Commun. Optim. Theory 2023 (2023) 18 https://doi.org/10.23952/cot.2023.18

MATHRES Communications in Optimization Theory
Available online at http://cot.mathres.org

ADAPTIVE PARTICLE-BASED APPROXIMATIONS OF THE GIBBS POSTERIOR FOR INVERSE PROBLEMS

ZILONG ZOU[1], SAYAN MUKHERJEE[1], HARBIR ANTIL[2,*], WILKINS AQUINO[1]

[1]Department of Civil and Environmental Engineering, Duke University, NC, USA
[2]Center for Mathematics and Artificial Intelligence and Department of Mathematical Sciences, George Mason University, VA, USA

This article is dedicated to the memory of our teacher and mentor Prof. Roland Glowinski

Abstract. In this paper, we adopt a general framework based on the Gibbs posterior to update belief distributions for inverse problems governed by partial differential equations (PDEs). The Gibbs posterior formulation is a generalization of standard Bayesian inference that only relies on a loss function connecting the unknown parameters to the data. It is particularly useful when the true data generating mechanism (or noise distribution) is unknown or difficult to specify. The Gibbs posterior coincides with Bayesian updating when a true likelihood function is known and the loss function corresponds to the negative log-likelihood, yet provides subjective inference in more general settings. We employ a sequential Monte Carlo (SMC) approach to approximate the Gibbs posterior using particles. To manage the computational cost of propagating increasing numbers of particles through the loss function, we employ a recently developed local reduced basis method to build an efficient surrogate loss function that is used in the Gibbs update formula in place of the true loss. We derive error bounds for our approximation and propose an adaptive approach to construct the surrogate model in an efficient manner. We demonstrate the efficiency of our approach through several numerical examples.

Keywords. Adaptive Sequential Monte Carlo; Bayesian framework; Error analysis; Gibbs posterior; Inverse problems with PDEs.

2020 Mathematics Subject Classification. 49J20, 35J05, 78M10, 65N20, 65N30.

1. INTRODUCTION

In stochastic inverse problems, we need to infer unknown system parameters from uncertain measurements of a system response. Such problems are ubiquitous in many application areas including medical imaging [15, 22], heat conduction [33], geosciences [6], atmospheric and oceanic sciences [2]. The Bayesian approach has been a foundation for performing such inference from noisy or incomplete observations while allowing us to quantify the uncertainty

*Corresponding author.
E-mail address: hantil@gmu.edu (H. Antil).
Received October 7, 2022; Accepted December 29, 2022.

of the inverse solution due to the inexactness of data [10, 31]. The solution of a Bayesian inverse problem is a probability distribution over the parameter space, which is referred to as the posterior distribution. Except for very limited settings, e.g., a linear model with Gaussian prior and Gaussian noise, the analytical form of the posterior distribution is rarely tractable. In most cases, we can only approximate the posterior either through sampling or parametrization.

The main contribution of this work is a framework to approximate the solution of stochastic inverse problems that involve solutions of PDEs or expensive numerical simulations without the need to construct explicit likelihood functions. To this end, the main computational advancement in our work is an adaptive reduced basis method that minimizes the number of expensive PDE evaluations. In addition, *a posteriori* error estimates are used to efficiently explore the posterior distribution, while certifying accuracy. Furthermore, we postulate the inverse problem in a general variational statement that relies on the use of a loss function instead of likelihood function. The use a loss function is a more natural approach to many inverse problems where the noise generating mechanism is not known.

Markov chain Monte Carlo (MCMC) [17] is the most well-known and versatile method for Bayesian inverse problems [31]. MCMC requires only pointwise evaluation of the likelihood function to generate a stream of samples from the posterior distribution that can be subsequently used to compute the statistics of the posterior. To accelerate MCMC, one popular approach is to employ an inexpensive surrogate model to approximate the likelihood evaluation in the sampling procedure. Plenty of research efforts have been devoted to construct efficient surrogate models for such a purpose. For example, the stochastic spectral method is used in [28], Gaussian process regression is used in [21] and projection-based model reduction is employed in [12, 25]. The surrogate models are typically constructed to be accurate over the support of the prior distribution [16, 25, 26, 27, 28] and are thought to be "globally accurate". However, thanks to the information contained in the data, the posterior distribution typically concentrate on a much smaller portion of the support of the prior. In this respect, requiring a "globally accurate" surrogate model seems unnecessary and inefficient. Several recent studies have exploited such information (or posterior) and build adaptive and data-driven surrogate models that are more efficient and accurate on the support of the posterior [8, 9, 12, 24].

Recently, sequential Monte Carlo (SMC) methods, or particle filters [3, 13, 14], have been applied in the setting of Bayesian inverse problems by a few researchers [4, 20]. In SMC, weighted samples, or particles, are generated and evolved to approximate a sequence of probability distributions which interpolate from the prior to the posterior. In [20] in particular, the authors employed a novel SMC method with a dimension-independent MCMC sampler [11, 23] as the mutation kernel to invert the initial conditions for Navier-Stokes equations. In [4], the authors enhanced the SMC method in [20] and provided a proof of the dimension-independent convergence property of the SMC methods for inverse problems. Both works have demonstrated the computational efficiency of SMC methods for high-dimensional inverse problems. In addition, the versatility and self-adaptivity of SMC methods provide a natural framework for the adaptive construction of surrogate models that can be used to further speedup the computations for inverse problems.

Taking inspiration in the latter contributions, in this work we put forward an adaptive reduced bases approach with guaranteed accuracy that efficiently explores the support of the posterior.

Based on a SMC framework developed in [20], we present a method to progressive approximate the posterior by simultaneously evolving the particles and adapting the local RB surrogate model in a sequential manner. The emphasis of the local RB surrogate is navigated to a small fraction of the parameter space, i.e., the support of the posterior, automatically by the evolving particles that progressively cluster over the support of the posterior. Computational savings are achieved thanks to the local accuracy and the efficiency of the local RB method [35]. Indeed, once the local RB surrogate becomes accurate enough over the local support of the posterior, further evolution of the particles takes minimal cost. In addition, we derive error bounds for our approximation and demonstrate the computational efficiency of our approach through several numerical examples including advection-diffusion problems and elasticity imaging problems.

The majority of Bayesian methods for inverse problems rely on an exact noise model, typically assumed to be i.i.d. Gaussian, to perform inference. It is desirable to extend such inference to more general settings where a noise model is unavailable or modeling the data generating mechanism is challenging. The Gibbs posterior provides a way to update belief distributions in such general setting without the need of an explicit likelihood function. Instead, the Gibbs posterior are applicable where the unknown parameters are only connected to the data through a loss function [1, 5, 32]. In many inverse problems or inference problems, it can be a simpler task to specify a loss function than the true data generating mechanism, i.e., an explicit likelihood function, which is the biggest advantage of using the Gibbs posterior over the usual Bayesian approach.

The rest of the paper is organized as follows. We first present the problem statement and develop a Gibbs posterior formulation. We then describe our adaptive reduced basis approach for reducing computational cost. Next, we describe how the reduced bases method can be combined with a SMC approach. In addition, we support all our numerical approximations with error bounds. We end the paper with a set of numerical examples and conclusions.

2. Problem Statement

Consider the abstract variational problem: find $u(\xi) : \Xi \to U$ such that

$$\langle M(u(\xi);\xi), v\rangle_{V^*,V} = 0 \quad \forall v \in V, \quad \forall \xi \in \Xi.$$

where U is the trial space, V is the test space, V^* is the dual space of V, $\Xi \subseteq \mathbb{R}^M$ is the parameter space, and $M(\cdot;\xi) : U \to V^*$ is a bounded linear operator for all $\xi \in \Xi$. We will use $U \equiv V$ in the sequel. For the sake of simplicity, we will work with finite dimensional spaces arising from the discretization of an underlying PDE model.

We assume that we have access to imperfect observations $d \in \mathbb{R}^D$ from the system. Furthermore, we describe these observations as

$$d = \mathscr{F}(\xi^*) + \varepsilon, \tag{2.1}$$

where $\mathscr{F} : \Xi \to \mathbb{R}^D$ is a model representing the system that maps each parameter to an observation, and $\varepsilon \in \mathbb{R}^D$ is a random element representing noise. For instance, $\mathscr{F}(\xi^*) := \mathscr{G}(u(\xi^*))$, where $\mathscr{G}$ is a map (nonlinear in general) from PDE solutions to observables. We also define $d^* = \mathscr{F}(\xi^*)$ as the true data. Notice that in this setting, even if we have an additive noise decomposition in the state, i.e. $v = u(\xi^*) + \eta$, $d = \mathscr{G}(v(\xi))$, the nonlinear map makes an analytical representation of ε intractable, in general. Here, η represents noise in the state.

A salient aspect of our setting is that we do not assume any knowledge of the probability law of $\varepsilon \in \mathbb{R}^D$. However, we do assume that we have a prior belief about ξ^*, which can be expressed as a prior distribution $\rho_0(\cdot) : \Xi \to \mathbb{R}$. Since we do know the probability law of ε, and hence the likelihood, one of the main challenges in our work is how to integrate the information contained in the data into our belief about ξ^*. To this end, we turn to the variational framework that circumvents this challenge.

In the variational setting presented in [5], we do not need a likelihood function. Instead, we use a loss function $l(\cdot,\cdot) : \Xi \times \mathbb{R}^D \to \mathbb{R}$ that measures the discrepancy between the prediction and observations. For example, we could use

$$l(\xi, d) = \|\mathscr{F}(\xi) - d\|_{l_2}^2$$

as the loss function. Unlike a likelihood function that requires exact knowledge of the data generating mechanism (or noise model), loss functions are typically easier to specify for inverse problems.

Given a set of observations d_i, $i = 1, 2, \ldots, n$, we update our belief according to the following optimization problem

$$\rho(\xi) = \underset{\hat{\rho} \in P}{\operatorname{argmin}} \int_\Xi W \sum_{i=1}^n l(\xi, d_i)\hat{\rho}(\xi) d\xi + D_{KL}(\hat{\rho} \| \rho_0). \tag{2.2}$$

where W is a weight for the loss that is yet to be specified. For now, we assume W is a fixed positive constant and will describe possible methods to prescribe W later on. Furthermore, $D_{KL}(\rho\|\rho_0)$ is the Kullback-Leibler (KL) divergence between the posterior and prior distributions and P is the space of candidate posterior distributions of ξ. If we allow P to contain all distributions over Ξ, we have an explicit update formula for $\rho(\xi)$ [5] as

$$\rho(\xi) = \frac{\exp(-W \sum_{i=1}^n l(\xi, d_i))\rho_0(\xi)}{\int_\Xi \exp(-W \sum_{i=1}^n l(\xi, d_i))\rho_0(\xi) d\xi}. \tag{2.3}$$

This is a coherent update formula in the sense that the use of sequential data in a sequential manner yields the same distribution as if we used all the data simultaneously as in (2.3). In addition, we can see from (2.2) that the Gibbs posterior minimizes the expected loss with an added requirement that this posterior be close to the prior in the sense of the KL divergence. Also, notice that W weights our relative belief between the information provided by the data versus the information in the prior.

Also, we can see that the usual Bayes rule is a special case of (2.3) by using the negative log-likelihood as the loss function with $W = 1$. Indeed, if we use $l(\xi, d_i) = -\log(\pi(d_i|\xi))$ where $\pi(d_i|\xi)$ is the likelihood function, we get

$$\rho(\xi) = \frac{\prod_{i=1}^n \pi(d_i|\xi)\rho_0(\xi)}{\int_\Xi \prod_{i=1}^n \pi(d_i|\xi)\rho_0(\xi) d\xi}$$

which is the conventional Bayes rule.

Integrating more data points into the Gibbs update requires just a summation of the individual losses to form a cumulative loss $l(\xi, \{d_i\}_{i=1}^n) := \sum_{i=1}^n l(\xi, d_i)$. So, the dependence of the loss function on data is straightforward. Hence, without loss of generality, we denote a generic loss function $l(\xi)$ in the sequel for the sake of notation simplicity. In this case, the Gibbs update

formula becomes

$$\rho(\xi) = \frac{\exp(-Wl(\xi))\rho_0(\xi)}{\int_\Xi \exp(-Wl(\xi))\rho_0(\xi)d\xi}. \tag{2.4}$$

3. Surrogate approximation

In the update formula (2.4), evaluating $l(\xi)$ at each parameter ξ requires an evaluation of a potentially expensive PDE model. In this work, we use a computationally inexpensive surrogate model $\bar{l}(\xi)$ to approximate the loss function $l(\xi)$. Although our exposition is generally applicable to any method used for building surrogate models, we will focus later on the integration of an adaptive local reduced basis approach [35] with the current Gibbs posterior framework.

Using $\bar{l}(\xi)$ for the Gibbs update in (2.4) results in an approximate Gibbs posterior that can be sampled (approximated) at a low computational cost. However, it is important to understand the error in such an approximation in order to build an effective surrogate model with controlled accuracy. To this end, we first define the approximate Gibbs posterior $\overline{\rho}(\xi)$ as

$$\overline{\rho}(\xi) = \frac{\exp(-W\bar{l}(\xi))\rho_0(\xi)}{\int_\Xi \exp(-W\bar{l}((\xi))\rho_0(\xi)d\xi}. \tag{3.1}$$

To quantify the error introduced by using the surrogate $\bar{l}(\xi)$ in (3.1), we derive a bound for the discrepancy between the approximate posterior $\overline{\rho}(\xi)$ and $\rho(\xi)$. For this purpose, we first state the following boundedness assumption on the loss function $l(\xi)$ and its surrogate $\bar{l}(\xi)$.

Assumption 1. The loss functions $l(\xi)$ and $\bar{l}(\xi)$ are nonnegative and are uniformly bounded from above: $\exists\, C_l, C_{\bar{l}} > 0$ independent of $\xi \in \Xi$ such that for all $\xi \in \Xi$

$$0 \le l(\xi) \le C_l,\ 0 \le \bar{l}(\xi) \le C_{\bar{l}}. \tag{3.2}$$

To measure the distance between probability distributions, we use the metric

$$h(\rho_1, \rho_2) = \sup_{|f|_\infty \le 1} \sqrt{\mathbb{E}|\rho_1[f] - \rho_2[f]|^2},$$

where $\rho_1, \rho_2 \in P$ are two possibly random elements in P, the supremum is over all $f : \Xi \to \mathbb{R}$ such that $\sup_{\xi\in\Xi} |f(\xi)| \le 1$, and $\rho[f] = \int_\Xi f(\xi)\rho(\xi)d\xi$. The expectation is with respect to the randomness of ρ_1, ρ_2. In case where ρ_1 is determined, and ρ_2 is an approximation to ρ_1 through a randomized algorithm, e.g., Monte Carlo, the expectation is with respect to the randomness of the algorithm. Note that h is indeed a metric on P, in particular, it satisfies the triangle inequality [4, 29].

In addition, we define an e-feasible set as

$$\Xi_e := \{\xi \in \Xi : |l(\xi) - \bar{l}(\xi)| \le e\} \tag{3.3}$$

where $e > 0$ is some constant indicating the accuracy of the surrogate model $\bar{l}(\xi)$. We always assume that e is small, e.g., $We \ll 1$. The set Ξ_e contains all the parameters where the surrogate is accurate in the sense that the absolute difference between $l(\xi)$ and the $\bar{l}(\xi)$ is bounded by e. The complement of Ξ_e is denoted by $\Xi_e^\perp := \Xi \setminus \Xi_e$. Now, we state the following theorem regarding the accuracy of $\overline{\rho}(\xi)$,

Theorem 3.1. *Under Assumption 1, the following bound holds:*

$$h(\rho, \overline{\rho}) \le 2\exp(WC_l)CWe + 2\exp(WC_l + W\max\{C_l, C_{\bar{l}}\})\rho[\mathbb{1}_{\Xi_e^\perp}],$$

for some constants $C > 0$.

Proof. See Appendix A. □

Note that $\rho[\mathbb{1}_{\Xi_e^\perp}] = \int_{\Xi_e^\perp} \rho(\xi) d\xi$ is exactly the posterior measure of $\Xi_e^\perp$. Theorem 3.1 states that, given a prescribed e, if the posterior measure of the region where the surrogate model $\bar{l}(\xi)$ is inaccurate, i.e., $\rho[\mathbb{1}_{\Xi_e^\perp}]$, is small, the approximate posterior $\overline{\rho}$ is close to the true posterior ρ (depending on the prescribed accuracy e). This indicates that the local RB surrogate model only needs to be accurate over the "important region" where the majority of the posterior mass is contained.

Indeed, thanks to the information contained in the data, the posterior distribution typically concentrate on a much smaller portion of the prior support. Hence, we usually do not need a "globally accurate" surrogate model over the entire support of the prior. The local RB method, discussed next, is naturally tailored to provide locally accurate approximations as shown in [34, 35].

4. The local RB surrogate

To construct the surrogate model $\bar{l}(\xi)$, we employ the local RB method first introduced in [35]. We briefly describe the local RB method in this section.

We assume the loss function $l(\xi) = g(u(\xi))$ for some functional $g : U \to \mathbb{R}$, and the functional g is Hölder continuous. That is, there exists $K > 0$ and $\alpha > 0$ such that

$$|g(w) - g(w')| \leq K\|w - w'\|_U^\alpha \quad \forall w, w' \in U.$$

We use the local RB method to build a surrogate model $\overline{u}(\xi) : \Xi \to U$ and evaluate $\bar{l}(\xi)$ as $\bar{l}(\xi) = g(\overline{u}(\xi))$. Note that based on the above assumption, we have that

$$|l(\xi) - \bar{l}(\xi)| \leq K\|u(\xi) - \overline{u}(\xi)\|_U^\alpha \quad \forall \xi \in \Xi. \tag{4.1}$$

In the local RB method, we partition the parameter space into Voronoi cells, i.e., $\Xi = \cup_{k=1}^n \Xi_k$, seeded at m selected atoms ξ_k, $k = 1, \ldots, m$. Within each cell, we form a local basis for approximating the PDE solution $u(\xi)$ using, e.g., full-order PDE solutions at a fixed number of proximal atoms as well as the gradient of the solution at the given seed. For example, Figure 1 shows a partition of a parameter domain $\Xi \subseteq \mathbb{R}^2$ with 2,000 Monte Carlo samples of ξ in the background (e.g., drawn from a prior distribution). The surrogate solution at the large blue dot is computed using a basis consisting of full PDE solutions at the large solid red dots as well as the solution and gradient at the large black dot. The number of neighbors N_b for the local basis is usually chosen to be a fixed constant. In general, the number of neighbors is an algorithmic choice of the user, but can also be chosen adaptively depending on the desired accuracy of the local approximation.

The local RB surrogate model $\overline{u}(\xi)$ of $u(\xi)$ is given by

$$\overline{u}(\xi) = \sum_{k=1}^n \mathbb{1}_{\Xi_k}(\xi) u_k(\xi) \tag{4.2}$$

where $\mathbb{1}_{\Xi_k}$ denotes the characteristic function of the set Ξ_k, i.e., $\mathbb{1}_{\Xi_k}(\xi) = 1$ if $\xi \in \Xi_k$ and $\mathbb{1}_{\Xi_k}(\xi) = 0$ otherwise, and $u_k : \Xi_k \to U_k$ is the solution of the reduced problem

$$\langle M(u_k(\xi); \xi), v \rangle_{V^*, V} = 0 \quad \forall v \in V_k, \quad \forall \xi \in \Xi_k.$$

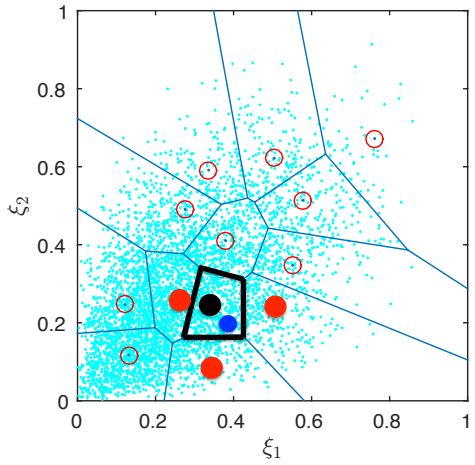


FIGURE 1. The local reduced basis method with two random parameters. The surrogate solution at the large blue dot is computed using a basis consisting of the solution at the large solid red dots as well as the solution and gradient at the large black dot. We plot 2,000 Monte Carlo samples of ξ in the background.

Here, Φ_k is a "local basis" within Ξ_k, e.g.,

$$\Phi_k = \left[u(\xi_k), \nabla_\xi u(\xi_k), u(\xi_{k_1}), u(\xi_{k_2}), \ldots, u(\xi_{k_{N_b}})\right],$$

$U_k = \text{span}(\Phi_k)$, and V_k is a finite-dimensional subspace of V. Also, $u(\xi_k)$ and $\nabla_\xi u(\xi_k)$ are the PDE solution and its gradient at the center of Cell k, while $u(\xi_{k_i}), i = 1...N_b$ are PDE solutions at the centers of neighboring cells. Since the cardinality of Φ_k is typically much smaller than the full discretization of the PDE, we often realize significant computational savings by using $\overline{u}(\xi)$ as a surrogate model for $u(\xi)$. In addition, due to the local nature of the approximation, the evaluation cost of $\overline{u}(\xi)$ at any $\xi \in \Xi$ does not increase as the number of atoms n increases.

To efficiently construct the local RB surrogate, we employ a greedy adaptive sampling procedure to select the atom set $\Theta := \{\xi_k\}_{k=1}^n$ (for details on this algorithm see [35]). The adaptive selection of Θ is guided by reliable a posteriori error indicators, denoted by $\varepsilon_u(\xi)$, i.e.,

$$\|\overline{u}(\xi) - u(\xi)\|_U \lesssim \varepsilon_u(\xi)$$

where $x \lesssim y$ denotes "x is less than or equal to a constant times y." That is, given k atoms, the next atom ξ_{k+1} is selected from the region of Ξ where the current surrogate error is the largest. The error indicators $\varepsilon_u(\xi)$ used in [35] are residual-based error estimates. In fact, we have shown in [35] that the error indicator $\varepsilon_u(\xi)$ can be further used to build more complex error indicators that are specifically targeted for the approximating quantities of interest such as risk measures.

For example, a possible error indicator for $\overline{l}(\xi)$ can be derived using Equation (4.1) as

$$|l(\xi) - \overline{l}(\xi)| \lesssim \varepsilon_l(\xi) := K\varepsilon_u(\xi)^\alpha \quad \forall \xi \in \Xi, \tag{4.3}$$

which we can use to tailor the local RB method to specifically approximate $l(\xi)$.

5. Particle based approximation with local RB surrogate

We now introduce a measure discretization that, along with an RB surrogate, will complete our proposed framework. To this end, consider the pairs $\{\xi_i, w_i\}_{i=1}^m$ and define the empirical measure

$$\rho^E(\xi) = \sum_{i=1}^m w_i \delta(\xi - \xi_i) \tag{5.1}$$

where $\delta(\cdot)$ is the Dirac delta function and the weights w_i satisfy $w_i \in [0,1]$ and $\sum_{i=1}^m w_i = 1$. This form of approximation has been used extensively in sequential Monte Carlo (SMC) methods [13, 14]. Here, we use such an approximation for the Gibbs posterior.

To approximate a given distribution $\rho(\xi)$, ξ_i and w_i need to be selected in some principled manner, e.g., by minimizing some distance between $\rho^E(\xi)$ and $\rho(\xi)$. For instance, a set of Monte Carlo (MC) samples drawn from $\rho(\xi)$ with equal weights is a particle approximation to $\rho(\xi)$. However, MC approximations typically display slow convergence and high variance. We mention in passing that there are more efficient methods for constructing particle approximations such as the Stochastic Reduced Order Model approach introduced in [18, 19].

Now, suppose we have a particle based approximation for the prior $\rho_0(\xi)$ based on the particle set $\{\xi_i, w_i^0\}_{i=1}^m$. Using the empirical measure $\rho_0^E(\xi) := \sum_{i=1}^m w_i^0 \delta(\xi - \xi_i)$ in (2.4), we have an update formula for the particle weights as

$$w_i = \frac{\exp(-Wl(\xi_i))w_i^0}{\sum_{k=1}^m w_k^0 \exp(-Wl(\xi_k))}. \tag{5.2}$$

where w_i is the posterior weight associated with ξ_i. By (5.1), $\{\xi_i, w_i\}_{i=1}^m$ defines a particle based approximation $\rho^E(\xi)$ to the Gibbs posterior distribution $\rho(\xi)$ in (2.4).

If we use a surrogate model $\bar{l}(\xi)$ for the loss function, the posterior weights are then approximated as

$$\overline{w}_i = \frac{\exp(-W\bar{l}(\xi_i))w_i^0}{\sum_{k=1}^m w_k^0 \exp(-W\bar{l}(\xi_k))}. \tag{5.3}$$

The particles $\{\xi_i, \overline{w}_i\}_{i=1}^m$ define a surrogate empirical posterior measure

$$\overline{\rho}^E(\xi) = \sum_{i=1}^m \overline{w}_i \delta(\xi - \xi_i) \tag{5.4}$$

that is an approximation to $\rho^E(\xi)$.

5.1. Accuracy of surrogate particle approximations.

We now undertake an error analysis of the above approximations based on a fixed set of particles. The evolution of particles will be addressed in the following section using the framework of SMC methods. The results shown here are cornerstones for the convergence of the SMC method described in a subsequent section. We first state a lemma regarding the particle based approximation.

Lemma 5.1. *Given a distribution $\rho_0(\xi)$, let $\{\xi_i\}_{i=1}^m$ be drawn independently from $\rho_0(\xi)$, w_i^0 be the associated probability weight, and $\rho_0^E(\xi)$ be the empirical distribution in (5.1), then we have that*

$$h(\rho_0^E, \rho_0) \leq \sqrt{\sum_{i=1}^m (w_i^0)^2}. \tag{5.5}$$

In particular, if $w_i^0 = \frac{1}{m}$*, as for the MC weights, we have that*

$$h(\rho_0^E, \rho_0) \leq \frac{1}{\sqrt{m}}. \tag{5.6}$$

Proof. See Appendix A. □

Lemma 5.2. *Denote the transformation of the Gibbs posterior formula in* (2.4) *as* $G_W : P \to P$*, such that* $\rho(\xi) = G_W \rho_0(\xi)$*. Under Assumption 1, we have*

$$h(G_W \rho_1, G_W \rho_2) \leq 2\exp(WC_l) h(\rho_1, \rho_2), \tag{5.7}$$

for $\forall\ \rho_1, \rho_2 \in P$*.*

Proof. See Appendix A. □

This result states that, for a fixed data set, G_W is continuous with respect to the prior measure in the metric h. Next we consider the surrogate approximation. We first state a counterpart of Lemma 5.2 when the surrogate model (3.1) is used instead of (2.4):

Lemma 5.3. *Denote the transformation of Gibbs posterior formula in* (3.1) *as* $\overline{G}_W : P \to P$*, such that* $\overline{\rho}(\xi) = \overline{G}_W \rho_0(\xi)$*. Under Assumption 1, we have*

$$h(\overline{G}_W \rho_1, \overline{G}_W \rho_2) \leq 2\exp(WC_{\overline{l}}) h(\rho_1, \rho_2), \tag{5.8}$$

for $\forall\ \rho_1, \rho_2 \in P$*.*

Proof. Same as Lemma 5.2. □

We impose the following additional assumption on the local RB surrogate loss function $\overline{l}(\xi)$.

Assumption 2. We assume that the surrogate loss function $\overline{l}(\xi)$ is accurate over the particle set, that is,

$$\sup_{i=1,\ldots,m} |l(\xi_i) - \overline{l}(\xi_i)| \leq e \tag{5.9}$$

for some $e > 0$ that indicates the error of the local RB approximation.

By the definition of the e-feasible set Ξ_e, in view of Assumption 2, we have that ξ_i belongs to Ξ_e. The bound (5.9) can be satisfied for any e by using the local RB Algorithm in [35] using $\{\xi_i\}_{i=1}^m$ as training samples and $\varepsilon_l(\xi)$ in Equation (4.3) as the error indicator. We have the following two lemmas quantifying the difference between ρ^E with weights computed by (5.2) and $\overline{\rho}^E$ in (5.4).

Lemma 5.4. *Under Assumptions 1 and 2, the following bound holds:*

$$D_{KL}(\overline{\rho}^E || \rho^E) \leq 2We.$$

Proof. See Appendix A. □

In words, the error between the empirical posterior obtained from full PDE evaluations and that obtained with a surrogate model is bounded up to a constant by the surrogate model error.

Lemma 5.5. *Under Assumptions 1 and 2, the following bound holds:*

$$h(\rho^E, \overline{\rho}^E) \leq 2\exp(WC_l) CWe,$$

for some constant $C > 0$*.*

Proof. See Appendix A. □

We now show one of the main analytical results of our work. Namely, we provide an error bound for a surrogate model, particle-based approximation to the Gibbs posterior.

Theorem 5.6. *For an empirical distribution ρ_0^E based on $\{\xi_i, w_i^0\}_{i=1}^m$, which is an approximation to the prior ρ_0, the approximate posterior $\overline{\rho}^E$ defined by (5.4) satisfies*

$$h(\overline{\rho}^E, \rho) \leq 2\exp(WC_l)CWe + 2\exp(WC_l)\sqrt{\sum_{i=1}^m (w_i^0)^2} \tag{5.10}$$

for the same constants $C > 0$ as in Lemma 5.5. In particular, if $w_i^0 = \frac{1}{m}$, i.e., the particles are MC samples of the prior, we have

$$h(\overline{\rho}^E, \rho) \leq 2\exp(WC_l)CWe + 2\exp(WC_l)\frac{1}{\sqrt{m}}.$$

Proof. By triangle inequality and Lemma 5.1, 5.2 and 5.5, we have that

$$\begin{aligned} h(\overline{\rho}^E, \rho) &\leq h(\overline{\rho}^E, \rho^E) + h(\rho^E, \rho) \\ &\leq h(\overline{\rho}^E, \rho^E) + 2\exp(WC_l)h(\rho_0^E, \rho_0) \\ &\leq 2\exp(WC_l)CWe + 2\exp(WC_l)\sqrt{\sum_{i=1}^m (w_i^0)^2}. \end{aligned}$$

This completes the proof. □

From these results, we can see that if we use an MC approximation to the prior, we can make $h(\overline{\rho}^E, \rho)$ arbitrarily small by decreasing e and increasing m. In practice, however, it is nontrivial to do both at the same time, as when the number of particles m increases, we require stronger global accuracy on our surrogate model $\overline{l}(\xi)$, which can only be achieved by globally refining the surrogate with more PDE solves. To efficiently represent the posterior distribution with a limited number of particles, we rely on a SMC method to progressively evolve the particles through a sequence of interpolating distributions from ρ_0 to ρ. The samples evolved through the local RB surrogate are automatically navigated to the support of the posterior in the process.

6. An adaptive Sequential Monte Carlo method for particle evolution

The above discussion deals with the asymptotic convergence of the particle approximation based on a fixed set of particles. In practice, using (5.3) as an approximation to the posterior weights with a fixed set of particles drawn from ρ_0 may lead to a poor approximation to ρ, especially when W is large. The reason for this is the potential loss of sample diversity, i.e., the posterior mass may concentrate over just a few particles.

In SMC, instead of computing the posterior weights at once, particles are evolved to approximate a sequence of intermediate distributions interpolating from the prior to posterior. The particles are resampled and mutated after each iteration to prevent degeneracy. We adopt such an SMC framework to approximate the Gibbs posterior distribution. In particular, our SMC method for Gibbs posterior follows closely to the work in [20] where the authors proposed an SMC method for high dimensional inverse problems.

6.1. **The Sequential Monte Carlo method.** In the context of Gibbs posterior, the sequence of the interpolating distributions are defined by

$$\rho_t = G_{W_t}\rho_0,\ 0 \le t \le N \tag{6.1}$$

where $0 = W_0 < W_1 < W_2 \cdots < W_t < \cdots < W_N = W$, and recall the definition of G_W as the Gibbs update formula defined in (2.4). we set $\rho_N = \rho$, which is the posterior distribution we want to approximate. Also, it is easy to show that we have the following property

$$\rho_t = G_{W_t - W_s}\rho_s,\ 0 \le s \le t \le N \tag{6.2}$$

by the Gibbs update formula (2.4). This property allows us to apply SMC methods and progressively approximate ρ.

As mentioned, the key idea of SMC is to start from a particle based approximation of ρ_0, i.e., ρ_0^E, which is easy to obtain, and gradually increase the weight W_t until it reaches W, adjusting the particles along the way. To this end, we denote the particle approximation to ρ_t as ρ_t^E,

$$\rho_t^E = \sum_{i=1}^m w_i^t \delta(\xi - \xi_i^t)$$

based on the particle set $\{\xi_i^t, w_i^t\}_{i=1}^m$.

The iteration $t+1$ of the SMC involves three steps: (i) update the weights of the current particle set $\{\xi_i^t\}_{i=1}^m$ by

$$w_i^{t+1} = \frac{\exp(-(W_{t+1} - W_t)l(\xi_i^t))w_i^t}{\sum_{k=1}^m w_k^t \exp(-(W_{t+1} - W_t)l(\xi_k^t))}. \tag{6.3}$$

The distribution based on $\{\xi_i^t, w_i^{t+1}\}_{i=1}^m$ is denoted by

$$\rho_{t+1,t}^E = \sum_{i=1}^m w_i^{t+1} \delta(\xi - \xi_i^t),$$

that is $\rho_{t+1,t}^E = G_{W_{t+1} - W_t}\rho_t^E$. By Lemma 5.2, we have that

$$h(\rho_{t+1,t}^E, \rho_{t+1}) \le 2\exp((W_{t+1} - W_t)C_l)h(\rho_t^E, \rho_t). \tag{6.4}$$

(ii) Resample the particles $\{\xi_i^t\}_{i=1}^m$ with replacement according to the weights $\{w_i^{t+1}\}_{i=1}^m$, i.e., resample according to $\rho_{t+1,t}^E$. This step effectively eliminates the particles with negligible weights and duplicate the particles with large weights. All resampled particles, including the duplicates, are denoted by $\{\xi_i^{t+1,t}\}_{i=1}^m$ and are assigned equal weights $1/m$. The resampled distribution is denoted by

$$\rho_{t+1,t}^{E,S} = \sum_{i=1}^m \frac{1}{m} \delta(\xi - \xi_i^{t+1,t}).$$

By Lemma 5.1, we have that

$$h(\rho_{t+1,t}^{E,S}, \rho_{t+1,t}^E) \le \frac{1}{\sqrt{m}}. \tag{6.5}$$

This resampling step alone does not prevent sample degeneracy, as only a few particles will survive and copy themselves. To preserve population diversity, a third step is required.

(iii) Apply a ρ_{t+1}-invariant mutation to the resampled set $\{\xi_i^{t+1,t}\}_{i=1}^m$ in step (ii). This can be achieved by evolving the particles $\{\xi_i^{t+1,t}\}_{i=1}^m$ independently by one or more steps using a

ρ_{t+1}-invariant Markov kernel $\mathscr{K}_{t+1}$ (i.e., $\rho_{t+1} = \rho_{t+1}\mathscr{K}_{t+1}$), e.g., an MCMC kernel with ρ_{t+1} as the stationary distribution. Note that the invariant property of $\mathscr{K}_{t+1}$ implies that [4, 29],

$$h(p\mathscr{K}_{t+1}, q\mathscr{K}_{t+1}) \leq h(p,q), \ \forall\ p,q \in P, \ \forall\ 0 \leq t \leq N-1. \tag{6.6}$$

The resulted particles from step (iii), denoted by $\{\xi_i^{t+1}\}_{i=1}^m$, with weights $1/m$, define the distribution

$$\rho_{t+1}^E = \sum_{i=1}^m \frac{1}{m}\delta(\xi - \xi_i^{t+1})$$

that is used to approximate ρ_{t+1} and is used for the next iteration of the SMC. We adopt the same MCMC mutation kernel proposed in [20] for this step, which has been shown to be efficient for high dimensional inverse problems.

We have the following theorem regarding the SMC method for the Gibbs posterior.

Theorem 6.1. *Assuming that the initial particles are a set of MC samples with equal weights $1/m$, then following the above outlined SMC method with the exact loss function $l(\xi)$, we have that for all iterations $t : 0 \leq t \leq N-1$,*

$$h(\rho_{t+1}^E, \rho_{t+1}) \leq \frac{1}{\sqrt{m}} \sum_{s=0}^{t+1} 6^{t+1-s} \exp((W_{t+1} - W_s)C_l) \tag{6.7}$$

in particular, we have a posterior error bound

$$h(\rho^E, \rho) \leq \frac{1}{\sqrt{m}} \sum_{s=0}^{N} 6^{N-s} \exp((W - W_s)C_l). \tag{6.8}$$

where $W = W_N$.

Proof. See Appendix A. □

When a local RB surrogate loss function $\bar{l}(\xi)$ is used, the sequence of distributions are defined by $\overline{\rho_t^E}$. The update in step (i) is replaced by

$$\overline{w_i^{t+1}} = \frac{\exp(-(W_{t+1} - W_t)\bar{l}(\xi_i^t))\overline{w_i^t}}{\sum_{k=1}^m \overline{w_k^t}\exp(-(W_{t+1} - W_t)\bar{l}(\xi_k^t))}. \tag{6.9}$$

which defines $\overline{\rho_{t+1,t}^E} = \sum_{i=1}^m \overline{w_i^{t+1}}\delta(\xi - \xi_i^t)$. That is, $\overline{\rho_{t+1,t}^E} = \overline{G}_{W_{t+1}-W_t}\overline{\rho_t^E}$. By Lemma 5.3, we have

$$h(\overline{\rho_{t+1,t}^E}, \overline{\rho_{t+1}}) \leq 2\exp((W_{t+1} - W_t)C_{\bar{l}})h(\overline{\rho_t^E}, \overline{\rho_t}), \tag{6.10}$$

where

$$\overline{\rho_t}(\xi) = \frac{\exp(-W_t\bar{l}(\xi))\rho_0(\xi)}{\int_\Xi \exp(-W_t\bar{l}((\xi))\rho_0(\xi)d\xi}. \tag{6.11}$$

In addition, the kernel mutation step requires evaluation of the loss function $l(\xi)$ at new parameters which can be accelerated by $\bar{l}(\xi)$ as well. To this end, we use a surrogate kernel $\overline{\mathscr{K}_{t+1}}$ that is invariant with respect to $\overline{\rho_{t+1}}$. The mutation with respect to $\overline{\mathscr{K}_{t+1}}$ only requires evaluation of $\bar{l}(\xi)$. We have the following theorem regarding the SMC method using $\bar{l}(\xi)$:

Theorem 6.2. *Assuming that the initial particles are a set of MC samples with equal weights $1/m$, then following the above outlined SMC method using the local RB surrogate $\bar{l}(\xi)$ for step (i) and (iii), we have that for all iterations $t : 0 \le t \le N-1$,*

$$h(\overline{\rho^E_{t+1}}, \rho_{t+1}) \le \frac{1}{\sqrt{m}} \sum_{s=0}^{t+1} 6^{t+1-s} \exp((W_{t+1} - W_s) C_{\bar{l}}) + 2\exp(W_{t+1} C_l) C_e W_{t+1} e \\ + 2\exp(W_{t+1} C_l + W_{t+1} \max\{C_l, C_{\bar{l}}\}) \rho_{t+1}[\mathbb{1}_{\Xi_e^\perp}] \tag{6.12}$$

In particular, we have the posterior error bound

$$h(\overline{\rho}^E, \rho) \le \frac{1}{\sqrt{m}} \sum_{s=0}^{N} 6^{N-s} \exp((W - W_s) C_{\bar{l}}) + 2\exp(W C_l) C_e W e \\ + 2\exp(W C_l + W \max\{C_l, C_{\bar{l}}\}) \rho[\mathbb{1}_{\Xi_e^\perp}], \tag{6.13}$$

where $W = W_N$.

Proof. The first term on the right-hand-side comes from a simple restatement of Theorem 6.1 for $h(\overline{\rho^E_{t+1}}, \overline{\rho_{t+1}})$. The remainders of the right-hand-side is due to the error bound in Theorem 3.1. □

From Theorem 6.2, we see how we should construct the surrogate model $\bar{l}(\xi)$. Given the number of particles m and the prescribed surrogate accuracy e, we need to build the surrogate $\bar{l}(\xi)$ so that $\rho[\mathbb{1}_{\Xi_e^\perp}]$, i.e. the posterior measure of the "unfeasible set" $\Xi_e^\perp$, is minimized. In terms of local RB surrogate, this requires concentration of local RB atoms and the accurate approximation of $l(\xi)$ over the support of the posterior. To this end, we progressively train the local RB surrogate using the sequence of particles $\{\xi_i^t\}_{i=1}^m$. As the particles gradually cluster over the support for the posterior through the SMC iterations, the focus of the local RB surrogate is automatically navigated to the support of the posterior as well, resulting in a decrease of the measure of the inaccurate set $\rho[\mathbb{1}_{\Xi_e^\perp}]$. In addition, notice that the support of the posterior typically corresponds to a small and local region of the support of the prior. Hence, once the local RB model becomes sufficiently accurate over that region, further evolution of the particles does not require expensive updates of the surrogate model, leading to computational savings.

Notice that the surrogate loss function $\bar{l}(\xi)$ is changing throughout the SMC iterations. We can recover consistency in (6.11) for all t by re-running the SMC algorithm from the beginning up to the current W_t using the latest $\bar{l}(\xi)$ before the next SMC iteration. This procedure is computationally inexpensive since the surrogate model samples do not need to evolve during the re-run and hence no full PDE solves are required. Therefore, we always assume that the update (6.11) is consistent for all iterations with the latest surrogate model $\bar{l}(\xi)$.

We now present the MCMC algorithm for the mutation step using the local RB surrogate $\bar{l}(\xi)$. To this end, we first define the following mean and variance of $\overline{\rho^E_{t+1,t}}$, which is the particle distribution after SMC step (i) and before resampling step (ii), for each parameter dimension $j \in \{1, 2, \ldots, M\}$ as

$$\overline{m^E_{t+1,t,j}} = \sum_{i=1}^m \overline{w_i^{t+1}} \xi_{i,j}^t, \quad \overline{\Sigma^E_{t+1,t,j}} = \sum_{i=1}^m \overline{w_i^{t+1}} (\xi_{i,j}^t - \overline{m^E_{t+1,t,j}})^2. \tag{6.14}$$

The above quantities provide estimates of the mean and variance of $\overline{\rho_{t+1}}$ along each individual parameter dimension at SMC iteration $t+1$, and will be used to facilitate the design a proposal distribution for the MCMC kernel $\overline{\mathscr{K}_{t+1}}$.

Based on the above definition, a proposal $\hat{\xi}_i^{t+1,t}$ for a particle $\xi_i^{t+1,t} \in \{\xi_i^{t+1,t}\}_{i=1}^m$ can be obtained by the following mutation

$$\hat{\xi}_{i,j}^{t+1,t} = \overline{m_{t+1,t,j}^E} + \gamma(\xi_{i,j}^{t+1,t} - \overline{m_{t+1,t,j}^E}) + \sqrt{1-\gamma^2}\Lambda_{t+1,t,j}, \quad 1 \le j \le M \tag{6.15}$$

where γ is an algorithmic constant and $\Lambda_{t+1,t,j}$ is a random variable with distribution $\mathcal{N}(0, \overline{\Sigma_{t+1,t,j}^E})$. Note that the scaling of the proposal distribution is tailored for each individual parameter dimension by the variance estimates $\overline{\Sigma_{t+1,t,j}^E}$ to improve mixing. In contrast to standard random walk proposals, the above proposal scales to high dimensional problems as shown in [20]. The transition probability associated with the proposal in (6.15) is given by

$$Q(\hat{\xi}_i^{t+1,t} | \xi_i^{t+1,t}) = \exp\left(-\frac{1}{2(1-\gamma^2)} \sum_{j=1}^{M} \frac{(\hat{\xi}_{i,j}^{t+1,t} - \overline{m_{t+1,t,j}^E} - \gamma(\xi_{i,j}^{t+1,t} - \overline{m_{t+1,t,j}^E}))^2}{\overline{\Sigma_{t+1,t,j}^E}} \right). \tag{6.16}$$

Algorithm 1 shows the $\overline{\rho_{t+1}}$-invariant mutation MCMC sampler.

Algorithm 1 The MCMC algorithm for $\overline{\rho_{t+1}}$-invariant mutation

For each $\xi(0) \in \{\xi_i^{t+1,t}\}_{i=1}^m$, evolve $\xi(0)$ independently for I steps with the following procedure

- **For** $i = 1, 2, \ldots, I$, **do**
  - Draw a proposal $\widehat{\xi(i)}$ using the proposal distribution in (6.15) based on $\xi(i-1)$.
  - Use $\bar{l}(\xi)$ to evaluate $\alpha = 1 \wedge \frac{\exp(-W_{t+1}\bar{l}(\widehat{\xi(i)}))\rho_0(\widehat{\xi(i)})Q(\xi(i-1)|\widehat{\xi(i)})}{\exp(-W_{t+1}\bar{l}(\xi(i-1))\rho_0(\xi(i-1))Q(\widehat{\xi(i)}|\xi(i-1))}$
  - With probability $1-\alpha$, reject $\widehat{\xi(i)}$ and set $\xi(i) = \xi(i-1)$, $i = i+1$. Go back to the proposal step.
  - If $\widehat{\xi(i)}$ not rejected, set $\xi(i) = \widehat{\xi(i)}$ and $i = i+1$. Go back to the proposal step.

  End
- Finally, return $\xi(I)$ as a $\overline{\rho_{t+1}}$-invariant mutation of $\xi(0)$.

6.2. **Adaptive selection of the SMC step size.** We now describe how the sequence of step size $0 = W_0 < W_1 < W_2 \cdots < W_t < \cdots < W_N = W$ can be selected adaptively. For each SMC iteration $t+1$, we would like to greedily apply all the residual weight $\Delta W = W_N - W_t$ to the particle distribution $\overline{\rho_t^E}$ from the previous iteration, so that we can directly approximate the posterior ρ. After applying the SMC step (i), i.e., updating the weights by Equation (6.9) using ΔW, we check a simple criteria called the effective sample size (ESS), which is used to measure the sample degeneracy of the current weights

$$\text{ESS} = \left(\sum_{k=1}^{m} \left(\overline{w_i^{t+1}} \right)^2 \right)^{-1}.$$

Note that ESS is small if the majority of the probability weights are pivoted on only a few particles, which indicates the loss of sample diversity. In this case, we reduce the incremental weights by a constant factor $\Delta W = \theta \Delta W$ with $\theta \in (0,1)$ and repeatedly backtrack and reevaluate Equation (6.9) and ESS until the latter variable is above some preset threshold. In this case, we accept ΔW, set $W_{t+1} = W_t + \Delta W$, move on to the step (ii) and (iii) and finish the current SMC iteration $t+1$. If the residual weight is not zero after iteration $t+1$, we set $t = t+1$ and move to the next SMC iteration. Otherwise, we have applied the total weight to the prior and obtained an approximation to the posterior.

Of course, the local RB surrogate model $\bar{l}(\xi)$ evolves as well by the adaptive training on the particles before applying the incremental weight in each SMC iteration. We require $\bar{l}(\xi)$ to satisfy Assumption 2 for each iteration t. As the particles gradually cluster over a small region in the parameter space, i.e., the support for the posterior, $\bar{l}(\xi)$ becomes accurate in that region as well, reducing the measure of the inaccurate set $\rho[\mathbb{1}_{\Xi_e^\perp}]$ as a result. In addition, as the particles become more compact in a local region, expensive refinements of the local RB model are less often triggered due to the local accuracy of $\bar{l}(\xi)$.

We first present the adaptive refinement of the local RB surrogate over a given particle set $\{\xi_i^t\}_{i=1}^m$ in Algorithm 2. We then show the complete adaptive SMC method in Algorithm 3. In Algorithm 2, we note that the accuracy parameter e_{thre} is prescribed beforehand and can be made adaptive as well. For example, we can further improve computational efficiency by setting a larger e_{thre} in the beginning stage of the SMC algorithm and gradually reduce e_{thre} throughout the iterations. This strategy leads to computational savings in the beginning stage when the particles are far from the support of the posterior and are less relevant for characterizing the posterior distribution. However, it is essential to set e_{thre} small enough such that each SMC iteration still leads to the particles moving towards the posterior.

One possible approach is to choose e_{thre} based on the range of variation of $\bar{l}(\xi)$ over the current particle set $\{\xi_i^t\}_{i=1}^m$, e.g., we can set e_{thre} to be a small fraction of the standard deviation of $\{\bar{l}(\xi_i^t)\}_{i=1}^m$ and compute e_{thre} automatically for each SMC iteration. With this approach, e_{thre} is large at initial stages of the SMC algorithm where particles are diverse and the range of variation of $\bar{l}(\xi)$ is large. In the latter stages where particles are more clustered, $\bar{l}(\xi)$ has a smaller range over the particles, which leads to a smaller e_{thre}.

Algorithm 2 Adaptive refinement of local RB surrogate

Given the current particle set $\Xi_P := \{\xi_i^t\}_{i=1}^m$, the current local RB surrogate model $\bar{l}(\xi)$ for $l(\xi)$, and a desired accuracy threshold e_{thre},

- Compute the local RB error indicator $\varepsilon_l(\xi)$ for each particle in Ξ_P and $e_{\max} = \max_{\xi \in \Xi_P} \varepsilon_l(\xi)$.
- **While** $e_{\max} > e_{\text{thre}}$, **do**
  - Select the particle $\xi_{\max} = \operatorname{argmax}_{\xi \in \Xi_P} \varepsilon_l(\xi)$.
  - Update $\bar{l}(\xi)$ by the PDE information at $\xi_{\max}$ via local RB method.
  - Update the error indicators $\varepsilon_l(\xi)$ for each particle in Ξ_P and $e_{\max}$.

  End
- Return the updated surrogate $\bar{l}(\xi)$.

Algorithm 3 The adaptive SMC method

Given initial particle approximation $\rho_0^E := \sum_{i=1}^m w_i^0 \delta(\xi - \xi_i^0)$ (and $\overline{w_i^0} := w_i^0$), the total loss weight W to be applied, set $W_{\text{current}} = 0$ and $t = 0$.

- **While** $W_{\text{current}} < W$, **do**
  - Run Algorithm 2 to possibly refine the local RB surrogate $\bar{l}(\xi)$ over the current particle set $\{\xi_i^t\}_{i=1}^m$.
  - Set $\Delta W = W - W_{\text{current}}$.
  - **While** TRUE, **do**
    - Compute the new weight $\overline{w_i^{t+1}}$ by Equation (6.9) using ΔW as the incremental weight.
    - Compute ESS of $\left\{\overline{w_i^{t+1}}\right\}_{i=1}^m$.
    - **If** ESS $>$ ESS$_{\text{thre}}$, **break**.
    - Backtrack: $\Delta W = \theta \Delta W$.

    End
  - Resample particles $\{\xi_i^t\}_{i=1}^m$ according to $\left\{\overline{w_i^{t+1}}\right\}_{i=1}^m$ to obtain $\{\xi_i^{t+1,t}\}_{i=1}^m$ with weights $1/m$.
  - Mutate each particle in $\{\xi_i^{t+1,t}\}_{i=1}^m$ with Algorithm 1 to obtain a new set of evolved particles $\{\xi_i^{t+1}\}_{i=1}^m$, set $\overline{w_i^{t+1}} = 1/m$, obtain the current particle approximation $\overline{\rho_{t+1}^E} = \sum_{i=1}^m \frac{1}{m} \delta(\xi - \xi_i^{t+1})$.
  - Set $W_{\text{current}} = W_{\text{current}} + \Delta W$, $t = t + 1$.

  End
- Report $\overline{\rho_{t+1}^E}$ as an approximation to the Gibbs posterior ρ.

7. Choosing the weight for the loss function

In this section we describe one approach to select W, the weights in the loss function. The importance of the weights in the Gibbs posterior formulation is to calibrate the loss, the calibration is automatic in classic Bayesian inference as the density of the data generation process is the calibration. For example, in the case of Gaussian noise the $\frac{1}{2\sigma^2}$ in the likelihood function can be interpreted as the weight, so when the noise level is high the likelihood is discounted as compared to the low noise setting. Methods to select W are generally subjective in the context of the Gibbs posterior, and often problem dependent as well [5, 32]. The work in [5] introduced several subjective ways to select W. In particular, one proposed method is to select W by balancing two isolated loss terms from the objective function (2.2). In settings where large data sets are available, one can also select W using methods like cross-validation to tune the predictive performance of the posterior.

For inverse problems, however, one typically has access to a rather limited number of observations, so without some assumption on the noise in the data it is hard to quantify uncertainty about the inverse solution. We will make some weak assumptions on the noise to provide a method to set the weight. We assume the noise are i.i.d with mean and standard deviation

$$\mathbb{E}[\varepsilon] = \varepsilon^M, \quad \varepsilon^D = \left(\mathbb{E}[\varepsilon^2] - \mathbb{E}[\varepsilon]^2\right)^{\frac{1}{2}}.$$

We adopt an approach that is in the same spirit as the Morozov's discrepancy principle [7, 30]. We select a weight for which the mean and standard deviation of residual of the posterior predictions will match the statistics on the observed data

$$W_{\text{opt}} = \text{argmin}_{W\in\boldsymbol{W}} \frac{\|\frac{1}{n}\sum_{i=1}^n \bar{\varepsilon}_i(W) - \varepsilon^M\| + \|\sqrt{\frac{1}{n-1}\sum_{i=1}^n (\bar{\varepsilon}_i(W) - \varepsilon^M)^2} - \varepsilon^D\|}{\|\varepsilon^D\|} \tag{7.1}$$

with $\bar{\varepsilon}_i(W) = \mathscr{F}(\mathbb{E}_\rho(\xi)) - d_i$ is the posterior predicted noise or residual for observation i, given weight W. Selecting the set $\boldsymbol{W}$ to optimize over in the above equation is nontrivial. In addition, when the sample size is small, Note that, even with only one observation, if multiple channels are available, one can still estimate the statistics of noise and use (7.1) to select W, however one would not have a great deal of trust in the estimate.

We know that the weight would be $\frac{1}{2(\varepsilon^D)^2}$ for the classic Bayesian setting with square loss and Gaussian noise. Using this information we provide a discrete grid of candidate weight values $\boldsymbol{W} := \left[\frac{1}{2(\varepsilon^D)^2 T}, T\frac{1}{2(\varepsilon^D)^2}\right]$, where $T > 1$ is a range parameter (e.g., $T = 50$). The discretization is for computational efficiency. To address the case where we may have a very small sample size we stabilize our weight estimate by modeling averaging with the standard Bayesian case

$$W = \frac{S}{S+n-1}\frac{1}{2(\varepsilon^D)^2} + \frac{n-1}{S+n-1}W_{\text{opt}} \tag{7.2}$$

for some $S \geq 1$ (e.g., $S = 10$). When n is small, we favor the empirical weight $\frac{1}{2(\varepsilon^D)^2}$, when n is large and the noise statistics can be computed with good accuracy and we favor the optimized weight W_{opt}, the above can be considered an empirical Bayes procedure.

We can take advantage of the sequential structure of the SMC procedure to efficiently evaluate the objective in (7.1) using intermediate computations from the SMC procedure. This allows us to efficiently compute W_{opt} upon finishing the SMC run and then compute the final weight by (7.2),.

It is worth further investigation to see if one can choose the weight W purely based on the data, instead of imposing additional assumptions of the noise. In addition, it is useful to understand if the use of further information about $\rho(\xi)$ beyond the posterior mean $\mathbb{E}_\rho(\xi)$ would help in determining W.

8. Numerical examples

Now, we present three numerical examples to show the behavior and computational efficiency of our SMC method.

8.1. 1D advection diffusion equation.

In the first example, we consider a 1D advection-diffusion problem. We show that our method recovers the usual Bayesian approach when a likelihood function is available and that we use the negative log-likelihood as the weighted loss function.

Let $D = (0,1)$ and consider the following boundary value problem

$$-\nu\frac{\partial^2 u}{\partial x^2}(x,\xi^*) + b(x,\xi^*)\frac{\partial u}{\partial x}(x,\xi^*) = f(x), \quad x \in D \tag{8.1a}$$

$$u(0,\xi^*) = u(1,\xi^*) = 0 \tag{8.1b}$$

The diffusivity, ν, and source, f, are known whereas the advection field, b, is a piecewise constant random field parametrized by two unknown parameters ξ_1^* and ξ_2^* as

$$b(x,\xi^*) = [b_1 + 2\xi_1^*]\,\mathbb{1}_{[0,0.5)}(x) + [b_2 + 2\xi_2^*]\,\mathbb{1}_{[0.5,1]}(x) \tag{8.2}$$

where $\mathbb{1}_S(x)$ is one if x is in the set S and is zero otherwise.

We are able to measure the solution at three different locations of $x = [0.1, 0.5, 0.9]$. Our noisy data is hence given by

$$d = \mathscr{D}u + \varepsilon$$

where $\mathscr{D}$ is an operator that maps the solution $u(x,\xi^*)$ to the measurement and ε is a noise vector that contains i.i.d entries. We assume the noise is drawn from a Gaussian distribution with standard deviation equal to 10% of the magnitude of the true data. In particular, we have $\varepsilon^D = 0.173$. To match the Gaussian likelihood, we use $W = \frac{1}{2(\varepsilon^D)^2} = 16.70$ and the loss

$$l(\xi, d) = \|\mathscr{D}u(x,\xi) - d\|_{l_2}^2.$$

The values of the known parameters are $\nu = 0.1,\ b_1 = -0.5,\ b_2 = -0.2,\ f(x) = 1$, while the true values of unknown parameters are $\xi_1^* = 0.2,\ \xi_2^* = 0.7$. For the prior distributions, we assume $\xi_1 \sim U[0,1],\ \xi_2 \sim U[0,1]$. We use Algorithm 3 to compute the Gibbs posterior with $m = 100$ evolving particles and local RB accuracy set to be $1e-3$. In addition, as reference, we perform the standard Random Walk Metropolis-Hastings algorithm with Gaussian likelihood to obtain $5,000$ samples from the posterior with $1,000$ burn-in steps.

We show the comparison of our SMC result with the MCMC reference in Figure 2. Clearly, the SMC method performs similarly to the reference in approximating the posterior distribution. In particular, the SMC method took just 3 iterations to reach a good approximation of the posterior. The evolution of the particles, the atoms of the local RB surrogate and the intermediate distributions are shown in Figure 3 to 6 for the various iterations. As can be seen, as the weight W is progressively increased, the particles cluster around the support of the posterior, while simultaneously leading the local RB surrogate to concentrate on the relevant region of the parameter space.

We report the accumulative number of PDE solves at each SMC iteration in Figure 7. Most of the computational effort corresponding to the construction of the local RB surrogate is spent the first iteration as the particles move the most towards the posterior support. Once the local RB becomes accurate over the posterior region, further evolution of the particles rarely triggers the refinement of the surrogate. The total number of PDE solves to obtain the shown posterior for this example was around 200, representing a significant computational saving over the MCMC method.

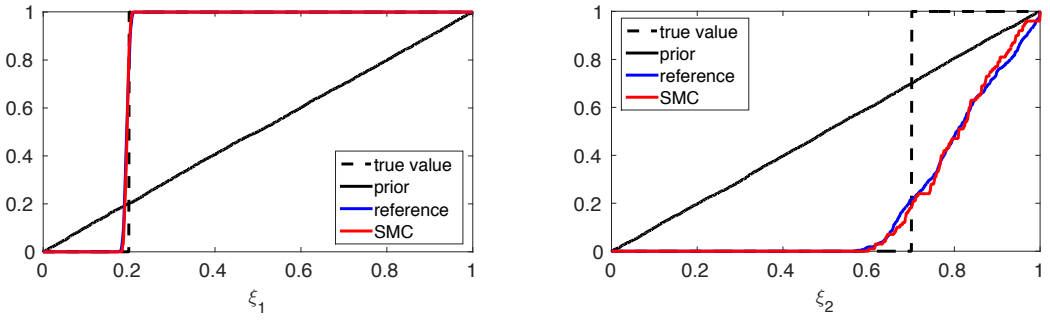
8.2. **2D advection diffusion equation.** In the second example, we consider the simultaneous identification of the diffusivity constant and unknown source for a 2D advection-diffusion problem. Let $D = (0,1)^2$. We consider the following problem,

$$-\nabla\cdot(\kappa(\xi^*)\nabla u(x,\xi^*)) + v(x)\cdot\nabla u(x,\xi^*) = f(x,\xi^*) \qquad x \in D \tag{8.3a}$$

$$u(x,\xi^*) = 0 \qquad x \in \Gamma_d \tag{8.3b}$$

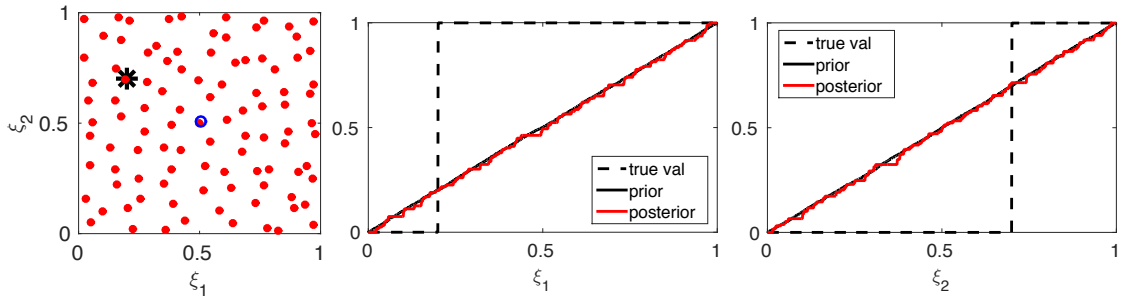
$$\kappa(\xi^*)\nabla u(x,\xi^*)\cdot n = 0 \qquad x \in \Gamma_n \tag{8.3c}$$

where $\Gamma_d := [0,1]\times\{0\}$ and $\Gamma_n := \partial D \setminus \Gamma_d$. The unknown parameters ξ^* are included in the diffusivity constant $\kappa(\xi^*)$ and the source term $f(x,\xi^*)$.



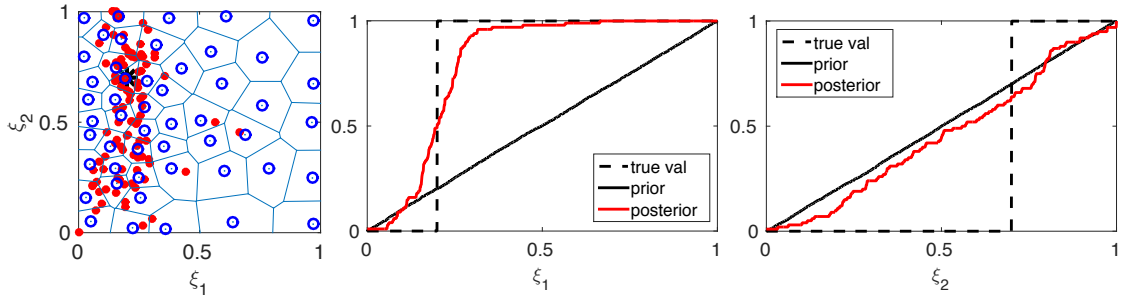
(A) CDF plot for ξ_1 (B) CDF plot for ξ_2

FIGURE 2. 1D advection diffusion equation: Comparison of the posterior distribution of the parameters computed by Algorithm 3 and the standard MCMC method with Gaussian likelihood function.



(A) Initial particles (B) CDF plot for ξ_1 (C) CDF plot for ξ_2

FIGURE 3. 1D advection diffusion equation: SMC iteration 0, with a loss weight $W_0 = 0$. In A) the true parameters are black, the particles are red and the atoms for local RB are blue.

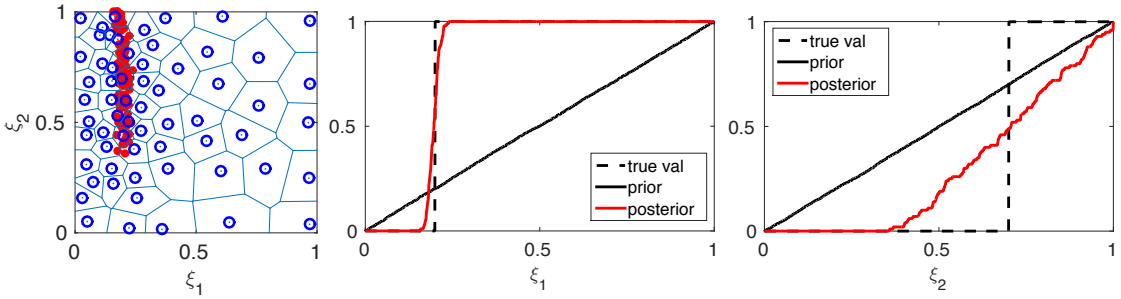


(A) Adapted particles (B) CDF plot for ξ_1 (C) CDF plot for ξ_2

FIGURE 4. 1D advection diffusion equation: SMC iteration 1, with a loss weight $W_1 = 0.0782$. In A) the true parameters are black, the particles are red and the atoms for local RB are blue.

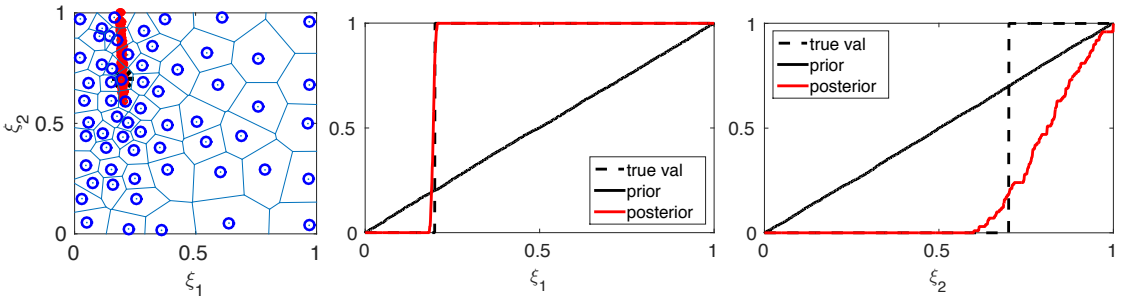
In particular, the diffusivity is modeled as

$$\kappa(\xi^*) = 0.02 + 0.98\xi_1^* \tag{8.4}$$



(A) Adapted particles (B) CDF plot for ξ_1 (C) CDF plot for ξ_2

FIGURE 5. 1D advection diffusion equation: SMC iteration 2, with a loss weight $W_2 = 1.43$. In A) the true parameters are black, the particles are red and the atoms for local RB are blue.



(A) Adapted particles (B) CDF plot for ξ_1 (C) CDF plot for ξ_2

FIGURE 6. 1D advection diffusion equation: SMC iteration 3, with a loss weight $W_3 = 16.7$. In A), the true parameters are black, the particles are red and the atoms for local RB are blue.

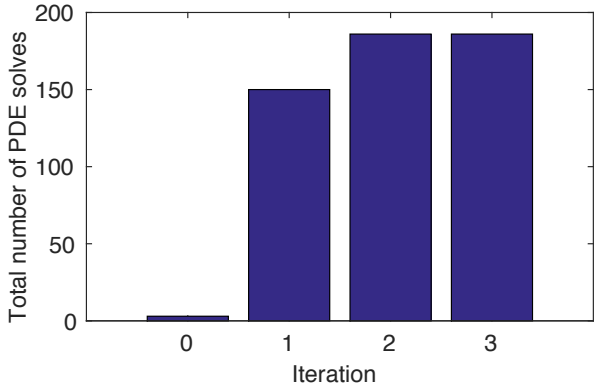


FIGURE 7. 1D advection diffusion equation: The accumulative number of PDE solves at each iteration of the SMC Algorithm 3 for the 1D advection diffusion problem.

The advection field is divergence free and is defined by

$$v(x) = 13\begin{pmatrix}1\\0\end{pmatrix} + 9\begin{pmatrix}-x_1\\x_2\end{pmatrix}. \tag{8.5a}$$

Finally, the forcing term f is modeled by two Gaussians function with unknown magnitudes, i.e.,

$$f(x,\xi^*) = 10\exp\left(\frac{-(x_1-0.25)^2-(x_2-0.5)^2}{0.25^2}\right)\xi_2^* + 5\exp\left(\frac{-(x_1-0.75)^2-(x_2-0.75)^2}{0.33^2}\right)\xi_3^* \tag{8.6}$$

The goal is to identify ξ^* from noisy measurements of the PDE solution $u(x,\xi^*)$. Again, we assume that the concentration field is measured over a uniform grid in the domain. Our noisy data is hence given by

$$d = \mathscr{D}u + \varepsilon$$

where $\mathscr{D}$ is an operator that maps the solution $u(x,\xi^*)$ to the measurement locations and ε is a noise vector that contains i.i.d entries. Notice that we are assuming that the concentration field has enough regularity as to allow for point-wise evaluations. We assume the noise is drawn from a Gaussian distribution with standard deviation equal to 20% of the magnitude of the true data. In particular, we have $\varepsilon^D = 0.0197$. For this problem, we use the following l_1 loss:

$$l(\xi,d) = \|\mathscr{D}u(x,\xi) - d\|_{l_1}.$$

The weight W was obtained using the approach outlined in Section 7. After estimating W, we compare the SMC method in Algorithm 3 to a Random Walk Metropolis-Hastings method using $\exp(-Wl(\xi))$ as the likelihood. Notice that the Gibbs posterior is invariant with respect to the MC transitions.

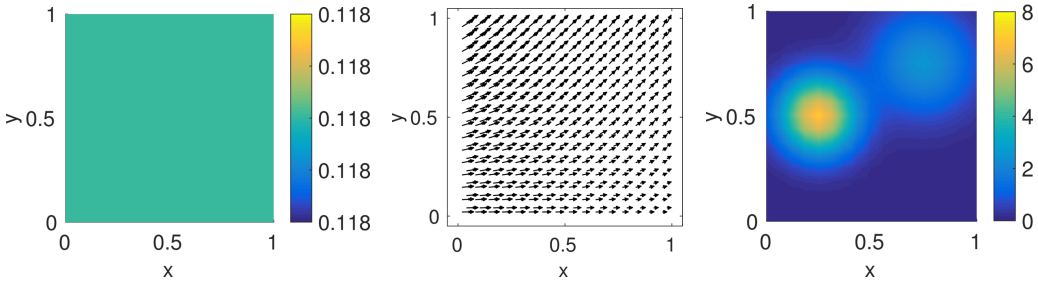
The true values of the unknown parameters are $\xi_1^* = 0.1,\ \xi_2^* = 0.7,\ \xi_3^* = 0.5$. For the prior, we assume $\xi_1 \sim \beta(1,2),\ \xi_2 \sim \beta(3,1),\ \xi_3 \sim \beta(3,1)$ and that they are independent. The final weight selected was $W = 25.8$, representing approximately $1/50$ of $\frac{1}{2(\varepsilon^D)^2}$. We use Algorithm 3 to compute the Gibbs posterior with $m = 100$ evolving particles. In addition, when training the local RB surrogate model at each SMC step t, we employ an adaptive accuracy e_{thre} that is equal to 2% of the standard deviation of $\{\bar{l}(\xi_i^t)\}_{i=1}^m$. We run the reference MCMC method to obtain $5,000$ samples from the posterior with $1,000$ burn-in steps.

In Figure 8, we show the true diffusivity, advection and source fields. In Figure 9 we show the noise-free PDE solution, the corrupted solution, and the measurement points. We show the comparison of our SMC result with the MCMC reference in Figure 10. Again, the SMC method performs similarly to the reference in approximating the posterior distribution. The random variable ξ_3 has the largest posterior uncertainty due to the fact that the solution, hence the data, has the least sensitivity with respect to this parameter. Notice that the source associated with ξ_3 is located near the top right corner of the domain and, hence, has limited impact on the concentration at most of the measurement points.

Only 3 iterations of our SMC algorithm were needed to reach the predefined tolerance in this example. Figure 11 shows the evolution of particles as well as the local RB atoms throughout the SMC iterations. Clearly, the particles and local RB atoms simultaneously evolve towards the support of the posterior, leading to an improved approximation of the posterior distribution. In addition, as the particles become more clustered, the variation of $\bar{l}(\xi)$ over the particles becomes lower, leading to smaller e_{thre} (higher accuracy requirement on the surrogate).

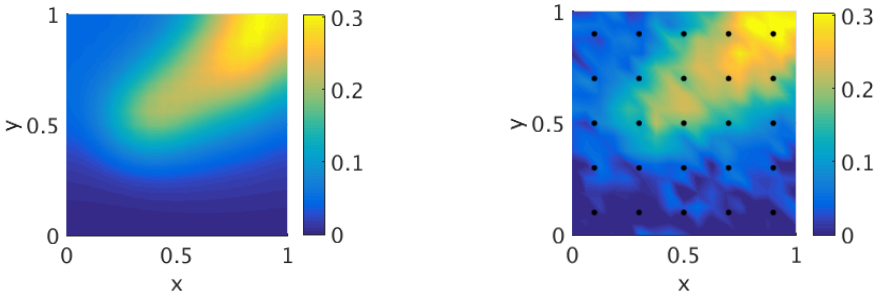
Finally, we show the accumulative number of PDE solves at each iteration in Figure 12. Note that we did not include the PDE solves in the preprocessing step to select W, and we reinitialized the local RB surrogate model before computing the posterior under the final weight. We do this

to demonstrate how the computational efforts corresponding to the construction of the local RB surrogate are distributed in the SMC iterations. The number of PDE solves (local RB atoms) depends critically on e_{thre} in each iteration. In the first iteration, because e_{thre} is relatively large, only about 100 PDE solves are incurred. In the latter iterations, e_{thre} becomes lower and the accuracy requirement imposed on $\bar{l}(\xi)$ becomes tighter as well. On the other hand, the particles become more compact in the latter iterations. Though the decreased e_{thre} demands more PDE solves to refine the surrogate model, the increased clustering of the particles makes it easier for the local RB surrogate to reach the accuracy requirement. These two competing factors jointly determine the number of additional refinements on the surrogates. Overall, only less than 400 PDE solves were incurred in the SMC method to obtain the approximate posterior, representing a significant computational saving over the MCMC reference.



(A) Constant diffusivity field (B) Advection field (C) Source

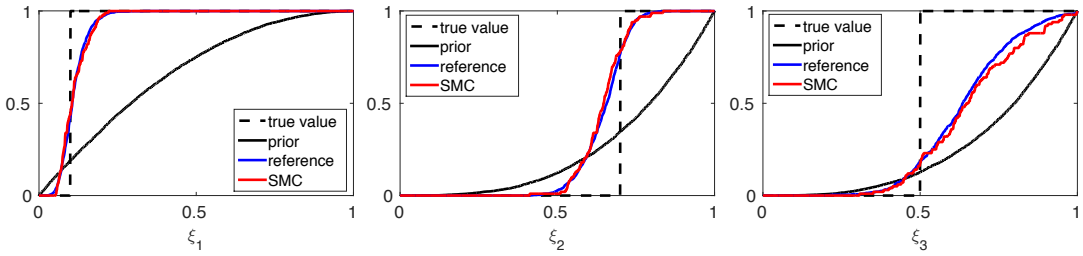
FIGURE 8. 2D advection diffusion equation: The diffusivity, advection and source fields of the 2D advection-diffusion equation.



(A) (B)

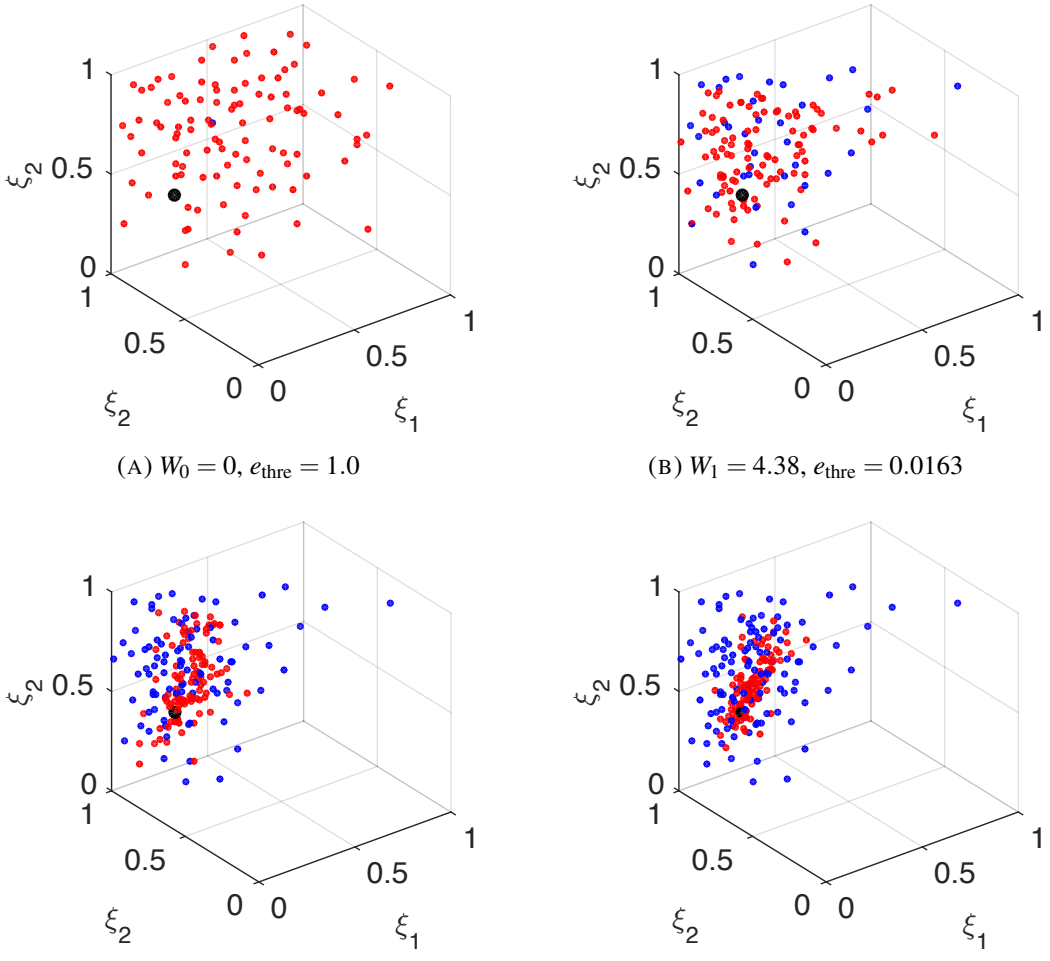
FIGURE 9. 2D advection diffusion equation: The noise-free solution and the noisy measurements.

8.3. **2D elasticity equation.** In the last example, we consider two simple elastography problems where we need to infer the distribution of mechanical properties given noisy displacement measurements under known loads. These problems are usually characterized by higher dimensionality than those in the previous examples, and hence, are more computationally expensive to solve.



(A) CDF plot for ξ_1 (B) CDF plot for ξ_2 (C) CDF plot for ξ_3

FIGURE 10. 2D advection diffusion equation: Comparison of the posterior distribution of the parameters computed by Algorithm 3 and the standard MCMC method.



(A) $W_0 = 0$, $e_{\text{thre}} = 1.0$ (B) $W_1 = 4.38$, $e_{\text{thre}} = 0.0163$

(C) $W_2 = 19.4$, $e_{\text{thre}} = 0.00395$ (D) $W_3 = 25.9$, $e_{\text{thre}} = 0.00188$

FIGURE 11. 2D advection diffusion equation: The evolution of the particles and the local RB atoms at each iteration of SMC algorithm. The true parameter is in black, the particles are in red and the local RB atoms are in blue.

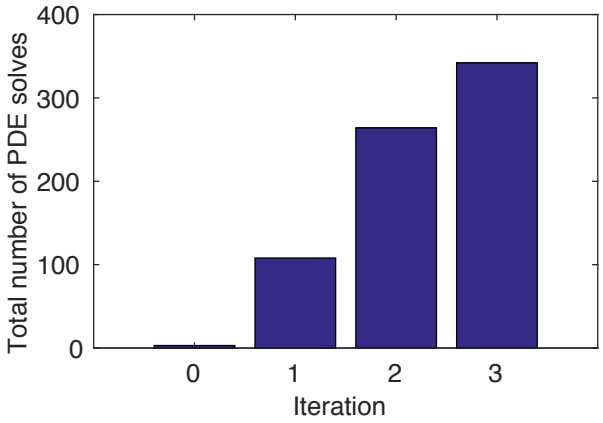


FIGURE 12. 2D advection diffusion equation: The accumulative number of PDE solves at each iteration of the SMC Algorithm 3 for the 2D advection diffusion problem.

Letting $D = (0,1)^2$, we consider the following linear elasticity problem.

$$-\nabla \cdot \sigma(x,\xi^*) + f = 0, \qquad x \in D \tag{8.7a}$$

$$\varepsilon(x,\xi^*) = \frac{1}{2}(\nabla u(x,\xi^*) + \nabla u(x,\xi^*)^T), \qquad x \in D \tag{8.7b}$$

$$\sigma(x,\xi^*) = C(\xi^*) : \varepsilon(x,\xi^*), \qquad x \in D \tag{8.7c}$$

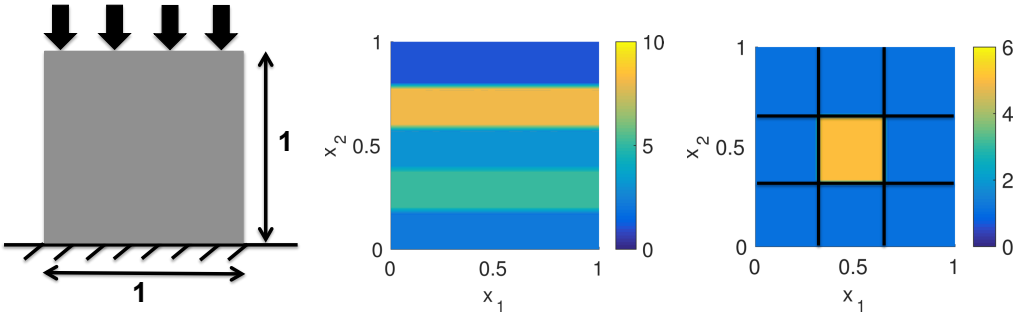
$$u(x,\xi^*) = 0 \qquad x \in \Gamma_d \tag{8.7d}$$

$$\sigma(x,\xi^*) \cdot n = \tau \qquad x \in \Gamma_n \tag{8.7e}$$

where $\Gamma_d := [0,1] \times \{0\}$ and $\Gamma_n := \partial D \setminus \Gamma_d$. The unknown parameters ξ^* are included in the modulus of the material, which is part of the elasticity tensor $C(\xi^*)$. We consider isotropic plane stress problems where we know the Poisson's ratio $\nu = 0.3$ and try to identify the unknown Young's modulus $E(\xi^*)$ from noisy measurements of $u(x,\xi^*)$. The setup of the problem as well as the two modulus models we used in this example are shown in Figure 13. The two problems have parameter dimensions of 5 and 9, respectively.

The displacement fields and the noisy measurements of the two models are shown in Figure 14 and 15, respectively. Note that we only use the noisy displacement data in the vertical, i.e., x_2, direction. We perturb the solution with 5%, 10% and 20% Gaussian noise and investigate the posterior mean and deviation computed by the SCM method. For the prior distribution, we assume the parameters are independent and follow a $\beta(1,3)$ distribution scaled to the range of $[0.1,10]$. We use a simple l_2 loss function and weights $W = \frac{1}{2(\varepsilon^D)^2}$, which corresponds to the Gaussian noise model exactly. In addition, when training the local RB surrogate model at each SMC step t, we employ an adaptive accuracy e_{thre} that is equal to 5% of the standard deviation of $\{\bar{l}(\xi_i^t)\}_{i=1}^m$.

We plot the inversion for the two models under different level of noise in Figure 16 and 17, respectively. The posterior mean gives reasonable approximations to the true modulus, and as the level of noise in the data increases, we have higher uncertainty about our inverse solution, as expected. Finally, we present the number of local RB atoms used in each of the models with each level of noise in Figure 18. When the noise level is low, i.e., the weight for the loss



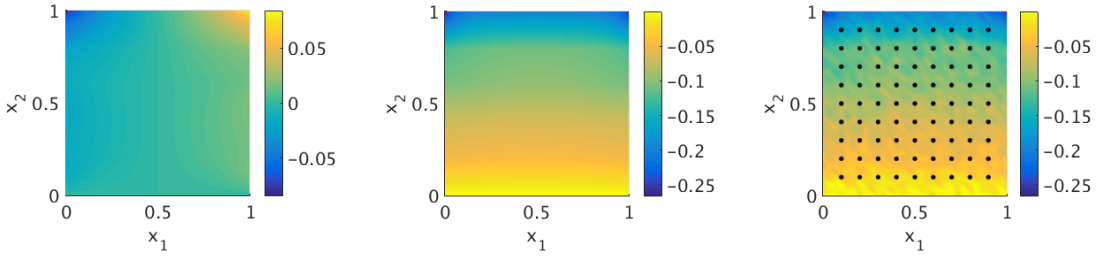
(A) Boundary conditions.

(B) Material with layered Young's modulus.

(C) Material with hard inclusion.

FIGURE 13. 2D elasticity equation: The boundary condition, and true material properties for the simple elastography problems.

is large, the posterior becomes increasingly concentrated in a small region within the support of the prior, and more SMC steps are needed to approximate the Gibbs posterior, leading to a larger number of refinements (atoms) for the local RB. Also noticed from the comparison is that when the dimension of the parameter space becomes high, as for the inclusion problem, higher computational cost is required to approximate the Gibbs posterior.



(A) Horizontal displacement

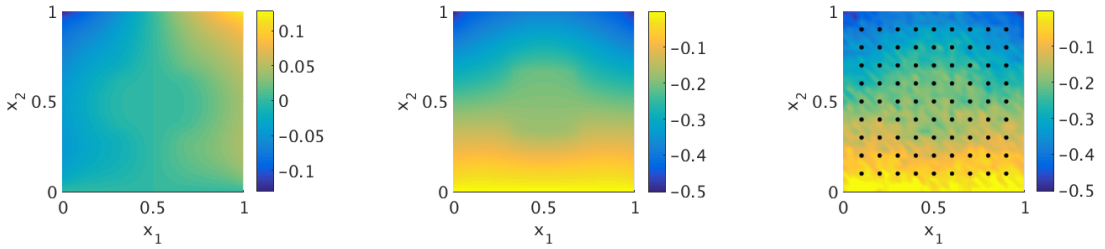
(B) Vertical displacement

(C) Measurement of the vertical displacement with 10% Gaussian noise

FIGURE 14. 2D elasticity equation: The displacement fields and the noisy measurements for the layered material.

9. CONCLUSION

In this work, we have proposed a particle-based approach with local RB surrogate model to approximate the Gibbs posterior for inverse problems. The Gibbs posterior has a particular advantage over the usual Bayesian approach, in the sense that it does not require an explicit model of the data generating mechanism (i.e., a likelihood function). The Gibbs posterior is

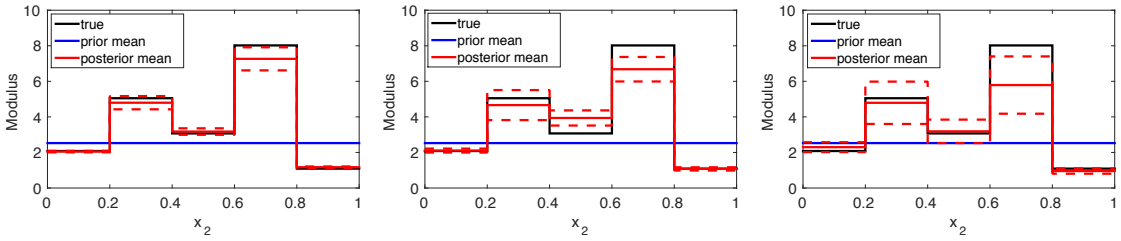


(A) Horizontal displacement

(B) Vertical displacement

(C) Measurement of the vertical displacement with 10% Gaussian noise

FIGURE 15. 2D elasticity equation: The displacement fields and the noisy measurements for the material with a hard inclusion.

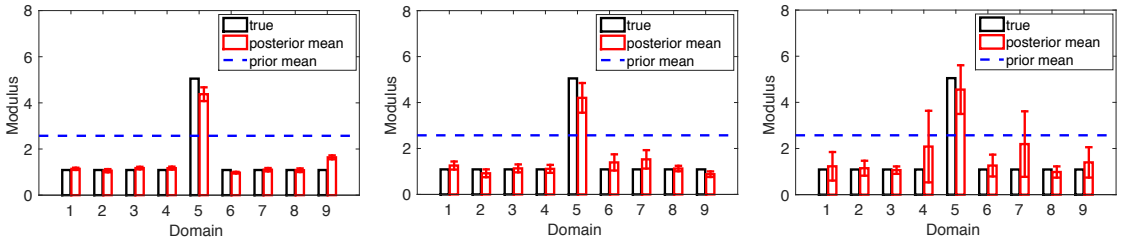


(A) 5% noise

(B) 10% noise

(C) 20% noise

FIGURE 16. 2D elasticity equation: The mean and standard deviation of Gibbs posterior computed using data with different levels of noise for the layered material.



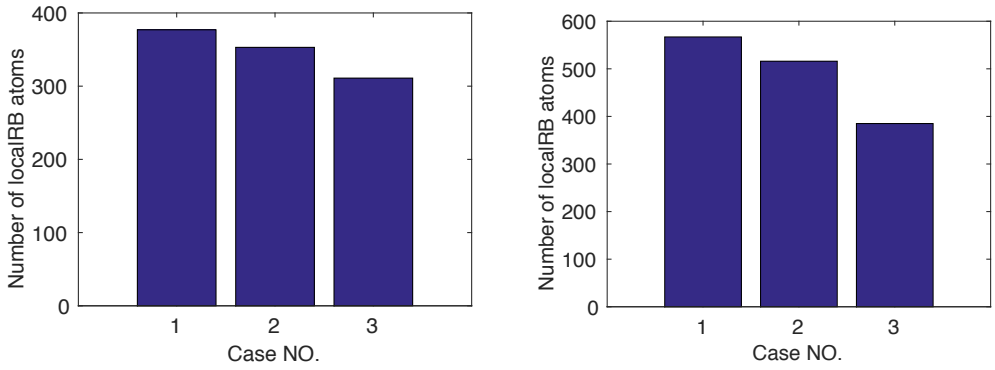
(A) 5% noise

(B) 10% noise

(C) 20% noise

FIGURE 17. 2D elasticity equation: The mean and standard deviation of Gibbs posterior computed using data with different levels of noise for the material with a hard inclusion.

applicable where the unknown parameters are connected to the data through a loss function. It provides a more general framework for updating belief distributions where the true data generating mechanism is unknown or difficult to specify. We employed the local RB method to approximate the loss function in the Gibbs update formula. Based on a Sequential Monte Carlo



(A) Layered material inversion (B) Material with a hard inclusion

FIGURE 18. 2D elasticity equation: The total number of local RB atoms upon solving the Gibbs posterior for both material models and with different levels of noise (Case NO.).

(SMC) framework, we presented a method to progressive approximate the Gibbs posterior by simultaneously evolving the particles and adapting the local RB surrogate model in a sequential manner. The emphasis of the local RB surrogate is navigated to a small fraction of the parameter space automatically by the evolving particles that progressively cluster over the support of the posterior. Computational savings are achieved thanks to the local accuracy and the efficiency of our local RB method. Indeed, once the local RB surrogate becomes accurate enough (specified by a parameter representing the approximation accuracy) over the local support of the posterior, further evolution of the particles takes minimal cost. Through several numerical examples that include advection-diffusion problems and elasticity imaging problems, we demonstrated the consistency of our method with the state-of-art Markov chain Monte Carlo (MCMC) method. Furthermore, we showed that significant computational savings can be achieved to approximate the Gibbs posterior using our proposed method.

Acknowledgements
HA was partially supported by NSF grants DMS-2110263 and DMS-1913004 and Air Force Office of Scientific Research under Award NO: FA9550-19-1-0036 and FA9550-22-1-0248 . WA and ZZ were partially supported by DARPA EQUiPS program, grant SNL 014150709. SM was partially supported by NSF grants DMS 1613261, DEB-1840223, DMS 1713012 and HFSP RGP0051/2017.

APPENDIX A. PROOFS TO THE MAIN RESULTS

Proof to Theorem 3.1. First, note that for $\forall\ \xi \in \Xi_e$, we have that $-e \leq l(\xi) - \bar{l}(\xi) \leq e$. For e sufficiently small, e.g., $We \ll 1$, we have

$$|\exp(-Wl(\xi)) - \exp(-W\bar{l}(\xi))| \leq \exp(-Wl(\xi))|1 - \exp(Wl(\xi) - W\bar{l}(\xi))| \leq CWe$$

for some $C > 0$. Let

$$Z_1 = \int_\Xi \exp(-Wl(\xi))\rho_0(\xi)d\xi, \quad Z_2 = \int_\Xi \exp(-W\bar{l}(\xi))\rho_0(\xi)d\xi. \tag{A.1}$$

Using Assumption 1 and the fact that $|f|_\infty \leq 1$, we have

$$Z_1 = \int_\Xi \exp(-Wl(\xi))\rho_1(\xi)d\xi \geq \exp(-WC_l),$$
$$Z_2 \geq \int_\Xi \exp(-W\bar{l}(\xi))\rho_2(\xi)|f(\xi)|d\xi.$$

In addition, we have

$$\begin{aligned}
|Z_1 - Z_2| &\leq \int_{\Xi_e} |\exp(-Wl(\xi)) - \exp(-W\bar{l}(\xi))|\rho_0(\xi)d\xi \\
&+ \int_{\Xi_e^\perp} |1 - \exp(Wl(\xi) - W\bar{l}(\xi))|\exp(-Wl(\xi))\rho_0(\xi)d\xi \\
&\leq CWe + \int_{\Xi_e^\perp} |1 - \exp(Wl(\xi) - W\bar{l}(\xi))|Z_1\rho(\xi)d\xi \\
&\leq CWe + \exp(W\max\{C_l, C_{\bar{l}}\})\rho[\mathbb{1}_{\Xi_e^\perp}].
\end{aligned}$$

Hence

$$\begin{aligned}
|\rho[f] - \overline{\rho}[f]| &= \left| \frac{\int_\Xi \exp(-Wl(\xi))\rho_0(\xi)f(\xi)d\xi}{\int_\Xi \exp(-Wl(\xi))\rho_0(\xi)d\xi} - \frac{\int_\Xi \exp(-W\bar{l}(\xi))\rho_0(\xi)f(\xi)d\xi}{\int_\Xi \exp(-W\bar{l}(\xi))\rho_0(\xi)d\xi} \right| \\
&\leq \frac{\int_\Xi |\exp(-Wl(\xi)) - \exp(-W\bar{l}(\xi))|\rho_0(\xi)|f(\xi)|d\xi}{Z_1} \\
&+ \frac{|Z_2 - Z_1| \int_\Xi \exp(-W\bar{l}(\xi))\rho_2(\xi)|f(\xi)|d\xi}{Z_1 Z_2} \\
&\leq \frac{CWe + \exp(W\max\{C_l, C_{\bar{l}}\})\rho[\mathbb{1}_{\Xi_e^\perp}]}{Z_1} + \frac{|Z_2 - Z_1|}{Z_1} \\
&\leq 2\exp(WC_l)CWe + 2\exp(WC_l + W\max\{C_l, C_{\bar{l}}\})\rho[\mathbb{1}_{\Xi_e^\perp}].
\end{aligned}$$

This completes the proof.

□

Proof to Lemma 5.1. For any f, we have

$$\rho_0^E[f] = \sum_{i=1}^m w_i^0 f(\xi_i).$$

Hence,

$$\rho_0^E[f] - \rho_0[f] = \sum_{i=1}^m w_i^0 (f(\xi_i) - \rho_0[f]).$$

Note that since the particles ξ_i are i.i.d. with distribution $\rho_0(\xi)$, we have $\mathbb{E}[f(\xi_i)] = \rho_0[f]\ \forall\ i = 1, \ldots, m$. Hence

$$\mathbb{E}[(f(\xi_i) - \rho_0[f])(f(\xi_j) - \rho_0[f])] = \delta_{ij}\mathbb{E}[|(f(\xi_i) - \rho_0[f])|^2].$$

Additionally, since $|f|_\infty \leq 1$, we have

$$\mathbb{E}[|(f(\xi_i) - \rho_0[f])|^2] = \mathbb{E}[|f(\xi_i)|^2] - \rho_0[f]^2 \leq 1.$$

Therefore,

$$\mathbb{E}[|\rho_0^E[f] - \rho_0[f]|^2] = \sum_{i=1}^m (w_i^0)^2 \mathbb{E}[|(f(\xi_i) - \rho_0[f])|^2] \leq \sum_{i=1}^m (w_i^0)^2,$$

which completes the proof.

□

Proof to Lemma 5.2. We have

$$\begin{aligned} G_W\rho_1[f] - G_W\rho_2[f] &= \frac{\int_\Xi \exp(-Wl(\xi))\rho_1(\xi)f(\xi)d\xi}{\int_\Xi \exp(-Wl(\xi))\rho_1(\xi)d\xi} - \frac{\int_\Xi \exp(-Wl(\xi))\rho_2(\xi)f(\xi)d\xi}{\int_\Xi \exp(-Wl(\xi))\rho_2(\xi)d\xi} \\ &= \frac{\int_\Xi \exp(-Wl(\xi))(\rho_1(\xi)-\rho_2(\xi))f(\xi)d\xi}{Z_3} \\ &+ \frac{(Z_4-Z_3)\int_\Xi \exp(-Wl(\xi))\rho_2(\xi)f(\xi)d\xi}{Z_3Z_4} \end{aligned}$$

where

$$Z_3 = \int_\Xi \exp(-Wl(\xi))\rho_1(\xi)d\xi, \quad Z_4 = \int_\Xi \exp(-Wl(\xi))\rho_2(\xi)d\xi. \tag{A.2}$$

Hence

$$\begin{aligned} |G_W\rho_1[f] - G_W\rho_2[f]| &\le \frac{|\int_\Xi \exp(-Wl(\xi))(\rho_1(\xi)-\rho_2(\xi))f(\xi)d\xi|}{Z_3} \\ &+ \frac{|Z_4-Z_3||\int_\Xi \exp(-Wl(\xi))\rho_2(\xi)f(\xi)d\xi|}{Z_3Z_4}. \end{aligned}$$

Note that since $|f|_\infty \le 1$,

$$\left|\int_\Xi \exp(-Wl(\xi))\rho_2(\xi)f(\xi)d\xi\right| \le \int_\Xi \exp(-Wl(\xi))\rho_2(\xi)d\xi = Z_4$$

Using Assumption 1, we have

$$Z_3 = \int_\Xi \exp(-Wl(\xi))\rho_1(\xi)d\xi \ge \exp(-WC_l)$$

hence we have

$$|G_W\rho_1[f] - G_W\rho_2[f]| \le \frac{|\int_\Xi \exp(-Wl(\xi))(\rho_1(\xi)-\rho_2(\xi))f(\xi)d\xi|}{\exp(-WC_l)} + \frac{|Z_4-Z_3|}{\exp(-WC_l)}.$$

Note that $|\exp(-Wl(\xi))|_\infty \le 1$ and $|\exp(-Wl(\xi))f(\xi)|_\infty \le 1$ for $\forall f$ such that $|f|_\infty \le 1$, hence

$$\left|\int_\Xi \exp(-Wl(\xi))(\rho_1(\xi)-\rho_2(\xi))f(\xi)d\xi\right| \le \sup_{|g|_\infty\le 1} |\rho_1[g]-\rho_2[g]|$$

and

$$|Z_4-Z_3| = \left|\int_\Xi (\rho_1(\xi)-\rho_2(\xi))\exp(-Wl(\xi))d\xi\right| \le \sup_{|g|_\infty\le 1} |\rho_1[g]-\rho_2[g]|.$$

Therefore,

$$|G_W\rho_1[f] - G_W\rho_2[f]| \le 2\exp(WC_l) \sup_{|g|_\infty\le 1} |\rho_1[g]-\rho_2[g]|$$

for $\forall f$ such that $|f|_\infty \le 1$. The lemma follows easily from the above inequality. □

Proof to Lemma 5.4. Based on Assumption 2, we have

$$\frac{\exp(-W\bar{l}(\xi_i))}{\exp(-Wl(\xi_i))} = \exp(Wl(\xi_i) - W\bar{l}(\xi_i)) \leq \exp(We) \ \ \forall\, i = 1,\dots,m$$
$$\frac{\exp(-Wl(\xi_i))}{\exp(-W\bar{l}(\xi_i))} = \exp(W\bar{l}(\xi_i) - Wl(\xi_i)) \leq \exp(We) \ \ \forall\, i = 1,\dots,m.$$

We first define

$$Z_5 = \sum_{k=1}^{m} w_k^0 \exp(-Wl(\xi_k)), \quad Z_6 = \sum_{k=1}^{m} w_k^0 \exp(-W\bar{l}(\xi_k)). \tag{A.3}$$

It is clear that

$$\log\left(\frac{Z_5}{Z_6}\right) \leq \log\left(\frac{\exp(We)\sum_{k=1}^m w_k^0 \exp(-W\bar{l}(\xi_k))}{\sum_{k=1}^m w_k^0 \exp(-W\bar{l}(\xi_k))}\right) = We.$$

Hence

$$\begin{aligned} D_{KL}(\overline{\rho}^E|\rho^E) &= \sum_{k=1}^{m} \frac{\exp(-W\bar{l}(\xi_k))w_k^0}{Z_6} \log\left(\frac{\exp(-W\bar{l}(\xi_k))w_k^0 Z_5}{\exp(-Wl(\xi_k))w_k^0 Z_6}\right) \\ &\leq \sum_{k=1}^{m} \frac{\exp(-W\bar{l}(\xi_k))w_k^0}{Z_6} \left[\log\left(\frac{Z_5}{Z_6}\right) + We\right] \\ &= \log\left(\frac{Z_5}{Z_6}\right) + We = 2We. \end{aligned}$$

□

Proof to Lemma 5.5. Recall the definition of Z_5 and Z_6 in (A.3), we have

$$\begin{aligned} \rho^E[f] - \overline{\rho}^E[f] &= \frac{\sum_{j=1}^m \exp(-Wl(\xi_j))w_j^0 f(\xi_j)}{\sum_{k=1}^m \exp(-Wl(\xi_k))w_k^0} - \frac{\sum_{j=1}^m \exp(-W\bar{l}(\xi_j))w_j^0 f(\xi_j)}{\sum_{k=1}^m \exp(-W\bar{l}(\xi_k))w_k^0} \\ &= \frac{\sum_{j=1}^m (\exp(-Wl(\xi_j)) - \exp(-W\bar{l}(\xi_j)))w_j^0 f(\xi_j)}{Z_5} \\ &\quad + \frac{(Z_6 - Z_5)\sum_{j=1}^m \exp(-W\bar{l}(\xi_j))w_j^0 f(\xi_j)}{Z_5 Z_6}. \end{aligned}$$

Note that

$$Z_5 = \sum_{k=1}^{m} \exp(-Wl(\xi_k))w_k^0 \geq \exp(-WC_l),$$

and that

$$\left|\sum_{j=1}^{m} \exp(-W\bar{l}(\xi_j))w_j^0 f(\xi_j)\right| \leq Z_6,$$

since $|f|_\infty \leq 1$. Also, since $-e \leq l(\xi_j) - \bar{l}(\xi_j) \leq e$, for e sufficiently small, e.g., $We \ll 1$, we have

$$|\exp(-Wl(\xi_j)) - \exp(-W\bar{l}(\xi_j))| \leq \exp(-Wl(\xi_j))|1 - \exp(Wl(\xi_j) - W\bar{l}(\xi_j))| \leq CWe$$

for some $C > 0$. Hence

$$|Z_5 - Z_6| \leq CWe.$$

Finally, we have

$$|\rho^E[f] - \overline{\rho}^E[f]| \leq 2\exp(WC_l)CWe,$$

which implies the Lemma.

□

Proof to Theorem 6.1. First, by Equation (6.5), (6.6) and the fact that $\rho_{t+1} = \rho_{t+1}\mathscr{K}_{t+1}$, we have

$$\begin{aligned} h(\rho^E_{t+1}, \rho^E_{t+1,t}) &= h(\rho^{E,S}_{t+1,t}\mathscr{K}_{t+1}, \rho^E_{t+1,t}) \leq h(\rho^{E,S}_{t+1,t}\mathscr{K}_{t+1}, \rho^E_{t+1,t}\mathscr{K}_{t+1}) + h(\rho^E_{t+1,t}\mathscr{K}_{t+1}, \rho^E_{t+1,t}) \\ &\leq h(\rho^{E,S}_{t+1,t}, \rho^E_{t+1,t}) + h(\rho^E_{t+1,t}\mathscr{K}_{t+1}, \rho_{t+1}\mathscr{K}_{t+1}) + h(\rho_{t+1}, \rho^E_{t+1,t}) \\ &\leq \frac{1}{\sqrt{m}} + 2h(\rho_{t+1}, \rho^E_{t+1,t}). \end{aligned}$$

Hence

$$\begin{aligned} h(\rho^E_{t+1}, \rho_{t+1}) &\leq h(\rho^E_{t+1}, \rho^E_{t+1,t}) + h(\rho_{t+1}, \rho^E_{t+1,t}) \leq \frac{1}{\sqrt{m}} + 3h(\rho_{t+1}, \rho^E_{t+1,t}) \\ &\leq \frac{1}{\sqrt{m}} + 6\exp((W_{t+1} - W_t)C_l)h(\rho^E_t, \rho_t) \end{aligned}$$

by Equation (6.4). Iterating gives

$$h(\rho^E_{t+1}, \rho_{t+1}) \leq \frac{1}{\sqrt{m}} \sum_{s=0}^{t+1} 6^{t+1-s}\exp((W_{t+1} - W_s)C_l),$$

which completes the proof.

□

REFERENCES

[1] Pierre Alquier, James Ridgway, and Nicolas Chopin. On the properties of variational approximations of Gibbs posteriors. *The Journal of Machine Learning Research*, 17(1):8374–8414, 2016.

[2] Andrew F Bennett. *Inverse modeling of the ocean and atmosphere*. Cambridge University Press, 2005.

[3] JM Bernardo, MJ Bayarri, JO Berger, AP Dawid, D Heckerman, AFM Smith, and M West. Sequential Monte Carlo for Bayesian computation. In *Bayesian Statistics 8: Proceedings of the Eighth Valencia International Meeting, June 2-6, 2006*, volume 8, page 115. Oxford University Press, USA, 2007.

[4] Alexandros Beskos, Ajay Jasra, Ege A Muzaffer, and Andrew M Stuart. Sequential Monte Carlo methods for Bayesian elliptic inverse problems. *Statistics and Computing*, 25(4):727–737, 2015.

[5] Pier Giovanni Bissiri, Chris C Holmes, and Stephen G Walker. A general framework for updating belief distributions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(5):1103–1130, 2016.

[6] Tan Bui-Thanh, Omar Ghattas, James Martin, and Georg Stadler. A computational framework for infinite-dimensional Bayesian inverse problems Part I: The linearized case, with application to global seismic inversion. *SIAM Journal on Scientific Computing*, 35(6):A2494–A2523, 2013.

[7] David Colton, Michele Piana, and Roland Potthast. A simple method using Morozov's discrepancy principle for solving inverse scattering problems. *Inverse Problems*, 13(6):1477, 1997.

[8] Patrick R Conrad, Youssef M Marzouk, Natesh S Pillai, and Aaron Smith. Accelerating asymptotically exact MCMC for computationally intensive models via local approximations. *Journal of the American Statistical Association*, 111(516):1591–1607, 2016.

[9] Patrick Raymond Conrad. *Accelerating Bayesian inference in computationally expensive computer models using local and global approximations*. PhD thesis, Massachusetts Institute of Technology, 2014.

[10] Simon L Cotter, Massoumeh Dashti, and Andrew M Stuart. Approximation of Bayesian inverse problems for PDEs. *SIAM Journal on Numerical Analysis*, 48(1):322–345, 2010.

[11] Simon L Cotter, Gareth O Roberts, Andrew M Stuart, and David White. MCMC methods for functions: modifying old algorithms to make them faster. *Statistical Science*, pages 424–446, 2013.

[12] Tiangang Cui, Youssef M Marzouk, and Karen E Willcox. Data-driven model reduction for the Bayesian solution of inverse problems. *International Journal for Numerical Methods in Engineering*, 102(5):966–990, 2015.

[13] Pierre Del Moral, Arnaud Doucet, and Ajay Jasra. Sequential Monte Carlo samplers. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(3):411–436, 2006.

[14] Arnaud Doucet, Nando De Freitas, and Neil Gordon. An introduction to sequential Monte Carlo methods. In *Sequential Monte Carlo Methods in Practice*, pages 3–14. Springer, 2001.

[15] Isabell M Franck and PS Koutsourelakis. Sparse Variational Bayesian approximations for nonlinear inverse problems: Applications in nonlinear elastography. *Computer Methods in Applied Mechanics and Engineering*, 299:215–244, 2016.

[16] M Frangos, Y Marzouk, K Willcox, and B van Bloemen Waanders. Surrogate and reduced-order modeling: a comparison of approaches for large-scale statistical inverse problems [chapter 7]. 2010.

[17] Dani Gamerman and Hedibert F Lopes. *Markov chain Monte Carlo: stochastic simulation for Bayesian inference*. Chapman and Hall/CRC, 2006.

[18] M Grigoriu. Reduced order models for random functions. application to stochastic problems. *Applied Mathematical Modelling*, 33(1):161–175, 2009.

[19] Mircea Grigoriu. A method for solving stochastic equations by reduced order models and local approximations. *Journal of Computational Physics*, 231(19):6495–6513, 2012.

[20] Nikolas Kantas, Alexandros Beskos, and Ajay Jasra. Sequential Monte Carlo Methods for High-Dimensional Inverse Problems: A Case Study for the Navier–Stokes Equations. *SIAM/ASA Journal on Uncertainty Quantification*, 2(1):464–489, 2014.

[21] Marc C Kennedy and Anthony O'Hagan. Bayesian calibration of computer models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(3):425–464, 2001.

[22] Phaedon-Stelios Koutsourelakis. A novel Bayesian strategy for the identification of spatially varying material properties and model validation: an application to static elastography. *International Journal for Numerical Methods in Engineering*, 91(3):249–268, 2012.

[23] Kody JH Law. Proposals which speed up function-space MCMC. *Journal of Computational and Applied Mathematics*, 262:127–138, 2014.

[24] Jinglai Li and Youssef M Marzouk. Adaptive construction of surrogates for the Bayesian solution of inverse problems. *SIAM Journal on Scientific Computing*, 36(3):A1163–A1186, 2014.

[25] Andrea Manzoni, Stefano Pagani, and Toni Lassila. Accurate solution of Bayesian inverse uncertainty quantification problems combining reduced basis methods and reduction error models. *SIAM/ASA Journal on Uncertainty Quantification*, 4(1):380–412, 2016.

[26] Youssef Marzouk and Dongbin Xiu. A stochastic collocation approach to Bayesian inference in inverse problems. 2009.

[27] Youssef M Marzouk and Habib N Najm. Dimensionality reduction and polynomial chaos acceleration of bayesian inference in inverse problems. *Journal of Computational Physics*, 228(6):1862–1902, 2009.

[28] Youssef M Marzouk, Habib N Najm, and Larry A Rahn. Stochastic spectral methods for efficient Bayesian solution of inverse problems. *Journal of Computational Physics*, 224(2):560–586, 2007.

[29] Patrick Rebeschini, Ramon Van Handel, et al. Can local particle filters beat the curse of dimensionality? *The Annals of Applied Probability*, 25(5):2809–2866, 2015.

[30] Otmar Scherzer. The use of Morozov's discrepancy principle for Tikhonov regularization for solving nonlinear ill-posed problems. *Computing*, 51(1):45–60, 1993.

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

[32] Nicholas Syring and Ryan Martin. Robust and rate-optimal Gibbs posterior inference on the boundary of a noisy image. *arXiv preprint arXiv:1606.08400*, 2016.

[33] Jingbo Wang and Nicholas Zabaras. A Bayesian inference approach to the inverse heat conduction problem. *International Journal of Heat and Mass Transfer*, 47(17-18):3927–3941, 2004.

[34] Zilong Zou, Drew Kouri, and Wilkins Aquino. An Adaptive Sampling Approach for Solving PDEs with Uncertain Inputs and Evaluating Risk. In *19th AIAA Non-Deterministic Approaches Conference*, page 1325, 2017.

[35] Zilong Zou, Drew Kouri, and Wilkins Aquino. An adaptive local reduced basis method for solving PDEs with uncertain inputs and evaluating risk. *Computer Methods in Applied Mechanics and Engineering*, 2018.