International Journal of Non-Linear Mechanics 94 (2017) 72–83

Contents lists available at ScienceDirect

International Journal of Non-Linear Mechanics

journal homepage: www.elsevier.com/locate/nlm

# Bayesian System Identification using auxiliary stochastic dynamical systems

Thomas A. Catanach, James L. Beck

*Computing and Mathematical Sciences, California Institute of Technology, United States*

ARTICLE INFO

*Keywords:*
System identification
Bayesian methods
Markov Chain Monte Carlo
Stochastic differential equations
Masing hysteresis model

ABSTRACT

Bayesian approaches to statistical inference and system identification became practical with the development of effective sampling methods like Markov Chain Monte Carlo (MCMC). However, because the size and complexity of inference problems has dramatically increased, improved MCMC methods are required. Dynamical systems based samplers are an effective extension of traditional MCMC methods. These samplers treat the posterior probability distribution as the potential energy function of a dynamical system, enabling them to better exploit the structure of the inference problem. We present an algorithm, Second-Order Langevin MCMC (SOL-MC), a stochastic dynamical system based MCMC algorithm, which uses the damped second-order Langevin stochastic differential equation (SDE) to sample a posterior distribution. We design the SDE such that the desired posterior probability distribution is its stationary distribution. Since this method is based upon an underlying dynamical system, we can utilize existing work to develop, implement, and optimize the sampler's performance. As such, we can choose parameters which speed up the convergence to the stationary distribution and reduce temporal state and energy correlations in the samples. We then apply this sampler to a system identification problem for a non-linear hysteretic structure model to investigate this method under globally identifiable and unidentifiable conditions.

## 1. Introduction

Bayesian methods for identification and estimation are critical to the robust understanding of a system because they allow us to quantify all of our uncertainty about the system using a probability distribution and to update this distribution with new information [1–9]. By taking the Bayesian approach, we are able to effectively capture our prior knowledge about a model and rigorously assess the plausibility of candidate model classes based on system data. Finally, we can then make robust probabilistic predictions that incorporate all uncertainties, allowing for better decision making and design. This robust approach is particularly relevant for system identification, where the inverse problems are often ill-posed and many candidate models exist to describe the behavior of a system.

We can broadly classify the posterior probability distributions that result from solving the inference problem into three types: globally identifiable, locally identifiable, and unidentifiable [1,10]. Globally identifiable probability distributions have a single pronounced peak around a unique maximum. Locally identifiable distributions have several separated peaks, each of approximately the same significance. Unidentifiable models do not have peaks, but instead have a manifold in the parameter space on which all values are approximately equally plausible based on the data and the selected prior information. When the problem results in a locally identifiable or unidentifiable distribution, Bayesian methods are essential since they can fully capture this complex distribution in a way optimization based system identification methods cannot. However, these types of problems still produce a significant challenge to computational Bayesian methods since it is often difficult to find and explore all the peaks or the manifold of plausible solutions.

Prior to the development of scientific computing, Bayesian method were restricted to inference problems where the posterior distribution could be expressed as a simple analytical distribution. While approximate methods exist, they are often have difficulty in handling locally identifiable or unidentifiable problems [4,11], where Bayesian methods are most needed. As a result, sampling methods are commonly used. The most common family of sampling methods is Markov Chain Monte Carlo (MCMC) [12], which creates a Markov chain defined by a transition rule, or kernel, whose stationary distribution is the desired posterior. In order to make estimates accurately, the samples must discretely capture the posterior distribution in a probabilistically appropriate way. This makes MCMC computational intensive, as often thousands to millions of model evaluations are needed to fully populate the high probability content of the posterior. While, by the central limit theorem, the estimate quality for the mean of a finite-variance stochastic variable scales independently of the dimension given independent samples, MCMC methods produce correlated samples, which can introduce poor high dimensional scaling. Many high dimensional problems where it is difficult to produce an efficient proposal distribution experience a “curse of dimensionality” because the sample correlation becomes very high. Thus, solving inference problems using Bayesian methods is often prohibitively expensive because sampling

http://dx.doi.org/10.1016/j.ijnonlinmec.2017.03.012
Received 5 August 2016; Received in revised form 25 February 2017; Accepted 11 March 2017
Available online 21 March 2017

high dimensional distributions efficiently is challenging.

One of the most successful methods for sampling high dimensional distributions is Hamiltonian Monte Carlo (HMC) [13,14]. HMC uses an auxiliary Hamiltonian dynamical system to propose samples far from the current sample in the parameter space but with similar probability. The position coordinates of this auxiliary system correspond to the inferred model parameters. The proposal trajectory conserves the Hamiltonian, which is related to the posterior probability. As a result, even though the candidate samples are far from the current sample, they will have high acceptance probability, thus reducing sample correlation. This is achieved by constructing a Hamiltonian dynamical system whose potential energy function is the negative log posterior probability density function (PDF), while the kinetic energy function is quadratic in the velocity coordinates of the auxiliary system, giving the corresponding momentum vector a Gaussian distribution. The marginal position distribution of this system is the desired posterior. An application of HMC to Bayesian updating of high-dimensional dynamic systems is given in [15].

This auxiliary dynamical systems approach can be extended to stochastic dynamical systems, described by a stochastic differential equation (SDE) whose stationary distribution corresponds to the posterior of the Bayesian inference problem. These SDE approaches can not only be used in a standard MCMC framework, but can also be used to approximate the distribution without Metropolis steps [16–18]. SDEs are an active area of research, so there is a great opportunity to leverage these results to study the properties of these algorithms, such as work in infinite spaces [19,20]. The damped second-order Langevin equation has recently been introduced as a sampler for Bayesian inference [21,17,19]. This SDE is an effective choice because it combines Hamiltonian dynamics with an Ornstein–Uhlenbeck process, which enables the state to both follow likely trajectories and to diffuse.

We introduce a new sampling method, Second-Order Langevin Monte Carlo (SOL-MC), which extends previous work on SDE samplers, and increases its applicability to system identification problems. This sampler combines a non-metropolized SDE optimized for convergence to the posterior manifold with a metropolized SDE which can effectively sample the posterior while reducing sample correlation. We then apply results from dynamical system and control theory to tune the parameters of the SDE to optimize the sampler's performance by balancing the influence of the “exploring” diffusion and “exploiting” Hamiltonian dynamics for Gaussian like distributions. Further, we are able to utilize new computational tools like automatic differentiation to make simulating this SDE tractable for system identification problems.

This paper is organized as follows. Section 2 discusses the Bayesian System Identification problem and standard computational methods. Section 3 introduces the SOL-MC algorithm and implementation. Section 4 discusses tuning SOL-MC to optimize performance. Section 5 presents an example system identification problem to investigate SOL-MC under different conditions. Finally, Section 6 presents concluding remarks.

## 2. Bayesian System Identification

### 2.1. The Bayesian framework

The Bayesian framework is a rigorous probabilistic method for representing our own uncertainty using probability distributions. This philosophy is rooted in probability as a logic [4,22–24]. Within this framework, probability distributions are used to quantify uncertainty due to insufficient information, regardless of whether that information is believed to exist but is currently not available (epistemic uncertainty), or it is believed not to exist because of postulated inherent randomness (aleatory uncertainty). This makes the Bayesian framework the appropriate framework for posing system identification problems. Therefore, we view system identification as updating a probability distribution that represents our beliefs about models of a system based on new information from system response data.

### 2.2. Problem formulation

We consider a continuous-time deterministic system model with state $x \in \mathbb{R}^{N_x}$, unknown fixed parameters $\theta \in \mathbb{R}^{N_p}$, and known inputs $u \in \mathbb{R}^{N_i}$ described by the differential equation:

$$\dot{x}(t) = f(x(t), u(t), \theta) \tag{1}$$

The initial state $x(0)$ can either be considered an unknown parameter and included in $\theta$ or a known parameter. The actual output of this system, $\hat{y}(t)$, for input $\hat{u}(t)$, is measured at discrete points $t_i$ for $i = 1 \ldots N$. Given the state $x(t_i)$, input $u(t_i)$, and parameters $\theta$, the predicted measured outputs are assumed to be subject to independent combined stochastic measurement and prediction errors $\nu(t, \theta)$ modeled by a parameterized distribution with possibly unknown parameters that are included in the model parameter vector $\theta$. Conditioned on $\nu(t_i, \theta)$, the output $y(t_i)$ is modeled as:

$$y(t_i) = h(x(t_i), u(t_i), \nu(t_i, \theta), \theta) \tag{2}$$

One formulation of Bayesian System Identification is then [4]: Given the input–output observation data $\mathscr{D} = \{\hat{u}(t_i), \hat{y}(t_i), i = 1 \ldots N\}$, and a system model class $\mathscr{M}$ consisting of (a) the prescribed functions $\{f, h\}$ in (1), (2), respectively, (b) a probability distribution for $\nu$, and (c) a prior distribution, $p(\theta|\mathscr{M})$, representing our initial beliefs about the relative plausibility of the possible values of the model parameter vector $\theta$, find the posterior distribution $p(\theta|\mathscr{D}, \mathscr{M})$, representing our updated beliefs. For this we employ Bayes' Theorem:

$$p(\theta|\mathscr{D}, \mathscr{M}) = \frac{p(\mathscr{D}|\theta, \mathscr{M}) p(\theta|\mathscr{M})}{p(\mathscr{D}|\mathscr{M})} \tag{3}$$

where the likelihood function, $p(\mathscr{D}|\theta, \mathscr{M})$ is given by substituting the data $\mathscr{D}$ into the model prediction of the measured system output, so it gives a measure of the plausibility of the data according to the model in $\mathscr{M}$ specified by $\theta$.

The normalizing factor in (3), $p(\mathscr{D}|\mathscr{M})$, is the evidence for $\mathscr{M}$:

$$p(\mathscr{D}|\mathscr{M}) = \int p(\mathscr{D}|\theta, \mathscr{M})\, p(\theta|\mathscr{M})\, d\theta \tag{4}$$

This integral over the whole domain of $\theta$ is usually analytically intractable and computationally challenging to evaluate numerically. As a result, typically only the ratio of posterior density for sample points can be computed. Markov Chain Monte Carlo uses these ratios to generate samples from the posterior distribution (3). These sample points can then be used to estimate the expectation of a function $g(\theta)$ with respect to the posterior distribution. Analogous to the central limit theorem for independent, identically distributed stochastic variables, the central limit theorem for Markov chains [25] implies that:

$$\mathbb{E}[g(\theta)|\mathscr{D}, \mathscr{M}] = \int g(\theta) p(\theta|\mathscr{D}, \mathscr{M}) d\theta \approx \frac{1}{N} \sum_{i=1}^{N} g(\theta_i) \tag{5}$$

where $\theta_i$ is a sample drawn from the posterior.

### 2.3. Computational methods

The basis for many MCMC methods is the Metropolis–Hastings algorithm, which produces a Markov chain with a desired stationary distribution, $\pi(\theta)$, by designing a transition kernel, $K(\theta'|\theta)$, such that the Markov chain is ergodic and reversible [25,26]. Reversibility is a sufficient condition for the existence of a stationary state defined as part of the detailed-balance condition:

$$\pi(\theta) K(\theta'|\theta) = \pi(\theta') K(\theta|\theta') \tag{6}$$

This means that we can choose any transition kernel $K(\theta'|\theta)$ and maintain the stationary distribution $\pi(\theta)$, as long as the condition (6) holds.