Effect of Stochastic Model Error on the Convergence and Accuracy of Markov Chains

by

Mandy Yao

B.Sc., McGill University, 2020

Project Submitted in Partial Fulfillment of the
Requirements for the Degree of
Master of Science

in the
Department of Statistics and Actuarial Science
Faculty of Science

SIMON FRASER UNIVERSITY
Summer 2022

Declaration of Committee

Name:	**Mandy Yao**
Degree:	**Master of Science**
Thesis title:	**Effect of Stochastic Model Error on the Convergence and Accuracy of Markov Chains**
Committee:	**Chair:** Derek Bingham Professor, Statistics and Actuarial Science
	Donald Estep Supervisor Professor, Statistics and Actuarial Science
	David Stenning Committee Member Assistant Professor, Statistics and Actuarial Science
	Liangliang Wang Committee Member Associate Professor, Statistics and Actuarial Science
	Derek Bingham Examiner Professor, Statistics and Actuarial Science

Abstract

We derive conditions that guarantee the stability of Markov Chains in the presence of stochastic model error. To do this, we adapt existing theory on the convergence of perturbed stable Markov Chains under roundoff error. We apply the results to Markov Chain Monte Carlo (MCMC) algorithms, which are widely used to construct a stochastic process whose limiting distribution is the unknown distribution of interest in a given problem. For example, they are used for Bayesian Calibration, which is the problem of determining a distribution for parameters in a physical model from noisy observations on the output of the model using a Bayesian approach. In practice, errors in the computer simulations that affect the MCMC samples, in particular having a significant effect on convergence and accuracy. Finally, we also model the error of perturbed MCMC samples using a time series model.

Keywords: Markov Chain Monte Carlo; Geometric ergodicity; Bayesian calibration; Self-Exciting Threshold AutoRegressive models; Stability of Markov Processes

Table of Contents

List of Figures

Chapter 1

Overview of the Problem

Roberts et al. study the convergence properties of Markov chains that are perturbed at each step by roundoff error [Roberts et al., 1998]. In particular, they study the preservation of geometric ergodicity under these small perturbations. We adapt these results to an error model in which perturbations are modelled as random samples from a probability distribution. The approach of modelling errors using random samples has been used in the past [Hennig et al., 2015]. It provides a way to model complex errors, which fits applications in which samples are computed from complex models that involve significant approximation.

We consider two kinds of perturbation. For each kind of perturbation, we adopt a probability model, in which perturbations are added to values by adding a random variable ϵ from a fixed probability space, e.g., $x \to x + \epsilon$. Under an iid assumption, we can represent the distribution of $x+\epsilon$ as a convolution of the distributions of x and ϵ. We apply this model in two cases. For the first case, we apply this to each sample x, and for the second case, we apply this to the coefficients of the probability transition matrix of the Markov chain. We adapt the theorems on stability of Markov chains under perturbation to find conditions under which convergence holds with these errors. In the "numerical investigation" section, we find that unbiased errors that are symmetric, mean zero, and have sufficiently small variance still allow convergence. These results verify the theorem.

Markov Chain Monte Carlo (MCMC) algorithms are useful for exploring a probability distribution ($\pi(\cdot)$) and for drawing inferences [Roberts et al., 1998]. However, they can only be used if the Markov Chain $P(x, \cdot)$ involved converges in distribution to $\pi(\cdot)$. Since computer simulations are used to implement the Markov chains, their convergence properties, convergence rate, and stationary distribution are affected due to the finite precision and range of computers, the used pseudo-randomness, the common use of algorithms that involve approximations, and the fact that computer simulations are approximate [Roberts et al., 1998]. As a result, we model the cumulative effect as the computer simulating a chain $\tilde{P}(x, \cdot)$ that is perturbed. Describing this by a general stochastic error model, we apply the results to MCMC algorithms to find conditions under which these algorithms still converge under this error model.

We also analyze the relationship between the original limiting distribution and the limiting distribution that we obtain after adding error, assuming that the chain converges. To do that, we fit time series models to the difference between the limiting distribution from a simulation with the added error and the mean of the limiting distributions of simulations without the added error. This gives insight into the effect of the perturbations.

Chapter 2

Application to Markov Chain Monte Carlo

2.1 MCMC Algorithms

We apply the theoretical result in this thesis to MCMC algorithms. MCMC is useful in Bayesian statistics because it is able to approximate a posterior distribution that is difficult to calculate directly. This is achieved by drawing samples from a proposal distribution subject to an acceptance criterion to ensure that we obtain samples from the correct stationary target distribution. These samples are correlated since each sample depends on (only) the previous one and they form a chain whose density approximates the target distribution [Chan, 2015].

2.1.1 Metropolis-Hastings Algorithm

The Metropolis-Hastings algorithm is a widely used version of MCMC. It is outlined below. For each step, we sample a proposed value by performing a random walk step (sampled from a Normal distribution) from the previous value in the Markov Chain. This proposed value is then subject to an acceptance criterion that is accepted when the density of the target distribution at the proposed value is larger than the density at the previous value. This criterion is based on the ratio of the density of the target distribution at these two values, so the integral in the denominator is cancelled out by division. The algorithm requires specification of the number of iterations n, the standard deviation σ of the Normal distribution being used to sample the proposed values, and a starting value s [Chan, 2015].

Algorithm 1 Metropolis-Hastings Algorithm

$x_0 \leftarrow s$
for i in $1 : n$ **do**
  $y \leftarrow x_{i-1} + v$ where $v \sim N(0, \sigma)$
  $\alpha \leftarrow min(1, \frac{\pi(y)}{\pi(x_{i-1})})$
  if $u < \alpha$ where $u \sim \text{Unif}(0, 1)$ **then**
    $x_i \leftarrow y$
  else
    $x_i \leftarrow x_{i-1}$
  end if
end for

2.2 Application to Bayesian Calibration

Bayesian Calibration applies to a computer model that connects input values (data and parameters) to output values (observations). It solves the problem of determining the distribution for parameters for the input of a physical model from noisy observations on the output of the model by using a Bayesian approach. MCMC is used to compute the Bayesian posterior on the parameters. We describe the Bayesian Calibration model using the notation from Kennedy and O'Hagan [2001].

The Bayesian Calibration model is written as

$$z_i = \xi(\mathbf{x}_i) + e_i = \rho\eta(\mathbf{x}_i, \theta) + \delta(\mathbf{x}_i) + e_i,$$

where z_i are the observations, $\xi(\cdot)$ is the true process, $\eta(\cdot, \cdot)$ is the computer model output, e_i is the observation error for the ith observation, ρ is an unknown regression parameter and $\delta(\cdot)$ is the model inadequacy function (which is independent of $\eta(\cdot, \cdot)$). The observation error e_i also includes the residual variability since the two sources of uncertainty are difficult to separate. We assume that $e_i \sim N(0, \lambda)$.

In the problem of Bayesian Calibration, there are two types of inputs into a physical model: calibration inputs and variable inputs. The calibration inputs $\theta = (\theta_1, ..., \theta_{q_2})$ are unknown and fixed for the observations used for the calibration and for the inputs of the process that we are predicting using the calibrated model. Variable inputs are known for the observations used for the calibration but may vary when we are using the calibrated model to make predictions.

For the computer model, we denote the variable inputs as $\mathbf{x} = (x_1, ..., x_{q_1})$ and the calibration inputs as $\mathbf{t} = (t_1, ..., t_{q_2})$ so that the output is given by $\eta(\mathbf{x}, \mathbf{t})$. $\mathbf{t}$ and θ are distinct since $\mathbf{t}$ is the known calibration inputs to the computer model and θ is the unknown calibration inputs for the process that we are predicting using the calibrated model. Before computing the computer model, we know all the inputs. After N computations of the

computer code, we obtain outputs $\mathbf{y} = (y_1, ..., y_N)^T$, where

$$y_j = \eta(\mathbf{x}_j^*, \mathbf{t}_j).$$

Here, $\mathbf{x}_j^*$ are the known variable inputs and $\mathbf{t}_j$ are the known calibration inputs.

For the calibration data, we have known variable inputs $\mathbf{x}_i$ that have errors associated with them. We obtain n observations $\mathbf{z} = (z_1, ..., z_n)^T$ where z_i is an observation of the true process $\xi(\mathbf{x}_i)$ at $\mathbf{x}_i$ with observation error e_i.

Finally, we obtain the full data $\mathbf{d}^T = (\mathbf{y}^T, \mathbf{z}^T)$, which comprises of the computer code outputs and the calibration data. In practice, N is likely to be much larger than n since it is usually much more expensive to gather more observations than to run more iterations of the computer model.

There are several possible sources of uncertainties in computer models that are important to address in Bayesian Calibration [Kennedy and O'Hagan, 2001]. First, there is model inadequacy which arises from the fact that no model can be perfect and we cannot perfectly predict output given an input using the model. It is defined as the difference between the true value of the process and the output of the model given the true input value. Second, there is residual variability which arises in situations where we can get different outputs for different runs even when the same inputs are fully specified each time. Finally, there is observation error which sometimes cannot be separated from residual variability in practice.

Gaussian processes are used to model computer code output and model inadequacy. If we define $f(\cdot)$ to be a function that maps an input $x \in \mathcal{X}$ to an output $y = f(x)$ then we can treat f as a random function in a Bayesian framework. We write x as a vector $\mathbf{x} = (x_1, x_2, ..., x_q)$ where q is the dimension of the input space $\mathcal{X}$. Then, we can represent prior knowledge about $f(\cdot)$ using a Gaussian process. If for every $n = 1, 2, 3, ...$ the joint distribution of $f(x_1), ..., f(x_n)$ is multivariate normal for all $x_1, ..., x_n \in \mathcal{X}$ then $f(\cdot)$ has a Gaussian process distribution. Then $f(\mathbf{x})$ is normally distributed for all $x \in \mathcal{X}$. The mean function of this distribution is defined as $m(\cdot)$, where $m(\mathbf{x}) = E\{f(\mathbf{x})\}$ and the covariance function is defined as $c(\cdot, \cdot)$, where $c(\mathbf{x}, \mathbf{x}') = \text{cov}\{f(\mathbf{x}), f(\mathbf{x}')\}$. We denote this situation, $f(\cdot)$, having a Gaussian process distribution with mean function $m(\cdot)$ and covariance function $c(\cdot, \cdot)$, by $f(\cdot) \sim N\{m(\cdot), c(\cdot, \cdot)\}$.

It can be useful to model $m(\cdot)$ and $c(\cdot, \cdot)$ hierarchically. For instance, we can approximate $m(\cdot)$ by a function of a class of shapes defined to be $\mathbf{h}(\cdot) = (h_1(\cdot), h_2(\cdot), ..., h_p(\cdot))^T$ which is a vector of functions over $\mathcal{X}$. Then we can use the linear model structure

$$m(\cdot) = \mathbf{h}(\cdot)^T \beta,$$

where $\beta = (\beta_1, \beta_2, ..., \beta_p)^T$ is a vector of unknown coefficients that are given prior distributions.

In practice, numerical errors in the simulations affect the MCMC samples and this can have a significant effect on convergence and accuracy. It is important to study this. To do so, we use some existing theory on the convergence of perturbed Markov Chains, which treats a "roundoff error" that arises from the finite precision of computer computation [Roberts et al., 1998].

Chapter 3

Preservation of Stability

3.1 Convergence of Perturbed Markov Chains

Roberts et al. [1998] study a "perturbed chain" $\tilde{P}(x, \cdot)$ defined by

$$\tilde{P}(x, \cdot) = P(x, h^{-1}(A)) \tag{3.1}$$

where $P(x, \cdot)$ is a family of transition probabilities for a Markov Chain on $\mathcal{X}$ (which is a measurable separable metric space with metric $\text{dist}(\cdot, \cdot)$), $h : \mathcal{X} \to \mathcal{X}$ is a "perturbation" function where $h(x)$ is close to x for each $x \in \mathcal{X}$, and A . Here, it is not $P(x, \cdot)$ that is perturbed, but the input to P is perturbed. As shown in equation (1), the perturbed chain is a composition of P with h^{-1}. The assumption of small computer errors is imposed by assuming that

$$\|h(x) - x\| \leq \delta, x \in \mathcal{X}, \tag{3.2}$$

where $\|x\|$ is the norm of $x \in \mathcal{X}$ and δ is some small constant.

Based on the Fundamental Limit Theorem for Ergodic Markov Chains, we require the Markov Chain to be ergodic, which is the case if it is irreducible, aperiodic, and positive recurrent. The reason is that for an ergodic Markov Chain $X_0, X_1, ...$, there exists a unique, positive, stationary distribution π, which is the limiting distribution of the chain. That is, $\pi_j = \lim_{n\to\infty} P_{ij}^n, \forall i, j$. This theorem is helpful because it allows an investigation of the impact of small errors on convergence by looking at the effect of perturbations on the coefficients of a matrix. A linear algebra proof for this theorem uses the Perron-Frobenius Theorem [Dobrow, 2016]:

Theorem 3.1.1. *Let M be a $k \times k$ positive matrix. Then*

1. *There is a positive real number λ^* which is an eigenvalue of M. For all other eigenvalues λ of M, $|\lambda| < \lambda^*$. The eigenvalue λ^* is called the Perron-Frobenius eigenvalue.*

2. *The eigenspace of eigenvectors associated with λ^* is one-dimensional.*

3. *There exists a positive right eigenvector v associated with λ^*, and a positive left eigenvector w associated with λ^*. Furthermore, $lim_{n\to\infty}\frac{1}{(\lambda^*)^n}M^n = vw^T$, where the eigenvectors are normalized so that $w^Tv = 1$.*

We actually require the stronger assumption that P is geometrically ergodic and the distribution $\pi(\cdot)$ is stationary, since Central Limit theorems are guaranteed to hold, which means that we can perform a better analysis on the performance of MCMC algorithms.

The Markov chain P is geometrically ergodic if it is Harris recurrent and if there exists a real number $r > 1$ such that

$$\sum_{n=1}^{\infty} r^n \left\| P^n(x,\cdot) - \pi(\cdot) \right| < \infty, x \in \mathcal{X}.$$

Recall that a set $A \in \mathcal{A}$ is ψ-positive when $\psi(A) > 0$, where ψ is a strictly positive measure on the state space $(\Omega, \mathcal{A})$, meaning $\psi(A) \geq 0$ for all $A \in \mathcal{A}$ and $\psi(\Omega) > 0$ [Geyer, 2020]. A Markov chain P is Harris recurrent if it is irreducible with maximal irreducibility measure ψ and for every ψ-positive set A the chain started at x hits A infinitely often with probability 1 [Geyer, 2020]. Harris recurrence is a stronger condition than standard recurrence since it guarantees that the chain returns to a state in A regardless of where it starts. In contrast, recurrence only ensures that the probability that the chain never returns to the initial state is zero.

We obtain an equivalent definition for geometric ergodicity using the following definition [Geyer, 2020]:

Definition 3.1.1 (ϕ-Irreducibility)**.** *A Markov chain P is ϕ-irreducible for some non-zero σ-finite measure ϕ if for each $x \in \mathcal{X}$, we have $\sum_n P^n(x, A) > 0$ whenever $\phi(A) > 0$*

An aperiodic, ϕ-irreducible Markov chain P is geometrically ergodic if there exists a π-a.e.-finite function $V : \mathcal{X} \to [1, \inf]$, a subset $C \subseteq \mathcal{X}$, and finite positive numbers β and b, such that the geometric drift condition

$$\Delta V(x) \leq -\beta V(x) + b\mathbb{1}_C(x), x \in \mathcal{X} \tag{3.3}$$

holds, where

$$\Delta V(x) \equiv PV(x) - V(x) \equiv \int V(y)P(x, dy) - V(x),$$

and where C is small for P, which means there is a non-zero measure v on $\mathcal{X}$ and a positive integer n_0 such that $P^{n_0}(x, A) \geq v(A)$ for all $x \in C$ and $A \subseteq \mathcal{X}$ [Roberts et al., 1998]. Here, we are looking for a geometric drift towards C. V is the drift function, which helps to impose directions on the set Ω, which does not have properties for direction [Geyer, 2020]. When studying transience and recurrence, we do not have much to work with without direction on the state space. For example, a chain is said to be transient when roughly speaking,

it moves off to infinity [Geyer, 2020]. But, in our state space, there is no direction toward infinity. To address this, the drift function compares the functions V and PV, where

$$(PV)(x) = \int P(x, dy)V(y) = E\{V(X_{n+1})|X_n = x.\}$$

Returning to the perturbation, it has been shown that when $\tilde{P}$ is close enough to P in a certain V-related sense, the property of geometric ergodicity is robust [Roberts et al., 1998]. This means that geometrically ergodic Markov chains remain geometrically ergodic after perturbation. This is summarized in the following theorem [Roberts et al., 1998]:

Theorem 3.1.2. *Let P be a geometrically ergodic Markov chain on $\mathcal{X}$, and let V and $\beta > 0$ satisfy (3.3) for P, for some small set C and $0 < b < \infty$. Let $\tilde{P}$ be a second Markov chain on $\mathcal{X}$, given by (3.1) for some roundoff function $h : \mathcal{X} \to \mathcal{X}$, and assume that $|\tilde{P}V - PV| < \delta V$ for some $\delta < \beta$. Then $\tilde{P}$ is geometrically ergodic.*

In the proof for this theorem, we use the fact that $|\tilde{P}V - PV| < \delta V$ for some $\delta < \beta$.

Another theorem provides another condition under which geometric ergodicity is preserved when roundoff error is small [Roberts et al., 1998]:

Theorem 3.1.3. *Let P be geometrically ergodic on $\mathcal{X}$, and let it satisfy (3.3) for some small set C and continuous function V, such that logV (or V) is uniformly continuous on $\mathcal{X}$. Then there is $\delta > 0$ such that, if $\tilde{P}$ is given by (3.1) with $M_h < \delta$, then $\tilde{P}$ is geometrically ergodic.*

Here, $M_h = \sup_{x \in \mathcal{X}} \text{dist}(x, h(x))$. This theorem is useful in practice since its condition for geometric ergodicity is weaker and easier to establish.

3.1.1 Example 1

For this example (taken from Geyer [2020]), we consider an AR(1) process

$$X_n = \phi X_{n-1} + \sigma Y_n \text{ for } n \in \mathbb{N}$$

where $\{Y_n\}$ is a sequence of iid random variables that have a standard normal distribution and are independent of X_0, $\sigma^2 < \infty$, and $\phi^2 < 1$.

We use the following theorem, which is a consequence of the geometric drift condition (3.3) Geyer [2020]:

Theorem 3.1.4. *The drift condition holds with a petite set C if and only if V is unbounded on non-petite sets and*

$$PV \leq \lambda V + L \tag{3.4}$$

for some $\lambda < 1, L < \infty$.

A subset C of the state space (Ω, A) is petite if there exists a nonzero positive measure ν on the state space and a subsampling distribution q such that

$$P_q(x, A) \geq \nu(A),\ A \in \mathcal{A} \text{ and } x \in C.$$

A subsampling distribution q is a nonnegative-integer valued random variable where

$$P_q(x, A) = \sum_{n=0}^{\infty} q_n P^n(x, A),$$

where P is a Markov kernel [Geyer, 2020]. P_q is also a Markov kernel, which gives a Markov chain that is created using the original chain by using subsampling. To check that V is unbounded on non-petite sets, we use the following theorem [Meyn and Tweedie, 1994]:

Theorem 3.1.5. *If Φ is a ψ-irreducible T-chain then every compact set is petite.*

A T-chain is a Markov chain that has a continuous component T such that $T(x, \Omega) > 0$ for all $x \in \mathbb{R}$, where the continuous component T of a kernel P having state $(\Omega, \mathcal{A})$ is a substochastic kernel such that the function $x \mapsto T(x, A)$ is lower semicontinuous (LSC) for any $A \in \mathcal{A}$ and there is a probability vector q, such that $P_q(x, A) \geq T(x, A), x \in A$ and $A \in \mathcal{A}$ [Geyer, 2020]. A kernel is substochastic if all its values are nonnegative and $K(x, \Omega) \leq 1, x \in \Omega$, and a function f is LSC if $\lim_{y \to x} \inf f(y) \geq f(x)\ \ \forall x$ [Geyer, 2020].

The AR(1) process is a T-chain since the conditional distribution of X_{n+1} given X_n is normal. Therefore the conditional probability density function for X_{n+1} given X_n is a continuous function of X_n and by Theorem 3.1.5, every compact set is petite. Then, if drift function V is defined $V(x) = 1 + x^2$, it is unbounded off petite sets. We have,

$$(PV)(x) = E(V(X_{n+1}|X_n = x) = E(1 + X_{n+1}^2|X_n = x) = 1 + \phi^2 x^2 + \sigma^2,$$

since $E(X_{n+1}|X_n = x) = \phi x$. This satisfies (3.4) with $\lambda = \phi^2$ and $L = 1 - \phi^2 + \sigma^2$. By theorem 3.1.4, the AR(1) process with $\phi^2 < 1$ is geometrically ergodic. In addition, if we have a perturbed chain $\tilde{P}$ given by (3.1), and $|\tilde{P}V - PV| < \delta V$ for some $\delta < \beta$ (from the geometric drift condition), then $\tilde{P}$ is also geometrically ergodic by Theorem 3.1.2.

3.1.2 Geometric Ergodicity and Floating-Point Roundoff Error

Breyer et al. [2001] also consider the case where computer errors are bounded by

$$\|h(x) - x\| \leq \delta \|x\|, x \in \mathcal{X}, \tag{3.5}$$

which provides for a nonuniform error.

For example, when we use floating point representations in computers, numbers (x) could be encoded as

$$x := \sigma \cdot (1 + k/2^N) \cdot 2^e,$$

where σ is the sign with 1 bit, k is the fractional part with N bits, and e is the exponent with M bits [Breyer et al., 2001]. The computer could be approximating x as

$$h(x) = \sigma \left\lfloor \frac{1}{2} + |x| 2^{N-1-e} \right\rfloor 2^{-(N-1-e)}.$$

so that

$$|h(x) - x| \leq 2^{-(N-1-e)} \leq 2^{-(N-1)}|x|.$$

This satisfies assumption (3.5) with $\delta = 2^{-(N-1)}$.

In order for geometric ergodicity to hold for these types of perturbations, we need new conditions. If we have a geometrically ergodic Markov chain P, with $\tilde{P}$ obtained using (3.1) for some roundoff function h satisfying (3.5) for some $\delta > 0$, we assume that the drift function V satisfies

$$V(y+u) - V(y) \leq \delta K V(y),\ \|u\| \leq \delta \|y\|,\ y \in \mathcal{X}, \tag{3.6}$$

for some $K < \infty$ [Breyer et al., 2001]. This condition requires that the gradient of log(V) decays sufficiently quickly. It can be shown that when (3.6), (3.3) and (3.5) hold with $\tilde{P}$ defined in (3.1), then geometric ergodicity holds provided that

$$\delta < K^{-1}(\lambda^{-1} - 1). \tag{3.7}$$

Then we have the following theorem [Breyer et al., 2001]:

Theorem 3.1.6. *Suppose a Markov chain P is geometrically ergodic and satisfies (3.3) for some V and C, where V satisfies (3.6) for some $K < \infty$. Suppose further that $\tilde{P}$ is obtained via (3.1), for some perturbation function h satisfying (3.5) and (3.7). Then $\tilde{P}$ is also geometrically ergodic.*

This theorem shows that when the drift function V satisfies a smoothness condition (3.6), geometric ergodicity is preserved under perturbations with magnitude proportional to that of x that satisfy (3.5). It is worthwhile to note that while the property of geometric ergodicity is preserved, we generally cannot expect the perturbed chain to have the same limiting distribution as the unperturbed matrix P.

The following theorem is useful since it connects Theorem 3.1.3 with Theorem 3.1.6 [Breyer et al., 2001]:

Theorem 3.1.7. *Suppose that $\tilde{X}$ is a perturbation of the state X with approximation function h. Let $\psi : \mathcal{X} \to \mathcal{X}'$ be a bi-measurable bijection. We also define a Markov chain $\{X_t^\psi\}$ on*

$\mathcal{X}'$ by $X_t^\psi = \psi(X_t)$, with corresponding transition kernel P^ψ. Then $\psi(\tilde{X})$ is a perturbation of X^ψ with approximation function $h^\psi \equiv \psi h \psi^{-1}$. In the special case that $\mathcal{X} = \mathbb{R}$, $\mathcal{X}' = \mathbb{R}^+$, and $\psi(x) = \exp(x)$, then h satisfies (3.2) if and only if $h^{\exp}$ satisfies (3.5). In that case, if P^{exp} is geometrically ergodic, then it is robust to perturbations of the kind satisfying (3.5) provided it has a drift function $V^{\exp}$ satisfying that $\log V^{\exp}(\exp(\cdot))$ is uniformly continuous.

3.1.3 Example 2

For this example taken from Breyer et al. [2001], suppose $\mathcal{X} = \mathbb{R}$ and $V(x) = C_1|x|^n + C_2$, with $C_1, C_2 \geq 0$. Then for $\delta' \leq \delta$,

$$V(y + \delta' y) - V(y) = C_1(|y + \delta' y|^n - |y|^n) \leq C_1(1 + \delta')^n |y|^n \leq (1 + \delta)^n V(y).$$

From this, we know that (3.6) holds with $K = (1 + \delta)^n/\delta$. Therefore, by Theorem 3.1.6, when a geometrically ergodic Markov chain has drift function $C_1|x|^n + C_2$, sufficiently small perturbations of the form (3.5) preserve geometric ergodicity.

3.1.4 Example 3

For this example taken from Breyer et al. [2001], we recall that the Gaussian Random Walk Metropolis algorithm uses a random walk proposal $Y_t = X_t + \sigma Z_t$ with Gaussian increment $Z_t \sim N(0, 1)$ with Gaussian target distribution $\pi(x) \propto \exp(-x^2/2)$, and Gaussian proposal increment with fixed standard deviation $\sigma = 1$. We write the Markov chain update $X_{t+1} = F_t(X_t)$ explicitly:

$$F_t(x) = \begin{cases} x + Z_t & \exp(x^2/2 - (x + Z_t)^2/2) > \xi_t \\ x & \text{otherwise,} \end{cases}$$

where $\xi_t \sim \text{Unif}[0, 1]$.

With the drift function $V(x) = \exp(|x|)$, this chain is geometrically ergodic [Breyer et al., 2001]. It is also robust to perturbations that satisfy (3.2). However, if the perturbations satisfy (3.5), the chain is no longer geometrically ergodic. For example, consider the roundoff function,

$$h(x) = \text{sign}(x) \cdot 2^{\lfloor \log_2 |x| \rfloor} \cdot (1 + 2^{-52} \lfloor 2^{52} (|x| 2^{-\lfloor \log_2 |x| \rfloor} - 1) \rfloor),$$

which maps to points whose spacing become larger as the inputs get further from zero. This means that at very large points, the chain can get "stuck" since the proposal chain is a random walk. Therefore, it cannot be geometrically ergodic.

Instead, we apply Theorem 3.1.7 and use a transformation in order to verify geometric ergodicity. If we have a Markov chain $X_t^\psi = \psi(X_t)$, where $\psi(x) = \exp(x)$ and $V^\psi(x) = V(\psi^{-1}(x)) = \exp(|\log x|) = \max\{x, x^{-1}\}$, then geometric ergodicity holds since the drift function is $V(x) = \exp(x)$ and $\max\{x, x^{-1}\}$ is uniformly continuous. The chain is then

written as $\tilde{X}_{t+1} = F^{\psi}(\tilde{X}_t)$, where $F^{\psi} = \psi \circ F \circ \psi^{-1}$ is given by

$$F^{\psi}(x') = \begin{cases} x'e^{Z_t} & -Z_t^2 - 2Z_t \log(x') > 2\log \xi_t \\ x' & \text{otherwise,} \end{cases}$$

so $\log \tilde{X}_t$ converges in distribution to the stationary target distribution π. We can invert the transformation in order to recover the target distribution.

3.2 A Stochastic Error Model

3.2.1 Two Ways to Introduce Perturbations

We consider two ways to introduce perturbations. First, we interpret the effect of the round-off error. The variable x acts like an index in the probability transition matrix. When x is perturbed to h(x), the index of the probability transition matrix changes so that we are looking at the probability of a perturbed value. Thus, $\pi(A)$ becomes $\pi(h(A)) = \tilde{\pi}(A)$ by having $P(x, A)\tilde{\pi}(A) = P(x, h^{-1}(A))\pi(A)$. In this case, the h function needs to be close to identity for $\tilde{P}$ to preserve geometric ergodicity, since $\tilde{P}$ needs to be sufficiently close to P for Theorem 3.1.2 to hold and $M_h = \sup_{x \in \mathcal{X}} \text{dist}(x, h(x))$ needs to be sufficiently small for Theorem 3.1.3 to hold.

In both cases, we add an error ϵ to quantity x, where ϵ is i.i.d. and independent from x. The distribution of the resulting variable $x + \epsilon$ is given by the convolution of the distribution of x and of ϵ. Therefore, we model the error with a convolution. The error can be written $\epsilon \sim (E, B, \mu)$ on a finite interval $E \subset \mathbb{R}$, where B is the Borel sigma algebra and μ is the probability distribution. $\tilde{\pi} = \mu * \pi$, so $P(x)\pi$ becomes $P(x)\tilde{\pi}$ using the properties of convolution.

3.2.2 Perturbation 1

The first perturbation we consider is the case when the distributions on state x are perturbed. Here, we take the distribution on x, take a convolution of the error distribution, and apply the matrix P.

We add the error so that $\pi_x P_x$ becomes $\pi_{x+\epsilon} P_x$. This means that there is a new distribution at every x, namely which can be written $\mu * \pi = C_x \pi$ for a matrix C_x since convolution can be written as matrix multiplication in the discrete case. So $P(x)\tilde{\pi} = P(x)C_x\pi$, with $\tilde{P}(x) = P(x)C_x$. Then we write $P(x)\tilde{\pi} = P(x)\mu * \pi = M * P(x)\mu$ because of the commutativity of convolution.

3.2.3 Perturbation 2

For the second perturbation, we consider the matrix $\tilde{P}$ obtained by taking the convolution of each coefficient with the error distribution μ, which is equivalent to adding an error ϵ to each coefficient of P.

We describe how to adapt the roundoff error method of Roberts et al. [1998]. $\pi_x P_x$ depends on x, and with the perturbation, $\pi_x P_x$ becomes $\pi_x \tilde{P}_x = \pi_x P_{h^{-1}(x)}$. It can be shown that this is equivalent to the first perturbation, where we obtain $\tilde{P}\pi$ by taking $\tilde{P}\pi = P\mu * \pi$. It is possible to do a convolution on P instead of on π because of the commutativity of convolution:

$$\mu * \pi = \int \mu(x-\epsilon)\pi(\epsilon)d\epsilon = \int \mu(\epsilon)\pi(x-\epsilon)d\epsilon.$$

This means that

$$P(x)\mu * \pi = P(x)\int \mu(x-\epsilon)\pi(\epsilon)d\epsilon$$

can be written as

$$\int P(x)\mu(x-\epsilon)\pi(\epsilon)d\epsilon = \int \mu(\epsilon)P(x)\pi(x-\epsilon)d\epsilon,$$

where $\tilde{P}(x) = \mu(\epsilon) * P(x)$. Now we can use the roundoff error results.

We write

$$\pi_{x+\epsilon} = \tilde{\pi}(x) = \int \mu(x-\epsilon)\pi(\epsilon)d\epsilon,$$

where we expand

$$\mu(x) - \mu'(x)\epsilon + \frac{1}{2}\mu''(x)\epsilon^2 + ...$$

Then $\tilde{\pi}(x)$ can be rewritten

$$\pi(x) - \mu'(x)\int \epsilon\pi(\epsilon)d\epsilon + \frac{1}{2}\mu''(x)\int \epsilon^2\pi(\epsilon)d\epsilon + ...$$

In this form, we see that the perturbation can be represented as a roundoff error, defined by keeping the first term of the Taylor expansion and dropping the higher order terms. We define a roundoff function h(x) for this by using the Implicit Function Theorem. From the Taylor expansion, we also see that we need kernels with bounded support.

3.3 Convergence Analysis for the Stochastic Error

We set up some notation taken from Boomgaard and Dorst [2021]. We use the pulse function δ that is zero everywhere except at zero, and satisfies

$$f(x) = \int_{\mathbb{R}^d} f(y)\delta(x-y)dy$$

for all functions f (which do not map to the empty set). In the discrete case, the discrete pulse function is defined as

$$\Delta(k) = \begin{cases} 1 & k = 0 \\ 0 & \text{elsewhere} \end{cases}.$$

We let T denote a translation over a vector x as

$$Tf(x) = T_t f(x) = f(x - t),$$

for all functions f. Then L is a linear translation invariant operator

$$L(T_t f) = T_t(Lf).$$

The convolution of a function f with a function ω can be written as a linear translation invariant operator L working on a function f:

$$(Lf)(x) = (f * w)(x) = \int_{\mathbb{R}^d} f(x - y)\omega(y)dy,$$

where $\omega = L\delta$.

In the discrete case, the convolution of a function f with a function W can be written as a linear translation invariant operator L working on a function F,

$$(LF)(k) = (F * W)(k) = \sum_l F(k - l)W(l),$$

where $W = L\Delta$.

Our error model is written by denoting the perturbed matrix $\tilde{P} = L_\epsilon P$. Then, we write

$$\int V(y)L_\epsilon P(x, dy) = \int V(y)\mu * P(x, dy).$$

By the associativity and commutativity of convolution, this can be rewritten as

$$\int (\mu * V)P(x, dy).$$

From (3.3), we have that the geometric drift condition holds for finite positive numbers β and b and a subset $C \in \mathcal{X}$ if there exists a π-a.e.-finite function $V : \mathcal{X} \to [1, \infty]$ such that

$$\Delta V(X) \leq -\beta V(x) + b\mathbb{1}_C(x), x \in \mathcal{X},$$

where C is small for P and

$$\Delta V(x) \equiv PV(x) - V(x) \equiv \int V(y)P(x, dy) - V(x).$$

The drift condition can be rewritten as

$$PV(x) \leq \lambda V(x) + b\mathbb{1}_C(x), x \in \mathcal{X}.$$

where $\lambda < 1$.

With the perturbed $\tilde{P}$, the left side of the above inequality can be written as:

$$\begin{aligned}\tilde{P}V(x) &= PV(x) + (\tilde{P} - P)V(x) \\ &= PV(x) + \int (\mu * V(y) - V(y))P(x, dy) \\ &\leq \lambda V(x) + b\mathbb{1}_C(x) + \int (\mu * V(y) - V(y))P(x, dy)\end{aligned}$$

We want $\mu * V(y) - V(y)$ to be close to 0. So if we write $\mu * V(y) - V(y) = \eta V(y)$, where $0 < \eta < 1$, then we get from the above inequality,

$$\tilde{P}V(x) \leq (1 + \eta)(\lambda V(x) + b\mathbb{1}_C(x)).$$

In order to satisfy the geometric drift condition, we need $\lambda(1 + \eta) < 1$, or

$$\eta < \frac{1}{\lambda} - 1. \tag{3.8}$$

If we can bound $\mu * V(y)$ so

$$\mu * V(y) \leq \alpha V(y),$$

then from the inequality, we obtain

$$(\alpha - 1)V(y) \leq \eta V(y)$$

so we just need

$$\alpha < \frac{1}{\lambda} \tag{3.9}$$

to satisfy the geometric drift condition. This puts a condition on the error distributions μ that we can use to perturb P.

We can also follow the analysis for floating-point roundoff error in Section 3.1.2 and assume that V satisfies

$$\tilde{V} - V \leq \delta K V(y).$$

Then if the geometric drift condition holds,

$$\tilde{P}V(x) \leq (1 + \delta K)(\lambda V(x) + b\mathbb{1}_C(x).$$

The proof is given by:

$$\begin{aligned}
\tilde{P}V(x) &= PV(x) + (\tilde{P} - P)V(x) \\
&= PV(x) + \int (\tilde{V}(y) - V(y))P(x, dy) \\
&\leq \lambda V(x) + b\mathbb{1}_C(x) + \int \delta K V(y) P(x, dy) \\
&\leq \lambda V(x) + b\mathbb{1}_C(x) + \delta K(\lambda V(x) + b\mathbb{1}_C(x)) \\
&\leq (1 + \delta K)(\lambda V(x) + b\mathbb{1}_C(x)
\end{aligned}$$

Then we have geometric ergodicity when $(1 + \delta K)\lambda < 1$, or when

$$\delta < K^{-1}(\lambda^{-1} - 1). \tag{3.10}$$

The inequalities in equations 3.8, 3.9, and 3.10 give limits to the size of the perturbations for the preservation of convergence. We perform simulations in Section 3.5 to see the perturbation sizes at which convergence is preserved in a Bayesian Calibration example.

3.4 Application to Different Error Distributions

We apply the analysis to some distributions. We found that meeting the conditions of the theorem requires the distributions of the perturbations to be symmetric around a mean of 0. In the simulations below, we see that violating these conditions can destroy convergence.

3.4.1 Uniform Error

If we have an error that is sampled from a uniform distribution μ with a support from $-\epsilon$ to ϵ, then

$$\mu * V = \frac{1}{2\epsilon} \int_{-\epsilon}^{\epsilon} V(y - x)dx.$$

By the Mean Value Theorem,

$$\mu * V = V(\xi_y)\frac{2\epsilon}{2\epsilon} = V(\xi_y),$$

where $\xi_y \in [y - \epsilon, y + \epsilon]$.

Using the assumption on V, we obtain geometric ergodicity if

$$\tilde{V}(y) - V(y) = \mu * V(y) - V(y) = V(\xi_y) - V(y) \leq \epsilon K V(y).$$

This holds since $|\xi_y - y| \leq \epsilon$ and we can use Theorem 3.1.6.

3.4.2 Quadratic Density

Consider an error sampled from a distribution in the shape of a convex quadratic function $-x^2 + c$ with support from $-\epsilon$ to ϵ. Then we want

$$\int_{-\epsilon}^{\epsilon} (c - x^2)dx = \frac{-2(\epsilon^3 - 3c\epsilon)}{3} = 1$$

so $c = \frac{2\epsilon^2+3}{6\epsilon}$, so the quadratic function is $-x^2 + \frac{2\epsilon^2+3}{6\epsilon}$. Then

$$\begin{aligned}\mu * V(y) &= \int_{-\epsilon}^{\epsilon} (\frac{2\epsilon^2+3}{6\epsilon} - x^2)V(y-x)dx \\ &= (\frac{2\epsilon^2+3}{3} - 2\epsilon\bar{x}^2)V(\xi_y),\end{aligned}$$

by the Mean Value Theorem, where $\xi_y \in [y-\epsilon, y+\epsilon]$, $\bar{x} \in [-\epsilon, \epsilon]$.

This means we have geometric ergodicity if

$$\begin{aligned}\tilde{V}(y) - V(y) = \mu * V(y) - V(y) &= \frac{2}{3}\epsilon^2 V(\xi_y) - 2\epsilon\bar{x}^2 V(\xi_y) + V(\xi_y) - V(y) \\ &\leq \frac{2}{3}\epsilon^2 V(\xi_y) - 2\epsilon\bar{x}^2 V(\xi_y) - V(y) \leq \epsilon K V(y).\end{aligned}$$

We can write

$$\begin{aligned}\frac{2}{3}\epsilon^2 V(\xi_y) - 2\epsilon\bar{x}^2 V(\xi_y) - V(y) &= \frac{2}{3}\epsilon^2 V(\xi_y) - 2\epsilon\bar{x}^2 V(\xi_y) - V(\xi_y) + V(\xi_y) - V(y) \\ &= (\frac{2}{3}\epsilon^2 - 2\epsilon\bar{x}^2 - 1)V(\xi_y) + V(\xi_y) - V(y).\end{aligned}$$

Since the maximum of $\frac{2}{3}\epsilon^2 - 2\epsilon\bar{x}^2 - 1$ is at $\bar{x} = 0$, the above is less than or equal to

$$(\frac{2}{3}\epsilon^2 - 1)V(\xi_y) + V(\xi_y) - V(y).$$

Since $V(\xi_y) - V(y) \leq K\epsilon V(y)$, for geometric ergodicity, we need $\frac{2}{3}\epsilon^2 - 1 \leq 0$ or

$$\epsilon \leq \sqrt{\frac{3}{2}}.$$

3.4.3 Beta Density

In order to get a beta density that is centred at 0 (instead of the usual domain of (0,1)), we use the non-standard beta density on (a,b)

$$\mu(x) = \frac{(x-a)^{\alpha-1}(b-x)^{\beta-1}}{B(\alpha,\beta)(b-a)^{\alpha+\beta-1}},$$

where B is the beta function. We set $a = -\epsilon$ and $b = \epsilon$, and to obtain a density that is symmetric and unimodal, we set $\alpha = \beta > 1$. Then

$$\mu(x) = \frac{(x+\epsilon)^{\alpha-1}(\epsilon - x)^{\alpha-1}}{B(\alpha,\alpha)(2\epsilon)^{2\alpha-1}},$$

and

$$\begin{aligned}\mu * V &= \frac{1}{B(\alpha,\alpha)(2\epsilon)^{2\alpha-1}} \int_{-\epsilon}^{\epsilon} (x+\epsilon)^{\alpha-1}(\epsilon - x)^{\alpha-1} V(y-x)dx \\ &= \frac{1}{B(\alpha,\alpha)(2\epsilon)^{2\alpha-1}} V(\xi_y) \int_{-\epsilon}^{\epsilon} (x+\epsilon)^{\alpha-1}(\epsilon - x)^{\alpha-1} dx\end{aligned}$$

by the special Mean Value Theorem, where $\xi_y \in [y-\epsilon, y+\epsilon]$ and we use the fact that $(x+\epsilon)^{\alpha-1}(\epsilon - x)^{\alpha-1} \geq 0$. The integral

$$\begin{aligned}I &= \int_{-\epsilon}^{\epsilon} (x+\epsilon)^{\alpha-1}(\epsilon - x)^{\alpha-1} dx \\ &= \epsilon^{2\alpha-2} \int_{-\epsilon}^{\epsilon} (1+\frac{x}{\epsilon})^{\alpha-1}(1-\frac{x}{\epsilon})^{\alpha-1} dx.\end{aligned}$$

Using the change of variables $y = \frac{x}{\epsilon}, dx = dy \cdot \epsilon$, we obtain

$$\begin{aligned}I &= \epsilon^{2\alpha-1} \int_{-1}^{1} (1+y)^{\alpha-1}(1-y)^{\alpha-1} dy \\ &= \epsilon^{2\alpha-1} \int_{-1}^{1} (1-y^2)^j dy,\end{aligned}$$

where we let $j = \alpha - 1$. Then

$$\begin{aligned}I &= \epsilon^{2\alpha-1} \frac{\sqrt{\pi}\Gamma(j+1)}{\Gamma(j+\frac{3}{2})} \\ &= \epsilon^{2\alpha-1} \frac{\sqrt{\pi}\Gamma(\alpha)}{\Gamma(\alpha+\frac{1}{2})} \\ &= \epsilon^{2\alpha-1} \frac{\sqrt{\pi}\Gamma(\alpha)(\alpha!)4^{\alpha}}{(2\alpha)!\sqrt{\pi}},\end{aligned}$$

since $\Gamma(\alpha + \frac{1}{2}) = \frac{(2\alpha)!\sqrt{\pi}}{(\alpha!)4^{\alpha}}$.

We use the fact that $B(\alpha,\alpha) = \frac{\Gamma(\alpha)\Gamma(\alpha)}{\Gamma(2\alpha)}$ to write

$$\begin{aligned}
\mu * V &= \frac{\Gamma(2\alpha)}{\Gamma(\alpha)\Gamma(\alpha)(2\epsilon)^{2\alpha-1}} V(\xi_y)\epsilon^{2\alpha-1}\frac{\sqrt{\pi}\Gamma(\alpha)(\alpha!)4^{\alpha}}{(2\alpha)!\sqrt{\pi}} \\
&= \frac{\Gamma(2\alpha)2(\alpha!)}{\Gamma(\alpha)(2\alpha)!}V(\xi_y) \\
&= \frac{(2\alpha-1)!2(\alpha!)}{(\alpha-1)!(2\alpha)!}V(\xi_y) \\
&= \frac{2\alpha}{2\alpha}V(\xi_y) \\
&= V(\xi_y).
\end{aligned}$$

This means that we obtain geometric ergodicity if

$$\tilde{V}(y) - V(y) = \mu * V(y) - V(y) = V(\xi_y) - V(y) \leq \epsilon K V(y).$$

This holds since $|\xi_y - y| \leq \epsilon$, and we can apply Theorem 3.1.6.
Now if we allow the beta density to be skewed, so $\alpha \neq \beta$ and $\alpha, \beta > 1$, then

$$\mu(x) = \frac{(x+\epsilon)^{\alpha-1}(\epsilon - x)^{\beta-1}}{B(\alpha,\beta)(2\epsilon)^{\alpha+\beta-1}},$$

so

$$\begin{aligned}
\mu * V &= \frac{1}{B(\alpha,\beta)(2\epsilon)^{\alpha+\beta-1}}\int_{-\epsilon}^{\epsilon}(x+\epsilon)^{\alpha-1}(\epsilon-x)^{\beta-1}V(y-x)dx \\
&= \frac{1}{B(\alpha,\beta)(2\epsilon)^{\alpha+\beta-1}}V(\xi_y)\int_{-\epsilon}^{\epsilon}(x+\epsilon)^{\alpha-1}(\epsilon-x)^{\beta-1}dx
\end{aligned}$$

by the special Mean Value Theorem, where $\xi_y \in [y-\epsilon, y+\epsilon]$ and we use the fact that $(x+\epsilon)^{\alpha-1}(\epsilon-x)^{\beta-1} \geq 0$. Then the integral

$$\begin{aligned}
I &= \int_{-\epsilon}^{\epsilon}(x+\epsilon)^{\alpha-1}(\epsilon-x)^{\beta-1}dx \\
&= \epsilon^{\alpha+\beta-2}\int_{-\epsilon}^{\epsilon}(1+\frac{x}{\epsilon})^{\alpha-1}(1-\frac{x}{\epsilon})^{\beta-1}dx.
\end{aligned}$$

Using the change of variables $y = \frac{x}{\epsilon}, dx = dy\cdot\epsilon$, we obtain

$$I = \epsilon^{\alpha+\beta-1}\int_{-1}^{1}(1+y)^{\alpha-1}(1-y)^{\beta-1}dy.$$

We assume $\alpha, \beta \in \mathbb{Z}$, consider the case where $\beta > \alpha$, then

$$I = \epsilon^{\alpha+\beta-1} \int_{-1}^{1} (1-y^2)^{\alpha-1}(1-y)^{\beta-\alpha} dy.$$

Using the special Mean Value Theorem and the facts that $(1-y^2)^{\alpha-1} \geq 0$ and $\int_{-1}^{1}(1+y^2)^{\alpha-1}dy = \frac{\Gamma(\alpha)(\alpha!)4^{\alpha}}{(2\alpha)!}$, we obtain

$$I = \epsilon^{\alpha+\beta-1}(1-\bar{y})^{\beta-\alpha}\frac{\Gamma(\alpha)(\alpha!)4^{\alpha}}{(2\alpha)!},$$

where $\bar{y} \in [-1, 1]$.

We use the fact that $B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$ to write

$$\begin{aligned}
\mu * V &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)(2\epsilon)^{\alpha+\beta-1}} V(\xi_y)\epsilon^{\alpha+\beta-1}(1-\bar{y})^{\beta-\alpha}\frac{\Gamma(\alpha)(\alpha!)4^{\alpha}}{(2\alpha)!} \\
&= \frac{\Gamma(\alpha+\beta)(\alpha!)}{\Gamma(\beta)(2\alpha)!} 2^{\alpha-\beta+1}(1-\bar{y})^{\beta-\alpha}V(\xi_y) \\
&= \frac{(\alpha+\beta-1)!\alpha!2}{(\beta-1)!(2\alpha)!}\frac{1}{2^{\beta-\alpha}}(1-\bar{y})^{\beta-\alpha}V(\xi_y)
\end{aligned}$$

Since $(1-\bar{y})^{\beta-\alpha} \leq 2^{\beta-\alpha}$,

$$\begin{aligned}
\mu * V &\leq \frac{(\alpha+\beta-1)!\alpha!2}{(\beta-1)!(2\alpha)!}V(\xi_y) \\
&= \frac{(2\alpha-1+(\beta-\alpha))!\alpha!2}{(\alpha-1+(\beta-\alpha))!(2\alpha)!}V(\xi_y) \\
&= \frac{(2\alpha-1+\beta-\alpha)(2\alpha-1+\beta-\alpha-1)\cdots(2\alpha-1)!\alpha!2}{(\alpha-1+\beta-\alpha)(\alpha-1+\beta-\alpha-1)\cdots(\alpha-1)!(2\alpha)!}V(\xi_y) \\
&= \frac{\beta+\alpha-1}{\beta-1}\cdot\frac{\beta+\alpha-2}{\beta-2}\cdots\frac{2\alpha}{\alpha}V(\xi_y) \\
&\leq 2^{\beta-\alpha-1}V(\xi_y)
\end{aligned}$$

Since $2^{\beta-\alpha-1}V(\xi_y) \geq V(\xi_y)$, we cannot use the argument above and apply Theorem 3.1.6 to find a condition for geometric ergodicity.

The same argument applies when $\alpha > \beta$, and we obtain $\mu * V \leq 2^{\alpha-\beta-1}V(\xi_y)$ for that case. Then, $2^{\alpha-\beta-1}V(\xi_y) \geq V(\xi_y)$ and we cannot use the argument above and apply Theorem 3.1.6 to find a condition for geometric ergodicity.

3.4.4 Normal Distribution

In order to adapt theorem 3.1.6 to fit an error sampled from a probability density, we require that the density has bounded support in order to use our proof techniques. As a result, we cannot show this result for adapting Theorem 3.1.6 to fit an error sampled from a Normal

distribution. Simulations show that convergence is not preserved when errors are sampled from a Normal distribution whose standard deviation is too large. This is because the tails of the distribution are too heavy, and we obtain errors that are too large.

3.5 Numerical Investigation

3.5.1 Setup of Problem

We illustrate the results using a Bayesian Calibration that is performed using 20 code observations and 21 field observations with design matrices generated using latin hypercubes. As usual, we use Gaussian processes to to model the computer model and the model inadequacy. This example is provided as part of the "calibrator" package in R [Hankin, 2005].

The field observations are written as

$$z(x) = \rho\eta(x, \theta) + \delta(x) + \epsilon,$$

where the observations are made over the known parameters x, θ are the unknown true parameters, $\eta(x, \theta)$ is a Gaussian process representing a computer model taking the parameters x and θ, $\delta(x)$ is a Gaussian model representing the model inadequacy, and ϵ is an observational iid error, where $\epsilon \sim N(0, \lambda^2)$.

In this problem, the true model is

$$x + y^2 + \sin(2\pi x) + \sin(2\pi y) + CE,$$

where CE is the correlated error. We obtain the field observation $z(x)$ by adding the uncorrelated observation error ϵ.

The computer model is

$$Ax + By^2 + CE.$$

where A, B are the unknown parameters. We set $A = 1$ and $B = 1$ as the true values. Thus, the vector of unknown parameters is $\theta = (A, B)$.

The model inadequacy term is

$$\sin(2\pi x) + \sin(2\pi y) + CE.$$

We use model basis functions to represent the computer model and the model inadequacy. The computer model is written

$$h_1(x, \theta)^\top \beta_1 + CE$$

where $h_1(x, \theta) = (1, x, A, Ax, By^2)$ is the model basis function. The true value of the coefficients are $\beta_1 = (0, 0, 0, 1, 1)$.

Similarly, the model inadequacy term is written

$$h_2(x)^\top \beta_2 + CE$$

where $h_2 = (\sin(2\pi x), \sin(2\pi y))$ is the model basis function. The true value of the coefficients are $\beta_2 = (1, 1)^\top$.

3.5.2 Brief Description of Calibrator

We are using the "calibrator" R package to perform the Bayesian calibration [Hankin, 2005]. It includes several useful functions for the calibration, including functions for estimating the hyperparameters and for obtaining expectations under different distributions.

The posterior PDF is computed in calibrator by the function p.eqn8.supp. We use a Metropolis Hastings algorithm to sample the posterior of the unknown parameters. The PDF provided in the supplement is only proportional to the true PDF, but that is enough for our purposes since the unknown constant cancels out in the calculation of the acceptance probability in the Metropolis Hastings algorithm.

3.5.3 Results Without Added Errors

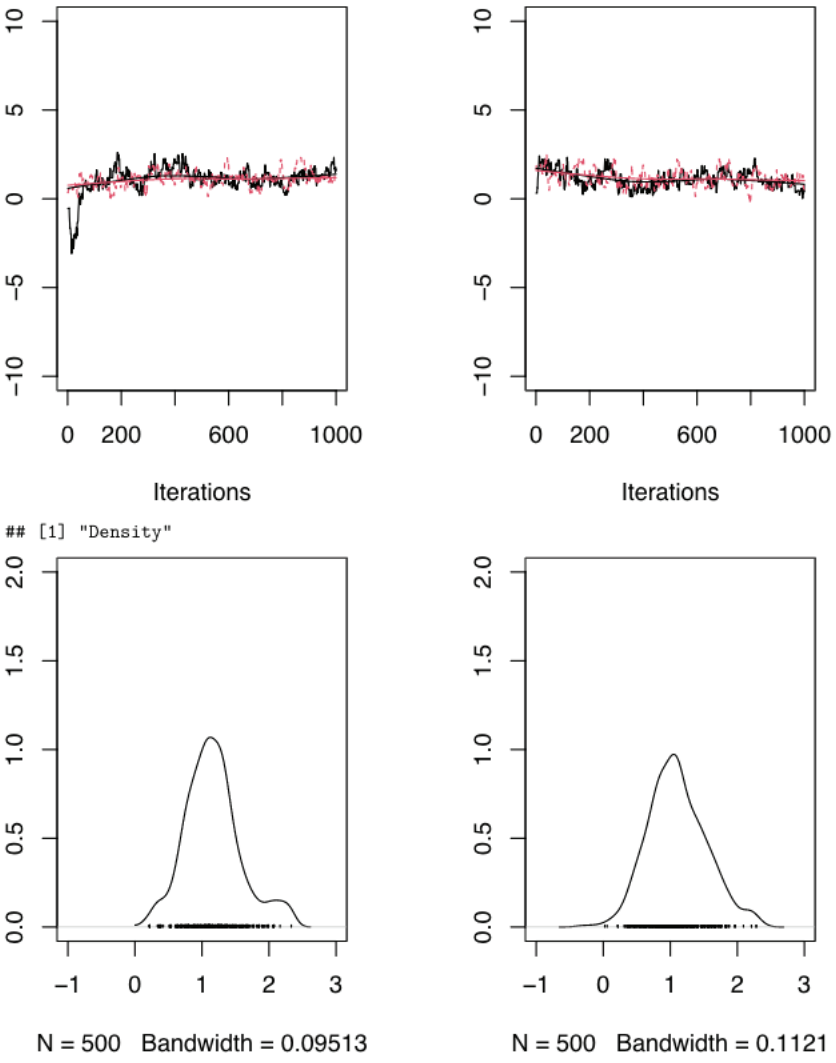


Figure 3.1: Markov chains and estimated densities for two computations of the Bayesian calibration example with two variables

When we perform the Bayesian calibration example without adding any errors, the Markov chains for both variables converge well. This is shown in the plots above, where two Markov Chains (one in red and one in black) are computed with different starting values, for each variable. Both chains quickly move near the true values of the variables (1 in each case). However, they exhibit some oscillatory behaviour and fluctuate around 1 instead of smoothly going near 1. This behaviour could be due to the fact that the algorithm still accepts proposal values that are not exactly 1, and is able to walk away from 1 for some distance. However,

when the chain is too far from 1, it no longer accepts values that are farther from 1 and starts walking back towards 1.

3.5.4 Uniform and Modified Beta Errors

We add errors sampled from the modified Beta(1,1) (uniform), Beta(2,2), and Beta(10,10) distributions, for $\epsilon = 1$, 0.5, and 0.1. The following figures show the probability densities for the modified Beta(1,1), Beta(2,2), and Beta(10,10) distributions for $\epsilon = 1$:

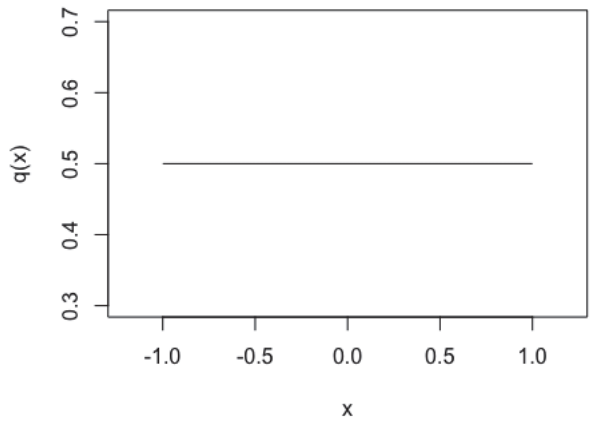


Figure 3.2: Probability density of the modified Beta(1,1) distribution

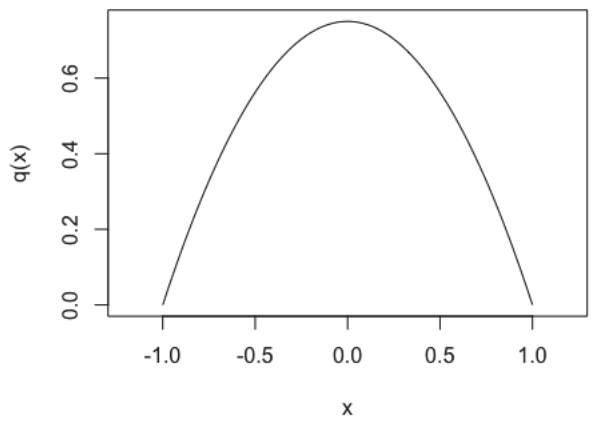


Figure 3.3: Probability density of the modified Beta(2,2) distribution

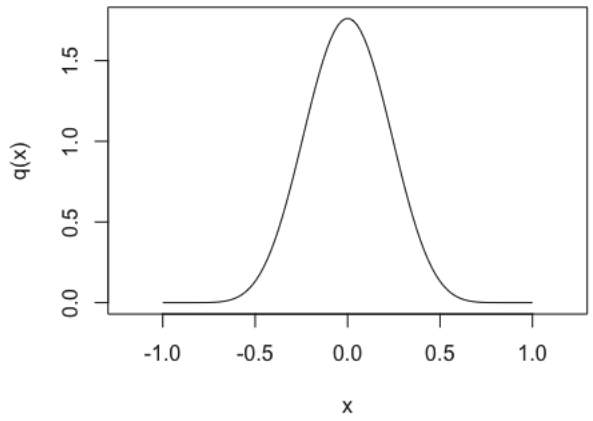


Figure 3.4: Probability density of the modified Beta(10,10) distribution

When we add errors sampled from these distributions, we find that the Markov Chains converge every time. 1 and 0.5 are large values for ϵ since the true value of the variables is 1. This shows that convergence is preserved not only when ϵ satisfies the condition in section 3.4.3, but also for larger values of ϵ. We show some chains for $\epsilon = 1$:

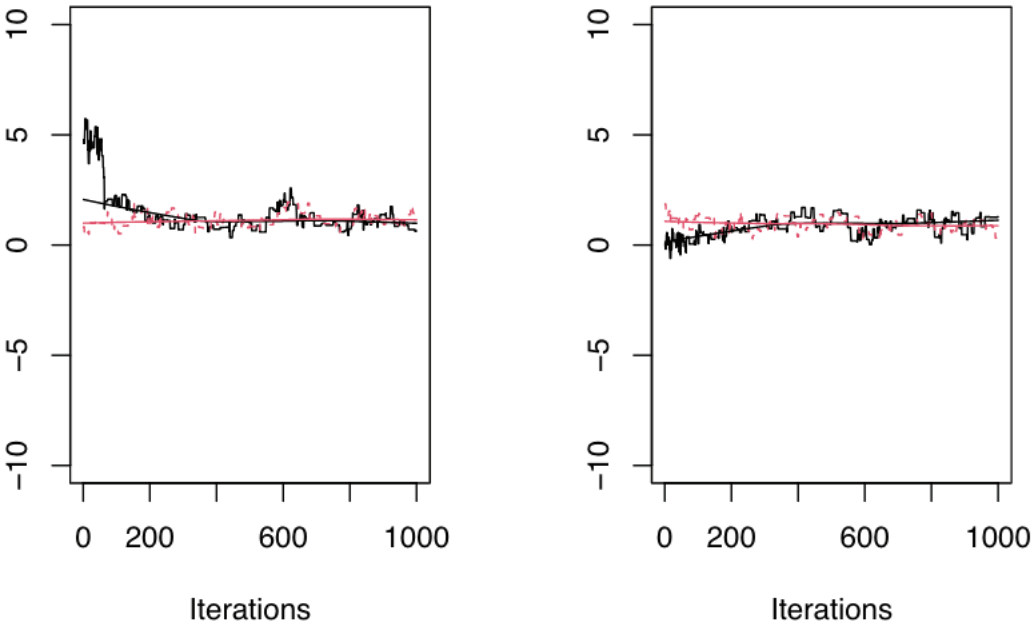


Figure 3.5: Markov chains and estimated densities for two computations of the Bayesian calibration example with two variables, with error sampled from the Uniform(-1,1) distribution

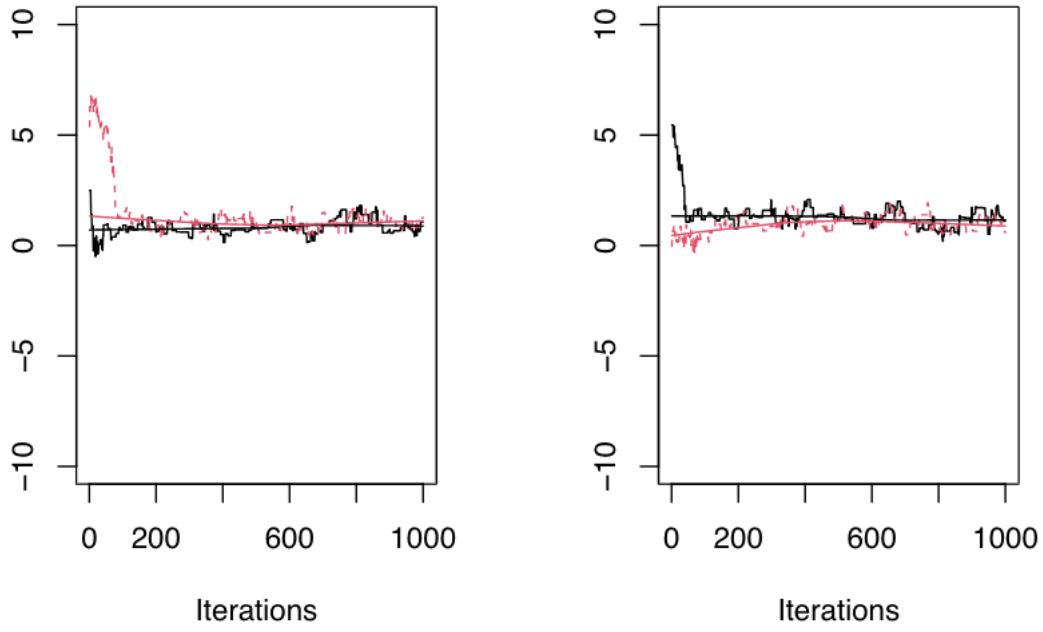


Figure 3.6: Markov chains and estimated densities for two computations of the Bayesian calibration example with two variables, with error sampled from the Modified Beta(2,2) distribution

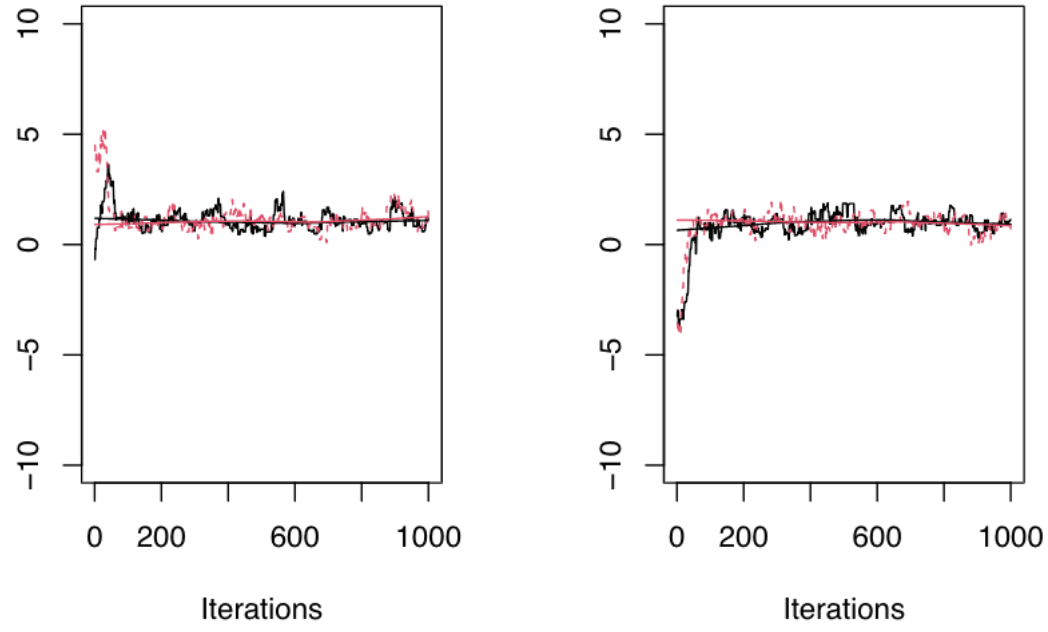


Figure 3.7: Markov chains and estimated densities for two computations of the Bayesian calibration example with two variables, with error sampled from the Modified Beta(10,10) distribution

There is a "blockiness" quality in some of these chains. For example, in Figures 3.5 and 3.6, since there are more iterations in the chains where they stay in the same place instead of accepting a new proposed value. This effect is reduced when ϵ is set to be smaller than 1, such as when $\epsilon = 0.1$. This may be because the acceptance rate in the Metropolis-Hastings is lower when ϵ is larger and we can get proposed values that are farther from the correct value. This means that the chains are not able to explore the state space as quickly since they spend more time staying in the same states. However, this is not an issue for this example since the chains are still able to move towards the true value 1 quite quickly.

We did find one chain that did not converge within 1000 iterations for the uniform(-0.5,0.5) error:

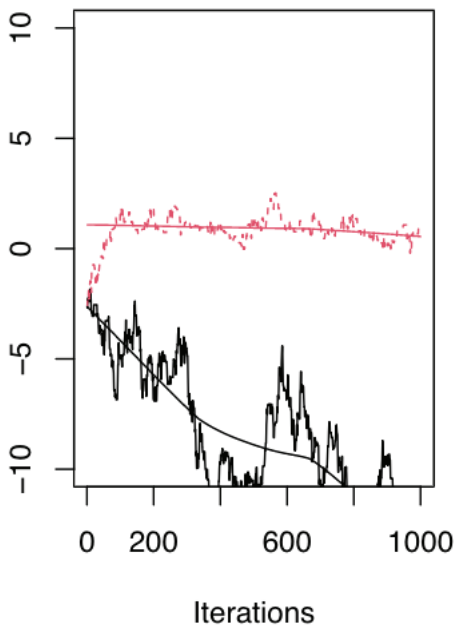


Figure 3.8: Markov chains for two computations of the Bayesian calibration example for one variable, with error sampled from the Uniform(-0.5,0.5) distribution

However, when we continue to run the chain for another 9000 iterations, we find that it converges after (approximately) 3000 more iterations:

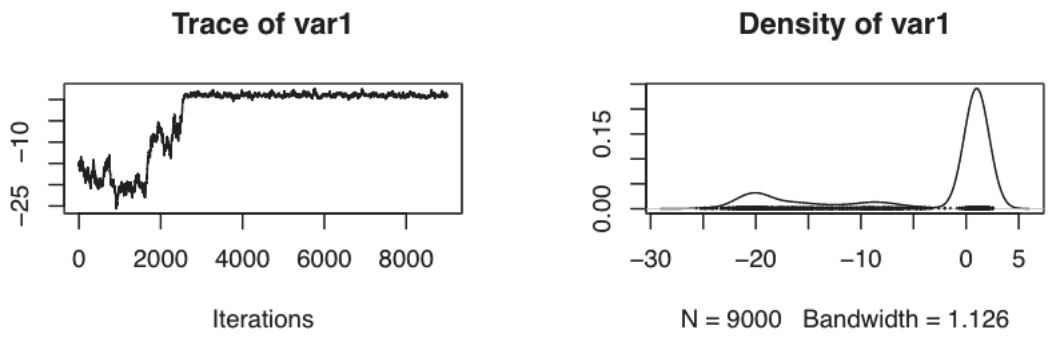


Figure 3.9: Markov chains up to 10000 iterations and estimated densities for two computations of the Bayesian calibration example for one variable, with error sampled from the Uniform(-0.5,0.5) distribution

The observation that there are runs where the chains take longer to converge motivates modelling the error of the chains using time series so that we have an idea of how the error behaves over time (see Chapter 4).

3.5.5 Variation: Non Symmetric Beta Error

To explore the limits of Theorem, we add non-symmetric modified beta errors for several different shape parameters with $\epsilon = 0.1$ and $\epsilon = 0.01$. We find that a few of the chains diverge, and that chains diverge more frequently for larger ϵ and for shape parameters that yield a distribution with more bias. For example, the modified Beta(2,10) error with $\epsilon = 0.1$ produces chains that diverge for 4 out of 10 runs.

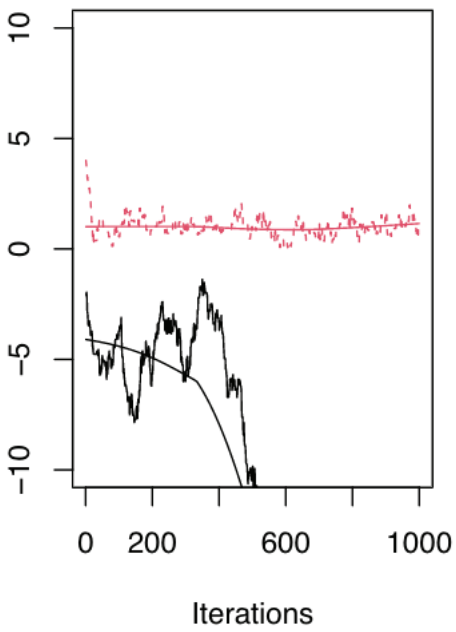


Figure 3.10: Markov chains for two computations of the Bayesian calibration example for one variable, with error sampled from the modified Beta(2,10) distribution

When a Markov Chain diverges, it tends to start walking further and further away from 1. In these simulations, the Markov Chains also do not tend to converge to a different value. This behaviour can be explained by the fact that the non-symmetric error yields perturbations in one direction more often, so that the chain gets "kicked" farther and farther away from the true value.

We also find that these chains diverge more frequently when the initial values of the chains are farther from the correct value. To investigate this, we run chains with starting values 5 and -5 with the modified Beta(2,10) error with $\epsilon = 0.1$ and find that for chains starting from -5, 9 out of the 10 chains diverge.

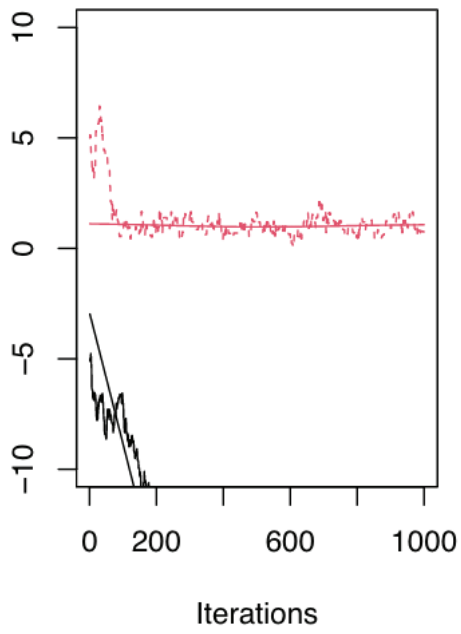


Figure 3.11: Markov chains for two computations of the Bayesian calibration example for one variable, with error sampled from the modified Beta(2,10) distribution and starting values 5 and -5

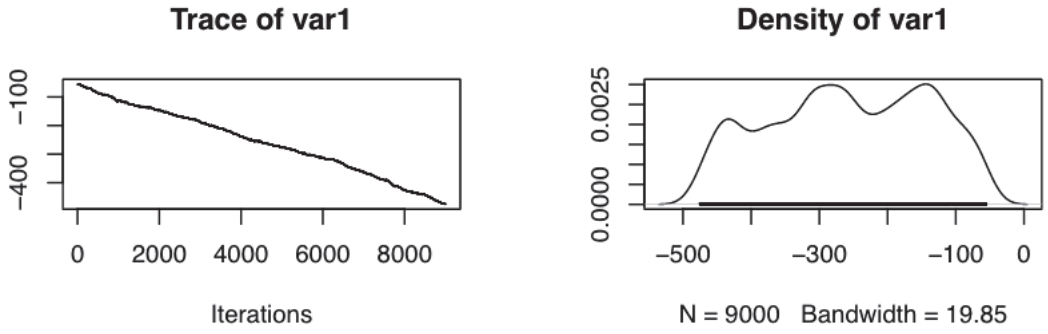


Figure 3.12: Markov chains up to 10000 iterations and estimated densities for two computations of the Bayesian calibration example for one variable, with error sampled from the modified Beta(2,10) distribution

When we run 9000 more iterations for these diverging chains, they all continue to diverge. This suggests that the asymmetry in the beta distribution for the error makes it so that convergence is no longer preserved.

Chapter 4

Modelling the Error

4.1 Modelling the Error Using Time Series

The results we have obtained so far address the preservation of convergence of Markov Chains in MCMC, but do not deal with the accuracy of the limiting distributions. Therefore, we look for a model of the error to assess the accuracy of the limiting distributions.

We fit some time series to the difference between the Markov Chains with error and an average Markov Chain with no error added to see if they fit in the same class of ARMA models. We did not fit them with chains that were clearly diverging. We first used the auto.arima() function in R to fit the chains from the examples with Uniform errors, but the results were not encouraging.

When we fit time series to the simulation data from the Markov chains, the residuals had oscillatory behaviour, so we tried to use some spectral methods. This did not give us any helpful results.

Further examination of the Markov chains suggested that their behaviour looks like random walks. Although it is not quite like a simple random walk (since the proposed values in the Markov chain are random samples from a distribution), it is useful to remember Polya's recurrence theorem [Mare, 2013]:

Theorem 4.1.1. *A simple random walk on a d-dimensional lattice is recurrent for* $d = 1, 2$ *and transient for* $d > 2$.

If it were true that the perturbed Markov Chains are simple random walks, then they should come back to the a point near the starting value, regardless of the behaviour at the beginning of the Markov Chain, since the Markov Chains are on a 2-dimensional lattice. If they are able to get near 1, they are likely to converge since the probability of staying in that area is high.

We plot first-order differences of the chains.

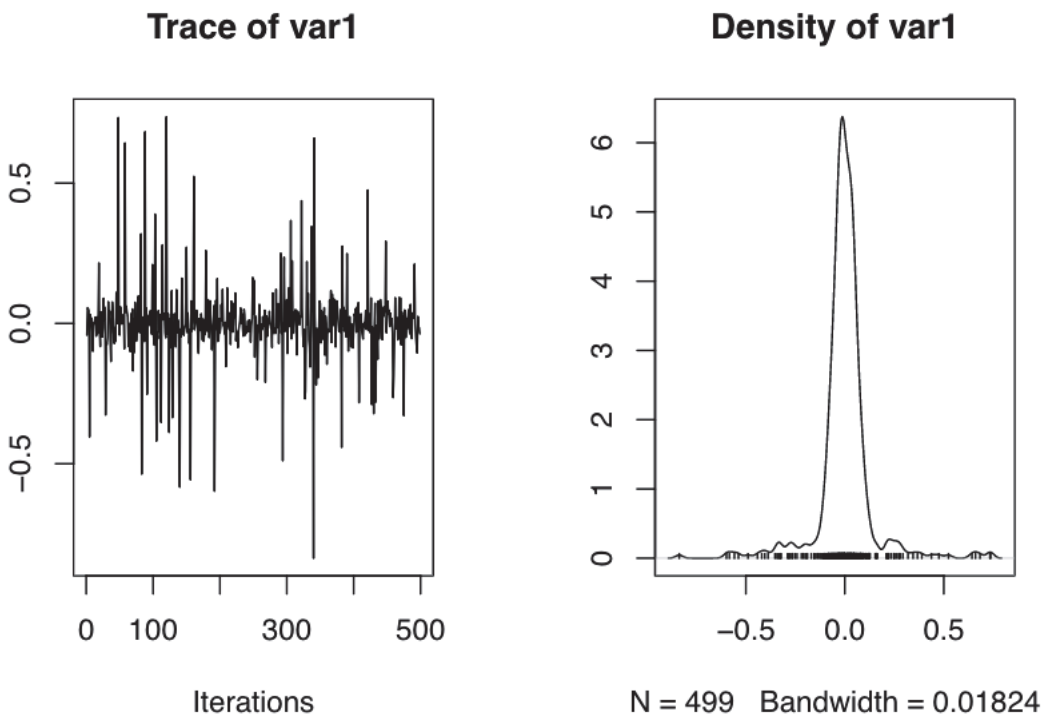


Figure 4.1: Markov chain and estimated density for first-order differences of the error, for the case of error sampled from the Uniform(-1,1) distribution

We find that these first-order differences look more like white noise, which suggests that the chains are similar to random walks.

We performed the Augmented Dickey-Fuller test on some of these chains and found that we have sufficient evidence to reject the null hypothesis that they are non-stationary. When we fit the time series model for Random Walks (ARIMA(0,1,0)) to the chains, the residuals still had oscillatory behaviour.

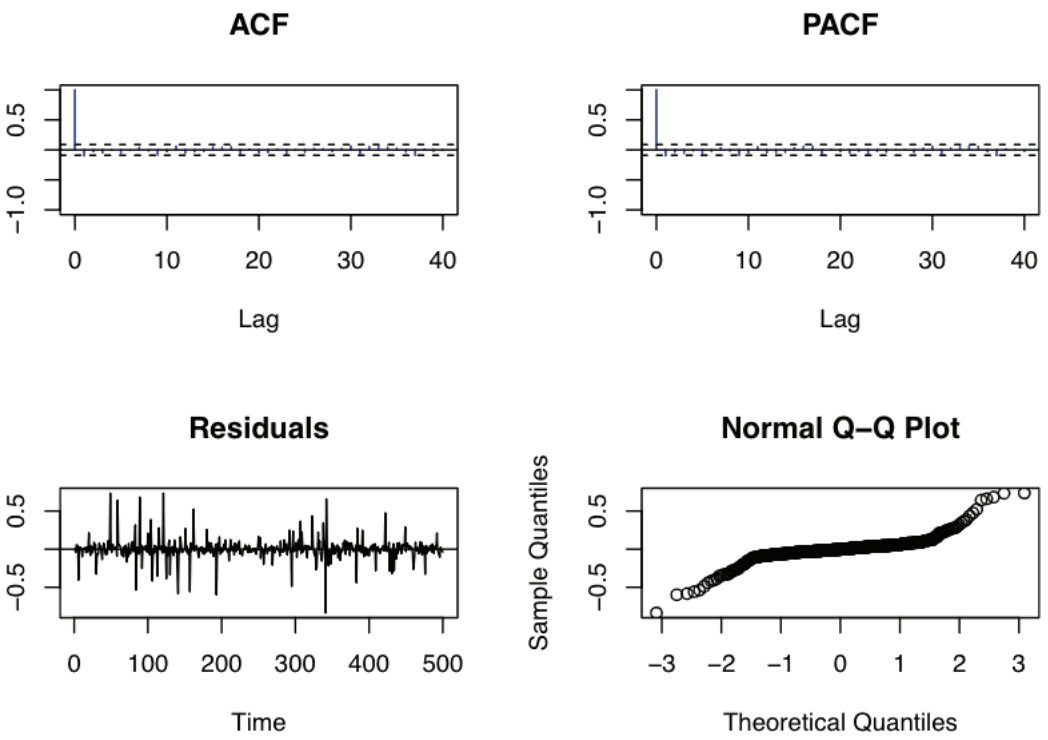


Figure 4.2: ACF, PACF, residuals, and QQ plots for an ARIMA(0,1,0) fit to the error

In addition, we found that the chains that converged oscillated quite close to their means. Because of this, we looked into whether we could fit a bounded random walk model.

4.1.1 Bounded Random Walk

Upon further investigation, we find that the chains that eventually converge behave like random walks that have upper and lower bounds. This may be the case because the chain cannot "walk" too far from the target value since the probability of the algorithm accepting a value that is too far is very small. Thus, the chain acts like a random walk until it reaches a "boundary" with too small of an acceptance probability, and then it has a reversion effect so that the chain walks back towards the mean. This prevents the chains from truly going on a random walk, potentially far away from the mean. In fact, these chains can be stationary and have ergodic distributions [Nicolau, 2002].

We explore some properties of the bounded random walk for the discrete case, since we are interested in fitting a time series in discrete time [Nicolau, 2002]. Since the bounded random walk behaves like a random walk most of the time, we need $E[\Delta X_t | X_{t-1} = x]$ (where $\Delta X_t = X_t - X_{t-1}$) is zero in some interval that corresponds to the "bounds" of the bounded random walk. We also need that $E[\Delta X_t | X_{t-1} = x]$ is positive when x is closer to 0, and negative when x is farther from 0. This is because the process is bounded in probability and mean-reverting to a value that we can call τ, which means that at

$x < \tau$, $E[\Delta X_t | X_{t-1} = x] > 0$ and at $x > \tau$, $E[\Delta X_t | X_{t-1} = x] < 0$. We also want $E[\Delta X_t | X_{t-1} = x]$ to be monotonic, since then the reversion effect is stronger the farther x is from the mean. Finally, we also want $E[\Delta X_t | X_{t-1} = x]$ to be differentiable so that the reversion effect can be smooth.

If we set $E[\Delta X_t | X_{t-1} = x] = e^k(e^{-\alpha_1(x-\tau)} - e^{\alpha_2(x-\tau)})$, with $\alpha_1 \geq 0, \alpha_2 \geq 0, k < 0$, then all the assumptions for $E[\Delta X_t | X_{t-1} = x]$ are satisfied [Nicolau, 2002]. We can define the bounded random walk model

$$X_t = X_{t-1} + e^k(e^{-\alpha_1(x-\tau)} - e^{\alpha_2(x-\tau)}) + \sigma_t \epsilon_t,$$

where ϵ_t is a sequence of independently and identically distributed (i.i.d.) random variables with $E(\epsilon_t) = 0$ and $Var(\epsilon_t) = 1$, and the volatility σ_t belongs to the information set $\mathcal{F}_{t-1} = \sigma(X_\tau : \tau \leq t-1)$ [Nicolau, 2002].

We examine the function

$$a(x) = e^k(e^{-\alpha_1(x-\tau)} - e^{\alpha_2(x-\tau)}).$$

First, $a(\tau) = 0$, so that the chain behaves like a random walk when X_{t-1} is at the mean. Otherwise, we can keep a(x) to be approximately 0 by controlling the parameters α_1, α_2, and k. If $\alpha_1 \geq 0, \alpha_2 \geq 0, k < 0$ and $|k|$ is large with respect to α_1 and α_2, then $a(x)$ would be approximately 0 if it is not too far away from the mean of the chain. This function also ensures that the "reversion" effect in the bounded random walk occurs when the chain moves too far from the mean. For example, if $X_{t-1} > \tau$, then $a(x) < 0$ and the probability that $X_t < X_{t-1}$ is higher, so that the chain is more likely to return closer to the mean.

When the parameter k is less than zero, it controls the range of the interval of the bounds on which the chain behaves like a random walk. This is because when $|k|$ is large, the range is larger, and when $|k|$ is small, the range is smaller. When the parameters $\alpha_1 \geq 0$ and $\alpha_2 \geq 0$ are equal, we expect the mean of the process to be τ. These parameters measure the strength of the reversion effect towards the mean, where α_1 corresponds to the reversion effect near the lower bound and α_2 corresponds to the reversion effect near the upper bound of the process. When these parameters are not equal, we have an asymmetrical effect of reversion at the two bounds.

In general, the function a(x) is a good choice since it is general enough to produce a large quantity of functions that can produce stationary bounded random walks (which are bounded in probability).

Finally, we select a Self-Exciting Threshold Autoregressive(3,1,1) (SETAR(3,1,1)) model, which is an alternative to the Bounded Random Walk that is easier to implement [Nicolau, 2002]. This model produces similar behaviour to the Bounded Random Walk. But the function $E[\Delta X_t | X_{t-1} = x]$ is not differentiable, and so at the threshold parameters, we have

a sudden transition of regimes, instead of the smooth transition seen in bounded random walks [Nicolau, 2002].

4.1.2 SETAR Models

SETAR models are an extension of Autoregressive (AR) models. $\{X_t\}$ is an AR process of order p if it is a sequence of random variables that is stationary, satisfying

$$X_t = \phi_0 + \phi_1 X_{t-1} + ... + \phi_p X_{t-p} + Z_t,\ t = 0, \pm 1, ...$$

where $\{Z_t\} \sim WN(0, \sigma^2)$, Z_t is uncorrelated with X_s for each $s < t$, and $\phi_0, ..., \phi_p$ are real-valued constants.

For the SETAR model, there is a threshold variable V_t that triggers changes in regime, based on past values of X. When the regime changes, we expect the behaviour of the time series to change as well. We call this model self-exciting since it is using past values of X. For these past values, we use X_{t-d}, where d is a delay parameter. The SETAR model can be written

$$X_t = Y_t \phi^{(j)} + \sigma^{(j)} Z_t \text{ if } r_{j-1} < V_t < r_j,$$

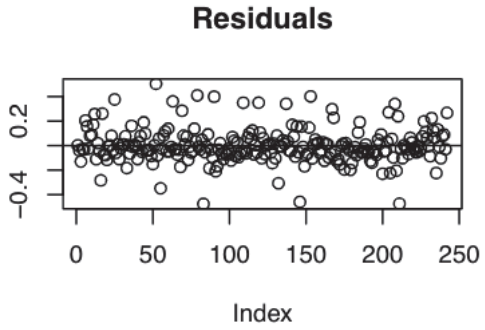
where $Y_t = (1, X_{t-1}, X_{t-2}, ..., X_{t-p})$ is a column vector of variables and $-\infty = r_0 < r_1 < ... < r_k = +\infty$ are k+1 non-trivial thresholds dividing the domain of V_t into k different regimes. This model has an AR process for each regime, so it has k AR parts. Each regime has its own set of coefficients to describe its behaviour.

The parameters of the SETAR(3,1,1) model indicate the number of regimes k, the delay parameter d, and the number of forecasting steps. In our case, we have 3 regimes since we have the random walk regime, the upper bound regime, and the lower bound regime. The latter two regimes display the reversion effect behaviour towards the mean.

4.2 Results

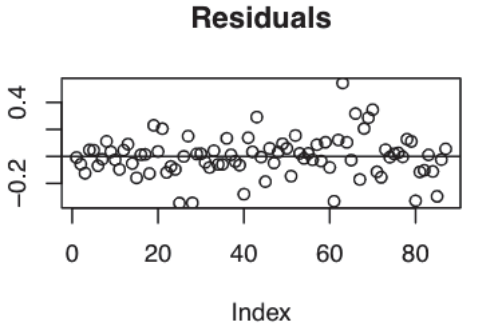
4.2.1 Modified Beta(2,2) Error

For the modified Beta(2,2) error, we find that the Setar(3,1,1) model fits well and that we could make good predictions for the error. We show the residuals and the tests for iid noise in the residuals for the three regimes in the model:



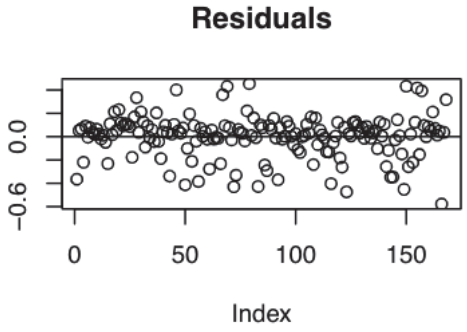

```
Null hypothesis: Residuals are iid noise.
Test                         Distribution Statistic   p-value
Ljung-Box Q                 Q ~ chisq(20)      17.8    0.6005
McLeod-Li Q                 Q ~ chisq(20)     14.47    0.8057
Turning points T    (T-160)/6.5 ~ N(0,1)       157    0.6462
Diff signs S      (S-120.5)/4.5 ~ N(0,1)       119    0.7389
Rank P        (P-14580.5)/629.4 ~ N(0,1)     14322    0.6813
```

Figure 4.3: Residuals and tests for iid noise for first regime of SETAR model (Modified Beta(2,2) Error)




```
Null hypothesis: Residuals are iid noise.
Test                         Distribution Statistic   p-value
Ljung-Box Q                 Q ~ chisq(20)     20.55    0.4243
McLeod-Li Q                 Q ~ chisq(20)     13.44    0.8575
Turning points T   (T-56.7)/3.9 ~ N(0,1)        53    0.3461
Diff signs S         (S-43)/2.7 ~ N(0,1)        41    0.4602
Rank P         (P-1870.5)/136.4 ~ N(0,1)      1954    0.5404
```

Figure 4.4: Residuals and tests for iid noise for second regime of SETAR model (Modified Beta(2,2) Error)



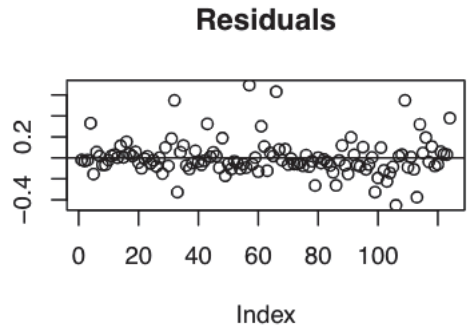

```
## Null hypothesis: Residuals are iid noise.
## Test                         Distribution Statistic   p-value
## Ljung-Box Q                 Q ~ chisq(20)     14.65     0.796
## McLeod-Li Q                 Q ~ chisq(20)     14.84    0.7857
## Turning points T  (T-110.7)/5.4 ~ N(0,1)       112    0.8062
## Diff signs S       (S-83.5)/3.8 ~ N(0,1)        76    0.0457 *
## Rank P           (P-7014)/364.5 ~ N(0,1)      6687    0.3697
```

Figure 4.5: Residuals and tests for iid noise for third regime of SETAR model (Modified Beta(2,2) Error)

Since the p-values for almost all the tests above are greater than 0.05, we do not have sufficient evidence to reject the null hypothesis that the residuals are iid noise. This provides evidence that the SETAR model fits well.

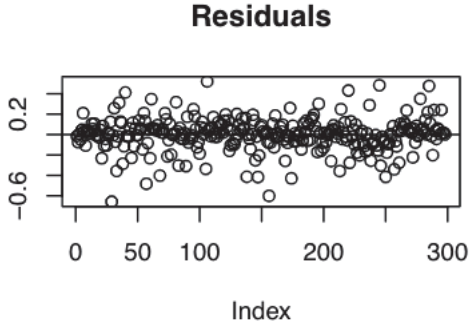
4.2.2 Modified Beta(10,10) Error

We find the same result with the modified Beta(10,10) error. We show the residuals for the three regimes in the model:



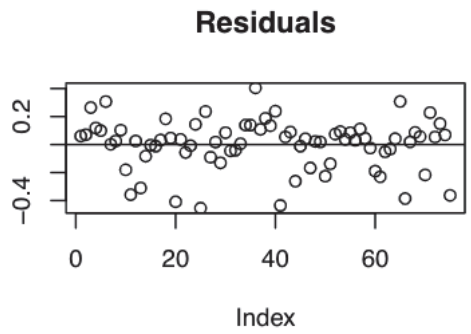

```
Null hypothesis: Residuals are iid noise.
Test                        Distribution Statistic   p-value
Ljung-Box Q                Q ~ chisq(20)     16.31    0.6971
McLeod-Li Q                Q ~ chisq(20)     13.08    0.8741
Turning points T   (T-81.3)/4.7 ~ N(0,1)        87     0.224
Diff signs S       (S-61.5)/3.2 ~ N(0,1)        61    0.8769
Rank P           (P-3813)/231.5 ~ N(0,1)      3516    0.1995
```

Figure 4.6: Residuals and tests for iid noise for first regime of SETAR model (Modified Beta(10,10) Error)




```
Null hypothesis: Residuals are iid noise.
Test                        Distribution Statistic   p-value
Ljung-Box Q                Q ~ chisq(20)     21.66    0.3592
McLeod-Li Q                Q ~ chisq(20)     24.38    0.2262
Turning points T  (T-197.3)/7.3 ~ N(0,1)       210    0.0809
Diff signs S        (S-148.5)/5 ~ N(0,1)       151    0.6165
Rank P        (P-22126.5)/859.5 ~ N(0,1)     21733    0.6471
```

Figure 4.7: Residuals and tests for iid noise for second regime of SETAR model (Modified Beta(10,10) Error)




```
## Null hypothesis: Residuals are iid noise.
## Test                         Distribution Statistic   p-value
## Ljung-Box Q                 Q ~ chisq(20)     19.28    0.5039
## McLeod-Li Q                 Q ~ chisq(20)     27.76    0.1153
## Turning points T   (T-48.7)/3.6 ~ N(0,1)        49    0.9264
## Diff signs S         (S-37)/2.5 ~ N(0,1)        40    0.2332
## Rank P         (P-1387.5)/109.3 ~ N(0,1)      1342    0.6772
```

Figure 4.8: Residuals and tests for iid noise for third regime of SETAR model (Modified Beta(10,10) Error)

Since the p-values for all the tests above are greater than 0.05, we do not have sufficient evidence to reject the null hypothesis that the residuals are iid noise. This provides evidence that the SETAR model fits well.

Chapter 5

Application of the Error Model to Stochastic Sensitivity Analysis for Differential Equations

We solve differential equations numerically and assess if we can fit a symmetric beta distribution to errors in the solutions coming from the variation of parameter values. We use the Lorenz equations for this example since it produces nontrivial error with time in the chaotic regime [Hirsch et al., 2004]. The Lorenz equations model what was thought to be ideal behaviour of the earth's atmosphere using three differential equations with parameters that we denote a, b, and c [Hirsch et al., 2004]. We start with using the parameters (a = -8/3, b = -10, c = 28) since these yield the chaotic case of the Lorenz equations [Hirsch et al., 2004]. With the starting state (X=1, Y=1, Z=1), we implement the Lorenz equations in R and use the deSolve package to solve these equations [Soetaert et al., 2010]. The numerical error is reasonably small over the time period we consider. We then take 100 random samples from intervals centred around each of the parameter values, with the minimum value of the intervals being 0.1 less than the centre values and the maximum value of the intervals being 0.1 greater than the centre values. We solve the differential equations with these sampled parameter values and calculate the difference between the solutions with the original parameters and the solutions with the new sampled parameters to find the error produced from the variation in the parameter values. We plot histograms for these errors and fit a non standard Beta distribution to them. In R, the betafunctions package contains the Beta.4p.fit function, which fits non-standard Beta distributions (which are not restricted to the domain (0,1)) to data [Haakstad, 2022].

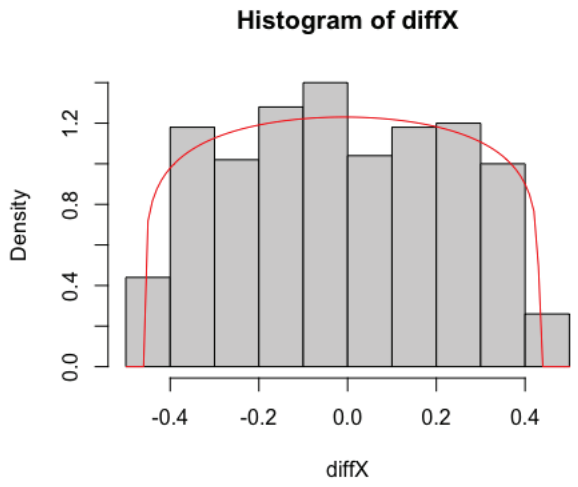


Figure 5.1: Histogram of errors for X produced from variation in the parameter values, with a fit of a non-standard Beta distribution

The fit gave the parameters $\alpha = 1.160154$ and 1.156269, with domain interval $(-0.4578563, 0.4306733)$.

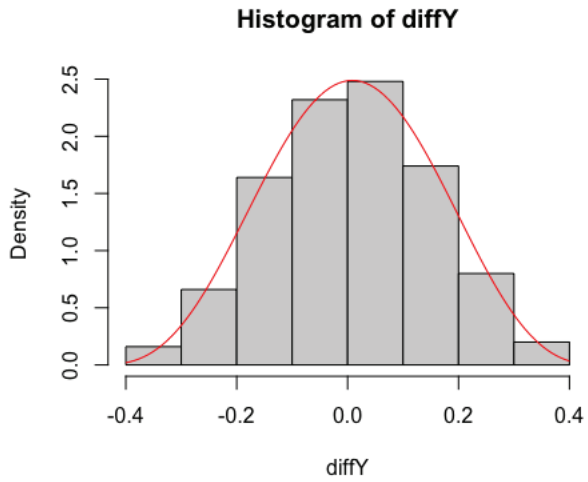


Figure 5.2: Histogram of errors for Y produced from variation in the parameter values, with a fit of a non-standard Beta distribution

The fit gave the parameters $\alpha = 4.563$ and 4.506922, with domain interval $(-0.4649549, 0.4744163)$.

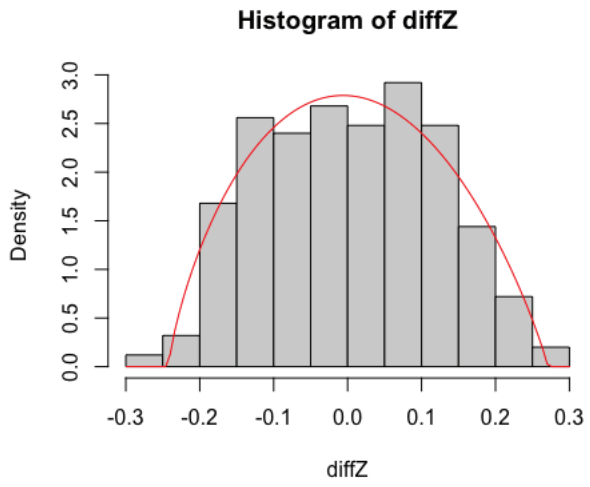


Figure 5.3: Histogram of errors for Z produced from variation in the parameter values, with a fit of a non-standard Beta distribution

The fit gave the parameters $\alpha = 1.763371$ and 1.897125, with domain interval $(-0.2419946, 0.2708269)$.

Since the α and β estimates were close for each of the fits, this means that it would be reasonable to fit symmetric beta distributions (modified to allow for flexible support centred at zero) to these errors.

Chapter 6

Conclusion

We have shown that the convergence of MCMC can be preserved under some conditions when an error sampled from a distribution is added to the values at each step. We apply our results to three distributions for error and illustrate with simulations. We also explore the modelling of the errors in perturbed Markov chains using a SETAR (3,1,1) time series model. Since we fit this model to the errors in the chains, this means that we know how the error behaves over time. We show that the stochastic error model is reasonable for the Lorenz differential equation. Finally, we apply our results to speeding convergence of MCMC by adding random noise.

Bibliography

Rein van den Boomgaard and Leo Dorst. 5.2. linear operators: Convolutions, 2021. URL https://staff.fnwi.uva.nl/r.vandenboomgaard/ComputerVision/LectureNotes/IP/LocalOperators/linearoperators_revised.html.

Laird Breyer, Gareth O. Roberts, and Jeffrey S. Rosenthal. A note on geometric ergodicity and floating-point roundoff error. *Statistics & Probability Letters*, 53(2):123–127, 2001. doi: 10.1016/s0167-7152(01)00054-2.

T. Butler, D. Estep, and L. Panda. A ramble through probability. how i learned to stop worrying and love measure theory.

Cliburn Chan. Markov chain monte carlo (mcmc), 2015. URL https://people.duke.edu/~ccc14/sta-663/MCMC.html.

Robert P. Dobrow. *Introduction to stochastic processes with R.* John Wiley & Sons, 2016.

Charles J. Geyer. Contents, 2020. URL https://www.stat.umn.edu/geyer/8501/markov.pdf.

Haakon Eidem Haakstad. *betafunctions: Functions for Working with Two- And Four-Parameter Beta Probability Distributions*, 2022. URL https://CRAN.R-project.org/package=betafunctions. R package version 1.7.0.

R. K. S. Hankin. Introducing bacco, an r bundle for bayesian analysis of computer code output. *Journal of Statistical Software*, 14, October 2005.

Philipp Hennig, Michael A. Osborne, and Mark Girolami. Probabilistic numerics and uncertainty in computations. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 471(2179):20150142, 2015. doi: 10.1098/rspa.2015.0142.

Morris William Hirsch, Stephen Smale, and Robert L. Devaney. *Differential equations, dynamical systems, and an introduction to Chaos.* Elsevier, 2004.

Matthew D. Hoffman. Roundoff error in metropolis-hastings accept-reject steps. 2020.

Marc C. Kennedy and Anthony O'Hagan. Bayesian calibration of computer models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(3):425–464, 2001. doi: 10.1111/1467-9868.00294.

Shrirang Mare. Polya's recurrence theorem - mathematics at dartmouth, 2013. URL https://math.dartmouth.edu/~pw/math100w13/mare.pdf.

Sean P. Meyn and Richard L. Tweedie. *Markov chains and stochastic stability*. Springer, 1994.

João Nicolau. Stationary processes that look like random walks— the bounded random walk process in discrete and continuous time. *Econometric Theory*, 18(1):99–118, 2002. doi: 10.1017/s0266466602181060.

Gareth O. Roberts, Jeffrey S. Rosenthal, and Peter O. Schwartz. Convergence properties of perturbed markov chains. *Journal of Applied Probability*, 35(01):1–11, 1998. doi: 10.1017/s0021900200014625.

Karline Soetaert, Thomas Petzoldt, and R. Woodrow Setzer. Solving differential equations in R: Package deSolve. *Journal of Statistical Software*, 33(9):1–25, 2010. doi: 10.18637/jss.v033.i09.

Appendix A

Convolution

The convolution of two probability distributions F_1 and F_2 associated with probability measures Q_1 and Q_2 is defined to be

$$F_1 * F_2(y) = F_2 * F_1(y) = \int_{\mathbb{R}} F_1(y - x_2) dQ_2((x_2) = \int_{\mathbb{R}} F_2(y - x_1) dQ_1(x_1).$$

This definition comes from the following theorem [Butler et al.]:

Theorem A.0.1. *Let X_1 and X_2 be independent random variables with probability distribution functions F_1 and F_2 and probability measures Q_1 and Q_2 respectively. Then*

$$P(\{\omega : X_1 + X_2 \leq y\}) = \int_{\mathbb{R}} F_1(y - x_2) dQ_2(x_2) = \int_{\mathbb{R}} F_2(y - x_1) dQ(x_1).$$

In particular, the convolution of two extended real valued functions f_1 and f_2 on $\mathbb{R}$ is defined to be

$$f_1 * f_2(x) = f_2 * f_1(x) = \int_{\mathbb{R}} f_1(x - x_2) f_2(x_2) d\mu_{\mathcal{L}}(x_2) = \int_{\mathbb{R}} f_2(x - x_1) f_1(x_1) d\mu_{\mathcal{L}}(x_1).$$

This definition comes from the following theorem [Butler et al.]:

Theorem A.0.2. *Assume X_1 and X_2 are independent random variables, where the probability distribution function of X_1 has density f_1 and X_2 has probability measure Q_2. Then, $X_1 + X_2$ has density*

$$f(x) = \int_{\mathbb{R}} f_1(x - x_2) dQ_2(x_2).$$

If the probability distribution function of X_2 has density f_2, then

$$f(x) = \int_{\mathbb{R}} f_1(x - x_2) f_2(x_2) d\mu_{\mathcal{L}}(x_2) = \int_{\mathbb{R}} f_2(x - x_1) f_1(x_1) d\mu_{\mathcal{L}}(x_1).$$

We add both biased (0 mean) and unbiased (nonzero mean) errors. If we are looking at the convolution of two independent random variables X and Y that both have a mean of zero, then it can be shown that the convolution is symmetric. If X and Y had symmetric densities

but had nonzero means μ_X and μ_Y, then we can take $\hat{X} = X - \mu_X$ and $\hat{Y} = Y - \mu_Y$ and show that $\hat{Z} := \hat{X} + \hat{Y} = (X + Y) - (\mu_X + \mu_Y) := Z - \mu_Z$ has a density that is symmetric about 0. As a result, $Z = X + Y$ has a density that is symmetric about $\mu_Z = \mu_X + \mu_Y$.

In general, we add errors sampled from a distribution to samples from a Normal distribution. Here, we look at the convolutions of two Normal distributions, an Exponential and Normal distribution, and a Uniform and Normal distribution. When using a convolution of two densities, the convoluted shape might necessitate searching continuously in order to find all the nooks and crannies.

A.1 Two Normal Distributions

If we assume that X_1 and X_2 are independent random variables with normal distributions, with parameters μ_1, σ_1 and μ_2, σ_2 respectively, then we can apply theorem A.0.2 and get that

$$f_{X_1+X_2}(x) = \frac{1}{2\pi\sqrt{\sigma_1\sigma_2}} \int_{\mathbb{R}} e^{-x_1^2/2\sigma_1} e^{-(x-x_1)^2/2\sigma_2} dx_1$$

$$= \frac{1}{2\pi\sqrt{\sigma_1\sigma_2}} e^{-x^2/2(\sigma_1+\sigma_2)} \int_{\mathbb{R}} e^{\frac{-\sigma_1+\sigma_2}{2\sigma_1\sigma_2}} (x_1 - \frac{\sigma_1}{\sigma_1+\sigma_2}x)^2 dx_1.$$

If we evaluate this integral, we get that $X_1 + X_2$ has a normal distribution with parameters $\mu_1 + \mu_2$ and $\sigma_1 + \sigma_2$.

A.2 A Normal and (Independent) Exponential Distribution

If X_1 has an exponential distribution with parameter λ and X_2 has a Normal distribution with parameters mu and σ, we can apply theorem A.0.2 and get that

$$f_{X_1+X_2}(x) = \frac{\lambda}{2} \exp\left(\frac{\lambda}{2}(2\mu + \lambda\sigma^2 - 2x)\right)\left(-1 + \frac{1}{\sigma} + Erfc\left(\frac{\mu + \lambda\sigma^2 - x}{\sigma\sqrt{2}}\right)\right),$$

where Erfc is the complementary error function $Erfc(z) = 1 - Erf(z) = 1 - \frac{2}{\sqrt{\pi}} \int_0^z e^{-t^2} dt$.

This distribution is a unimodal distribution, and it is quite close to a Normal distribution with mean $\mu + \frac{1}{\lambda}$ and variance $\sigma^2 + \frac{1}{\lambda^2}$, but it has non-zero skewness.

A.3 A Normal and Uniform Distribution

If X_1 has a Uniform distribution that is distributed uniformly in the region [a,b] but is zero everywhere else, and X_2 has a Normal distribution with mean 0 and variance σ^2, then we can apply theorem A.0.2 and get that

$$f_{X_1+X_2}(x) = \frac{\psi_0((x-a)/\sigma) - \psi_0((x-b)/\sigma)}{b-a},$$

where $\psi_0(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-t^2/2} dt$ is the distribution function of the standard normal distribution. This convolution gives us a distribution that almost looks uniform, but that has its edges smeared out.

The following figure shows the convolution of a Normal(0,0.1) distribution and a Uniform distribution for different support parameters:

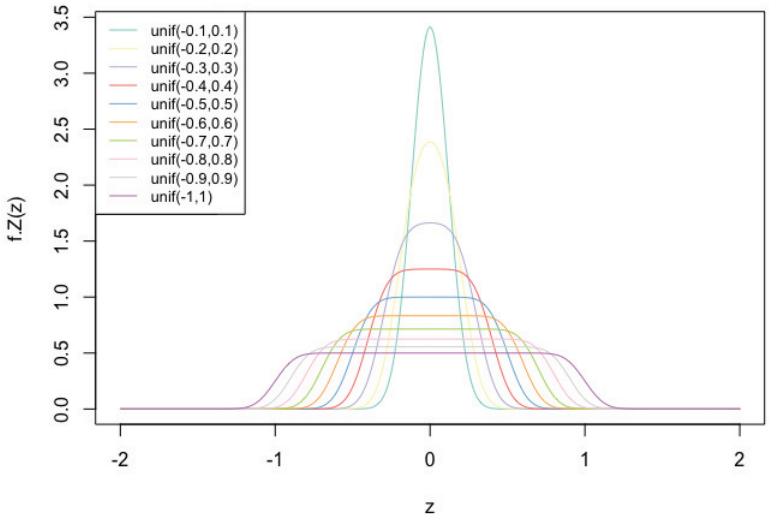


Figure A.1: Convolution of a Normal(0,0.1) distribution and a Uniform distribution for different support parameters

This can be compared to the following figure, which shows the convolution of a Normal(0,0.1) distribution and a modified Beta distribution with different support parameters. The modified Beta distribution is a non-standard Beta distribution that can be defined on any interval, and not just on (0,1).

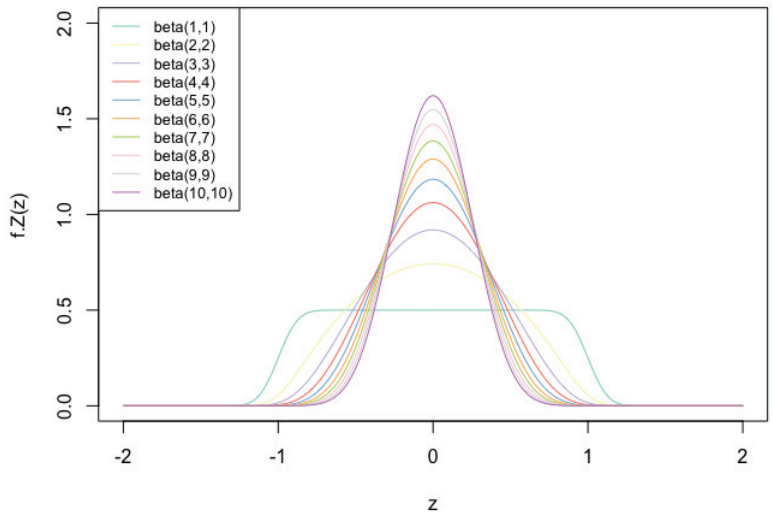


Figure A.2: convolution of a Normal(0,0.1) distribution and a modified Beta distribution with different support parameters

A.4 Taylor Expansion of Convolution

If the convolution is written as

$$f(x) = \int f_1(x - x_2) f_2(x_2) dx_2$$

and x_2 is small, then we can expand it using a straightforward Taylor expansion in the powers of x_2 and get that

$$f(x) = f_1(x) - \frac{\delta f_1(x)}{\delta x} \int x_2 f_2(x_2) dx_2 + \frac{1}{2} \frac{\delta^2 f_1(x)}{\delta x^2} \int x_2 f_2(x_2) dx_2 + ...$$

Then, if we multiply by $f_2(x_2)$ and integrate, we get that

$$f_1(x - x_2) = f_1(x) - f_1'(x) x_2 + \frac{1}{2} f_1''(x) x_2^2 + ...$$

Appendix B

Other Results

B.1 Polynomial Convergence

A more general rate of convergence is polynomial convergence. The results we obtained for geometric convergence of perturbed Markov chains are generalized in Breyer et al. [2001] to work with polynomial convergence.

A Markov chain with transition kernel P converges at the polynomial rate α if

$$\|P^n(x,\cdot) - \pi(\cdot)\| \leq C(x)n^{-(\alpha/1-\alpha)}, n \in \mathbb{N},$$

for some $0 < \alpha < 1$. The corresponding drift condition is

$$PV \leq V - aV^\alpha + b\mathbb{1}_C \tag{B.1}$$

for a function $V \geq 1$, a small set C and a constant $a > 0$. Now if we assume that for some non-negative constants c and β, our perturbation function h satisfies

$$\|h(x) - x\| \leq c\,\|x\|^\beta\,, \tag{B.2}$$

then for some constants $\gamma \leq \min(1,\alpha), \epsilon \geq \beta$, and $\delta \geq c$, polynomial convergence is preserved given that

$$V(y+u) - V(y) \leq \delta K(V(y))^\gamma, \|u\| \leq \delta\,\|y\|^\epsilon\,, y \in \mathcal{X}. \tag{B.3}$$

Then we have the following proposition [Breyer et al., 2001]:

Proposition B.1.0.1. *Consider a Markov chain transition kernel P. Suppose that P is polynomial ergodic with polynomial rate alpha, and satisfies (eq) for some small set C and some drift function V which satisfies (eq) for some constants* $\gamma \leq \min(1,\alpha)$ *and* $\epsilon > 0$. *Define* $\tilde{P}$ *by (eq), and assume that (eq) holds for some* $\beta \leq \epsilon$ *and* $c \leq \delta$. *Then*

$$\tilde{P}V \leq V - aV^\alpha + b'\mathbb{1}_C + cKV^\alpha,$$

for some $b' < \infty$. In particular, if $c < a/K$, then the chain defined by $\tilde{P}$ is also polynomially ergodic, with the same polynomial rate α.

B.2 Roundoff Error in Accept-Reject Steps of Metropolis-Hastings Algorithm

In this section, we summarize the findings in Hoffman [2020] where we look at the case where we feed log-density calculations to the Metropolis-Hastings algorithm.

When using exact arithmetic, the Metropolis-Hastings calculate the correct acceptance probability. However, computers cannot do that on real numbers. Usually, there is no issue with using floating-point arithmetic, but there are two scenarios in which F32 arithmetic presents problems in calculating the acceptance probability.

First, since we have a limited number of mantissa bits to represent our numbers, we have problems calculating acceptance probability when the magnitude of our log-densities are very large. For example, since the number 10000000.51 cannot be represented as an F32, it gets rounded to 10000001.0. Now, if $\log p(\theta') + \log q(\theta|\theta') = 10000000.49$ and $\log p(\theta) + \log q(\theta'|\theta) = 10000000.51$, then the computer would give an acceptance probability of $e^{-1} \approx 0.37$ if these values had been rounded prior to subtraction. The true acceptance probability is $e^{-0.02} \approx 0.98$, which is quite far from 0.37.

Second, small errors accumulate when they are summed. This issue applies when $\log p(\theta)$ is a sum of many terms. Clearly, this is problematic for calculating acceptance probabilities. With these large errors, some proposed values that would have been rejected would now be accepted, and some proposed values that would have been accepted would now be rejected.

In order to analyze the effect of roundoff errors, we write the log-density with roundoff error as

$$\log \hat{p}(\theta) = \log p(\theta) + \epsilon(\theta).$$

We also have the normalizing constant $Z = \int_\theta \hat{p}(\theta) = \mathbb{E}_p[e^{\epsilon(\theta)}]$ since $\epsilon(\theta)$ is bounded (we obtain it from rounding bounded floating-point numbers). With the roundoff error, the Metropolis-Hastings algorithm would leave a perturbed distribution $\frac{\hat{p}}{Z}$ invariant. The asymptotic bias for this algorithm would be small as long as expectations of interest with respect to $\frac{\hat{p}}{Z}$ are close to their values under p. However, the average acceptance rate could get very small if the variance of $\epsilon(\theta)$ is large. This is because a lot of the mass in $\hat{p}$ would be concentrated on values of θ for which $e^{\epsilon(\theta)}$ are very large if $e^{\epsilon(\theta)}$ has heavy tails. As a result, the algorithm would be slowed down.

If we look at our first problem where we had an accurately computed log-density that had to be rounded, we have roundoff errors where ϵ follows a Uniform distribution. We can calculate the exact acceptance rate $\hat{\alpha}^*$ when $\epsilon \sim Uniform([\mu - \sigma, \mu + \sigma])$ as

$$\hat{\alpha}^* = \frac{1}{\sigma} + 1 - \frac{1}{\tanh(\sigma)}.$$

When σ is small, we get that $\hat{\alpha}^* = 1 - \frac{\sigma}{3} + O(\sigma^3)$. When it is large, we get that $\hat{\alpha}^* = \frac{1}{\sigma}$. This shows that with large roundoff errors, we can end up with a much lower acceptance rate.

If we look at our second problem where many independent roundoff errors are being summed, we have roundoff errors where ϵ follows a Normal distribution. We can calculate the exact acceptance rate $\hat{\alpha}^*$ when $\epsilon \sim N(\mu, \sigma)$ as

$$\hat{\alpha}^* = 2\Phi(-\sigma/\sqrt{2}).$$

When σ is close to zero, we get that $\hat{\alpha}^* = 1 - \frac{1}{\sqrt{\pi}}\sigma + O(\sigma^3)$. This shows that with large errors, we end up with a much lower acceptance rate, and that the degradation in acceptance rate is worse than in the case where ϵ follows a Uniform distribution. For example, when $\sigma = 2$, we get that $\hat{\alpha}^* \approx 0.16$, but when $\sigma = 4$, we get that $\hat{\alpha}^* \approx 0.005$.