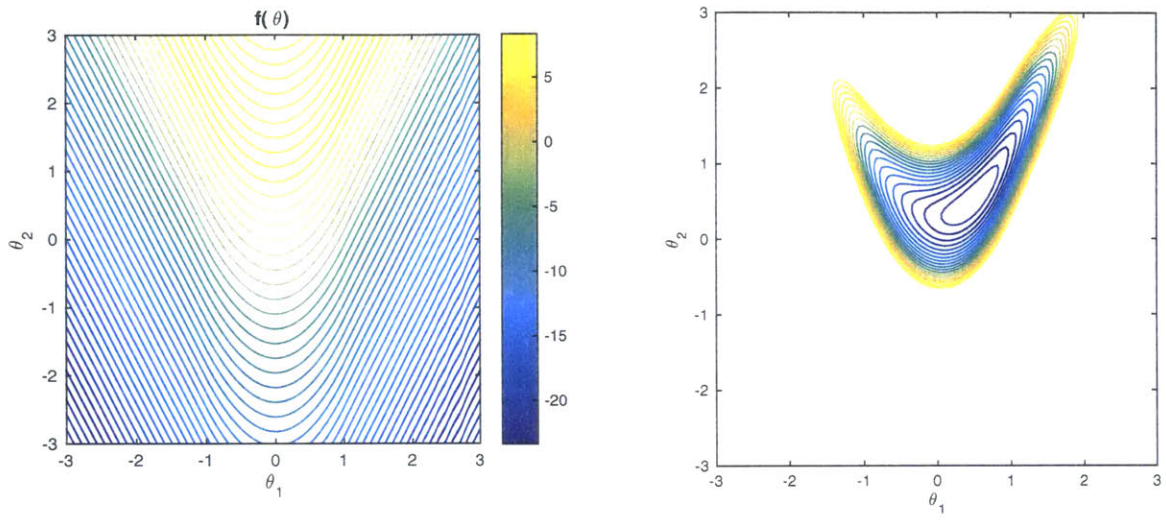

Figure 3-4: Boomerang example problem. (Left) Forward model. (Right) Posterior.

When we naïvely implement RTO, we observe several proposal samples that do not converge, in other words the residuals do not decrease to zero. As a result, these samples cluster around a ridge in $\mathbb{R}^2$, shown in Fig. 3-6. This is caused by the fact that $\bar{Q}^T J(\theta)$ is not invertible at those values of θ along the ridge. In addition, since the proposal formula prescribes a very low proposal density near the ridge, independence Metropolis-Hastings almost always rejects the samples that do not lie on the ridge. The result is a Markov chain that is not distributed according to the posterior.

When we change $\bar{Q}$ so that RTO's assumptions hold, for example when $\bar{Q}$ corresponds to the prior, we obtain the correct samples from the posterior; see Fig. 3-5. Adaptive RTO selects a proposal distribution that satisfies RTO's conditions and is also efficient. In Fig. 3-7, the autocorrelation from adaptive RTO decays faster than that from using the prior as a proposal. For this test case, we show that adaptive RTO succeeds in sampling from the posterior when RTO does not.

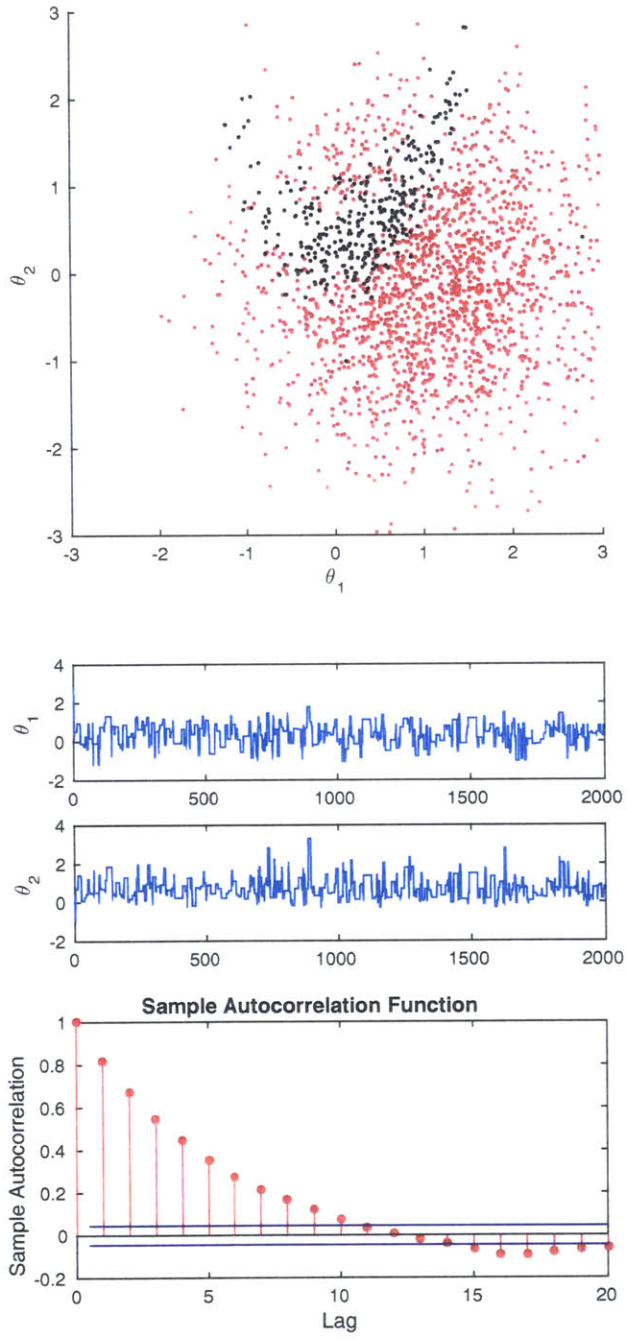

Figure 3-5: Boomerang example: prior results. (Top) Scattered proposal samples, in red, and posterior samples, in black. (Middle) Markov chains after independence Metropolis-Hastings. (Bottom) Autocorrelation function of θ_2.

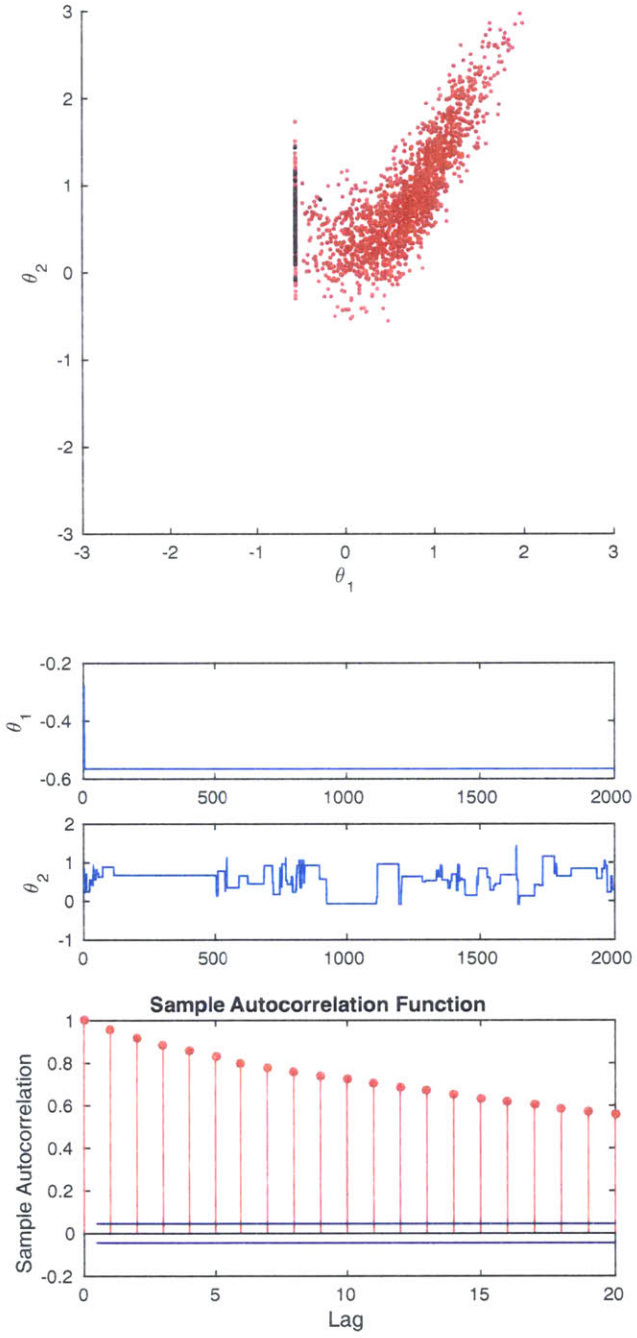

Figure 3-6: Boomerang example: RTO results. (Top) Scattered proposal samples, in red, and posterior samples, in black. (Middle) Markov chains after independence Metropolis-Hastings. (Bottom) Autocorrelation function of θ_2.

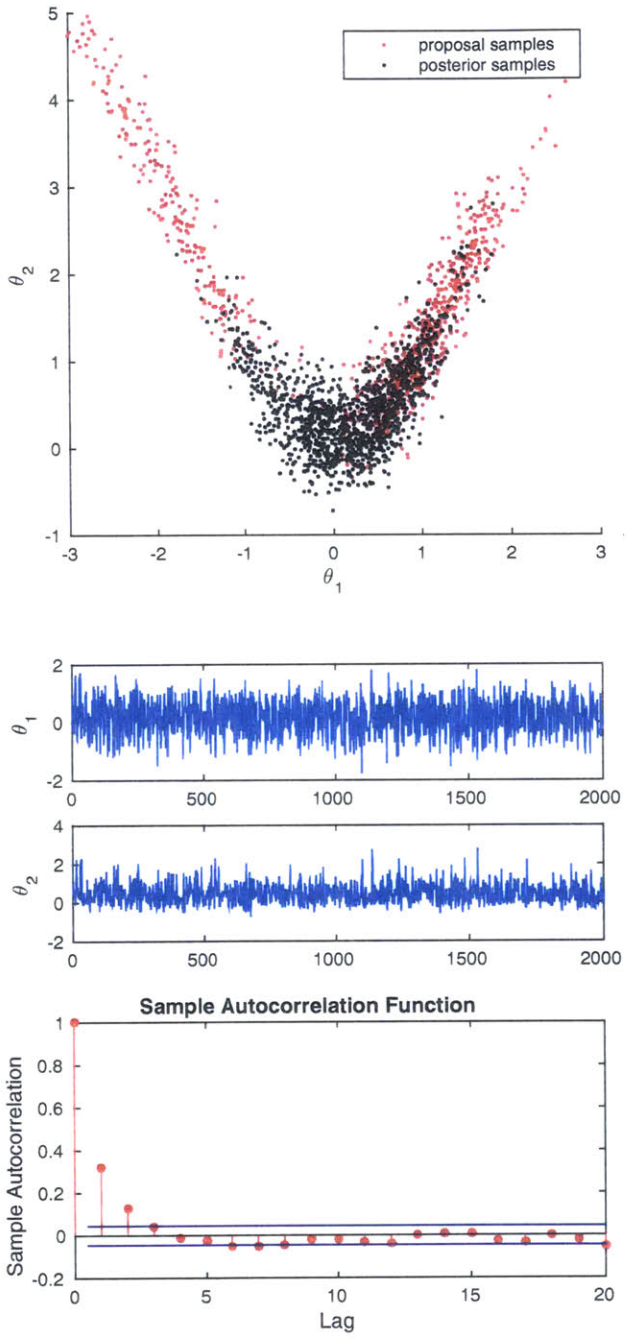

Figure 3-7: Boomerang example: adaptive RTO results. (Top) Scattered proposal samples, in red, and posterior samples, in black. (Middle) Markov chains after independence Metropolis-Hastings. (Bottom) Autocorrelation function of θ_2.

3.3.2 Cubic example

Similar to the previous example, we have two parameters and one observation. The forward model is

$$f(\theta) = 10\theta_2 - 10\theta_1^3 + 5\theta_1^2 + 6\theta_1$$

and is shown in Fig. 3-8. We observe $y = 1$, and the prior is a Gaussian with mean $[1, 0]^T$ and identity covariance. The resulting posterior is also shown in Fig. 3-8.

For this example, both RTO and adaptive RTO sample relatively efficiently from the true posterior; see Fig. 3-9 and 3-10. The adaptive algorithm has a faster decaying autocorrelation function. Adaptive RTO has a higher acceptance ratio: 0.76, compared to RTO's 0.55. This indicates that adapting the proposal can improve sampling efficiency.

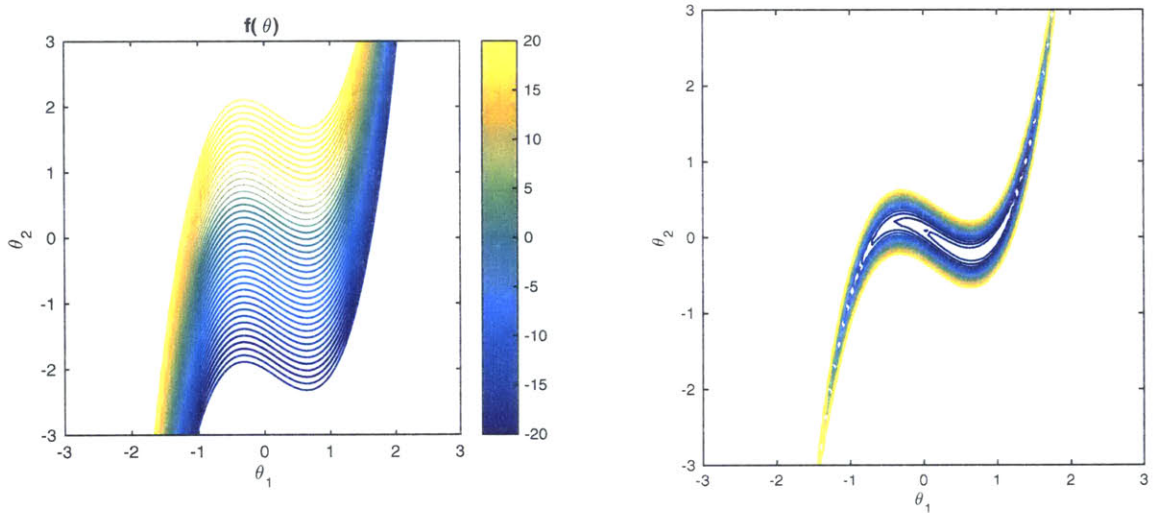

Figure 3-8: Cubic example problem. (Left) Forward model. (Right) Posterior.

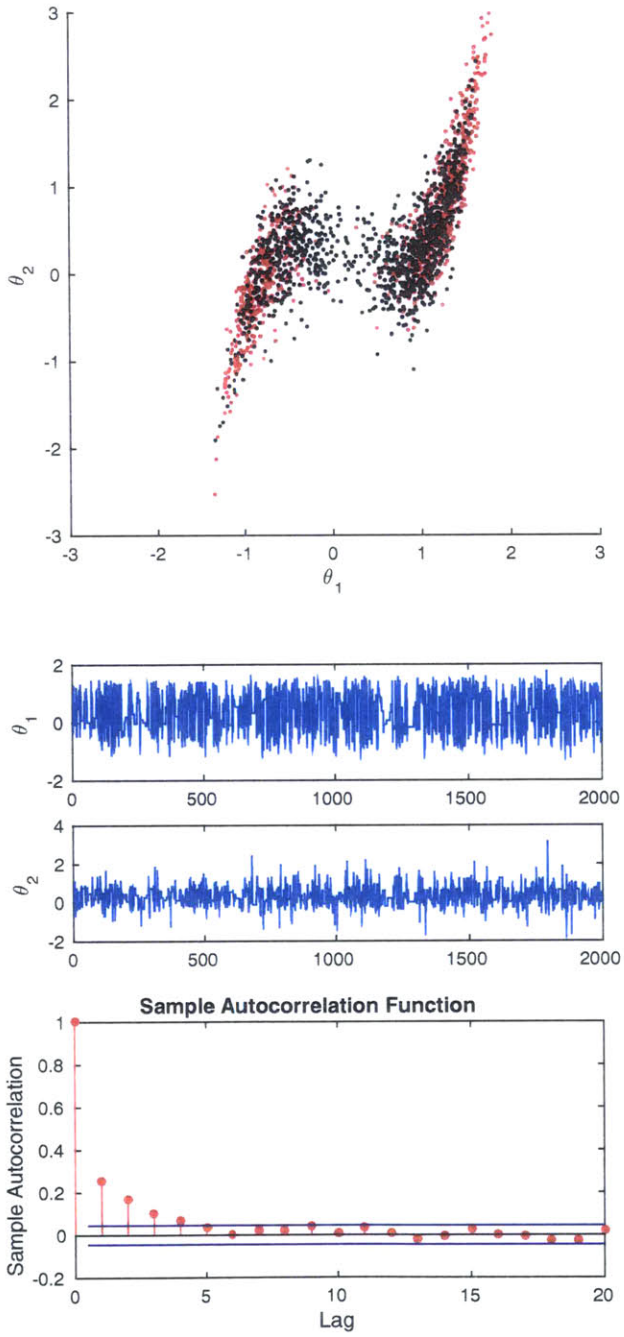

Figure 3-9: Cubic example: RTO results. (Top) Scattered proposal samples, in red, and posterior samples, in black. (Middle) Markov chains after independence Metropolis-Hastings. (Bottom) Autocorrelation function of θ_2.

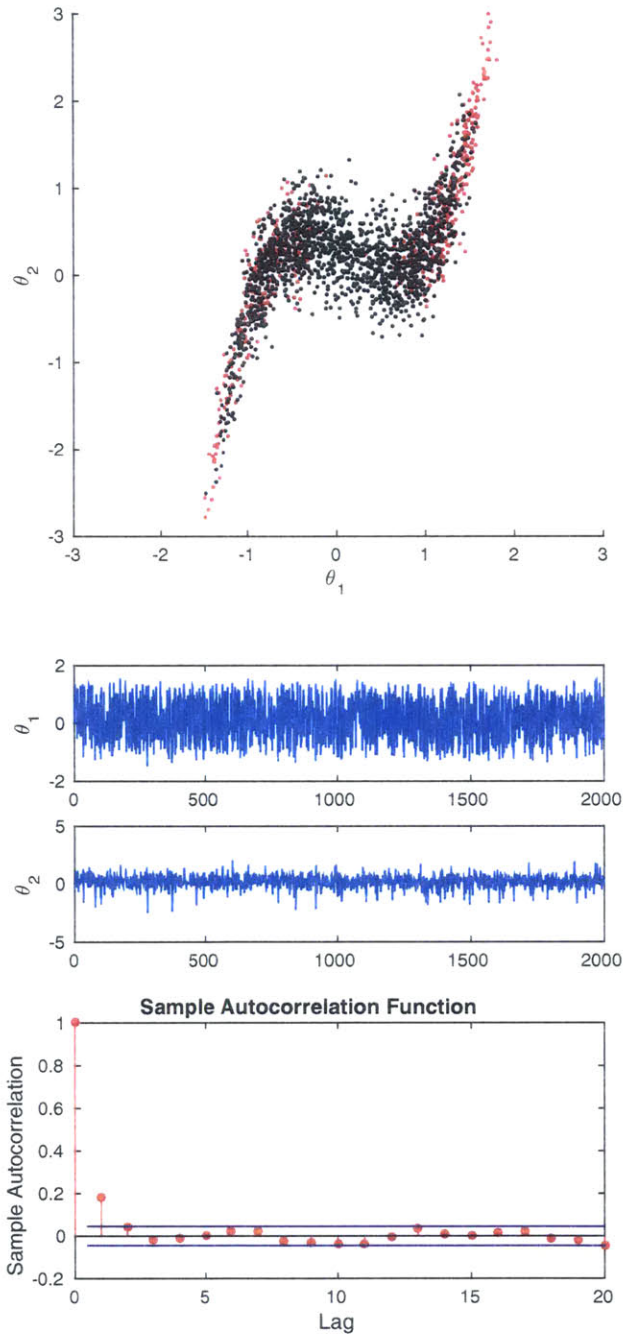

Figure 3-10: Cubic example: adaptive RTO results. (Top) Scattered proposal samples, in red, and posterior samples, in black. (Middle) Markov chains after independence Metropolis-Hastings. (Bottom) Autocorrelation function of θ_2.

3.4 Concluding Remarks

Motivated from the geometric perspective of RTO, we present an adaptive version of RTO that is more robust and efficient, through optimizing for the best proposal in the family of RTO-like proposals. This algorithm simultaneously samples from the posterior and determines the best proposal to do so. Numerical examples show that adaptive RTO improves upon the original version.

Chapter 4

Prior Transformations: How to use RTO with Non-Gaussian Priors

The methodology and algorithms described the previous two chapters are well defined for Bayesian inverse problems with Gaussian priors and observational noise. The resulting posterior density functions from such inverse problems can be written in least-squares form, which is required for RTO. However, in many applications, the prior distribution may be non-Gaussian, and the posterior may not be rewritten in least-squares form. In such cases, we are unable to directly apply RTO or its adaptive variant.

We address this issue by transforming the inverse problem to one that is amenable to RTO. In doing so, we demonstrate that RTO can be applicable to a broader range of priors. A new random variable, the reference parameter, is constructed to have predetermined one-to-one mapping[1] to the physical parameter. The reference parameter is designed to have a Gaussian prior. The inverse problem is then redefined over the reference parameter and solved using RTO. The resulting posterior samples of the reference parameter, are then transformed back to corresponding posterior samples of the physical parameter.

This chapter illustrates the above procedure using a concrete application: L1-type

[1]The prior transformation mapping is used as a change of variables for the inverse problem. Do not confuse this with the RTO's approximate map in Sec. 2. RTO's approximate map describes the steps within RTO that change Gaussian samples to proposal samples.

priors. L1-type priors, which include Total Variation (TV) priors and $B^s_{1,1}$ Besov space priors, are often used to infer blocky, discontinuous signals. We build up the analytical prior mapping for L1-type priors and interpret its effect on the posterior; then, we conclude with a numerical example, a reconstruction of a blocky signal from noisy and blurred measurements using a TV prior. In the numerical example, we show that this approach yields the correct answer[2] with *orders of magnitude* increases in efficiency compared to other state-of-the-art sampling algorithms.

This chapter is organized as follows: Section 4.1 defines L1-type priors and describes their use; Section 4.2 broadly sketches the procedure to sample from the posterior with a non-Gaussian prior using RTO and a prior transformation; Section 4.3 describes the prior transformation for one parameter with a Laplace prior; Section 4.4 describes the prior transformation for multiple parameters with a joint L1-type prior, and Section 4.6 presents the numerical example in detail.

4.1 L1-type Priors

L1-type priors are defined to be priors that rely on the L1 norm of the parameters[3]. The L1-type priors considered in this chapter are those that have independent Laplace distributions on an invertible transformation of the parameters. These priors can be written in the following form

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

where $D \in \mathbb{R}^{n\times n}$ is an invertible matrix and $\lambda \in \mathbb{R}$ is a hyper-parameter of the prior.

Three examples of L1-priors are: total variation (TV) priors, $B^s_{1,1}$ Besov space priors, and impulse priors. The first two are used to specify the a priori knowledge that the parameter field[4] is blocky or has discontinuous jumps. The impulse prior is used to specify the a priori knowledge that the parameter values themselves are

[2]This approach will obtain the correct answer, samples from the true posterior, under the same assumptions as in RTO.

[3]L1-type priors may also rely on the L1 norm of multiple linear combinations of the parameters.

[4]The parameter field is the random field for which we wish to solve the inverse problem.

sparse[5].

TV priors are related to the much more well-known TV regularization, wherein the total variation of a signal is used as a regularization term in optimization to promote a blocky, discontinuous solution in signal processing. [23] introduced constrained optimization on the total variation to remove noise from images. [1] used total variation regularization to reconstruct blocky images. However, when we use a TV prior in Bayesian inference, many have observed that the samples are not discontinuous or blocky in nature even though the posterior mode may be thus. The TV prior can be thought of as multiple independent Laplace distributions on the differences between neighboring parameter node values.

TV priors are criticized for not being discretization-invariant[6]. [7] proposes an alternative prior, the Besov space prior. The Besov space prior is discretization-invariant and also promotes blocky solutions. The $B^s_{1,1}$ Besov space priors can be thought of as independent Laplace distributions on the wavelet coefficients of the signal[7]. Finally, the impulse prior is an independent Laplace distribution on the parameter values themselves.

The exploration of L1-type priors and experimental verification of theoretical results require sampling from high dimensional, irregular posteriors. Sampling from such posteriors using traditional MCMC techniques is difficult. The posteriors are not differentiable even with differentiable forward models. Typically, derivative-free MCMC algorithms, such as random-walk Metropolis-Hastings or single-component adaptive Metropolis, are used. [15] derived an over-relaxed Gibbs sampler to sample from the such a posterior when the forward model is linear.

[5]A sparse parameter vector has many components equal to or very close to zero.

[6]Loosely, a prior distribution is discretization-invariant when the mean and mode of the posterior that describes the uncertainty in the parameter field converge to an infinite dimensional answer, as we refine the discretization; and that the infinite dimensional answer captures the intended a priori knowledge.

[7]The wavelet coefficients of the parameter field are expected to be sparse.

4.2 RTO with Prior Transformations

We propose an alternative technique to sample from the posterior by means of a variable transformation. The transformation changes the L1-type prior defined on the physical parameter $\theta \in \mathbb{R}^n$ to a standard Gaussian defined on a reference parameter $u \in \mathbb{R}^n$. This transformation simplifies the prior and pushes the complexity to the likelihood term of the posterior. In particular, the transformed forward model is the original forward model composed with the nonlinear mapping function. After such a transformation, the posterior density is in least-squares form[8] and we are able to use RTO. In addition, when the forward model is differentiable, the posterior also remains differentiable and we can use gradient-based sampling techniques such as the Metropolis-Adjusted Langevin Algorithm (MALA).

Alg. 6 broadly sketches the procedure for sampling from a posterior on $\theta \in \mathbb{R}^n$ using RTO and the prior transformation.

Algorithm 6 RTO with prior transformation

1: Determine the prior mapping function $g : \mathbb{R}^n \to \mathbb{R}^n$ where $\theta = g(u)$.
2: Find the mode $\bar{u} \in \mathbb{R}^n$ of transformed posterior.
3: Calculate $\bar{Q} \in \mathbb{R}^{n \times m+n}$ in transformed posterior, whose columns form a basis of the tangent space at $\bar{u}$.
4: **for** $i = 1, \cdots, n_{\text{samps}}$ **do** in parallel
5: Sample a standard Gaussian for $\xi^{(i)}$
6: Optimize using RTO to get samples $\hat{u}^{(i)}$ distributed according to RTO's proposal defined on u.
7: **for** $i = 1, \cdots, n_{\text{samps}}$ **do** in series
8: Use independent Metropolis-Hastings to obtain $u^{(i)}$ distributed according to the posterior defined on u.
9: Transform samples back using $\theta^{(i)} = S(u^{(i)})$ to obtain samples distributed to the posterior defined on θ.

[8]The transformed posterior is in least-squares form with the additional assumption of a Gaussian noise.

The prior transformation described is very similar to the variable transformation presented in [12]. In the paper, a variable transformation is used to obtain a target distribution with super-exponentially light tails, so that sampling using random-walk Metropolis is geometrically ergodic. In a similar fashion, to perform a prior transformation, we use a variable transformation to obtain a posterior distribution in least-squares form so that RTO can be applied.

In the following two sections, we describe the prior transformation of one parameter $g_{1D}(\cdot) : \mathbb{R} \to \mathbb{R}$ for the Laplace prior, and the prior transformation of multiple parameters $g(\cdot) : \mathbb{R}^n \to \mathbb{R}^n$ for the L1-type prior, in more detail.

4.3 Prior Transformation of One Parameter with a Laplace Prior

Suppose we have a single observation and a single parameter in a Bayesian inference setting.

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

where $\theta \in \mathbb{R}$ is a random parameter, $f(\cdot) : \mathbb{R} \to \mathbb{R}$ is a nonlinear forward model, $\epsilon \in \mathbb{R}$ is an additive Gaussian noise, and $y \in \mathbb{R}$ is the observation. With the Laplace prior on θ,

$$p(\theta) \propto \exp\left(-\lambda|\theta|\right),$$

the posterior is

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

Due to the Laplace prior, the posterior density cannot directly be written in least-squares form and is also not differentiable. We can use gradient-free sampling techniques such as DRAM [11] or, if the forward model is linear, the Gibbs sampler for L1-type priors found in [15].

The alternative we propose is to transform the problem in order to change the prior to a standard Gaussian distribution and, as a result, change the posterior to

be in least-squares form and be differentiable. Let us define a one-to-one mapping function $g(\cdot) : \mathbb{R} \to \mathbb{R}$ that relates a reference, a priori Gaussian distributed random variable $u \in \mathbb{R}$ to the physical, a priori Laplace distributed parameter $\theta \in \mathbb{R}$:

$$\theta = g(u)$$

One invertible, monotone transformation that achieves this is

$$g(u) \equiv \mathcal{L}^{-1}(\varphi(u)) = -\frac{1}{\lambda}\text{sgn}\left(\varphi(u) - \frac{1}{2}\right)\log\left(1 - 2\left|\varphi(u) - \frac{1}{2}\right|\right) \tag{4.1}$$

where $\mathcal{L}(\cdot)$ is the cumulative distribution function (cdf) of the Laplace distribution and $\varphi(\cdot)$ is the cdf of the standard Gaussian distribution. To prove that the reference random variable is indeed the standard Gaussian, we find its cdf.

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

Hence, this mapping function indeed transforms a standard Gaussian reference random variable u to the Laplace distributed parameter θ, and thus,

$$p(u) \propto \exp\left(-\frac{1}{2}u^2\right).$$

Fig. 4-1 and Fig. 4-2 depicts the analytical mapping function and its derivative.

The transformation is a one-to-one analytical mapping between θ and u. We can perform Bayesian inference on u and transform the posterior samples of u to posterior samples of θ using the mapping. We now derive the posterior density on u.

In only the following few lines, we will change notation for clarity. Let $\pi_\Theta(w) \equiv p(\theta)|_{\theta=w}$ be the prior density on θ evaluated at $\theta = w$, and $\pi_U(w) \equiv p(u)|_{u=w}$ be the prior density on u evaluated at $u = w$, and so forth for the posterior densities. First, note that

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

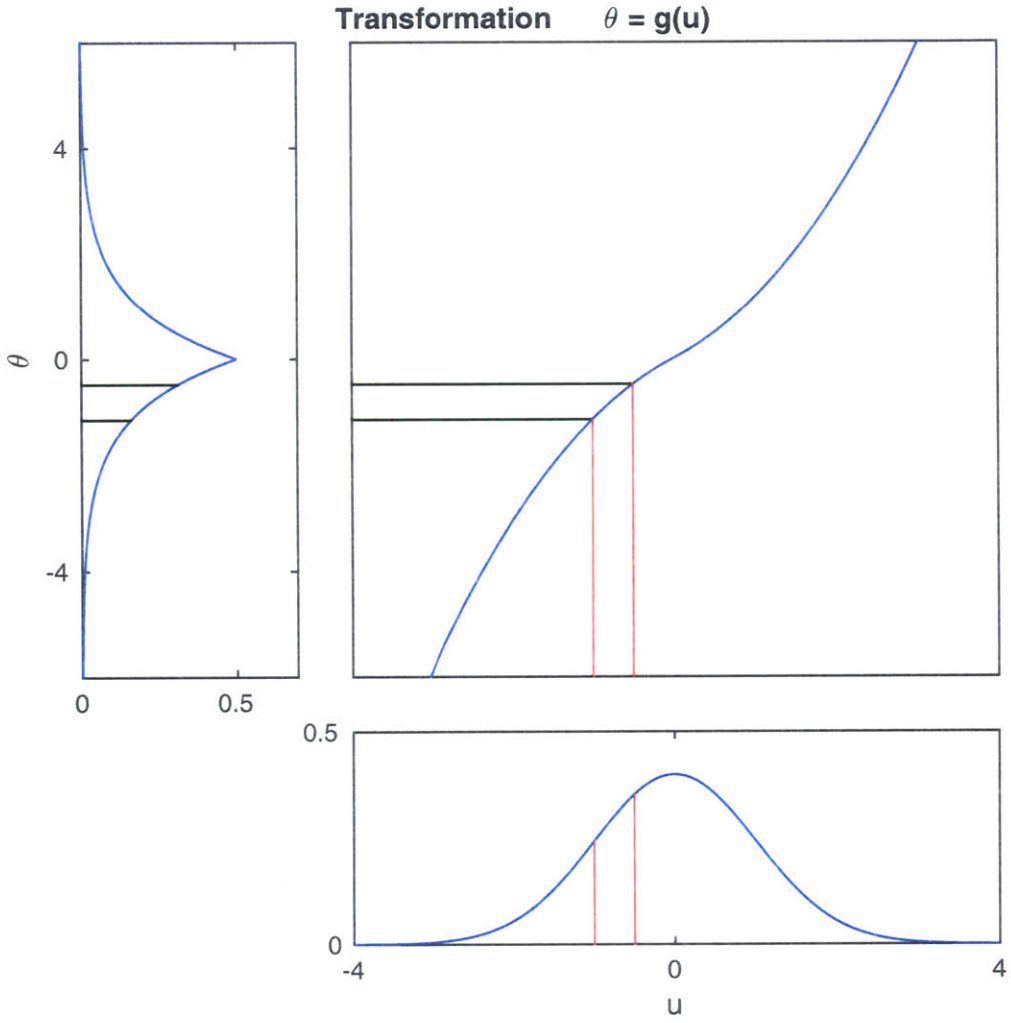

Figure 4-1: Transformation from the standard Gaussian to a Laplace distribution (with $\lambda = 1$). The probability mass between the two red lines is equal to that between the two green ones.

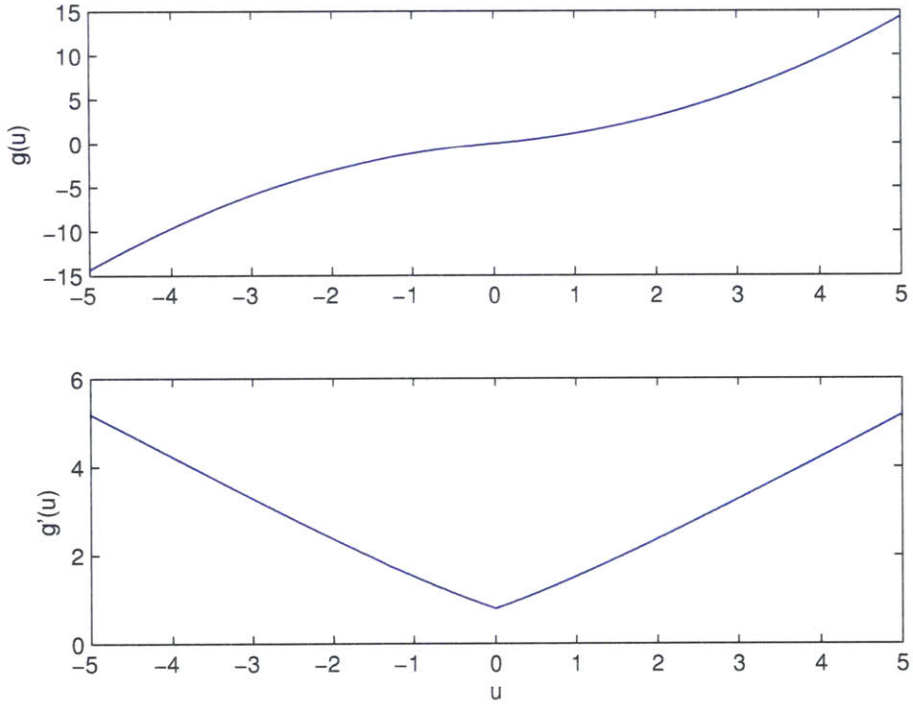

Figure 4-2: Mapping function $g_{1D}(\cdot)$ and its derivative. The mapping function is smooth and its derivative is continuous.

and thus

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

After the transformation, the prior over the new variables simplifies to a standard Gaussian, and the forward model becomes more complex. In particular, the transformed forward model is the original forward model composed with the nonlinear mapping. The mapping $g(\cdot)$ is differentiable. Its derivative is

$$g'(u) = \begin{cases} -\frac{\varphi'(u)}{\lambda\varphi(u)} & u \geq 0 \\ -\frac{\varphi'(u)}{\lambda(1-\varphi(u))} & u < 0 \end{cases}$$

where $\varphi'(u)$ is the pdf of the standard Gaussian distribution,

$$\varphi'(u) = \frac{1}{\sqrt{2\pi}}\exp\left(\frac{1}{2}u^2\right).$$

The mapping $g(\cdot)$ from standard Gaussian to Laplace is smooth and monotonic, and its derivative is continuous, as seen in Fig. 4-1 and Fig. 4-2. The resulting posterior is in least-squares form with a Gaussian prior and observational noise. The resulting structure allows us to use RTO. This transformation also changes the posterior to be differentiable and allows us to use derivative-based sampling techniques for Bayesian inference.

4.4 Prior Transformation of Multiple Parameters with an L1-type Prior

This section introduces the prior transformation for multiple parameters and multiple observations. Let n be the number of unknown parameters and m the number of observations.

We have multiple parameters of interest with the following vector of observations

$$y = f(\theta) + \epsilon \qquad \epsilon \sim N(0, \Gamma_{\text{obs}}),$$

where $\theta \in \mathbb{R}^n$ is a vector of parameters, $f(\cdot) : \mathbb{R}^n \to \mathbb{R}^m$ is a nonlinear differentiable forward model with a vector of outputs, $\epsilon \in \mathbb{R}^m$ is a vector of additive Gaussian noise with covariance matrix $\Gamma_{\text{obs}} \in \mathbb{R}^{m \times m}$, and $y \in \mathbb{R}^m$ is a vector of observations. We impose the following L1-type prior on θ

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

where $D \in \mathbb{R}^{n \times n}$ is an invertible matrix[9] and $(D\theta)_i$ denotes the ith element of vector $D\theta$. The posterior on θ is then

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

Each element of the vector $D\theta$ is a priori independent and identically distributed (i.i.d.) Laplace. Reference random variables that are *a priori* i.i.d. Gaussian can be transformed to each Laplace-distributed element of $D\theta$ using the 1-D transformation $g_{\text{1D}}(\cdot)$ defined by Eq. 4.1. Using such a transformation, the posterior over the new

[9]TV priors (with appropriate boundary conditions) in one dimension and $B^s_{1,1}$ Besov space priors can both be cast into the L1-type prior form shown with an invertible matrix D.

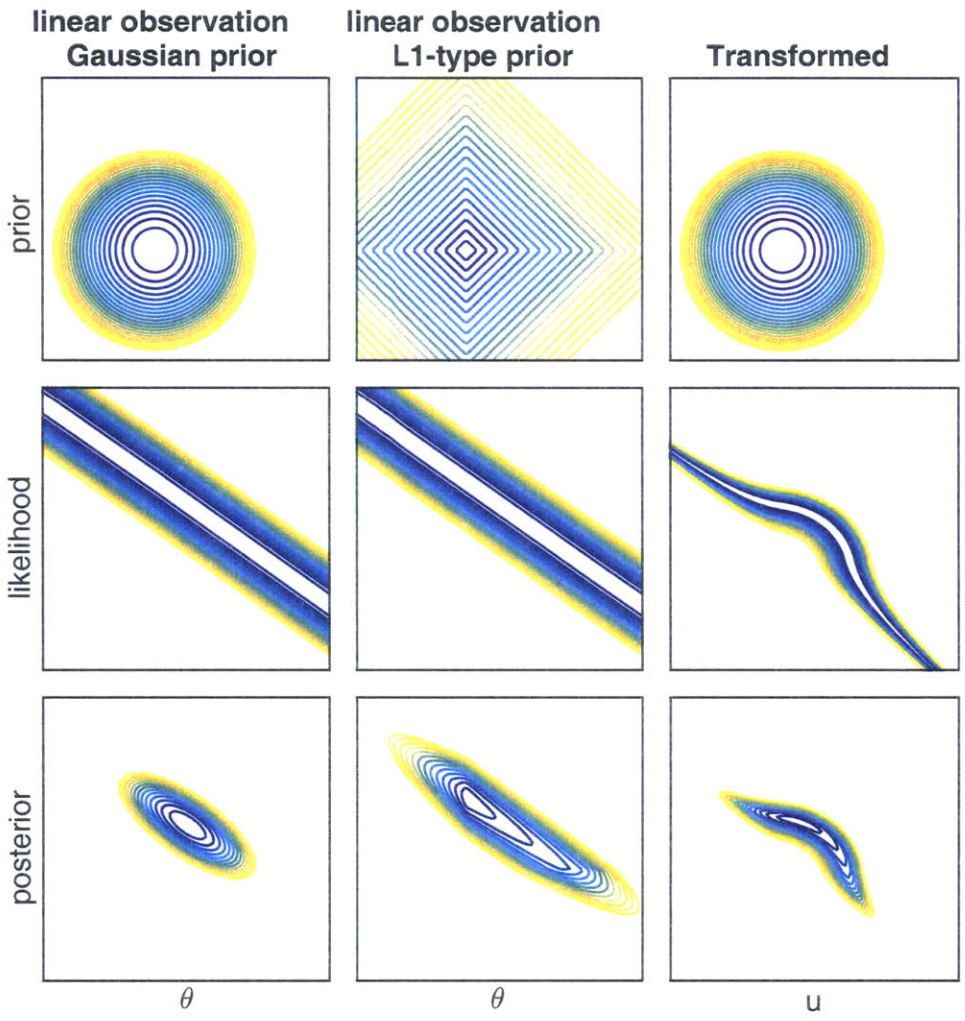

Figure 4-3: Posteriors from Bayesian inference of two parameters. The plots depict the log-prior, log-likelihood and log-posterior. The three cases shown are a Gaussian prior with linear likelihood (left), Laplace prior with the same likelihood (middle), and transformed Laplace prior with transformed likelihood (right). The transformation changes the prior to a Gaussian and makes the likelihood more complex. The posterior on the right is smooth and can be written in least-squares form.

variables is then

$$p(u|y) \propto \exp\left[-\frac{1}{2}\big(f\big(D^{-1}g(u)\big)-y\big)^T\Gamma_{\text{obs}}^{-1}\big(f\big(D^{-1}g(u)\big)-y\big)\right]\exp\left(-\frac{1}{2}u^Tu\right)$$
$$= \exp\left[-\frac{1}{2}\Big(\tilde{f}(u)-y\Big)^T\Gamma_{\text{obs}}^{-1}\Big(\tilde{f}(u)-y\Big)\right]\exp\left(-\frac{1}{2}u^Tu\right),$$

where $u \in \mathbb{R}^n$ is a vector of new variables, $g(\cdot) : \mathbb{R}^n \to \mathbb{R}^n$ is a vector valued function that transforms each element of u using the Gaussian-to-Laplace transformation, and $\tilde{f}(u) \equiv f(D^{-1}g(u))$ is effectively the *transformed* forward model. Samples of u can be transformed back to samples of θ by

$$\theta = D^{-1}g(u).$$

The transformation changes the prior to a standard multivariate Gaussian and makes the likelihood term more complex. In particular, a linear forward model will have a nonlinear likelihood due to the transformation. See Fig. 4-3.

In addition to having a posterior in least-squares form, a second benefit of solving the inverse problem on the transformed variables u rather than the original parameters θ is that the posterior has a continuous gradient, provided that the original non-linear forward model is C_1 continuous. This allows us to use sampling algorithms that require derivative information, such as MALA.

To perform the optimization steps in RTO and to evaluate the proposal density of RTO, we need the Jacobian of $\tilde{f} \in \mathbb{R}^n \to \mathbb{R}^m$ with respect to u. Let us denote this Jacobian, evaluated at u_*, by $J_{\tilde{f}}(u_*)$, where $J_{\tilde{f}}(\cdot) : \mathbb{R}^n \to \mathbb{R}^{m\times n}$. By the chain rule, this Jacobian is

$$J_{\tilde{f}}(u_*) = J_f\big(D^{-1}g(u_*)\big)D^{-1}J_g(u_*)$$

where $J_f(\cdot) : \mathbb{R}^n \to \mathbb{R}^{m\times n}$ is the Jacobian of the original forward model and $J_g(\cdot) : \mathbb{R}^n \to \mathbb{R}^{n\times n}$ is the Jacobian of g, which is a diagonal matrix that contains the

derivatives of the scalar function g_{1D} defined by Eq. 4.1.

$$J_g(u_*) = \begin{bmatrix} g'_{1D}(u_{*,1}) & & & \\ & g'_{1D}(u_{*,2}) & & \\ & & \ddots & \\ & & & g'_{1D}(u_{*,n}) \end{bmatrix}$$

4.5 Validity of RTO with Linear Forward Models and Prior Transformations

In this section, we explore the cases when RTO is valid after a prior transformation. Suppose the original forward model is linear. Let $f(\theta) = A\theta$, then $\tilde{f}(u) = f(D^{-1}g(u)) = AD^{-1}g(u)$ and the posterior on u is

$$\begin{aligned} p(u|y) &\propto \exp\left[-\frac{1}{2}(AD^{-1}g(u) - y)^T\Gamma_{\text{obs}}^{-1}(AD^{-1}g(u) - y)\right]\exp\left(-\frac{1}{2}u^Tu\right) \\ &= \exp\left(-\frac{1}{2}\left\|\begin{bmatrix} u \\ \Gamma_{\text{obs}}^{-1/2}AD^{-1}g(u)\end{bmatrix} - \begin{bmatrix} 0 \\ \Gamma_{\text{obs}}^{-1/2}y\end{bmatrix}\right\|^2\right) \\ &= \exp\left(-\frac{1}{2}\|F(u) - Y\|^2\right) \end{aligned}$$

where

$$F(u) = \begin{bmatrix} u \\ \Gamma_{\text{obs}}^{-1/2}AD^{-1}g(u)\end{bmatrix} \qquad Y = \begin{bmatrix} 0 \\ \Gamma_{\text{obs}}^{-1/2}y\end{bmatrix} \qquad J(u) = \begin{bmatrix} I \\ \Gamma_{\text{obs}}^{-1/2}AD^{-1}J_g(u)\end{bmatrix}$$

From Sec. 2.3, RTO samples from the true posterior after independence M-H, when the following two conditions hold

(i) the matrix $\bar{Q}^TJ(u)$ is invertible for all u in the domain of F.

(ii) there is a u such that $\bar{Q}^T(F(u) - (Y + \Xi)) = 0$ for any Ξ.

For the prior transformation defined in Sec. 4.4, we show that (i) is true and leave (ii)

as an assumption. For any $u_1 \in \mathbb{R}^n$ and $u_2 \in \mathbb{R}^n$,

$$\begin{aligned}
\det\left[J(u_1)^T J(u_2)\right] &= \det\left(\begin{bmatrix} I \\ \Gamma_{\text{obs}}^{-1/2} A D^{-1} J_g(u_1) \end{bmatrix}^T \begin{bmatrix} I \\ \Gamma_{\text{obs}}^{-1/2} A D^{-1} J_g(u_2) \end{bmatrix}\right) \\
&= \det\left[I + J_g(u_1) D^{-T} A^T \Gamma_{\text{obs}} A D^{-1} J_g(u_2)\right] \\
&= \det J_g(u_1) \det\left[J_g(u_1)^{-1} J_g(u_2)^{-1} + D^{-T} A^T \Gamma_{\text{obs}} A D^{-1}\right] \det J_g(u_2).
\end{aligned}$$

$J_g(u)$ is a positive diagonal matrix for any u, and $D^{-T} A^T \Gamma_{\text{obs}} A D^{-1}$ is symmetric positive semi-definite. Then, $J_g(u_1)^{-1} J_g(u_2)^{-1} + D^{-T} A^T \Gamma_{\text{obs}} A D^{-1}$ is symmetric positive definite and the final expression in the equation above is positive. Thus, $\det\left[J(u_1)^T J(u_2)\right] > 0$ and $J(u_1)^T J(u_2)$ is invertible.

$\bar{Q}$ is from the compact QR-decomposition of $J(\bar{u})$, where $\bar{u}$ is the posterior mode on u. It follows that $J(u)^T \bar{Q} = J(u)^T \bar{Q} \bar{R} \bar{R}^{-1} = J(u)^T J(\bar{u}) \bar{R}^{-1}$ is invertible for any $u \in \mathbb{R}^n$. This proves condition (i) holds.

4.6 Numerical Example: Deconvolution

We use a deconvolution of a square signal as a numerical experiment. In this example, a square pulse, discretized on 128 grid points, is convolved with a Gaussian kernel to obtain a blurred signal. 32 noisy measurements of the convolved signal are used as observations. The inverse problem is to reconstruct the original signal from the noisy and blurred measurements. Fig. 4-4 depicts the true signal, the convolved signal and the measurements. A TV prior is used.

$$\pi(\theta) \propto \exp\left(-\lambda |D\theta|\right)$$

where $\theta \in \mathbb{R}^{128}$, $D \in \mathbb{R}^{128\times128}$, $\lambda = 1$ and

$$D = \begin{bmatrix} 1 & & & \\ -1 & 1 & & \\ & \ddots & \ddots & \\ & & -1 & 1 \end{bmatrix}.$$

Note that here, we fix the left boundary (the first row in matrix D) to make D invertible, which is needed to define a one-to-one transformation of an independently Laplace distributed random variable to the parameter value on the grid.

Posterior samples are obtained using over-relaxed Gibbs[10] [15], MALA[11] and RTO. In MALA and RTO, we use a prior transformation to redefine the problem on reference parameters described in the previous section.

[10]Over-relaxed Gibbs uses the code provided by the original author of [15] with a systematic scan and a relaxation parameter of 7.

[11]MALA uses the Hessian at the posterior mode for preconditioning.

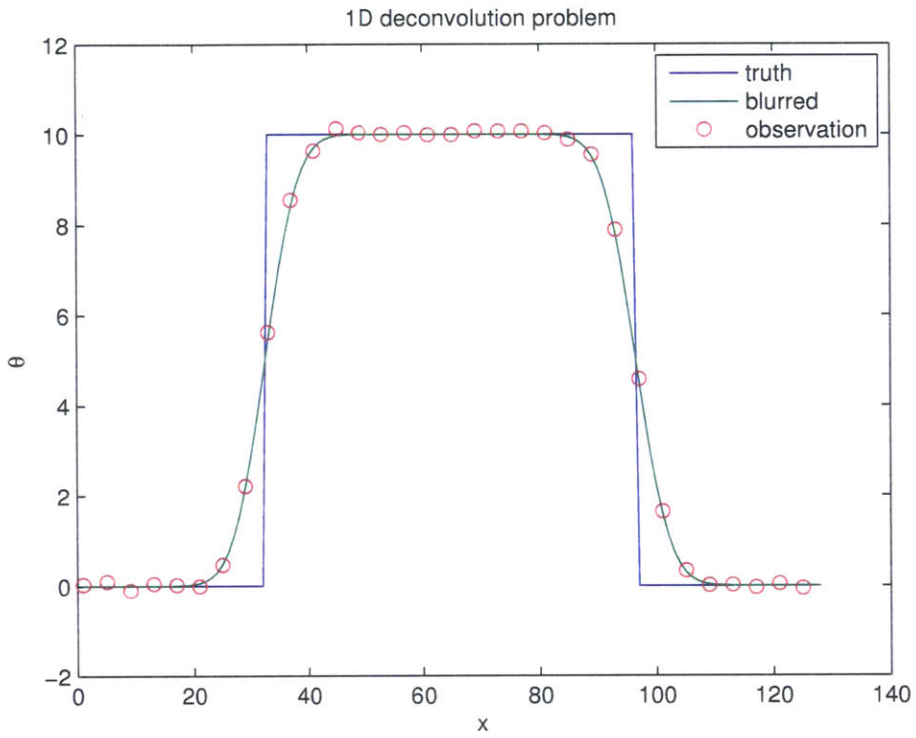

Figure 4-4: 1D deconvolution problem. The unknown true signal (blue), the signal after convolution (green) and noisy observations (red). The inverse problem is to determine the true signal from the noisy observations.

As shown in Fig. 4-5, 4-6 and 4-7, the posterior distribution using the three samplers match. The posterior mean, covariance, and marginals agree. This is the expected behavior, as all algorithms should sample from the true posterior as the number of samples obtained approaches infinity.

A visual comparison of the sampling efficiency of the three algorithms can be done by examining the autocorrelation decay as a function of the number of forward model evaluations. For this problem, RTO took an average of 7.4 function and derivative evaluations to generate a new point in its MCMC chain, and Gibbs requires an update in each dimension with 128 function evaluations for each new point in its MCMC chain. Fig. 4-8 shows the autocorrelation over function evaluations for the three methods. Table 4.1 lists the number of function and derivative evaluations required per effective sample size for the three algorithms. The sampling efficiency of Gibbs and MALA are comparable, whereas RTO is four orders of magnitude more efficient.

Table 4.1: Function evaluations per effective sample size for the three sampling algorithms.

Method	Function/derivative evaluations per ESS
Gibbs	2 518 000
prior transformation + MALA	1 886 000
prior transformation + RTO	164

This numerical example demonstrates that RTO, with a prior transformation, can be an extremely efficient algorithm for problems on which other state-of-the-art sampling algorithms struggle.

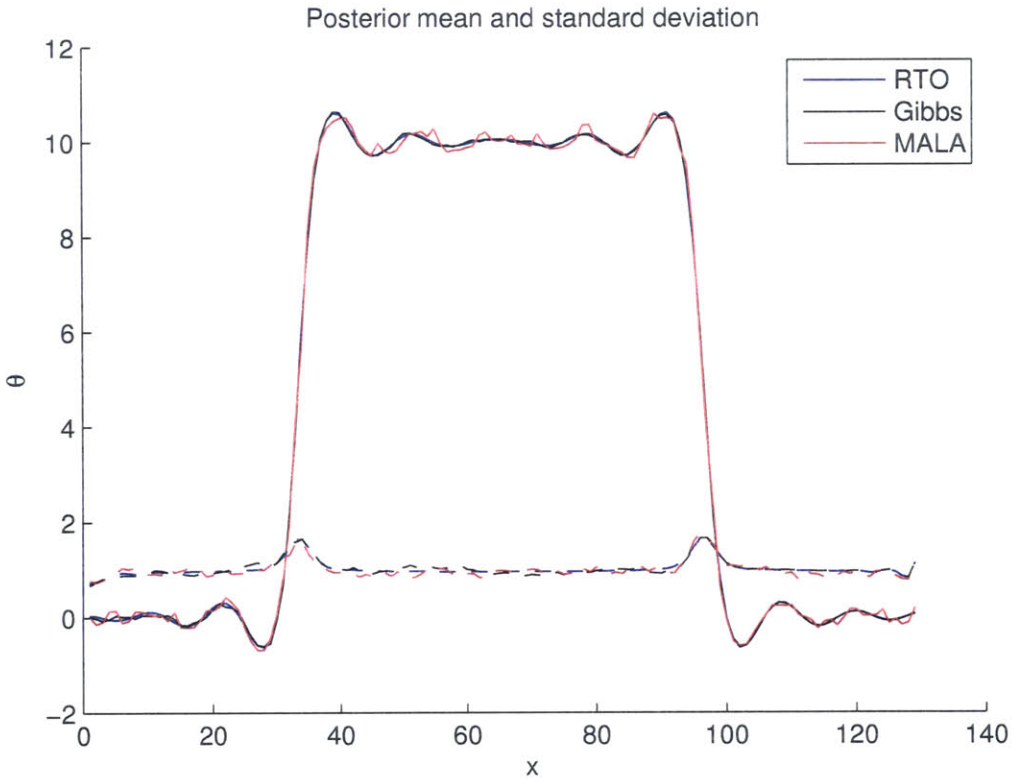

Figure 4-5: Posterior mean and standard deviation. The posterior mean and standard deviation from all three sampling techniques match.

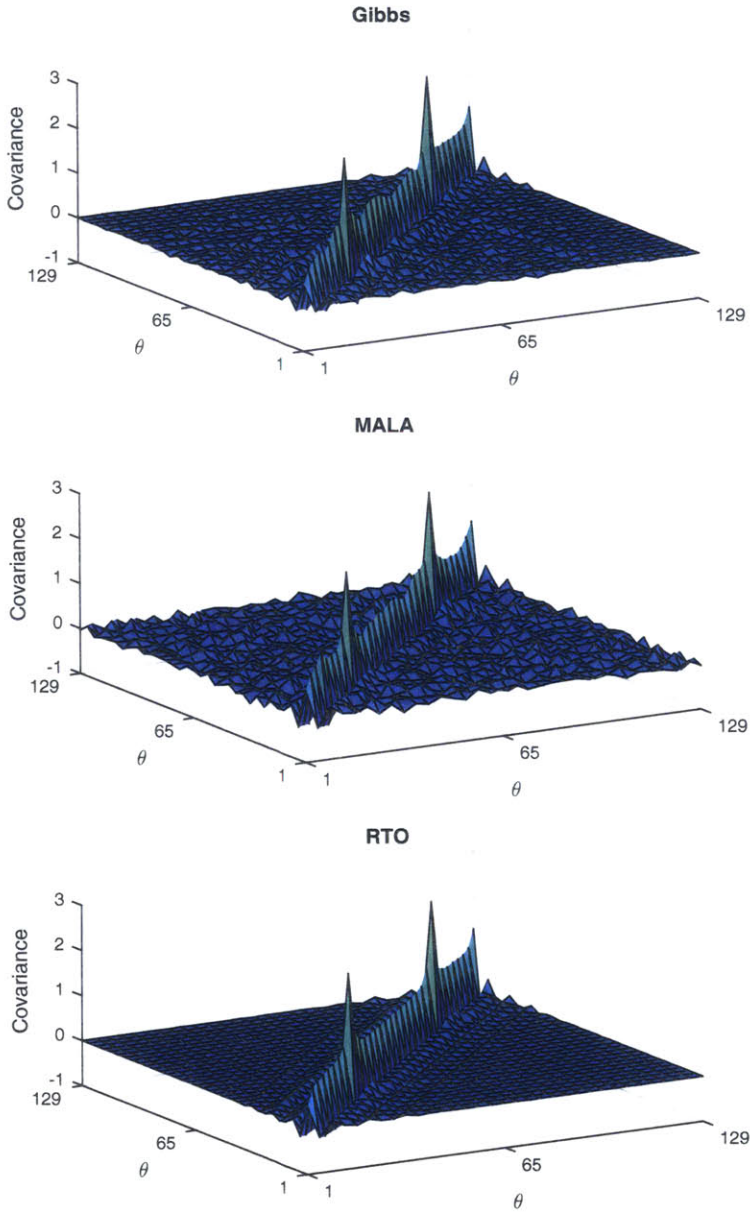

Figure 4-6: Posterior covariance over parameters from the three sampling techniques. Gibbs (top), MALA (middle), RTO (bottom). The posterior covariances from all three MCMC chains agree.

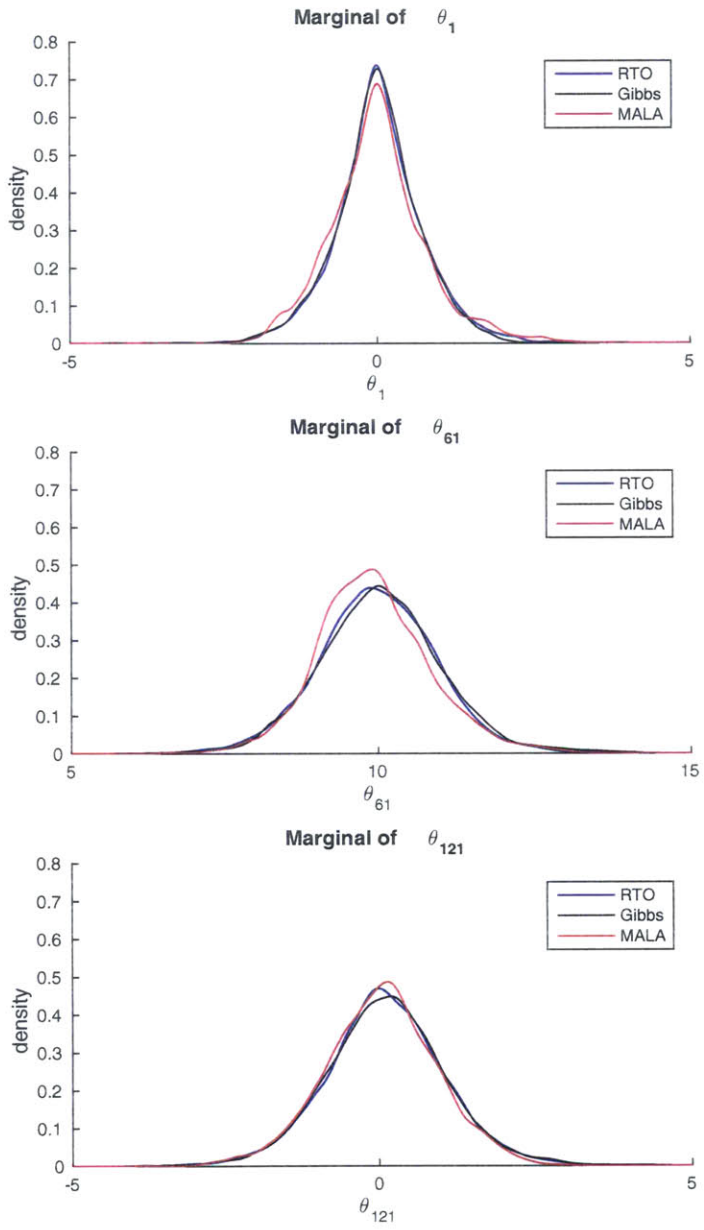

Figure 4-7: Posterior marginals of the 1^{st}, 61^{st} and 121^{st} parameters. All sampling techniques show similar marginals.

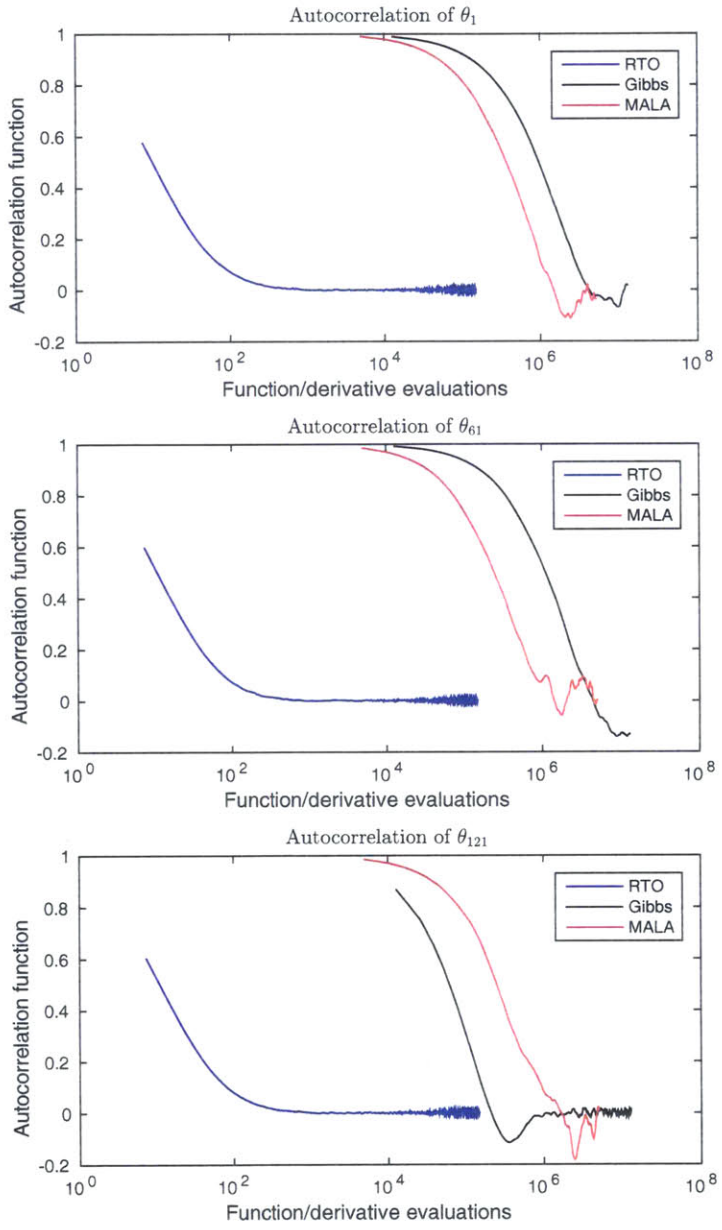

Figure 4-8: Autocorrelation over function evaluations of the 1^{st}, 61^{st} and 121^{st} parameters using the three sampling techniques. RTO requires requires significantly fewer function evaluations to reduce its autocorrelation.

4.7 Concluding Remarks

In this chapter, we demonstrate that RTO can be applied to non-Gaussian priors. We derive a prior transformation for L1-type priors and outline the steps to use the prior transformation with RTO. In the numerical example, RTO with this transformation is substantially more efficient than other sampling techniques.

For other types of priors, we would need to define its application-specific mapping function. A few general frameworks that may help are: the Rosenblatt transformation [22] and optimal transport maps [18] [19].

The sampling efficiency for the different algorithms is problem-dependent. In cases where the Jacobian of the forward model can not be computed cheaply, RTO may need to use finite differences to determine its Jacobian, and would require more function evaluations for each new point in its chain.

Chapter 5

Conclusions and Future Work

This thesis presents three main contributions that improve the performance and the understanding of the randomize-then-optimize (RTO) algorithm. The geometric interpretation helps visualize and understand RTO. It connects RTO to two other MCMC methods in literature. The interpretation is also used in constructing an adaptive version of RTO that is demonstrated to be more robust and efficient than the original. Finally, a prior transformation technique is presented that allows RTO to be applied to a broader range of inverse problems. Using the transformation in combination with RTO results in a method shown to be several orders of magnitude more efficient than other state-of-the-art algorithms when applied to a challenging problem.

As optimization routines and differentiation techniques[1] are improved, the computational cost for RTO will decrease. In addition, high performance computing with distributed processing promotes algorithms that scale well in parallel, which is true of RTO. Thus, with future developments in mind, this relatively new algorithm has the potential to become one of the go-to sampling tools for Bayesian inference. An understanding of RTO and other optimization-based sampling algorithms may become essential within the uncertainty quantification academic community.

With that said, there are many avenues for future development. A few are listed

[1]Adjoints in finite element solvers is one example where the derivative is obtained as cheaply as one simulation.

below.

- Building upon the connection to optimal transport maps, we may be able to ascertain the conditions under which sampling from RTO will be produce a uniformly ergodic Markov chain. In particular, [16] discusses conditions under which Metropolis-Hastings converges geometrically to the true posterior.

- In terms of implementation and performance, the proposal sample generating procedure in RTO requires repeated optimization of a cost function that depends directly on the forward model. There is a possibility that information from previous optimizations can be stored and used to help future optimizations. This would reduce the computational cost of RTO.

- Multi-modal posteriors remain difficult to sample from using a single RTO-like proposal. Mixtures of RTO-like proposals may be used to tackle these particularly challenging distributions.

- Extending RTO sampling to very high dimensions remain a challenge. Incorporating dimension-reduction techniques such as likelihood informed subspaces [6] may be helpful.

There appears to be much work remaining between the current state of the RTO sampling algorithm for Bayesian inference and its full potential. This thesis is a small, but sure step forward in improving its efficiency and applicability.

Bibliography

[1] Johnathan M. Bardsley and Aaron Luttman. Total variation-penalized Poisson likelihood estimation for ill-posed problems. *Advances in Computational Mathematics*, 31:35–59, October 2009.

[2] Johnathan M. Bardsley, Antti Solonen, Heikki Haario, and Marko Laine. Randomize-then-optimize: a method for sampling from posterior distributions in nonlinear inverse problems. *SIAM Journal on Scientific Computing*, 36(4):A1895–A1910, August 2014.

[3] Nicolas Boumal, Bamdev Mishra, P.-A. Absil, and Rodolphe Sepulchre. Manopt, a matlab toolbox for optimization on manifolds. *Journal of Machine Learning Research*, 15:1455–1459, 2014.

[4] Steve Brooks, Andrew Gelman, Galin L. Jones, and Xiao-Li Meng. *Handbook of Markov Chain Monte Carlo*. Handbooks of Modern Statistical Methods. Chapman & Hall/CRC, 2004.

[5] Alexandre J. Chorin, Matthias Morzfeld, and Xuemin Tu. Implicit particle filters for data assimilation. *Communications in Applied Mathematics and Computational Science*, 5(2), May 2010.

[6] Tiangang Cui, James Martin, Youssef M. Marzouk, Antti Solonen, and Alessio Spantini. Likelihood-informed dimension reduction for nonlinear inverse problems. *Inverse Problems*, 30:114015, 2014.

[7] Masoumeh Dashti, Stephen Harris, and Andrew Stuart. Besov priors for Bayesian inverse problems. *Inverse Problems and Imaging*, 6(2):183–200, 2012.

[8] Alan Edelman, T. A. Arias, and Steven T. Smith. The geometry of algorithms with orthogonality constraints. *SIAM Journal of Matrix Analysis and Applications*, 20(2):303 –353, 1998.

[9] W. R. Gilks, S. Richardson, and D. J. Spiegelhalter. *Markov Chain Monte Carlo in Practice*. Chapman & Hall/CRC, 1996.

[10] Jonathan Goodman, Kevin K. Lin, and Matthias Morzfeld. Small-noise analysis and symmetrization of implicit Monte Carlo samplers. *ArXiv e-prints*, October 2014. arXiv:1410.6151 [math.NA].

[11] Heikki Haario, Marko Laine, Antonietta Mira, and Eero Saksman. DRAM: Efficient adaptive MCMC. *Statistics and Computing*, 16:339–354, 2006.

[12] Leif T. Johnson and Charles J. Geyer. Variable transformation to obtain geometric ergodicity in the random-walk Metropolis algorithm. *The Annals of Statistics*, 40(6):3050 – 3076, 2012.

[13] Ruben Juanes, Bradford H. Hager, Thomas A. Herring, and Youssef M. Marzouk. Coupled flow and reservior geomechanics: From data to decisions. Technical report, MIT, eni E&P Division, January 2013. A collaborative eni-MIT Project.

[14] Jari Kaipio and Erkki Somersalo. *Statistical and Computational Inverse Problems.* Applied Mathematical Sciences. Springer, 2005.

[15] Felix Lucka. Fast Markov chain Monte Carlo sampling for sparse Bayesian inference in high-dimensional inverse problems using l1-type priors. *Inverse Problems*, 28:125012, September 2012.

[16] K. L. Mengersen and R. L. Tweedie. Rates of convergence of Hastings and Metropolis algorithms. *The Annals of Statistics*, 24(1):101 – 121, 1996.

[17] Matthias Morzfeld, Xuemin Tu, Ethan Atkins, and Alexandre J. Chorin. A random map implementation of implicit filters. *Journal of Computational Physics*, 231:2049–2066, November 2011.

[18] Tarek A. Moselhy and Youssef M. Marzouk. Bayesian inference with optimal maps. *Journal of Computational Physics*, 231:7815–7850, August 2012.

[19] Matthew Parno and Youssef Marzouk. Transport map accelerated markov chain monte carlo. *ArXiv e-prints*, December 2014. arXiv:1412.5492 [stat.CO].

[20] Christian P. Robert and George Casella. *Monte Carlo Statistical Methods Second Edition.* Springer Texts in Statistics. Springer, 2 edition, 2004.

[21] Gareth O. Roberts and Richard L. Tweedie. Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli*, 2(4):341–363, December 1996.

[22] Murray Rosenblatt. Remarks on a multivariate transformation. *The Annals of Mathematical Statistics*, 23(3):470–472, September 1952.

[23] Leonid I. Rudin, Stanley Osher, and Emad Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D*, 60:259–268, 1992.

[24] Andrew M. Stuart. Inverse problems: a Bayesian perspective. *Acta Numerica*, 19:451 – 559, May 2010.