Technische Universität München

Department of Matheematics

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, April 28, 2018

Place, Date

Signature

Abstract

Titel auf Englisch wiederholen.

Es folgt die englische Version der Kurzfassung.

Contents

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) (NOTE: need ref) , 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* $\theta \in \Theta$ is unknown or uncertain. To model the behaviour of the system, we introduce *forward response operators* $\mathcal{G} : \Theta \to Y$ 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 $\theta_{true} \in \Theta$, and assume that *observations* $y \in Y$ 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 follwoing approximate model

$$y \approx \mathcal{G}(\theta_{true}), \tag{1}$$

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

In this thesis, we assume a specific structure on the forward response operator and on the data. The systems are assumed to be time-dependent and described as the solution of a deterministic initial value problem of the form

$$\begin{aligned} \frac{\mathrm{d}x}{\mathrm{d}t} &= f(x;\theta) \\ x(0) &= x_0, \end{aligned} \tag{2}$$

where $\theta \in \Theta$ is the parameter of the model, and the solution $x(t;\theta) \in X$ is assumed to exist for every time $t \geq 0$. We further assume the measurements of the system to be sequentially taken at times $0 \leq t_1 < \ldots < t_N$. We use an *observational operator* $\mathcal{O} : X \to Y$ to model the measurement procedure, mapping states of the system to observations. We can then define a sequence of forward response operator $\mathcal{G}_i : \Theta \to Y$ all given by $\mathcal{G}_i = \mathcal{O} \circ x(t_i;\cdot)$. Similarly to the inverse problem, we get an approximate model of the form

$$y_i \approx \mathcal{G}_i(\theta_{true}). \tag{3}$$

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.

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

$$\frac{\mathrm{d}^2 x}{\mathrm{d}t^2} = -\frac{g}{l}\sin(x). \tag{4}$$

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 $x(0) = 5\frac{\pi}{180}$ 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 discussing frameworks for expressing such problems, we define the useful concept of *well-posedness* first given by Hadamard for describing properties of models of physical phenomena.

Definition 2.1 (Well-posedness)**.** 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 $\hat{\theta} \in \Theta$ that solves the inverse problem *as well as possible*. This is done by replacing the inverse problem by the optimization problem

$$\hat{\theta} = \operatorname*{argmin}_{\theta \in \Theta} \|y - \mathcal{G}(\theta)\|_Y .$$

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 problems 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 but rather tuning parameters of the optimization problem. Then, assuming that the optimization algorithm does provide an estimate $\hat{\theta}$, this point estimate does not contain any information about the *uncertainty* around this estimation. We proceed to present the Bayesian approach for solving inverse problem, and how it can be extended to recursively solve filtering problems as well.

In the Bayesian framework, the inverse problem is treated as a statistical problem. The model is reformulated as a probabilistic model in which all parameters are expressed as *random variables*, and the model equation is made exact by explicitly incorporating noise assumption in the model, leading to a new model

$$y = \mathcal{G}(\theta) + \eta. \tag{5}$$

In this model, θ and η are independent random variable, where θ is distributed according to μ_0, called the *prior measure*, and η is a zero-mean random variable, commonly following a Gaussian distribution. The observation $y = (y_1, \ldots, y_n)$ are treated as realizations of the probabilistic model for $\theta = \theta_{true}$.

3 Proofs

Theorem 3.1. 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

$$\frac{\mathrm{d}\nu^y}{\mathrm{d}\mu^y}(x) = \begin{cases} \frac{1}{c(y)}\phi(x, y) & \text{if } c(y) > 0, \text{ and} \\ 1 & \text{else,} \end{cases} \tag{6}$$

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

Proof. See [Dud02]. □

Theorem 3.2 (Generalized Bayes' Rule)**.** Assume that $\mathcal{G} : \Theta \to Y$ is continuous, that η has a density ρ with support equal to Y and that $\mu_0(\Theta) = 1$. Then $\theta|y$ 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

$$\frac{\mathrm{d}\mu^y}{\mathrm{d}\mu_0}(\theta) = \frac{1}{Z_y}\rho(y - \mathcal{G}(\theta)), \tag{7}$$

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

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

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

Further define two measures ν_0, ν on $Y \times \Theta$ by

$$\begin{aligned} \nu_0(\mathrm{d}y, \mathrm{d}\theta) &= \mathbb{Q}_0(\mathrm{d}y) \otimes \mu_0(\mathrm{d}\theta) \\ \nu(\mathrm{d}y, \mathrm{d}\theta) &= \mathbb{Q}(\mathrm{d}y|\theta)\mu_0(\mathrm{d}\theta). \end{aligned}$$

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

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

Since ν_0 is a product measure over $Y \times \Theta$, the random variables y and θ are independent giving $\theta | y = \theta$. This implies that the conditional distribution of $\theta | y$ under ν_0 is then $\nu_0^y = \mu_0$. We further note that since $\rho > 0$ we have $C(y) > 0$ and $\rho(y - \mathcal{G}(\theta)) > 0$ giving

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

Thus, by Theorem 3.1, the conditional distribution of $\theta | y$ under ν, denoted μ^y, exists and its Radon-Nikodym derivative with respect to $\nu_0^y = \mu_0$ is

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

where $Z_y = \int_\Theta \rho(y - \mathcal{G}(\theta))\mu_0(\mathrm{d}\theta)$. □

Lemma 3.1. 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 not 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 $u^{(1)}, \ldots, u^{(M)}$ are i.i.d. random variables distributed according to μ. From the randomness of the samples $u^{(1)}, \ldots, u^{(M)}$, it follows that $S^M \mu$ is a random variable with values in $P(E)$. The operator $\mu \mapsto S^M \mu$ is called the *sampling operator*.

Lemma 3.2. 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 $u^{(1)}, \ldots, u^{(M)}$ 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

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

Since we are interested in the expected value of the initial 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 $\mathbb{E}[f_i] = \mathbb{E}[f(u^{(i)})] - \mu f = 0$, giving $\mathbb{E}[f_i f_j] = 0$ for $i \neq j$. Furthermore, since $|f|_\infty \leq 1$ we have

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

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

$$\mathbb{E}[((S^M\mu)f - \mu f)^2] = \frac{1}{M^2}\sum_{i=0}^{N} \mathbb{E}[f_i^2] \leq \frac{1}{M}.$$

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

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.

[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.

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