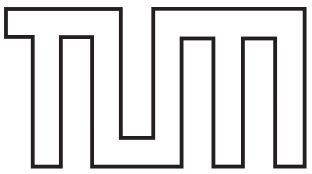

Technische Universität München
Department of Mathematics

Bachelor's Thesis

# Sequential Monte Carlo for time-dependent Bayesian Inverse Problems

Matthieu Bulté

| | |
|---|---|
| Supervisor: | Prof. Dr. Elisabeth Ullmann |
| Advisor: | M.Sc. Jonas Latz |
| 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, June 7, 2018
Place, Date

Signature

# Abstract

The goal of this paper is to present a sound solution for solving inverse problems. We focus on the scenario where data becomes available over time, called filtering problem. Formulated in the Bayesian framework, we prove that the inverse and filtering problems are well-posed under mild assumptions. We study the Sequential Monte Carlo algorithm to approximate the solution to Bayesian filtering problems and compare it to several related approximation algorithms. The use of the Sequential Monte Carlo algorithm is validated through proofs of convergence and numerical experiments. We also illustrate the theoretical and numerical concepts presented in the paper by a concrete filtering problem, in which we attempt to estimate the parameter of a dynamical system from measurements of the system.



# Zusammenfassung

In dieser Arbeit beschäftigen wir uns damit, eine korrekte und effiziente Lösung zu inversen Problemen zu finden. Wir konzentrieren uns auf Filtrierungsprobleme, wo Messpunkte im Laufe der Zeit zur Verfügung gestellt werden. Mithilfe des Bayesschen Formalismus beweisen wir dass, unter schwachen Vorausetzungen, inverse Probleme und Filtrierungsprobleme gut gestellt sind. Wir studieren den Sequential Monte Carlo Algorithmus zur Approzimierung der Lösung des Bayesschen Filtrierungsproblems und vergleichen es zu ähnlichen Approximierungsalgorithmen. Wir bestätigen die Wahl des Sequential Monte Carlo Algorithmus durch Konvergenzbeweise und numerische Experimente. Die theoretischen und numerischen Konzepte werden auch durch ein konkretes Filtrierungsproblem dargestellt, indem wir versuchen aus Messpunkten, Parameter eines dynamischen Systems zu schätzen.

# 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 *least squares* methods and the *Bayesian* methods. While both classes rely on a *likelihood function* that measures how likely it is to see some observations given a certain value of the parameter, the two methods exploit this information differently. In least squares methods, the solution of the inverse problem is given as the value of the parameter maximixing the likelihood of the observed data. In Bayesian methods, the system is re-modeled probabilistically with random variables. This solution is given as a conditionning of the model by the observed data using Bayes' rule, 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, Chapter 1] 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*. In Bayesian filtering, we are interested in sequentially updating our knowledge about the model parameters as more observations becomes available. We will demonstrate that under weak model assumptions, the solution to the filtering problem is *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 are required to compute these integrals with respect to a probability measure. We present approximations making use of weighted pseudo-random values to approximate the hard integral. 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, the numerical solution given by the MCMC algorithm cannot be extended as more data becomes available and requires to re-run the entire computation, making it unsuitable for filtering problems. A simpler method to operate is *importance sampling* (IS) [RC05, Chapter 3], 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. This method can be extended to be used to approximate sequences of distributions such as those found in filtering problems, giving the *Sequential Importance Sampling* (SIS) algorithm. However, choosing auxiliary distributions that are close to the target distributions is not always possible, and failing to do so results in a poor estimation of the distributions of interest.

*Sequential Monte Carlo* (SMC) [DMDJ06] is a method merging ideas of MCMC and IS 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 (BIP) 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 (BFP). 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 ordinary 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 define inverse and filtering problems, and describe the Bayesian framework as an alternative to classical minimization based solutions. We define well-posedness and prove that Bayesian inverse problems are well-posed under mild assumptions. We also describe the Pendulum problem and its Bayesian formulation and show that it is a well-posed problem. In Section 3, we show how the IS algorithm can be extended to the SIS and later SMC algorithms to efficitently approximate the solution of BFPs. We conclude with proving that these approximations provide uniform convergence bounds to the real solution of the BFPs. Section 4 then presents and analyses the numerical solution to the pendulum problem provided by the SMC algorithm and compares it to the solution given by the MCMC algorithm. Finally, Section 5 concludes the thesis with a discussion of other application areas and current research on SMC algorithms.

# 2 Bayesian Filtering

This section introduces and describes the Bayesian approach for solving inverse problems, laying out the theoretical foundations of the taken approach. In Section 2.1 we present the class of problems we will be considering in this thesis: finite-dimensional inverse and filtering problems. In Section 2.2 we present a filtering problem based on a simple dynamical system estimation problem which will be used to illustrate concepts presented throughout the thesis. In Section 2.3 we define desired characteristics of solutions to inverse problems. We mention the challenges encountered by the classical optimization approach for solving inverse problems in practical situations. We then present the Bayesian framework, and show how it uses prior information about the structure of the problem to construct a solution to the inverse and filtering problems. In Section 2.4 we prove well-posedness of this constructed solution under some light assumptions about the model. Finally, Section 2.5 illustrates Bayesian modelling by describing the process of constructing a probabilistic model for the pendulum problem, and show that the problem is well-posed and thus admits a stable solution.

## 2.1 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 a *forward response operator* $\mathcal{G} : X \to Y$ 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 $\mathcal{G}$. We obtain the following model

$$y^{obs} = \mathcal{G}(u^*), \tag{2.1}$$

where the equality is not exact due to uncertainties. Our initial question then becomes: *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 incrementally building up knowledge about the unknown parameter $u^*$. 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. Firstly, it allows to perform inference on running dynamical systems and update our knowledge as more measurements become available. Secondly, it can also be used in situations where all the data is available to transform one large 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 $y^{obs} = (y_1^{obs}, \ldots, y_T^{obs})^\top$. We also decompose the forward response operator $\mathcal{G}$ in a sequence of operators $\mathcal{G}_1, \ldots, \mathcal{G}_T$ such that

$$y_t^{obs} = \mathcal{G}_t(u^*).$$

This allows us to reformulate the question from above as *How can we use new observations to update our knowledge about $u^*$?* Answering this new question is called the *filtering problem*. This thesis will focus on solving inverse and filtering problems.

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

$$\begin{aligned} x'(t) &= f(x(t); u) \\ x(0) &= x_0, \end{aligned} \tag{2.2}$$

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 $0 \leq t_1 < \ldots < t_T$. We further define two additional kinds of operators: *solution* and *observational* operators. The solution operator $G$ maps elements of the parameter space to the solution of the associated initial value problem by $G(u) = x(\cdot, u)$. Observational operators $\mathcal{O}_i$ then map this solution to the respective observation, here we can define for each time $t_i$ the observational operator $\mathcal{O}_i(x(\cdot, u)) = x(t_i, u)$. The forward response operators can then be written as the composition of these two operators $\mathcal{G}_i = \mathcal{O}_i \circ G$. We provide an example of the modelling of a inverse problem in the next section.

## 2.2 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. We start by modelling the behaviour of the pendulum using a differential equation parametrized by the gravitational acceleration $g$.

We choose the *simple pendulum model*, a simplified mathematical model that ignores the mass of the string of the pendulum and ignores the forces of friction acting on the hanging mass. It also assumes that the movement of the pendulum only happens in two dimensions. This simplification allows us 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 second ordinary order differential equation.

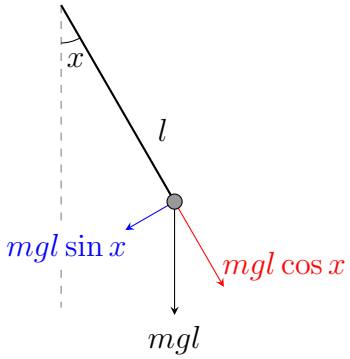


Figure 1: Pendulum model and forces applied to the mass, the vertical vector represents the gravitational force and the red and blue vectors represent the same force decomposed into its components, parallel and perpendicular to the motion of the pendulum. The dashed line represents the 0 angle at which measurements are taken.

$$\begin{aligned} x''(t) &= -\frac{g}{l}\sin(x(t)) \\ x'(0) &= x'_0 \\ x(0) &= x_0. \end{aligned} \tag{2.3}$$

In this model, $g$ is the Earth's gravitational acceleration and $l$ is the length of the string holding the hanging mass. A representation of this model is given in Figure 1.

We then run an experiment in which the pendulum is let go from an initial angle of $x_0 = 5\pi/180$ (5°) and no initial velocity $x'_0 = 0$. Using a stopwatch, we collect $T = 11$ time measurements at which the pendulum was aligned with the vertical axis, corresponding to a null angle of the pendulum.

## 2.3 Bayesian filtering

The previous sections defined the inverse and filtering problems. Before starting to discuss solutions to these problems, we present the concept of *well-posedness* first given by Hadamard in [Had02] 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 respect to the data.

A problem failing to satisfy these conditions is said to be *ill-posed*. In the context of inverse problems, the third property states that small perturbations in the data should result in proportionally small perturbations in the solution $u$ of the inverse problem.

A possible way to solve inverse problems is to try to find a value $\hat{u} \in X$ that solves the inverse problem *as well as possible*. This is done by replacing the inverse problem by the following non-linear least squares problem

$$\hat{u} = \underset{u \in X}{\operatorname{argmin}} \frac{1}{2} \left\| y^{obs} - \mathcal{G}(u) \right\|_Y^2 .$$

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 addressed by *regularization*, two issues remain unresolved. Firstly, regularization and the choice of the minimized norm are *ad hoc* decisions that are not part of the modelling process. Secondly, assuming that the optimization algorithm does provide an estimate $\hat{u}$, this point estimate does not contain information about the quality, nor *uncertainty*, of this estimation. For these reasons, we chose to study a different approach for solving inverse problems: the Bayesian framework.

In the Bayesian framework, the model is treated as an encoding of the relation between the knowns and the unknowns of the system, as opposed to being treated as an equation that has to be inverted. 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 to collecting the data is incorporated in the distribution of $u$, called the *prior distribution* with measure $\mu_0$. The remaining variable $\eta$ is used to represent the uncertainty in $y$, such as measurement noise and modelling error. These uncertainties are often modelled using a mean-zero Gaussian distribution.

The coupling of these random variables given by (2.4) allows us to answer a wide range of questions about the model through the use of conditioning. For instance, we can consider the probability density of the conditioned random variable $y|u$, called the *likelihood function*. If $\rho$ denotes the density function of $\eta$, the likelihood function is

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

The observations are then realizations of the *data-generating measure* with density $\ell(y|u^*)$, where $u^*$ is fix. Motivated by the wide use of the Gaussian distribution for error modelling, we assume throughout the thesis that the likelihood function can be written as

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

where $\Phi$ 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

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

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

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

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

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

$$\mathrm{d}\mu^y(u) \propto \ell(y|u)\mathrm{d}\mu_0(u), \tag{2.8}$$

where the $\propto$ symbol denotes proportionality up to a constant factor $Z_y$ only dependent on the data $y$, called the *normalizing constant* or *model evidence*, and given by

$$Z_y = \int_X \ell(y|u)\mathrm{d}\mu_0(u). \tag{2.9}$$

This framework for solving inverse problems can be used to express filtering problems, and thus also provide a solution to the latter.

Extending the proposed formulation of inverse problems to filtering problems is done in an intuitive way. The use of a prior distribution to model existing knowledge about the model paramater remains unchanged. However, more structure has to be given to the uncertainty

since it now requires a sequence of independent random variables $\eta_1, \dots, \eta_T$ for each step of the filtering. This provides us with a sequence of potentials $\Phi_1, \dots, \Phi_T$ and likelihoods $\ell_1, \dots, \ell_T$, which in turn provide us with a sequence of partial solutions

$$\mathrm{d}\mu_t^y(u) \propto \ell_t(u|y_{1:t})\mathrm{d}\mu_0(u).$$

However, when no natural sequence of intermediate measures exists, we need to create this sequence artificially. One common way to create this sequence of artificial states is to consider an increasing subset of the observations for each of these measures. Given the forward response operator $\mathcal{G} : X \to Y^T$, we consider the sequence of potentials

$$\Phi_t(u; y) = \|y_{1:t} - \mathcal{G}(u)_{1:t}\|_{Y^t}^2 = \sum_{i=1}^{t} \|y_i - \mathcal{G}_i(u)\|_Y^2 . \tag{2.10}$$

This method has two main advantages. Firstly, it does not require to have all observations available at each step of the filtering, allowing to perform inference on a running experiment. Secondly, the sequence of posteriors acts as an interpolation between the prior and the posterior, as illustrated in the right panel of Figure 2 of page 15. This property will be shown to be central in the design of the numerical approximations presented later.

While the argument presented above provides the correct result, the extension of Bayes' rule from events to probability measures is not valid. Moreover, it is not clear which assumptions were made about the model to ensure existence the solutions, and more generally, well-posedness has not been discussed. We present a more rigorous proof and characterization of well-posed inverse problems in the following section.

## 2.4 Well-posedness of Bayesian inverse problems

In this section, we will focus on well-posedness of Bayesian inverse problems. However, as shown in the previous section, a filtering problem can be interpreted as a sequence of inverse problems defined by forward response operators and likelihoods $(\mathcal{G}_1, \ell_1), \dots, (\mathcal{G}_T, \ell_T)$. A filtering problem is then said to be well-posed if every intermediate inverse problem is well-posed.

We start by stating the following assumptions proposed by Stuart [Stu10]

**Assumption 2.2.** 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. These assumptions provide us with upper and lower bounds on $\Phi$, as well as its Lipschitz continuity with respect to the data $y$ and the parameter $u$. However, since potential functions are often of the form of (2.7), we are interested in refining these assumptions to the shared structure of such problems.

**Assumption 2.3.** 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 2.4.** Assume that $\mathcal{G} : X \to \mathbb{R}^N$ satisfies Assumption 2.3. Then, for any covariance matrix $\Gamma$, the potential function given by (2.7) satisfies Assumption 2.2 with $(Y, \|\cdot\|_Y) = (\mathbb{R}^N, \|\cdot\|_\Gamma)$.

*Proof.* Let $\mathcal{G} : X \to \mathbb{R}^N$ be a forward operator satisfying Assumption 2.3, $\Gamma$ be a covariance matrix and $\Phi$ be the potential function given by (2.7), i.e. $\Phi(u; y) = \|y - \mathcal{G}(u)\|_\Gamma^2$. We now prove $\Phi$ satisfies Assumptions 2.2(1-4).

1. Let $\epsilon > 0$ and $r > 0$, then for $M = 0$ and all $u \in X$, $y \in Y$ with $\|y\|_Y < r$ we have by positivity of the norms $\|\cdot\|_\Gamma$ and $\|\cdot\|_X$

$$\Phi(u; y) = \|y - \mathcal{G}(u)\|_\Gamma^2 \geq 0 \geq 0 - \epsilon \|u\|_X^2 = M - \epsilon \|u\|_X^2 .$$

2. Let $r > 0$, then from Assumption 2.3(1) there is for $\epsilon = 1$ an $M \in \mathbb{R}$ such that $\|\mathcal{G}(u)\|_\Gamma \leq \exp(\|u\|_X^2 + M)$ for every $u \in X$. Let $K = r^2 + \exp(2r^2 + 2M) > 0$, then for all $u \in X$ and $y \in Y$ with $\max \|u\|_X , \|y\|_\Gamma < r$ we have by

$$\Phi(u; y) = \|y - \mathcal{G}(u)\|_\Gamma^2 \leq \|y\|_\Gamma^2 + \|\mathcal{G}(u)\|_\Gamma^2 \leq \|y\|_\Gamma^2 + \exp(\|u\|_X^2 + M)^2 \leq K.$$

3. Let $r > 0$, from Assumption 2.3(2) there is a $K > 0$ such that for every $u_1, u_2 \in X$ with $\max\{\|u_1\|_X , \|u_2\|_X\} < r$ we have $\|\mathcal{G}(u_1) - \mathcal{G}(u_2)\|_\Gamma \leq K \|u_1 - u_2\|_X$. Let $L = 2rK^2$, for all $u_1, u_2 \in X$ and $y \in Y$ with $\max\{\|u_1\|_X , \|u_2\|_X , \|y\|_\Gamma\} < r$ we obtain by definition of $L$ and reverse triangle inequality

$$\begin{aligned} |\Phi(u_1; y) - \Phi(u_2; y)| &= \left| \|y - \mathcal{G}(u_1)\|_\Gamma^2 - \|y - \mathcal{G}(u_2)\|_\Gamma^2 \right| \leq \|(y - \mathcal{G}(u_1)) - (y - \mathcal{G}(u_2))\|_\Gamma^2 \\ &= \|\mathcal{G}(u_1) - \mathcal{G}(u_2)\|_\Gamma^2 \leq K^2 \|u_1 - u_2\|_X^2 \leq L \|u_1 - u_2\|_X . \end{aligned}$$

4. Let $\epsilon > 0$, $r > 0$ and $y_1, y_2 \in Y$ with $\max\{\|y_1\|_\Gamma , \|y_2\|_\Gamma\} < r$ and $u \in X$ be arbitrary. Since $\Phi$ is continuously differentiable with respect to $y$, we obtain

$$|\Phi(u; y_1) - \Phi(u; y_2)| \leq \sup_{\|y\|_\Gamma \leq r} \|\nabla_y \Phi(u; y)\|_\Gamma \|y_1 - y_2\|_\Gamma .$$

Additionally, since $\nabla_y \Phi(u; y) = 2\langle y - \mathcal{G}(u), \cdot \rangle_\Gamma$ and making use of the Cauchy-Schwarz inequality, we have for any $y \in Y$

$$\begin{aligned} \|\nabla_y \Phi(u; y)\|_\Gamma &= \sup_{\|w\|_\Gamma = 1} |\nabla_y \Phi(u; y)[w]| = \sup_{\|w\|_\Gamma = 1} 2|\langle y - \mathcal{G}(u), w \rangle_\Gamma| \\ &\leq \sup_{\|w\|_\Gamma = 1} 2 \|y - \mathcal{G}(u)\|_\Gamma \|w\|_\Gamma = 2 \|y - \mathcal{G}(u)\|_\Gamma . \end{aligned}$$

Furthermore, using Assumption 2.3(2) there is an $M \in \mathbb{R}$ such that $\|G(u)\|_\Gamma \leq \exp(\epsilon \|u\|_X^2 + M)$ for all $u \in X$, giving for any $y \in Y$ with $\|y\|_\Gamma \leq r$

$$\begin{aligned} \|y - \mathcal{G}(u)\|_\Gamma &\leq \|y\|_\Gamma + \|\mathcal{G}(u)\|_\Gamma \leq r + \exp(\epsilon \|u\|_X^2 + M) \\ &= \exp(\epsilon \|u\|_X^2 + M)(r \exp(-\epsilon \|u\|_X - M) + 1) \\ &\leq \exp(\epsilon \|u\|_X^2 + M)(r \exp(-M) + 1) \\ &= \exp(\epsilon \|u\|_X^2 + C), \end{aligned}$$

where $C = M + \log(r \exp(-M) + 1)$. All together, we obtain the desired bound

$$|\Phi(u; y_1) - \Phi(u; y_2)| \leq \exp(\epsilon \|u\|_X^2 + C) \|y_1 - y_2\| .$$

□

Using these assumptions, we can now proceed to provide a formal proof of Bayes' rule for continuous random variables and of equation (2.8), which we recall gives the following relation for the posterior distribution

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

The following theorem will play a central role in our proof

**Theorem 2.5.** Let $\mu, \nu$ 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 $\mu$ has Radon-Nikodym derivative $\phi$ with respect to $\mu$. Further assume that the conditional distributions of $x|y$ under $\nu$, denoted by $\nu^y(\mathrm{d}x)$, exists. Then, the conditional distribution of $x|y$ under $\mu$, 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

$$\frac{\mathrm{d}\nu^y}{\mathrm{d}\mu^y}(x) = \begin{cases} \frac{1}{c(y)}\phi(x,y) & \text{if } c(y) > 0 \\ 1 & \text{otherwise} \end{cases} \tag{2.11}$$

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 in Section 10.2 of Dudley [Dud02]. □

We now prove the infinitesimal version of Bayes' rule as stated in (2.8), and prove it using the previously stated theorem.

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

$$\frac{\mathrm{d}\mu^y}{\mathrm{d}\mu_0}(u) = \frac{1}{Z_y}\exp(-\Phi(u; y)). \tag{2.12}$$

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

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

where $C(y) := 1/\rho(y)$ is well defined since Assumption 2.2(2) gives an upper bound on $\Phi$ thus also giving a strictly positive lower bound on $\rho$. We further define two measures $\nu_0, \nu$

on $Y \times X$ by

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

Since $\mathcal{G}$ is continuous and $\mu_0(X) = 1$, it is also $\mu_0$-measurable. Thus, $\nu$ is well-defined and continuous with respect to $\nu_0$ with Radon-Nikodym derivative

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

Since $\nu_0$ 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 $\nu_0$ is then $\nu_0^y = \mu_0$. In addition, again using Assumption 2.2(2), we have a strictly positive lower bound on $\rho$ 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 2.5, the conditional distribution of $u|y$ under $\nu$, denoted $\mu^y$, 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 show that the problem is well-posed, we still need to prove that the solution is continuous with respect to the data $y$. Proving continuity of the solution requires us to define a metric over the space of probability measures, we make use of the following metric

**Definition 2.7.** Let $\mu$ and $\mu'$ be two measures absolutely continuous to an arbitrary measure $\nu$. The *Hellinger distance* between $\mu$ and $\mu'$ is then defined as

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

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

**Definition 2.8.** A probability measure $\mu$ 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 us to use many common distributions, such as Gaussians by Fernique's theorem 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 2.9.** Let $\Phi$ satisfy Assumptions 2.2 and $\mu_0$ be a light-tailed probability measure with support equals to $X$. Then $\mu^y$ given in 2.8 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 respective model evidences. We start by showing that both $Z$ and $Z'$ are bounded by above and below by positive constants. Using Assumption 2.2(2) we have for any $r > 0$

$$Z = \int_X \exp(-\Phi(u;y))\mathrm{d}\mu_0(u) \geq \int_{\|u\|_X \leq r} \exp(-C)\mathrm{d}\mu_0(u) = \exp(-C)\mu_0(\|u\|_X \leq r) > 0,$$

where the last inequality holds because $\mu_0$ has support equal to $X$. This shows that $Z$, and analogously $Z'$, have a constant positive lower bound. We introduce the notation $a \wedge b = \min\{a, b\}$ and using the Lipschitz continuity with Lipschitz constant 1 of the exponential function, we have for any $u \in X$

$$\begin{aligned}
&|\exp(-\Phi(u;y)) - \exp(-\Phi(u;y'))| \\
=&\exp\left(-\frac{\Phi(u;y)+\Phi(u;y')}{2}\right)\left|\exp\left(\frac{\Phi(u;y')-\Phi(u;y)}{2}\right) - \exp\left(\frac{\Phi(u;y)-\Phi(u;y')}{2}\right)\right| \\
\leq&\exp\left(-[\Phi(u;y)\wedge\Phi(u;y')]\right)|\Phi(u;y)-\Phi(u;y')|.
\end{aligned}$$

Integrating on both sides of the equation allows us to obtain the following upper bound

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

Further, since $\mu_0$ 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 $\epsilon/2$ for Assumption 2.2(1) and (4), we get

$$\begin{aligned} |Z - Z'| &\leq \int_X \exp(\frac{1}{2}\epsilon \|u\|_X^2 - M) \exp(\frac{1}{2}\epsilon \|u\|_X^2 + C) \|y - y'\|_Y \, \mathrm{d}\mu_0(u) \\ &= C \|y - y'\|_Y \int_X \exp(\epsilon \|u\|_X^2) \mathrm{d}\mu_0(u) \\ &\leq C \|y - y'\|_Y . \end{aligned} \tag{2.13}$$

From the definition of the Hellinger distance, we further have

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

By using Assumptions 2.2(1) and (4), and the same reasoning as above, we get

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

This, together with the constant postivie lower bound on $Z$, shows $I_1 \leq C \|y - y'\|_Y$. Additionally, we can bound the $I_2$ term using the following result

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

where the last inequality holds because both $Z$ and $Z'$ are strictly greater than 0 and using bound (2.13). Moreover using Assumption 2.2(2) and that $\mu_0$ is light-tailed, we can easily

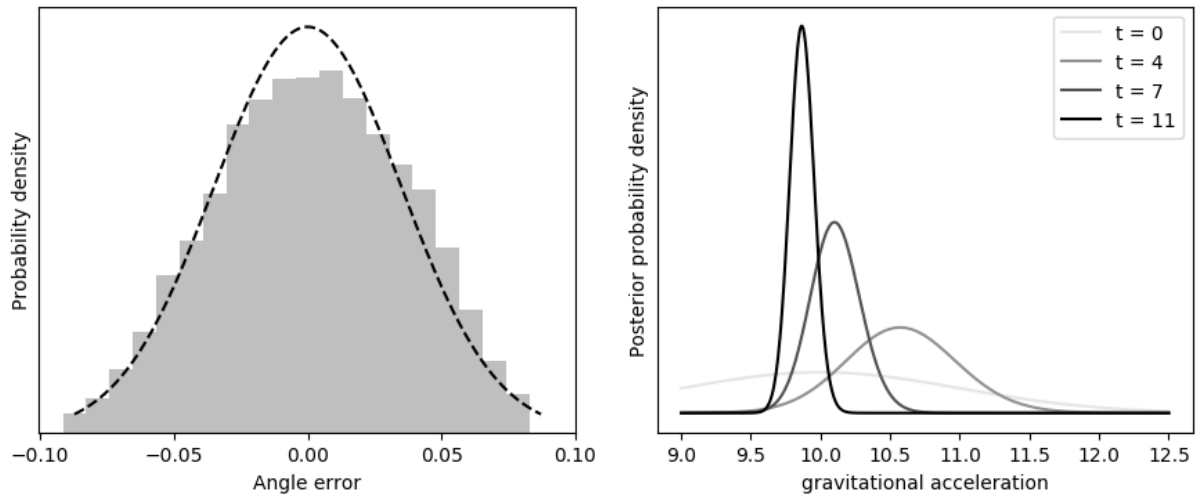


Figure 2: Left: numerical forward propagation of time uncertainties through the pendulum model. Dashed line is the density function of the Gaussian used to model the uncertainty in our probabilistic model. Right: Sequence of posterior densities after observing $t = 0, 3, 6, 9$ data points from light to dark, the lightest and darkest being respectively the prior and full posterior density. We see how the sequence of posteriors interpolate between the prior and full posterior distributions.

show that $\int_X \exp(-\Phi(u; y'))\mathrm{d}\mu_0(u)$ is bounded from above by a constant. This shows that $I_2 \leq C \left\|y - y'\right\|_Y^2$, hence completing the proof. □

We have shown that the Bayesian inverse problem is well-defined under light conditions on $\Phi$, given in Assumption 2.2, and provided that the prior measure $\mu_0$ 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 Another look at the Pendulum problem

To model the pendulum problem in the Bayesian framework, we need to define a prior measure for the value of the parameter $g$ and find a distribution to model the measurements error.

To choose our prior measure, we first note that gravity attracts objects towards the center of the Earth. Given the coordinates system we have chosen, we know that the gravitational acceleration must be a positive value. We can also convince ourselves, by prior experiments or high-school knowledge, that the value of $g$ is close to 10 and lower than 20. Provided no additional information, the natural choice for the prior distribution to incorporate the knowledge stated above is to use a Gaussian distribution centered in 10 with variance 1, truncated to the interval $[0, 20]$.

For modelling the uncertainty in our measurements, we assume that measurement errors are symmetrical around 0, and further assume an error in the measurements of approximately half a second. In order to obtain an approximation of the angle uncertainty, we propagate the time uncertainty through the forward response operator of the pendulum. This gives us an estimate of the uncertainty in the angle measurements. The forward propagation of the error is numerically estimated using the Monte Carlo method, described in Section 3, in which simulated time measurements were perturbated by a zero-mean Gaussian variable of standard deviation 0.5 and used as new time measurements for the model. The result of this experiment leads us to model the uncertainty in the angles measurements with a zero-mean Gaussian of variance 0.035. The left panel of Figure 2 illustrates the result of the numerical experiment.

We consider the estimation of the gravitational acceleration as a filtering problem, in which the sequence of likelihoods is defined using the construction proposed in equation (2.10)

$$\ell_t(x|g) := \exp\left(-\frac{1}{2}\sum_{i=1}^{t}|x_i - \mathcal{G}_i(g)|^2\right), \tag{2.14}$$

where $\mathcal{G}_i(g)$ is the solution of the equation (2.3) at time $t = t_i$ with parameter $g$. The un-normalized posteriors resulting from this model are illustrated in the right panel of Figure 2.

While the choice of prior is trivially light-tailed, we still need the following lemma to show that the pendulum problem is well-posed.

**Lemma 2.10.** The pendulum problem satisfies Assumption 2.3.

*Proof.* We first prove that the solution $x(t)$ of the initial value problem is bounded, which is a stronger result than Assumption 2.3(1). Let $g \in [0, 20]$ be fixed, we start by defining the Hamiltonian

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

By simple calculations, 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 $t^*$ such that $x(t^*) = \pi$, giving

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

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 for every $g \in [0, 20]$ and $t > 0$ implying $\mathcal{G}_t$ is bounded as well, which in turn implies Assumption 2.3(1). 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. □

We have proven that the pendulum problem is well posed, and can now use the Bayesian solution to the inverse problem to estimate the true value of $g$ from our measurements of the system. The next section will study numerical approximation methods to estimate this solution.

# 3 Sequential Monte Carlo

The last section was motivated by validating the use of Bayesian methods for solving inverse problems. However, the solution of Bayesian filtering problems, given by the sequence of posterior measures, often does not have a closed form solution. This requires us to construct numerical estimates of the sequence of posterior measures.

The class of estimators we will be considering for this task are *sampling estimators*. Such estimators are constructed by using a set of *particles* representing weighted points of the parameter space $X$ in order to estimate the distribution of interest. A building block of such estimators is the following operator

**Definition 3.1.** Let $\mathcal{P}(X)$ denote the set of probability measures on $X$. Let $M \in \mathbb{N}$, for any measure $\mu \in \mathcal{P}(X)$ 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 $\mu$. From the randomness of $U_1, \dots, U_M$, it follows that $S^M\mu$ is a not a probability measure, but a random variable taking values in $\mathcal{P}(X)$. The operator $\mu \mapsto S^M\mu$ is called *sampling operator*.

The sampling operator is the simplest building block we will use in the construction of particle algorithms, we will denote it by

$$\mu^y \xrightarrow{\text{sample}} S^M\mu^y.$$

When only using this single building block, we obtain the *Monte Carlo* estimate, or *empirical measure*, of $\mu^y$. One can easily check that the Monte Carlo estimate is unbiased, and we will later prove that it has a standard deviation of $\mathcal{O}(M^{-1/2})$, regardless of the dimension of $X$. This algorithm alone can only be used in specific situations where direct sampling from the distribution of interest is possible. In Bayesian filtering however, the sequence of posterior distributions $\mu_t^y$ is often too complex to be directly sampled from, and a simple Monte Carlo estimate cannot be constructed. This requires us to extend the Monte Carlo method to let us approximate non-trivial distributions as well. *Sequential Monte Carlo* (SMC) is such a method.

The rest of the section will be structured as follows. In Section 3.2, we will present more building blocks for constructing estimation algorithms and show how we can use them to construct several popular algorithms. These iterations over the simple Monte Carlo estimate will then finally lead us to the SMC algorithm. In Section 3.3, we prove convergence properties of different algorithm blocks, and use these properties to show that

the SMC algorithm provides a good estimate of each distribution of the sequence of filtering posterior distributions.

## 3.1 Sequential Monte Carlo

**Importance sampling** An additional basic building block of SMC samplers is the *Importance Sampling* (IS) method. The IS method can be used to estimate probability measures $\mu^y$ of the form of (2.8) for which the Radon-Nikodym derivative with respect to the prior $\mu_0$ is only known up to an unknown normalizing constant. To create an estimate, the IS method relies on using an *importance distribution* $\nu$ from which we can easily generate samples. If $\mu^y \ll \nu \ll \mu_0$, by the chain rule for Radon-Nikodym derivatives and substituting the expression of the posterior distribution (2.8), we have

$$\begin{aligned}\frac{\mathrm{d}\mu^y}{\mathrm{d}\nu}(u) &= \frac{\mathrm{d}\mu^y}{\mathrm{d}\mu_0}(u)\left(\frac{\mathrm{d}\nu}{\mathrm{d}\mu_0}(u)\right)^{-1} \\ &= \frac{1}{Z_y}\ell(u|y)\left(\frac{\mathrm{d}\nu}{\mathrm{d}\mu_0}(u)\right)^{-1} \\ &=: w(u)/Z_y.\end{aligned}$$

Where $w$ is the *importance weight* function. This gives us for any $\mu^y$-integrable function $f$ the following equality

$$\begin{aligned}\mu^y(f) &= \nu(w \cdot f)/Z_y, \\ Z_y &= \nu(w).\end{aligned}$$

These two equalities allow us to use a Monte Carlo estimate of $\nu$ to create an estimate of $\mu^y$. Given i.i.d. random variables $V_1, \ldots, V_M$ distributed according to $\nu$, we consider the Monte Carlo estimate $S^M\nu$ to define the estimates

$$\hat{Z}_y = S^M\nu(w) = \frac{1}{M}\sum_{i=1}^{M} w(V_i),$$

$$\hat{\mu}^y(f) = \frac{S^M\nu(w \cdot f)}{\hat{Z}_y} = \frac{1}{M\hat{Z}_y}\sum_{i=1}^{M} w(V_i)f(V_i) = \sum_{i=1}^{M} W_i f(V_i),$$

where $W_i$ are the *normalized importance weights* given by

$$W_i = \frac{w(V_i)}{\sum_{i=1}^{M} w(V_i)}.$$

Before discussing properties of IS, we first give a slightly different formulation of the IS method to allow us using it as a building block for other algorithms. Expressing $\mu^y$ in terms of $\nu$ can be encapsulated in a non-linear *reweighting operator* $L^w$ given by

$$(L^w \nu)(\mathrm{d}u) = \frac{w \cdot \nu(\mathrm{d}u)}{\nu(w)},$$

which, from the definition of $w$, gives us the equality $\mu^y = L^w \nu$. We can thus define a reweighting transition that can be used to construct algorithms as

$$\nu \xrightarrow{\text{reweigh}} L^w \nu = \mu^y.$$

Having this new building block, we can combine the reweighting operator with the sampling operator to obtain

$$\nu \xrightarrow{\text{sample}} S^M \nu \xrightarrow{\text{reweigh}} L^w S^M \nu.$$

We can then verify that the IS algorithm, written as a composition of the two blocks presented, delivers the same result as before

$$(L^w S^M \nu)(f) = \frac{S^M \nu(w \cdot f)}{S^M \nu(w)} = \frac{\sum_{i=1}^M w(V_i) f(V_i)}{\sum_{i=1}^M w(V_i)} = \sum_{i=1}^M W_i f(V_i) = \hat{\mu}^y(f).$$

This formulation is thus identical to the previous formulation. However, we will later see how this modular definition of algorithms makes it easier to extend them and study their properties.

The IS method provides an unbiased estimate for the normalizing constant $Z_y$ and a biased estimate for $\mu^y(f)$ with both bias of order $\mathcal{O}(M^{-1})$ and variance of order $\mathcal{O}(M^{-1/2})$ [APSAS17]. However, while this method provides asymptotic convergence results, we are interested in the non asymptotic regime and the constants involved. In this situation were $M$ is fixed, the variance of the estimation is proportional to the variance of the importance weights [APSAS17]. This means that the estimator is only useful if the importance distribution is *close enough* to the distribution we want to sample from. In practice, finding good importance distributions can be hard for non-standard target distributions, which makes it hard or impossible to directly use IS. Yet, this does not imply that ideas of IS should be completely discarded, and we now show how IS can be extended to sample filtering distributions.

Considering the sequence of distributions induced by the potential functions given in (2.10), we can expect consecutive distributions to be approximately close to each other. This

means that each distribution $\mu_{t-1}^y$ should be a good importance distribution to use importance sampling on the next distribution $\mu_t^y$. This gives us for each importance sampling step $t$ from $\mu_{t-1}^y$ to $\mu_t^y$ an importance weight function $w_{t-1}$ and a reweighting operator $L_{t-1} = L^{w_{t-1}}$. We can thus iterate reweighting steps to map the prior distribution to any arbitrary distribution of the sequence of posteriors giving $\mu_t^y = (L_{t-1} \circ \ldots \circ L_0)\mu_0$ and

$$\mu_0 \xrightarrow{\text{reweigh}} L_0\mu_0 = \mu_1^y \xrightarrow{\text{reweigh}} \ldots \xrightarrow{\text{reweigh}} L_{t-1}\mu_{t-1}^y = \mu_t^y.$$

Similar to the IS method, we can use this equality to create an algorithm by applying the sequence of operators to a Monte Carlo estimate of $\mu_0$, giving the algorithm

$$\begin{aligned} \mu_0 \xrightarrow{\text{sample}} S^M\mu_0 &\xrightarrow{\text{reweigh}} L_0S^M\mu_0 \\ &\ldots \\ &\xrightarrow{\text{reweigh}} (L_{t-1} \circ \ldots \circ L_0)S^M\mu_0. \end{aligned}$$

We note that in the common case where the posterior distributions are induced by potentials $\Phi_t$ given by (2.7), the importance weight $w_{t-1}$ is then given by

$$w_{t-1}(u) = \frac{l_t(u|y)}{l_{t-1}(u|y)} = \exp\left(\Phi_{t-1}(u;y) - \Phi_t(u;y)\right) = \exp(-\left|y_t - \mathcal{G}_t(y)\right|_\Gamma^2), \tag{3.1}$$

this implies that the reweighting operator can be efficiently implemented by only requiring $M$ evaluations of the forward response operator.

While this appears to be a valid algorithm for estimating the sequence of distributions, it does not work in general. A simple counter-example would be to consider estimating a sequence of Gaussian distributions shifting away from $\mu_0$ as $t$ increases. Since the initial set of samples will, with high probability, be concentrated around the mean of $\mu_0$, as $t$ increases the distributions will shift away from where the samples are located. This will cause the weights of each particle to decrease as $t$ increases, which in turn will increase the variance of the estimate and make each estimate worse than the previous one.

**Sequential Importance Sampling** To avoid this, we introduce an additional step after applying the reweighting operator to *correct* the position of the samples by getting them closer to the target distribution. This should then result in a better importance distribution for the next step. To do this correction, we construct for each posterior distribution $\mu_t^y$ an ergodic Markov chain with stationary distribution $\mu_t^y$ characterized by its Markov kernel $K_t$ from which we know how to sample the next state given the current state, i.e. for every $u \in X$ we can sample from $K_t(u, \cdot)$. Since the Markov chain is ergodic, the limiting distribution of the Markov chain is equal to its stationary distribution $\mu_t^y$ regardless of the

initial distribution [RC05, Chapter 6]. Considering the operator induced by the Markov kernel $K_t$ given by

$$(K_t\mu)(f) = \mu(K_t(f)) = \int_X \int_X f(u)K_t(x, \mathrm{d}u)\mu(\mathrm{d}x),$$

the theorem stated in the previous sentence implies that applying the operator $K_t$ to any distribution will get it closer $\mu_t^y$, the stationary distribution of the Markov chain. This lets us define a new correction building block

$$\mu \xrightarrow{\text{correct}} K_t\mu.$$

We can thus replace each step of reweighting in the previous algorithm by the operator $\Psi_t = K_t L_{t-1}$, giving for every $\mu$

$$(\Psi_t\mu)(\mathrm{d}u) = (K_t L_{t-1}\mu)(\mathrm{d}u) = \frac{\mu(w_{t-1} \cdot K_t(\mathrm{d}u))}{\mu(w_{t-1})},$$

which in turn, for every $\mu_t^y$-integrable function $f$ gives

$$(\Psi_t\mu)(f) = \frac{\mu(w_{t-1} \cdot K_t)(f)}{\mu(w_{t-1})} = \frac{1}{\mu(w_{t-1})} \int_X \int_X f(u)K_t(x, \mathrm{d}u)w_{t-1}(x)\mu(\mathrm{d}x).$$

We can then verify that the operator $\Psi_t$ maps $\mu_{t-1}^y$ to the distribution $\mu_t^y$ by using that $\mu_t^y$ is invariant under $K_t$

$$(\Psi_t\mu_{t-1}^y)(\mathrm{d}u) = (K_t L_{t-1}\mu_{t-1}^y)(\mathrm{d}u) = (K_t\mu_t^y)(\mathrm{d}u) = \mu_t^y(\mathrm{d}u),$$

which gives us the transitions from any posterior distribution $\mu_{t-1}^y$ to $\mu_t^y$ by

$$\mu_{t-1}^y \xrightarrow{\text{reweigh}} L_{t-1}\mu_{t-1}^y \xrightarrow{\text{correct}} \Psi_t\mu_{t-1}^y = \mu_t^y.$$

As done previously, we can apply the sequence of operators to map $\mu_0$ to any posterior distribution of the sequence $\mu_t^y$ giving $\mu_t^y = (\Psi_t \circ \ldots \circ \Psi_1)\mu_0$ and

$$\begin{aligned} \mu_0 &\xrightarrow{\text{reweigh}} L_0\mu_0 \xrightarrow{\text{correct}} \Psi_1\mu_0 = \mu_1^y \\ &\ldots \\ &\xrightarrow{\text{reweigh}} L_{t-1}\mu_{t-1}^y \xrightarrow{\text{correct}} \Psi_t\mu_{t-1}^y = \mu_t^y. \end{aligned}$$

Transforming this into an algorithm by performing a step of sampling on $\mu_0$ before applying the sequence of operators $\Psi_t \circ \ldots \circ \Psi_1$ gives the *Sequential Importance Sampling* (SIS) algorithm

$$\begin{aligned}\mu_0 \xrightarrow{\text{sample}} S^M\mu_0 \xrightarrow{\text{reweigh}} L_0 S^M\mu_0 \xrightarrow{\text{correct}} \Psi_1 S^M\mu_0 \\ \ldots \\ \xrightarrow{\text{reweigh}} L_{t-1}(\Psi_{t-1} \circ \ldots \circ \Psi_1)S^M\mu_0 \xrightarrow{\text{correct}} (\Psi_t \circ \ldots \circ \Psi_1)S^M\mu_0.\end{aligned}$$

It remains to discuss the contruction of the Markov Kernels $K_1, \ldots, K_T$. The optimal choice for each step $t$ is the Markov chains generating i.i.d. samples of the posterior distribution $\mu_t^y$. However, this choice requires us being able to sample from $\mu_t^y$, which is not possible otherwise we could just use a Monte Carlo approximation. Thankfully, much work has been put in designing Markov kernels with specific invariant distribution from which it is easy to sample.

**Sequential Monte Carlo** We have defined with SIS an algorithm that incrementally updates a population of weighted independent samples to fit a sequence of distributions. However, even if the SIS algorithm tries to correct the position of the particles to match the sequence of posterior distributions, the transitions applied by the Markov kernels only guarantee convergence in the asymptotic regime with respect to the number of transitions and particles. Since we are clearly not in this regime, we need to measure the impact of the re-weighting procedure on the quality of the Monte Carlo estimate given by these weighted particles.

If we consider a Monte Carlo estimate induced by $M$ i.i.d. random variables $Y_1, \ldots, Y_M$ of variance $\sigma^2$, the variance of the uniformly weighted sum of these variables would be equal to $\sigma^2/M$. If we instead consider a weighted sum of these variables, such as the one given by the IS and SIS algorithms of the form

$$S = \frac{\sum_{i=1}^M w_i Y_i}{\sum_{i=1}^M w_i}, \tag{3.2}$$

what would now be the variance of this sum? If each particle has an approximately equal importance weight, and thus equal participation in the estimate, we would have a variance close to the uniform case $\sigma^2/M$. We know this might not be the case and we want to measure the effective number of independent random variables $M_{eff}$ used in this sum, called the *effective sample size* (ESS). We assume that $S$ was generated by summing $M_{eff}$ equally weighted i.i.d. samples and thus has a variance of $\sigma^2/M_{eff}$. This lets us solve for $M_{eff}$ by using the variance of the weighted sum $S$ given in (3.2). The resulting effective sample size is then

$$M_{eff} = \frac{\left(\sum_{i=1}^{M} w_i\right)^2}{\sum_{i=1}^{M} w_i^2}. \tag{3.3}$$

We can verify that this value is indeed bounded by the actual number of samples $M$

$$M_{eff} = M^2 \frac{\left(\frac{1}{M}\sum_{i=1}^{M} w_i\right)^2}{\sum_{i=1}^{M} w_i^2} \leq M^2 \frac{\frac{1}{M}\sum_{i=1}^{M} w_i^2}{\sum_{i=1}^{M} w_i^2} = M,$$

where the inequality is the weighted Jensen's inequality. This tells us that if the weights are too imbalanced, the effective size of our sample is going to be much smaller than the number of particles we have, which results in a Monte Carlo estimate of higher variance.

We are interested in modifying our algorithm to control the variance of the weights within our population, and thus keep the ESS over a certain level $M_{thresh}$ close to $M$. To achieve this, the Sequential Monte Carlo (SMC) algorithm controls the ESS after each step of reweighting and performs a step of *resampling* if the ESS is too low. The resampling will reset the population by applying the sampling operator to the current estimate. This will reset every weight to $M^{-1}$ and discard with high probability the particles with lower weights. Specifically, the resampling step maps the current population estimate of the posterior distribution $\mu_t^y$ to a new population with uniform weights,

$$\sum_{i=1}^{M} W_t^i \delta_{U_i} \xrightarrow{\text{resample}} S^M \left(\sum_{i=1}^{M} W_t^i \delta_{U_i}\right) = \sum_{i=1}^{M} \frac{1}{M} \delta_{V_i},$$

where $V_i = U_j$ with probability $W_t^j$. This finally gives us the sequence of measures $\nu_0^M, \ldots, \nu_T^M$ created by the SMC algorithm to approximate the sequence of posterior measures $\mu_0, \ldots, \mu_T^y$

$$\begin{aligned} \nu_0 &= S^M \mu_0 \\ \nu_{t+1}^M &= \begin{cases} \Psi_{t+1} \nu_t^M & \text{if } M_{eff} > M_{thresh}, \\ S^M \Psi_{t+1} \nu_t^M & \text{otherwise.} \end{cases} \end{aligned}$$

The final form of the standard Sequential Monte Carlo algorithm is given in Algorithm 1. We now proceed to proving convergence properties of the SMC algorithm.

**Algorithm 1:** Sequential Monte Carlo

*(1)* $\mu_0 \xrightarrow{sample} S^M\mu_0 = \nu_0^M$
Sample each $U_0^i$ from $\mu_0$ and set each weight to $W_0^i = M^{-1}$.

**for** $t = 1, \dots, T$ **do**
  *(2.1)* $K_t\nu_{t-1}^M \xrightarrow{reweigh} \Psi_t\nu_{t-1}^M$
  Adjust weights $\hat{W}_t^i = w_t(U_{t-1}^i)W_{t-1}^i$ and normalize them according to

$$W_t^i = \hat{W}_t^i \Big/ \left(\sum_{j=1}^{M} \hat{W}_t^j\right).$$

  *(2.2)* $\nu_{t-1}^M \xrightarrow{correct} K_t\nu_{t-1}^M$
  Sample each $U_t^i$ from $K_t(U_{t-1}^i, \cdot)$.

  *(2.3)* $\Psi_t\nu_{t-1}^M \xrightarrow{resample} \nu_t^M$
  Compute $M_{eff}$ according to (3.3).
  **if** $M_{eff} < M_{thresh}$ **then**
    Sample each $U_t^i$ from $\Psi_t\nu_{t-1}^M$ and set each weight to $W_t^i = M^{-1}$.
  **end**
**end**

## 3.2 Convergence of the SMC algorithm

In this section, we will prove properties of the different building blocks used in the construction of the SMC algorithm and use them to prove convergence of the algorithm. Since the output of the SMC algorithm is not deterministic, but instead a sequence of random probability measures, we will show that each of these approximations converges weakly in the number of particles $M$ to the approximated probability measures.

We first define the following metric over the space of random probability measures

**Definition 3.2.** For every $\mu$ and $\nu$ random variables with values in $\mathcal{P}(X)$, we define

$$d(\mu, \nu) := \sup_{|f|_\infty \leq 1} \sqrt{\mathbb{E}\left[|\mu(f) - \nu(f)|^2\right]}.$$

Then $d$ is a metric over the space of random measures on $X$.

Note that this metric bounds the variance of the unbiased estimate of an integral. Indeed, if we wish to estimate $\mu(f)$ by approximating $\mu$ using the unbiased estimate $\nu$ coming from a randomized algorithm, the variance of the estimate would be given by

$$\mathbb{V}(\nu(f)) = \mathbb{E}\left[|\nu(f) - \mathbb{E}[\nu(f)]|^2\right] = \mathbb{E}\left[|\nu(f) - \mu(f)|^2\right] \leq d(\nu, \delta_\mu)^2.$$

Hence, many of the results given in the coming lemma will not only show prove weak convergence of our approximations, but also bound the variance of the estimates as the number of particles increases.

We start by showing that the sampling operator can approximate any probability measure to an arbitrary precision using enough samples.

**Lemma 3.3.** The sampling operator satisfies

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

*Proof.* Let $\mu$ be an element of $\mathcal{P}(X)$ and $U_1, \ldots, U_M$ be i.i.d. random variables distributed according to $\mu$. For every measurable $f$ 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 $\mathbb{E}[f_i] = \mathbb{E}[f(U_i)] - \mu(f) = 0$, giving $\mathbb{E}[f_i f_j] = 0$ for $i \neq j$. Furthermore, since $|f|_\infty \leq 1$ we have

$$\mathbb{E}\left[f_i^2\right] = \mathbb{V}[f(U_i)] = \mathbb{E}\left[f(U_i)^2\right] - \mathbb{E}\left[f(U_i)\right]^2 \leq 1.$$

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. □

Before proving the main result, we formalize the assumption that consecutive posterior distributions from a filtering problem should be close to each other.

**Assumption 3.4.** The importance weights $w_0, \dots, w_{T-1}$ have the following property. There exists a $\kappa > 0$ such that for every $t \in \mathbb{N}$ and $u \in X$

$$\kappa \leq w_t(u) \leq 1/\kappa.$$

For completeness, we show that if the inverse problem is well-posed and the sequence of posterior distributions given by the potential functions (2.10) do satisfy the previous assumption on any bounded parameter space $X$.

**Lemma 3.5.** Assume that the potentials functions $\Phi_1, \dots, \Phi_T$ are given by (2.10), that the forward response operators $\mathcal{G}_1, \dots, \mathcal{G}_T$ satisfy Assumptions 2.3 and that $X$ is bounded. Then, given some fix observation $y \in Y^T$, the importance weights $w_0, \dots, w_T$ given in (3.1) satisfy Assumption 3.4.

*Proof.* We first show that $w_{t-1}(u)$ has an upper bound for each $t = 1, \dots, T$ by

$$w_{t-1}(u) = \frac{\ell_t(u|y)}{\ell_{t-1}(u|y)} = \exp\left(-\|y_t - \mathcal{G}_t(u)\|_Y^2\right) \leq 1$$

Then, using Assumption 2.3(1), there exists an $M \in \mathbb{R}$ such that

$$\|\mathcal{G}_t(u)\|_Y^2 \leq \exp(\|u\|_X^2 + M).$$

If further $C = \sup_{u \in X} \|u\|_X^2$, we have

$$\begin{aligned} w_{t-1}(u) &= \exp\left(-\|y_t - \mathcal{G}_t(u)\|_Y^2\right) \\ &\geq \exp\left(-\|y_t\|_Y^2 - \|\mathcal{G}_t(u)\|_Y^2\right) \\ &\geq \exp\left(-\|y_t\|_Y^2 - \exp(C + M)\right). \end{aligned}$$

The proof is complete by setting $\kappa = \min\{1, \exp\left(-\|y_t\|_Y^2 - \exp(C + M)\right)\}$. □

We now show that the composition of reweighting and correction in the SIS and SMC algorithm is Lipschitz continuous. This property is necessary to ensure that the asymptotic quality of the sampling approximation will not be lost by this additional step.

**Lemma 3.6.** Assume that the importance weights $w_0, \dots, w_{T-1}$ satisfy Assumption 3.4. Then for any $t \in \mathbb{N}$ and $\mu, \nu \in \mathcal{P}(X)$,

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

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

$$\begin{aligned}
(\Psi_t\mu - \Psi_t\nu)(f) &= (K_t L_{t-1}\mu - K_t L_{t-1}\nu)(f) \\
&= \frac{\mu(w_{t-1}K_t(f))}{\mu(w_{t-1})} - \frac{\nu(w_{t-1}K_t(f))}{\nu(w_{t-1})} \\
&= \frac{\mu - \nu}{\mu(w_{t-1})}(w_{t-1}K_t(f)) + \frac{\nu(w_{t-1}K_t(f))}{\mu(w_{t-1})} - \frac{\nu(w_{t-1}K_t(f))}{\nu(w_{t-1})} \\
&= \frac{\mu - \nu}{\mu(w_{t-1})}(w_{t-1}K_t(f)) + \nu(w_{t-1}K_t(f)) \left( \frac{1}{\mu(w_{t-1})} - \frac{1}{\nu(w_{t-1})} \right) \\
&= \frac{\mu - \nu}{\mu(w_{t-1})}(w_{t-1}K_t(f)) + \frac{\nu(w_{t-1}K_t(f))}{\mu(w_{t-1})\nu(w_{t-1})}(\nu - \mu)(w_{t-1}).
\end{aligned}$$

By Minkowski's inequality, we then have

$$\begin{aligned}
\sqrt{\mathbb{E}\left[|(\Psi_t\mu - \Psi_t\nu)(f)|^2\right]} \leq \mathbb{E}&\left[ \left| \frac{\mu - \nu}{\mu(w_{t-1})}(w_{t-1}K_t(f)) \right|^2 \right]^{\frac{1}{2}} \\
&+ \mathbb{E}\left[ \left| \frac{\nu(w_{t-1}K_t(f))}{\mu(w_{t-1})\nu(w_{t-1})}(\nu - \mu)(w_{t-1}) \right|^2 \right]^{\frac{1}{2}}.
\end{aligned}$$

In the second term, since $\frac{\nu(w_{t-1}K_t(f))}{\nu(w_{t-1})} = (\Psi_t\nu)(f)$ is the expectation of $f$ under the probability measure $\Psi_t\nu$ and $|f|_\infty \leq 1$, we have

$$\frac{\nu(w_{t-1}K_t(f))}{\nu(w_{t-1})} = (\Psi_t\nu)(f) \leq |f|_\infty (\Psi_t\nu)(X) \leq 1.$$

Further, since $w_{t-1}(u) \geq \kappa$ for all $u$, it holds that $\mu(w_{t-1}) \geq \kappa$ and $1/\mu(w_{t-1}) \leq 1/\kappa$. From these two bounds, we obtain

$$\begin{aligned}
\sqrt{\mathbb{E}\left[|(\Psi_t\mu - \Psi_t\nu)(f)|^2\right]} \leq \frac{1}{\kappa}&\mathbb{E}\left[|(\mu - \nu)(w_{t-1}K_t(f))|^2\right]^{\frac{1}{2}} \\
&+ \frac{1}{\kappa}\mathbb{E}\left[|(\mu - \nu)(w_{t-1})|^2\right]^{\frac{1}{2}}.
\end{aligned}$$

Finally, using that $0 \leq w_{t-1} \leq 1/\kappa$, we have $|\kappa w_{t-1}|_\infty \leq 1$ and similarly $[w_{t-1}K_t(f)](u) = \int_X f(y) w_{t-1}(u) K_t(u, \mathrm{d}y) \leq |f w_{t-1}|_\infty K_t(X, u) \leq 1/\kappa$, thus making $|\kappa w_{t-1}K_t(f)|_\infty \leq 1$. Bounding our expression by the supremum gives us

$$\sqrt{\mathbb{E}\left[|(\Psi_t\mu - \Psi_t\nu)(f)|^2\right]} \leq \frac{2}{\kappa^2} \sup_{|g|_\infty \leq 1} \sqrt{\mathbb{E}\left[|\mu(g) - \nu(g)|^2\right]} = \frac{2}{\kappa^2} d(\delta_\mu, \delta_\nu).$$

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

**Theorem 3.7.** Assume that the importance weights $w_0, \ldots, w_{T-1}$ satisfy Assumption 3.4 and consider the SMC algorithm with $M_{thresh} = M$. Then, for any $t \in \mathbb{N}$,

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

*Proof.* We prove this result by induction over $t$. For $t = 0$, the result holds by direct application of Lemma 3.3. For $t > 0$ we have using the triangle inequality, Lemmata 3.3 and 3.6, and using the induction hypothesis

$$\begin{aligned}
d(\delta_{\nu_t}, \nu_t^M) &= d(\Psi_t\nu_{t-1}, S^M\Psi_t\nu_{t-1}^M) \\
&\leq d(\Psi_t\nu_{t-1}, \Psi_t\nu_{t-1}^M) + d(\Psi_t\nu_{t-1}^M, S^M\Psi_t\nu_{t-1}^M) \\
&\leq \frac{2}{\kappa^2} d(\delta_{\nu_{t-1}}, \nu_{t-1}^M) + \frac{1}{\sqrt{M}} \\
&\leq \frac{2}{\kappa^2} \frac{1}{\sqrt{M}} \sum_{j=0}^{t-1} \left(\frac{2}{\kappa^2}\right)^j + \frac{1}{\sqrt{M}} \\
&= \frac{1}{\sqrt{M}} \sum_{j=0}^{t} \left(\frac{2}{\kappa^2}\right)^j.
\end{aligned}$$

□

This theorem shows that the SMC algorithm has an asymptotical standard deviation of $\mathcal{O}(d(\delta_{\mu_t}, \mu_t^M)) = \mathcal{O}(M^{-1/2})$. Furthermore, since the SMC algorithm is at its core an IS algorithm, it also has an asymptoticaly vanishing bias of order $\mathcal{O}(M^{-1})$. Since the mean square error (MSE) can be decomposed as $MSE = bias^2 + variance$, the mean square error of the SMC algorithm is dominated by the variance of the estimate, which is of $\mathcal{O}(M^{-1})$.

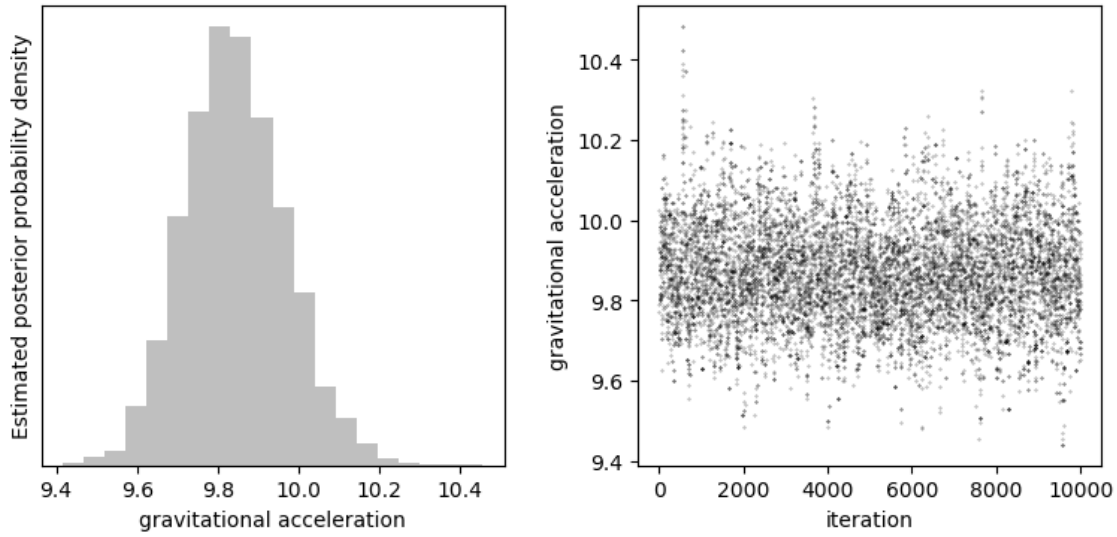


Figure 3: Left: estimated posterior distribution using $M = 10,000$ samples from the MCMC algorithm with burn-in period of 500 samples. Right: samples of the Markov chain for one run on the MCMC algorithm. We see that the burn-in period allows to have all samples in the zone of high density and converge.

# 4 Numerical solution to the pendulum problem

In this section, we present and analyze the numerical solution of the filtering problem associated to the pendulum experiment. We then run the SIS and SMC algorithms to approximate the expected value of the full posterior distribution and analyze their empirical convergence properties. As a reference, we also present the result of the MCMC approximation to the complete data set containing the following time measurements:

| $t_1$ | $t_2$ | $t_3$ | $t_4$ | $t_5$ | $t_6$ | $t_7$ | $t_8$ | $t_9$ | $t_{10}$ | $t_{11}$ |
|---|---|---|---|---|---|---|---|---|---|---|
| 1.29s | 4.04s | 6.64s | 9.66s | 12.42s | 15.61s | 18.33s | 21.31s | 24.04s | 26.94s | 29.98s |

In the first experiment, we create a reference estimate to the solution to the Bayesian inverse problem by considering all the available data collected during the physical experiment. We use the MCMC algorithm for $10,000$ steps after a burn-in period of 500 steps and construct the Markov chain using Metropolis-Hasting kernel with Gaussian proposals of variance 0.1. The burn-in period runs the Markov Chain for 500 steps without collecting the generated samples in order to reach the region of high probability without biasing the estimate. After 5 runs of the algorithm, we find an mean estimate of $\mathbb{E}[u|y] = 9.86 \pm \mathcal{O}(10^{-3})$ and a variance of $\mathcal{O}(10^{-6})$. This result is consistent with the unnormalized likelihoods presented in Figure 2, page 15. A typical run of the algorithm is given in Figure 3.

However, we are interested in approximating the sequence of filtering posterior distributions arising from the pendulum problem where the data only becomes available one observation

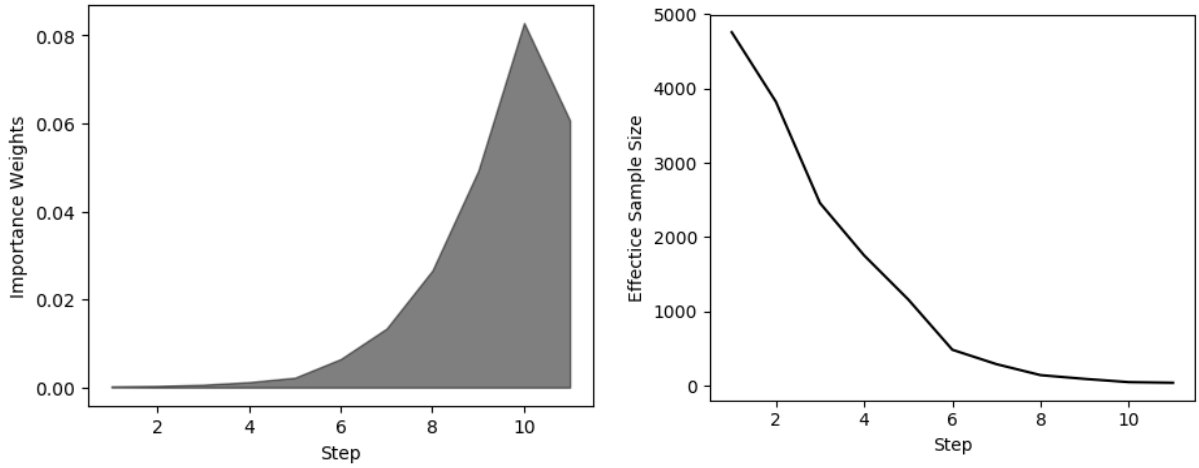


Figure 4: Left: the evolution of the importance weights of the particle estimates. The upper and lower bounds of the shaded area represent the value of the largest and smaller importance weight of the population. We can see that close to the end, a single particle contain almost all mass of the estimated probability measure. Right: evolution of the ESS over time, we see that after half the total number of iterations, the ESS is already only almost 1.

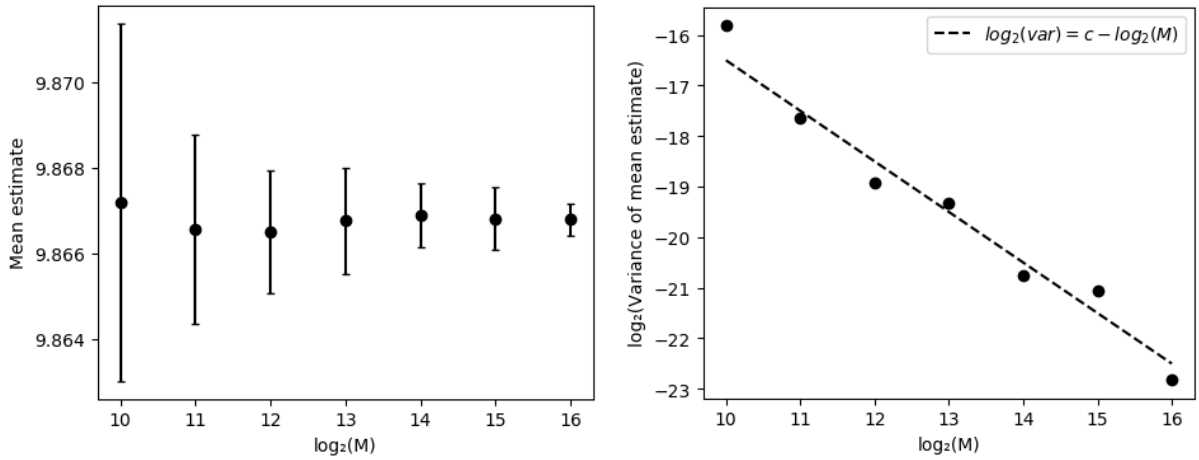


Figure 5: Both plots consider 40 runs of $M$-particles SMC estimates of the posterior expected value. Left: Each point is the mean and variance over the 40 SMC runs for each $M$. Right: Logarithmic plot of the variance of 40 estimates. Line illustrates the asymptotical $\mathcal{O}(M^{-1})$ behaviour of the variance proved in Theorem 3.7.

at a time. In the following experiments, we use the sequence of importance weights given in (2.14). Since the sequence of potential functions satisfy the assumptions stated in Lemma 3.5, the sequence of importance weights satisfy Assumption 3.4 and the estimates provided by the SIS and SMC algorithms thus weakly converge in $M$ to the true posteriors. In both algorithms, we use Metropolis-Hastings kernels with Gaussian proposals of variance 0.05.

In the second experiment, we illustrate why the step of resampling is crucial in the SMC algorithm. We run the SIS algorithm with $M = 5,000$ particles and monitor the ESS as well as the importance weights over time. As expected, the experiment shows that the variance of the weights increases with both very large and very small weights, meaning that most of the probability mass of the particle estimate is contained in only a few samples. Monitoring the ESS confirms this is also shown by the convergence of the ESS to 1. These two results are shown in Figure 4.

In a last experiment, we run the SMC algorithm with $M_{thresh} = M/2$ for several values of $M$ and repeat the experiment 40 times per configuration. We use 5 MCMC updates for the correction step to improve the approximations, this claim is theoretically justified by [MS18]. Figure 5 reports the result of the experiment, showing the convergence of the posterior expected value estimate.

To conclude this section, we compare computational costs of the MCMC and SMC algorithms. Each time new data becomes available, the MCMC algorithm needs to sequentially re-sample $M$ values from the associated Markov chain. This in turn leads to $M$ evaluations of the likelihood function, requiring to numerically solve $M$ initial value problems. While the SMC algorithm also needs to perform $M$ evaluations of the likelihood function to update the particles, since each particle is independent these evaluations do not need to be made sequentially. This allows to parallelize the evaluation of the likelihood function and potentially strongly reduce the cost of updating an existing estimate as shown by [LYG$^+$10].

# 5 Conclusion

We have presented inverse problems and filtering problems for cases where more data becomes available over time. We have shown how the Bayesian framework provides a well-posed solution to the inverse problem and how Bayesian filtering gives a natural and sound formalism to incorporate new data to the existing knowledge about the unknown parameters. We have then shown how common numerical estimation algorithms fail at efficiently approximating the sequence of filtering posterior distributions and have constructed the SMC algorithm to provide an estimator that is not only asymptotically correct, but also solves many practical difficulties encountered when running the algorithm with finitely many particles available for approximation.

However, many areas could be explored to extend this work. Firstly, we have used a very simple model of the pendulum. We could define a new model to take into account air friction and treat the friction coefficient of the pendulum's mass as an additional parameter of the model. Since the pendulum was let go by a human operator in the experiments, there is also uncertainty in the initial angle of the pendulum which could be taken into account in the model. These additional modeling choice could not only improve the quality of our estimate, but would also be of interest to test the SMC algorithm on higher-dimensional models with latent variables. Secondly, we did not elaborate about the choice on Markov kernels for the MCMC, SIS and SMC algorithms. While Metropolis-Hastings kernels are a very common choice, contructing such Markov chains has been a very active research area from which it could be possible to use more recent results. Also, the use of Markov kernels was only justified by empirical observations, and we have not given any theoretical proof of the impact of these kernels on the convergence properties of the algorithms.

# List of Figures

# 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.

[APSAS17] Sergios Agapiou, Omiros Papaspiliopoulos, Daniel Sanz-Alonso, and Andrew M Stuart. Importance sampling: Intrinsic dimension and computational cost. *Statistical Science*, 32(3):405–431, 2017.

[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] Richard M Dudley. *Conditional Expectations and Martingales*. Cambridge Studies in Advanced Mathematics. Cambridge University Press, 2nd 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.

[LYG+10] Anthony Lee, Christopher Yau, Michael B Giles, Arnaud Doucet, and Christopher C Holmes. On the utility of graphics cards to perform massively parallel simulation of advanced Monte Carlo methods. *Journal of computational and graphical statistics*, 19(4):769–789, 2010.

[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.

[MS18] Joseph Marion and Scott C Schmidler. Finite Sample Complexity of Sequential Monte Carlo Estimators. *arXiv preprint arXiv:1803.09365*, 2018.

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

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