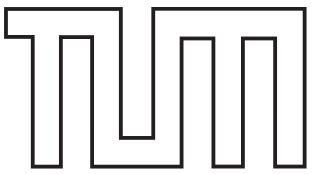

Technische Universität München
Department of Mathematics

Bachelor's Thesis

Sequential Monte Carlo for time-dependent Bayesian Inverse Problems

Matthieu Bulté

Supervisor:	Ullmann, Elisabeth; Prof. Dr. rer. nat.
Advisor:	Latz, Jonas; M.Sc
Submission Date:	15. June 2018

With my signature below, I assert that the work in this thesis has been composed by myself independently and no source materials or aids other than those mentioned in the thesis have been used.

München, May 5, 2018

Place, Date

Signature

Abstract

Titel auf Englisch wiederholen.

Es folgt die englische Version der Kurzfassung.

Contents

t hings that could be done:

1. add examples after definitions
2. more structure in first chapter

1 Introduction

The study of complex systems is often done through mathematical modelling, allowing the simulation, analysis and prediction of their behaviour. These mathematical models require input parameters, for which only limited or no information is known. Finding these parameters from measurements of the system is called the *inverse problem*. Since measurements are often noisy or sparse, and the mathematical models can be complex and expensive to evaluate, developing sound and efficient mathematical frameworks to treat the inverse problem is a complicated task.

Two prominent classes of methods for attempting to address this problem are the *maximum likelihood* methods and the *Bayesian* methods. In maximum likelihood methods, the solution of the inverse problem is given as the maximizer of the likelihood of the observed data. In Bayesian methods, the system is re-modeled probabilistically with random variables. This solution is given as a marginalization of the model by the observed data using Bayes' formula, as first developed by Laplace [Lap20]. A theoretical and practical comparison of the two methods is given by Kaipio and Somersalo [KS06], including a broad introduction to solving inverse problems found in science and engineering.

The Bayesian approach is a very general modelling and inference framework allowing to address very different kinds of statistical problems. The work by Gelman et al. [GCS+14] gives a broad introduction to the field of *Bayesian data analysis*. Based on the framework presented by Stuart [Stu10] on Bayesian methods for inverse problems, we focus on the application of Bayesian inference to time-dependent inverse problems, called *Bayesian filtering*. We will demonstrate that under weak model assumptions, the solution can be shown to be *well-posed*, using a definition of well-posedness similar to Hadamard's [Had02].

Very often, the solution of an inverse problem given in the Bayesian framework does not admit any analytical solution. Since one is interested in obtaining summarized statistics about the solution of the inverse problem, such as mean and variance, numerical approximations will involve computing integrals over the parameter space. Since volumes grow exponentially with the number of dimensions, classical methods of numerical integration cannot be used for high-dimensional problems. This phenomenon is called the *curse of dimensionality*.

Fortunately, other numerical approximations were developed that do not suffer from the curse of dimensionality. Typically, such approximations work by generating pseudo-random values distributed according to the posterior distribution and use them to approximate the hard integral. Some variation of the law of large numbers will then provide dimensionality-free error bounds, making these methods suited for high-dimensional problems. A common class of algorithms falling in this category are the *Markov Chain Monte Carlo* (MCMC) methods, presented by Metropolis et al. [MRR+53] for a specific class of problems, and later extended to the general case by Hastings [Has70]. While having dimensionality-free error bounds, MCMC algorithms often need a lot of knowledge and tunning to properly operate.

A simpler method to operate is *importance sampling* (IS) [RC05], where the sampling is done by choosing an auxiliary distribution that is similar to the target distribution, but from which direct sampling is easier. The discrepancy between the generated samples and a sample generated from the posterior distribution is then corrected by assigning correction weights to the values of the sample. However, choosing an auxiliary distribution that is close to the target distribution is not always possible, and failling to do so results in a poor estimation of the posterior distribution.

Sequential Monte Carlo (SMC) [DMDJ06] is a method merging ideas of MCMC and IS samplers in an attempt to solve major problems found in these other two methods. This sampler was created to approximate sequences of distributions, such as those found in data assimilation problems. However, it can also be used on an artificial sequence of distributions to interpolate between a simple initial auxiliary distribution and the true posterior. This is done by Beskos et al. [BJMS15] for approximating the solution of a Bayesian inverse problem associated to elliptic PDEs. By drawing parallels to particle physics, Del Moral [DM13, DM04] provides convergence results of the algorithm that will be presented in this thesis.

In their work, Allmaras et al. [ABL$^+$13] give a case study-based introduction to the whole process of Bayesian techniques for solving inverse problems. This thesis will follow a similar approach, by structuring itself around a simple time-dependent Bayesian filtering problem. The system studied is the simple pendulum, an idealized model for a pendulum in which the mass of the pendulum and the air friction are ignored. It can be described by a second-order, non-linear differential equation with a parameter representing the *gravitational acceleration*. We will describe the model of the pendulum, together with the inversion task of estimating the gravitational acceleration from a set of measurements taken in an experiment.

The rest of the thesis is structured as follows. In Section 2, we describe time-dependent inverse problems and Bayesian filtering. Moreover, we show how to formulate the pendulum problem in the Bayesian framework. In Section 3, we present the construction of the SMC algorithm and show the parallels to IS and MCMC. We also present a proof of convergence of the algorithm and discuss possible extensions. We conclude the section by computing and comparing numerical solutions to the pendulum problem. Finally, Section 4 discusses other application areas and current research on SMC algorithms.

2 Bayesian Filtering

2.1 Overview

This section will introduce and describe the Bayesian for solving time-dependent inverse problems, laying out the theoretical foundations of the taken approach. In Section 2.3 we reformulate the definition of inverse problems for the finite-dimensional time-dependent case. We then present the pendulum problem, which will be studied along the whole thesis to illustrate important ideas. In Section 2.4 we present first present the classical approach for solving inverse problems and the challenges it encounters with noisy data. We next present the Bayesian approach, and show how it incorporates prior information about the structure of the problem to address uncertainty. We then adapt the definition of the pendulum problem to the Bayesian framework. Finally, Section 2.5 presents important results, including a characterization of the class of well-posed inverse problems. We conclude the section by proving that the pendulum problem is well-posed.

2.2 Set-Up

We study parametrized models for which the value of the *parameter* $u \in X$ is unknown or uncertain. To model the behaviour of the system, we introduce *forward response operators* $\mathcal{G} : X \to Y$ mapping values of the *parameter space* to the *data space*, assuming both spaces to be finite-dimensional vector spaces.

We consider a real system described by $\mathcal{G}$ and the true parameter $u^* \in X$, and assume that *observations* $y_{obs} \in Y$ of the system are available from measurements. The data y is then the image of the true parameter under the mapping $\mathcal{G}$. Due to noise present in the data, we obtain the following approximate model

$$y_{obs} \approx \mathcal{G}(u^*), \tag{2.1}$$

and the question *To which extent can we find the inverse of the data y_{obs} under the forward response operator $\mathcal{G}$?* Answering this question is known as the *inverse problem.*

In this thesis, we are interested in solving the inverse problem by incrementaly building up knowledge about the unknown. To do this, we decompose the data and consider a growing sequence of observations made available for analysis. The advantages of this approach are twofold. First, it allows to perform inference on running dynamical systems and update our knowledge as more measurements become available. Second, it can also be used for non-dynamical problems in the hope to transform one hard problem in a sequence of smaller problems that can hopefully be solved more efficiently.

In this sequential formulation, the data y is decomposed in a finite sequence of observations $y_{obs} = (y_{obs}^{(1)}, \dots, y_{obs}^{(N)})$ (NOTE: maybe it's relevant to cite [Cho02] here) . We also decompose the forward response oeprator $\mathcal{G}$ in a sequence of operators $\{\mathcal{G}_i\}_{i=1}^N$ such that

$$y_{obs}^{(i)} \approx \mathcal{G}_i(u^*).$$

This allows us to reformulate the question from above as *How can we use new observations to update our knowledge about θ_{true}?* Answering this new question is called the *filtering problem*, and solving inserve and filtering problems will be the focus of this thesis.

One situation where this decomposition can naturally be used is when the system is described by a initial value problem of the form

$$\begin{aligned} \frac{\mathrm{d}x}{\mathrm{d}t} &= f(x; u^*) \\ x(0) &= x_0, \end{aligned} \tag{2.2}$$

where $u^* \in X$ is the parameter of the model, and the solution $x(t; u^*)$ is assumed to exist for every time $t \geq 0$. The sequence of observations then correspond to measurements at times $0 \leq t_1 < \ldots < t_N$. We further use an *observational operator* $\mathcal{O}$ to model the measurement procedure, mapping states of the system to observations. The sequence of forward response operator is $\mathcal{G}_i : X \to Y$ are then given by $\mathcal{G}_i = \mathcal{O} \circ x(t_i; \cdot)$. We illustrate this in the next section.

2.3 Pendulum problem

We now introduce a filtering problem that will guide the rest of the thesis. Using a pendulum, we would like to estimate the value of the Earth's gravitational acceleration. To do this, we first had to model the behaviour of the pendulum using a model parametrized by the gravitational acceleration g. We chose to use the *simple pendulum model*, a simplified model that ignores the mass of the hanging mass and of the string, ignores the forces of friction present on the hanging mass and assumes the movement on the pendulum is only happening on one plane. This model is illustrated in Figure 1. This simplification allows to model the state of the pendulum with a single value $x(t)$ representing the angle of the pendulum to the resting point, described by the following differential equation

$$\frac{\mathrm{d}^2 x}{\mathrm{d}t^2} = -\frac{g}{l} \sin(x). \tag{2.3}$$

In this model, g is the Earth's gravitational acceleration and l is the length of the string holding the hanging mass.

We proceeded to run an experiment, in which the pendulum was let go from an initial angle of 5° and no initial velocity. We then measured the first $N = 11$ times at which the pendulum was aligned with the vertical axis, indicating a null angle.

2.4 Bayesian filtering

The previous paragraphs focused on giving a definition of inverse and filtering problems. Before starting to discuss solutions to these problems, we present the concept of *well-posedness* first given by Hadamard for describing properties of models of physical phenomena.

Definition 2.1. A problem is said to be *well-posed* if it satisfies the following conditions:

1. a solution exists,
2. the solution is unique,
3. the solution changes continuously with the initial condition.

A problem failing to satisfy these conditions is said to be *ill-posed*. In the context of inverse problems, the 3rd property should be understand as continuity of the solution of the problem with respect to the data.

A possible way to solve inverse problems is to try to find a value $u^* \in X$ that solves the inverse problem *as well as possible*. This is done by replacing the inverse problem by the optimization problem

$$u^* = \underset{u \in X}{\operatorname{argmin}} \left\| y_{obs} - \mathcal{G}(u) \right\|_Y .$$

However, finding a global minimum in the presence of noise is often a difficult task since it might not exist, or the minimized function might admit multiple local minima. Solving the inverse problem by minimization is thus an ill-posed problem. While some of these difficulties can be addresses by *regularization*, two issues remain unresolved. First, regularization and the choice of the minimized norm are *ad hoc* decisions that are not part of the modeling process. Then, assuming that the optimization algorithm does provide an estimate $\hat{u}$, this point estimate does not contain information about the quality, or *uncertainty*, of this estimation. We chose to study a different approach for solving inverse problems: the Bayesian framework.

In the Bayesian framework, the model is not anymore treated as an equation that has to be inverted, but rather as an encoding the relation between the knowns and the unknowns of the system. In this encoding, we represent every variable using *random variables* and refine the model to include every available information of the system. This allows us to rewrite the standard inverse problem as follows

$$y = \mathcal{G}(u) + \eta. \tag{2.4}$$

Here, the variable y models the measurements of the system, and the observations y_{obs} are treated as realizations of this random variable. Existing knowledge about the parameter *prior* collecting the data is incorporated in the distribution of u, called the *prior distribution* with measure μ_0. The remaining variable η is used to represents the uncertainty in the observations y. Commonly, η is used to incorporate measurement noise or modelling error, and is often model using a mean-zero Gaussian distribution.

The coupling of these random variables allows to answer a wide range of questions about the model through the use of conditioning. For instance, in a situation where the real parameter $u = u^*$ is known, we can consider the conditioned random variable $y|u$. The probability density of this conditioned random variable is refered to as the *likelihood function*. If ρ denotes the density function of η, the likelihood function is

$$\ell(y|u) := \rho(y - \mathcal{G}(u)). \tag{2.5}$$

Motivated by the wide use of the Gaussian distribution for error modeling, we assume throughout the thesis that the likelihood function can be written as

$$\ell(y|u) = \exp(-\Phi(u; y)), \tag{2.6}$$

where Φ is called a *potential function*, or *negative log-likelihood*. In the case where the uncertainty is modeled by a multivariate Gaussian distribution $\eta \sim \mathcal{N}(0, \Gamma)$, the potential function is given by

$$\Phi(u; y) = \left\| \Gamma^{-1/2}(y - \mathcal{G}(u)) \right\|_Y^2. \tag{2.7}$$

(NOTE: reference to the link between regularization and error modeling?)

When solving the inverse problem, we are interested in learning how observed data update our prior believes about the model's parameters. This question can be naturally translated into studying the conditional distribution of u under the observations $y = y_{obs}$. The new knowledge about the parameter u is then contained in the distribution of $u|y$, called the *posterior distribution* with measure μ^y. Intuitively, *Bayes' rule* can be used to find the posterior distribution in terms of the prior distribution and the likelihood.

Given a probability triplet $(\Omega, \mathcal{F}, \mathbb{P})$, and two events $A, B \in \mathcal{F}$ with $\mathbb{P}(B) > 0$, Bayes' rule gives the distribution of the conditioned event $\mathbb{P}(A|B)$ by

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(B|A)\mathbb{P}(A)}{P(B)}.$$

An informal extension of this theorem to infinitesimal events gives the following relation for the posterior distribution

$$\mu^y(\mathrm{d}u) \propto \ell(y|u)\mu_0(\mathrm{d}u),$$

where the $\propto$ symbol denotes proportionality up to a constant factor. While this argument does provide the correct result, the extension of Bayes' rule from events to probability measures is not valid. Moreover, it is not clear which assumptions are made about the model to ensure existence the solution, and more generally, well-posedness has not been discussed. A more rigorous proof and characterization of well-posed inverse problems will now be presented.

We start by stating the following assumptions

Asssumption 2.1. The function $\Phi : X \times Y \to \mathbb{R}$ has the following properties.

1. For every $\epsilon > 0$ and $r > 0$ there is an $M \in \mathbb{R}$ such that, for all $u \in X$ and all $y \in Y$ with $\|y\|_Y < r$,
$$\Phi(u; y) \geq M - \epsilon \|u\|_X^2 .$$

2. For every $r > 0$ there is a $K > 0$ such that, for all $u \in X$ and $y \in Y$ with $\max\{\|u\|_X , \|y\|_Y\} < r$,
$$\Phi(u; y) \leq K.$$

3. For every $r > 0$ there is a $L > 0$ such that, for all $u_1, u_2 \in X$ and $y \in Y$ with $\max\{\|u_1\|_X , \|u_2\|_X , \|y\|_Y\} < r$,
$$|\Phi(u_1; y) - \Phi(u_2; y)| \leq L \|u_1 - u_2\|_X .$$

4. For every $\epsilon > 0$ and $r > 0$ there is a $C \in \mathbb{R}$ such that, for all $y_1, y_2 \in Y$ with $\max\{\|y_1\|_Y , \|y_2\|_Y\} < r$, and for all $u \in X$,
$$|\Phi(u; y_1) - \Phi(u; y_2)| \leq \exp(\epsilon \|u\|_X^2 + C) \|y_1 - y_2\|_Y .$$

These mild assumptions happen to be rather easy to fulfill for many practical problems. They provide with upper and lower bounds, and Lipshitz continuity with respect to the data y and the parameter u. However, since potential functions often are of the form of (2.7), we are interested in refining these assumptions to the shared structure of such problems.

Asssumption 2.2. The function $\mathcal{G} : X \to \mathbb{R}^N$ has the following properties.

1. For every $\epsilon > 0$ there is an $M \in \mathbb{R}$ such that for all $u \in X$,

$$|\mathcal{G}(u)|_\Gamma \leq \exp(\epsilon \|u\|_X^2 + M).$$

2. For every $r > 0$ there is a $K > 0$ such that, for all $u_1, u_2 \in X$ with $\max\{\|u_1\|_X, \|u_2\|_X\} < r$,

$$|\mathcal{G}(u_1) - \mathcal{G}(u_2)|_\Gamma \leq K \|u_1 - u_2\|_X .$$

We can naturally use these assumptions on the forward response operator $\mathcal{G}$ to derive properties of the potential function using the following lemma.

Lemma 2.1. Assume that $\mathcal{G} : X \to \mathbb{R}^N$ satisfies Assumption 2.2. Then, for any covariance matrix Γ, the potential function given by (2.7) satisfies Assumption 2.1 with $(Y, \|\cdot\|_Y) = (\mathbb{R}^N, \|\cdot\|_\Gamma)$.

Proof. (NOTE: This is trivial, can I skip this proof?) □

Using these assumptions, we can now proceed to provide a formal proof of Bayes' rule for continuous random variables and (2.4). The following theorem will play a central role in proving

Theorem 2.1. Let μ, ν be probability measures on $S \times T$, where $(S, \mathcal{A})$ and $(T, \mathcal{B})$ are measurable spaces. Let $(x, y) \in S \times T$. Assume that $\mu \ll \nu$ and that μ has Radon-Nikodym derivative ϕ with respect to μ. Assume further that the conditional distributions of $x|y$ under ν, denoted by $\nu^y(\mathrm{d}x)$, exists. Then the conditional distribution of $x|y$ under μ, denoted $\mu^y(\mathrm{d}x)$, exists and $\mu^y(\mathrm{d}x) \ll \nu^y(\mathrm{d}x)$. The Radon-Nikodym derivative is given by

$$\frac{\mathrm{d}\nu^y}{\mathrm{d}\mu^y}(x) = \begin{cases} \frac{1}{c(y)}\phi(x, y) & \text{if } c(y) > 0, \text{ and} \\ 1 & \text{else,} \end{cases} \tag{2.8}$$

where $c(y) = \int_S \phi(x, y)\mathrm{d}\mu^y(x)$ for all $y \in T$.

Proof. The proof of this theorem is given by Dudley [Dud02]. □

We now state Bayes' rule for continuous distribution and prove it using the previously stated theorem.

Theorem 2.2 (Generalized Bayes' Rule)**.** Assume that the likelihood function is given as in (2.6) where Φ satisfies Assumptions 2.1 and that $\mu_0(X) = 1$. Then $u|y$ is distributed according to the measure μ^y, with $\mu^y \ll \mu_0$ and Radon-Nikodym derivative with respect to μ_0 given by

$$\frac{\mathrm{d}\mu^y}{\mathrm{d}\mu_0}(u) = \frac{1}{Z_y}\exp(-\Phi(u;y)), \tag{2.9}$$

where Z_y is a constant that only dependends on y and not on u, called the *model evidence*.

Proof. Let $\mathbb{Q}_0(\mathrm{d}y) = \rho(y)\mathrm{d}y$ and $\mathbb{Q}(\mathrm{d}y|u) = \rho(y - \mathcal{G}(u))\mathrm{d}y$. Since both measures have a Radon-Nikodym derivative with respect to the Lebesgue measure λ, we have

$$\frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\mathbb{Q}_0}(y|u) = \frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\lambda}(y|u)\left(\frac{\mathrm{d}\mathbb{Q}_0}{\mathrm{d}\lambda}(y)\right)^{-1} = \frac{\rho(y-\mathcal{G}(u)}{\rho(y)} = C(y)\rho(y-\mathcal{G}(u)),$$

where $C(y) := 1/\rho(y)$ is well defined since Assumption 2.1(2) gives an upper bound on Φ thus also giving a strictly positive lower bound on ρ. We further define two measures ν_0, ν on $Y \times X$ by

$$\nu_0(\mathrm{d}y, \mathrm{d}u) = \mathbb{Q}_0(\mathrm{d}y) \otimes \mu_0(\mathrm{d}u)$$
$$\nu(\mathrm{d}y, \mathrm{d}u) = \mathbb{Q}(\mathrm{d}y|u)\mu_0(\mathrm{d}u).$$

Since $\mathcal{G}$ is continuous and $\mu_0(X) = 1$, it is also μ_0-measurable. Thus, ν is well-defined and continuous with respect to ν_0 with Radon-Nikodym derivative

$$\frac{\mathrm{d}\nu}{\mathrm{d}\nu_0}(y,u) = C(y)\rho(y-\mathcal{G}(u)).$$

Since ν_0 is a product measure over $Y \times X$, the random variables y and u are independent, giving $u|y = u$. This implies that the conditional distribution of $u|y$ under ν_0 is then $\nu_0^y = \mu_0$. In addition, again using Assumption 2.1(2), we have a strictly positive lower bound on ρ which in turn shows that

$$c(y) := \int_X C(y)\rho(y-\mathcal{G}(u))\mu_0(\mathrm{d}u) > 0$$

Thus, by Theorem 2.1, the conditional distribution of $u|y$ under ν, denoted μ^y, exists and its Radon-Nikodym derivative with respect to $\nu_0^y = \mu_0$ is

$$\frac{\mathrm{d}\mu^y}{\mathrm{d}\mu_0}(u) = \frac{1}{c(y)}C(y)\rho(y-\mathcal{G}(u)) = \frac{1}{Z_y}\rho(y-\mathcal{G}(u)) = \frac{1}{Z_y}\exp(-\Phi(u;y)).$$

where $Z_y = \int_X \exp(-\Phi(u;y))\mu_0(\mathrm{d}u)$. □

The previous theorems provide a well defined solution to the bayesian inverse problem. In order to consider the problem well-posed, we still need to prove that the solution is

continuous with respect to the data y. In order to prove this continuity of the solution, we need to define a metric over the space of probability measures.

Definition 2.2. The *Hellinger distance* between μ and μ' is

$$d_{Hell}(\mu, \mu') = \sqrt{\frac{1}{2} \int \left(\frac{\mathrm{d}\mu}{\mathrm{d}\nu} - \frac{\mathrm{d}\mu'}{\mathrm{d}\nu} \right)^2 \mathrm{d}\nu},$$

where ν is an arbitrary measure to which both μ and μ' are absolutely continuous.

Moreover, we will require the prior measure to have exponentially bounded tails. This is formalized in the following definition.

Definition 2.3. A probability measure μ on a Banach space X is called *light-tailed* if there exists an $\alpha > 0$ such that

$$\int_X \exp(\alpha \|x\|_X^2) \mu(\mathrm{d}x) < \infty.$$

The restriction on light-tailed measures still allows to use many common distributions, such as Gaussians and distributions over compact sets, and is thus not too restrictive in practical situations. We can now complete the proof of well-posedness by the following theorem.

Theorem 2.3. Let Φ satisfy Assumptions 2.1 and μ_0 be light-tailed with support equals to X. Then μ^y given in 2.4 is Lipschitz continuous with respect to the data y in the Hellinger distance.

Proof. Since many different temporary constants are being used throughout the proof, we use C as a placeholder for a non-negative constant, and may change its value from term to term.

Let $y, y' \in Y$ and $\mu^y, \mu^{y'}$ be the measures obtained by application of Bayes' rule and Z, Z' be the respecting model evidences.

$$
\begin{aligned}
|Z - Z'| &= \left| \int_X \exp(-\Phi(u;y)) - \exp(-\Phi(u;y')) \mathrm{d}\mu_0(u) \right| \\
&= \left| \int_X \exp(-\Phi(u;y')) \left(\exp(\Phi(u;y') - \Phi(u;y)) - 1\right) \mathrm{d}\mu_0(u) \right| \\
&\leq \int_X \exp(-\Phi(u;y')) \left|\exp(\Phi(u;y') - \Phi(u;y)) - 1\right| \mathrm{d}\mu_0(u) \\
&\leq \int_X \exp(-\Phi(u;y')) \exp(\Phi(u;y') - \Phi(u;y)) \mathrm{d}\mu_0(u) \\
&\leq \int_X \exp(-\Phi(u;y')) \left|\Phi(u;y') - \Phi(u;y)\right| \mathrm{d}\mu_0(u)
\end{aligned}
$$

Since μ_0 is light-tailed, there is an $\epsilon > 0$ such that $\int_X \exp(\epsilon \|u\|_X^2) \mathrm{d}\mu_0(u) = C < \infty$. Using this $\epsilon/2$ for Assumption 2.1(1) and (4), we get

$$
\begin{aligned}
|Z - Z'| &\leq \int_X \exp(\frac{1}{2}\epsilon \|u\|_X^2 - M) \exp(\frac{1}{2}\epsilon \|u\|_X^2 + C) \|y - y'\|_Y \mathrm{d}\mu_0(u) \\
&= C \|y - y'\|_Y \int_X \exp(\epsilon \|u\|_X^2) \mathrm{d}\mu_0(u) \\
&\leq C \|y - y'\|_Y .
\end{aligned}
$$

From the definition of the Hellinger distance, we have by triangle inequality

$$
\begin{aligned}
2d_{Hell}(\mu^y, \mu^{y'}) &= \int_X \left[Z^{-1/2} \exp\left(-\frac{1}{2}\Phi(u;y)\right) - (Z')^{-1/2} \exp\left(-\frac{1}{2}\Phi(u;y')\right) \right]^2 \mathrm{d}\mu_0(u) \\
&\leq \frac{2}{Z} \int_X \left[\exp\left(-\frac{1}{2}\Phi(u;y)\right) - \exp\left(-\frac{1}{2}\Phi(u;y')\right) \right]^2 \mathrm{d}\mu_0(u) \\
&+ 2 \left| Z^{-1/2} - (Z')^{-1/2} \right|^2 \int_X \exp\left(-\Phi(u;y')\right) \mathrm{d}\mu_0(u) \\
&= I_1 + I_2.
\end{aligned}
$$

By using Assumptions 2.1(1) and (4), and the same reasoning as above, we get

$$
\begin{aligned}
\frac{Z}{2} I_1 &\leq \int_X \frac{1}{4} \exp(\epsilon \|u\|_X^2 - M) \exp(2\epsilon \|u\|_X^2 + 2C) \|y - y'\|_Y^2 \mathrm{d}\mu_0(u) \\
&\leq C \|y - y'\|_Y^2 .
\end{aligned}
$$

And we can further show that Z greater than zero using Assumption 2.1(2), since for every $r > 0$

$$
Z \geq \int_{\|u\|_X \leq r} \exp(-C) \mathrm{d}\mu_0(u) = \exp(-C) \mu_0(\|u\|_X \leq r) > 0.
$$

Where the last inequality holds because μ_0 has support equal to X. This together with the previously given upper bound on $Z/2I_1$ gives us $I_1 \leq C \|y - y'\|$.

Additionally, we have

$$\begin{aligned}\left|Z^{-1/2} - (Z')^{-1/2}\right|^2 &= \left|\frac{\sqrt{Z} - \sqrt{Z'}}{\sqrt{ZZ'}}\right|^2 \leq \left|\frac{\sqrt{Z} - \sqrt{Z'}}{Z \wedge Z'}\right|^2 \\ &= (Z \wedge Z')^{-3} \left|\sqrt{Z \wedge Z'}(\sqrt{Z} - \sqrt{Z'})\right|^2 \\ &\leq (Z \wedge Z')^{-3} |Z - Z'|^2 \leq C \|y - y'\|_Y^2 .\end{aligned}$$

Where the last inequality holds because both Z and Z' are strictly greater than 0 and the bound proved earlier. Moreover using Assumption 2.1(2) and that μ_0 is light-tailed, we can easily show that $\int_X \exp(-\Phi(u; y'))\mathrm{d}\mu_0(u)$ is bounded by a constant. This shows that $I_2 \leq C \|y - y'\|_Y^2$, hence completing the proof. □

We have shown that the Bayesian inverse problem is well-defined under light conditions on Φ, given in Assumption 2.1, and provided that the prior measure μ_0 is light-tailed. In the next section, we will model the pendulum problem in the probabilistic framework presented earlier, and will show that it is a well-posed problem.

2.5 A new take at the Pendulum problem

3 Proofs

3.1 well-posedness

Lemma 3.1. The pendulum problem satisfies Assumption 2.1.

Proof. We first prove that the solution $x(t)$ of the initial value problem is bounded, which is a stronger property than property i). We start by defining the Hamiltonian

$$H(x, x') = \frac{1}{2}x'^2 - \frac{g}{l}\cos(x).$$

By simple calculation, one can show that H is constant along the solution of the initial value problem. Since the initial values are $x_0 \in (0, \pi/2)$ and $x'_0 = 0$, we have $H(x(t), x'(t)) = -\frac{g}{l}\cos(x_0) \in (-\frac{g}{l}, 0)$ for all $t > 0$. Assuming that $x(t)$ is not bounded, since x is continuous and $x_0 \leq \pi/2$, there is a time $t^\star$ such that $x(t^\star) = \pi$, giving

$$H(x(t^\star), x'(t^\star)) = \frac{1}{2}x'(t^\star)^2 - \frac{g}{l}cos(x(t^\star)) \geq -\frac{g}{l}\cos(\pi) = \frac{g}{l} > 0.$$

This contradicts $H(x(t), x'(t)) = H(x_0, x'_0) \in (\frac{g}{l}, 0)$.Thus the solution of the initial value problem is bounded and so is $\mathcal{G}$, proving i). Furthermore, we know that the solution of the initial value problem is continuously differentiable with respect to g, it is thus locally Lipschitz continuous everywhere, and so is $\mathcal{G}$, thus completing the proof. □

3.2 smc convergence

Lemma 3.2. Let $P(E)$ denote the set of all probability measures on E. For every μ and ν random variables with values in $P(E)$, we define

$$d(\mu, \nu) := \sup_f \sqrt{\mathbb{E}\left[|\mu f - \nu f|^2\right]},$$

where the supremum is taken over all $f : E \to \mathbb{R}$ with $|f|_\infty \leq 1$, and μf denotes the integral of f under μ. Then d is a metric over the space of random measures over E.

Proof. Trivial enough to skip the proof? □

Definition 3.1. Let $M \in \mathbb{N}$, for every $\mu \in P(E)$ we define $S^M\mu$ by

$$S^M\mu = \frac{1}{M}\sum_{i=1}^{M}\delta_{u_i},$$

where $u^{(1)}, \dots, u^{(M)}$ are i.i.d. random variables distributed according to μ. From the randomness of the samples $u^{(1)}, \dots, u^{(M)}$, it follows that $S^M\mu$ is a random variable with values in $P(E)$. The operator $\mu \mapsto S^M\mu$ is called *sampling operator*.

Lemma 3.3. The sampling operator satisfies

$$\sup_{\mu \in P} d(S^M\mu, \mu) \leq \frac{1}{\sqrt{M}}$$

Proof. Let μ be an element of $P(E)$ and $u^{(1)}, \dots, u^{(M)}$ be i.i.d. random variables distributed according to μ. For every f with $|f|_\infty \leq 1$ we have

$$(S^M\mu)f - \mu f = \frac{1}{M}\sum_{i=1}^{M} f(u^{(i)}) - \mu f = \frac{1}{M}\sum_{i=1}^{M} f_i,$$

where $f_i = f(u^{(i)}) - \mu f$. This gives

$$\left|(S^M\mu - \mu)f\right|^2 = \left(\frac{1}{M}\sum_{i=1}^{M} f_i\right)^2 = \frac{1}{M^2}\sum_{i,j=1}^{M} f_i f_j.$$

Since we are interested in the expected value of this term, we now consider $\mathbb{E}[f_i f_j]$. For $i \neq j$, f_i and f_j are independent random variables, thus $\mathbb{E}[f_i f_j] = \mathbb{E}[f_i]\,\mathbb{E}[f_j]$, and since $u^{(i)} \sim \mu$, we have $\mathbb{E}[f_i] = \mathbb{E}\left[f(u^{(i)})\right] - \mu f = 0$, giving $\mathbb{E}[f_i f_j] = 0$ for $i \neq j$. Furthermore, since $|f|_\infty \leq 1$ we have

$$\mathbb{E}\left[f_i^2\right] = \mathrm{Var}[f(u^{(i)})] = \mathbb{E}\left[f(u^{(i)})^2\right] - \mathbb{E}\left[f(u^{(i)})\right]^2 \leq 1.$$

By linearity of the expected value, we then have for every f with $|f|_\infty \leq 1$

$$\mathbb{E}\left[\left|(S^M\mu)f - \mu f\right|^2\right] = \frac{1}{M^2}\sum_{i=1}^{M} \mathbb{E}\left[f_i^2\right] \leq \frac{1}{M}.$$

Taking the square root on both sides of the equation and the supremum over all such f yields the desired result. □

Asssumption 3.1. There exists a $\kappa > 0$ such that for every $n \in \mathbb{N}$ and $\theta \in \Theta$

$$\kappa \leq l_n(\theta) \leq 1/\kappa.$$

This assumption is weaker than (NOTE: assumption 2.6 in Stuart 2010) if Θ is compact.

(NOTE: Φ_n is defined in Beskos 2015, as $\Phi_n = K_n L_{n-1}$ where K_n is a transition kernel and L_{n-1} is the Bayes' rule update, meaning $(L_{n-1}\mu)(du) = \frac{l_{n-1}(u)}{\mu(l_{n-1})}\mu(du)$, where l_i are the likelihoods of the sequence. These values will be properly defined in the thesis and another notation will probably be used.)

Lemma 3.4. Assume that the likelihoods l_i satisfy Assumption 3.1. Then for any $n \in \mathbb{N}$ and $\mu, \nu \in P(E)$,

$$d(\Phi_n\mu, \Phi_n\nu) \leq \frac{2}{\kappa^2} d(\mu, \nu).$$

Proof. Let $f : E \to \mathbb{R}$ be a measurable function with $|f|_\infty \leq 1$, we then have

$$\begin{aligned}
(\Phi_n\mu - \Phi_n\nu)f &= (K_n L_{n-1}\mu - K_n L_{n-1}\nu)f \\
&= \frac{\mu(l_{n-1}K_n(f))}{\mu(l_{n-1})} - \frac{\nu(l_{n-1}K_n(f))}{\nu(l_{n-1})} \\
&= \frac{\mu - \nu}{\mu(l_{n-1})}(l_{n-1}K_n(f)) + \frac{\nu(l_{n-1}K_n(f))}{\mu(l_{n-1})} - \frac{\nu(l_{n-1}K_n(f))}{\nu(l_{n-1})} \\
&= \frac{\mu - \nu}{\mu(l_{n-1})}(l_{n-1}K_n(f)) + \nu(l_{n-1}K_n(f))\left(\frac{1}{\mu(l_{n-1})} - \frac{1}{\nu(l_{n-1})}\right) \\
&= \frac{\mu - \nu}{\mu(l_{n-1})}(l_{n-1}K_n(f)) + \frac{\nu(l_{n-1}K_n(f))}{\mu(l_{n-1})\nu(l_{n-1})}(\nu - \mu)(l_{n-1}).
\end{aligned}$$

By Minkowski's inequality, we then have

$$\begin{aligned}
\sqrt{\mathbb{E}\left[|(\Phi_n\mu - \Phi_n\nu)f|^2\right]} \leq \mathbb{E}&\left[\left|\frac{\mu - \nu}{\mu(l_{n-1})}(l_{n-1}K_n(f))\right|^2\right]^{\frac{1}{2}} \\
&+ \mathbb{E}\left[\left|\frac{\nu(l_{n-1}K_n(f))}{\mu(l_{n-1})\nu(l_{n-1})}(\nu - \mu)(l_{n-1})\right|^2\right]^{\frac{1}{2}}.
\end{aligned}$$

In the second term, we observe that $\frac{\nu(l_{n-1}K_n(f))}{\nu l_{n-1}}$ is the expectation of f under the probability measure $\Phi_n\nu$. Together with $|f|_\infty \leq 1$, this gives us $\frac{\nu(l_{n-1}K_n(f))}{\nu l_{n-1}} \leq 1$. Similarly, since $l_{n-1}(u) \geq 1/\kappa$ for all u, it holds that $\mu(l_{n-1}) \geq \kappa$ and $1/\mu(l_{n-1}) \leq 1/\kappa$. From these two bounds, we obtain

$$\begin{aligned}
\sqrt{\mathbb{E}\left[|(\Phi_n\mu - \Phi_n\nu)f|^2\right]} \leq \frac{1}{\kappa}&\mathbb{E}\left[|(\mu - \nu)(l_{n-1}K_n(f))|^2\right]^{\frac{1}{2}} \\
&+ \frac{1}{\kappa}\mathbb{E}\left[|(\mu - \nu)(l_{n-1})|^2\right]^{\frac{1}{2}}.
\end{aligned}$$

Finally, using that $0 \leq l_{n-1} \leq 1/\kappa$, we have $|\kappa l_{n-1}|_\infty \leq 1$ and similarly $[l_{n-1}K_n(f)](u) = \int_\Theta f(y) l_{n-1}(u) K_n(u, \mathrm{d}y) \leq |f l_{n-1}|_\infty K_n(\Theta, u) \leq 1/\kappa$, thus making $|\kappa l_{n-1} K_n(f)|_\infty \leq 1$. We can thus bound our expression by the supremum over all g with $|g|_\infty \leq 1$ and get

$$\sqrt{\mathbb{E}\left[|(\Phi_n\mu - \Phi_n\nu)f|^2\right]} \leq \frac{2}{\kappa^2} \sup_g \sqrt{\mathbb{E}\left[|\mu g - \nu g|^2\right]} = \frac{2}{\kappa^2} d(\mu, \nu).$$

Since this result is valid for all f with $|f|_\infty \leq 1$, taking the supremum on the left side yields the desired result. □

(NOTE: $\nu_n^M = S^M \Phi_n \nu_{n-1}^M$ and for $n = 0$ we dirac approximate with an initial sample of ν_0, this will be better defined in the presentation of the algorithm.)

Theorem 3.1. Assume that the likelihoods l_i satisfy Assumption 3.1 and consider the SMC algorithm with $M_{thresh} = M$. Then, for any $n \in \mathbb{N}$,

$$d(\nu_n, \nu_n^M) \leq \sum_{j=0}^{n} \left(\frac{2}{\kappa^2}\right)^j \frac{1}{\sqrt{M}}.$$

Proof. For $n = 0$, the result hold by direct application of Lemma 3.3. For $n > 0$ we have by the triangle inequality and Lemmata 3.3 and 3.4

$$\begin{aligned} d(\nu_n, \nu_n^M) &= d(\Phi_n \nu_{n-1}, S^M \Phi_n \nu_{n-1}^M) \\ &\leq d(\Phi_n \nu_{n-1}, \Phi_n \nu_{n-1}^M) + d(\Phi_n \nu_{n-1}^M, S^M \Phi_n \nu_{n-1}^M) \\ &\leq \frac{2}{\kappa^2} d(\nu_{n-1}, \nu_{n-1}^M) + \frac{1}{\sqrt{M}}. \end{aligned}$$

Iterating on n gives the desired result. □

4 Case Study

5 Conclusion

A Appendix

List of Figures

List of Tables

References

[ABL+13] Moritz Allmaras, Wolfgang Bangerth, Jean Marie Linhart, Javier Polanco, Fang Wang, Kainan Wang, Jennifer Webster, and Sarah Zedler. Estimating parameters in physical models through bayesian inversion: A complete example. *SIAM Review*, 55(1):149–167, 2013.

[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.

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

[DM04] Pierre Del Moral. *Feynman-kac formulae*. Springer, 2004.

[DM13] Pierre Del Moral. *Mean field simulation for Monte Carlo integration*. Chapman and Hall/CRC, 2013.

[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.

[Dud02] R. M. Dudley. *Conditional Expectations and Martingales*, chapter 10, pages 336–384. Cambridge Studies in Advanced Mathematics. Cambridge University Press, 2 edition, 2002.

[GCS+14] Andrew Gelman, John B Carlin, Hal S Stern, David B Dunson, Aki Vehtari, and Donald B Rubin. *Bayesian data analysis*, volume 2. CRC press Boca Raton, FL, 2014.

[Had02] Jacques Hadamard. Sur les problèmes aux dérivés partielles et leur signification physique. *Princeton University Bulletin*, 13:49–52, 1902.

[Has70] W Keith Hastings. Monte carlo sampling methods using markov chains and their applications. *Biometrika*, 57(1):97–109, 1970.

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

[Lap20] Pierre Simon Laplace. *Théorie analytique des probabilités*. Courcier, 1820.

[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.

[RC05] Christian P. Robert and George Casella. *Monte Carlo Statistical Methods (Springer Texts in Statistics)*, chapter 3, pages 79,122. Springer-Verlag, Berlin, Heidelberg, 2005.

[Stu10] A. M. Stuart. Inverse problems: A bayesian perspective. *Acta Numerica*, 19:451–559, 2010.