Bayesian System Identification via Markov Chain Monte Carlo Techniques

Brett Ninness [a,1] Soren Henriksen [a],

[a] *School of Electrical Engineering and Computer Science, The University of Newcastle, Australia*

Abstract

The work here explores new numerical methods for supporting a Bayesian approach to parameter estimation of dynamic systems. This is primarily motivated by the goal of providing accurate quantification of estimation error that is valid for arbitrary, and hence even very short length data records. The main innovation is the employment of the Metropolis–Hastings algorithm to construct an ergodic Markov chain with invariant density equal to the required posterior density. Monte–Carlo analysis of samples from this chain then provide a means for efficiently and accurately computing posteriors for model parameters and arbitrary functions of them.

Key words: Parameter Estimation; System Identification; Bayesian Methods; Maximum Likelihood.

1 Introduction

A dominant force in both the practice and underlying understanding of modern methods for system identification of dynamic systems has been work within the context of the maximum likelihood and prediction error frameworks [4,23,44]. In particular, the software developed as part of this latter effort [24] has become an industry standard.

A key aspect of these approaches is that any quantification of the accuracy of the associated system estimates relies on employing asymptotic in data length expressions as if they applied for finite data lengths. For example, the Gaussian distribution commonly achieved by estimates in the infinite data limit, is usually assumed to hold for whatever finite data length is available.

While these techniques enjoy widespread acceptance, there has arisen a body of work that has sought to derive methods and supporting theory which apply for arbitrarily short data records. To give some examples, the so-called 'bounded error', or 'set estimation' techniques [31,27,46] were developed to provide non-asymptotic error bounds in situations where measurement corruptions were of constrained magnitude.

More recent work has examined how bounds applicable to finite data situations may be computed for prediction error methods [47,48], correlation methods [7] and least squares techniques with model structure linear in the parameter vector [9], and has also turned to concepts from machine learning theory [6,5,45], which have originated largely within the Computer Science community [28].

This paper is directed at the same issue of dynamic system identification in a finite data record setting, but takes a different approach to the problem. In particular, the perspective here is that especially for very short data lengths, it is sensible to take a Bayesian approach to quantifying the manner in which prior knowledge and data-based information are combined to yield posterior information about system properties.

For example, if one accepts the Bayesian interpretation of probability, then the posterior density of system parameters given observed data provides a complete and concise description of all knowledge that can be extracted from the observed data about those parameters [18,22]. Furthermore, except in special cases, alternatives approaches such as maximum likelihood and prediction error methods deliver estimates whose associated probability densities cannot be formulated except via approximate methods involving first order estimates whose validity depends on data records of substantial length being available.

[⋆] This work was supported by the Australian Research Council.

Email addresses: brett.ninness@newcastle.edu.au (Brett Ninness), soren.henriksen@newcastle.edu.au (Soren Henriksen).

[1] Corresponding author.

Due to these and other attractive theoretical and philosophical aspects, interest in a Bayesian approach to parameter estimation has a very long history, stretching at least as far back as [34]. However, it is not commonly used (save for the linearly parametrized Gaussian case where Kalman filter-based Bayesian estimation is routinely employed), mainly due to the computational difficulties of computing posterior densities, marginals of them, and associated functions such as posterior means. See [20] for a recent discussion of these issues.

However, over the last decade, there has been great progress made within the statistics community in overcoming this computational burden and making Bayesian analysis tractable for a wide range of complicated models such as those inherent to demographic and population studies, image processing, and drug response modeling [14] to name but a few cases. The central tool has been to construct a Markov chain (i.e. a vector random number generator) with invariant, and hence stationary density equal to the desired posterior density of interest.

Monte–Carlo analysis of realisations from this chain then provides a means for efficiently computing required posteriors of parameters, and also very general functions of them. The computational load is therefore lessened by seeking only an approximate answer, with a smooth tradeoff between accuracy and computation time being provided by the choice of Markov chain simulation duration.

Despite the widespread success and interest in these so called 'Markov chain Monte–Carlo' (MCMC) methods in the statistics literature, there appears to have been no study of their potential for dynamic system parameter estimation problems of interest to the control and signal processing communities, aside from a preliminary study by the current authors [30]. Indeed, it appears they were first suggested for control relevant applications at all only quite recently in [40] where state (but not parameter) estimation was considered.

There have, however, been previous control relevant contributions using other types of Monte–Carlo analysis. For example, the recent works [42,3,10] use a 'bootstrap' technique, also having its roots on the statistics literature[12] to generate random samples of parameter estimates for subsequent Monte–Carlo analysis. Unlike the methods considered here, each bootstrap iteration requires re-calculation of parameter estimates, and the goal is to compute densities of prediction error estimates (or in the case of [3], subspace estimates) as opposed to the posterior densities studied here. Furthermore underpinning convergence analysis in these applications is not straightforward, and hence has not yet been studied.

As another example of Monte–Carlo based methods, the application of 'particle filters' has recently been studied in several areas of signal processing and state estimation [43,8], for which purpose it appears to have been first introduced in the control literature [15]. However, particle filtering, or more properly 'sequential importance resampling' (SIR) methods are specifically designed to approximate the integrations involved with the Chapman–Kolmogorov equations for measurement and time update of optimal state estimates.

As such, they share little with the MCMC methods studied here, which are designed to compute almost arbitrary posterior densities, with no underlying constraint of a Chapman–Kolmogorov relationship. However, there appears to be some confusion on this point due to the fact that MCMC methods are sometimes employed as a component (the so-called 'resampling step') of some SIR algorithms. Nevertheless, in the specific case of model structures which are linear in the parameters, so that the problem of estimating them can be cast as a filtering problem, SIR techniques have been studied for parameter estimation applications [20,38].

With this as background, this paper studies the application of MCMC methods for Bayesian estimation of parameters and, perhaps more importantly, *arbitrary functions of them.* The latter can be the posterior density of system frequency response, achieved closed loop gain/phase margin, or achieved sensitivity function infinity norm for a given candidate controller, as just three among an essentially arbitrary number of examples.

Associated with this, the paper is designed to give a self contained introduction to the Metropolis–Hastings class of MCMC algorithms and their theoretical foundations in terms of convergence analysis. Since this depends crucially on techniques and results of Markov chain theory on non-countable spaces, which many readers may not be familiar with, a brief overview of this material is provided in an appendix.

2 Problem Setting

This paper addresses parametric modeling of an observed data record $Y \triangleq \{y_1, \cdots, y_N\}$ according to a stochastic model

$$Y \sim p(Y \mid \theta) \tag{1}$$

where $\theta \in \mathcal{X} \subseteq \mathbf{R}^n$ is a vector of model parameters, and $p(Y \mid \theta)$, which is completely described by θ, is the joint probability density function [2] for the elements of Y.

Commonly, such a model is employed to form an estimate $\widehat{\theta}_N$ of the true parameters θ_* via a maximum-

[2] Importantly, the form of the density $p(\theta \mid Y)$ may depend on further factors, such as the model structure that θ parametrizes, or an observed exogenous input sequence. In the interests of readability, and as is common practice [2, Equations (3.1), (3.2)], [23, Page 216] this will be taken as understood and not explicitly acknowledged via notational embellishment.

likelihood (ML) approach

$$\widehat{\theta}_N \triangleq \arg\max_\theta p(Y \mid \theta) = \arg\max_\theta \log p(Y \mid \theta). \qquad (2)$$

In many applications of interest, via a Kalman filter or some other method of solving the associated Chapman–Kolmogorov equations, the model (1) permits a decomposition

$$y_t = \widehat{y}_{t|t-1}(\theta) + \varepsilon_t. \qquad (3)$$

Here $\widehat{y}_{t|t-1}(\theta)$, is a one step ahead predictor (Wiener filter) of y_t based on observations strictly prior to time t, and ε_t is a zero mean i.i.d. innovations process that is independent of $\widehat{y}_{t|t-1}(\theta)$. To provide a concrete example, the standard reference [23] considers the model structure

$$y_t = G(q,\theta)u_t + H(q,\theta)\varepsilon_t \qquad (4)$$

as [23, page 45] "*the basic description of a linear system subject to additive random disturbances*" where $\{u_t\}$ is an observed exogenous input, and $G(q,\theta)$, $H(q,\theta)$ are transfer functions, rational in the forward shift operator q, and completely described by the elements of the parameter vector θ. In this case, with $H(q,\theta)$ being constrained to be monic, the predictor $\widehat{y}_{t|t-1}(\theta)$ is well known to be given as

$$\widehat{y}_{t|t-1}(\theta) = H^{-1}(q,\theta)G(q,\theta)u_t + [1 - H^{-1}(q,\theta)]y_t \quad (5)$$

under the assumption that both $H^{-1}(q,\theta)G(q,\theta)$ and $H^{-1}(q,\theta)$ are stable. Returning to the more general case, via Bayes' rule, equation (3) implies that $p(Y \mid \theta)$ may be computed according to

$$p(Y \mid \theta) = p(y_0 \mid \theta)\prod_{t=1}^{N} p_\varepsilon(y_t - \widehat{y}_{t|t-1}(\theta)) \qquad (6)$$

where $p_\varepsilon(\cdot)$ is the probability density function for the stationary innovations $\{\varepsilon_t\}$ and $p(y_0 \mid \theta)$ encapsulates assumptions on initial conditions.

In this case, for a very broad variety of model structures, the widely used Maximum–Likelihood approach (2) obeys the asymptotic distributional convergence

$$\sqrt{N}(\widehat{\theta}_N - \theta_*) \xrightarrow{\mathcal{D}} \mathcal{N}(0,P) \qquad \text{as } N \to \infty \qquad (7)$$

where $\mathcal{N}(0,P)$ denotes a zero mean multivariable Gaussian density with covariance matrix P. The latter is defined by ($\mathbf{E}\{\cdot\}$ denotes expectation)

$$P^{-1} \triangleq \lim_{N\to\infty} -\frac{\mathrm{d}^2}{\mathrm{d}\theta\mathrm{d}\theta^T}\frac{1}{N}\mathbf{E}\{\log p(Y \mid \theta)\}\bigg|_{\theta=\theta_*}. \qquad (8)$$

Typically, in the absence of an alternative, quantification of the error $\widehat{\theta}_N - \theta_*$ is achieved by assuming the convergence (7) has effectively occurred for the finite data length N available. For example, based on this, it is commonly approximated that (χ^2_n denotes an n-degree of freedom chi-squared distribution)

$$N(\widehat{\theta}_N - \theta_*)^T P^{-1}(\widehat{\theta}_N - \theta_*)^T \sim \chi^2_n \qquad (9)$$

as a means of generating ellipsoidal confidence regions for $\widehat{\theta}_N$. Furthermore, when necessary to quantify the estimation error $\widehat{\theta}^i_N - \theta^i_*$ in the i'th element of the parameter estimate, a further common technique is to assume from (7) that ($[P]_{i,j}$ is the i,j'th element of P)

$$\widehat{\theta}^i_N - \theta^i_* \sim \mathcal{N}\left(0, \frac{1}{N}[P]_{i,i}\right) \qquad (10)$$

and hence (for example) that a 'two sigma' bound $2N^{-1}[P]_{i,i}$ is an effective 95% confidence interval for $\widehat{\theta}^i_N - \theta^i_*$. Note that using (10) involves a further approximation of assuming independence between the elements $\widehat{\theta}^i_N$ to arrive at the marginal (10).

Often, it is of more interest to quantify the error in a function $f(\widehat{\theta}_N)$ of the parameter estimate. For example, the frequency response of a model parametrized by $\widehat{\theta}_N$. To address this requirement, a further approximating step is usually introduced whereby the functional relationship is reduced to a first order expansion, and Gauss' approximation formula [21] is then used to estimate that ($\cdot^\star$ denotes conjugate transpose)

$$f(\widehat{\theta}_N) - f(\theta_*) \sim \mathcal{N}\left(0, \frac{1}{N}\frac{\mathrm{d}f(\widehat{\theta}_N)}{\mathrm{d}\theta}^\star P^{-1}\frac{\mathrm{d}f(\widehat{\theta}_N)}{\mathrm{d}\theta}\right). \qquad (11)$$

Despite the several approximations in the error quantifications (9)-(11), they have proved effective and accurate in many applications when N is reasonably large and hence $\|\widehat{\theta}_N - \theta_*\|$ is small so that first order approximations to functions of $\widehat{\theta}_N$ are informative.

Equally, there are many applications where the available estimation data is limited. In these cases, it is simultaneously more important to accurately quantify estimation error and hence model quality, but also more difficult to do so as the assumptions underlying (9)-(11) become problematic [13].

3 A Bayesian Approach to Estimation and Error Quantification

In order to address the problem of accurate estimation error quantification for arbitrarily small data lengths, this paper examines a Bayesian approach. This involves using the posterior density $p(\theta \mid Y)$ as a means of quantifying all information that can be extracted from a combination of the observed data and prior information.

In principle, the required posterior may be computed using Bayes' rule applied to (6) to provide

$$p(\theta \mid Y) = \frac{p(Y \mid \theta)p(\theta)}{p(Y)} \qquad (12)$$

where $p(Y \mid \theta)$ is computed via (6), $p(\theta)$ is the a-priori distribution of θ and $p(Y)$ is a normalising factor (constant with respect to θ) given as

$$p(Y) = \int p(Y \mid \theta) p(\theta)\, \mathrm{d}\theta. \tag{13}$$

In this case, a Bayesian maximum a-posterior (MAP) estimate $\widehat{\beta}_N$ may be taken as

$$\widehat{\beta}_N \triangleq \arg\max_{\theta} \log p(\theta \mid Y) \tag{14}$$

$$= \arg\max_{\theta} \left[\log p(Y \mid \theta) + \log p(\theta)\right]. \tag{15}$$

When there is little prior information on θ, and hence $p(\theta)$ is diffuse (constant) over a wide range, then the point estimates $\widehat{\beta}_N$ in (15) and $\widehat{\theta}_N$ in (2) will co-incide. In other cases, provided $p(\theta)$ is sufficiently regular (for example, smooth), then existing algorithms such as gradient based search used to compute the ML estimate (2) can equally well be applied to compute the MAP estimate (14). These principles and ideas, as mentioned in the introduction, are old and well known.

However, the further steps of using the posterior $p(\theta \mid Y)$ as a means of computing estimate accuracy, or using the alternative Bayesian point estimate

$$\widehat{\beta}_N = \mathbf{E}\left\{\theta \mid Y\right\} \tag{16}$$

are not routinely employed due to the associated high computational burdens.

For example, if the marginal density of only a particular i'th parameter element θ^i is required, then this requires numerical computation of the multidimensional integral

$$p(\theta^i \mid Y) = \int p(\theta \mid Y) \mathrm{d}\theta^1 \cdots \mathrm{d}\theta^{i-1} \mathrm{d}\theta^{i+1} \cdots \mathrm{d}\theta^n. \tag{17}$$

This is problematic, since if k points on the density curve $p(\theta^i \mid Y)$ are required to represent it, then (17) implies the numerical evaluation of k multidimensional integrals, and this latter dimension could be relatively large: a modest fifth order transfer function model would imply a nine dimensional integral in order to obtain the marginal on just one parameter.

In situations where the residual ε_t in (3) has bounded support on a region Δ

$$\varepsilon_t \notin \Delta \;\Rightarrow\; p_\varepsilon(\varepsilon_t) = 0 \tag{18}$$

then one approach to coping with these difficulties is to seek only the support Θ of the posterior $p(\theta \mid Y)$ which in the case of uniform and diffuse prior $p(\theta)$ is clearly given by (12), (6) as

$$\Theta = \bigcap_{t=1}^{N} \left\{\theta : y_t - \widehat{y}_{t|t-1}(\theta) \in \Delta\right\}. \tag{19}$$

This domain Θ is precisely the 'feasible parameter set' that is studied (albeit via different motivation) in the 'bounded error' estimation literature, and for which (depending on model structure) numerous efficient methods for computation have been developed [32,31,27,46].

While this provides a partial solution in cases satisfying (18), there are further frequent requirements it does not address. For example, suppose as is common, interest is centered not on the parameter estimate itself, but on a function of it such as (for example) system phase margin $\phi_m(K)$ for a given closed loop controller $K(q)$. Then, even ignoring computational load issues, it is not at all clear how one might tractably compute the posterior density

$$p(\phi_m(K) \mid Y). \tag{20}$$

4 A Markov Chain Monte–Carlo Solution

As a solution to the difficulties just profiled of computing posterior densities of parameters and functions of them such as (20), this paper explores a Markov-chain Monte–Carlo (MCMC) technique.

This approach involves numerically computing the required densities by a strategy of first generating a random sequence of realisations $\{\theta_k\}$ with each individual element $\theta_k \in \mathcal{X} \subseteq \mathbf{R}^n$ and with limiting distribution equal to the desired posterior density; viz.

$$\lim_{k\to\infty} p(\theta_k = \theta \mid \theta_0) = p(\theta \mid Y) \qquad \forall \theta_0 \in \mathcal{X} \subseteq \mathbf{R}^n. \tag{21}$$

The simulated realisation $\{\theta_k\}$ is then used as if it were a random sample from $p(\theta \mid Y)$. Provided (as will be shown presently) that the required distributional convergence holds, then via a law of large numbers based argument, this will lead to consistent estimates of various quantities. For example, it allows the numerical computation and consistent estimation of the conditional expectation $\mathbf{E}\left\{f(\theta) \mid Y\right\}$ as

$$\mathbf{E}\left\{f(\theta) \mid Y\right\} = \int_{\mathcal{X}} f(\theta) p(\theta \mid Y) \mathrm{d}\theta \tag{22}$$

$$\approx \frac{1}{M} \sum_{k=1}^{M} f(\theta_k) \triangleq \widehat{f}_M \tag{23}$$

where f is an arbitrary Borel measurable function. It also allows the computation of rather arbitrary posterior densities via sample histograms:

$$p(f(\theta) \in A \mid Y) \approx \frac{1}{M} \sum_{k=1}^{M} I_{f^{-1}(A)}(\theta_k). \tag{24}$$

Here, $I_X(\xi)$ is the indicator function for $\xi \in X$ defined as

$$I_X(\xi) = \begin{cases} 1 \; ; \xi \in X \\ 0 \; ; \text{Otherwise} \end{cases} \tag{25}$$

and A is any f-measurable set.

While, this may seem like a reasonable approach in principle, it may also appear to be practically infeasible due to the implied requirement of a vector random number generator with given joint density $p(\theta \mid Y)$.

Perhaps surprisingly, this difficulty can be overcome quite straightforwardly by recognising that the related task of simply *evaluating* the function $p(\theta \mid Y)$ for a given θ is usually straightforward.

For example, in the common situation where the general model structure (4) is employed, then via (12) and (6), the evaluation of $p(\theta \mid Y)$ for a given θ reduces to computing the prediction error $\varepsilon_t(\theta) = y_t - \widehat{y}_{t|t-1}(\theta)$ using the standard formula (5) and then evaluating the density $p_\varepsilon(\cdot)$ for the innovations ε_t in (4) at this point. Use of this process is illustrated in section 8.

This leads to a process which first simulates an arbitrary Markov chain with convenient transition density $\gamma(\theta_k \mid \theta_{k-1})$, and then appropriately modifying the resulting samples via *function evaluations* $p(\theta_k \mid Y)$ in order to yield vector realisations distributed, upon convergence, according to $p(\theta \mid Y)$. The details of this process are provided by the following algorithm.

Algorithm 4.1 ***(Markov Chain Monte Carlo Sampler)***

(1) Initialise θ_0 at some value such that $p(\theta_0 \mid Y) > 0$ and set $k = 1$;

(2) At iteration k, consider a candidate value ξ_k for θ_k which is drawn from a ***proposal*** *density $\gamma(\xi_k \mid \theta_{k-1})$. That is, find a possible realisation for θ_k as*

$$\xi_k \sim \gamma(\cdot \mid \theta_{k-1}); \tag{26}$$

(3) Compute the acceptance probability

$$\alpha(\xi_k \mid \theta_{k-1}) = \min\left\{1,\ \frac{p(\xi_k \mid Y)}{p(\theta_{k-1} \mid Y)} \cdot \frac{\gamma(\theta_{k-1} \mid \xi_k)}{\gamma(\xi_k \mid \theta_{k-1})}\right\}; \tag{27}$$

(4) Accept the proposed ξ_k and set $\theta_k = \xi_k$ with probability $\alpha(\xi_k \mid \theta_{k-1})$, otherwise leave θ_k unchanged by setting $\theta_k = \theta_{k-1}$;

(5) Increment k and return to step 2. □

Note that step 4 may be simply implemented by drawing a random variable $z \sim U_{[0,1]}(\cdot)$ from a uniform distribution on $[0, 1]$ and setting $\theta_k = \xi_k$ if $z < \alpha(\xi_k \mid \theta_{k-1})$.

Algorithm 4.1 is, in fact, an instance of the so-called 'Metropolis–Hastings' algorithm, which was developed in [25] and generalised in [16] to the form specified above. Although the introduction to this paper mentioned the use of ideas currently popular in the statistics community, this algorithm is in fact also widely used in physics, chemistry and biology, as profiled in [17] where it is listed at first place in a survey of great algorithms of scientific computing.

An important specialisation of this method occurs when the 'proposal density' γ is symmetric in that $\gamma(\xi \mid \theta) = \gamma(\theta \mid \xi)$. In this situation

$$\frac{\gamma(\theta_{k-1} \mid \xi_k)}{\gamma(\xi_k \mid \theta_{k-1})} = 1, \tag{28}$$

and hence the acceptance probability (27) simplifies to

$$\alpha(\xi_k \mid \theta_{k-1}) = \min\left\{1,\ \frac{p(\xi_k \mid Y)}{p(\theta_{k-1} \mid Y)}\right\}. \tag{29}$$

In this case, Algorithm 4.1 is known as the *Metropolis Algorithm*. To see how this symmetric proposal scenario might occur, consider the obvious 'random walk' proposal density implied by

$$\xi_k = \theta_{k-1} + \nu_k \tag{30}$$

with ν_k being a random perturbation. In the special, but common, case in which the mean of ν_k is zero so that the probability density $p_\nu(\cdot)$ governing ν_k is symmetric ($p_\nu(x) = p_\nu(-x)$), then clearly

$$\gamma(\xi \mid \theta) = p_\nu(\xi - \theta) = p_\nu(\theta - \xi) = \gamma(\theta \mid \xi). \tag{31}$$

The Metropolis specialisation (29) most clearly exposes an intuitive explanation of Algorithm 4.1. Namely, realisations θ_k convergent (in a distributional sense) to a target distribution $p(\theta \mid Y)$ are obtained by first drawing a random proposal ξ_k. In the Metropolis case (29), if it so happens that ξ_k is more likely to be a realisation drawn from the density $p(\cdot \mid Y)$ than the previous iteration θ_{k-1}, in that $p(\xi_k \mid Y) > p(\theta_{k-1} \mid Y)$, then ξ_k is definitely used as a new realisation $\theta_k = \xi_k$. On the other hand, if the randomly drawn proposal ξ_k is less likely, then *it is not necessarily thrown away*. Instead, even though it is less likely, it may be retained. Whether or not this happens is randomly decided. More specifically, a proposed ξ_k less likely than θ_{k-1} is retained (in that $\theta_k = \xi_k$) with probability $\alpha(\xi_k \mid \theta_{k-1}) = p(\xi_k \mid Y)/p(\theta_{k-1} \mid Y)$.

For readers wishing a more detailed discussion of the Metropolis–Hastings algorithm complete with its history, tutorial examples, and application to state estimation, the authors recommend the survey [40]. For those wishing to explore the underlying theory and extensions of MCMC methods in more depth, the seminal work [41] and the monograph [35] are recommended.

5 Convergence Analysis

A particularly attractive feature of the MCMC ideas embodied in Algorithm 4.1 is that they implement a Markov chain, for which an extensive theory is available [26,33]. This facilitates a rigorous convergence analysis to establish the validity of the approach.

Balancing this advantage is the fact that the details of this theory and its application are scattered in the literature, and somewhat inaccessible to the non-specialist. What follows here and in Appendix A is designed to redress this by providing a streamlined and self contained convergence analysis of Algorithm 4.1 applied to the posterior $p(\theta \mid Y)$.

To proceed with this, it is fundamental to observe that according to Algorithm 4.1, the mechanism of generating a new sample θ_k is a time-homogeneous Markov chain with transition density $K(\theta_k \mid \theta_{k-1})$ given as the product of the probability $\gamma(\xi \mid \theta)$ of proposing a move ξ, times the probability $\alpha(\xi \mid \theta)$ of accepting it:

$$K(\theta_k = \xi_k \mid \theta_{k-1}) = \alpha(\xi_k \mid \theta_{k-1})\gamma(\xi_k \mid \theta_{k-1}) I_{\mathcal{X}_{\theta_{k-1}}}(\xi_k) + \delta(\xi_k - \theta_{k-1}) r(\theta_{k-1}) \qquad (32)$$

where $\mathcal{X}_{\theta_{k-1}} = \{\xi \in \mathcal{X} : \xi \neq \theta_{k-1}\}$ and

$$r(\theta_{k-1}) = 1 - \int_{\mathcal{X}_{\theta_{k-1}}} \alpha(\xi \mid \theta_{k-1})\gamma(\xi \mid \theta_{k-1})\, \mathrm{d}\xi$$

is the probability of no change in the value of θ_k from one iteration to another. In (32) the delta function is of the Dirac type. Where necessary

$$\mathbf{P}(A \mid \theta) = \int_A K(\xi \mid \theta)\, \mu(\mathrm{d}\xi) \qquad (33)$$

will denote the associated transition distribution with μ being Lebesgue measure, and A being any μ-measurable set. Furthermore, in what follows, for brevity $\mu(\mathrm{d}\xi)$ will be shortened to simply $\mathrm{d}\xi$ (or equivalent).

Now, suppose that θ_{k-1} is drawn randomly according to a probability density function $\pi_{k-1}(\theta)$. Then clearly, the probability density function $\pi_k(\theta)$ for an ensuing element θ_k in this Markov chain is given by the law of total probability as

$$\pi_k(\theta_k) = \int_{\mathcal{X}} K(\theta_k \mid \theta_{k-1})\pi_{k-1}(\theta_{k-1})\, \mathrm{d}\theta_{k-1}. \qquad (34)$$

Therefore, if the realisations $\{\theta_k\}$ generated by Algorithm 4.1 are to converge in a distributional sense to realisations having some constant density $\pi(\theta)$, then that density must satisfy

$$\pi(\theta) = \int_{\mathcal{X}} K(\theta \mid \xi)\pi(\xi)\, \mathrm{d}\xi \qquad (35)$$

in which case $\pi(\theta)$ is termed an *invariant* (or stationary) density with respect to the transition kernel $K(\theta_k \mid \theta_{k-1})$.

With this in mind, we now establish that Algorithm 4.1 is targeted at the posterior $p(\theta \mid Y)$ of interest.

Lemma 5.1 *The posterior density $p(\theta \mid Y)$ defined by (12) is an invariant density of the Markov chain realised by Algorithm 4.1.*

PROOF. Assume first that $\xi \neq \theta$. Then according to the formulation (32) for the transition density of the Markov chain realised by Algorithm 4.1, and using the expression (27)

$$\begin{aligned} p(\theta \mid Y)K(\xi \mid \theta) &= p(\theta \mid Y)\gamma(\xi \mid \theta) \times \\ &\quad \min\left\{1, \frac{p(\xi \mid Y)}{p(\theta \mid Y)} \cdot \frac{\gamma(\theta \mid \xi)}{\gamma(\xi \mid \theta)}\right\} \\ &= \min\left\{p(\theta \mid Y)\gamma(\xi \mid \theta), p(\xi \mid Y)\gamma(\theta \mid \xi)\right\}. \end{aligned} \qquad (36)$$

Similarly,

$$\begin{aligned} p(\xi \mid Y)K(\theta \mid \xi) &= p(\xi \mid Y)\gamma(\theta \mid \xi) \times \\ &\quad \min\left\{1, \frac{p(\theta \mid Y)}{p(\xi \mid Y)} \cdot \frac{\gamma(\xi \mid \theta)}{\gamma(\theta \mid \xi)}\right\} \\ &= \min\left\{p(\xi \mid Y)\gamma(\theta \mid \xi), p(\theta \mid Y)\gamma(\xi \mid \theta)\right\}. \end{aligned} \qquad (37)$$

Therefore, comparing (36) and (37) and noting that the $\min\{\cdot,\cdot\}$ operation is symmetric implies that algorithm 4.1 yields a Markov chain for which the so-called 'reversibility condition'

$$p(\theta \mid Y)K(\xi \mid \theta) = p(\xi \mid Y)K(\theta \mid \xi) \qquad (38)$$

holds when $\xi \neq \theta$. Similarly, considering now the case of $\xi = \theta$, then (38) trivially holds simply by substitution of $\xi = \theta$ into the left hand side of (38).

Therefore, (38) holds for all possible transitions. Substituting $p(\cdot \mid Y)$ for $\pi(\cdot)$ into the right hand side of (35) and using (38) then implies

$$\begin{aligned} \int_{\mathcal{X}} K(\theta \mid \xi)p(\xi \mid Y)\, \mathrm{d}\xi &= \int_{\mathcal{X}} K(\xi \mid \theta)p(\theta \mid Y)\, \mathrm{d}\xi \\ &= p(\theta \mid Y)\int_{\mathcal{X}} K(\xi \mid \theta)\, \mathrm{d}\xi \\ &= p(\theta \mid Y) \end{aligned} \qquad (39)$$

where the transition to the last line follows since $K(\xi \mid \theta)$ is a probability density function and hence integrates to one. □

Therefore, the desired posterior density $p(\theta \mid Y)$ is a candidate for any density that realisations of Algorithm 4.1 might converge to. In fact, under further mild assumptions, it is the only candidate. To establish this and further results related to convergence, it is useful to define the associated invariant distribution

$$\varphi(A) = \int_A p(\xi \mid Y)\, \mathrm{d}\xi \qquad (40)$$

and also define the supports Θ, Γ of $p(\theta \mid Y)$, $\gamma(\xi \mid \theta)$ as

$$\Theta \triangleq \{\theta \in \mathcal{X} : p(\theta \mid Y) > 0\}, \qquad (41)$$

$$\Gamma \triangleq \{\xi \in \mathcal{X} : \gamma(\xi \mid \theta) > 0, \forall \theta \in \Theta\}. \qquad (42)$$

This allows the definition of certain assumptions that will be required in the results to follow, which are now collected together for convenience of reference.

Assumptions 5.1

(1) The support Γ of the proposal $\gamma(\xi \mid \theta)$ contains the support Θ of the posterior $p(\theta \mid Y)$ (required in all following results):

$$\Theta \subseteq \Gamma; \tag{43}$$

(2) The target posterior is bounded and its support is connected (required in all following results):

$$p(\theta \mid Y) < \kappa < \infty, \qquad \Theta \textit{ is a connected set}; \tag{44}$$

(3) The proposal density is bounded (required for Theorem 5.2 and all results following it):

$$\gamma(\xi \mid \theta) < \kappa < \infty, \quad \forall \theta \in \Theta; \tag{45}$$

(4) There exists some $\epsilon > 0$ such that (required for Theorems 6.1,6.2)

$$\gamma(\xi \mid \theta) > \epsilon\, p(\xi \mid Y) \qquad \forall \theta \in \Theta. \tag{46}$$

The first three of these assumptions can be considered quite mild restrictions. For example, regarding the assumption (43), it is intuitively clear that any sensible implementation of Algorithm 4.1 must incorporate a density $\gamma(\xi \mid \theta)$ that can generate proposals $\{\xi_k\}$ throughout the entire support Θ. Assuming boundedness of posterior and proposal densities is also relatively unrestrictive. The final assumption (46) is only required to generate overbounds on convergence rate, and is not required for convergence itself to occur. It will be commented on further when it is employed in section 6.

Using the first two of these assumptions, the uniqueness of $p(\theta \mid Y)$ as a candidate for what Algorithm 4.1 might converge to is established as follows.

Lemma 5.2 *Suppose that assumptions* (43) *and* (44) *hold. Then the invariant density $p(\theta \mid Y)$ of the Markov chain realised by Algorithm 4.1 is unique. That is, it is the only density satisfying (35).*
PROOF. According to (32), for $\xi \neq \theta$

$$\begin{aligned} K(\xi \mid \theta) &= \alpha(\xi \mid \theta)\gamma(\xi \mid \theta) && (47)\\ &= \min\left\{\gamma(\xi \mid \theta), \frac{p(\xi \mid Y)}{p(\theta \mid Y)}\gamma(\theta \mid \xi)\right\}. && (48) \end{aligned}$$

Therefore under the assumption (43), $K(\xi \mid \theta) > 0$ for any $\xi, \theta \in \Theta$ and hence via (33), (40) and (41), for any $\theta \in \Theta$

$$\varphi(A) > 0 \;\Rightarrow\; \mathbf{P}(A \mid \theta) > 0. \tag{49}$$

Therefore, the Markov chain realised by Algorithm 4.1 is φ-irreducible. Therefore, by Proposition 10.1.1 of [26] the chain is recurrent, and hence by Theorem 10.2.1 of [26], the invariant measure φ, and hence (by boundedness) the density $p(\theta \mid Y)$ are unique. □
Therefore, if the realisations from the Markov chain implemented by Algorithm 4.1 converge to having a stationary density, the only possibility for that density is the desired posterior $p(\theta \mid Y)$.

To address the issue of actual convergence, recall the definition (33) of the transition distribution $\mathbf{P}(A \mid \theta)$ of the associated Markov chain. Further define by $\mathbf{P}^n(A \mid \theta_0)$ the distribution of the n'th iteration of the chain when started at θ_0 and note that it can be computed recursively according to [26, Theorem 3.4.2]

$$\begin{aligned} \mathbf{P}^n(A \mid \theta_0) &= \int \mathbf{P}^{n-1}(A \mid \xi)\mathbf{P}(\mathrm{d}\xi \mid \theta_0), && (50)\\ \mathbf{P}^1(A \mid \xi) &\triangleq \mathbf{P}(A \mid \xi). && (51) \end{aligned}$$

With this formulation of $\mathbf{P}^n$ in place, its convergence to the invariant distribution φ defined by (40) can be established as follows.

Theorem 5.1 *Under the assumptions of Lemma 5.2*

$$\lim_{n\to\infty} \sup_{A\in\sigma(\mathcal{X})} |\mathbf{P}^n(A \mid \theta_0) - \varphi(A)| = 0 \tag{52}$$

for φ almost all $\theta_0 \in \Theta$ where φ is defined by (40) and $\sigma(\mathcal{X})$ is the Borel sigma algebra of sets on $\mathcal{X}$.
PROOF. See Appendix B. □
However, this paper proposes the use an estimate $\widehat{f}_M$ of the form (23), (24), so while convergence in distribution is of interest, it is equally important to consider convergence of sample averages.

Theorem 5.2 *Under the assumptions of Lemma 5.2 and the further conditions that* (45) *holds and that $f : \mathcal{X} \to \mathbf{R}$ satisfies*

$$\int_{\mathcal{X}} |f(\theta)|\, p(\theta \mid Y)\, \mathrm{d}\theta < \infty, \tag{53}$$

then for the sequence $\{\theta_k\}$ generated by Algorithm 4.1

$$\lim_{M\to\infty} \frac{1}{M}\sum_{k=1}^{M} f(\theta_k) = \int_{\mathcal{X}} f(\theta)p(\theta \mid Y)\, \mathrm{d}\theta \tag{54}$$

with probability one.
PROOF. See Appendix C. □
Therefore, with the choice $f(\theta) = I_{f^{-1}(A)}(\theta)$ being the indicator function for $f(\theta) \in A$ with A being an arbitrary (measurable) subset of the range of f, the above theorem establishes that the sample histogram is a strongly consistent estimator of the underlying true posterior density. That is

$$\lim_{M\to\infty} \frac{1}{M}\sum_{k=1}^{M} I_{f^{-1}(A)}(\theta_k) = p(f(\theta) \in A \mid Y) \tag{55}$$

with probability one. In particular, if $f(\theta) = I_{\theta^i \in A}(\theta)$ so that f determines the presence of only the i'th element of θ in a set A, then (55) can be used to compute the posterior *marginal* density $p(\theta^i \mid Y)$ which, as explained in section 3, is generally intractable by any other means.

Furthermore if a point estimate of θ is required, then since the posterior expectation $\mathbf{E}\{\theta \mid Y\}$ is the minimum variance estimate [1, Theorem 3.2, page 30],[4, Theorem

1.2, page 124], it is a reasonable choice. Again, via Theorem 5.2, this can be estimated in a strongly consistent fashion with the choice $f(\theta) = \theta$ to provide

$$\lim_{M\to\infty} \frac{1}{M}\sum_{k=1}^{M} \theta_k = \mathbf{E}\{\theta \mid Y\} \tag{56}$$

with probability one.

6 Convergence Rate

Having established mild conditions sufficient for convergence of Algorithm 4.1, this section turns to the question of convergence rate, which is of important practical interest.

For this purpose, it is necessary to strengthen the assumption of $\Theta \subseteq \Gamma$ to be being the final condition (46) of Assumption set 5.1 holding. Again, this is not considered a strong requirement.

For example, it is common that due to considerations of what is physically reasonable for the problem at hand, a prior $p(\theta)$ is imposed that limits the range of feasible θ to a bounded set. Since $p(\theta \mid Y) \propto p(Y \mid \theta)p(\theta)$ by (12), this implies that commonly, the posterior $p(\theta \mid Y)$ will be a truncated version of $p(Y \mid \theta)$. Due to the normalising constant $p(Y)$ (again recall (12)), this further implies that in many practical applications, the posterior $p(\theta \mid Y)$ is bounded away from zero on its support Θ, making (46) feasible for sufficiently small $\epsilon > 0$.

While this common situation of $p(\theta \mid Y)$ having bounded support is sufficient (46) to hold, it is not necessary. For example, in the simple scalar case where $p(\theta \mid Y) = \mathcal{N}(0, \sigma_\theta^2)$ and $\gamma(\xi \mid \theta) = \mathcal{N}(0, \theta_\xi^2)$, then a straightforward calculation establishes that when $\sigma_\xi^2 > \sigma_\theta^2$, the choice of $\epsilon < \sqrt{\sigma_\theta^2/\sigma_\xi^2}$ leads to (46) holding even though $p(\theta \mid Y)$ has unbounded support.

With this in mind, in situations where the strengthened requirement (46) does hold, an exponential bound on the distributional convergence first addressed in Theorem 5.1 can be established.

Theorem 6.1 *Under the assumptions of Theorem 5.2, and the further condition* (46)

$$\sup_{A\in\sigma(\mathcal{X})} |\mathbf{P}^n(A \mid \theta_0) - \varphi(A)| \le (1-\epsilon)^n \tag{57}$$

for any $\theta_0 \in \Theta$.

PROOF. See Appendix D. □

Note that due to the constraint that both $\gamma(\xi \mid \theta)$ and $p(\xi \mid Y)$ are unit area, then as ϵ is increased away from zero, the condition (46) can only be satisfied if the functional form of $\gamma(\xi \mid \theta)$ becomes close to that of $p(\xi \mid Y)$. Therefore, (57) indicates that the closer the proposal $\gamma(\xi \mid \theta)$ is to the desired posterior $p(\xi \mid Y)$, then the faster the exponential convergence of the distribution of realisations $\{\theta_k\}$ generated by Algorithm 4.1.

While the rapid exponential distributional convergence (57) is reassuring, it does not imply fast convergence of a sample averages $\widehat{f}_M$ such as (23), (24) which are the main focus of this paper. This is because the latter depends additionally on the correlation between samples.

To address this, denote a centered (i.e. zero mean) version $\widetilde{f}_k$ of $f(\theta_k)$ as

$$\widetilde{f}_k \triangleq f(\theta_k) - \mathbf{E}(f(\theta_k) \mid Y) \tag{58}$$

where $f : \mathcal{X} \to \mathbf{R}$ is a function satisfying the conditions of Theorem 5.2 together with $|f| < \infty$. Associated with this, define the conditional asymptotic variance σ_f^2 of the estimate $\widehat{f}_M$ (see (23)) as

$$\sigma_f^2 \triangleq \lim_{M\to\infty} \frac{1}{M}\mathbf{E}\left\{ \left(\sum_{k=1}^{M} \widetilde{f}_k\right)^2 \middle| Y \right\}. \tag{59}$$

A quantification of the rate of convergence of $\widehat{f}_M$ to $\mathbf{E}\{f(\theta) \mid Y\}$ is then possible via the following result.

Theorem 6.2 *Under the conditions of Theorem 6.1, the limit in (59) exists and is finite and for any $\lambda < 1/2$*

$$\lim_{M\to\infty} \frac{1}{M^{1-\lambda}} \sum_{k=1}^{M} [f(\theta_k) - \mathbf{E}\{f(\theta \mid Y)\}] = 0 \tag{60}$$

with probability one. Furthermore, if σ_f^2 defined by (59) is non-zero, then

$$\sqrt{M}\left(\frac{1}{M}\sum_{k=1}^{M} f(\theta_k) - \mathbf{E}\{f(\theta \mid Y)\}\right) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \sigma_f^2) \tag{61}$$

as $M \to \infty$.

PROOF. The convergence of (59) and the asymptotic distributional result (61) follows by direct application of the result (57) and thence Theorem A.3. Furthermore, since (59) holds, then for some $C < \infty$

$$\mathbf{E}\left\{ \left(\sum_{k=1}^{M} \widetilde{f}_k\right)^2 \middle| Y \right\} < C\,M. \tag{62}$$

The result (60) is then obtained by the application of Theorem 2.1 of [29]. □

Therefore, with the definition (23) for the sample average estimate $\widehat{f}_M$, the result (60) provides the convergence rate bound

$$\widehat{f}_M - \mathbf{E}\{f(\theta) \mid Y\} = o\left(\frac{1}{M^\lambda}\right), \qquad \lambda < 1/2 \tag{63}$$

which applies with probability one. This $M^{-1/2}$ bound on the convergence rate is further sharpened by the distributional result (61). In particular, while (63) shows that the convergence rate is at least (arbitrarily close to) $M^{-1/2}$, the convergence (61) indicates (with respect to weak convergence of distributions) it is no faster than $M^{-1/2}$.

7 Choice of Proposal Distribution

The main design variable in the implementation of Algorithm 4.1 is the choice of the proposal density $\gamma(\xi \mid \theta)$. This involves a tradeoff between convergence rate and complexity. For example, as the correlation between the realisations $\{\theta_k\}$ decreases, the convergence rate of the sample average $\widehat{f}_M$ to $\mathbf{E}\{f(\theta) \mid Y\}$ increases.

However, as correlation is minimised, algorithm complexity may increase. At one end of the scale, substituting the choice $\gamma(\xi \mid \theta) = p(\theta \mid Y)$ into (27) implies an acceptance probability $\alpha(\xi \mid \theta) = 1$, so that realisations of Algorithm 4.1 are independent realisations from $p(\theta \mid Y)$, and hence $\widehat{f}_M$ converges maximally fast.

This is clearly infeasible, since the entire premise of Markov chain Monte–Carlo method is that they are employed because sampling from $p(\theta \mid Y)$ is computationally impossible. Therefore, it is necessary to consider suboptimal choices for $\gamma(\xi \mid \theta)$ that are reasonable while not overly sacrificing convergence rate.

One of the simplest proposal choices is the random walk (30) with $\nu_k \sim \mathcal{N}(0, \sigma_\nu^2 I)$, which leaves the variance σ_ν^2 as a single design variable. If chosen too small, then almost all proposals will be accepted, and the correlation between samples will be very high. This will manifest in realisations $\{\theta_k\}$ very slowly trawling the range where $p(\theta \mid Y)$ is non-zero, resulting in very slow convergence. Vice versa, if σ_ν^2 is chosen too high, then overly large jumps in θ_k will be proposed that are rarely in regions where $p(\theta \mid Y)$ is significant, and hence rarely accepted. Again, the correlation between samples will be very high (realisations θ_k will remain the same over long periods of rejected proposals) and convergence will be slow.

In consideration of this, the authors have found it effective to adaptively modulate σ_ν^2 in order for an observed acceptance rate a_L to achieve a given target $\alpha_\star$. Here, the rate α_L is defined as the sample average proportion of acceptances over a window of width L (below δ is the Kronecker delta):

$$a_L \triangleq \frac{1}{L} \sum_{k=1}^{L} \delta(\theta_k - \xi_k). \tag{64}$$

There are many ways to implement this principle, and one found to be successful is to initialise σ_ν^2 and L at some small values (say 10^{-6} and 10 respectively) and then for some $\lambda > 1$ (say, 1.2) perform the following steps

```
$k = 1$
while $k <$ maxvalue do
    if $k \bmod L = 0$ then
        if $\alpha_L < \alpha_\star$ then
            $\sigma_\nu^2 \mapsto \lambda \sigma_\nu^2$
        else
            $\sigma_\nu^2 \mapsto \sigma_\nu^2 / \lambda$ .
        end if
        $L \mapsto 2L$
    end if
    $k \mapsto k + 1$
end while
```

Clearly, this is quite heuristic, but has been found effective in practice (as illustrated in the following section 8) with a choice of $\alpha_\star \in [0.2, 0.4]$. Note that under the strong simplifying assumption of the components θ^i of θ being conditionally (on Y) independent, theoretical analysis is available [37] to conclude that $\alpha_\star = 0.234$ provides the fastest convergence rate in the sense given by the left hand side of (57).

8 Simulation Study

To illustrate the application of these ideas, this section considers the case of a linear and time invariant system model of the output error (OE) form

$$y_t = G(q, \theta) u_t + \varepsilon_t, \qquad G(q, \theta) = \frac{B(q, \theta)}{A(q, \theta)}, \tag{65}$$

where

$$A(q, \theta) = 1 + a_1 q^{-1} + a_2 q^{-2} + \cdots + a_{m_a} q^{-m_a}, \tag{66}$$

$$B(q, \theta) = b_0 + b_1 q^{-1} + b_2 q^{-2} + \cdots + b_{m_b} q^{-m_b}, \tag{67}$$

$$\theta^T = [a_1, \cdots, a_{m_a}, b_0, \cdots, b_{m_b}]. \tag{68}$$

In this case, to evaluate the likelihood (6) the required predictor is given by (5) as simply $\widehat{y}_{t|t-1}(\theta) = G(q, \theta) u_t$ so that via (12) and (6) the posterior $p(\theta \mid Y)$ can be evaluated for any value of θ as

$$p(\theta \mid Y) = k \cdot p(\theta) \prod_{t=1}^{N} p_\varepsilon \left(y_t - G(q, \theta) u_t \right). \tag{69}$$

Here k is a constant independent of θ, that will not be required an any subsequent calculations (since it will cancel), but is included in (69) to ensure it has unit total probability.

Furthermore, in the example to follow, ε_t of variance $\mathrm{Var}\{\varepsilon_t\} = \sigma^2$ will have a uniform distribution $\varepsilon_t \sim \mathcal{U}[-1.5\sigma^{2/3}, 1.5\sigma^{2/3}]$ so that $p_\varepsilon(\cdot)$ will be the indicator function $I_{[-1.5\sigma^{2/3}, 1.5\sigma^{2/3}]}(\cdot)$ and hence the product term in (69) will be either one or zero. Similarly, in what is to follow, the prior $p(\theta)$ will be uniform and hence zero or not.

A further specialisation in the example presented in this section is the employment (in the MCMC Algorithm 4.1) of the random walk proposal (30)

$$\xi_k = \theta_{k-1} + \nu_k, \qquad \nu_k \sim \mathcal{N}(0, \sigma_\nu^2 I) \tag{70}$$

whereby the previous iteration θ_{k-1} is perturbed by a Gaussian distributed random amount ν_k. Recall, that as explained via (31), in this situation the acceptance probability $\alpha(\xi \mid \theta)$ simplifies to the Metropolis form (29). As a result, in the example we present here, the MCMC Algorithm 4.1 is implemented as follows.

Algorithm 8.1 ***(MCMC for OE Model)***

(1) Initialise θ_0 at some value such that according to (69) *the probability $p(\theta_0 \mid Y) > 0$, and set $k = 1$;*
(2) At iteration k, generate a candidate value ξ_k computed according to the random walk proposal (70)*;*
(3) Substitute the ξ_k obtained in step 2 and the θ_{k-1} from the previous iteration into (69) *in order to compute the acceptance probability; viz.*

$$\alpha(\xi_k \mid \theta_{k-1}) = \min\left\{1, \frac{p(\xi_k \mid Y)}{p(\theta_{k-1} \mid Y)}\right\}; \tag{71}$$

(4) Generate a realisation $z \sim \mathcal{U}[0,1]$, where $\mathcal{U}[0,1]$ represents a uniform distribution on the interval $[0,1]$.
(5) Set $\theta_k = \xi_k$ if

$$z < \alpha(\xi_k \mid \theta_{k-1}), \tag{72}$$

otherwise set $\theta_k = \theta_{k-1}$.
(6) Increment k and return to step 2.

□

Within this setting, the case studied here is the simplest possible first order one of

$$m_a = 1, \quad m_b = 0, \quad a_1 = -0.8, \quad b_0 = 0.2 \tag{73}$$

with $\{\varepsilon_t\}$ a zero mean i.i.d. process of variance $\sigma^2 = \mathbf{E}\{\varepsilon_t^2\} = 0.01$. It is then supposed that the available data from this system consists of only $N = 20$ samples of $\{y_t\}$ and $\{u_t\}$ being a sampled step response transiting $1 \mapsto 0$ at the data record midpoint. This is illustrated in Figure 1, where the solid line is the noise free response, and the samples around this line are the noise corrupted data assumed to be available.

In the case where the density $p_\varepsilon(\cdot)$ governing ε_t is uniform, and with prior distribution on $\theta = [a_1, b_0]$ being one that assigns zero weight to $b_0 < 0$ and $|a_1| > 1$, then the posterior distributions for these parameters given the data realisation shown in Figure 1 are illustrated in Figure 2.

There, the solid line shows the marginal posterior density for b_0 and a_1 computed via Algorithm 4.1 with the random walk proposal (30) using perturbations $\{\nu_k\}$ which are i.i.d. zero mean Gaussian with variance tuned via the algorithm of the preceding section to deliver an $\alpha_\star = 0.3$ acceptance rate. These marginals were produced from histograms based on 10^5 iterations of Algorithm 4.1, which were then smoothed using standard kernel density estimation methods [39].

By way of comparison, the marginals obtained by numerical integration using Simpson's rule over 220 bins to evaluate (17) by 'brute force' are shown as a dashed line in Figure 2. They are virtually indistinguishable from the solid line, indicating the accuracy of the MCMC approach in this example.

As further comparison, the error quantification (10) associated with a least squares estimate obtained from the data in Figure 1 is shown as the dash-dot Gaussian curves in Figure 2. While these quantifications are not strictly comparable to the posterior densities, since they evaluate different quantities, it would still seem interesting to compare the two in terms of their utility for informing a user of what system information can be extracted from the available data, particularly in view of the very widespread use of the least squares method and (asymptotic based) associated error quantification.

In relation to this, note that since $p_\varepsilon(\cdot)$ is uniform in this example, then via (6) the likelihood is constant on the region Θ defined in (19), and hence has no uniquely defined maximum, so that a maximum likelihood estimator cannot be profiled.

To illustrate further potential uses of the Markov chain methods proposed here, suppose that it is necessary to design a closed loop PI controller $K(q)$ for the system responsible for the observations in Figure 1 and under the hypothesis that the system is first order.

The choice

$$K(q) = 2 + \frac{0.1}{q-1} \tag{74}$$

achieves a phase margin $\phi_m = 99.3°$ and gain margin $g_m = 5.56$ on the afore-mentioned least squares estimate.

However, it is of course important to gauge the likely performance of the controller (74) on the real system. As argued in this paper, a Bayesian approach addresses this question by computing the posteriors

$$p(\phi_m \mid Y), \qquad p(g_m \mid Y). \tag{75}$$

Due to the implicit way in which ϕ_m and g_m are defined (i.e. they do not obey a closed form formula) this would be a daunting (if not impossible) task if approached from an analytical point of view, or via standard asymptotic approximation methods such as (10), (11).

In contrast, it is straightforward to compute the required marginals (75) using the Monte–Carlo approach of this paper. Each realisation of $\{\theta_k\}$ provided by Algorithm 4.1, implies an associated ϕ_m and g_m which is straightforward to compute. Furthermore, they (or histograms of their values) may each be thought of as arising from a bounded mapping $f : \Theta \to \mathbf{R}$ so that Theorem 5.2 assures convergence of associated sample averages.

The results of this approach are shown in Figure 3. Their accuracy is ensured by the demonstrated accuracy of the distribution of the marginals of $p(\theta \mid Y)$ in Figure 2. Clearly, there seems to be good evidence from the data that the the controller (74) will achieve a phase margin greater than 95° and a gain margin greater than 3.7.

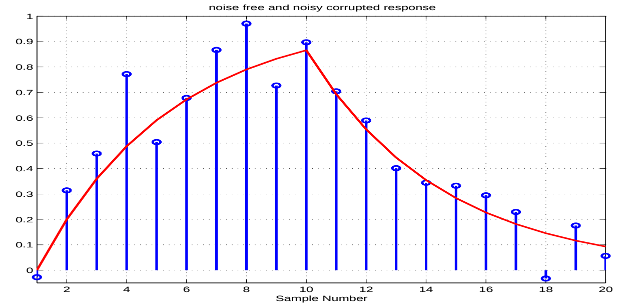
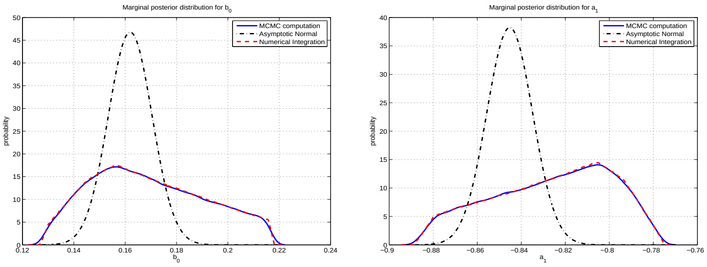


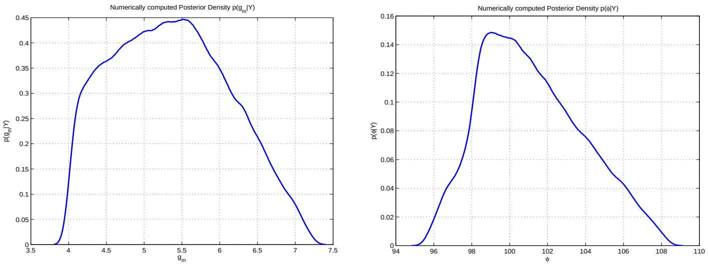
Fig. 1. *First order system response: Solid line is noise free, sampled dots are the noise corrupted measurements available.*



(a) Posterior $p(b_0 \mid Y)$

(b) Posterior $p(a_1 \mid Y)$

Fig. 2. *Posterior marginal densities of parameters computed via Algorithm 4.1 shown as a solid line together with (dashed line) another evaluation of the posterior marginals computed via numerical integration of the joint posterior, and (dash-dot line), the parameter information that would be inferred from the data via a least squares approach with the asymptotic in N distribution approximation (10).*



(a) Gain Margin Posterior

(b) Phase Margin Posterior

Fig. 3. *Posterior distributions for phase margin ϕ_m and gain margin g_m for a given PI controller.*

9 Discussion

While the Bayesian/MCMC approach presented here has some attractive aspects, such the ability to compute otherwise intractable quantities, it has limitations which are important to recognise.

The Bayesian approach itself (independent of a MCMC or other computational solution) has the disadvantage of requiring more prior information than some other approaches, such as maximum-likelihood (ML) or prediction error (PE). For example, Bayesian and ML solutions require knowledge of the probability density function of the measurement noise.

Furthermore, neither the PE or ML methods require specification of a prior density $p(\theta)$ on the parameters as the Bayesian approach does. Of course, in some situations, the ability to incorporate prior knowledge (for example, system stability, or positivity of physical quantities) is actually a strength. Nevertheless, this also raises the important consideration of the sensitivity of the ensuing Bayesian estimate (either the MAP or conditional mean) to the prior. This, and indeed the idea of a prior itself, has been at the heart of strong debate between so-called frequentist and Bayesian schools within the statistics community for at least a century; see [18,36,11] for recent perspectives and comments on the past, and in particular see [34] for a discussion relevant to system identification. The debate seems far from settled, and this paper does not pretend to make a contribution on these fundamentals.

Rather, the scope here is to suggest that since MCMC methods now provide a tool for computing Bayesian estimates, it is worth considering and evaluating their utility for addressing system identification problems.

Turning to limitations of the MCMC-based solution, the most obvious one is the computational burden involved in generating a sufficiently large number of realisations $\{\theta_k\}$ for the sample average approximation (24) to be an accurate evaluation of a required posterior density. This is undeniably a key weakness, but there are aspects mitigating it.

First, while a number of realisations are required, the computation involved with each one is modest, since it is dominated by evaluating $p(\xi \mid Y)$ which is almost identical to computing the cost $V_N(\theta)$ associated with a prediction error approach.

Second, the MCMC approach is eminently parallelisable. Five MCMC runs initialised independently an run for ten thousand iterations will take 20% of the time of one run of fifty thousand iterations but can provide results of equivalent accuracy.

Finally, and related to the above issues of convergence, an important limitation of the MCMC approach is that it computes only an *approximation* to the desired posterior $p(\theta \mid Y)$. This approximation can be made arbitrarily accurate by increasing the simulation length M (without, it must be stressed, requiring any increase in the length N of the observed data record Y). While this approximation can be very accurate, as illustrated in Figure 2, it will never be formally exact.

10 Conclusion

This paper has presented new Monte–Carlo based techniques to support a Bayesian approach to dynamic system estimation, with particular emphasis on the potential for accurate estimation error quantification from short data records. While both theoretical and empirical analysis has been presented to establish algorithm capabilities, there remain many issues worthy of further study.

A Markov Chains on Non-countable Spaces

The theoretical analysis of this paper relies on ideas and results pertaining to Markov chains on non-countable spaces. Since these tools are unlikely to be familiar to the general reader, this section gives a brief synopsis of the necessary material.

Let $\{\theta_k\} \in \mathcal{X} \subseteq \mathbf{R}^n$ be a sequence of random variables where the collection has a joint probability density function $p(\theta_0, \cdots, \theta_N)$ which can be decomposed according to Bayes' rule as

$$p(\theta_1, \cdots, \theta_N) = p(\theta_N \mid \theta_{N-1}, \cdots, \theta_0)p(\theta_{N-1}, \cdots, \theta_0). \tag{A.1}$$

This sequence $\{\theta_k\}$ is termed a *Markov chain* if the conditional probability density $p(\theta_N \mid \theta_{N-1}, \cdots, \theta_0)$ in (A.1) satisfies

$$p(\theta_N \mid \theta_{N-1}, \cdots, \theta_0) = p(\theta_N \mid \theta_{N-1}). \tag{A.2}$$

Associated with any such Markovian density, is its distribution measure $\mathsf{P}(\cdot \mid \cdot)$ defined for any $A \in \sigma(\mathcal{X})$ (where $\sigma(\mathcal{X})$ is the Borel sigma algebra) and any $\theta \in \mathcal{X}$ as

$$\mathsf{P}(A \mid \theta) = \int_A p(\xi \mid \theta)\, \mathrm{d}\mu(\xi) \tag{A.3}$$

where μ denotes Lebesgue measure. This distribution, also called a *transition probability kernel*, completely characterises the Markov chain.

Now consider an arbitrary measure ψ on $\mathcal{X}$. A Markov chain $\mathsf{P}(A \mid \theta)$ is then termed *ψ-irreducible* if for each $\theta \in \mathcal{X}$ and each $A \in \sigma(\mathcal{X})$ such that $\psi(A) > 0$, there exists a $k < \infty$ such that [26, Page 88]

$$\mathsf{P}^k(A \mid \theta) > 0. \tag{A.4}$$

That is, for any set that is non-trivial with respect to the measure ψ, there is a non-zero probability of eventually entering that set.

Furthermore, a ψ-irreducible Markov chain $\mathsf{P}(A \mid \theta)$ is termed *recurrent* if for any set $A \in \sigma(\mathcal{X})$ with $\psi(A) > 0$, the conditions [41, Section 3.1],[33, Definition 3.5]

$$\mathsf{P}_\theta\{\theta_n \in A \text{ infinitely often}\} > 0 \quad \text{for all } \theta, \tag{A.5}$$

and

$$\mathsf{P}_\theta\{\theta_n \in A \text{ infinitely often}\} = 1 \quad \text{for } \psi \text{ almost all } \theta, \tag{A.6}$$

are both satisfied, where the notation P_θ denotes '*the probability of events conditional on the chain beginning with $\theta_0 = \theta$*'.

An important point is that any irreducible chain that is capable of converging to a stationary distribution is recurrent. To state this precisely, suppose that a particular state θ_k has distribution function $\varphi_k(\cdot)$, and that we seek the distribution function $\varphi_{k+1}(\cdot)$ of θ_{k+1}. Then clearly

$$\varphi_{k+1}(A) = \int \mathsf{P}(A \mid \xi)\varphi_k(\mathrm{d}\xi). \tag{A.7}$$

Therefore, if the chain is to have any hope of converging in the sense of the distribution functions $\{\varphi_k(\cdot)\}$ converging, say to the distribution $\varphi(\cdot)$, then this latter distribution must satisfy

$$\varphi(A) = \int \mathsf{P}(A \mid \xi)\varphi(\mathrm{d}\xi). \tag{A.8}$$

A distribution which satisfies (A.8) is termed an *invariant* distribution for the chain [26, Page 235].

If a chain $\mathsf{P}(A \mid \theta)$ is ψ-irreducible for some measure ψ, and also admits an invariant probability measure φ, then $\mathsf{P}(A \mid \theta)$ is called a *positive chain* [26, Page 235]. Two important results are that a positive chain is necessarily recurrent [26, Proposition 10.1.1], and that the invariant measure φ is unique [26, Theorem 10.2.1],[41].

A strengthened version of (A.5) (A.6) leads to a stronger version of recurrence. Namely, if a ψ-irreducible recurrent chain $\mathsf{P}(A \mid \theta)$ also satisfies [26, Page 204]

$$\mathsf{P}_\theta\{\theta_n \in A \text{ infinitely often}\} = 1 \quad \text{for all } \theta. \tag{A.9}$$

then it is termed *Harris recurrent.* This definition is important as a necessary condition for a key ergodic result.

Theorem A.1 (Law of Large Numbers) *Suppose that $\{\theta_k\}$ is a realisation of a positive Harris recurrent chain with invariant distribution φ. Suppose that $f : \mathcal{X} \to \mathbf{R}$ is such that*

$$\int |f(\xi)|\varphi(\mathrm{d}\xi) < \infty. \tag{A.10}$$

Then

$$\lim_{N\to\infty} \frac{1}{N}\sum_{k=1}^{N} f(\theta_k) = \int f(\xi)\varphi(\mathrm{d}\xi), \quad \forall \theta_0 \in \mathcal{X} \tag{A.11}$$

with probability one.

PROOF. Follows by Theorem 17.1.7 of [26]. □

Establishing Harris recurrence, and hence ergodicity is achieved in this paper by consideration of what are termed harmonic functions. Namely, a function $h : \mathcal{X} \to \mathbf{R}$ satisfying

$$h(\theta) = \int h(\xi)\mathbf{P}(\mathrm{d}\xi \mid \theta), \quad \forall \theta \in \mathcal{X} \tag{A.12}$$

is termed *harmonic* with respect to the measure $\mathbf{P}(A \mid \theta)$. If $\mathbf{P}(A \mid \theta)$ is recurrent then necessarily $h(x) = \overline{h}$ a constant, for ψ almost all x [33, Proposition 3.13]. Furthermore, $\mathbf{P}(A \mid \theta)$ is Harris recurrent if, and only if, every bounded harmonic function with respect to $\mathbf{P}(A \mid \theta)$ is a constant everywhere [41, Theorem 2].

Finally, a ψ-irreducible Markov chain $\mathbf{P}(A \mid \theta)$ is termed *aperiodic* if there exists a set $C \in \mathcal{X}$ and a sequence of positive measures $\nu_k(\cdot)$ such that for any $\theta \in C$ and for any set $A \in \mathcal{X}$ [26, Page 121]

$$\text{g.c.d.}\left\{k \geq 1 : \mathbf{P}^k(A \mid \theta) > \nu_k(A)\right\} = 1 \tag{A.13}$$

where g.c.d. stands for 'greatest common denominator'. That is, if the chain is initialised at $\theta_0 \in C$, then it can visit any other state at any time. Via a non-trivial argument [26, Page 120], the aperiodicity of a ψ-irreducible chain is invariant to the initial set C chosen, except possibly for ones which differ on a set which is measure zero with respect to ψ.

The importance of aperiodicity is that when it holds in addition to recurrence, then a chain converges to a stationary distribution which is its invariant measure.

Theorem A.2 *Suppose that $\mathbf{P}(A \mid \theta)$ is a ψ-irreducible and aperiodic recurrent chain with invariant measure φ. Then for φ almost all θ*

$$\lim_{n\to\infty} \sup_{A\in\sigma(\mathcal{X})} |\mathbf{P}^n(A \mid \theta) - \varphi(A)| = 0. \tag{A.14}$$

Furthermore, if the chain is Harris recurrent, then (A.14) holds for any initial condition $\theta \in \mathcal{X}$.

PROOF. The first part not requiring Harris recurrence, is Theorem 1 of [41]. The latter part requiring Harris recurrence is established via [26, Proposition 13.0.1]. □

Finally, if the distributional convergence in (A.14) is sufficiently rapid, then sample averages of realisations $\{\theta_k\}$ from the Markov chain $\mathbf{P}(A \mid \theta)$ obey a Central Limit Theorem.

Theorem A.3 *Suppose that $\mathbf{P}(A \mid \theta)$ is a ψ-irreducible aperiodic Harris recurrent chain with invariant measure φ and suppose that for some $M < \infty$, $r \in (0, 1)$*

$$\sup_{A\in\sigma(\mathcal{X})} |\mathbf{P}^n(A \mid \theta) - \varphi(A)| \leq Mr^n. \tag{A.15}$$

Let $f : \mathcal{X} \to \mathbf{R}$ be such that $|f| < \kappa < \infty$ on $\mathcal{X}$. Then the limit

$$\sigma_f^2 \triangleq \lim_{N\to\infty} \frac{1}{N}\mathbf{E}_\varphi\left\{\left[\sum_{k=1}^{N} f(\theta_k) - \mathbf{E}_\varphi\{f\}\right]^2\right\} \tag{A.16}$$

exists and is finite, where $\mathbf{E}_\varphi\{\cdot\}$ denotes expectation with respect to the measure φ. Furthermore, if $\sigma_f^2 > 0$ then

$$\sqrt{N}\left(\frac{1}{N}\sum_{k=1}^{N} f(\theta_k) - \mathbf{E}_\varphi\{f\}\right) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \sigma_f^2) \tag{A.17}$$

as $N \to \infty$.

PROOF. See Theorem 17.0.1 of [26]. □

B Proof of Theorem 5.1

PROOF. In the proof of Lemma 5.2 the Markov chain $\mathbf{P}(A \mid \theta)$ has been established as positive recurrent with invariant measure φ given by (40). Furthermore, via (49), (50) $\mathbf{P}^n(A \mid \theta) > 0$ for any set A satisfying $\varphi(A) > 0$ and hence the chain $\mathbf{P}(A \mid \theta)$ is aperiodic for any starting $\theta_0 \in \Theta$. Application of Theorem A.2 then completes the proof. □

C Proof of Theorem 5.2

PROOF. The proof which follows is due to Tierney [41]. However, it is presented there in a very compact form with many details and steps obvious only to a specialist. In order to provide a self contained treatment of the theory underlying and justifying a Markov-chain Monte–Carlo approach, what follows is the authors interpretation of the method of Tierney [41], with the arguments expanded so as to be explicit for a non-specialist in Markov chain theory.

Firstly, as established in Lemma 5.1,

$$\varphi(A) = \int_A p(\theta \mid Y)\,\mathrm{d}\theta$$

is an invariant distribution of the Markov chain $\mathbf{P}(A \mid \theta)$ realised by Algorithm 4.1, and hence $\mathbf{P}(A \mid \theta)$ is positive recurrent and φ irreducible. Therefore, if there exists a bounded function $h : \mathcal{X} \to \mathbf{R}$ that is harmonic with respect to $\mathbf{P}(A \mid \theta)$ in that

$$h(\theta) = \int_{\mathcal{X}} h(\xi)\,\mathbf{P}(\mathrm{d}\xi \mid \theta) \tag{C.1}$$

then it must hold that $h(\theta) = \overline{h}$ a constant for φ almost all θ. Now, define

$$A^+ \triangleq \{\theta : p(\theta \mid Y) > 0\}, \qquad F \triangleq \{\theta : h(\theta) \neq \overline{h}\} \tag{C.2}$$

so that $\varphi(F)$=0 and $h(\theta) = \overline{h} \ \forall \ \theta \in F^c$. Furthermore, by (38)

$$\alpha(\xi \mid \theta)\gamma(\xi \mid \theta)p(\theta \mid Y) = \alpha(\theta \mid \xi)\gamma(\theta \mid \xi)p(\xi \mid Y). \tag{C.3}$$

Therefore, by the boundedness of γ and since by definition $\alpha(\cdot \mid \cdot) \leq 1$, it holds that for $\theta \in A^+$

$$\begin{aligned}&\int_F \alpha(\xi \mid \theta)\gamma(\xi \mid \theta)\,\mathrm{d}\mu(\xi)\\ =&\frac{1}{p(\theta \mid Y)}\int_F \alpha(\theta \mid \xi)\gamma(\theta \mid \xi)p(\xi \mid Y)\,\mathrm{d}\mu(\xi)\\ \leq&\frac{\kappa}{p(\theta \mid Y)}\int_F p(\xi \mid Y)\,\mathrm{d}\mu(\xi) = \frac{\kappa}{p(\theta \mid Y)}\,\varphi(F) = 0\end{aligned} \tag{C.4}$$

where the restriction of $\theta \in A^+$ is necessary due to the division by $p(\theta \mid Y)$ above. Therefore, since h is bounded and since $\alpha(\xi \mid \theta), \gamma(\xi \mid \theta) \geq 0$, then $\forall \theta \in A^+$

$$\int_F \alpha(\xi \mid \theta)\gamma(\xi \mid \theta)h(\xi)\,\mathrm{d}\mu(\xi) = 0. \tag{C.5}$$

Furthermore, via (32)

$$K(\xi \mid \theta) = \alpha(\xi \mid \theta)\gamma(\xi \mid \theta)I_{\mathcal{X}_\theta}(\xi) + r(\theta)\delta(\xi - \theta) \tag{C.6}$$

so that by the defining property (A.12) of a harmonic function, for $\theta \in A^+$,

$$\begin{aligned}h(\theta) =& \int_{\mathcal{X}} K(\xi \mid \theta)h(\xi)\,\mathrm{d}\mu(\xi)\\ =& \int_F \alpha(\xi \mid \theta)\gamma(\xi \mid \theta)I_{\mathcal{X}_\theta}(\xi)h(\xi)\,\mathrm{d}\mu(\xi) +\\ & \int_{F^c} \alpha(\xi \mid \theta)\gamma(\xi \mid \theta)I_{\mathcal{X}_\theta}(\xi)h(\xi)\,\mathrm{d}\mu(\xi) + r(\theta)h(\theta)\\ =& \, 0 + \overline{h}\int_{\mathcal{X}_\theta} \alpha(\xi \mid \theta)\gamma(\xi \mid \theta)\,\mathrm{d}\mu(\xi) + r(\theta)h(\theta)\\ =& \, \overline{h}[1 - r(\theta)] + r(\theta)h(\theta).\end{aligned}$$

Therefore

$$[1 - r(\theta)][\overline{h} - h(\theta)] = 0, \qquad \forall \theta \in A^+ \tag{C.7}$$

and hence since $\mathbf{P}(A \mid \theta)$ is φ-irreducible, then $r(\theta) < 1 \ \forall \theta \in \mathcal{X}$, so that

$$h(\theta) = \overline{h} \ \forall \ \theta \in A^+. \tag{C.8}$$

Finally, by the design of the algorithm, it is initialised with $\theta \in A^+$ and all accepted proposals lie in A^+. Therefore, the transition distribution must satisfy

$$1 = \int_{A^+} \mathbf{P}(\mathrm{d}\xi \mid \theta), \qquad 0 = \int_{A^{+c}} \mathbf{P}(\mathrm{d}\xi \mid \theta) \tag{C.9}$$

for any θ. Therefore, for $\theta \notin A^+$ and again by the boundedness of h

$$h(\theta) = \int_{A^+} h(\xi)\,\mathbf{P}(\mathrm{d}\xi \mid \theta) + \int_{A^{+c}} h(\xi)\,\mathbf{P}(\mathrm{d}\xi \mid \theta) = \overline{h} + 0. \tag{C.10}$$

Therefore, if $h(\xi)$ is harmonic to $\mathbf{P}(\mathrm{d}\xi \mid \theta)$, then it is a constant $h(\xi) = \overline{h}$ for all $\xi \in \mathcal{X}$. Hence as discussed in Appendix A, by [41, Theorem 2], the Markov chain $\mathbf{P}(\xi \mid \theta)$ is Harris recurrent. Application of Theorems A.1 and A.2 then completes the proof. □

D Proof of Theorem 6.1

PROOF. The proof which follows is drawn from original arguments in [19,35] which are below adapted to the specific problem settings and assumptions of this paper. To begin with, define the sets

$$R_\theta \triangleq \left\{\xi \in \mathcal{X} : \frac{p(\xi \mid Y)}{p(\theta \mid Y)} \cdot \frac{\gamma(\theta \mid \xi)}{\gamma(\xi \mid \theta)} < 1\right\}, \quad T_\theta \triangleq \mathcal{X}\backslash R_\theta \tag{D.1}$$

Then for any $\theta \in \Theta$, and with I denoting the indicator function (25)

$$\begin{aligned}K(\xi \mid \theta) \geq& \alpha(\xi \mid \theta)\gamma(\xi \mid \theta)\\ =& \frac{p(\xi \mid Y)}{p(\theta \mid Y)}\,\gamma(\theta \mid \xi)I_{R_\theta}(\xi) + \gamma(\xi \mid \theta)\,I_{T_\theta}(\xi)\\ \geq& \,\epsilon\, p(\xi \mid Y)I_{R_\theta}(\xi) + \epsilon\, p(\xi \mid Y)I_{T_\theta}(\xi) = \epsilon\, p(\xi \mid Y).\end{aligned}$$

Therefore,

$$r(\xi|\theta) \triangleq \frac{K(\xi|\theta) - \epsilon\, p(\xi \mid Y)}{1 - \epsilon}$$

is a bona-fide probability density function. Making the transition density $K(\xi \mid \theta)$ the subject of the above equation allows it to be written in 'split' form as

$$K(\xi \mid \theta) = \epsilon\, p(\xi \mid Y) + (1 - \epsilon)r(\xi \mid \theta). \tag{D.2}$$

This provides an alternate means for drawing realisations from the Markov chain with transition probability $K(\xi \mid \theta)$. Namely

```
if $\delta_k \sim \mathrm{Ber}(\epsilon) = 0$ then
  $\theta_{k+1} \sim r(\cdot \mid \theta_k)$
else
  $\theta_{k+1} \sim p(\cdot \mid Y)$.
end if
```

where Ber(ϵ) denotes a Bernoulli density that delivers a 1 with probability ϵ, and 0 with probability $1-\epsilon$. Consider now an additional chain $\{\beta_k\}$, which is initialised by drawing from the stationary distribution of the chain

$$\beta_0 \sim p(\cdot \mid Y) \tag{D.3}$$

and is propagated by the above alternate means for realising $K(\xi \mid \theta)$ according to

```
if $\delta_k \sim \mathrm{Ber}(\epsilon) = 0$ then
  $\theta_{k+1} \sim r(\cdot \mid \theta_k), \quad \beta_{k+1} \sim r(\cdot \mid \beta_k)$
else
  $\theta_{k+1} = \beta_{k+1} \sim p(\cdot \mid Y)$ .
```

```
end if
```

and once the algorithm reaches $\theta_{k+1} = \beta_{k+1} \sim p(\cdot \mid Y)$, then all all future draws are made so that $\theta_k = \beta_k$ is preserved.

The essential point is that since the chain $\{\beta_k\}$ is initialised from the stationary distribution, then for any set $A \in \mathcal{X}$ and any $k \geq 0$

$$p(\beta_k \in A) = \int_A p(\xi \mid Y)\,\mathrm{d}\xi = \varphi(A) \tag{D.4}$$

and hence, since the drawings of θ_{k+1} and β_{k+1} from $r(\cdot \mid \cdot)$ above are independent

$$\begin{aligned}
&|\mathbf{P}^n(A \mid \theta_0) - \varphi(A)| = |p(\theta_n \in A) - p(\beta_n \in A)| \\
&= |p(\theta_n \in A, \theta_n = \beta_n) + p(\theta_n \in A, \theta_n \neq \beta_n) \\
&- p(\beta_n \in A, \theta_n = \beta_n) - p(\beta_n \in A, \theta_n \neq \beta_n)| \\
&= |p(\theta_n \in A, \theta_n \neq \beta_n) - p(\beta_n \in A, \theta_n \neq \beta_n)| \\
&\leq \max\{p(\theta_n \in A, \theta_n \neq \beta_n), p(\beta_n \in A, \theta_n \neq \beta_n)\} \\
&\leq p(\theta_n \neq \beta_n) \leq p(T > n)
\end{aligned}$$

where T is the 'coupling time' which is defined to be the random time at which $\{\theta_k\}$ and $\{\beta_k\}$ come together. However, since the drawings of δ_k are independently $\text{Ber}(\epsilon)$, then $p(T = n) = \epsilon(1-\epsilon)^{n-1}$ so that $P(T > n) = (1-\epsilon)^n$. □

References

[1] B. Anderson and J. Moore, *Optimal Filtering*, Prentice Hall, 1979.

[2] K. Åström, *Maximum likelihood and prediction error methods*, Automatica, 16 (1980), pp. 551–574.

[3] S. Bittanti and M. Lovera, *Bootstrap-based estimates of uncertainty in subspace identification methods*, Automatica J. IFAC, 36 (2000), pp. 1605–1615.

[4] P. Caines, *Linear Stochastic Systems*, John Wiley and Sons, New York, 1988.

[5] M. C. Campi and P. R. Kumar, *Learning dynamical systems in a stationary environment*, Systems Control Lett., 34 (1998), pp. 125–132. Learning theory.

[6] M. C. Campi and M. Vidyasagar, *Learning with prior information*, IEEE Trans. Automat. Control, 46 (2001), pp. 1682–1695.

[7] M. C. Campi and E. Weyer, *Guaranteed non-asymptotic confidence regions in system identification*, Automatica, 41 (2005), pp. 1751–1764.

[8] P. Djurić and S.J. Godsill (Guest Editors), *Special issue on Monte Carlo methods for statistical signal processing*, IEEE Transactions on Signal Processing, 50 (2002).

[9] S. G. Douma and P. M.J. Van den Hof, *Probabilistic model uncertainty bounding:An approach with finite-time perspectives*, in Preprints of the 14th IFAC Symposium on System Identification, 2006, pp. 1021–1026.

[10] W. J. Dunstan and R. R. Bitmead, *Empirical estimation of parameter distributions in system identification*, in Proceedings of the 13th IFAC Symposium on System Identificatin, The Netherlands, 2003.

[11] B. Efron, *Bayesians, frequentists and scientists*, Am. Stat. Assoc. Presidential Address. Available electroncially via `http://www-stat.stanford.edu/ brad/papers/`, (2004).

[12] B. Efron and R. Tibshirani, *Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy*, Statist. Sci., 1 (1986), pp. 54–77. With a comment by J. A. Hartigan and a rejoinder by the authors.

[13] S. Garatti, M. C. Campi, and S. Bittanti, *Assessing the quality of identified models through the asymptotic theory—when is the result reliable?*, Automatica J. IFAC, 40 (2004), pp. 1319–1332.

[14] W. Gilks, S. Richardson, and D. Spiegelhalter, *Markov Chain Monte Carlo in practice*, Chapman and Hall, 1996.

[15] J. E. Handschin and D. Q. Mayne, *Monte Carlo techniques to estimate the conditional expectation in multi-stage nonlinear filtering*, Internat. J. Control (1), 9 (1969), pp. 547–559.

[16] W. Hastings, *Monte Carlo sampling methods using Markov chains and their applications*, Biometrika, 57 (1970), pp. 97–109.

[17] J.Dongarra and F. (eds), *The top ten algorithms - The Metropolis algorithm*, Computing in Science and Engineering, 2 (2000), pp. 65–69.

[18] R. Jeffrey, *Subjective Probability:The real thing*, Cambridge University Press, 2004.

[19] G. L. Jones and J. P. Hobert, *Honest exploration of intractable probability distributions via Markov chain Monte Carlo*, Statist. Sci., 16 (2001), pp. 312–334.

[20] A. L. Juloski, S. Weiland, and W. P. M. H. Heemels, *A Bayesian approach to identification of hybrid systems*, IEEE Trans. Automat. Control, 50 (2005), pp. 1520–1533.

[21] E. Lehmann, *Theory of Point Estimation*, John Wiley & Sons, 1983.

[22] D. Lindley, *Bayesian Statistics:A Review*, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, 1971.

[23] L. Ljung, *System Identification: Theory for the User, (2nd edition)*, Prentice-Hall, Inc., New Jersey, 1999.

[24] ———, *MATLAB System Identification Toolbox Users Guide, Version 6*, The Mathworks, 2004.

[25] M. Metropolis, A. Rosenbluth, M. Rosenbluth, A. Teller, and E. Teller, *Equations of state calculations by fast computing machines*, Journal of Chemical Physics, 21 (1953), pp. 1087–1091.

[26] S. P. Meyn and R. L. Tweedie, *Markov chains and stochastic stability*, Springer-Verlag, London, 1993.

[27] M. Milanese and A. Vicino, *Optimal inner bounds of feasible parameter set in linear estimation with bounded noise*, IEEE Transactions on Automatic Control, 36 (1991), p. 759.

[28] ———, *Information based complexity and nonparametric worst-case system identification*, Journal of Complexity, 9 (1993), pp. 427–446.

[29] B. Ninness, *Strong laws of large numbers under weak assumptions with application.*, IEEE Trans. Automatic Control, 45 (2000), pp. 2117–2122.

[30] B. Ninness and S. Henriksen, *A computational Bayesian approach to system identification*, in Proceedings 13th IFAC Symposium on System Identification, Rotterdam, August 2003.

[31] J. Norton, *Identification and application of bounded parameter models*, Automatica, 23 (1987), pp. 497–507.

[32] ———, *Identification of parameter bounds of ARMAX models from records with bounded noises*, International Journal of Control, 42 (1987), pp. 375–390.

[33] E. Nummelin, *General irreducible Markov chains and nonnegative operators*, Cambridge University Press, Cambridge, 1984.

[34] V. Peterka, *Bayesian system identification*, Automatica—J. IFAC, 17 (1981), pp. 41–53.

[35] C. Robert and G. Casella, *Monte Carlo Statistical Methods*, Springer, 1999.

[36] C. P. Robert, *The Bayesian Choice*, Springer Verlag, 2 ed., 2001.

[37] G. O. Roberts and J. S. Rosenthal, *Optimal scaling for various metropolis–hastings algorithms*, Statistical Science, 16 (2001), pp. 351–367.

[38] T. Schön and F. Gustafsson, *Particle filters for system identification of state-space models linear in either parameters or states*, in Proceedings of the 13th IFAC Symposium on System Identification, Rotterdam, The Netherlands, Sep 2003, pp. 1287–1292.

[39] B. W. Silverman, *Density estimation for statistics and data analysis*, Monographs on Statistics and Applied Probability, Chapman & Hall, London, 1986.

[40] J. Spall, *Estimation via Markov Chain Monte–Carlo*, IEEE Control Systems Magazine, 23 (2003), pp. 34–45.

[41] L. Tierney, *Markov chains for exploring posterior distributions*, Ann. Statist., 22 (1994), pp. 1701–1762. With discussion and a rejoinder by the author.

[42] F. Tjärnström and L. Ljung, *Estimating the variance in case of undermodeling using bootstrap*, IEEE Trans. Automatic Control, AC-47 (2002), pp. 395–398.

[43] J.-Y. Tournerat and Olivier Cappe (Guest Editors), *Special issue on Markov Chain Monte Carlo (MCMC) methods for signal processing*, Signal Processing, 81 (2001).

[44] T.Söderström and P.Stoica, *System Identification*, Prentice Hall, New York, 1989.

[45] M. Vidyasagar, *A Theory of Learning and Generalization*, Springer-Verlag, 1997.

[46] E. Walter and H.Piet-Lahanier, *Exact recursive polyhedral description of the feasible parameter set for bounded-error models*, IEEE Transactions on Automatic Control, AC-34 (1989), pp. 911–914.

[47] E. Weyer, *Finite sample properties of system identification of ARX models under mixing conditions*, Automatica J. IFAC, 36 (2000), pp. 1291–1299.

[48] E. Weyer, R. C. Williamson, and I. M. Y. Mareels, *Finite sample properties of linear model identification*, IEEE Trans. Automat. Control, 44 (1999), pp. 1370–1383.