Sequential Ensemble Transform for Bayesian Inverse Problems

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

September 22, 2020

Abstract

We present the Sequential Ensemble Transform (SET) method, an approach for generating approximate samples from a Bayesian posterior distribution. The method explores the posterior distribution by solving a sequence of discrete optimal transport problems to produce a series of transport plans which map prior samples to posterior samples. We prove that the sequence of Dirac mixture distributions produced by the SET method converges weakly to the true posterior as the sample size approaches infinity. Furthermore, our numerical results indicate that, when compared to standard Sequential Monte Carlo (SMC) methods, the SET approach is more robust to the choice of Markov mutation kernels and requires less computational efforts to reach a similar accuracy when used to explore complex posterior distributions. Finally, we describe adaptive schemes that allow to completely automate the use of the SET method.

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

[*]Institute for Computational Engineering & Sciences, The University of Texas at Austin, Austin, TX 78712, USA.
[†]Sanchez Oil & Gas, Houston, TX
[‡]Department of Aerospace Engineering & Engineering Mechanics , The University of Texas at Austin, Austin, TX 78712, USA
[§]Department of Satistics and Applied Probability, National University of Singapore, Singapore

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) 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 $\mathcal{O}(N^{-1/2})$. 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, similarly 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], or more generally when there is a large discrepancy between the proposal and target distributions. 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 mutation 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 when compared to 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 (such as low dimensional parameter spaces or closed-to-Gaussian posteriors) when the design of Markov kernels with good mixing properties is not challenging, our proposed method may not provide significant computational savings over more standard SMC or MCMC methods.

Our main contributions are as follows. We propose the Sequential Ensemble Transform (SET) algorithm, an interacting particle methodology inspired from the data-assimilation literature [Rei13], for Bayesian inversion. Unlike most interacting particle methods that rely on resampling approaches, the SET method is based on optimal transportation. We demonstrate empirically that this leads to an algorithm that is less affected by particle degeneracy, and requires less computational effort to converge, than more standard SMC approaches when used in complex settings where designing efficient Markov mutation kernels is not trivial. We make SET practical, especially for complex applications, by providing several adaptation strategies for automating the choice of tuning parameters. Finally, we establish conditions under which, in the limit when the number of particles approaches infinity, the SET method is provably consistent, i.e., the sequence of particle approximations produced by the SET converges weakly towards the underlying target distribution.

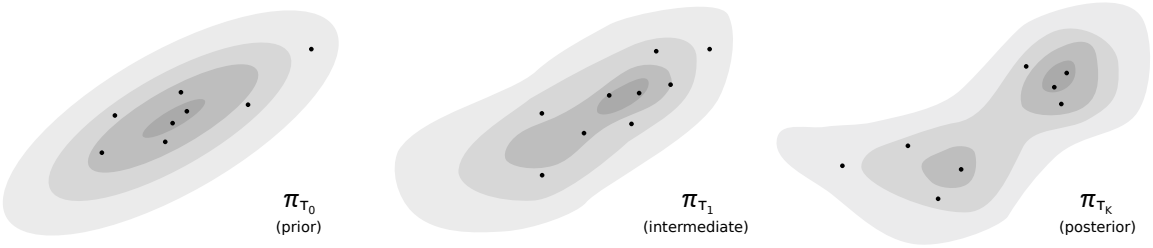

Figure 1: A representation of the SET method using optimal transport to move particles in parameter space as to represent the posterior

The article is structured as follows. In Section 2, PDE-constrained Bayesian inverse problems are briefly described. An overview of particle methods and importance sampling is presented in Section 3. Section 4 presents the concept of optimal transport and describes the main components of SET method, as well as their asymptotic properties. Section 5 describes the SET methods in details, as well as several adaptive strategies that can be used to automate several aspects of

the method. Finally, Section 6 presents various numerical results, including a Bayesian inverse problem with a non-linear forward map. 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 σ-algebra. The notations μ and ν (along with any use of super- or sub-scripts) denote probability distributions. A sequence of probability distributions $\{\mu^N\}_{N\geq 1}$ on $\mathcal{X}$ converges weakly towards the distribution μ, denoted as $\mu^N \xrightarrow{\text{w}} \mu$, if for any bounded and continuous test function $\varphi : \mathcal{X} \to \mathbb{R}$ we have that $\int \varphi(u)\, \mu^N(du) \to \int \varphi(u)\, \mu(du)$ as $N \to \infty$. Similarly, a sequence of random probability distribution μ_ω^N almost surely converges weakly towards μ if, for $\mathbb{P}$-almost every ω, we have that $\mu_\omega^N \xrightarrow{\text{w}} \mu$. The set of probability distributions on a state space $\mathcal{X}$ is denoted as $\mathcal{P}(\mathcal{X})$. For a set S, the notation $\mathbb{1}_S$ 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 Bayesian treatment of inverse problems. We are interested in estimating a field $u \in \mathcal{X}$, where $\mathcal{X}$ denotes a space of functions, from a finite set of observations contaminated by additive Gaussian noise,

$$\mathbf{d} = \mathcal{G}(u) + \boldsymbol{\eta},$$

where $\mathbf{d} = [d_1, \ldots, d_D]^\top \in \mathcal{Y}$ and $\boldsymbol{\eta} \sim \mathcal{N}(0, \mathbf{L})$ is centred Gaussian vector with covariance matrix $\mathbf{L}$. The operator $\mathcal{G} : \mathcal{X} \to \mathcal{Y}$ describes the mapping from the parameters to observables. In Bayesian inverse problems and as illustrated in Section 6, estimating the quantity $\mathcal{G}(u)$ typically involves solving a set of partial differential equations. In order to estimate the uncertainty associated to the necessarily imperfect reconstruction of the parameter $u \in \mathcal{X}$, the Bayesian approach postulates a prior distribution μ_{prior} that describes the information available on the parameter u prior to any data collection. Under mild assumptions [Stu10], the Bayesian posterior distribution μ_{post} is defined through the change of measure formula

$$\frac{d\mu_{\text{post}}}{d\mu_{\text{prior}}}(u) \propto \exp\left\{-\frac{1}{2}\left|\mathbf{d} - \mathcal{G}(u)\right|_{\mathbf{L}}^2\right\} \tag{1}$$

where $|\cdot|_{\mathbf{L}} \equiv \left|\mathbf{L}^{-\frac{1}{2}}\cdot\right|$ denotes the $\mathbf{L}^{-\frac{1}{2}}$-weighted Euclidean norm. In situations when the mapping $\mathcal{G}$ is non-linear, the posterior distribution is typically intractable and numerical methods such as MCMC are required to estimate expectations (and other statistics) of observables with respect to the posterior μ_{post}.

3 Particle Methods

Particle methods approximate probability distributions with weighted mixtures of Diracs, also referred to as *particle approximation* in this text. 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. μ_0, and the posterior distribution μ_K. A standard choice for μ_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

$$\frac{d\mu_k}{d\mu_{k-1}}(u) = \frac{1}{Z_k}\,\Psi_k(u), \tag{2}$$

for a μ_{k-1}-integrable potential function $\Psi_k : \mathcal{X} \to (0, \infty)$ and (typically unknown) normalization constant $Z_k > 0$. The SMC algorithm recursively constructs particle approximations

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

where $N \geq 1$ denotes the number of particles, by iterating *re-weighting*, *resampling*, and *mutation* operations that are described below. In the remaining of this text, we make use of the following notations that are standard in the Monte-Carlo literature and compactly allow to describe expectations with respect to probability distributions and Markov kernels. For a probability distribution μ on the state space $\mathcal{X}$ and a μ-integrable test function $\varphi : \mathcal{X} \to \mathbb{R}$, set $\mu(\varphi) \equiv \int \varphi(u)\,\mu(du)$. Similarly, for a Markov kernel $M(u, dv)$, define $(M\varphi)(u) \equiv \int\, \varphi(v)\,M(u, dv)$.

3.1 Re-weighting

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

$$\frac{d\nu}{d\mu}(u) = \frac{1}{Z}\,\Psi(u) \tag{3}$$

for a μ-integrable potential function $\Psi : \mathcal{X} \to (0, \infty)$ 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 μ as $N \to \infty$. Under mild assumptions, the sequence of *self-normalized importance sampling* weighted particle approximations ν^N defined as

$$\nu^N \equiv \sum_{i=1}^N w_i^N\,\delta(u_i^N) \tag{4}$$

for normalized weights

$$w_i^N \equiv \frac{\Psi(u_i^N)}{[\Psi(u_1^N) + \ldots + \Psi(u_N^N)]}$$

converges weakly to ν. For concreteness, define the mapping from μ^N to ν^N as $\nu^N = \mathscr{B}_\Psi(\mu^N)$ where $\mathscr{B}_\Psi$ is the so-called Bayes operator that transforms a probability distribution μ into the probability distribution $\mathscr{B}_\Psi(\mu)$ that satisfies $\mathscr{B}_\Psi(\mu)(\varphi) = \mu(\Psi\,\varphi)/\mu(\Psi)$ for any test function φ. The following proposition shows that, under a mild *uniform integrability* condition, the convergence $\mathscr{B}_\Psi(\mu^N) \xrightarrow{\text{w}} \mathscr{B}_\Psi(\mu)$ holds.

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

$$\lim_{t\to\infty} \limsup_{N\to\infty} \mu^N(\mathcal{E} \times \mathbb{1}_{\mathcal{E}>t}) = 0, \tag{5}$$

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

1. *for any (potentially unbounded) continuous test function φ such that $|\varphi| \leq \mathcal{E}$,*

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

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

Remark 2. *The technical condition Equation (5) means that if ζ_N is a sequence of random variables such that $\zeta_N \sim \mu^N$, the sequence of scalar random variables $\overline{\zeta}_N \equiv \mathcal{E}(\zeta_N)$ is uniformly integrable [Wil91].*

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

$$\mathscr{B}_\Psi(\mu^N)(\varphi) = \frac{\mu^N(\Psi\,\varphi)}{\mu^N(\Psi)} \qquad \text{and} \qquad \mathscr{B}_\Psi(\mu)(\varphi) = \frac{\mu(\Psi\,\varphi)}{\mu(\Psi)},$$

and $\mu^N(\Psi) \to \mu(\Psi)$ as well as $\mu^N(\Psi\,\varphi) \to \mu(\Psi\,\varphi)$ for any bounded and continuous test function φ. Let us now prove the first assertion. Since $\mathcal{X}$ is a metric space and $\mathcal{E}$ 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 $\rho_t(u) = 1$ on the set $\{u \in \mathcal{X} : \mathcal{E}(u) \leq t - 1\}$ and $\rho_t(u) = 0$ on the set $\{u \in \mathcal{X} : \mathcal{E}(u) \geq t\}$. Since $\mathcal{E}$ is μ-integrable and $|\varphi| \leq \mathcal{E}$ μ-almost everywhere, then for any $\varepsilon > 0$ there exists $T_\varepsilon \geq 0$ such that $|\mu(\varphi) - \mu(\varphi\,\rho_t)| < \varepsilon$ for any $t \geq T_\varepsilon$. Furthermore, since the function $\varphi\,\rho_t$ is bounded and continuous and $\mu^N \xrightarrow{\text{w}} \mu$, we have that $\mu^N(\varphi\,\rho_t) \to \mu(\varphi\,\rho_t)$. It follows that for any $t > T_\varepsilon$

$$\begin{aligned}
\limsup_{N\to\infty} |\mu^N(\varphi) - \mu(\varphi)| &\leq \limsup_{N\to\infty} |\mu^N(\varphi\,\rho_t) - \mu(\varphi)| + \limsup_{N\to\infty} |\mu^N(\varphi\,(1 - \rho_t))| \\
&\leq \limsup_{N\to\infty} |\mu^N(\varphi\,\rho_t) - \mu(\varphi)| + \limsup_{N\to\infty} \mu^N(\mathcal{E} \times \mathbb{1}_{\mathcal{E}>t-1}) \\
&\leq \varepsilon + \limsup_{N\to\infty} \mu^N(\mathcal{E} \times \mathbb{1}_{\mathcal{E}>t-1}).
\end{aligned}$$

Equation (5) gives the conclusion. □

Note that if the potential Ψ is bounded, Proposition 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 μ, more precise estimates are available. The distributions μ^N and ν^N are random and one can readily check that

$$|||\mu^N - \mu||| \leq \frac{1}{\sqrt{N}}, \tag{6}$$

where we have used the norm defined as

$$|||\mu^N - \mu|||^2 \equiv \sup_{\|\varphi\|_\infty < 1} \mathbb{E}\left[\left(\mu^N(\varphi) - \mu(\varphi)\right)^2\right] \tag{7}$$

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

$$|||\mu^N - \mu||| \leq \frac{2}{\sqrt{N}} \frac{\mu(\Psi^2)^{\frac{1}{2}}}{\mu(\Psi)}.$$

The sequence of approximations μ^N converges at Monte-Carlo rate towards μ.

3.2 Resampling 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 μ 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

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

where $\{u_{i,\mathrm{IS}}^N\}_{i=1}^N$ are i.i.d. samples from μ^N. Equation (6) states that the norm between a distribution and an equally weighted mixture of Dirac masses centred at N i.i.d samples from that distribution is less than $1/\sqrt{N}$. Applying this remark and the fact that μ_{IS}^N is precisely an equally weighted mixture of Dirac masses centred at N i.i.d samples from μ^N, it follows that $|||\mu_{\mathrm{IS}}^N - \mu^N||| \leq 1/\sqrt{N}$. 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 $\mathscr{R}$ the resampling operator that maps a weighted particle approximation to an equally weighted one. Note that for a given weighted particle approximation μ^N, the quantity $\mathscr{R}(\mu^N)$ is in general a random probability distribution. The resampling scheme $\mathscr{R}$ is called *consistent* if it maps μ^N, a possibly random sequence of distributions that almost surely converges weakly towards μ, into another sequence $\mathscr{R}(\mu^N)$ that almost surely converges weakly towards μ. 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 $\{\mu_k\}_{k=0}^K$ of distributions interpolating between a tractable distribution μ_0 and the posterior distribution μ_K such that for any index $1 \leq k \leq K$ we have

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

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

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

of the initial distribution μ_0. 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 $\mu_k^N = \mathscr{R} \circ \mathscr{B}_{\Psi_k}(\mu_{k-1}^N)$ converges in an appropriate sense towards μ_k as $N \to \infty$. For example, Proposition 1 shows that, if the potential Ψ_k are bounded and the resampling scheme $\mathscr{R}$ is consistent, as soon as μ_0^N almost surely converges weakly towards μ_0 the sequence μ_k^N also almost surely converges weakly towards μ_k as $N \to \infty$.

In most realistic scenarios, though, the particle approximation μ_K^N, as an approximation to μ_K, is worse than the direct importance sampling particle approximation $\mathscr{B}_{\Psi_1\Psi_2\ldots\Psi_K}(\mu_0^N)$ from μ_0 to μ_K where $(d\mu_K/d\mu_0)(u) \propto [\Psi_1\Psi_2\ldots\Psi_K](u)$. It is because in that case 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 $\{u_{0,i}\}_{i=1}^N$ are located in regions of the parameter space where the distribution μ_K does not have much probability mass, the approximation μ_K^N to μ_K can be very poor. For importance sampling to work well in high-dimensional situations, the proposal distributions need to be chosen very judiciously, and adaptive importance sampling (AIS) [OB92, CMMR12, 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 μ_k in the interpolating sequence of distributions, consider a (mutation) Feller Markov kernel $M_k(u, d\widehat{u})$ that leaves the distribution μ_k invariant. Consider the operator $\mathscr{M}_k$ that transforms a particle approximation $\mu_k^N = (1/N)\sum_{i=1}^N \delta(u_{k,i}^N)$ into $\mathscr{M}_k(\mu_k^N) = (1/N)\sum_{i=1}^N \delta(v_{k,i}^N)$ where, conditionally upon $\{u_{k,i}^N\}_{i=1}^N$, the samples $\{v_{k,i}^N\}_{i=1}^N$ are independent realizations of $M_k(u_{k,i}^N, d\widehat{u})$. The following lemma shows that, as soon as the sequence μ_k^N almost surely converges weakly to μ_k, the sequence $\mathscr{M}_k(\mu_k^N)$ also almost surely converges weakly to μ_k.

Lemma 3. *Let μ be a probability distribution on a locally compact and σ-compact metric space $\mathcal{X}$. Consider $M(u, d\widehat{u})$ a μ-invariant Feller Markov kernel. For each $N \geq 1$, let $\{u_i^N\}_{i=1}^N \subset \mathcal{X}$ be such that*

$$\frac{1}{N}\sum_{i=1}^N \delta(u_i^N) \overset{w}{\to} \mu.$$

For independent random variables $V_i^N \sim M(u_i^N, d\widehat{u})$, 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 σ-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 φ 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. □

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

$$\mu_k^N = \mathscr{M}_k \circ \mathscr{R} \circ \mathscr{B}_{\Psi_k}(\mu_{k-1}^N). \tag{8}$$

The Markov mutations ensure that, in general, the particles $\{u_{k,i}^N\}_{i=1}^N$ do not form a subset of $\{u_{0,i}^N\}_{i=1}^N$. The particle algorithm resulting from (8) 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 or fir exploring distribution with complex dependency structures.

4 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 μ and ν 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 ν out of a particle approximation of μ. 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 μ and ν, let $\mathcal{P}(\mu,\nu)$ be the set of probability couplings between μ and ν, i.e. the convex set of probability distributions on $\mathcal{X}\times\mathcal{X}$ that admit μ and ν as marginals. For a cost function $\mathbf{c} : \mathcal{X}\times\mathcal{X} \to [0,\infty)$, the optimal transportation problem seeks

to minimize the transport cost $\mathbb{E}_\gamma[\mathbf{c}(\hat{u},\hat{v})]$, for $(\hat{u},\hat{v}) \sim \gamma$, over the set of all possible couplings $\gamma \in \mathcal{P}(\mu,\nu)$,

$$\gamma^{\mathrm{OT}} \;=\; \mathbf{argmin}\Big\{\gamma \mapsto \mathbb{E}_\gamma[\mathbf{c}(\hat{u},\hat{v})] \quad \text{with} \quad \gamma \in \mathcal{P}(\mu,\nu)\Big\}. \tag{9}$$

On an Euclidean space, a standard choice is the quadratic cost function $\mathbf{c}(u,v) = \|u-v\|^2$. For cost functions of the type $\mathbf{c}(u,v) = h(v-u)$ for a strictly convex function h, Brenier's theorem [Bre91] states that, if μ is compactly supported and has a density with respect to the Lebesgue measure, there exists a deterministic map $\mathbf{T} : \mathcal{X} \to \mathcal{X}$, uniquely defined on the support of μ, such that the optimal coupling γ^{OT} is obtained by pushing-forward the distribution μ through the deterministic function $(\mathbf{Id},\mathbf{T}) : \mathcal{X} \mapsto \mathcal{X}\times\mathcal{X}$. That is, for a test function $\varphi : \mathcal{X}\times\mathcal{X} \to \mathbb{R}$, the quantity $\gamma^{\mathrm{OT}}(\varphi)$ can also be expressed as $\mathbb{E}_\mu[\varphi(\hat{u},\mathbf{T}(\hat{u}))]$ for $\hat{u} \sim \mu$. For more general cost functions, the situation is more delicate [EG99, TW01, CFM02, Amb03].

4.1 Approximation of the Bayes operator

Consider a weighted particle approximation $\mu^N = \sum_{i=1}^N \alpha_i\,\delta(u_i^N)$ of the distribution μ and, for a potential function $\Psi : \mathcal{X} \to (0;\infty)$, the probability distribution

$$\mathscr{B}_\Psi(\mu^N) \equiv \sum_{i=1}^N \beta_i\,\delta(u_i^N) \equiv \nu^N, \tag{10}$$

with $\beta_i = \alpha_i\,\Psi(u_i^N)/[\alpha_1\,\Psi(u_1^N)+\ldots+\alpha_N\,\Psi(u_N^N)]$. The optimal coupling $\gamma^{\mathrm{OT},N}$ between μ^N and ν^N is supported on the finite set $\{(u_i^N,u_j^N)\}_{1\le i,j\le N}$ and can thus be expressed as

$$\gamma^{\mathrm{OT},N} = \sum_{i,j=1}^N \mathbf{C}_{ij}^{\mathrm{OT},N}\,\delta(u_i^N)\otimes\delta(u_j^N).$$

where $\delta(u)\otimes\delta(v)$ denotes the Dirac mass centred at $(u,v)\in\mathcal{X}\times\mathcal{X}$. Here, the coupling matrix $\mathbf{C}^{\mathrm{OT},N} \in \mathbb{R}_+^{N,N}$ is the solution of the linear programming problem that consists in minimizing the matrix functional

$$C \mapsto \sum_{i,j=1}^N C_{i,j}\times\mathbf{c}(u_i^N,u_j^N) \equiv \langle C,\mathbf{D}\rangle_{\mathrm{F}} \qquad \text{with} \qquad \mathbf{D}_{i,j} = \mathbf{c}(u_i^N,u_j^N) \tag{11}$$

over the convex set $\mathcal{P}(\alpha,\beta)$ of matrices with marginals α and β, i.e. the set of matrices $C \in \mathbb{R}_+^{N,N}$ such that $\sum_j C_{i_0,j} = \alpha_{i_0}$ and $\sum_i C_{i,j_0} = \beta_{j_0}$ for all $1 \le i_0, j_0 \le N$. In Equation (11), the quantity $\langle C,\mathbf{D}\rangle_{\mathrm{F}} = \sum_{i,j} C_{i,j}\,\mathbf{D}_{i,j}$ is the Frobenius inner product between the coupling matrix C and the *cost matrix* $\mathbf{D} \in \mathbb{R}^{N,N}$. More details are given at the end of this section.

We now describe how, once the coupling matrix $\mathbf{C}^{\mathrm{OT},N}$ has been computed, a particle approximation of the distribution $\mathscr{B}_\Psi(\mu^N)$ can be constructed: we stress that, in order to implement this method, the coupling matrix $\mathbf{C}^{\mathrm{OT},N}$ is the only quantity that needs to be computed.

For motivating the methodology, assume that the optimal coupling $\gamma^{\mathrm{OT}} \in \mathcal{P}(\mu, \nu)$ is described by a deterministic map $\mathbf{T} : \mathcal{X} \to \mathcal{X}$ and consider a test function $\varphi : \mathcal{X} \to \mathbb{R}$. Since μ^N is a particle approximation to μ, the quantity $\mu^N(\varphi \circ \mathbf{T}) = \sum_{i=1}^N \alpha_i \, \delta(\mathbf{T}(u_i^N))$ is expected to be an approximation of $\mu(\varphi \circ \mathbf{T}) = \nu(\varphi)$. Consequently, it is reasonable to expect

$$\sum_{i=1}^N \alpha_i \, \delta(\mathbf{T}(u_i^N))$$

to be a particle approximation of ν. Although the optimal transformation $\mathbf{T}$ is generally computationally intractable (i.e. it is never actually computed in our proposed method) one can resort to an approximation scheme. Note that the quantity $\mathbf{T}(u_i^N)$ can be expressed as a conditional expectation

$$\mathbf{T}(u_i^N) \;=\; \mathbb{E}[\hat{v} \mid \hat{u} = u_i^N] \qquad \text{for} \qquad (\hat{u}, \hat{v}) \sim \gamma^{\mathrm{OT}},$$

since the pair $(\hat{u}, \hat{v})$ has the same distribution as $(\hat{u}, \mathbf{T}(\hat{u}))$ for $\hat{u} \sim \mu$. This motivates the approximation

$$\mathbf{T}(u_i^N) \;\approx\; \mathbb{E}\big[\hat{v}^N \mid \hat{u}^N = u_i^N\big] = \frac{\sum_{j=1}^N \mathbf{C}_{ij}^{\mathrm{OT},N} \, u_j^N}{\sum_{j=1}^N \mathbf{C}_{ij}^{\mathrm{OT},N}} = \frac{1}{\alpha_i} \sum_{j=1}^N \mathbf{C}_{ij}^{\mathrm{OT},N} \, u_j^N \tag{12}$$

with $(\hat{u}^N, \hat{v}^N) \sim \gamma^{\mathrm{OT},N}$. The newly created particles $\{u_i^{\mathrm{OT},N}\}_{i=1}^N$ defined as

$$u_i^{\mathrm{OT},N} \equiv \frac{1}{\alpha_i} \sum_{j=1}^N \mathbf{C}_{ij}^{\mathrm{OT},N} \, u_j^N \tag{13}$$

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. In summary, the computational optimal transport executed in the SET algorithm proceeds by first solving for $\mathbf{C}^{\mathrm{OT},N}$ given the constraints described by Equation 11. In a second stage, the coupling matrix $\mathbf{C}^{\mathrm{OT},N}$ is then used to transport the particles following Equation 13. For concreteness and in accordance with the previous sections, we denote by $\mathscr{T}_\Psi$ the operator that realizes the mapping

$$\mathscr{T}_\Psi \left(\sum_{i=1}^N \alpha_i \, \delta(u_i^N) \right) \equiv \sum_{i=1}^N \alpha_i \, \delta(u_i^{\mathrm{OT},N}) = \sum_{i=1}^N \alpha_i \, \delta \left(\frac{1}{\alpha_i} \sum_{j=1}^N \mathbf{C}_{ij}^{\mathrm{OT},N} \, u_j^N \right). \tag{14}$$

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

Algorithm 1: Optimal Transportation operator $\mathscr{T}_\Psi$

Weights computation: define

$$\beta_i = \alpha_i \, \Psi(u_i^N) / [\alpha_1 \, \Psi(u_1^N) + \ldots + \alpha_N \, \Psi(u_N^N)].$$

Cost matrix: build the matrix $\mathbf{D} \in \mathbb{R}^{N,N}$ defined in (11).
Optimal Transport: compute $\mathbf{C}^{\mathrm{OT},N} = \mathbf{argmin}_{\mathcal{P}(\alpha,\beta)} C \mapsto \langle C, \mathbf{D} \rangle_{\mathrm{F}}$.
Transportation: set $u_i^{\mathrm{OT},N} = (1/\alpha_i) \sum_{j=1}^N \mathbf{C}_{ij}^{\mathrm{OT},N} \, u_j^N$ and define

$$\mathscr{T}_\Psi(\mu^N) = \sum_{i=1}^N \alpha_i \, \delta(u_i^{\mathrm{OT},N})$$

Algorithm 1 summarizes the optimal transport approach to approximating the Bayes operator that transforms a particle approximation $\mu^N = \sum_{i=1}^N \alpha_i \, \delta(u_i^N)$ of a distribution μ into a particle approximation $\mathscr{T}_\Psi(\mu^N)$ of the distribution $\nu = \mathscr{B}_\Psi(\mu)$,

$$\mu^N = \sum_{i=1}^N \alpha_i \, \delta(u_i^N) \quad \xrightarrow[\mathscr{T}_\Psi]{\text{Optimal Transport}} \quad \sum_{i=1}^N \alpha_i \, \delta(u_i^{\mathrm{OT},N}) \equiv \mathscr{T}_\Psi(\mu^N).$$

The only potentially computationally expensive step is the computation of the coupling matrix $\mathbf{C}^{\mathrm{OT},N}$. The computational costs are discussed at the start of Section 6 and we refer the reader to [PC+19] for a book-length treatment of the computational aspects associated to optimal transportation problems.

4.2 Consistency

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

Assumption 4 (Unique Deterministic Coupling)**.** *The optimal transport problem between μ and $\mathscr{B}_\Psi(\mu)$ with cost function $\mathbf{c}$ admits a unique solution γ that can be realized by a deterministic transport map $\mathbf{T} : \mathcal{X} \to \mathcal{X}$.*

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 μ and ν is weakly compact and the functional $\mu \mapsto \mathbf{E}_\mu[\mathbf{c}(u, v)]$ 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 (Regularity of the Transport Map)**.** *Let Assumption 4 holds for a deterministic map $\mathbf{T} : \mathcal{X} \to \mathcal{X}$. For any bounded and Lipschitz function $\varphi : \mathcal{X} \to \mathbb{R}$ and sequence μ^N that converges weakly to μ, we have that $\mu^N(\varphi \circ \mathbf{T}) \to \mu(\varphi \circ \mathbf{T})$.*

The continuous mapping theorem [MW43] shows that Assumption 5 is satisfied provided that the set of discontinuities of $\mathbf{T}$ has zero measure under μ. In particular, Assumption 5 holds in the case when the optimal map $\mathbf{T}$ is continuous. Theorem 6 below shows that, under mild growth and regularity assumptions on the optimal transport map $\mathbf{T} : \mathcal{X} \to \mathcal{X}$, the optimal transport scheme $\mathscr{T}_\Psi$ is consistent as the number of particles $N \geq 1$ approaches infinity.

Theorem 6. *Consider a potential function $\Psi : \mathcal{X} \to (0; \infty)$ and two probability distributions μ and $\nu = \mathscr{B}_\Psi(\mu)$ on the state space $\mathcal{X}$. Assume that Assumptions 4 and 5 are satisfied for a deterministic optimal map $\mathbf{T} : \mathcal{X} \to \mathcal{X}$. 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 μ, and such that $\mathscr{B}_\Psi(\mu^N)$ converges weakly to $\mathscr{B}_\Psi(\mu)$. If the growth assumption

$$\limsup_{N \to \infty} \quad \mu^N(u \mapsto |\mathbf{T}(u)|^p) + \mathscr{B}_\Psi(\mu^N)(u \mapsto |u|^p) < \infty, \tag{15}$$

is satisfied for some exponent $p > 1$, we have that

$$\mathscr{T}_\Psi(\mu^N) \xrightarrow{\text{w}} \mathscr{B}_\Psi(\mu) \equiv \nu. \tag{16}$$

Proof. Let $\gamma^{\text{OT},N} = \sum_{i,j} \mathbf{C}_{i,j}^N \, \delta(u_i^N) \otimes \delta(u_j^N)$ be the optimal coupling between μ^N and $\mathscr{B}_\Psi(\mu^N)$. By assumption, $\mu^N \xrightarrow{\text{w}} \mu$ and $\nu^N \equiv \mathscr{B}_\Psi(\mu^N) \xrightarrow{\text{w}} \mathscr{B}_\Psi(\mu) \equiv \nu$ and there is a unique optimal coupling γ^{OT} between μ and ν. By compactness (see, e.g. [Vil08, Corollary 5.21]), we have that $\gamma^{\text{OT},N} \xrightarrow{\text{w}} \gamma$ as $N \to \infty$.

To show the weak convergence of $\mathscr{T}_\Psi(\mu^N)$ towards ν, it suffices to prove that for any Lipschitz and bounded test function φ we have that $\mathscr{T}_\Psi(\mu^N) \to \nu(\varphi)$. Assumption 5 implies $\mu^N(\varphi \circ \mathbf{T}) \to \mu(\varphi \circ \mathbf{T}) = \nu(\varphi)$. Consequently, it suffices to show that the difference $\mathscr{T}_\Psi(\mu^N) - \mu^N(\varphi \circ \mathbf{T})$ converges to zero as $N \to \infty$, i.e.,

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

Since φ is Lipschitz, and $\sum_{j=1}^N \mathbf{C}_{ij}^N = \alpha_i^N$, it is sufficient to show that

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

Note that $\sum_{i,j}^N \mathbf{C}_{ij}^N \left|u_j^N - \mathbf{T}(u_i^N)\right| = \gamma^{\mathrm{OT},N}(F)$ with $F(u,v) = |v - \mathbf{T}(u)|$. Since $F^p(u,v) \lesssim |v|^p + |\mathbf{T}(u)|^p$, assumption (15) yields that $\limsup_N \gamma^{\mathrm{OT},N}(F^p) < \infty$. Since $\gamma^{\mathrm{OT},N} \xrightarrow{\mathrm{w}} \gamma$, the bound $\limsup_N \gamma^{\mathrm{OT},N}(F^p) < \infty$ implies that the sequence $\gamma^{\mathrm{OT},N}(F)$ converges towards $\gamma(F)$. Since $\gamma(F) = 0$, the conclusion follows. □

5 Sequential Ensemble Transform

In this section, we describe our proposed methodology, the *Sequential Ensemble Transform* (SET), prove that it is consistent in the limit of infinitely many particles, and discuss adaptation strategies that are important for practical implementations of the method.

5.1 High-level description and consistency

As in Section 3, consider a sequence $\{\mu_i\}_{i=0}^K$ of distributions that interpolates between a distribution μ_0 and the posterior distribution μ_K. For any index $1 \leq k \leq K$ we have that $(d\mu_k / d\mu_{k-1})(u) = (1/Z_k)\, \Psi_k(u)$ for a μ_{k-1}-integrable and continuous potential function $\Psi_k : \mathcal{X} \to (0, \infty)$. In this section, we assume the following.

Assumption 7. *The sequence of probability distributions $\{\mu_k\}_{k=0}^K$ is such that:*

1. *for any $0 \leq k \leq K$, the support of μ_k is bounded,*
2. *for any $1 \leq k \leq K$, the pair of distributions (μ_{k-1}, μ_k) satisfies Assumptions 4 and 5.*

Instead of constructing a sequence of particle approximations to the intermediate distributions μ_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 μ_0. Define the equally weighted particle approximations μ_k^N through the recursion formula

$$\mu_k^N = \mathscr{M}_k \circ \mathscr{T}_{\Psi_k}(\mu_{k-1}^N), \tag{17}$$

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

Theorem 8 (Consistency of the SET algorithm)**.** *Let $\{\mu_k\}_{k=0}^K$ be a sequence of distributions that satisfies Assumption 7 and consider $\{u_{0,i}^N\}_{i=1}^N \subset \mathbb{R}^d$ such that*

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

Then, for any index $1 \leq k \leq K$, the sequence of equally weighted particle approximations μ_k^N defined recursively through Equation (17) weakly converges to μ_k almost surely.

Proof. One can proceed by induction. It suffices to prove that if $\mu_{k-1}^N \xrightarrow{\text{w}} \mu_{k-1}$ almost surely then $\mathscr{M}_k \circ \mathscr{T}_{\Psi_k}(\mu_{k-1}^N) \equiv \mu_k^N \xrightarrow{\text{w}} \mu_k$ almost surely. Under Assumption (7), the support of the distribution μ_{k-1} is bounded: one can find a bounded and continuous function V_k that dominates Ψ_k and invoke Proposition 1 to see that $\mathscr{B}_{\Psi_k}(\mu_{k-1}^N) \xrightarrow{\text{w}} \mu_k$ almost surely. Furthermore, under Assumption 7 the pair (μ_{k-1}, μ_k) satisfies Assumptions (4)-(5) as well as Equation 15. Theorem 6 shows that $\mathscr{T}_{\Psi_k}(\mu_{k-1}^N) \xrightarrow{\text{w}} \mu_k$ almost surely. Finally, since the Feller Markov process M_k lets μ_k invariant, Lemma 3 yields that $\mathscr{M}_k \circ \mathscr{T}_{\Psi_k}(\mu_{k-1}^N) \xrightarrow{\text{w}} \mu_k$ 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 strategies described in Section 5.2 can be used within the SET method. In Section 6, we compare the SET approach to more standard SMC approaches.

5.2 Adaptive tempering

In complex scenarios such as Bayesian inverse problems, it is a nontrivial task to specify a sequence of distributions (2) that interpolates between a distribution μ_0 that is straightforward to sample from and the posterior distribution. Instead, we consider an adaptive annealing scheme [DBR00, MDMM10, JSDT11, ZJA16, NSPD16, SC13, KBJ14]. The reader is referred to [BJKT15, GDM+17] for a theoretical analysis of adaptive annealing methods. For notational convenience, we identify distributions with their densities, and assume that the posterior distribution μ_{post} is absolutely continuous with respect to μ_0, i.e. $d\mu_{\text{post}}/d\mu_0(u) \propto \exp[V(u)]$ for some potential function $V : \mathbb{R}^d \to \mathbb{R}$. Consider the sequence $\{\mu_k\}_{k=0}^K$ defined as

$$\frac{d\mu_k}{d\mu_0}(u) \;\propto\; \exp\left[\tau_k\, V(u)\right] \tag{18}$$

for an (inverse) temperature parameter τ_k that interpolates between $\tau_0 = 0$ and $\tau_K = 1$. In practice, it can be difficult to choose the number $K \geq 1$ of temperatures (i.e. the number of interpolating densities) and the corresponding temperatures. The adaptive scheme proceeds as follows. Assume that the particle approximation

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

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

$$\text{ESS}_k(\tau) \equiv \frac{1}{N}\frac{\left(\sum_{i=1}^N \exp\left[(\tau-\tau_k)\, V(u_{k,i}^N)\right]\right)^2}{\sum_{i=1}^N \exp\left[(\tau-\tau_k)\, V(u_{k,i}^N)\right]^2} \in [0,1]. \tag{19}$$

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

$$\tau_{k+1} = \inf\{\tau > \tau_k \ : \ \mathrm{ESS}_k(\tau) \leq \xi_{\mathrm{ESS}}\}, \tag{20}$$

the procedure stops as soon as τ_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 ξ_{ESS} close to one leads to a slow annealing, which may be computationally wasteful. On the other hand, taking ξ_{ESS} close to zero can lead to an annealing scheme that is too rapid, ultimately leading to a poor particle approximation of the posterior distribution. Except stated otherwise, we choose $\xi_{\mathrm{ESS}} = 1/2$ in the numerical experiments of Section 6.

5.3 Adaptive mutation kernels

Choosing a-priori a sequence of well-mixing Markov mutation kernels is, in most realistic scenarios, not feasible. A standard approach consists in exploiting the population $\{u_{k,i}^N\}_{i=1}^N$ of particles at temperature τ_k to estimate summary statistics of the distribution μ_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 and that lets the distribution μ_k invariant. In high-dimensional settings, this adaptive tuning of the mutation kernel is often crucial to obtaining satisfying performances. In this section, we concentrate on two classes of proposals, namely *autoregressive proposals* that do not make use of any derivative information and *Preconditioned Crank-Nicholson Langevin proposals* that can make use of gradient information for enhanced mixing properties. We refer the reader to [CLM16] and the references therein for more advanced adaptation strategies especially designed to tackle high-dimensional Bayesian inverse problems.

Autoregressive Proposals: for a mean vector $\mathbf{m} \in \mathbb{R}^d$ and a positive definite covariance matrix $\mathbf{\Gamma} \in \mathbb{R}^{d,d}$, the Markovian proposal $u \mapsto \widehat{u}$ defined as

$$\widehat{u} = \mathbf{m} + \rho\,(u - \mathbf{m}) + (1 - \rho^2)^{1/2}\,\mathcal{N}(0, \mathbf{\Gamma}) \tag{21}$$

for some scaling factor $\rho \in (0, 1)$ is reversible with respect to the Gaussian distribution with mean $\mathbf{m}$ and covariance $\mathbf{\Gamma}$. This proposal mechanism, also sometimes called the *Preconditioned Crank-Nicholson* proposal [CRSW13], can consequently be used within a standard Metropolis-Hastings scheme to efficiently explore distributions that are well approximated by a Gaussian distribution with mean $\mathbf{m}$ and covariance $\mathbf{\Gamma}$. This remark can be used to design an adaptation strategy [KBJ14] for automatically tuning the mutation kernels within SMC methods or the SET algorithm. At iteration k, right after the resampling step of a SMC method, or right after the transportation step of the SET, consider a set $\{\widetilde{u}_{k,i}^N\}_{i=1}^N$ of particles whose (equally weighted) empirical distribution approximates the distribution μ_k. In order to use an autoregressive Markov kernel (21), one can use the particles $\{\widetilde{u}_{k,i}^N\}_{i=1}^N$ to compute an approximation $\mathbf{m}_k^N$ of the mean of μ_k as well as an

approximation $\mathbf{\Gamma}_k^N$ of its covariance matrix. In high-dimensional settings, or when the number of particles is low when compared to the dimensionality of the state-space, it is customary to only consider diagonal approximations of the covariance structure: the approximate covariance matrix $\mathbf{\Gamma}_k^N$ is diagonal, with the empirical marginal variances on its diagonal. The scaling factor ρ_k^N can also be chosen adaptively. Values of $\rho_k^N \approx 1^-$ lead to conservative proposals while values of $\rho_k^N \approx 0^+$ are more likely to be rejected. Given two fixed thresholds $0 < \xi_- < \xi_+ < 1$, the scaling factor ρ_k^N can be adapted based upon the acceptance rate of the Metropolis-Hastings proposals (21). Specifically, we set $\rho_k^N = \min(1, [1+\varepsilon]\,\rho_{k-1}^N)$ if the proportion of accepted proposals falls below ξ_-, set $\rho_k^N = [1-\varepsilon]\,\rho_{k-1}^N$ if the proportion of accepted proposals is above ξ^+ and set $\rho_{k+1}^N = \rho_k^N$ otherwise. In other words, the scaling factor is augmented or decreased by an proportion $\varepsilon \in (0,1)$ depending on the acceptance rate of the MCMC proposals. In experiments presented in Section 6, we use $\xi_- = 20\%$ and $\xi_+ = 85\%$ and $\varepsilon = 20\%$.

Preconditioned Crank-Nicholson Langevin proposals: one potential drawback of the autoregressive proposals (21) is that no derivative information is exploited. Instead, Markovian proposals $u \mapsto \widehat{u}$ of the type

$$\widehat{u} = u + (1-\rho)\,\Gamma\,\nabla \log \mu_k(u) + (1-\rho^2)^{1/2}\,\mathcal{N}(0, \mathbf{\Gamma}) \tag{22}$$

can be used within a Metropolis-Hastings scheme for exploring the target density μ_k. Here, $\mathbf{\Gamma}$ is still an approximation of the covariance matrix of μ_k and $\rho \in (0,1)$ is a scaling factor. Indeed, in the case where the target density is Gaussian, this reduces to the autoregressive proposal (21). Both the scaling factor ρ and the covariance matrix $\mathbf{\Gamma}$ can be adapted throughout the evolution of a SMC or SET method. In the non-linear-PDE example of Section 6.3, we describe how gradient/Hessian information can be leveraged to adapt the covariance structure Γ.

5.4 Adaptive number of Mutations

In challenging scenarios, it is important to apply several steps of Markovian mutation at each temperature level. Nevertheless, choosing a sensible number of mutation steps a-priori is often difficult. In this section, we present an adaptive procedure for automatically selecting the appropriate number of mutation steps, inspired by the methodology first proposed in [KBJ14]. Consider the SET approach when implemented to approximate a target distribution on the state-space $\mathcal{X} \equiv \mathbb{R}^d$. Furthermore, consider $S \geq 1$ summary statistics, i.e. functions $\mathcal{S}_s : \mathcal{X} \to \mathbb{R}$ for $1 \leq s \leq S$.

At iteration $k \geq 0$, right after the resampling step of a SMC method, or right after the transportation step of the SET approach, consider a set $\{\widetilde{u}_{k,i}^N\}_{i=1}^N$ of particles whose empirical distribution approximates the distribution μ_k. Before applying the mutation kernel M_k, the summary statistics are computed, i.e. $\widetilde{s}_{k,i}^{N,s} = \mathcal{S}_s(\widetilde{u}_{k,i}^N)$, for $1 \leq i \leq N$ and $1 \leq s \leq S$. The mutation kernel M_k is then applied to the particles until the correlation along the summary statistics has fallen under a pre-determined threshold $0 < \xi_{\text{stat}} < 1$. In other words, the particles are mutated by defining $\{\widetilde{u}_{k,i}^N[p]\}_{p=0}^{p_k}$ where $\widetilde{u}_{k,i}^N[p] \sim M_k\big(\widetilde{u}_{k,i}^N[p-1], d\widehat{u}\big)$, initialized at $\widetilde{u}_{k,i}^N[p=0] = \widetilde{u}_{k,i}^N$, and the

number of mutation steps p_k is set as the smallest index $p \geq 1$ such that

$$\mathrm{Corr}\big(\{\mathcal{S}_s(\widetilde{u}^N_{k,i}[0])\}_{i=1}^N, \{\mathcal{S}_s(\widetilde{u}^N_{k,i}[p])\}_{i=1}^N\big) \;\leq\; \xi_{\mathrm{stat}} \qquad \text{for all} \qquad 1 \leq s \leq S \tag{23}$$

or when the number of iteration $p_k \geq 1$ reaches a maximum threshold. The index p_k is referred to as the adaptive *number of mutation steps*. The final mutated particles can thus be described as

$$u^N_{k,i} \sim M_k^{p_k}\big(\widetilde{u}^N_{k,i}, d\widehat{u}\big).$$

A similar approach has been employed in [KBJ14] in which the low-frequencies of a Fourier expansion is used as summary statistics. In Section 6.3, we use as summary statistics the projection of the particles along likelihood-informed directions and a threshold of $\xi_{\mathrm{stat}} = 80\%$.

5.5 Sequential Ensemble Transform: practical implementations

For completeness, we now described in more details the SET methodology when used in conjunction with the adaptation strategies discussed in Sections 5.2 and 5.3. As in Section 5.2, consider an initial distribution μ_0 that is straightforward to sample from, and the posterior distribution μ_{post} that can be expressed as $d\mu_{\mathrm{post}}/d\mu_0(u) \propto \exp[V(u)]$ for some potential $V : \mathbb{R}^d \to \mathbb{R}$. We consider tempered distributions μ_k defined as $d\mu_k/d\mu_0(u) \propto \exp[\tau_k\, V(u)]$ for a temperature parameter $\tau_k \in [0, 1]$ that is found adaptively. The entire method is summarised in Algorithm 2.

Algorithm 2: Sequential Ensemble Transform

Inputs: initial and final distributions μ_0 and μ_{post}

Output: particle approximation $(1/N)\sum_{i=1}^N \delta(u^N_{i,\star})$ of the distribution μ_{post} Set $k = 0$ and $\tau_0 = 0$ and initialize $\{u^N_{0,i}\}_{i=1}^N$ as samples from μ_0.

while $\tau_k < 1$ **do**

Evaluate $V(u^N_{k,i})$ for $1 \leq i \leq N$.

Find the next temperature τ_{k+1} through Equation (20)

Define the probability weights $w^N_{k+1,i} \propto \exp[(\tau_{k+1} - \tau_k)\, V(u^N_{k,i})]$.

Compute the cost matrix $\mathbf{D}_{i,j} = \mathbf{c}(u^N_{k,i}, u^N_{k,j})$

Compute $\mathbf{C}^{\mathrm{OT},N} = \mathbf{argmin}\; C \mapsto \langle C, \mathbf{D}\rangle_{\mathrm{F}} \in \mathbb{R}_+^{N,N}$ under the constraint

$$\sum_{j=1}^N \mathbf{C}^{\mathrm{OT},N}(i,j) = \frac{1}{N} \qquad \text{and} \qquad \sum_{i=1}^N \mathbf{C}^{\mathrm{OT},N}(i,j) = w^N_{k+1,j}.$$

Transport the particles by setting: $\widetilde{u}^N_{k+1,i} = N\, \sum_{j=1}^N \mathbf{C}^{\mathrm{OT},N}(i,j)\, u^N_{k,j}$

Use $\{\widetilde{u}^N_{k+1,i}\}_{i=1}^N$ to tune a μ_{k+1}-invariant Markov kernel $M_{k+1}(u, d\widehat{u})$.

Set $\widetilde{u}^N_{k+1,i}[0] = \widetilde{u}^N_{k+1,i}$ and $p_k = 0$.

while *criterion* (23) *not satisfied* **do**

Set $p_k \leftarrow p_k + 1$

Define: $\widetilde{u}^N_{k+1,i}[p_k] \sim M_{k+1}(\widetilde{u}^N_{k+1,i}[p_k - 1], d\widehat{u})$

Set $u^N_{k+1,i} = \widetilde{u}^N_{k+1,i}[p_k]$ and $k \leftarrow k + 1$.

6 Numerical Experiments

For PDE-constrained Bayesian inverse problems, the overall cost of the SET algorithm is dominated by PDE solves [DHJ+03]. Estimating the matrices $\mathbf{C}^{\mathrm{OT},N}$ requires solving an optimal transport problem: the standard simplex method or interior point method [PW09] directly applied to the linear program (11) scales as $\mathcal{O}(N^3)$. Faster and approximate methods are available: for example, the entropic relaxation of [Cut13] computes an ε-approximation with the cost of $\mathcal{O}(N^2/\varepsilon^3)$ [ANWR17]. Our numerical experiments show that even with $N = \mathcal{O}(10^4)$ particles, the computational overheads associated with solving optimal transport problems to obtain the optimal transport matrix $\mathbf{C}^{\mathrm{OT},N}$ is negligible when compared to the cost of computing the forward PDE solves. Consequently, for all the numerical simulations presented in this section, the approximate but more scalable methods such as the ones described in [GCPB16, Cut13] for computing discrete optimal transport schemes were not employed. Instead, the optimal transport matrices were computed through a standard simplex solver [FC17]. To operate, the SET method requires $\mathcal{O}(N \times M)$ PDE-solves where M is the total number of Markov mutations applied to each particle. In this section, we adopt the strategies described in Sections 5.2 and 5.3 and 5.4 for automatically adapting the sequence of temperatures, the Markov mutation kernels, and the number of times these Markov mutation kernels were applied. We compare the SET approach to the state-of-the-art adaptive SMC approach of [KBJ14, BJKT15]. In this section, we present three numerical experiments with increasing complexity. The first experiment investigates the influence of the mixing properties of the mutation kernels: for this purpose, the adaptive schemes used for adapting the Markov kernels, temperature ladder, and number of mutations at each temperature are switched off. The second experiment looks into the effect of the number of mutations at each temperature level. Finally, the last experiment is a relatively challenging Bayesian inverse problem. It illustrates the robustness and efficiency of the SET method when used in conjunction with automated adaptation strategies; to the best of our knowledge, the scheme using the averaged Gauss-Newton Hessian for adapting the PCNL covariance structure is new.

6.1 Scalar Target Distribution

In this section, we investigate the influence of the mixing properties of the Markov mutation kernels. We consider a one-dimensional Gaussian target distribution $\mu(du)$ defined as

$$\frac{d\mu}{d\mu_0}(u) \propto \exp\left\{-\frac{1}{\sigma^2_{\text{noise}}}(u-1/2)^2\right\} \equiv \exp\left[V(u)\right] \tag{24}$$

for a "prior" distribution μ_0 chosen as a centred Gaussian with unit variance $\sigma_0 = 1$. In the experiments presented in this section, we chose $\sigma_{\text{noise}} = 10^{-3}$: although all the quantities are Gaussian, this setting is challenging since $\sigma_{\text{noise}} \ll \sigma_0$. In order to focus on the mixing properties of the Markov mutation kernels, we fix a sequence of intermediate temperatures equally spaced on a logarithmic scale $\{\tau_k\}_{k=1}^K$ with $K = 30$. In other words, the adaptive tempering scheme presented in Section 5.2 is not used. Denote by σ_k the standard deviation of the Gaussian intermediate distribution μ_k defined as $d\mu_k/d\mu_0(u) \propto \exp[\tau_k V(u)]$. At temperature $\tau_k > 0$, the Markov mutation kernel is chosen as a Random Walk Metropolis (RWM) kernel with Gaussian

perturbations with standard deviation $\rho\,\sigma_k$, where $\rho > 0$ is used to control the mixing properties of the mutation kernels. For $\rho \ll 1$, the mutation kernels are inefficient while for $\rho \approx 1$ the mutation kernels are close to optimal.

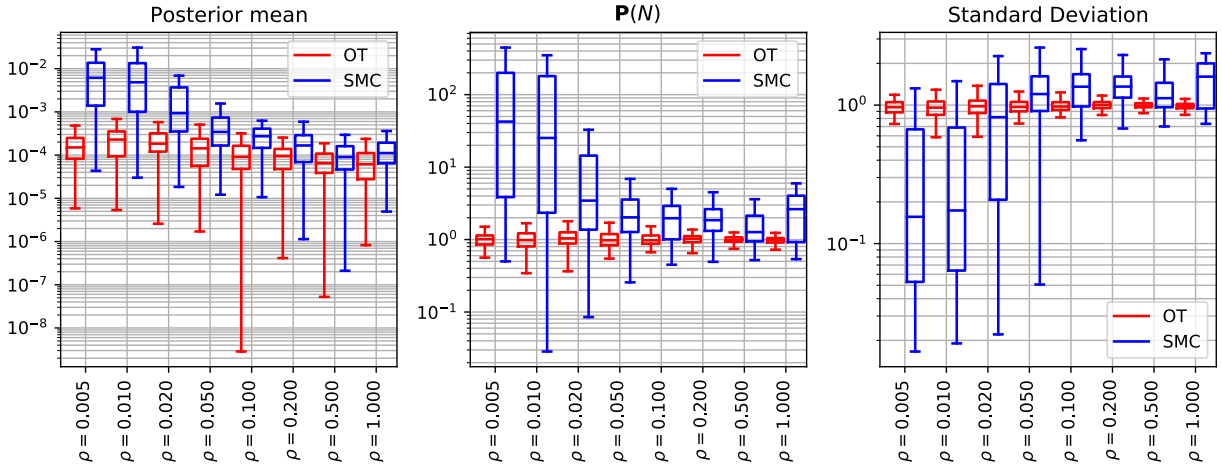

Figure 2: Target distribution (24) with $N = 10^2$ particles, no adaptation and a ladder of $K = 30$ temperatures equally spaced on a logarithmic scale. Each experiment is executed and averaged over $n = 100$ times. The scaling parameter $\rho > 0$ quantifies the quality of the Markovian mutations. **Left:** distribution of $|\widehat{m}^N_{\text{post}} - m_{\text{post}}|$ **Middle:** distribution of the quantity $\mathbf{P}(N)$ defined in (25) **Right:** distribution of the ratio $\widehat{\sigma}^N_{\text{post}}/\sigma_{\text{post}}$

Set m_{post} and σ_{post} the mean and standard deviation of the target distribution $\mu(du)$. For a particle approximation $\mu^N = (1/N)\sum_{i=1}^N \delta(u_i^N)$, we take $\widehat{m}^N_{\text{post}}$ and $\widehat{\sigma}^N_{\text{post}}$ as its mean and standard deviation. We also consider the quantity $\mathbf{P}(N)$ that equals, up to irrelevant additive and multiplicative constants, the negative log-posterior,

$$\mathbf{P}(N) \equiv \frac{1}{N}\sum_{i=1}^{N} \frac{(u_i^N - m_{\text{post}})^2}{\sigma^2_{\text{post}}}. \tag{25}$$

In this Gaussian setting and in the idealized situation when the samples $\{u_i^N\}_{i=1}^N$ are i.i.d samples from μ_{post} and $N \to \infty$, the quantity (25) converges to one. Figure 2 reports the quality of the approximation of the mean, standard deviation, and the quantity (25) when the SET and SMC methods are employed with $N = 10^2$ particles and identical conditions. For each value of ρ, the same experiment is executed $n = 100$ times. For quantifying the quality of the approximation of the posterior mean, the absolute difference $|\widehat{m}^N_{\text{post}} - m_{\text{post}}|$ is reported. For quantifying the approximation of the standard deviation, the ratio $\widehat{\sigma}^N_{\text{post}}/\sigma_{\text{post}}$ is reported. Finally, the quantity (25) is also reported: values closer to one indicate a better calibrated approximation. In this

setting, the SET approach outperforms the SMC method over all the metrics. Furthermore and as expected, as $\rho \to 0$, i.e. as the mixing of the mutation kernels gets worse, the efficiency of the SMC approach degrades. Although the theoretical results described in Section 4.2 does not explain this phenomenon, the SET method appears to continue to perform well in the regime $\rho \to 0$ in that example.

6.2 Multivariate Gaussian Target Distribution

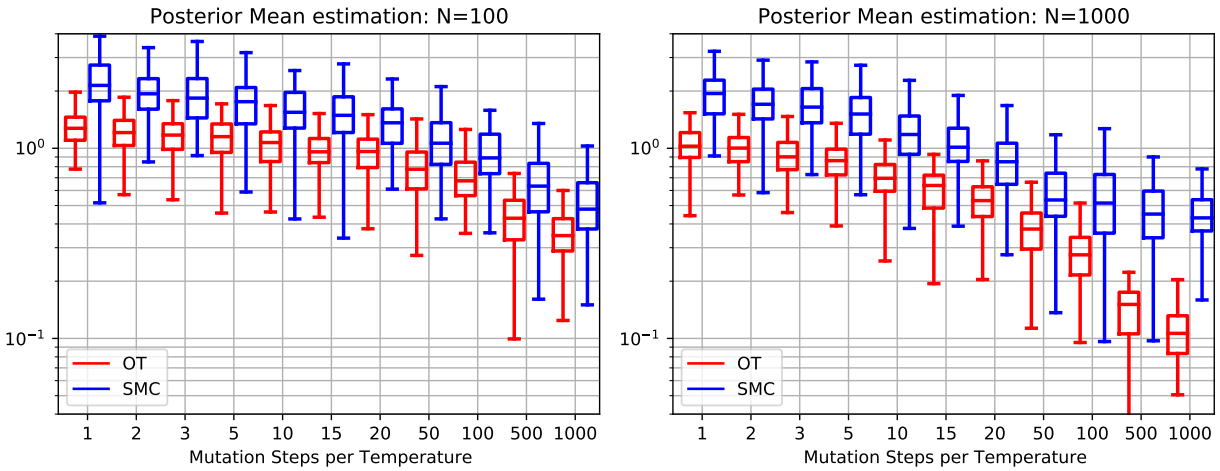

Figure 3: Estimation of the posterior mean of distribution (26) with $N = 10^2$ **(left)** and $N = 10^3$ **(right)** particles. The error $\|\widehat{m}^N_{\text{post}} - m_{\text{post}}\|$ is plotted against the number of mutation steps $p \geq 1$ at each temperature level. Each experiment is averaged over $n = 50$ runs.

In this section, we study the influence of the number of mutation steps at each temperature in a more challenging scenario. As opposed to Section 6.1, we employ the adaptive tempering scheme described in Section 5.2. Let μ_0 be a centered Gaussian distribution in $\mathbb{R}^D$ with identity covariance matrix. The Gaussian target distribution μ is defined through the change of probability measure

$$\frac{d\mu}{d\mu_0}(u) \;\propto\; \exp\left\{-\frac{1}{2}\,\langle u, \Gamma^{-1}u\rangle\right\}. \tag{26}$$

The covariance matrix $\Gamma \in \mathbb{R}^{D,D}$ is 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 $\ell > 0$. In the numerical experiments of this section, we chose $\sigma = 1$ and $\ell = 4$ and $D = 20$. Although the target distribution

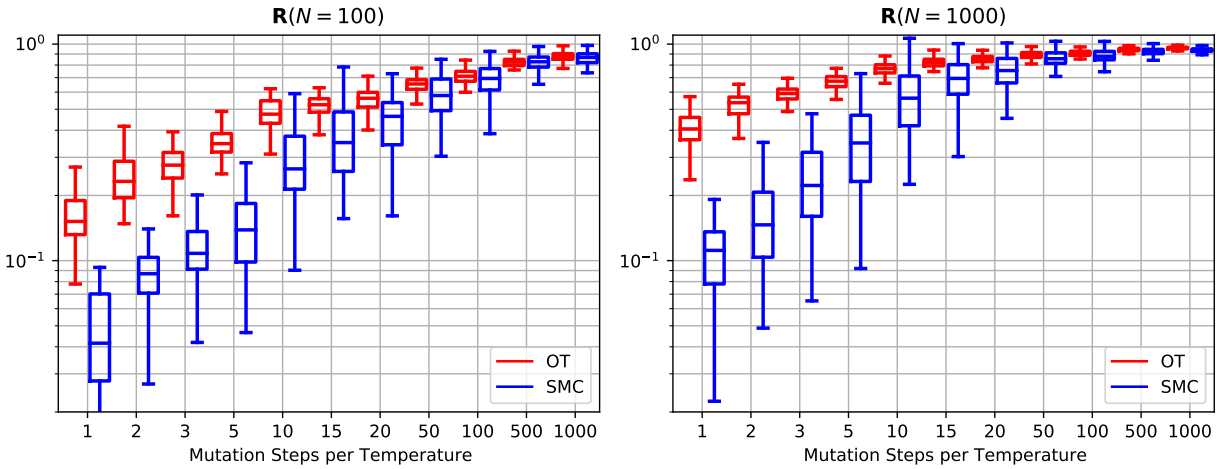

Figure 4: Estimation of the posterior standard deviation of distribution (26) with $N = 10^2$ **(left)** and $N = 10^3$ **(right)** particles. The quantity $\mathbf{R}(N)$ defined in Equation (27) is plotted against the number $p \geq 1$ of mutation steps at each temperature level. Each experiment is repeated $n = 50$ times.

is Gaussian, it is a challenging scenario since it is already relatively high-dimensional ($D = 20$) and the covariance matrix of the posterior distribution is highly ill-conditioned (and hence the posterior probability mass is concentrated in low dimensional spaces dictated by dominant eigenvectors of the covariance matrix). The SET and SMC methods have been implemented with a number of particles $N \in \{10^2, 10^3\}$ and an effective sample size threshold (20) is set to $\xi_{\text{ESS}} = 1/2$. Furthermore, we used autoregressive MCMC proposals as defined in Equation 21 with mean and covariance structure empirically estimated from the population of particles. In particular, the covariance matrix of the autoregressive proposals is assumed to be diagonal with empirical marginal variances on the diagonal (see [KBJ14] for a similar approach in the SMC context). As described in Section 5.2 the scaling parameter $\rho > 0$ was chosen adaptively to maintain MCMC mutations with acceptance rates in between the thresholds $\xi_- = 20\%$ and $\xi_+ = 80\%$.

We compared the performance of the SET and SMC methods when used with a fixed number $p \geq 1$ of mutation steps at each temperature level. Experiments were carried out for a number of mutations as low as $p = 1$ and as high as $p = 10^3$. As in Section 6.1, we report the quality of the posterior mean and posterior standard deviation. In Figure 3, for each value of the number of mutation steps $p \geq 1$, $\|\widehat{m}^N_{\text{post}} - m_{\text{post}}\|$ is averaged over 50 runs, where $\widehat{m}^N_{\text{post}} \in \mathbb{R}^D$ denotes the posterior mean estimated from the population of particles. Similarly, in Figure 4 we report the averaged ratio $\mathbf{R}(N)$ between the estimated standard deviations and the theoretical ones,

$$\mathbf{R}(N) \equiv \frac{1}{D} \sum_{d=1}^{D} \frac{\widehat{\sigma}^N_{\text{post},d}}{\sigma_{\text{post},d}}, \tag{27}$$

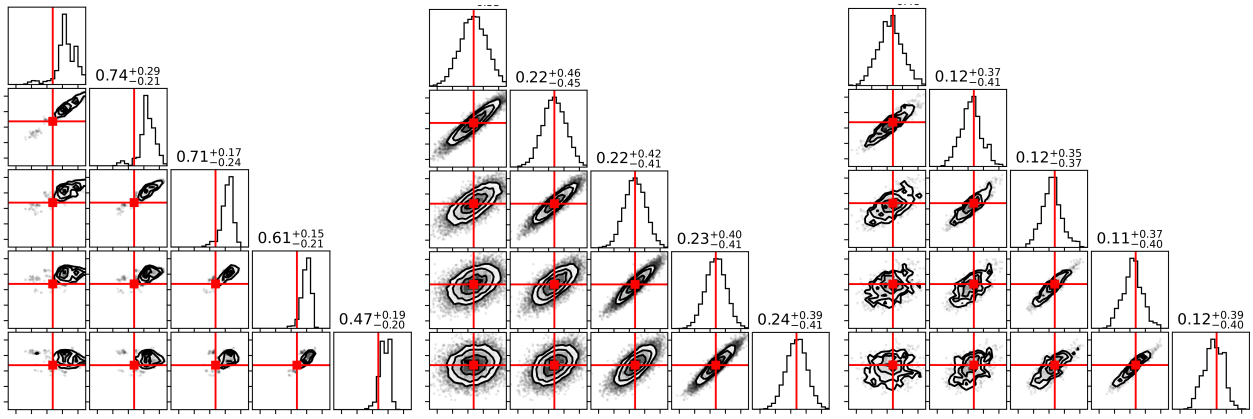

Figure 5: Density plots of the four first coordinates of the 20-dimensional target distribution (26) for SMC **(left)** and SET **(right)** with $N = 10^4$ particles and $p = 20$ mutations at each temperature. **Middle** is the density plot of $N = 10^4$ independent samples from the target distribution.

where $\sigma_{\text{post},d}$ is the marginal standard deviation in the d-th dimension and $\widehat{\sigma}^N_{\text{post},d}$ is its estimate obtained from the population of particles. Similarly to Section 6.1, we observe that the SET method appears to be more robust when the number of mutations p is very low. As p increases, the difference between the two methods progressively disappears and $\mathbf{R}(N) \to 1$ for both methods.

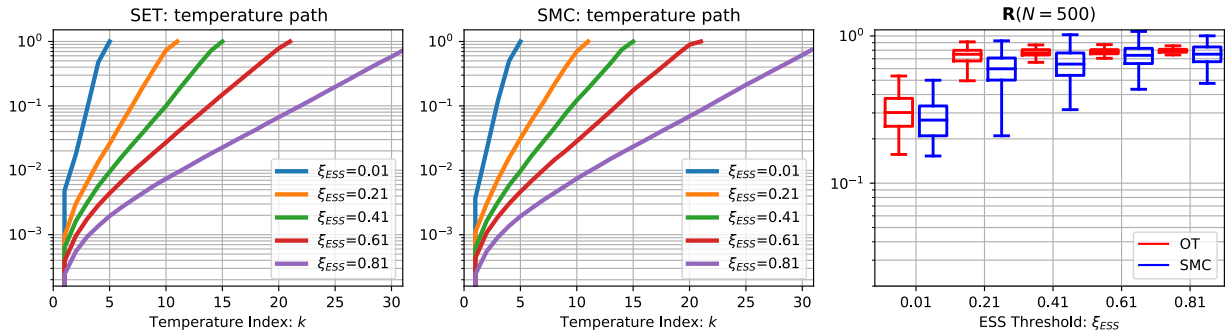

Figure 6: Temperature trajectories for the target distribution (26) using the SET **(left)** and the SMC **(center)** methods with $N = 500$ particles and $p = 20$ mutations at each temperature: each trajectory corresponds to a different effective sample size threshold $\xi_{\text{ESS}} \in \{1\%, 21\%, 41\%, 61\%, 81\%\}$. The plot on the **(right)** displays the the $\mathbf{R}(N)$ statistics, which quantifies the approximation of the posterior mean, for each value of ξ_{ESS}.

Figure 5 shows the result with $N = 10^4$ particles and $p = 20$ mutations for each temperature. The marginal pairwise distribution of the first four dimensions are displayed: although for $p = 20$

neither SMC nor SET produces an entirely satisfactory approximation of the target distribution, it is qualitatively visible that the SET produces an approximation that is closer to the correct distribution.

Finally, in order to gain some understanding of the influence of the effective sample size threshold ξ_{ESS} on the sequence of temperatures, as well as to study the sequence of temperatures adaptively chosen by the SMC and SET methods, Figure 6 displays the temperature paths for the SMC and SET methods. As expected, larger values of the effective sample size threshold ξ_{ESS} lead to a slower increase of the (inverse) temperature parameter. Furthermore, low values of the effective sample size threshold results in a loss of accuracy. This phenomenon is well understood for SMC methods since lowering ξ_{ESS} exacerbates particle degeneracy but more investigations are required for SET to understand in more details the mechanisms. Figure 6 also shows that, except at the very start of the algorithm, the (inverse) temperature increases roughly linearly on a logarithmic scale. Furthermore, when the number of mutations per temperature is fixed, as is done in this example, the SET and SMC temperature trajectories are very close to each other. This remark is important since it ensures that the numerical results presented in Figures 3 and 4 are fair, that is, for each number $p \geq 1$ mutations per temperature, the computational budgets used by the SMC and SET methods are equivalent.

6.3 Bayesian Inverse Problem

In this section, we test the SET method for inference in a Bayesian inverse problem governed by a Partial Differential Equation (PDE). More specifically, we consider the following Poisson PDE on the unit disk $\Omega \subset \mathbb{R}^2$,

$$\nabla \cdot (e^z \nabla f)(x) = h(x) \qquad \text{for} \qquad x \in \Omega, \tag{28}$$

for a known source term $h : \Omega \to \mathbb{R}$, Dirichlet boundary conditions $f(x) = 0$ for $x \in \partial\Omega$ and a temperature field $f : \Omega \to \mathbb{R}$. We are interested in reconstructing $z : \Omega \to \mathbb{R}$, the log-conductivity field, from noisy observations collected at locations $x_1, \ldots, x_K \in \Omega$ modelled as $d_i = f(x_i) + \eta_i \in \mathbb{R}$ with independent Gaussian random noises $\eta_1, \ldots, \eta_K$ centered at 0 and with variance σ^2_{noise}. We assume a Gaussian prior distribution μ^z_{prior} on the log-conductivity field $z : \Omega \to \mathbb{R}$ with Matern covariance structure [LRL11]. Draws from this prior distribution can be generated by solving the elliptic PDE

$$\big(\kappa^2 - \Delta\big) z(x) = w(x) \qquad \text{for} \qquad x \in \Omega \tag{29}$$

with vanishing Dirichlet boundary conditions. The right-hand-side of Equation (29) is the realization $w : \Omega \to \mathbb{R}$ of a Gaussian white noise process on the domain Ω. For the numerical experiments, the scale parameter κ was set to 10^{-1}. The posterior distribution $\mu^z_{\text{post}}(dz)$ on the log-conductivity fields reads

$$\frac{d\mu^z_{\text{post}}}{d\mu^z_{\text{prior}}}(z) \;\propto\; \exp\left\{ -\frac{1}{2\,\sigma^2_{\text{noise}}} \sum_{i=1}^{K} \big(d_i - \mathcal{F}(z)[x_i]\big)^2 \right\} \tag{30}$$

where $\mathcal{F}$ is the parameter-to-observable map that associates to each log-conductivity field z the corresponding temperature field $f = \mathcal{F}(z)$ obtained by solving the PDE (28). In our experiments, we assume that the standard deviation of the additive Gaussian noise is known, i.e. $\sigma_{\text{noise}} = 10^{-2}$. The location of the observations, the ground truth temperature and log-conductivity fields are depicted in Figure 7. Here, the ground truth log-conductivity field was obtained as a draw from the prior distribution (29).

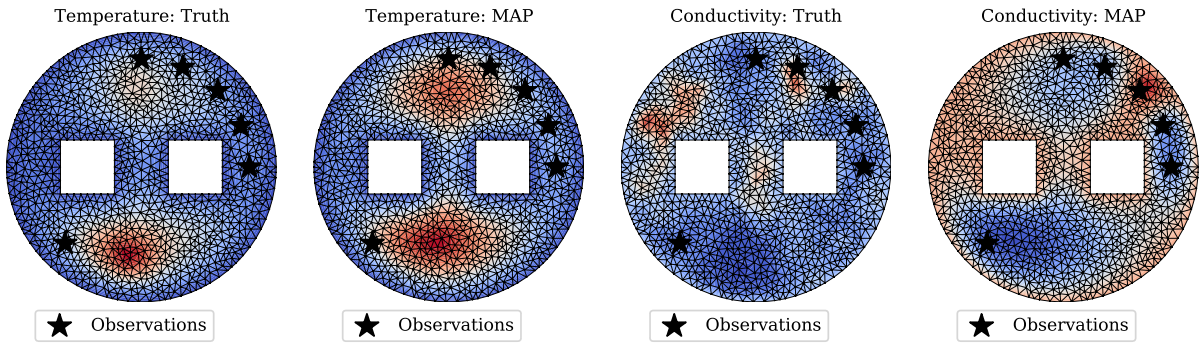

Figure 7: Ground truth and MAP estimates of the temperature field $f : \Omega \to \mathbb{R}$ and log-conductivity field $z : \Omega \to \mathbb{R}$ in the inverse problem (28). The prior distribution on the log-conductivity field is a Gaussian field with vanishing mean and Mattern covariance structure (29).

6.3.1 Discretization and parametrization

The PDE (28) was discretized with the Finite Element Method (FEM) implemented on a mesh $\mathcal{M}$ with $M = 1170$ nodes, as shown in Figure 7, using *FEniCS* [DHJ+03]. In the remainder of this section, we consequently approximated all functions defined on the domain Ω with their projection on the finite element function space with piecewise linear basis functions $e_i : \Omega \to \mathbb{R}$ for $1 \leq i \leq M$. In other words, the function space (infinite dimensional) Bayesian posterior described in Equation (30) is approximated by a M-dimensional posterior distribution. To avoid the notational burden, we use the same notations to refer to the original (infinite dimensional) quantities and their FEM approximations. Similarly, the notation $\mathcal{F}$ refers to both the original forward operator and its discretized version. If $\mathbf{M}$ denotes the mass matrix associated to the FEM basis $\{e_i\}_{i=1}^M$, the discretization of a Gaussian white noise on Ω can be realized as $w(x) = w_1 e_1(x) + \ldots + w_M e_M(x)$ where $\boldsymbol{w} = (w_1, \ldots, w_M) \in \mathbb{R}^M$ is a realization of a centered Gaussian random variable with covariance matrix $\mathbf{M}^{-1}$. We used a (sparse) Cholesky decomposition $\mathbf{M} = \mathbf{L}\mathbf{L}^\top$ and expressed the white noise vector as solution of the linear system $L^\top \boldsymbol{w} = \mathbf{u}$ where $\mathbf{u} \in \mathbb{R}^M$ is the realization of a centered standard Gaussian distribution with identity covariance matrix. For convenience, denote by $\Phi : \mathbb{R}^M \to \mathbb{R}^M$ the (linear) operator that maps $\mathbf{u}$ to the corresponding log-conductivity field. In other words, the function $z(x) = z_1 e_1(x) + \ldots + z_M e_M(x)$, with $(z_1, \ldots, z_M) = \mathbf{z} = \Phi(\mathbf{u}) \in \mathbb{R}^M$, is

the solution obtained through the FEM of the PDE (29) with right-hand-side represented by $\boldsymbol{w} = (L^\top)^{-1}\,\mathbf{u}$. When implementing the SET and SMC methods, the log-conductivity field is parametrized through the quantity $\mathbf{u}$. This parametrization has the advantage of corresponding to a standard isotropic Gaussian prior distribution with identity covariance matrix and a posterior distribution that is close to a standard Gaussian distribution except along a few data-informed directions [FWA+11, BTBG+12, BTGMS13a, CLM16]. These properties lead to Markov mutation kernels that are easier to tune and automatically adapted. In this parametrization, the $\mathbb{R}^M$-valued posterior density $\mu^{\mathbf{u}}_{\text{post}}(\mathbf{u})$ reads

$$\mu^{\mathbf{u}}_{\text{post}}(\mathbf{u}) \propto \exp\left\{-\frac{1}{2}\|\mathbf{u}\|^2 - \frac{1}{2\,\sigma^2_{\text{noise}}}\sum_{i=1}^{K}\left(d_i - \mathcal{F}(z)[x_i]\right)^2\right\} \propto \mu^{\mathbf{u}}_0(\mathbf{u})\,\exp\left[V(\mathbf{u})\right] \tag{31}$$

where $\mu^{\mathbf{u}}_0$ is the density of a centered standard isotropic Gaussian distribution in $\mathbb{R}^M$ and $V(\mathbf{u}) = -(1/2)\sigma^{-2}_{\text{noise}}\sum_{i=1}^{K}\left(d_i - \mathcal{F}(z)[x_i]\right)^2$ is the negative of the *data-misfit* functional.

For implementing the SET method, a cost matrix is needed. In order to take into account the geometry of the problem, when the SET method is implemented with N particles $\{\mathbf{u}^N_i\}_{i=1}^N$, the costs matrix $\mathbf{D} \in \mathbb{R}^{N,N}_+$ is defined as follows. The entry $\mathbf{D}_{i,j}$ is set to the squared $L^2(\Omega)$ distance between the log-conductivity fields associated to the particles $\mathbf{u}_i$ and $\mathbf{u}_j$,

$$\mathbf{D}_{i,j} = \langle \mathbf{u}^N_j, \mathbf{M}\,\mathbf{u}^N_j \rangle. \tag{32}$$

6.3.2 Adaptive scheme

To automate the choice of the number of mutation steps at each temperature, the adaptive scheme presented in Section 5.4 is used. In order to obtain meaningful summary statistics, we considered a data-driven approach. First, the *maximum a posterior* (MAP) estimate $\mathbf{u}_{\text{MAP}}$ was obtained by minimizing the negative log-posterior density $\mathbf{u} \mapsto -\log \mu^{\mathbf{u}}_{\text{post}}(\mathbf{u})$: gradients were computed with the adjoint method and a standard L-BFGS minimization procedure was used. Figure 7 displays the MAP estimate as well as the ground truth. Then, we formed a Gauss-Newton approximation[1] $\mathcal{H}_{\text{GN}}(\mathbf{u}_{\text{MAP}})$ of the Hessian to $-\log \mu^{\mathbf{u}}_{\text{post}}$ at the MAP,

$$\mathcal{H}_{\text{GN}}(\mathbf{u}_{\text{MAP}}) = \mathbf{I} + \frac{1}{\sigma^2_{\text{noise}}}\sum_{i=1}^{K}\left(\nabla_{\mathbf{u}}\mathcal{F}(z)[x_i]\right)\left(\nabla_{\mathbf{u}}\mathcal{F}(z)[x_i]\right)^\top \in \mathbb{R}^{M,M}. \tag{33}$$

On the right-hand-side of (33), all the gradient terms are evaluated at $\mathbf{u} = \mathbf{u}_{\text{MAP}}$. The eigenvectors $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_K \in \mathbb{R}^M$ corresponding to the K dominating eigenvalues of the Gauss-Newton Hessian $\mathcal{H}_{\text{GN}}(\mathbf{u}_{\text{MAP}})$ span the directions along which the collected data are the most informative and along which the posterior distribution differs most from the prior distribution [BWG+08, BTG12]. Finally, we chose $S = K$ summary statistics defined as $\mathcal{S}_k(\mathbf{u}) \equiv \langle \boldsymbol{v}_k, u \rangle$. Figure 6.3.2 shows the $K = 6$ directions $\Phi(\boldsymbol{v}_k)$ for $1 \leq k \leq K$.

[1] Note that the Gauss-Newton approximation includes a term corresponding to the Gaussian prior.

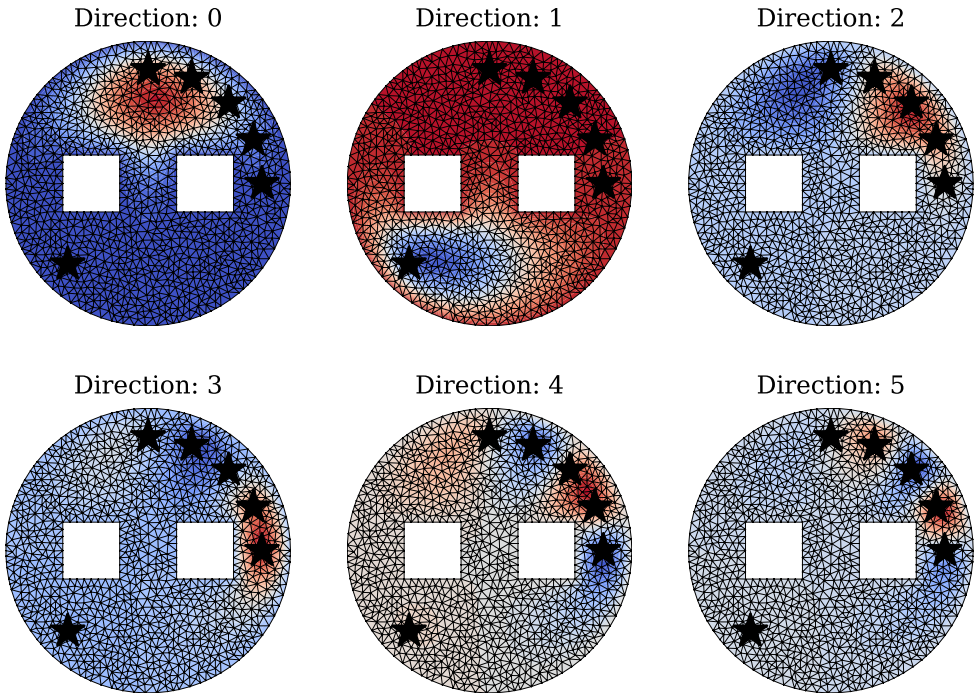

Figure 8: Directions $\{\Phi(\boldsymbol{v}_k)\}_{k=1}^K$ associated to the $K = 6$ dominating eigenvectors $\{\boldsymbol{v}_k\}_{k=1}^K$ of the Gauss-Newton Hessian $\mathcal{H}_{\mathrm{GN}}(\mathbf{u}_{\mathrm{MAP}})$ defined in Equation (33).

We used Preconditioned Crank-Nicholson Langevin (PCNL) proposals, as described in Section 5.3, for mutating the particles: the scaling parameter $\rho \in (0, 1)$ was adapted so as to maintain an acceptance probability in between $\xi_- = 20\%$ and $\xi_+ = 80\%$. The structure of the covariance Γ of the PCNL proposals was also chosen adaptively. When exploring the density $\mu_k(\mathbf{u}) \propto \mu_0^{\mathbf{u}}(\mathbf{u}) \exp[\tau_k V(\mathbf{u})]$ at temperature τ_k, the particle system $\{\mathbf{u}_i^N\}_{i=1}^N$ was used to approximate the averaged Gauss-Newton Hessian $(1/N)\sum_{i=1}^N \mathcal{H}_{\mathrm{GN}}^{\tau_k}(\mathbf{u}_i^N)$ where

$$\mathcal{H}_{\mathrm{GN}}^{\tau_k}(\mathbf{u}) = \mathbf{I} + \frac{\tau_k}{\sigma^2_{\mathrm{noise}}} \sum_{i=1}^K (\nabla_{\mathbf{u}} \mathcal{F}(z)[x_i])(\nabla_{\mathbf{u}} \mathcal{F}(z)[x_i])^\top \in \mathbb{R}^{M,M}. \tag{34}$$

In the setting when the prior distribution is Gaussian and the forward map is linear, the posterior distribution is also Gaussian and the averaged Gauss-Newton Hessian equals the precision of this Gaussian posterior distribution. This motivates the use of the averaged Gauss-Newton Hessian—which is guaranteed to be positive definite—as the inverse covariance structure for the

noise used within the PCNL proposals. To summarize, at temperature τ_k and right after the transportation step when using the SET approach, or right after the resampling step when using SMC, the covariance $\widehat{\Gamma}_k$ used within the PCNL proposals was defined as

$$\widehat{\Gamma}_k = \left\{ \frac{1}{N} \sum_{i=1}^{N} \mathcal{H}_{\mathrm{GN}}^{\tau_k}(\mathbf{u}_i^N) \right\}^{-1} \tag{35}$$

where $\{\mathbf{u}_i^N\}_{i=1}^N$, again, denotes the current particle population.

6.3.3 Results

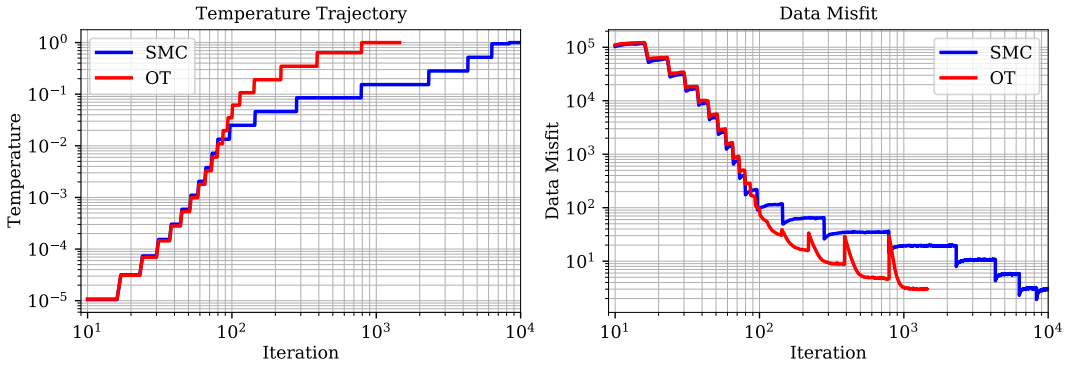

Figure 9: Bayesian Inverse Problem (28): Trajectories for Temperature **(left)** and averaged data-misfit functional **(right)** using the SET and SMC methods with $N = 2.10^3$ particles, adaptive PCNL mutation kernels, an adaptive temperature scheme, and an adaptive number of mutation steps at each temperature.

We implemented the SET and SMC approaches with $N = 2.10^3$ particles with identical conditions on a server with 20 computing cores, one for each particle to be computed in parallel. The initial distribution was chosen as the prior, i.e. $\mu_0^{\mathbf{u}}(\mathbf{u}) \propto \exp\left[-(1/2)\|\mathbf{u}\|^2\right]$. Furthermore, the schemes presented in Sections 5.2, 5.3 and 5.4 were used to automatically adapt the ladder of temperatures, the Markov mutation kernels, and the number of mutations.

The ground truth was obtained by running 20 Preconditioned Crank-Nicholson Langevin MCMC simulations (in parallel) for $L = 10^7$ iterations, each of the runs was initialized from independent draws using the Gauss-Newton approximation described in Section 6.3.2. Convergence was checked by verifying that the marginal means, variances, and summary statistics described in Section 6.3.2 agreed among the 20 chains.

Figure 9 (left) shows the trajectories of the temperatures for both the SET and SMC methods as a function of the total number of mutation steps. In this example, the Markov mutation steps are more computationally expensive, by at least an order of magnitude, than all the other computational overheads. Consequently, the total number of mutation steps is roughly proportional

to the wall-clock compute time. Note that the SET method requires almost an order magnitude less iterations to converge since, as numerically observed in Sections 6.1 and 6.2, it is less affected by particle degeneracy. Consequently, it is able to more efficiently adapt the Markov mutation kernels through the adaptation strategy (35). Figure 9 (right) displays the averaged value of the data-misfit functional $(-1/N)\sum_{j=1}^{N} V(\mathbf{u}_j^N)$ as a function of the total number of mutation steps. As displayed in Figure 10, the estimation of the posterior marginal mean and standard deviation produced by the SMC and SET methods are equally good, and agree well with the MCMC simulations.

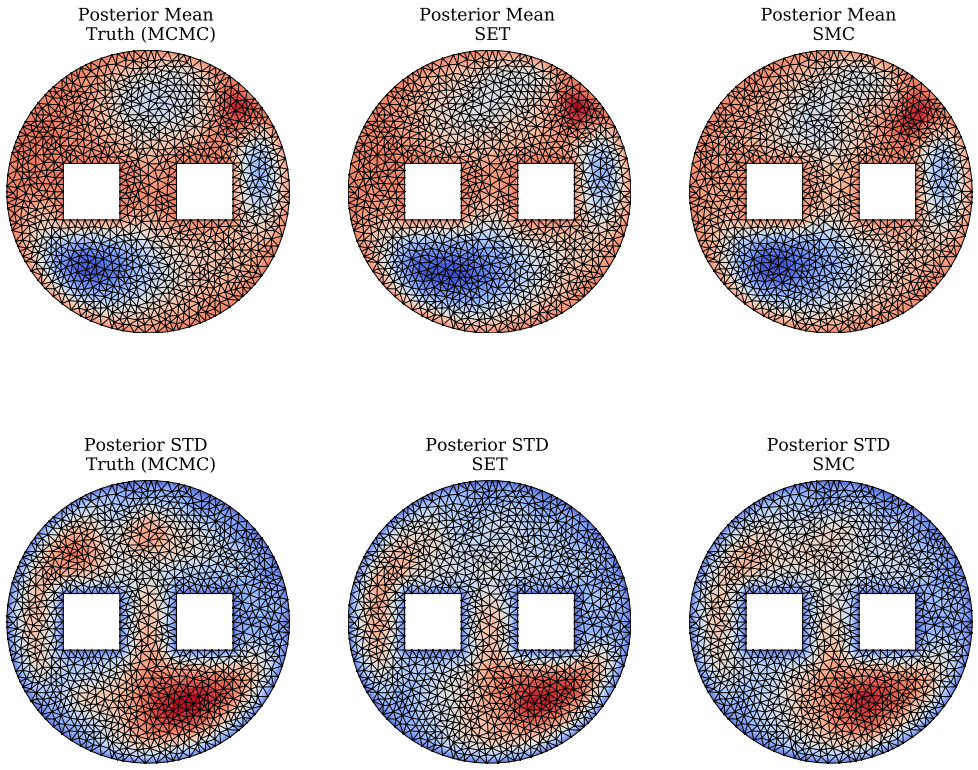

Figure 10: Bayesian Inverse Problem (28). **(First row:)** posterior mean obtained from MCMC **(left)** and SET **(center)** and SMC **(right)**. **(Second row:)** posterior marginal standard deviations obtained from MCMC **(left)** and SET **(center)** and SMC **(right)**. The SET and SMC methods were used with $N = 2.10^3$ particles, adaptive PCNL mutation kernels, an adaptive temperature scheme, and an adaptive number of mutation steps at each temperature.

We conclude this section with a brief discussion of effective sample size computations. Although there have been a few recent and important methodological advances in this area [CL+13, LW18, OD+19, DG19], it is fair to say that it is still difficult reliably to evaluate the effective sample size for interacting particles methods such as SMC or the SET. It is worth emphasizing that the effective sample size functional defined in Equation (19) is only used for adapting the temperature

ladder: it is not designed, nor should be used, to provide a reliable estimate of the variability of the quantities derived from a particle system.

7 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. Our numerical results indicate that the SET method, by relying on transportation methods instead of a resampling scheme, is less affected than SMC by particles degeneracy and is able to better exploit the particles system to adapt the mutation kernels. Although our theoretical results provide consistency guarantees, they do not quantify nor explain the empirical gains observed when comparing the SET to standard SMC approaches.

Acknowledgement: AM and TBT 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), the Defense Threat Reduction Agency (grant HDTRA1-18-1-0020), and a 2018 ConTex award. AM 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.

[ANWR17] Jason Altschuler, Jonathan Niles-Weed, and Philippe Rigollet. Near-linear time approximation algorithms for optimal transport via sinkhorn iteration. In *Advances in neural information processing systems*, pages 1964–1974, 2017.

[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.

[BTBG+12] Tan Bui-Thanh, Carsten Burstedde, Omar Ghattas, James Martin, Georg Stadler, and Lucas C. Wilcox. Extreme-scale UQ for Bayesian inverse problems governed by PDEs. In *SC12: Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, 2012.

[BTG12] Tan Bui-Thanh and Omar Ghattas. Analysis of the Hessian for inverse scattering problems: I. Inverse shape scattering of acoustic waves. *Inverse Problems*, 28(5):55001, 2012.

[BTGMS13a] T 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.

[BWG+08] O. Bashir, K. Willcox, O. Ghattas, B. van Bloemen Waanders, and J. Hill. Hessian-based model reduction for large-scale systems with initial condition inputs. *International Journal for Numerical Methods in Engineering*, 73:844–868, 2008.

[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.

[CL+13] Hock Peng Chan, Tze Leung Lai, et al. A general theory of particle filters in hidden markov models and some applications. *The Annals of Statistics*, 41(6):2877–2904, 2013.

[CLM16] Tiangang Cui, Kody JH Law, and Youssef M Marzouk. Dimension-independent likelihood-informed mcmc. *Journal of Computational Physics*, 304:109–137, 2016.

[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.

[CRSW13] 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*, 28(3):424–446, 2013.

[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.

[DDFG01] 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.

[DG19] Qiming Du and Arnaud Guyader. Variance estimation in adaptive sequential monte carlo. *arXiv preprint arXiv:1909.13602*, 2019.

[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] Rémi Flamary and Nicolas Courty. POT python optimal transport library, 2017.

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

[FWA+11] H. P. Flath, L. C. Wilcox, V. Akçelik, J. Hill, B. van Bloemen Waanders, and O. Ghattas. Fast Algorithms for Bayesian Uncertainty Quantification in Large-Scale Linear Inverse Problems Based on Low-Rank Partial Hessian Approximations. *SIAM Journal on Scientific Computing*, 33:407–432, 2011.

[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.

[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.

[KBJ14] 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.

[LRL11] Finn Lindgren, Håvard Rue, and Johan Lindström. An explicit link between gaussian fields and gaussian markov random fields: the stochastic partial differential equation approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(4):423–498, 2011.

[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.

[LW18] Anthony Lee and Nick Whiteley. Variance estimation in the particle filter. *Biometrika*, 105(3):609–625, 2018.

[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.

[OD+19] Jimmy Olsson, Randal Douc, et al. Numerically stable online estimation of variance in particle filters. *Bernoulli*, 25(2):1504–1535, 2019.

[PC+19] Gabriel Peyré, Marco Cuturi, et al. Computational optimal transport: With applications to data science. *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.

[PW09] Ofir Pele and Michael Werman. Fast and robust earth mover's distances. In *2009 IEEE 12th International Conference on Computer Vision*, pages 460–467. IEEE, 2009.

[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.

[Wil91] David Williams. *Probability with Martingales*. Cambridge University Press, 1991.

[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.