Sequential Monte Carlo for Bayesian inverse problems

1 Introduction

Understanding physical phenomenon is often done through mathematical models, which parameters allow to predict and understand the behaviour of the system. Estimating these parameters from a set of measurements is called the *Inverse Problem*. Solving the Inverse Problem in a sound manner has led to the development of different mathematical frameworks.

Bayesian inversion is one of those approaches, which reformulates the inverse problem as a *quest for information* [5]. In mathematics, a common way of modelling information and uncertainty is through probabilities, in which randomness represents the degree of uncertainty. Bayesian inversion thus models quantities of the system as random variables, and defines their relations using a so-called forward model.

By formulating the inverse problem in this probabilistic framework, its solution is no longer a single value from the parameter space, but rather a probability measure over this state, called the *posterior measure*. Finding this solution can then be shown to be well-posed with respect to the set of observations.

The scope of the Bachelor thesis will be to provide a detailed presentation of the Bayesian inverse problem, with a focus on selected numerical solutions, in particular the *Sequential Monte Carlo* algorithm. The next sections will review Baysian inversion and provide an overview of the algorithms that will be studied in the Bachelor thesis.

2 Bayesian Inversion

In the study of engineering and physical systems, the behaviour of the system is often modelled by an *observation operator* $\mathcal{G} : X \to Y$, mapping values from the *parameter space* X to the *data space* Y. Given a set of observations $y \in Y$, solving the *inverse problem* is finding a solution $x \in X$ to the equation

$$y = \mathcal{G}(x). \tag{1}$$

However, even without even considering observational noise, finding x is usually *ill-posed* in the sense of Hadamard [4] with respect to y: there may be

no solution to the problem or the solution to the problem may not be unique or may depend sensitively on y. It is for this reason that we consider the Bayesian approach, which as shown in [6] doesn't suffer from the presented ill-posedness.

In the Bayesian formulation of the inverse problem and its *quest for information*, we replace $x \in X$ by a square-integrable X-values random variable χ distributed according to μ_0, called the *prior measure*. The prior measure is a known measure encoding any existing information about the solution of the problem. We then use the observation y to *update* our knowledge about χ, leading to the solution of the Bayesian inverse problem, the *posterior measure* μ^y, given by

$$\mu^y \stackrel{\text{def}}{=} \mathbb{P}(\cdot \mid \mathcal{G}(\chi) + \eta = y) \tag{2}$$

where η is a known Y-valued random variable with a zero mean and density ρ, refered to as the *observational noise*.

In the case of finite dimensional X and Y, when considering the p.d.f.s π_0 and π^y of μ_0, respectively μ^y, we obtain by Bayes' formula

$$\pi^y(x) = \frac{\Phi(x)\pi_0(x)}{\int_X \Phi(x)\pi_0(x)dx} \tag{3}$$

in which Φ is called the *data likelihood*, given by

$$\Phi(x) \stackrel{\text{def}}{=} \rho(y - \mathcal{G}(x)). \tag{4}$$

Under sufficient conditions on Φ and μ_0, it can be shown [6] that μ^y is Lipschitz continuous with respect to y, implying that the Bayesian inverse problem is well-posed. It is of interest to note, that the conditions required for the proof of Lipschitz continuity happen to be fulfilled by Gaussian prior and observational noise. Since Gaussian distributions are common in probabilistic modelling, this theoretical result validates the use of Bayesian methods for a large set of applications.

Despite having theoretically validated the Bayesian approach, it is typically not possible to analytically compute μ^y as defined in (3). This limitation requires us to study algorithms enabling practitioners to answer quantitative questions about our solution, the most common one being computing integrals with respect to the posterior measure. The Monte Carlo method is a common approach to approximating high-dimensional integrals, in which given a sequence of points $\{x_n\}_{n=1}^N$ distributed according to μ^y, we get the approximate

$$\mu^y \approx \frac{1}{N}\sum_{n=1}^{N} \delta_{x_n}. \tag{5}$$

Since sampling directly from μ^y is often not possible, we have translated our approximation problem into a sampling problem, for which we now present two algorithms.

3 Markov Chain Monte Carlo

The idea behind Markov Chain Monte Carlo (MCMC) is to construct a Markov chain distributed according to a distribution of interest μ. If the chain is constructed to be ergodic, it can be shown that the empirical distribution of the chain converges weakly to μ. Since this does not describe how to construct the Markov chain, many different algorithms where developed to construct and analyse such Markov chains. Here, we focus on the Metropolis-Hastings method.

The Metropolis-Hastings method is a widely used method for constructing Markov chains for which the target distribution has a density only known up to a constant, i.e. for densities π of the form

$$\pi(x) = \frac{1}{Z}\Phi(x), \tag{6}$$

where Z is an untrackable *normalization constant*, such as the high-dimensional integral in (3), and Φ is a known non-negative function.

The Metropolis-Hastings algorithm, presented in Algorithm 1, is composed of two building blocks: a Markov kernel $K(x, \cdot)$, with density function $k(x_n, x_{n+1})$, from which one can sample *proposals* for the next state of the Markov chain from the current state x_n of the chain, and an *acceptance-rejection rule* based on μ used to "thin down" the chain to the most relevant proposed. The acceptance rule relies on an *acceptance rate*, defining the probability of accepting a given proposal, it is usually defined as following

$$\alpha(x, y) \overset{\text{def}}{=} \frac{k(x, y)\pi(y)}{k(y, x)\pi(x)} = \frac{k(x, y)\frac{1}{Z}\Phi(y)}{k(y, x)\frac{1}{Z}\Phi(x)} = \frac{k(x, y)\Phi(y)}{k(y, x)\Phi(x)}. \tag{7}$$

This definition can be shown to have the desired properties regarding conservation of the distriubution μ, but also has the advantage to remove the dependency on the constant Z, which in the case of Bayesian inversion is usually very hard to compute.

While MCMC algorithms have a very wide range of applications and often provides a satisfactory solution for sampling from complex posteriors, it also suffers from very practical issues. First, it is usually difficult to predict and evaluate whether the Markov chain has reach its stationary regime, thus making it hard to know when to stop the sampling algorithm. Another important issue with MCMC algorithms is that, despite having the property to asymptotically converge to the posterior distribution, it can easily get stuck in local maxima of the posterior, restraining the chain to reach its stationary regime. Finally, when using MCMC for sampling from a posterior distribution, one is required to use the entire dataset, which might be slow, or even only partially available, thus excluding online inference. For these reasons, we will study an alternative sampling class of algorithms: the Sequential Monte Carlo methods.

Algorithm 1: Metropolis-Hastings

Initialize: $x_0 \sim \mu_0$
for $n = 1, 2, \ldots$ **do**
  Propose: $x_{prop} \sim K(x_n, \cdot)$
  Acceptance Probability: $\alpha = \min\{1, \alpha(x_n, x_{prop})\}$
  $u \sim Uniform(0, 1)$
  if $u < \alpha$ **then**
    Accept proposal: $x_{i+1} = x_{prop}$
  else
    Reject proposal: $x_{i+1} = x_i$
  end
end

4 Sequential Monte Carlo

While presenting similarities to the MCMC methods, Sequential Monte Carlo (SMC) solves a different kind of sampling problem in a rather different way.

Considering a sequence of probability measures $\{\mu_t\}_{t=1}^T$ over a common space E, each of them admiting densities $\{\pi_t\}_{t=1}^T$. We are then interested in *sequentially* sampling from this sequence of distributions, that is, first sampling from μ_1, then μ_2 up to μ_T. To do this, SMC algorithms evolves a set of *particles* and *weights* $\{X_n^{(t)}, W_n^{(t)}\}_{n=1}^N (t = 1, \ldots, T; \sum_{n=1}^N W_n^{(t)} = 1)$ whose empirical distribution converges asymptotically in N to μ_t, i.e. for any μ_t-integrable function ϕ

$$\sum_{n=1}^{N} W_n^{(t)} \phi(X_n^{(t)}) \xrightarrow{a.s.} \mathbb{E}_{\mu_t}(\phi). \tag{8}$$

The particles are evolved by alternating steps of *importance sampling* (IS) and *resampling*. While being historically developed to estimate dynamic parameters, Chopin [2] shows that SMC methods can be used to estimate static parameters in Bayesian models by exploring partial posteriors distributions $\mu_t \overset{\text{def}}{=} \mathbb{P}(\cdot|y_1, \ldots, y_t)$.

To understand SMC, we first present the ideas behind IS. As before, we consider distributions μ_t from which the density π_t is only known up to a normalization constant Z_t with $\pi_t(x) = \Phi_t(x)/Z_t$. We then consider auxiliary distributions $\eta_t(dx)$ with densities $\eta_t(x)$ called *importance distributions*, and use the equality

$$\mathbb{E}_{\mu_t}(\phi) = \mathbb{E}_{\eta_t}(\phi \frac{\pi_t}{\eta_t}) \tag{9}$$

by further using the notation $w_t(x) = \frac{\pi_t(x)}{\eta_t(x)}$, and assuming that we can sample from η_t, we can use the Monte Carlo estimate

$$\mu_t^N = \sum_{n=1}^{N} w_t(X_n^{(t)})\delta_{X_n^{(t)}}. \tag{10}$$

In order for (10) to deliver a good approximation, one should select η_t to be close to μ_t. However, this can be a difficult task when μ_t is a non-standard high-dimensional distribution. Nonetheless, it is still possible to use ideas of IS to generate good approximations of μ_t. If we assume that every two consecutive distributions μ_t and μ_{t+1} are similar to each other, a particle approximation of μ_t could be updated to approximate μ_{t+1}. Thus, if we can construct the sequence of distributions to start with a distribution easy to approximate through IS, alternating IS steps and perturbations steps could deliver a sequence of approximations of the distributions up to μ_T. This approach is called the Sequential Importance Sampling (SIS) method.

Formally, given a sequence of N particles $\{X_n^{(t)}\}$ at time t and a Markov kernel $K_{t+1}(x, \cdot)$ with density $k_{t+1}(x_n, x_{n+1})$, we can marginalize π_t to define an importance distribution for time $t+1$,

$$\eta_{t+1}(x') = \int \pi_t(x)k_{t+1}(x, x')dx. \tag{11}$$

And then use IS to get the next estimate μ_{t+1}^N. The question of the choice of the kernels K_n remains unanswered. While [3] presents several possibilities, we will focus on having K_n be a MCMC kernel of invariant distribution μ_n, which will happens to simplify some further results.

The SIS method thus provides approximations of the sequence of distributions of interest. However, it requires computing recursive intergrals, leading to integrals of very high-dimensions, which can have a prohibitive cost that will be described further in the Bachelor thesis. Instead, we present now the Sequential Monte Carlo (SMC) methods, allowing to compute similar approximations than the SIS method while preserving an $O(N)$ computing cost.

While being the main topic of the Bachelor thesis, the main ideas of the SMC algorithm are similar to those of the SIS algorithm. We thus chose here to limit the presentation of SMC to the differences with SIS and the algorithm. Of course, the Bachelor thesis will contain a more detailed description of the algorithm's construction and will attempt to prove convergence properties for the inference of static parameters.

In order to solve the problems presented in the previous paragraph, one studies a growing state space of particle paths rather than single particles. This is done by introducing backward Markov kernels $L_{t-1}(x, \cdot)$ with densities $l_{t-1}(x_t, x_{t-1})$, and studying the distribution $\tilde{\mu}_t$ of density proportional (up to normalization) to $\tilde{\Phi}_t(x_{1:t}) = \Phi_t(x_t)\prod_{i=1}^{t-1} l_i(x_i, x_{i+1})$. We then use a SIS method, in which a sequence of Markov kernels K_t are used to extend the path of each particle, to produce a sequence of approximations of the paths distributions $\{\tilde{\mu}_t\}$. The Bachelor thesis will show that, by using MCMC kernels, the following expression gives a good post-normalization approximation of the importance weights

$$w_t(x_{1:t}) \approx w_{t-1}(x_{1:t-1}) \frac{\Phi_t(x_t)}{\Phi_{t-1}(x_{t-1})}. \tag{12}$$

Another improvement brought by the SMC algorithm is to add a step of resampling. This step solves the problem of *particle degeneracy*, a phenomenon in which only a few particles have a significant weight leading to a reduced *effective sample size* (ESS). To avoid this, the ESS is monitored and a resampling is performed using the current approximation of μ_t, giving particles with low weight a high probability of not being resampled.

We conclude this section by noting that the SMC method with MCMC updates as presented in Algorithm 2, not only reduces the complexity of the approximation of μ^y to $O(N)$, but is also trivially parallelizable.

Algorithm 2: Sequential Monte Carlo

(1) initializing
for $n = 1, \ldots, N$ **do**
  $X_0^{(n)} \sim \mu_0$
  $W_0^{(n)} = w_0(X_0^{(n)})$
end
Normalize weights
(2) updating
for $t = 1, \ldots, T$ **do**
  (2.1) resampling
  if *ESS < threshold* **then**
    Resample $\{X_t^{(n)}\}$
    $W_t^{(n)} = \frac{1}{N}$
  end
  (2.2) sampling
  for $n = 1, \ldots, N$ **do**
    $X_t^{(n)} \sim K(X_{t-1}^{(n)}, \cdot)$
    $W_t^{(n)} = w_t(X_{1:t}^{(n)})$, as given in (12)
  end
  Normalize weights
end

5 Conclusion

The goal of this document was to give an overview of the subject and possible construction of the Bachelor thesis. Only some selected topics were presented in this proposal, and the Bachelor thesis will put more focus on several points such as the formal derivation of the SMC algorithm, as well as convergence analysis

of the algorithm. The Bachelor thesis will also walk through solving an inference problem to illustrate and compare the different approximation methods presented above, in the style of [1].

References

[1] 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.

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

[3] 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.

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

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

[6] A. M. Stuart. Inverse problems: A bayesian perspective. *Acta Numerica*, 19:451559, 2010.