Sequential Ensemble Transform for Bayesian Inverse Problems

Aaron Myers *, Alexandre H. Thiery §, Kainan Wang †, Tan Bui-Thanh *‡

September 23, 2019

Abstract

We present the Sequential Ensemble Transform (SET) method, a new approach for generating approximate samples from a Bayesian posterior distribution. The method explores the posterior by solving a sequence of discrete linear optimal transport problems to produce a series of transport plans which map prior samples to posterior samples. We show that the sequence of Dirac mixture distributions produced by the SET method converges weakly to the true posterior as the sample size approaches infinity. Our numerical results indicate that, as opposed to more standard Sequential Monte Carlo (SMC) methods used for inference in Bayesian inverse problems, the SET approach is more robust to the choice of Markov mutation kernel steps.

Keywords: Optimal transport, Bayesian inverse problems, Particle methods.

1 Introduction

Inverse problems enable integration of observational and experimental data, simulations and/or mathematical models to make scientific predictions. We focus on inverse problems in which the goal is to determine a parameter of interest from indirect and imprecise observations. The relationship between the parameter and the noise-free observations, the forward map, is often provided through the solution of a complex mathematical model—the forward problem.

The Bayesian approach formulates the inverse problem as a statistical inference problem [MT95, Stu10, KS06]. Given noisy observational data, the governing forward problem, and a prior probability distribution, the solution of the Bayesian inverse problem is the posterior probability distribution over the parameters. The prior distribution encodes knowledge or assumptions about the parameter space before data are observed. The posterior distribution incorporates both the prior knowledge and the observations. Non-linearity of the forward map leads to posterior distributions that are typically not Gaussian, even in situations when both the prior and observational noise probability distributions are Gaussian.

Exploring a high dimensional non-Gaussian posterior is computationally challenging. Indeed, evaluating the posterior density typically requires evaluating the forward map which, for problems governed by partial differential equations (PDEs), dominates the computational cost. Standard numerical quadrature methods routinely used for estimating statistical quantities of interest (e.g. statistical moments, probability of rare event, etc.) are infeasible in these high-dimensional settings.

The Markov chain Monte Carlo (MCMC) algorithm [Has70, MRR+53] is a popular approach for exploring the posterior distribution in Bayesian inverse problems. Estimates obtained from standard MCMC methods often require a large number of samples to be meaningful, especially in high dimensional settings. In Bayesian inverse problems, generating each MCMC sample requires an evaluation of the posterior density, which relies on evaluating the computationally expensive forward map.

Sequential Monte Carlo (SMC) methods are computational techniques widely used in engineering, statistics, and many other fields [GSS93, DDFG01, Del04, DJ09, DMDJ06] to approximate a sequence of probability distributions,
usually of increasing complexity or dimension. A standard approach in Bayesian inverse problems consists of introducing a sequence of distributions that interpolates between a distribution that is easy to sample from (e.g. the prior distribution, or a Gaussian approximation of the posterior distribution) and the posterior distribution. Through a combination of importance sampling, Markovian mutations and resampling procedures, the SMC method iteratively constructs a sequence of particle approximations of this sequence of distributions. Under very mild assumptions, SMC methods are consistent in the limit when the number \( N \) of particles goes to infinity and converge at Monte-Carlo rate \( O(N^{-1/2}) \). Furthermore, methods are available for implementing this class of algorithms on parallel architectures [WLH+16, VDDMM15, LW16, ST19].

In this article, inspired by recent developments in the data-assimilation literature [Rei13, CR13], we exploit algorithms based on the concept of optimal transport [Mon81, Vil08, Vil03, PC+19]. Our approach, the Sequential Ensemble Transform (SET) method, combines the SMC framework with the use of optimal transport to efficiently build particle approximations of the posterior distribution in high-dimensional Bayesian inverse problems (see figure 1). We refer the readers to [MM12, HDP15, PM14, SBM18] for other Monte-Carlo methods based on transportation concepts. Unlike SMC methods, the SET approach, similar to the algorithm of [Rei13], uses an optimal transport scheme instead of the usual resampling procedure. The main advantage of the proposed method is its robustness with respect to the choice of mutation kernel steps. Indeed, without mutation kernel, the SMC method is a variant of the standard importance sampling procedure, which is known to behave poorly in high-dimensional settings [BBL+08]. Consequently, good mutation kernels are often crucial to the successful implementation of SMC methods in Bayesian inverse problems [BCJ14]. Unfortunately, it is notoriously difficult to design Markov kernels with good mixing properties in high-dimensional settings that are common in Bayesian inverse problems [BTGMS13b, BJMS15, KBJ14]. Adaptive SMC procedures [Cho02, DMDJ12, JSDT11, BJKT15] can help mitigate this issue by automatically tuning the mutation kernels and the interpolating sequence of distributions. Our numerical studies presented in Section 6 show that the SET approach performs favorably, even in the extreme case when no mutation kernel is employed, when compared to more standard SMC methods. Furthermore, although approximate methods [GCPB16, Cut13] are available for efficiently solving discrete optimal transport problems, we have found that in most realistic Bayesian inverse problems and for a typical number of particles \( N \lesssim 10^4 \), the computational cost of (exactly) solving the discrete optimal transport problems is negligible when compared to the computational burden associated with the forward-solves necessary to implement the SET/SMC algorithms. Finally, it should be mentioned that in situations when the design of Markov kernels with good mixing properties is not an issue, our proposed method typically does not bring significant computational savings over standard SMC or MCMC methods. That is for example the case when the dimension of the posterior distribution is low.

![Figure 1: A representation of the SET method using optimal transport to move particles in parameter space as to represent the posterior](image)

The article is structured as follows. In Section 2, PDE-constrained Bayesian inverse problems are briefly introduced. An overview of particle methods and importance sampling is presented in Section 3. Section 5 introduces the concept of optimal transport and describes the SET method, as well as its asymptotic properties. Finally, Section 6 presents various numerical results including a Gaussian case and an inverse problem with non-linear forward maps. Section 7 concludes the paper and discusses future work.
Notations and conventions

Unless stated otherwise, all the state spaces are endowed with a metric and the associated Borel $\sigma$-algebra. For a probability distribution $\mu$ on the state space $\mathcal{X}$ and a $\mu$-integrable test function $\varphi : \mathcal{X} \to \mathbb{R}$, we make use of the notation $\mu(\varphi) = \int \varphi(x) \mu(dx)$. Similarly, for a Markov kernel $M(u, dv)$, we have that $(M \varphi)(u) \equiv \int \varphi(v) M(u, dv)$. The Markov kernel $M$ has the Feller property if $M \varphi$ is continuous when $\varphi$ is continuous and bounded. A sequence of probability distributions $\{\mu_N\}_{N \geq 1}$ on $\mathcal{X}$ converges weakly towards the distribution $\mu$, denoted as $\mu_N \to^{\text{w}} \mu$, if for any bounded and continuous test function $\varphi : \mathcal{X} \to \mathbb{R}$ we have that $\mu_N(\varphi) \to \mu(\varphi)$ as $N \to \infty$. Similarly, a sequence of random probability distribution $\mu^N_\omega$ almost surely converges weakly towards $\mu$ if, for $\mathbb{P}$-almost every $\omega$, we have that $\mu^N_\omega \to^{\text{w}} \mu$. The set of probability distributions on a state space $\mathcal{X}$ is denoted as $\mathcal{P}(\mathcal{X})$. For a set $S$, the notation $1_S$ refers to the indicator function of $S$, i.e., the function that equals one for $x \in S$ and zero otherwise. For $u \in \mathcal{X}$, the Dirac probability distribution $\delta(u)$ is the distribution that puts all its probability mass at $u$.

2 Problem Statement

Although the methods described in this article are general, for illustration purposes, we focus on the following prototype inverse problem. For an open and bounded domain $\Omega \subset \mathbb{R}^d$, with $d \in \{1, 2, 3\}$, consider the elliptic PDE

\[
\begin{aligned}
-\nabla \cdot (e^u \nabla w) &= 0 \quad \text{in } \Omega \subset \mathbb{R}^d, \\
-e^u \nabla w \cdot n &= B_i \cdot w \quad \text{on } \partial \Omega \setminus \Gamma_R, \\
-e^u \nabla w \cdot n &= -1 \quad \text{on } \Gamma_R,
\end{aligned}
\]  

(1)

where $w$ is the state variable (e.g. temperature field), $u$ the parameter of interest (e.g. logarithm of the thermal conductivity) on $\Omega$, $n$ the unit outward normal on $\partial \Omega$, and $B_i$ a constant (e.g. the Biot number).

The forward problem consists of computing the field $w$ given a description of the parameter field $u$. The inverse problem is the task of reconstructing the field $u$, and the associated statistical uncertainties, given some (possibly non-linear) noisy and incomplete observations of the field $w$. A standard setup is when the temperature field $w$ is only discretely observed at a finite set of locations inside the domain $\Omega$. For simplicity, assume the following additive noise-corrupted point-wise observation model,

\[d_j \equiv w(x_j) + \eta_j, \quad j = 1, \ldots, D\]

(2)

where $\{x_j\}_{j=1}^D$ denotes the set of points at which the field $w$ is observed, $\eta_j$ the additive noise, and $d_j$ the actual noisy observations. Concatenating all the observations, Equation (2) can be succinctly expressed as

\[d \equiv \mathcal{G}(u) + \eta.\]

(3)

The quantity $\mathcal{G} \equiv [w(x_1), \ldots, w(x_D)]^T$ denotes the mapping from the parameters to observables and the random variable $\eta$ is assumed to be a Gaussian with mean zero and bounded covariance matrix $L$, i.e. $\eta \sim \mathcal{N}(0, L)$. The vector $d = [d_1, \ldots, d_D]^T$ summarizes the observed data. To keep the exposition as simple as possible, although the SET methodology straightforwardly extends to much more general settings, we postulate a Gaussian prior measure on the parameter $u$,

\[\mu_{\text{prior}} = \mathcal{N}(u_0, C),\]

with mean $u_0$ and covariance operator,

\[C \equiv (\delta I - \gamma \Delta)^{-s} \equiv A^{-s}\]

(4)

where regularization parameters $\delta > 0$ and $\gamma > 0$ control the variance and correlation length of the covariance operator, respectively. The operator $A$ is well-defined on its domain

\[\mathcal{D}(A) \equiv \left\{ u \in H^2(\Omega) : \frac{\partial u}{\partial n} = 0 \text{ on } \partial \Omega \right\}.\]
where $H^2(\Omega)$ is the usual Sobolev space. If $u_0$ lives in the Cameron-Martin space of $C$ and $s > 1$, the prior distribution $\mu_{\text{prior}}$ is well-defined and realizations from it are almost surely in the Hölder space $\mathcal{X} \equiv C^{0,\beta}(\Omega)$ with $0 < \beta < s/2$ [Stu10]. Additional motivation for the selection of this prior can be found in [Stu10, BTGMS13a].

With these modeling assumptions, the Bayesian posterior measure $\mu_{\text{post}}$ is well-defined and given by the change of measure formula

$$\frac{d\mu_{\text{post}}}{d\mu_{\text{prior}}}(u) \propto \exp \left\{ -\frac{1}{2} |d - G(u)|_L^2 \right\} .$$  (5)

Here, $|\cdot|_L \equiv |\cdot|_{L^{-\frac{1}{2}}}$ denotes the $L^{-\frac{1}{2}}$-weighted Euclidean norm.

## 3 Particle Methods

Particle methods approximate probability distributions with weighted mixtures of Diracs. To construct a particle approximation of the posterior distribution, the SMC and SET approaches proceed by introducing a sequence $\{\mu_k\}_{k=0}^K$ of distributions that interpolates between a distribution that is easy to sample from, i.e. $\mu_0$, and the posterior distribution $\mu_K$. A standard choice for $\mu_0$ is the prior distribution, or a Gaussian approximation of the posterior distribution obtained through efficient deterministic methods. For any index $1 \leq k \leq K$, set

$$\frac{d\mu_k}{d\mu_{k-1}}(u) = \frac{1}{Z_k} \Psi_k(u),$$  (6)

for a $\mu_k$-integrable potential function $\Psi_k : \mathcal{X} \rightarrow (0, \infty)$ and (typically unknown) normalization constant $Z_k > 0$.

The SMC algorithm recursively constructs particle approximations

$$\mu_k^N = \frac{1}{N} \sum_{i=1}^{N} \delta(u_{i,k}^N) \approx \mu_k,$$

where $N \geq 1$ denotes the number of particles, by iterating re-weighting, resampling, and mutation operations.

### 3.1 Re-weighting

Consider two probability distributions $\mu$ and $\nu$ defined on the same state space $\mathcal{X}$ and related by a change of measure (Radon-Nikodym derivative),

$$\frac{d\nu}{d\mu}(u) = \frac{1}{Z} \Psi(u)$$  (7)

for a $\mu$-integrable potential function $\Psi : \mathcal{X} \rightarrow (0, \infty)$ and a possibly unknown normalization constant $Z > 0$.

Suppose that, for any integer $N \geq 1$, it is possible to generate a set of $N$ particles $\{u_{i}^N\}_{i=1}^{N} \subset \mathcal{X}$ such that the sequence of equally weighted particle approximations,

$$\mu^N \equiv \frac{1}{N} \sum_{i=1}^{N} \delta(u_{i}^N),$$

converges weakly towards $\mu$ as $N \rightarrow \infty$. Under mild assumptions, the sequence of self-normalized importance sampling weighted particle approximations $\nu^N$ defined as

$$\nu^N \equiv \sum_{i=1}^{N} w_{i}^N \delta(u_{i}^N)$$  (8)

for normalized weights

$$w_{i}^N \equiv \frac{\Psi(u_{i}^N)}{[\Psi(u_{1}^N) + \ldots + \Psi(u_{N}^N)]}$$
converges weakly to $\nu$. For concreteness, define the mapping from $\mu^N$ to $\nu^N$ as $\nu^N = \mathcal{B}_\Psi(\mu^N)$ where $\mathcal{B}_\Psi$ is the so-called Bayes operator that transforms a probability distribution $\mu$ into the probability distribution $\mathcal{B}_\Psi(\mu)$ that satisfies $\mathcal{B}_\Psi(\mu)(\phi) = \mu(\Psi \phi)/\mu(\Psi)$ for any test function $\phi$. The following proposition shows that, under a mild uniform integrability condition, the convergence $\mathcal{B}_\Psi(\mu^N) \xrightarrow{\text{w}} \mathcal{B}_\Psi(\mu)$ is guaranteed.

**Proposition 3.1.** Consider a probability distribution $\mu$ and a continuous and positive $\mu$-integrable function $\Psi$. Assume that there exists a continuous $\mu$-integrable function $V : \mathcal{X} \to [1, \infty)$ such that

$$\lim_{t \to \infty} \limsup_{N \to \infty} \mu^N(V \times 1_{V > t}) = 0,$$

and $\Psi(u) \leq V(u)$ for $\mu$-almost every $u \in \mathcal{X}$. We have that:

1. for any (potentially unbounded) continuous test function $\phi$ such that $|\phi| \leq V$,

$$\lim_{N \to \infty} \mu^N(\phi) = \mu(\phi).$$

2. the sequence $\mathcal{B}_\Psi(\mu^N)$ converges weakly towards $\mathcal{B}_\Psi(\mu)$.

**Proof.** The second assertion is a direct consequence of the first one since

$$\mathcal{B}_\Psi(\mu^N)(\phi) = \frac{\mu^N(\Psi \phi)}{\mu^N(\Psi)} \quad \text{and} \quad \mathcal{B}_\Psi(\mu)(\phi) = \frac{\mu(\Psi \phi)}{\mu(\Psi)},$$

and $\mu^N(\Psi) \to \mu(\Psi)$ as well as $\mu^N(\Psi \phi) \to \mu(\Psi \phi)$ for any bounded and continuous test function $\phi$. Let us now prove the first assertion. Since $\mathcal{X}$ is a metric space and $V$ is continuous, for any threshold $t \geq 0$ there exists (Urysohn’s lemma) a separating continuous function $\rho_t : \mathcal{X} \to [0, 1]$ (Urysohn’s function) such that $\rho_t$ equals one on the set $\{u \in \mathcal{X} : V(u) \leq t - 1\}$ and zero on the set $\{u \in \mathcal{X} : V(u) \geq t\}$. Since $V$ is $\mu$-integrable and $|\phi| \leq V$ $\mu$-almost everywhere, then for any $\epsilon > 0$ there exists $T_\epsilon \geq 0$ such that $|\mu(\phi) - \mu(\phi \rho_t)| < \epsilon$ for any $t \geq T_\epsilon$. Furthermore, since the function $\phi \rho_t$ is bounded and continuous and $\mu^N \xrightarrow{\text{w}} \mu$ we have that $\mu^N(\phi \rho_t) \to \mu(\phi \rho_t)$. It follows that for any $t > T_\epsilon$

$$\limsup_{N \to \infty} \left| \mu^N(\phi) - \mu(\phi) \right| \leq \limsup_{N \to \infty} \left| \mu^N(\phi \rho_t) - \mu(\phi) \right| + \limsup_{N \to \infty} \left| \mu^N(\phi (1 - \rho_t)) \right|$$

$$\leq \limsup_{N \to \infty} \left| \mu^N(\phi \rho_t) - \mu(\phi) \right| + \limsup_{N \to \infty} \mu^N(V \times 1_{V > t - 1})$$

$$\leq \epsilon + \limsup_{N \to \infty} \mu^N(V \times 1_{V > t - 1}).$$

Equation (9) gives the conclusion. \qed

Note that if the potential $\Psi$ is bounded, Proposition 3.1 always applies. In the standard Monte-Carlo setting where $u_i^N = u_i$ for i.i.d samples $\{u_i\}_{i \geq 0}$ from the distribution $\mu$, more precise estimates are available. The distributions $\mu^N$ and $\nu^N$ are random and one can readily check that

$$\left\| \mu^N - \mu \right\| \leq \frac{1}{\sqrt{N}}, \quad (10)$$

where we have used the norm

$$\left\| \mu^N - \mu \right\|^2 = \sup_{\|\phi\|_\infty < 1} \mathbb{E} \left[ (\mu^N(\phi) - \mu(\phi))^2 \right], \quad (11)$$

to measure the discrepancy between two random measures. Furthermore, [APSAS15, Theorem 2.1] states that

$$\left\| \mu^N - \mu \right\| \leq \frac{2}{\sqrt{N}} \frac{\mu(\Psi^2)}{\mu(\Psi)}. \quad (12)$$

The sequence of approximations $\mu^N$ converges at Monte-Carlo rate towards $\mu$. 


3.2 Re-sampling schemes

In standard SMC methods, as well as the SET method described in this article, one needs to transform a weighted particle approximation of a distribution \( \mu \) into an equally weighted particle approximation of the same distribution. The multinomial resampling scheme approximates \( \mu^N = \sum_{i=1}^N w_i^N \delta(u_i^N) \) by the equally weighted particle approximation

\[
\mu_{\text{IS}}^N = \frac{1}{N} \sum_{i=1}^N \delta(u_{\text{IS},i}^N)
\]

where \( \{u_{\text{IS},i}^N\}_{i=1}^N \) are i.i.d. samples from \( \mu^N \). Equation (11) shows that \( \| \mu_{\text{IS}}^N - \mu^N \| \leq 1/\sqrt{N} \). There are more sophisticated approaches, such as the stratified [HSG06] and systematic [DC05] resampling methods, to generate equally weighted particle approximations. We refer the reader to [GCW17] for a recent study of theoretical properties of these typically more statistically efficient resampling schemes. Unless otherwise stated, all the numerical simulations presented in this article use the stratified resampling scheme.

For concreteness, we denote by \( \mathcal{R} \) the resampling operator that maps a weighted particle approximation to an equally weighted one. Note that for a given weighted particle approximation \( \mu^N \), the quantity \( \mathcal{R}(\mu^N) \) is in general a random probability distribution. The resampling scheme \( \mathcal{R} \) is called consistent if it maps \( \mu^N \), a possibly random sequence of distributions that almost surely converges weakly towards \( \mu \), into another sequence \( \mathcal{R}(\mu^N) \) that almost surely converges weakly towards \( \mu \). It has long been known [CD02] that the multinomial resampling scheme is consistent in finite dimensional Euclidean spaces. As investigated in [HSG06], the situation is much more delicate for the stratified and systematic resampling methods.

3.3 Mutation

Consider a sequence \( \{\mu_k\}_{k=0}^N \) of distributions interpolating between a tractable distribution \( \mu_0 \) and the posterior distribution \( \mu_k \) such that for any index \( 1 \leq k \leq K \)

\[
\frac{d\mu_k}{d\mu_{k-1}}(u) = \frac{1}{Z_k} \Psi_k(u)
\]

for a \( \mu_{k-1} \)-integrable potential function \( \Psi_k : \mathcal{X} \rightarrow (0, \infty) \). For technical reasons, we also assume that \( \Psi_k \) is continuous. Given a particle approximation

\[
\mu_0^N = \frac{1}{N} \sum_{i=1}^N \delta(u_{i,0}^N)
\]

of the initial distribution \( \mu_0 \), it is straightforward to build a particle approximation of the posterior distribution. Under mild assumptions, the sequence of equally weighted distributions \( \mu_k^N = (1/N) \sum_{i=1}^N \delta(u_{k,i}^N) \) recursively defined as \( \mu_k^N = \mathcal{R} \circ \mathcal{R}_{\psi_k}(\mu_{k-1}^N) \) converges in an appropriate sense towards \( \mu_k \) as \( N \to \infty \). For example, Proposition 3.1 shows that, if the potential \( \Psi_k \) are bounded and the re-sampling scheme \( \mathcal{R} \) is consistent, as soon as \( \mu_0^N \) almost surely converges weakly towards \( \mu_0 \) the sequence \( \mu_k^N \) also almost surely converges weakly towards \( \mu_k \) as \( N \to \infty \).

In most realistic scenarios, though, the distribution \( \mu_k^N = \mathcal{R}_{\psi_1, \psi_2, \ldots, \psi_k}(\mu_0^N) \), as an approximation to \( \mu_k \), is worse than the direct importance sampling estimate from \( \mu_0 \) to \( \mu_k \). The main reason for the inefficiency of recursive importance sampling is that the particles \( \{u_{k,i}^N\}_{i=1}^N \) form a subset of \( \{u_{0,i}^N\}_{i=1}^N \). Consequently, if the initial set of particles \( \{u_{0,i}^N\}_{i=1}^N \) are located in regions of the parameter space where the distribution \( \mu_k \) does not have much probability mass, the approximation \( \mu_k^N \) to \( \mu_k \) can be very poor. For importance sampling to work well in high-dimensional situations, the proposal distribution needs to be chosen very judiciously, and adaptive importance sampling (AIS) [OB92, CMR12, CDG+08, FT19] can partially remedy this issue. A standard approach to mitigate this issue is to introduce mutation steps, which we now describe. For each distribution \( \mu_k \) in the interpolating sequence of distributions, consider a (mutation) Feller Markov kernel \( M_k(u, du) \) that leaves the distribution \( \mu_k \) invariant. Consider the operator \( \mathcal{H}_k \) that transforms a particle approximation \( \mu_k^N = (1/N) \sum_{i=1}^N \delta(u_{k,i}^N) \) into
\( \mathcal{M}_k(\mu_k^N) = (1/N) \sum_{i=1}^N \delta(u_{k,i}^N) \) where, conditionally upon \( \{u_{k,i}^N\}_{i=1}^N \), the samples \( \{v_{k,i}^N\}_{i=1}^N \) are independent realizations of \( M_k(u_{k,i}^N, du) \). The following lemma shows that, as soon as the sequence \( \mu_k^N \) almost surely converges weakly to \( \mu_k \), the sequence \( \mathcal{M}_k(\mu_k^N) \) also almost surely converges weakly to \( \mu_k \).

**Lemma 3.2.** Let \( \mu \) be a probability distribution on a locally compact and \( \sigma \)-compact metric space \( \mathcal{X} \). Consider \( M(u, du) \) a \( \mu \)-invariant Feller Markov kernel. For each \( N \geq 1 \), let \( \{u_i^N\}_{i=1}^N \subset \mathcal{X} \) be such that

\[
\frac{1}{N} \sum_{i=1}^N \delta(u_i^N) \xrightarrow{w} \mu \in \mathcal{P}(\mathbb{R}^d).
\]

For independent random variables \( V_i^N \sim M(u_i^N, du) \), we have that, almost surely,

\[
\frac{1}{N} \sum_{i=1}^N \delta(V_i^N) \xrightarrow{w} \mu.
\]

**Proof.** Since \( \mathcal{X} \) is a locally compact and \( \sigma \)-compact metric space, there exists a countable and dense (for the supremum norm) subset \( \mathcal{H} \) of the set of continuous functions with compact support in \( \mathcal{X} \). One needs to prove that for any \( \varphi \in \mathcal{H} \) we have that \( \lim_{N \to \infty} (1/N) \sum_{i=1}^N \varphi(v_i^N) = \mu(\varphi) \) almost surely. Since the function \( M\varphi \) is continuous and bounded,

\[
\lim_{N \to \infty} \mathbb{E} \left[ \frac{1}{N} \sum_{i=1}^N \varphi(v_i^N) \right] = \lim_{N \to \infty} \frac{1}{N} \sum_{i=1}^N (M\varphi)(u_i^N) = \mu(M\varphi) = \mu(\varphi).
\]

Since \( \varphi \) is bounded, the moment of order four of the ergodic sum \( \frac{1}{N} \sum_{i=1}^N [\varphi(v_i^N) - (M\varphi)(u_i^N)] \) is upper bounded by a constant multiple of \( N^{-2} \). The Borel-Cantelli lemma gives the conclusion. \( \Box \)

Leveraging these Markov mutation kernels, we now define the sequence of equally weighted particle approximations \( \{\mu_k^N\}_{k=0}^K \) recursively as

\[
\mu_k^N = \mathcal{M}_k \circ R \circ \mathcal{B}_{\varphi_k}(\mu_{k-1}^N).
\]

(12)

The Markov mutations ensure that, in general, the equally weighted particle approximation \( \mu_k^N = (1/N) \sum_{i=1}^N \delta(u_{k,i}^N) \) is such that the particles \( \{u_{k,i}^N\} \) do not form a subset of \( \{u_{i,0}^N\} \). The particle algorithm resulting from (12) is a special case of Sequential Monte Carlo (SMC) samplers [DMDJ06]. Note that, in Bayesian inverse problems, simulating from the Markovian kernel \( M_k \) typically requires evaluating the computationally expensive forward map. Moreover, as explained in the introduction, whilst well-designed Markovian kernels can greatly enhance the statistical efficiency of the resulting algorithm, it is notoriously difficult to design well-mixing mutation kernels in high-dimensional settings.

### 4 Adaptive procedure

In complex scenarios such as Bayesian inverse problems, it is a nontrivial task to specify a sequence of distributions (6) that interpolates between a distribution, which is easy to sample from, and the posterior distribution. Similarly, choosing a-priori a sequence of Markov mutation kernels is typically not feasible. Instead, we consider an adaptive annealing scheme [DBR00, MDMM10, JSDT11, ZJA16, NSPD16, SC13, KBJ14]. The reader is referred to [BJKT15, GDM+] for theoretical analysis of adaptive annealing methods. Recall definition (5) of the posterior distribution. After a finite element discretization, all the quantities of interest are finite dimensional and the distributions can be described by their densities with respect to the Lebesgue measure in \( \mathbb{R}^m \). For notational convenience, we identify distributions with their densities, and assume that the posterior distribution \( \mu_{\text{post}} \) is absolutely continuous with respect to \( \mu_0 \), i.e. \( d\mu_{\text{post}}/d\mu_0(u) \propto \exp\{V(u)\} \) for some potential function \( V : \mathbb{R}^m \to \mathbb{R} \). Consider the sequence \( \{\mu_k\}_{k=0}^K \) defined as

\[
\frac{d\mu_k}{d\mu_0}(u) \propto \exp \{ \tau_k V(u) \}
\]
for an (inverse) temperature parameter \( \tau_k \) that interpolates between \( \tau_0 = 0 \) and \( \tau_K = 1 \). For the case where \( \mu_0 = \mu_{\text{prior}} \), the potential \( V \) is the log-likelihood. In practice, it is not trivial to choose the number of temperatures \( K \) (i.e. the number of interpolating densities) and the corresponding temperature values. The adaptive scheme proceeds as follows. Assume that the particle approximation

\[
\mu_k^{N} = \frac{1}{N} \sum_{i=1}^{N} \delta(u_{i,k}^{N})
\]

to the density \( \mu_k \) has already been constructed. For a predetermined threshold \( 0 < \xi < 1 \), the next temperature \( \tau_{k+1} \) is defined as the smallest temperature \( \tau > \tau_k \) such that \( \text{ESS}_k(\tau) \leq \xi \). Here, The Effective Sample Size (ESS) functional is defined as

\[
\text{ESS}_k(\tau) \equiv \frac{1}{N} \left( \sum_{i=1}^{N} \exp \left( \frac{(\tau - \tau_k) V(u_{i,k}^{N})}{\tau_k} \right) \left( \frac{\tau - \tau_k}{V(u_{i,k}^{N})} \right)^2 \right)^{1/2} \in [0, 1].
\]

Clearly, \( \text{ESS}_k(\tau_k) = 1 \). Lemma 3.1 of [BJKT15] states that the function \( \tau \mapsto \text{ESS}_k(\tau) \) is decreasing for \( \tau \in (\tau_k, \infty) \) so that \( \tau_{k+1} \) can very efficiently be found by a bisection method. Finding \( \tau_{k+1} \) typically does not require evaluating the forward map since, in standard implementations of the SMC or SET methods, the quantities \( V(u_{i,k}^{N}) \) would have already been computed at previous steps. Starting from \( \tau_0 = 0 \) and setting

\[
\tau_{k+1} = \inf \{ \tau > \tau_k : \text{ESS}_k(\tau) \leq \xi \},
\]

the procedure stops as soon as \( \tau_k \) is greater or equal to one. One thus sets \( K = \inf \{ k \geq 1 : \tau_k \geq 1 \} \) and defines \( \tau_K = 1 \). Note that taking \( \xi \) close to one leads to a slow annealing, which may be computationally wasteful. On the other hand, taking \( \xi \) close to zero can lead to an annealing scheme that is too rapid, ultimately leading to a poor particle approximation of the posterior distribution. We choose \( \xi = 1/2 \) in the numerical experiments of Section 6.

Choosing a-priori a sequence of Markov mutation kernels is, in most realistic scenarios, not feasible. A standard approach consists in exploiting the population \( \{u_{i,k}^{N}\}_{i=1}^{N} \) of particles at temperature \( \tau_k \) to estimate summary statistics of the distribution \( \mu_k \). These summary statistics estimates (e.g. mean and covariance matrix) can then be leveraged to design a Markov kernel \( M_k \) with reasonable mixing properties. For instance, one can estimate the mean and covariance matrix of \( \mu_k \) in order to design a (variation of the) random-walk Metropolis-Hastings Markov kernel that is reversible with respect to \( \mu_k \).

## 5 Optimal Transport

For technical simplicity, we assume in this section that the state space \( \mathcal{X} \) is a finite dimensional Euclidean space with norm denoted by \( \| \cdot \| \). For two distributions \( \mu \) and \( \nu \) related by a change of probability \( d\nu/d\mu(u) \propto \Psi(u) \), the Monge-Kantorovich optimal transport approach provides an alternate methodology for building a particle approximation of a distribution \( \nu \) out of a particle approximation of \( \mu \). To the best of our knowledge, the idea was first proposed in [Rei13], and further developed in [GCR16, CRR16, GT19], in the context of data-assimilation of dynamical systems.

For two probability distributions \( \mu \) and \( \nu \), let \( \mathcal{P}(\mu, \nu) \) be the set of probability couplings between \( \mu \) and \( \nu \), i.e. the convex set of probability distributions on \( \mathcal{X} \times \mathcal{X} \) that admit \( \mu \) and \( \nu \) as marginals. For a cost function \( c : \mathcal{X} \times \mathcal{X} \rightarrow [0, \infty) \), the optimal transportation problem seeks to minimize the transport cost \( \mathbb{E}_{\gamma}[c(\hat{u}, \hat{v})] \), for \( (\hat{u}, \hat{v}) \sim \gamma \), over the set of all possible couplings \( \gamma \in \mathcal{P}(\mu, \nu) \).

\[
\gamma^{\text{OT}} = \arg\min_{\gamma} \left\{ \gamma \mapsto \mathbb{E}_{\gamma}[c(\hat{u}, \hat{v})] \mid \gamma \in \mathcal{P}(\mu, \nu) \right\}.
\]

On an Euclidean space, a standard choice is the quadratic cost function \( c(u, v) = \|u - v\|^2 \). For cost functions of the type \( c(u, v) = h(v - u) \) for a strictly convex function \( h \), Brenier’s theorem [Bre91] states that, if \( \mu \) is compactly supported and has a density with respect to the Lebesgue measure, there exists a deterministic map \( T : \mathcal{X} \rightarrow \mathcal{X} \),
NEWLY CREATED PARTICLES

\[ \mu \]

\[ \text{Consider a weighted particle approximation} \]

\[ \text{situation is more delicate [EG99, TW01, CFM02, Amb03].} \]

\[ \gamma X \]

\[ \text{approximation to} \]

\[ \mu T \]

\[ \text{is described by a deterministic map} \]

\[ \text{it is reasonable to expect} \]

\[ \text{one can resort to an approximation scheme. Note that the quantity} \]

\[ \gamma X \]

\[ \text{through the deterministic function} \]

\[ \text{on the finite set} \]

\[ \text{with} \]

\[ \Psi : X \rightarrow X \]

\[ \text{the probability distribution} \]

\[ \text{on the finite set} \]

\[ \text{with} \]

\[ \sum P \]

\[ \text{over the convex set} \]

\[ \text{on the /finite set} \]

\[ \text{with} \]

\[ C \]

\[ \Psi : X \rightarrow \]

\[ \text{approximation to} \]

\[ \mu \]

\[ T \]

\[ \text{is expected to be an approximation of} \]

\[ \mu \]

\[ \text{to be a particle approximation of} \]

\[ \text{Although the optimal transformation} \]

\[ \text{is generally computationally intractable,} \]

\[ \text{one can resort to an approximation scheme. Note that the quantity} \]

\[ \text{can be expressed as a conditional expectation} \]

\[ \text{since the pair} \]

\[ \text{has the same distribution as} \]

\[ \text{This motivates the approximation} \]

\[ \text{with} \]

\[ \gamma X \]

\[ \text{Note that the denominator in the right-hand side of (16) satisfies} \]

\[ \text{The newly created particles} \]

\[ u_i^{OT,N} \]

\[ \text{defined as,} \]

\[ \frac{1}{\alpha_i} \sum_{j=1}^N C_{ij}^{OT,N} u_j^N. \]
are convex combinations of the original particles \(\{u_1^N, \ldots, u_N^N\}\) and thus all lie in the convex hull of the set of original particles. For concreteness and in accordance with the previous sections, we denote by \(\mathcal{T}_\Psi\) the operator that realizes the mapping

\[
\mathcal{T}_\Psi \left( \sum_{i=1}^N \alpha_i \delta(u_i^N) \right) = \sum_{i=1}^N \alpha_i \delta(u_i^{OT,N}) = \sum_{i=1}^N \alpha_i \left( \frac{1}{\alpha_i} \sum_{j=1}^N c_{ij}^{OT,N} u_j^N \right),
\]

(17)

Similar to the operator \(\mathcal{R} \circ \mathcal{B}_\Psi\), the operator \(\mathcal{T}_\Psi\) maps an equally weighted particle approximation of a probability distribution \(\mu\) into an equally weighted particle approximation of \(\mathcal{B}_\Psi(\mu)\). However, unlike \(\mathcal{R} \circ \mathcal{B}_\Psi\), the support of the particle approximation \(\mu^N\) and \(\mathcal{T}_\Psi(\mu^N)\) are typically disjoint.

5.2 Consistency

Consider a potential function \(\Psi : \mathcal{X} \to (0, \infty)\) and two distributions \(\mu\) and \(\nu = \mathcal{B}_\Psi(\mu)\). In this section, we generalize and extend Theorem 1 of [Rei13] to prove that, under mild assumptions, the optimal transport operator \(\mathcal{T}_\Psi\) transforms a sequence \(\mu^N\) to \(\mu\) into a sequence \(\mathcal{T}_\Psi(\mu^N)\) that converges weakly to \(\mathcal{B}_\Psi(\mu)\).

**Assumption 5.1** (Unique Deterministic Coupling). The optimal transport problem between \(\mu\) and \(\mathcal{B}_\Psi(\mu)\) with cost function \(c\) admits a unique solution \(\gamma\) that can be realized by a deterministic transport map \(T : \mathcal{X} \to \mathcal{X}\).

The problem of existence and uniqueness of the solution to an optimal transport problem is well-studied. Under mild assumptions (see McCann’s main theorem [McC95]), the set of couplings between \(\mu\) and \(\nu\) is weakly compact and the functional \(\mu \mapsto E_{\mu}[c(u, v)]\) is continuous in the appropriate topologies, ensuring the existence of an optimal coupling. The uniqueness and regularity properties of the optimal transport map are more delicate to establish and we refer to [Cav15] for recent developments. To proceed to the main result of this section we further assume the following.

**Assumption 5.2** (Regularity of the Transport Map). Let Assumption 5.1 hold for a deterministic map \(T : \mathcal{X} \to \mathcal{X}\). For any bounded and Lipschitz function \(\varphi : \mathcal{X} \to \mathbb{R}\) and sequence \(\mu^N\) that converges weakly to \(\mu\), we have that \(\mu^N(\varphi \circ T) \to \mu(\varphi \circ T)\).

The continuous mapping theorem [MW43] shows that Assumption 5.2 is satisfied provided that the set of discontinuities of \(T\) has zero measure under \(\mu\). In particular, Assumption 5.2 holds in the case when the optimal map \(T\) is continuous. Theorem 5.3 below shows that, under mild growth and regularity assumptions on the optimal transport map \(T : \mathcal{X} \to \mathcal{X}\), the optimal transport scheme \(\mathcal{T}_\Psi\) is consistent as the number of particles \(N \geq 1\) approaches \(\infty\).

**Theorem 5.3.** Consider a potential function \(\Psi : \mathcal{X} \to (0, \infty)\) and two probability distributions \(\mu\) and \(\nu = \mathcal{B}_\Psi(\mu)\) on the state space \(\mathcal{X}\). Assume that Assumptions 5.1 and 5.2 are satisfied for a deterministic optimal map \(T : \mathcal{X} \to \mathcal{X}\). Consider further a sequence of weighted particle approximations \(\mu^N = \sum_{i=1}^N \alpha_i^N \delta(u_i^N)\) that converges weakly to \(\mu\), and such that \(\mathcal{B}_\Psi(\mu^N)\) converges weakly to \(\mathcal{B}_\Psi(\mu)\). If the growth assumption

\[
\limsup_{N \to \infty} \mu^N(u \mapsto |T(u)|^p) + \mathcal{B}_\Psi(\mu^N)(u \mapsto |u|^p) < \infty,
\]

(18)

is satisfied for some exponent \(p > 1\), we have that

\[
\mathcal{T}_\Psi(\mu^N) \xrightarrow{\ast} \mathcal{B}_\Psi(\mu) \equiv \nu.
\]

(19)
Proof. Let $\gamma^{OT,N} = \sum_{i,j} C_{i,j}^{N} \delta(u_i^N) \otimes \delta(u_j^N)$ be the optimal coupling between $\mu^N$ and $\mathcal{F}_\Psi(\mu^N)$. By assumption, $\mu^N \xrightarrow{w} \mu$ and $\nu^N \equiv \mathcal{F}_\Psi(\mu) \xrightarrow{w} \mathcal{F}_\Psi(\mu) \equiv \nu$ and there is a unique optimal coupling $\gamma^{OT}$ between $\mu$ and $\nu$. By compactness (see, e.g. [Vil08, Corollary 5.20]), we have that $\gamma^{OT,N} \xrightarrow{w} \gamma$ as $N \to \infty$.

To show the weak convergence of $\mathcal{F}_\Psi(\mu^N)$ towards $\nu$, it suffices to prove that for any Lipschitz and bounded test function $\varphi$ we have that $\mathcal{F}_\Psi(\mu^N) \to \nu(\varphi)$. Assumption 5.2 implies $\mu^N(\varphi \circ T) \to \nu(\varphi)$. Consequently, it suffices to show that the difference $\mathcal{F}_\Psi(\mu^N) - \mu^N(\varphi \circ T)$ converges to zero as $N \to \infty$, i.e.,

$$
\lim_{N \to \infty} \sum_{i=1}^{N} \alpha_i^N \left| \varphi \left( \frac{1}{\alpha_i^N} \sum_{j=1}^{N} C_{i,j}^{N} u_j \right) - \varphi(T(u_i)) \right| = 0.
$$

Since $\varphi$ is Lipschitz, and $\sum_{j=1}^{N} C_{i,j}^{N} = \alpha_i^N$, it is sufficient to show that

$$
\lim_{N \to \infty} \sum_{i,j} \left| u_j - T(u_i) \right| = 0.
$$

Note that $\sum_{i,j} C_{i,j}^{N} |u_j - T(u_i)| = \gamma^{OT,N}(F)$ with $F(u,v) = |v - T(u)|$. Since $F(p,u,v) \leq |v|^p + |T(u)|^p$, assumption (18) yields that $\limsup_{N} \gamma^{OT,N}(F^p) < \infty$. Since $\gamma^{OT,N} \xrightarrow{w} \gamma$, the bound $\limsup_{N} \gamma^{OT,N}(F^p) < \infty$ implies that the sequence $\gamma^{OT,N}(F)$ converges towards $\gamma(F)$. Since $\gamma(F) = 0$, the conclusion follows. \hfill \Box

5.3 Sequential Ensemble Transform

As in Section 3, consider a sequence $\{\mu_{i}\}^{K}_{i=0}$ of distributions that interpolates between a distribution $\mu_0$ and the posterior distribution; for any index $0 \leq k \leq K$ we have that $(d\mu_k/d\mu_{k-1})(u) = (1/Z_k) \Psi_k(u)$ for a $\mu_{k-1}$-integrable and continuous potential function $\Psi_k : \mathcal{X} \to (0, \infty)$. In this section, we assume the following.

Assumption 5.4. The sequence of probability distributions $\{\mu_{k}\}^{K}_{k=0}$ is such that:

1. for any $0 \leq k \leq K$, the support of $\mu_k$ is bounded,
2. for any $1 \leq k \leq K$, the pair of distributions $(\mu_{k-1}, \mu_k)$ satisfies Assumptions 5.1 and 5.2.

Instead of constructing a sequence of particle approximations to the intermediate distributions $\mu_k$ through importance sampling-resampling methods, consider the following approach that leverages optimal transport. Let $\mu_0^N = (1/N) \sum_{i=0}^{N} \delta(u_0^{i,N})$ be an equally-weighted particle approximation of the initial distribution $\mu_0$. Define the equally weighted particle approximations $\mu_k^N$ through the recursion formula

$$
\mu_k^N = \mathcal{M}_k \circ \mathcal{F}_\Psi(\mu_{k-1}^N),
$$

(20)

where $\mathcal{M}_k$ is the operator associated to a $\mu_k$-invariant Markov mutation kernel $M_k$.

Theorem 5.5 (Consistency of the Sequential Ensemble Transport (SET) algorithm). Let $\{\mu_{k}\}^{K}_{k=0}$ be a sequence of distributions that satisfies Assumption 5.4 and consider $\{u_{0,1}^{i,N}\}^{N}_{i=1} \subset \mathbb{R}^m$ such that

$$
\mu_0^N \equiv \frac{1}{N} \sum_{i=1}^{N} \delta(u_{0,1}^{i,N}) \xrightarrow{w} \mu_0.
$$

Then, for any $1 \leq k \leq K$, the sequence of equally weighted particle approximations $\mu_k^N$ defined recursively through Equation (20) weakly converges to $\mu_k$ almost surely.

Proof. One can proceed by induction. It suffices to prove that if $\mu_{k-1}^N \xrightarrow{w} \mu_{k-1}$ almost surely then $\mathcal{M}_k \circ \mathcal{F}_\Psi(\mu_{k-1}^N) \equiv \mu_k^N \xrightarrow{w} \mu_k$ almost surely. Under Assumption (5.4), the support of the distribution $\mu_{k-1}$ is bounded: one can find a
bounded and continuous function $V_k$ that dominates $\Psi_k$ and invoke Proposition 3.1 to see that $\mathcal{R}_{\Psi_k}(\mu_{k-1}^N) \xrightarrow{w} \mu_k$ almost surely. Furthermore, under Assumption 5.4 the pair $(\mu_{k-1}, \mu_k)$ satisfies Assumptions (5.1)-(5.2) as well as Equation 18. Theorem 5.3 thus shows that $\mathcal{R}_{\Psi_k}(\mu_{k-1}^N) \xrightarrow{w} \mu_k$ almost surely. Finally, since the Feller Markov process $M_k$ lets $\mu_k$ invariant, Lemma 3.2 yields that $M_k \circ \mathcal{R}_{\Psi_k}(\mu_{k-1}^N) \xrightarrow{w} \mu_k$ almost surely.

As previously mentioned, one of the advantages of relying on optimal transportation instead of sampling-resampling techniques is that, as illustrated in Section 6, the resulting algorithm is much less sensitive to the mixing properties of the Markov mutation kernels $M_k$. Moreover, the adaptive tempering strategy described in Section 4 can straightforwardly be used within the SET method. In the next section, we compare the SET approach to more standard SMC approaches.

6 Numerical Results

For PDE-constrained Bayesian inverse problems, the overall cost of the SET algorithm is dominated by PDE solves [DHJ+03]. Indeed, we have found that, even with a number of particles $N = \mathcal{O}(10^4)$, the computational cost of exactly solving the discrete optimal transportation problems [FC17] is negligible when compared to the cost of computing the forwards solves. Consequently, in all the numerical simulations presented in this section, the recently developed approximate, but more scalable, methods [GCPB16, Cut13] for computing discrete optimal transport are not employed. To operate, the SET method requires $\mathcal{O}(K \times N \times (p + 1))$ PDE-solves where $K$ is the number of intermediate temperatures and $p$ is the number of Markov mutations per level. In this section, we adopt the method described in Section 4 for automatically adapting the sequence of temperatures and the Markov mutation kernels. We compare the SET approach to the state-of-the-art adaptive SMC approach of [KBJ14, BJKT15].

6.1 Gaussian toy example

Let $\mu_0$ be a centered and isotropic Gaussian distribution in $\mathbb{R}^D$. The target distribution $\mu$ is also centered, but with covariance matrix $\Gamma \in \mathbb{R}^{D, D}$ given by

$$\Gamma_{i,j} = \sigma^2 \exp \left\{ -\frac{(j-i)^2}{2\ell^2} \right\}$$

for a variance parameter $\sigma^2 > 0$ and length-scale parameter $\ell > 0$. Note that the marginal variance of each coordinate equals $\sigma^2 > 0$. We have that

$$\frac{d\mu}{d\mu_0}(u) \propto \exp \left\{ -\frac{1}{2} \left( \langle u, \Gamma^{-1} u \rangle - \|u\|^2 \right) \right\}.$$  

We apply the (adaptive) SET and (adaptive) SMC methods with the same adaptation strategy: the temperature parameters are obtained adaptively as described in Section 4. We follow the approach of [KBJ14] for the adaptation of the Markov mutation kernels. From the set of particles $\{u_{i,k}^N\}_{i=1}^N$ approximating the distribution $\mu_k$, one can compute an empirical estimate $\mathbf{m}_k^N$ of the mean of $\mu_k$, and an empirical estimate $\mathbf{\Gamma}_k^N$ of its covariance matrix. Since estimating full-covariance matrices in high-dimensional settings from a small number of samples is unstable, similarly to [KBJ14] we use a diagonal approximation with empirical marginal variances on the diagonal. Finally, we use a Metropolis-Hastings Markov kernel with autoregressive proposals of the type

$$\tilde{u}_{k,i}^N = \mathbf{m}_k^N + \rho_k^N (u_{i,k}^N - \mathbf{m}_k^N) + (1 - |\rho_k^N|^2)^{1/2} \mathcal{N}(0, \mathbf{\Gamma}_k^N).$$  

The scaling factor $\rho_k^N \in (0, 1)$ is also chosen adaptively. Values of $\rho_k^N$ close to one lead to conservative proposals while values of $\rho_k^N$ close to zero are more likely to be rejected. Given two thresholds $0 < \xi_0 < \xi_1 < 1$, we adapt $\rho_k^N$ based upon the acceptance rate of the Metropolis-Hastings proposals (21). Specifically, we set $\rho_{k+1}^N = \min(1, 2 \rho_k^N)$ if the proportion of accepted proposals falls below $\xi_0$; we set $\rho_{k+1}^N = \rho_k^N / 2$ if the proportion of accepted proposals
is above ξ⁺; we set \( \rho_{k+1}^N = \rho_k^N \) otherwise. In practice, we use \( \xi_- = 15\% \) and \( \xi_+ = 85\% \). At each intermediate temperature, we apply a fixed number of times the Metropolis-Hastings Markov kernel described by proposal (21). This is referred to as the number of mutations.

Figure 2: Dimension \( D = 20 \). RMSE of the mean of \( \mu \) for the Optimal Transport (OT) and Sequential Monte Carlo (SMC) based approaches for different number of mutations at each intermediate temperature based on \( n = 50 \) independent runs.

Figure 2 shows the Root Means Square Error (RMSE) for the estimation of the mean of \( \mu \) using the (adaptive) SET approach and the (adaptive) SMC approach for different number of mutations at each intermediate temperature. When no mutation are used, the SET approach outperforms the standard SMC method by an order of magnitude or more. As the number of mutations increases, the SMC and SET methods perform roughly equivalently for estimating the posterior mean.

Figure 3: Dimension \( D = 20 \). Estimation of the marginal variances for the Optimal Transport (OT) and Sequential Monte Carlo (SMC) based approaches for different number of mutations at each intermediate temperature. Results are based on \( n = 50 \) independent runs. The quantity \( R(N) \), as defined in Equation (22), is reported.
To investigate the quality of the uncertainty-quantification estimates, we consider the estimation of the marginal variances. Recall that the posterior particle approximation reads
\[ \frac{1}{N} \sum_{i=1}^{N} \delta(u_i) \] where \( u_i = (u_{i,1}, \ldots, u_{i,D}) \in \mathbb{R}^D \) represents the \( i \)-th particle. We measure the discrepancy between the empirical and exact posterior variance through the quantity \( R(N) \), the average of the ratios between empirical and exact marginal variances:
\[ R(N) \equiv \frac{1}{D} \sum_{d=1}^{D} \frac{\hat{\sigma}_d^2(N)}{\sigma_d^2} . \tag{22} \]

The quantity \( \hat{\sigma}_d^2(N) \) is the empirical estimate of the variance of the \( d \)-th coordinate obtained from the \( N \geq 2 \) particles,
\[ \hat{\sigma}_d^2(N) \equiv \frac{1}{N} \sum_{i=1}^{N} u_{i,d}^2 - \left( \frac{1}{N} \sum_{i=1}^{N} u_{i,d} \right)^2 . \]

The quantity \( \sigma_d^2 = \mu(X_d^2) - \mu(X_d)^2 \) denotes the exact posterior variance of the \( d \)-th coordinate. The closer to one the quantity \( R(N) \) is, the better the uncertainty quantification. Values \( R(N) \ll 1 \) indicate an underestimation of the uncertainty. We have established in the previous sections that, under mild assumptions, \( R(N) \to 1 \) as \( N \to \infty \).

Figure 3 displays the quantity \( R(N) \) in different settings. If no mutation at all is used at each temperature, the SMC severely underestimate the uncertainty. Interestingly, even when no mutation is used, the SET method is able to provide a reasonable uncertainty quantification. As expected, for both the SET and SMC approaches, the quality of the uncertainty-quantification increases as the number of mutations per temperature is increased. The SET estimates remain an order of magnitude more accurate even with five mutations per temperature. The SET method is able to produce a significantly better uncertainty-quantification than the adaptive SMC methodology. The SMC estimate consistently under-estimates the uncertainty.

### 6.2 Non-linear PDE-constrained Bayesian inverse problem

In this section, we test the SET method for inference in a Bayesian inverse problem governed by a system of non-linear PDEs. For an open and bounded domain \( \Omega \subset \mathbb{R}^d \) and as previously mentioned in Section 2, consider the elliptic PDE
\[ \begin{align*}
-\nabla \cdot (e^u \nabla w) &= 0 \quad \text{in } \Omega \subset \mathbb{R}^d, \\
-e^u \nabla w \cdot n &= \text{Bi} \cdot w \quad \text{on } \partial \Omega \setminus \Gamma_R, \\
-e^u \nabla w \cdot n &= -1 \quad \text{on } \Gamma_R. 
\end{align*} \tag{23} \]

As described in Section 2, the posterior distribution is described by Equation (5). We perform comparisons in a \( d = 2 \) setting. The PDEs (23) are discretized with the finite element method and implemented using \textsc{FEniCS} \cite{FEniCS}. Figure 4 compares the (adaptive) SET method with the (adaptive) SMC approach and reports the relative \( L^2 \)-norm of the error in estimating the posterior mean. We use a MCMC simulation of length \( L = 10^6 \) using the DRAM sampler \cite{DRAM} for approximating accurately the mean and covariance structure of the posterior distribution. A uniform mesh is generated for a unit-square with 10 nodal points. The synthetic solution is a standard 2D sine wave, the forward problem is governed by Equation 23 and prior covariance operator defined by Equation 4. The results in Figure 4 show the error in estimating the posterior mean. Not surprisingly, each method improves as the number of mutation steps is increased.

More importantly, Figure 5 shows the quality of the estimation of the uncertainty in the same settings: the metric (22) is used. This figure indicates that, even without mutation, the SET method can produce a reasonable uncertainty-quantification while the adaptive SMC methodology collapses (red dots in Figure 5). The SMC method consistently under-estimate the uncertainty. In all settings considered, the estimations of the marginal variances is more than two orders of magnitude worse than the ones produced by the SET methodology.
Conclusions

We have introduced the SET method, an optimal-transport based approach for performing inference in high-dimensional Bayesian inverse problems. The SET methodology is, under mild assumptions, provably consistent in the large-particle regime. Our numerical simulations indicate that, in complex high-dimensional scenarios such as PDE-constrained Bayesian inverse problems where it is typically difficult to design efficient Markov mutation kernels, the SET method performs favourably when compared to other particle-based approaches such as modern adaptive SMC methodologies.

Acknowledgement: Myers and Bui-Thanh are partially supported by the Department of Energy (grant DE-SC0018147), the National Science Foundation (grants NSF-DMS1620352, Early Career NSF-OAC1808576, and NSF-OAC1750863), and the Defense Threat Reduction Agency (grant HDTRA1-18-1-0020). Myers thanks Nick Alger for
many useful discussions on Bayesian inverse problems. AHT acknowledges support from a National University of Singapore (NUS) Young Investigator Award Grant (R-155-000-180-133) and a Singapore Ministry of Education Academic Research Funds Tier 2 (MOE2016-T2-2-135).

References

[Amb03] Luigi Ambrosio. Lecture notes on optimal transport problems. LECTURE NOTES IN MATHEMATICS-SPRINGER VERLAG-, pages 1–52, 2003.

[APSAS15] S. Agapiou, O. Papaspiliopoulos, D. Sanz-Alonso, and A. M. Stuart. Importance Sampling: Intrinsic Dimension and Computational Cost. 2015.

[BBL+08] Thomas Bengtsson, Peter Bickel, Bo Li, et al. Curse-of-dimensionality revisited: Collapse of the particle filter in very large scale systems. In Probability and statistics: Essays in honor of David A. Freedman, pages 316–334. Institute of Mathematical Statistics, 2008.

[BCJ14] Alexandros Beskos, Dan Crisan, and Ajay Jasra. On the Stability of Sequential Monte Carlo Methods in High Dimensions. Adv Applied Probability, 46(4), 2014.

[BJKT15] Alexandros Beskos, Ajay Jasra, Nikolas Kantas, and Alexandre Thiery. On the convergence of adaptive sequential monte carlo methods. Annals of Applied Probability, 26(2):1111–1146, 2015.

[BJMS15] 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.

[Bre91] Yann Brenier. Polar factorization and monotone rearrangement of vector-valued functions. Communications on pure and applied mathematics, 44(4):375–417, 1991.

[BTGMS13a] 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 …., 35(6):2494–2523, 2013.

[BTGMS13b] 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.

[Cav15] Martin Cavalletti, Fabio and Huesmann. Existence and uniqueness of optimal transport maps. Annales de l’Institut Henri Poincare (C) Non Linear Analysis, 32(6):1367–1377, nov 2015.

[CD02] Dan Crisan and Arnaud Doucet. A survey of convergence results on particle filtering methods for practitioners. IEEE Transactions on Signal Processing, 50(3):736–746, 2002.

[CDG+08] Olivier Cappé, Randal Douc, Arnaud Guillin, Jean-Michel Marin, and Christian P Robert. Adaptive importance sampling in general mixture classes. Statistics and Computing, 18(4):447–459, 2008.

[CFM02] Luis Caffarelli, Mikhail Feldman, and Robert McCann. Constructing optimal maps for monge transport problem as a limit of strictly convex costs. Journal of the American Mathematical Society, 15(1):1–26, 2002.

[Cho02] Nicolas Chopin. A sequential particle filter method for static models. Biometrika, 89(3):539–552, 2002.

[CMMR12] JEAN-MARIE CORNUET, JEAN-MICHEL MARIN, Antonietta Mira, and Christian P Robert. Adaptive multiple importance sampling. Scandinavian Journal of Statistics, 39(4):798–812, 2012.

[CR13] Yuan Cheng and Sebastian Reich. A McKean optimal transportation perspective on Feynman-Kac formulae with application to data assimilation. arXiv preprint arXiv:1311.6300, 2013.
[CRR16] Nawinda Chustagulprom, Sebastian Reich, and Maria Reinhardt. A hybrid ensemble transform particle filter for nonlinear and spatially extended dynamical systems. *SIAM/ASA Journal on Uncertainty Quantification*, 4(1):592–608, 2016.

[Cut13] M. Cuturi. Sinkhorn distances: Lightspeed computation of optimal transportation. In *Advances in Neural Information Processing Systems 26*, pages 2292–2300, 2013.

[DBR00] Jonathan Duetscher, Andrew Blake, and Ian Reid. Articulated body motion capture by annealed particle filtering. In *cvpr*, page 2126. IEEE, 2000.

[DC05] Randal Douc and Olivier Cappé. Comparison of resampling schemes for particle filtering. In *Image and Signal Processing and Analysis, 2005. ISPA 2005. Proceedings of the 4th International Symposium on*, pages 64–69. IEEE, 2005.

[DFFG01] 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.

[Del04] Pierre Del Moral. *Feynman-Kac Formulae*. Springer, 2004.

[DHJ+03] T. Dupont, J. Hoffman, C. Johnson, R. Kirby, M. Larson, A. Logg, and R. Scott. The FEniCS project. Technical report, 2003.

[DJ09] Arnaud Doucet and Adam M Johansen. A tutorial on particle filtering and smoothing: Fifteen years later. *Handbook of nonlinear filtering*, 12(656-704):3, 2009.

[DMDJ06] 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.

[DMDJ12] Pierre Del Moral, Arnaud Doucet, and Ajay Jasra. An adaptive sequential monte carlo method for approximate bayesian computation. *Statistics and Computing*, 22(5):1009–1020, 2012.

[EG99] Lawrence C Evans and Wilfrid Gangbo. *Differential equations methods for the Monge-Kantorovich mass transfer problem*, volume 653. American Mathematical Soc., 1999.

[FC17] R’emi Flamary and Nicolas Courty. POT python optimal transport library, 2017.

[FT19] Matthew M Graham and Alexandre H Thiery. On the relationship between variational inference and adaptive importance sampling. *arXiv preprint arXiv:1907.10477*, 2019.

[GCPB16] Aude Genevay, Marco Cuturi, Gabriel Peyré, and Francis Bach. Stochastic optimization for large-scale optimal transport. In *Advances in Neural Information Processing Systems*, pages 3440–3448, 2016.

[GCR16] Alastair Gregory, Colin J Cotter, and Sebastian Reich. Multilevel ensemble transform particle filtering. *SIAM Journal on Scientific Computing*, 38(3):A1317—A1338, 2016.

[GCW17] Mathieu Gerber, Nicolas Chopin, and Nick Whiteley. Negative association, ordering and convergence of resampling methods. *arXiv preprint arXiv:1707.01845*, 2017.

[GDM+17] François Giraud, Pierre Del Moral, et al. Nonasymptotic analysis of adaptive and annealed feynman–kac particle models. *Bernoulli*, 23(1):670–709, 2017.

[GSS93] Neil J Gordon, David J Salmond, and Adrian FM Smith. Novel approach to nonlinear/non-gaussian bayesian state estimation. In *IEE Proceedings F-radar and signal processing*, volume 140, pages 107–113. IET, 1993.

[GT19] Matthew M Graham and Alexandre H Thiery. A scalable optimal-transport based local particle filter. *arXiv preprint arXiv:1906.00507*, 2019.
[Has70] W. Keith Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109, 1970.

[HDP15] Jeremy Heng, Arnaud Doucet, and Yvo Pokern. Gibbs flow for approximate transport with applications to bayesian computation. *arXiv preprint arXiv:1509.08787*, 2015.

[HLMS06] Heikki Haario, Marko Laine, Antonietta Mira, and Eero Saksman. Dram: efficient adaptive mcmc. *Statistics and computing*, 16(4):339–354, 2006.

[HSG06] Jeroen D Hol, Thomas B Schon, and Fredrik Gustafsson. On resampling algorithms for particle filters. In *Nonlinear Statistical Signal Processing Workshop, 2006 IEEE*, pages 79–82. IEEE, 2006.

[JSDT11] Ajay Jasra, David A Stephens, Arnaud Doucet, and Theodoros Tsagaris. Inference for levy-driven stochastic volatility models via adaptive sequential monte carlo. *Scandinavian Journal of Statistics*, 38(1):1–22, 2011.

[KB14] 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.

[KS06] Jari Kaipio and Erkki Somersalo. *Statistical and computational inverse problems*, volume 160. Springer Science & Business Media, 2006.

[LW16] Anthony Lee and Nick Whiteley. Forest resampling for distributed sequential monte carlo. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 9(4):230–248, 2016.

[Mcc95] Robert J McCann. Existence and Uniqueness of Monotone Measure-Preserving Maps. *Duke Math. J.*, 80(2):309–323, mar 1995.

[MDMM10] Pierre Minvielle, Arnaud Doucet, Alan Marrs, and Simon Maskell. A bayesian approach to joint tracking and identification of geometric shapes in video sequences. *Image and Vision Computing*, 28(1):111–123, 2010.

[MM12] Tarek A. El Moselhy and Youssef M. Marzouk. Bayesian inference with optimal maps. *Journal of Computational Physics*, 231(23):7815 – 7850, 2012.

[Mon81] Gaspard Monge. Mémoire sur la théorie des déblais et des remblais. *Histoire de l’Académie Royale des Sciences*, pages 666–704, 1781.

[MRR+53] Nicholas Metropolis, Arianna W. Rosenbluth, Marshall N. Rosenbluth, Augusta H. Teller, and Edward Teller. Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6):1087–1092, 1953.

[MT95] Klaus Mosegaard and Albert Tarantola. Monte carlo sampling of solutions to inverse problems. *Journal of Geophysical Research: Solid Earth*, 100(B7):12431–12447, 1995.

[MW43] H. B. Mann and A. Wald. On stochastic limit and order relationships. *Ann. Math. Statist.*, 14(3):217–226, 09 1943.

[NSPD16] Thi Le Thu Nguyen, François Septier, Gareth W Peters, and Yves Delignon. Efficient sequential monte-carlo samplers for bayesian inference. *IEEE Transactions on Signal Processing*, 64(5):1305–1319, 2016.

[OB92] Man-Suk Oh and James O Berger. Adaptive importance sampling in monte carlo integration. *Journal of Statistical Computation and Simulation*, 41(3-4):143–168, 1992.

[PC+19] Gabriel Peyré, Marco Cuturi, et al. Computational optimal transport. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.
[PM14] Matthew Parno and Youssef Marzouk. Transport map accelerated Markov chain Monte Carlo. *ArXiv*, pages 1–48, 2014.

[Rei13] Sebastian Reich. A nonparametric ensemble transform method for Bayesian inference. *SIAM J. Sci. Comput.*, 35(4):A2013–A2024, 2013.

[SBM18] Alessio Spantini, Daniele Bigoni, and Youssef Marzouk. Inference via low-dimensional couplings. *The Journal of Machine Learning Research*, 19(1):2639–2709, 2018.

[SC13] Christian Schäfer and Nicolas Chopin. Sequential monte carlo on large binary sampling spaces. *Statistics and Computing*, 23(2):163–184, 2013.

[ST19] Deborshee Sen and Alexandre H Thiery. Particle filter efficiency under limited communication. *arXiv preprint arXiv:1904.09623*, 2019.

[Stu10] Andrew M M Stuart. Inverse problems: a Bayesian perspective. *Acta Numerica*, 19:451–559, 2010.

[TW01] Neil S Trudinger and Xu-Jia Wang. On the monge mass transfer problem. *Calculus of Variations and Partial Differential Equations*, 13(1):19–31, 2001.

[VDDMM15] Christelle Vergé, Cyrille Dubarry, Pierre Del Moral, and Eric Moulines. On parallel implementation of sequential monte carlo methods: the island particle model. *Statistics and Computing*, 25(2):243–260, 2015.

[Vil03] Cédric Villani. *Topics in optimal transportation*. Number 58. American Mathematical Soc., 2003.

[Vil08] Cédric Villani. *Optimal transport: old and new*, volume 338. Springer Science & Business Media, 2008.

[WLH+16] Nick Whiteley, Anthony Lee, Kari Heine, et al. On the role of interaction in sequential monte carlo algorithms. *Bernoulli*, 22(1):494–529, 2016.

[ZJA16] Yan Zhou, Adam M Johansen, and John AD Aston. Toward automatic model comparison: an adaptive sequential monte carlo approach. *Journal of Computational and Graphical Statistics*, 25(3):701–726, 2016.