BAYESIAN INFERENCE FOR STOCHASTIC PROCESSES

Sean James Malory, MMath.

Submitted for the degree of Doctor of Philosophy in Statistics
Lancaster University
Supervised by Dr. Chris Sherlock

Dedicated to my family.

ABSTRACT

This thesis builds upon two strands of recent research related to conducting Bayesian inference for stochastic processes.

Firstly, this thesis will introduce a new residual-bridge proposal for approximately simulating conditioned diffusions formed by applying the modified diffusion bridge approximation of Durham and Gallant, 2002 to the difference between the true diffusion and a second, approximate, diffusion driven by the same Brownian motion. This new proposal attempts to account for volatilities which are not constant and can, therefore, lead to gains in efficiency over recently proposed residual-bridge constructs (Whitaker et al., 2017) in situations where the volatility varies considerably, as is often the case for larger inter-observation times and for time-inhomogeneous volatilities. These gains in efficiency are illustrated via a simulation study for three diffusions; the Birth-Death (BD) diffusion, the Lotka-Volterra (LV) diffusion, and a diffusion corresponding to a simple model of gene expression (GE).

Secondly, this thesis will introduce two new classes of Markov Chain Monte Carlo samplers, named the Exchangeable Sampler and the Exchangeable Particle Gibbs Sampler, which, at each iteration, use exchangeablility to simulate multiple, weighted proposals whose weights indicate how likely the chain is to move to such a proposal. By generalising the Independence Sampler and the Particle Gibbs Sampler respectively, these new samplers allow for the locality of moves to be controlled by a *scaling* parameter which can be tuned to optimise the mixing of the resulting MCMC procedure, while still benefiting from the increase in acceptance probability that typically comes with using multiple proposals. These samplers can lead to chains with better mixing properties, and, therefore, to MCMC estimators with smaller variances than their corresponding algorithms based on independent proposals. This improvement in mixing is illustrated, numerically, for both samplers through simulation studies, and, theoretically, for the Exchangeable Sampler through a result which states that, under certain conditions, the Exchangeable Sampler is geometrically ergodic even when the *importance* weights are unbounded and, hence, in scenarios where the Independence Sampler cannot be geometrically ergodic. To provide guidance in the practical implementation of such samplers, this thesis derives asymptotic expected squared-jump distance results for the Exchangeable Sampler and the Exchangeable Particle Gibbs Sampler. Moreover, simulation studies demonstrate, numerically, how the theory plays out in practice when d is finite.

ACKNOWLEDGMENTS

I would firstly like to thank my supervisor, Dr. Chris Sherlock, for his patience and support throughout my time as a PhD student. His knowledge, enthusiasm, ideas, and capacity for communicating complex material have made writing this thesis a pleasure; and its quality has been improved immensely by his guidance.

I would also like to thank my viva examiners, Dr. Chris Nemeth and Dr. Adam Johansen, not only for taking the time to read and review this thesis, but also for their reasoned guidance throughout the examination process.

My PhD has been supported by the EPSRC-funded Statistics and Operational Research (STOR-i) Doctoral Training Centre. I am very grateful to the Centre for the support I have received over the years, and for the opportunities it gave me to expand my skillset.

Finally, I would like to express my immense gratitude to my wife, Beth. Her unfailing support and strength through some very challenging times, along with her absolute dedication to raising our children, made finishing this thesis achievable. I could not have done it without her.

DECLARATION

I declare that the work in this thesis has been done by myself and has not been submitted elsewhere for the award of any other degree.

Sean James Malory

CONTENTS

1 THE INTRODUCTION 1

2 BACKGROUND MATERIAL 3
- 2.1 Concepts in Probability 3
  - 2.1.1 Sequences of Random Variables 3
  - 2.1.2 Exchangeability 5
  - 2.1.3 Markov Processes 7
  - 2.1.4 Diffusion Processes 9
- 2.2 Bayesian Inference 11
- 2.3 Monte Carlo Algorithms 12
  - 2.3.1 Idealised Algorithm 13
  - 2.3.2 Importance Sampling 15
  - 2.3.3 Normalized Importance Sampling 16
  - 2.3.4 Effective Sample Size 17
  - 2.3.5 Markov Chain Monte Carlo (MCMC) Algorithms 18
  - 2.3.6 Propose and Accept-Reject MCMC Algorithms 28
    - 2.3.6.1 The Independence Sampler 30
    - 2.3.6.2 The Random-Walk Sampler 32
    - 2.3.6.3 Optimal Scaling 37
- 2.4 The Filtering Problem 42
  - 2.4.1 The Particle Filter 44
    - 2.4.1.1 Sequential Importance Sampling 45
    - 2.4.1.2 Sequential Importance Resampling 46

3 SIMULATING CONDITIONED DIFFUSIONS 53
- 3.1 The Introduction 53
  - 3.1.1 The Birth-Death Diffusion 56
  - 3.1.2 The Lotka-Volterra Diffusion 56
  - 3.1.3 A Diffusion for a Simple Gene Expression Model 57
- 3.2 Simulating Conditioned Diffusions 57
  - 3.2.1 Absolute Continuity of Proposals 59
  - 3.2.2 Forward Simulation 59
  - 3.2.3 The Modified Diffusion Bridge 61
  - 3.2.4 The Fearnhead and Lindström Bridges 62
  - 3.2.5 Bridges Based on Residual Processes 66
  - 3.2.6 Bridges Based on Guided Proposals 70
- 3.3 New Bridges Based on Residual Processes 72
  - 3.3.1 Computational Considerations 75
  - 3.3.2 A Simulation Study 76
  - 3.3.3 Results 79
  - 3.3.4 Absolute Continuity 82
- 3.4 Summary 85

4 EXCHANGEABLE PARTICLE MCMC 87
- 4.1 The Introduction 87
- 4.2 Particle MCMC Algorithms 89
  - 4.2.1 The Pseudo-Marginal MH Sampler 92

4.2.2 Conditional Sequential Monte Carlo 96
4.2.3 Particle MH Samplers 103
4.2.4 The Particle Gibbs Sampler 108
4.3 The Exchangeable Sampler 114
4.3.1 Optimal Scaling 131
4.3.2 A Simulation Study 145
4.3.3 Results 150
4.4 The Exchangeable Particle Gibbs Sampler 167
4.4.1 Optimal Scaling 178
4.4.2 A Simulation Study 186
4.4.3 Results 191
4.5 Summary 197
5 CONCLUSION AND FURTHER WORK 203

BIBLIOGRAPHY 207
A PROOFS 217
B TECHNICAL LEMMATA 249
C EXTRA RESULTS 255

NOMENCLATURE

Abbreviations

SDE Stochastic Differential Equation

MC Monte Carlo

MCMC Markov Chain Monte Carlo

ESJD Expected Squared Jump Distance

PMCMC Particle Markov Chain Monte Carlo

PGS Particle Gibbs Sampler

xPGS Exchangeable Particle Gibbs Sampler

CPsMMCMC Correlated Pseudo-Marginal Markov Chain Monte Carlo

PGAS Particle Gibbs with Ancestor Sampling

EM Euler-Marayama

FS Forward Simulation

MDB Modified Diffusion Bridge

LNA Linear Noise Approximation

BD Birth-Death

LV Lotka-Volterra

GE Gene Expression

MSE Mean Squared Error

SMC Sequential Monte Carlo

xSMC Exchangeable Sequential Monte Carlo

CSMC Conditional Sequential Monte Carlo

CxSMC Conditional Exchangeable Sequential Monte Carlo

PIMH Particle Independent Metropolis-Hastings

PsMMH Pseudo-Marginal Metropolis-Hastings

PMMH Particle Marginal Metropolis-Hastings

Notation

$|A|$ Denotes the cardinality of the set A.

$A \subseteq B$ A is a subset of B.

$\mathcal{P}(B)$ Denotes the power set of a set B; that is, $\mathcal{P}(B) := \{A : A \subseteq B\}$.

A^c Denotes the complement of a set $A \subseteq \Omega$; that is,

$$A^c := \{\omega \in \Omega : \omega \notin A\} .$$

$x_{1:\infty}$ Shorthand for $x_1, x_2, \ldots$. Similarly $x^{(1:\infty)}$ is shorthand for $x^{(1)}, x^{(2)}, \ldots$, $x_{1:n}$ is shorthand for $x_1, x_2, \ldots, x_n$ etc. For dual scripts, the superscript takes precedent over the subscript so that, for instance, $x_{1:n}^{(1:m)}$ is shorthand for $x_1^{(1:m)}, \ldots, x_n^{(1:m)}$.

$\mathbb{N}$ The set of natural numbers not including 0.

μ - a.e. Denotes almost everywhere with respect to the measure μ defined on some measurable space $(\mathcal{X}, \mathcal{G}_\mathcal{X})$; that is, if $\phi(x)$ is some logical statement about some variable $x \in \mathcal{X}$, then $\phi(x)$, μ - a.e. if $\{x \in \mathcal{X} : \phi(x) \text{ is false.}\} \subseteq A \in \mathcal{G}_\mathcal{X}$ with $\mu(A) = 0$.

$\sigma(\Omega)$ Denotes the σ-algebra generated by the sets of Ω; that is the smallest σ-algebra containing all elements of Ω.

$\mathcal{B}(\mathbb{R}^d)$ Denotes the Borel sets of $\mathbb{R}^d$; that is,

$$\mathcal{B}(\mathbb{R}^d) := \sigma(\{(a, b] : a, b \in \mathbb{R}^d \text{ and } a < b\}) .$$

$x_n \downarrow x$ Denotes convergence from above; thats is, $x_n \to x$ as $n \to \infty$ and, for any $n \in \mathbb{N}$, $x_n \geq x$.

$a \leq b$ For any two vectors $a = (a_1, \ldots, a_d)$, $b = (b_1, \ldots, b_d) \in \mathbb{R}^d$, $a \leq b$ if and only if $a_i \leq b_i$ for each $i = 1, \ldots, d$. Similarly for other inequalities. The definition of intervals such as $[a, b]$ obey this ordering.

$f^{-1}(A)$ Denotes the pre-image of a set A with respect to the function $f : \mathcal{X} \to \mathcal{Y}$; that is $f^{-1}(A) := \{x \in \mathcal{X} : f(x) \in A\}$.

$\mathbb{1}_A(x)$ Denotes the indicator function corresponding to a set A; that is,

$$\mathbb{1}_A(x) := \begin{cases} 0 & \text{if } x \notin A , \\ 1 & \text{if } x \in A . \end{cases}$$

$\mathbb{P}(\phi(X))$ Denotes, for a logical statement $\phi(x)$ about a random variable X defined on $(\mathcal{X}, \mathcal{G}_\mathcal{X}, \mathbb{P})$,

$$\mathbb{P}(\{w \in \mathcal{X} : \phi(X(w))\}) .$$

$\underset{n\to\infty}{\text{dlim}}$ Denotes the distributional limit of the sequence of random variables $X_{1:\infty}$.

$\underset{n\to\infty}{\text{plim}}$ Denotes the probabilistic limit of the sequence of random variables $X_{1:\infty}$.

$\underset{n\to\infty}{\text{aslim}}$ Denotes the almost sure limit of the sequence of random variables $X_{1:\infty}$.

Permutation A permutation, σ, of a finite set, $\Omega := \{1, \ldots, N\}$, is a one-to-one mapping from Ω onto itself.

$[A]_{i,j}$ Denotes the (i, j)-th entry of the matrix A.

$\text{diag}(x_{1:d})$ Denotes a d-dimensional diagonal matrix with entries $x_{1:d}$; that is, a matrix $X \in \mathbb{R}^{d\times d}$ such that $[X]_{i,j} = 0$ for any $i \neq j$ and $[X]_{i,i} = x_i$ for any $i \in \{1, \ldots, d\}$.

$\text{T}(\nu)$ Denotes the T-distribution with ν degrees of freedom; that is, a continuous distribution with density

$$\frac{\Gamma(0.5(\nu+1))}{\sqrt{\nu\pi}\Gamma(0.5\nu)}\left(1+\frac{x^2}{\nu}\right)^{-0.5(1+\nu)}, \quad x \in \mathbb{R}\,.$$

Δy_k Denotes $y_{k+1} - y_k$.

$\text{Exp}(\lambda)$ Denotes the Exponential distribution with rate λ; that is a continuous distribution with density

$$\lambda \exp(-\lambda x)\,, \quad x \in [0, \infty)\,.$$

$o(h(t))$ Denotes little order; that is, $f(t) = g(t) + o(h(t))$ if

$$\lim_{t\downarrow 0}[f(t) - g(t)]/h(t) = 0\,.$$

A^T and A^* Denote the transpose of the matrix A.

$\text{N}_k(\mu, \Sigma)$ Denotes the k-dimensional Normal distribution with mean μ and variance matrix Σ (in one dimension the mean and variance will both be scalars); that is a continuous distribution with density

$$(2\pi)^{-k/2}|\det(\Sigma)|^{-1/2}\exp[-(x-\mu)^T\Sigma^{-1}(x-\mu)/2]\,, \quad x \in \mathbb{R}^k\,.$$

$\pi[h]$ Denotes the expectation of the function $h : \mathbb{R}^d \to \mathbb{R}^r$ with respect to the distribution π; that is the r-dimensional vector whose i-th component is given by

$$\pi[h_i] = \int_{\mathbb{R}^d} h_i(x)\pi(\mathrm{d}x)\,.$$

If $r = 1$ then $\pi[h]$ denotes a scalar.

$\mathrm{Var}_\pi(h(X))$ Denotes the variance of the function $h : \mathbb{R}^d \to \mathbb{R}^r$ with respect to the distribution π; that is the matrix of size $r \times r$ whose (i, j)-th component is

$$\mathrm{Var}_\pi(h(X))_{ij} = \pi[h_i h_j] - \pi[h]_i \pi[h]_j \ .$$

If $r = 1$ then $\mathrm{Var}_\pi(h(X))$ denotes a scalar.

$\mathrm{supp}(f)$ Denotes the support of a function f; that is,

$$\mathrm{supp}(f) = \{x : f(x) \neq 0\} \ .$$

$\mathcal{I}_d$ Denotes the d-dimensional identity matrix; that is the matrix whose (i, j)-th element is 1 if $i = j$ and 0 otherwise.

$\mu_1 \ll \mu_2$ Denotes, for two probability measures, μ_1 and μ_2 defined on the same measurable space $(\mathcal{X}, \mathcal{G}_\mathcal{X})$, that μ_1 is absolutely continuous with respect to μ_2.

$\|\mu_1 - \mu_2\|$ Denotes, for two probability measures, μ_1 and μ_2 defined on the same measurable space $(\mathcal{X}, \mathcal{G}_\mathcal{X})$, the total variation distance between the two measures; that is,

$$\|\mu_1 - \mu_2\| := \sup_{A \in \mathcal{G}_\mathcal{X}} |\mu_1(A) - \mu_2(A)| \ .$$

$\mathrm{Cov}(h(X), g(Y))$ Denotes the covariance of the functions $h : \mathbb{R}^d \to \mathbb{R}$ and $g : \mathbb{R}^d \to \mathbb{R}$ of two random variables, X and Y; that is, if X, Y have a joint distribution π and marginal distributions π_X and π_Y respectively, then

$$\mathrm{Cov}(h(X), g(Y)) = \pi[hg] - \pi_X[h]\pi_Y[g] \ .$$

$\mathrm{Corr}(h(X), g(Y))$ Denotes the correlation of the functions $h : \mathbb{R}^d \to \mathbb{R}$ and $g : \mathbb{R}^d \to \mathbb{R}$ of two random variables, X and Y; that is, if X, Y have a joint distribution π and marginal distributions π_X and π_Y respectively, then

$$\mathrm{Corr}(h(X), g(Y)) = \mathrm{Cov}(h(X), g(Y)) / \sqrt{\mathrm{Var}_{\pi_X}(h(X))\mathrm{Var}_{\pi_Y}(g(Y))} \ .$$

$\langle \cdot, \cdot \rangle_\pi$ Denotes the inner-product associated with the Hilbert space of functions which are square-integrable with respect to π, which is a measure on some measurable space $(\mathcal{X}, \mathcal{G}_\mathcal{X})$; that is, for any $f, g \in \mathrm{L}^2(\pi)$ where

$$\mathrm{L}^2(\pi) := \{f : \mathcal{X} \to \mathbb{R} \text{ such that } \pi[f^2] < \infty\} \ ,$$

define

$$\langle f, g \rangle_\pi := \int_{\mathcal{X}^2} \pi(\mathrm{d}x) f(x) g(x) \ .$$

$\text{Unif}(\mathcal{A})$ Denotes the discrete uniform distribution over the finite set $\mathcal{A}$; that is, the distribution with probability mass function, f, such that, for any $a \in \mathcal{A}$, $f(a) = 1/|\mathcal{A}|$.

$\text{Unif}(a, b)$ Denotes the Uniform distribution on the interval $(a, b]$; that is, a continuous distribution with density

$$(b-a)^{-1} , \quad x \in (a, b] .$$

$\text{Gamma}(\alpha, \beta)$ Denotes the Gamma distribution with shape α and rate β; that is, a continuous distribution with density

$$\frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} \exp(-\beta x) \, \mathrm{d}x , \quad x \in [0, \infty) .$$

$a \wedge b$ Denotes the minimum of a and b; that is $a \wedge b = b$ if $a \geq b$ and $a \vee b = a$ otherwise.

$a \vee b$ Denotes the maximum of a and b; that is $a \vee b = a$ if $a \geq b$ and $a \vee b = b$ otherwise.

$\lceil x \rceil$ Denotes the *ceiling* of $x \in \mathbb{R}$; that is,

$$\lceil x \rceil = \min\{k \in \mathbb{Z} : k \geq x\} .$$

$\lfloor x \rfloor$ Denotes the integer part of $x \in \mathbb{R}$; that is,

$$\lfloor x \rfloor := \max\{k \in \mathbb{Z} : k \leq x\} .$$

δ_x Denotes, for a measurable space $(\mathcal{X}, \mathcal{G}_\mathcal{X})$, the Dirac measure centred on x.

1 THE INTRODUCTION

Inference for processes which are stochastic in nature, or, by virtue of incomplete information, can be modelled as such, is an increasingly common task in many academic disciplines, including, but not limited to, biology (see, for instance, Wilkinson, 2006; Kuhner, 2006; Boys, Wilkinson, and Kirkwood, 2008), epidemiology (Neal and Roberts, 2004; Ball and Neal, 2008; Jewell, Keeling, and Roberts, 2008), physics (see, for example, Cancès, Legoll, and Stoltz, 2007; Akeret et al., 2015; Lelièvre and Stoltz, 2016), and economics (see, for instance, Glasserman, 2010; Rambharat and Brockwell, 2010; Sen, Jasra, and Zhou, 2017). The arrival of cheap computational resources at the end of the twentieth century has enabled practitioners to conduct statistical inference for more complex, and arguably, therefore, more realistic stochastic processes, leading to significant advancements in the field.

This thesis will concentrate on Bayesian inference for stochastic processes, with a specific focus on the challenges involved in conducting inference for *diffusions* driven by Stochastic Differential Equations (SDEs). In particular, this thesis is concerned with the construction of stochastic approximations to expectations, $\mathbb{E}[f(x)]$, defined with respect to a *target* distribution, π, based on *averaging*, or, more formally, Monte Carlo (MC) techniques (see, for example, Liu, 2001; Andrieu et al., 2003; Glasserman, 2010; Doucet et al., 2001; Robert and Casella, 2004). Such techniques can be favourable over their deterministic counterparts since, firstly, under appropriate conditions, they retain an *error* independent of the dimension of the probability space on which π resides (see Section 2.3 of this thesis, or, for example, Corollary 2.1, Roberts and Rosenthal, 1997, Section 1.3.1, Doucet et al., 2001, or Section 1.7, Brooks et al., 2011), secondly, they are, in a sense that shall be made clear in Section 2.3 (and is clear from the aforementioned references), independent of the function f[1], and, thirdly, recent approaches, which utilise the sequential structure of certain distributions, have vastly extended the scope of such techniques (see, for example, Andrieu, Doucet, and Holenstein, 2010; Lindsten, Jordan, and Schön, 2014; Chopin and Singh, 2015).

After introducing the necessary background material in Chapter 2, this thesis will build upon two strands of recent research related to conducting Bayesian inference for stochastic processes. Firstly, in Chapter 3, this thesis will introduce a new residual-bridge proposal for approximately simulating conditioned diffusions formed by applying the modified diffusion bridge approximation of Durham and Gallant, 2002 to the difference between the true diffusion and a second, approximate, diffusion driven by the same Brownian motion. This new proposal attempts

1 It is the logic of the algorithm itself that is independent of f, not the error of the approximation.

to account for volatilities which are not constant and can, therefore, lead to gains in efficiency over recently proposed residual-bridge constructs (Whitaker et al., 2017) in situations where the volatility varies considerably, as is often the case for larger inter-observation times and for time-inhomogeneous volatilities. These gains in efficiency are illustrated via a simulation study (see Section 3.3.2).

Secondly, in Chapter 4, this thesis will introduce two new classes of Markov Chain Monte Carlo samplers named the Exchangeable Sampler and the Exchangeable Particle Gibbs Sampler, which, at each iteration, use exchangeablility to simulate multiple, weighted proposals whose weights indicate how likely the chain is to move to such a proposal. By generalising the Independence Sampler and the Particle Gibbs Sampler respectively, these new samplers allow for the locality of moves to be controlled by a *scaling* parameter which can be tuned to optimise the mixing of the resulting MCMC procedure, in a manner reminiscent of MCMC algorithms based on random walks (see, for instance, Roberts, Gelman, and Gilks, 1997; Roberts and Rosenthal, 1998b, 2001; Sherlock and Roberts, 2009), while still benefiting from the increase in acceptance probability that typically comes with using multiple proposals. As a result, these samplers can lead to chains with better mixing properties, and, therefore, to MCMC estimators with smaller variances than their corresponding algorithms based on independent proposals. This improvement in mixing is illustrated, numerically, for both samplers through simulation studies (see Sections 4.3.2 and 4.4.2), and, theoretically, for the Exchangeable Sampler through Corollary 4.3.7 which proves that, under certain conditions, the Exchangeable Sampler is geometrically ergodic even when the *importance* weights are unbounded and, hence, in scenarios where the Independence Sampler cannot be geometrically ergodic (see, for example, Mengersen and Tweedie, 1996; Atchadé and Perron, 2007 for proofs that bounded weights is necessary for geometric ergodicity). Moreover, to provide guidance in the practical implementation of such samplers, this thesis will investigate the optimal *scaling* parameter for both the Exchangeable Sampler and the Exchangeable Particle Gibbs Sampler in a manner similar to optimal scaling results in the literature (see, for instance, Roberts, Gelman, and Gilks, 1997; Roberts and Rosenthal, 1998b, 2001; Sherlock and Roberts, 2009). In particular, Theorem 4.3.17 of Section 4.3.1, and Theorem 4.4.7 of Section 4.4.1 derive an asymptotic (as the dimension, d, of the state space tends towards infinity) expected squared-jump distance result for the Exchangeable Sampler and the Exchangeable Particle Gibbs Sampler respectively. Moreover, Section 4.3.1 and Section 4.4.1 provide simulation studies numerically demonstrating how the theory plays out in practice when d is finite.

2 BACKGROUND MATERIAL

2.1 CONCEPTS IN PROBABILITY

The ideas in this thesis rely on foundational concepts in probability, which, for completeness, are briefly introduced in this section. In Section 2.1.1 we introduce the concept of a sequence of random variables, define several notions of convergence of such a sequence, and also state Bayes' theorem. We introduce, in Section 2.1.2, the concept of exchangeability- a concept which will be fundamental when we discuss a novel generalization of current state of the art Markov Chain Monte Carlo methods in Chapter 4. In Section 2.1.3 we introduce the idea of a random (stochastic) process whose future behaviour is dependent on the past only through the present and, finally, in Section 2.1.4, the diffusion process, whose inference is the focus of this thesis, is introduced and characterised.

2.1.1 *Sequences of Random Variables*

Throughout this thesis it will be convenient to discuss a sequence of random variables, defined via a sequence of conditional random variables, without mentioning the underlying probability space on which it resides. Such discussions only make sense if such a space exists. To this end, let $X_1 \sim \pi_1$ be a d_1-dimensional random variable defined on the space $(\mathbb{R}^{d_1}, \mathcal{B}(\mathbb{R}^{d_1}), \pi_1)$, where $\mathcal{B}(\mathbb{R}^{d_1})$ denotes the Borel sets of $\mathbb{R}^{d_1}$. For each $j \in \mathbb{N}$, and each $x_{1:j} \in \mathbb{R}^{d_1} \times \ldots \times \mathbb{R}^{d_j}$, let $X_{j+1}|X_{1:j} = x_{1:j} \sim \pi_{X_{j+1}|x_{1:j}}$ be a d_{j+1}-dimensional random variable, defined on the measurable space $(\mathbb{R}^{d_{j+1}}, \mathcal{B}(\mathbb{R}^{d_{j+1}}))$. Suppose further that the mapping $\pi_{X_{j+1}|X_{1:j}} : \mathbb{R}^{d_1} \times \ldots \times \mathbb{R}^{d_j} \times \mathcal{B}(\mathbb{R}^{d_{j+1}}) \to [0,1]$, defined by $\pi_{X_{j+1}|X_{1:j}}(x_{1:j}, A) := \pi_{X_{j+1}|x_{1:j}}(A)$, is measurable in the first argument for any fixed $A \in \mathcal{B}(\mathbb{R}^{d_{j+1}})$. Then, for any $j \in \mathbb{N}$, one can talk about a sequence of random variables $X_1, \ldots, X_j$ and a probability measure $\mathbb{P}$ such that for any $A^j \in \sigma(\mathcal{B}(\mathbb{R}^{d_1}) \times \ldots \times \mathcal{B}(\mathbb{R}^{d_j}))$,

$$\mathbb{P}(X_{1:j} \in A^j) = \int_{\mathbb{R}^{d_1}} \pi_1(\mathrm{d}x_1) \int_{\mathbb{R}^{d_2}} \pi_{X_2|x_1}(\mathrm{d}x_2) \ldots \int_{\mathbb{R}^{d_j}} \mathbb{1}_{A^j}(x_{1:j}) \pi_{X_j|x_{1:j-1}}(\mathrm{d}x_j) \,.$$

A formal statement and proof of the existence of such a measure on an appropriately defined probability space follows from the Infinite-Dimensional Product Measure Theorem (see, for example, Theorem 2.7.2, Ash and Doléans-Dade, 2000). With this *joint* measure in place, one can derive a natural definition of the marginal random variables, X_j, for each $j \in \mathbb{N}$. Indeed, for any $j \in \mathbb{N}$, let $X_j \sim \pi_j$ be the d_j-

dimensional random variable which is defined, for any $A_j \in \mathcal{B}(\mathbb{R}^{d_j})$, by

$$\begin{aligned}\pi_j(A_j) &:= \mathbb{P}\left(\left\{w \in \prod_{r=1}^{\infty} \mathbb{R}^{d_r} : w_j \in A_j\right\}\right) \\ &= \int_{\mathbb{R}^{d_1}} \pi_1(\mathrm{d}x_1) \int_{\mathbb{R}^{d_2}} \pi_{X_2|x_1}(\mathrm{d}x_2) \ldots \int_{A_j} \pi_{X_j|x_{1:j-1}}(\mathrm{d}x_j) \; .\end{aligned}$$

Thus, $X_{1:\infty}$ is a sequence of random variables all defined on the common probability space constructed only by specifying the distribution π_1 of a random variable X_1, and, for each $j \in \mathbb{N}$ and each $(x_{1:j}) \in \mathbb{R}^{d_1} \times \ldots \times \mathbb{R}^{d_j}$, the distribution of a d_{j+1}-dimensional random variable denoted by $X_{j+1}|X_{1:j} = x_{1:j}$, and thought of as the random variable X_{j+1} conditioned on the knowledge that $X_{1:j} = x_{1:j}$.

Given the density π_1 of a random variable X_1 and the conditional density $\pi_{X_2|X_1=x_1}$ corresponding to the conditional random variable $X_2|X_1 = x_1$, one can, via Bayes' Theorem (see, for instance, Section 7.3, Papoulis, Pillai, and Pillai, 2002), derive the density corresponding to the conditional random variable $X_1|X_2 = x_2$;

$$g(x_1|x_2) = \frac{f(x_2|x_1)\pi_1(x_1)}{\pi_2(x_2)} \; .$$

Finally, it will be useful to introduce several notions of convergence, (see, for instance, Ash and Doléans-Dade, 2000, Durrett, 2010, or Capinski and Kopp, 2013) that will be discussed throughout this thesis;

DEFINITION 2.1.1 (Convergence of Random Variables). *A sequence of random variables, not necessarily defined on a common probability space, $X_{1:\infty}$, with distribution functions $F_1, F_2, \ldots$ is said to converge in distribution to a random variable, X, with distribution function F, written $\operatorname{dlim}_{n\to\infty} X_n = X$, if and only if, for every $x \in \mathbb{R}^d$ at which F is continuous,*

$$\lim_{n\to\infty} F_n(x) = F(x) \; .$$

Suppose now that the random variables, $X_{1:\infty}$, X, are defined on a common probability space $(\mathcal{X}, \mathcal{G}_{\mathcal{X}}, \mathbb{P})$. Then, the sequence $X_{1:\infty}$ is said to converge in probability to X, written $\operatorname{plim}_{n\to\infty} X_n = X$ if and only if, for every $\epsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}(|X_n - X| \geq \epsilon) = 0 \; .$$

The sequence $X_{1:\infty}$ is said to converge almost surely to X, written $\operatorname{aslim}_{n\to\infty} X_n = X$, if and only if

$$\lim_{n\to\infty} X_n(w) = X(w) \; , \quad \mathbb{P}\text{-}a.e. \; .$$

2.1.2 *Exchangeability*

A sequence of random variables is said to be exchangeable if their joint distribution is independent of their order (see Definition 2.1.2 for a definition in the case of finite sequences). While exchangeability is closely related to independence (see, for instance, de-Finetti's Theorem, de Finetti, 1931, and its extensions in Hewitt and Savage, 1955, and Diaconis and Freedman, 1980), the extra flexibility afforded by exchangeable random variables allows one to simulate identically distributed random variables whose *closeness* can be parametrized (see, for example, Lemma 2.1.3 and Algorithm 1). It is this property which allows one to generalise current state of the art Markov Chain Monte Carlo methods, which rely on independent proposals, to methods which utilise exchangeability and, therefore, can benefit from tailored *jump-sizes* (see Chapter 4 of this thesis).

DEFINITION 2.1.2. *A sequence $X_{1:N}$ of random variables, $X_{1:N} \in \mathcal{X}^N$ is said to be exchangeable if, for any $A_{1:N} \in \mathcal{G}_{\mathcal{X}}^N$,*

$$\mathbb{P}(X_1 \in A_1, \ldots, X_N \in A_N) = \mathbb{P}(X_{\sigma(1)} \in A_1, \ldots, X_{\sigma(N)} \in A_N) \,,$$

for any permutation[1], σ, of the set $\{1, \ldots, N\}$. In other words, if the joint probability density function, π, exists, then

$$\pi(x_{1:N}) = \pi(x_{\sigma(1)}, \ldots, x_{\sigma(N)}) \,,$$

for any permutation σ of the set $\{1, \ldots, N\}$.

It is fairly clear that independent and identically distributed random variables are also exchangeable. Moreover, exchangeable random variables are identically distributed, but not necessarily independent. These results are essentially consequences of de-Finetti's Theorem and the extensions thereof (see, for instance, de Finetti, 1931, Hewitt and Savage, 1955, and Diaconis and Freedman, 1980). Intuitively, a sequence of random variables is exchangeable if the order they are simulated in does not matter. The results of de Finetti, 1931 and Hewitt and Savage, 1955 show, not only that exchangeability can be achieved through independence given some underlying random variable, but also, in some sense, that this is the only way of obtaining an exchangeable sequence. Indeed, let f_0 be a prior distribution for a random variable $\Theta \in \mathcal{T}$, and, given $\Theta = \theta$, let f^* be a joint distribution for a sequence of independent and identically distributed random variables $X_{1:N}$ each with marginal distribution f. Then, the sequence $X_{1:N}$, whose joint distribution is π^*, say, is exchangeable but not independent. Indeed,

$$\pi^*(x_{1:N}) = \int_{\mathcal{T}} f_0(\theta) \prod_{i=1}^{N} f(x_i|\theta) \, \mathrm{d}\theta \,, \tag{1}$$

1 A permutation, σ, of a finite set, $\Omega := \{1, \ldots, N\}$, is a one-to-one mapping from Ω onto itself.

which is clearly exchangeable due to the independence of the x_i within the integral. Moreover, the marginal distributions are given by

$$\pi(x) = \int_{\mathcal{T}} f_0(\theta) f(x|\theta) \, \mathrm{d}\theta \, ,$$

and, in general,

$$\pi^*(x_{1:N}) \neq \prod_{i=1}^{N} \int_{\mathcal{T}} f_0(\theta) f(x_i|\theta) \, \mathrm{d}\theta \, .$$

That is, the sequence is, in general, not independent. The representation given by (1) provides a way of simulating N exchangeable random variables. First, simulate a θ from $f_0(\cdot)$, then simulate N independent and identically distributed random variables $X_{1:N}$ such that, for any $i \in \{1, \ldots, N\}$, $X_i \sim f(\cdot|\theta)$. A particularly useful choice of f_0 and f allows for the generation of exchangeable normal random variables via the simulation of other normal random variables. Indeed, let $\Theta \sim \mathrm{N}_d(\mu, \Sigma)$ for some $\mu \in \mathbb{R}^d$ and $\Sigma \in \mathbb{R}^{d \times d}$, and, for any $i \in \{1, \ldots, N\}$, $X_i|\theta \sim \mathcal{N}_d(A\theta + b, C)$ where $A \in \mathbb{R}^{d \times d}$, $b \in \mathbb{R}^d$, and $C \in \mathbb{R}^{d \times d}$ is a diagonal matrix with diagonal entries $c_{1:d}$; that is, $[C]_{i,j} = 0$ for any $i \neq j$ and $[C]_{i,i} = c_i$ for any $i \in \{1, \ldots, d\}$. Then, $X_i \sim \mathrm{N}_d(A\mu + b, A\Sigma A^T + C)$. The usefulness of this choice stems from the fact that normal random variables are well understood and easy to simulate. Moreover, the expected squared Euclidean distance between the X_i is tractable;

LEMMA 2.1.3. *Let $X_{1:N}|\Theta = \theta$ be a sequence of independent random variables given some underlying random variable θ such that, for any $i \in \{1, \ldots, N\}$, $X_i|\Theta = \theta \sim \mathrm{N}_d(A\theta + b, C)$ where $A \in \mathbb{R}^{d \times d}$, $b \in \mathbb{R}^d$, and $C = \mathrm{diag}(c_{1:d}^2) \in \mathbb{R}^{d \times d}$ is a diagonal matrix with entries $c_{1:d}^2$; that is, $[C]_{i,j} = 0$ for any $i \neq j$ and $[C]_{i,i} = c_i^2$ for any $i \in \{1, \ldots, d\}$. Then, for any $i \neq j$, and any $k \in \{1, \ldots, N\}$,*

$$\mathbb{E}[\|X_i^{(k)} - X_j^{(k)}\|^2] = 2c_k^2 \, ,$$

where $X^{(k)}$ denotes the k-th component of the random variable X.

Proof. See A.1. □

As a result of this lemma, one can simulate a sequence of exchangeable d-dimensional normal random variables whose expected *closeness* in dimension k can be controlled. Indeed, in one dimension, the following procedure (Algorithm 1), which is at the heart of the exchangeable Markov Chain Monte Carlo methods introduced in Chapter 4 of this thesis, demonstrates how one can simulate what this thesis terms an ϵ-close exchangeable sequence; that is, a sequence $Z_{1:N}$ of standard normal random variables whose *closeness*, in terms of the square-root expected squared distance is equal to a *jump-size*, $\epsilon \in [0, 1]$;

Algorithm 1 Simulate an ϵ-Close Sequence of Exchangeable Standard Normal Random Variables

1: Let $\epsilon \in [0,1]$ and set $\delta = \epsilon/\sqrt{2}$.
2: Sample θ from a $\mathrm{N}(0,1)$ distribution.
3: **for** $i = 1, \ldots, N$ **do**
4: Sample $\hat{z}_i$ from a $\mathrm{N}(0,1)$ distribution.
5: Set $z_i = \theta\sqrt{1-\delta^2} + \delta\hat{z}_i$.
6: **end for**

REMARK 1. *The procedure outlined in Algorithm 1 is a multiple-sample extension of the preconditioned Crank-Nicolson proposal introduced in Cotter et al., 2013.*

Consider, for any $i \neq j$, Z_i and Z_j simulated by this procedure. By Lemma 2.1.3,

$$\sqrt{\mathbb{E}[\|Z_i - Z_j\|^2]} = \sqrt{2}\delta = \epsilon \, .$$

Trivially, if $X_{1:N}$ are exchangeable; that is, the order in which they are simulated does not matter, then, for any measurable function $f : \mathcal{X} \to \mathbb{R}^d$, the sequence $f(X_1), \ldots, f(X_N)$ is exchangeable; for one can simulate the X_i in any order and apply the function f to each X_i independently[2]. In one-dimension, this fact can be used in conjunction with Theorem 2.3.2 to simulate a sequence of exchangeable random variables, $X_{1:N}$, where each $X_i \sim \pi$, provided π is a distribution whose corresponding distribution function, $F(x) := \pi((-\infty, x])$, is invertible for every $x \in \mathbb{R}$. Indeed, by Theorem 2.3.2, $\Phi(Z_i) \sim \mathrm{Unif}(0,1)$. Hence,

Algorithm 2 Simulation of General Exchangeable Sequences

1: Let Φ denote the distribution function of a standard normal random variable.
2: Simulate $z_{1:N}$ via Algorithm 1.
3: Set $x_i = F^{-1}(\Phi(z_i))$ for each $i \in \{1, \ldots, N\}$.

$X_i := F^{-1}(\Phi(Z_i)) \sim \pi$. In general, the X_i will not be ϵ-close. However, a *closeness* of at most ϵ could be enforced if the function $F^{-1}(\Phi(\cdot))$ were suitably smooth. For example, if $F^{-1}(\Phi(\cdot))$ were Lipschitz continuous with Lipschitz constant $a > 0$, then, for any $i \neq j$,

$$\mathbb{E}[\|X_i - X_j\|^2] \leq a^2\mathbb{E}[\|Z_i - Z_j\|^2] = a^2\epsilon^2 \, .$$

Therefore, if an ϵ/a-close sequence $Z_{1:N}$ were simulated via Algorithm 1, then the *closeness* of the X_i would be at most ϵ.

2.1.3 *Markov Processes*

Processes which exhibit random movements over time; that is, stochastic processes, are formalised by a collection of random variables indexed by time;

2 This is, again, essentially a consequence of de-Finetti's Theorem and the extensions thereof. See, for instance, de Finetti, 1931, Hewitt and Savage, 1955, and Diaconis and Freedman, 1980

DEFINITION 2.1.4 (Stochastic Process). *A d-dimensional stochastic process (henceforth, process) is a collection of random variables, $\{X_t : t \in \mathcal{T}\}$ defined on a common probability space $(\mathcal{X}, \mathcal{G}_\mathcal{X}, \mathbb{P})$, where, for the purposes of this thesis, $\mathcal{T}$ is either $\{1, \ldots, T\} \subseteq \mathbb{N}$ or $[0, T] \subseteq \mathbb{R}_+$ for some, potentially infinite, T. For any $\omega \in \mathcal{X}$, the function $t \to X_t(\omega)$ is called a sample path of the process.*

As highlighted in Section 2.1.1, it is possible, using the Infinite-Dimensional Product Theorem (see, for example, Theorem 2.7.2, Ash and Doléans-Dade, 2000), to demonstrate the existence of a common underlying probability space on which a sequence of random variables, $X_{1:\infty}$, and, therefore, a stochastic process with a countable index set, $\mathcal{T} = \{1, \ldots, T\}$, is defined. Provided a *consistent* set of finite-dimensional probability measures exists, Kolmogorov's Extension Theorem (see, for instance, Theorem 2.7.5, Ash and Doléans-Dade, 2000) provides a similar result for stochastic processes with an uncountable index set, $\mathcal{T} = [0, T]$. The sets $\mathbb{N}$ and $[0, \infty)$ are ordered. Thus, stochastic processes are naturally ordered and, therefore, for any time $t \in \mathcal{T}$, the history of the process up to that time makes sense as a concept and, formally, is encapsulated in the natural filtration;

DEFINITION 2.1.5 (Natural Filtration). *The natural filtration corresponding to a d-dimensional process $\{X_t : t \in \mathcal{T}\}$ is defined to be the collection of sets $\{\mathcal{F}_t^X : t \in \mathcal{T}\}$ such that, for any $t \in \mathcal{T}$,*

$$\mathcal{F}_t^X := \sigma(\{X_s^{-1}(\mathcal{B}(\mathbb{R}^d)) : \ s \leq t\}) \ .$$

The natural filtration for a process $\{X_t : t \in \mathcal{T}\}$, therefore, is a sequence of σ-algebras, $\{\mathcal{F}_t^X : t \in \mathcal{T}\}$, such that, for any $t \in \mathcal{T}$, $\mathcal{F}_t^X$ is the smallest σ-algebra that ensures that X_s is measurable with respect to $\mathcal{F}_t^X$, for any $s \leq t$. Many processes of interest are such that their future behaviour depends only on their most recent *accessible* value, and these processes are said to be weakly Markov. Specifically, the behaviour of a weakly Markov process at time $t \in \mathcal{T}$ depends on its behaviour up to time $s < t$ only through the processes value at time s;

DEFINITION 2.1.6 (Weakly Markov Process). *A process $\{X_t : t \in \mathcal{T}\}$ is weakly Markov (henceforth, Markov) if, for any $U \in \mathcal{B}(\mathbb{R}^d)$ and $(s, t) \in \mathcal{T}^2$, with $s < t$,*

$$\mathbb{P}(X_t^{-1}(U)|\mathcal{F}_s^X) = \mathbb{P}(X_t^{-1}(U)|\sigma(X_s^{-1}(\mathcal{B}(\mathbb{R}^d)))) \ .$$

The support of a process is the space of all possible values that the process could take. Formally,

DEFINITION 2.1.7. *The support of a process $\{X_t : t \in \mathcal{T}\}$ is defined by*

$$\bigcup_{t \in \mathcal{T}} \{y \in \mathbb{R}^d : \text{there exists an open } B \in \mathcal{B}(\mathbb{R}^d) \text{ with } y \in B \text{ and } \mathbb{P}(X_t^{-1}(B)) > 0\} \ .$$

Henceforth, for brevity, a Markov process $\{X_t : t \in \{1, \ldots, T\}\}$ will be called a Markov chain and the term Markov process will be exclusively reserved for a process $\{X_t : t \in [0, T]\}$. In either case, since the index set is implicit, the chain/process will simply be referred to as X_t.

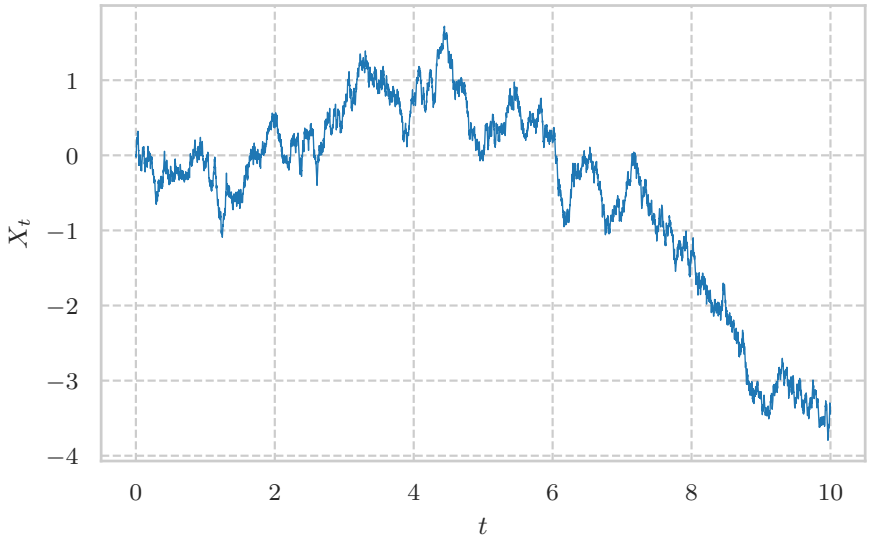

Figure 1: A sample path of a Wiener process.

2.1.4 *Diffusion Processes*

The fundamental diffusion process is the Wiener process (see Wiener, 1923, or, for example, Definition 9.2.1, Ash and Doléans-Dade, 2000, Definition 3.3.1, Shreve, 2004, and Section 8.1, Durrett, 2010), a sample path of which can be seen in Figure 1;

DEFINITION 2.1.8. *A one-dimensional Wiener process is a process, W_t, with the following properties;*

(W1) $W_0 = 0$.

(W2) For any $0 \leq t_1 < t_2$, $W_{t_2} - W_{t_1} \sim \mathrm{N}(0, t_2 - t_1)$.

(W3) For any $0 \leq t_1 < t_2 \leq t_3 < t_4$, *the random variables* $W_{t_4} - W_{t_3}$ *and* $W_{t_2} - W_{t_1}$ *are independent.*

(W4) Sample paths of W_t *are continuous.*

A d-dimensional Wiener process is a vector of d one-dimensional independent Wiener processes.

Following Section 7.1, Durrett, 2010, a process satisfying properties (W1), (W2), and (W3) can be constructed using the Kolmogorov Extension Theorem. Indeed, for any $n \in \mathbb{N}$, consider a sequence of times $t_{1:n} \in [0, T]^n$ such that $0 := t_0 \leq t_1 < t_2 < \ldots < t_N$. Define, for any permutation τ of $\{1, \ldots, n\}$, the probability measures $P_{t_{\tau(1)},\ldots,t_{\tau(n)}}$, to be such that, for any $B \in \sigma(\mathcal{B}(\mathbb{R})^n)$,

$$P_{t_{\tau(1)},\ldots,t_{\tau(n)}}(B) := \int_B \prod_{k=0}^{n-1} (2\pi\Delta t_k)^{-1/2} \exp\left(-\frac{(\Delta w_{t_k})^2}{2\Delta t_k}\right) \mathrm{d}w_{t_1} \ldots \mathrm{d}w_{t_n} \ ,$$

where $\Delta t_k := t_{k+1} - t_k$ and $\Delta w_k := w_{k+1} - w_k$. The existence of the Wiener process, W_t, then follows by Kolmogorov's Continuity Theorem (see, for instance, Theorem 2.1.6, Stroock and Varadhan, 1997), and the fact that, for any $s < t$, $W_t - W_s \sim \mathrm{N}(0, t-s)$.

A different view of the Wiener process (known as Donsker's invariance principle; see for example, Donsker, 1951, Theorem 37.8, Billingsley, 1995, or Theorem 12.9, Kallenberg, 1997), which will motivate the general diffusion process, can be seen by considering, for a small, fixed, time increment, Δt, the behaviour of the process at times $t_k := k\Delta t$ for $k \in \mathbb{N}$. Indeed, consider the one-dimensional *random walk* Markov chain, Y_t, defined by $Y_0 = 0$, and, for any $k \in \{1, 2, \ldots\}$,

$$Y_{k+1} = Y_k + \Delta W_{k+1} \,, \tag{2}$$

where $\Delta W_{k+1} := W_{t_{k+1}} - W_{t_k}$. Define the process $W_t^{\Delta t} := Y_{\lfloor t/\Delta t \rfloor}$, then, in a sense which, for the purposes of this thesis, is intentionally left vague, $W_t^{\Delta t}$ converges to the Wiener process as $\Delta t \downarrow 0$. Therefore, defining the integral with respect to the Wiener process as an appropriate limit of the sum of Wiener increments as $\Delta t \downarrow 0$,

$$W_t^{\Delta t} \to W_t = \int_0^t \mathrm{d}W_s \leftarrow \sum_{j=1}^{\lfloor t/\Delta t \rfloor} \Delta W_j = W_t^{\Delta t} \,. \tag{3}$$

A general one-dimensional diffusion process can be motivated by considering a generalization of the random walk Markov chain which allows for a non-zero initial state, a non-zero, time-inhomogenous, change in mean, and a non-unit, time-inhomogenous variance; namely, a Markov chain X_t, such that $X_0 = x_0$, and, for any $k \in \mathbb{N}$,

$$X_{k+1} = X_k + \Delta t \mu(X_k, t_k) + \zeta(X_k, t_k)\Delta W_{k+1} \,, \tag{4}$$

where the change in mean, $\mu : \mathbb{R} \times [0, \infty) \to \mathbb{R}$, and the variance, $\zeta^2 : \mathbb{R} \times [0, \infty) \to \mathbb{R}$, are, respectively, known as the drift and volatility of the process. Extrapolating (4) gives

$$X_{k+1} = x_0 + \Delta t \sum_{j=0}^{k} \mu(X_j, t_j) + \sum_{j=0}^{k} \zeta(X_j, t_j)\Delta W_{j+1} \,.$$

Again, taking an appropriate limit as $\Delta t \downarrow 0$,

$$X_t = x_0 + \int_0^t \mu(X_s, s)\, \mathrm{d}s + \int_0^t \zeta(X_s, s)\, \mathrm{d}W_s \,. \tag{5}$$

This is often written in shorthand as a stochastic differential equation (SDE);

$$\mathrm{d}X_t = \mu(X_t, t)\, \mathrm{d}t + \zeta(X_t, t)\, \mathrm{d}W_t \,. \tag{6}$$

A similar approach can be taken to motivate a general d-dimensional process, except in this case $\mu : \mathbb{R}^d \times [0, \infty) \to \mathbb{R}^d$, $\zeta : \mathbb{R}^d \times [0, \infty) \to \mathbb{R}^{d \times r}$,

and W_t is an r-dimensional Wiener process. The construction or, indeed, existence of a general diffusion process which satisfies Equation (5) is beyond the scope of this thesis[3]. In this thesis the continuous version of the Markov chain defined by (4), namely, the diffusion driven by (6), is only needed to motivate some of the algorithmic approaches taken when simulating conditioned diffusions as described in Chapter 3. As shall be highlighted in Chapter 3, inference conducted in this thesis will be for the Markov chain, defined by (4), for small Δt. Since diffusion processes are, themselves, approximations to real-world phenomena, this approach is valid, provided a suitably small increment, Δt, is chosen so that the chain exhibits similar behaviour to the real-world process being modelled. Of course, in practice, when conducting any data analysis, such verifications should always take place.

2.2 BAYESIAN INFERENCE

This section will briefly described the paradigm of Bayesian inference. For a more detailed introduction see, for example, Gelman et al., 2003. Given a set of data, assumed to be a sample from some pre-specified stochastic model, Bayesian inference seeks to infer the distribution, or properties thereof, of the underlying random variables (or parameters) driving the behaviour of the model. By specifying a prior distribution on the parameters of the model, Bayes' Theorem can be used to derive the conditional distribution of the parameters given the set of data. The construction of this prior distribution should be based upon sensible considerations of the dataset likely to be seen, without reference to any particular dataset that may have been collected; that is, the prior should be decided upon a-priori to seeing the dataset that will be used for analysis.

Formally, suppose there is a stochastic model $\mathbb{P}(X \in A|\Theta = \theta)$ with density $f(x|\theta)$ driving the generation of some observed dataset $X = x$ given a set of parameters $\Theta = \theta$. Suppose, further, that the parameters have a prior density $f_0(\theta)$. By Bayes' Theorem (see, for instance, Section 7.3, Papoulis, Pillai, and Pillai, 2002), the density of Θ given $X = x$; that is the posterior, is given by

$$g(\theta|x) = \frac{f(x|\theta)f_0(\theta)}{\int\limits_{\mathcal{X}} f(x|\theta)f_0(\theta)\,\mathrm{d}\theta}\,, \tag{7}$$

which encapsulates all the properties of the parameters Θ given the dataset $X = x$. Often, in practice, and with a sensible, *honest* f_0, the denominator of Equation (7), which is independent of θ, is intractable and, thus, the posterior is known only up to a *constant of proportionality*. Deriving properties from this posterior, or, in general, densities only known up to a constant of proportionality will be the subject of the remainder of this thesis.

3 For a rigorous construction, see, for example, Ethier and Kurtz, 1986; Stroock and Varadhan, 1997; Rogers and Williams, 2000a.

2.3 MONTE CARLO ALGORITHMS

Of primary interest in Bayesian inference problems and, therefore, this thesis, is the expectation of π-integrable functions h;

$$\pi[h] := \mathbb{E}_\pi[h(X)] = \int_{\mathcal{X}} h(x)\pi(x)\,\mathrm{d}x\,. \tag{8}$$

The d-dimensional random variable $X \sim \pi$ is encapsulated by such expectations; indeed π itself could be reconstructed from such expectations, since, for any measurable set A,

$$\pi(A) = \int_A \pi(x)\,\mathrm{d}x = \int_{\mathcal{X}} \mathbb{1}_A(x)\pi(x)\,\mathrm{d}x = \pi[\mathbb{1}_A]\,. \tag{9}$$

Loosely speaking, for moderately large d, $d \geq 5$ say, deterministic, grid-based methods for numerically approximating (8) have drawbacks which are undesirable for conducting Bayesian inference in many real-world applications. Firstly, the approximation error for a fixed computational cost for naive grid-based methods (see, for instance, Chapters 7 and 10 of Süli and Mayers, 2003, Sobolev and Vaskevich, 2013, or Hinrichs et al., 2014) typically scales exponentially poorly with dimension, and, secondly, adaptive grid-based approaches (see, for example, Dooren and Ridder, 1976, Berntsen, Espelid, and Genz, 1991, or Genz, 1991) rely heavily on the function h and, therefore, are computationally burdensome if $\pi[h]$ needs to be calculated for different functions h. The key problem with grid-based algorithms is that they consider the integral in (8) in the Riemann sense and, therefore, suffer from the exponentially increasing number of terms in the approximating Riemann sum. An alternative way of looking at $\pi[h]$, which, for suitably well-behaved functions h, partially [4] alleviates the curse of dimensionality, is as the integral of the function h with respect to the probability measure π,

$$\pi[h] = \int_{\mathcal{X}} h(x)\pi(\mathrm{d}x)\,.$$

Therefore, the contributions to $\pi[h]$ which are the most important are those areas of the space which π has relatively large mass. This suggests that an alternative approach to approximating $\pi[h]$ is to consider sets with fixed measure π, or, as shown in Section 2.3.2, for importance sampling, with fixed measure q for a suitably chosen q. This consideration leads to the idea of stochastic approximations formed by *averaging*, formally Monte Carlo (MC), techniques, that attempt to sample from

4 The *curse of dimensionality* is not entirely avoided. Indeed, in general, the volume of the regions of equal π measure increase as the dimension increases and, therefore, if the function h is not sufficiently smooth, the approximation of the integral over those regions deteriorates. It is clear, then, that the error of such an approximation has a subtle dependence on the dimension of the space through the smoothness of h.

the distribution π, and, thus, concentrate on representing h accurately in the regions where π has the most mass.

This section will introduce the core ideas of the Monte Carlo approach underpinning the algorithms of this thesis. In particular, in Section 2.3.1, the idealized Monte Carlo estimate which uses independent samples from π to construct a MC approximation to $\pi[h]$ will be introduced and its properties will be described. Further, two algorithms for simulating independent and identically distributed samples from $\pi[h]$, namely the inverse transform and rejection sampling, will be discussed. Section 2.3.2 will then describe importance sampling, which, similar to rejection sampling, relies on being able to sample from a suitably chosen proposal distribution, q. However, unlike rejection sampling, importance sampling uses all the simulated proposals when constructing the MC approximation. In Section 2.3.3, the normalised importance sampler, which is a MC algorithm which avoids the need to be able to calculate π exactly and only relies upon being able to calculate π up to a constant of proportionality, will be introduced. Section 2.3.5 will introduce Markov Chain Monte Carlo (MCMC) algorithms and the machinery needed to analyse the properties of the resulting MCMC estimator (given by (18)). Section 2.3.6 will apply this machinery to propose and accept-reject MCMC algorithms, highlighting several *ergodicity* results detailed in the literature. These results will then be discussed in the context of the Importance Sampler (Section 2.3.6.1) and the Random-Walk Sampler (Section 2.3.6.2).

2.3.1 *Idealised Algorithm*

For any function $h : \mathbb{R}^d \to \mathbb{R}$ the idealised Monte Carlo approximation of $\pi[h]$ (see, for example, Section 1.1, Liu, 2001, Section 1.1.1, Doucet et al., 2001, or Section 1.3.1, Glasserman, 2010) relies on a sequence of independent samples, $X_{1:N}$, from π;

$$I_N^{\mathrm{IA}}(X_{1:N}; h) := \frac{1}{N} \sum_{i=1}^{N} h(X_i) \, . \tag{10}$$

Such an approximation has several desirable properties which Theorem 2.3.1 collects from several sources in the literature (see, for example, Williams, 1991; Liu, 2001; Doucet et al., 2001; Robert and Casella, 2004);

THEOREM 2.3.1. *Let $X_{1:\infty}$ be a sequence of independent d-dimensional random variables each having distribution π and let $h : \mathbb{R}^d \to \mathbb{R}$ be any function such that $\pi[|h|] < \infty$ (so that $\pi[h]$ exists). For any $N \in \mathbb{N}$ let $I_N^{\mathrm{IA}}(X_{1:N}; h)$ be defined by (10). Then,*

(U) The estimator is unbiased; that is, for any $N \in \mathbb{N}$,

$$\mathbb{E}[I_N^{\mathrm{IA}}(X_{1:N}; h)] = \pi[h] \, .$$

(SL) The estimator is strongly consistent; that is,

$$\operatorname*{aslim}_{N \to \infty} I_N^{\mathrm{IA}}(X_{1:N}; h) = \pi[h] \, .$$

Moreover, if $\sigma^2 := \mathrm{Var}_\pi[h(X)] < \infty$*, then the following properties also hold;*

(CR) The estimator concentrates at the rate $N^{-1/2}$*; that is, for any* $\epsilon > 0$*,*

$$\mathbb{P}(|I_N^{\mathrm{IA}}(X_{1:N}; h) - \pi[h]| > \epsilon N^{-1/2}) \leq \sigma^2/\epsilon^2 \, .$$

(CLT) The estimator obeys a Central Limit Theorem with limiting variance σ^2*; that is,*

$$\operatorname*{dlim}_{N\to\infty} \sqrt{N}[I_N^{\mathrm{IA}}(X_{1:N}; h) - \pi[h]] = N_1(0, \sigma^2) \, . \tag{11}$$

This theorem highlights the appeal of the Monte Carlo estimator (10), since, although the approximation is stochastic, it is unbiased (property U), and, therefore, on average does not over-estimate or under-estimate $\pi[h]$, it converges to $\pi[h]$ with probability 1 (by property SL), and, the rate of convergence, that is, the Monte Carlo error, can, not only be characterised probabilistically (via properties CR and CLT), but is also independent of the dimension d (although, see the footnote on the previous page for a discussion about the subtle fact that, in general, $\mathrm{Var}[h(X)]$ will increase as the dimension increases). One way to simulate samples from π is to use the inverse transform (see, for example, Theorem 2.1, Devroye, 1986, Lemma 2.1.1, Liu, 2001, and Section 2.2.1, Glasserman, 2010), which relies on being able to simulate independent samples from the uniform distribution on the interval $(0, 1)$[5] and being able to invert the cumulative distribution associated with π:

THEOREM 2.3.2. *Let* $U \sim \mathrm{Unif}(0, 1)$ *and suppose that* π *is a distribution function such that the corresponding cumulative distribution function,* $F(x) = \pi((-\infty, x])$*, is invertible for every* $x \in \mathbb{R}$*. Then,* $X := F^{-1}(U) \sim \pi$*. Suppose, instead that* $X \sim \pi$*. Then,* $F(X) \sim \mathrm{Unif}(0, 1)$*.*

Unfortunately, for more complicated distributions π, inverting the cumulative distribution function, either exactly through an explicit form, or approximately via an algorithm, is either too computationally intensive or impossible. One alternative solution is to use rejection sampling (see, for instance, Section 3, Chapter 2, Devroye, 1986, Section 2.2, Liu, 2001, or Section 2.2.2, Glasserman, 2010) to indirectly simulate samples from π by simulating samples from some other, appropriately chosen proposal density, q, and accepting them as samples from π with probability proportional to π/q. The drawback of this approach is highlighted in Theorem 4.2, Owen, 2013, and the discussion thereafter which shows that, on average, the computational cost of simulating N samples from π, and, therefore, obtaining a Monte Carlo error of order $N^{-1/2}$, using this procedure is NM, where M is the bound on the ratio π/q. Hence,

5 Simulating pseudo-random independent samples from a $\mathrm{U}(0, 1)$ distribution is a non-trivial task which has received a lot of attention in the literature (see, for instance, L'Ecuyer, 1994). This thesis will assume that there exists, at hand, a sequence $U_{1:\infty}$ of independent samples from a $\mathrm{U}(0, 1)$ distribution.

not only is it necessary to choose a proposal density, q, such that an M exists, but, for a computationally efficient algorithm, it is also necessary to choose one such that M is small, thus restricting the range of applicability of rejection sampling.

2.3.2 *Importance Sampling*

Importance sampling (see, for example, Section 2.5, Liu, 2001, Section 1.3.2, Doucet et al., 2001, or Section 4.6, Glasserman, 2010) is an algorithm which utilises a change of measure argument to construct a Monte Carlo approximation to $\pi[h]$ using all of the samples proposed from some *proposal* distribution q, which is in constrast to the rejection sampling procedure of the previous section. By weighting proposals from q by the Radon-Nikodym derivative, one can quantify how *likely* such a proposal is from π. The identity

$$\pi(A) = \int_A \frac{\pi(x)}{q(x)} q(x) \, \mathrm{d}x$$

motivates the following importance sampling estimator of $\pi[h]$,

$$I_N^{\mathrm{IS}}(X_{1:N}; h) := \frac{1}{N} \sum_{i=1}^{N} w(X_i) h(X_i) \,, \tag{12}$$

where $w(x) := \pi(x)/q(x)$. This estimator shares the same desirable properties as the *idealized* MC estimator, (10), but under different conditions. Like Theorem 2.3.1, Theorem 2.3.3 collects these properties from several sources in the literature (see, for example, Williams, 1991; Liu, 2001; Doucet et al., 2001; Robert and Casella, 2004);

THEOREM 2.3.3. *Let $X_{1:\infty}$ be a sequence of independent d-dimensional random variables each having distribution q, where the corresponding density, also denoted by q, is such that $\mathrm{supp}(\pi) \subseteq \mathrm{supp}(q)$, and let $h : \mathbb{R}^d \to \mathbb{R}$ be any function such that $\pi(|h|) < \infty$ (so that $\pi[h]$ exists). For any $N \in \mathbb{N}$ let $I_N^{\mathrm{IS}}(X_{1:N}; h)$ be defined by (12). Then,*

(U) The estimator is unbiased; that is, for any $N \in \mathbb{N}$,

$$\mathbb{E}[I_N^{\mathrm{IS}}(X_{1:N}; h)] = \pi[h] \,.$$

(SL) The estimator is strongly consistent; that is,

$$\operatorname*{aslim}_{N \to \infty} I_N^{\mathrm{IS}}(X_{1:N}; h) = \pi[h] \,.$$

Moreover, if $\sigma^2 := \mathrm{Var}_q[w(X)h(X)] < \infty$, then the following properties also hold;

(CR) The estimator concentrates at the rate $N^{-1/2}$; that is, for any $\epsilon > 0$,

$$\mathbb{P}(|I_N^{\mathrm{IS}}(X_{1:N}; h) - \pi[h]| > \epsilon N^{-1/2}) \leq \sigma^2/\epsilon^2 \,.$$

(CLT) The estimator obeys a Central Limit Theorem with limiting variance σ^2; that is,

$$\operatorname*{dlim}_{N\to\infty} \sqrt{N}[I_N^{\mathrm{IS}}(X_{1:N}; h) - \pi[h]] = N_1(0, \sigma^2) \; . \tag{13}$$

The Importance Sampling estimator depends crucially on the choice of the proposal distribution q. Choosing a *good* proposal is, in general, a non-trivial task, and is discussed throughout the literature (see, for instance, Oh and Berger, 1992; Owen and Zhou, 2000; Richard and Zhang, 2007). The following Theorem (see Kahn and Marshall, 1953) gives the form of the optimal proposal density, in the sense of minimizing the variance of the resulting Importance Sampling estimator;

THEOREM 2.3.4. *The proposal density q which minimises* $\mathrm{Var}_q(w_q(X)h(X))$, *where* $w_q(x) := \pi(x)/q(x)$ *is* $q(x) \propto \pi(x)|h(x)|$.

Typically the optimal choice can not be implemented in practice. However, it can help guide the construction of implementable proposals which lead to estimators with a relatively small variance. Ideally, the choice of q would depend on h and would be chosen such that $\mathrm{Var}_q(w(X)h(X)) < \infty$. Often, however, estimators for various functions h are wanted, and, in such cases, choosing different proposals for different functions h is, in many cases, prohibitively costly, and, when the functions h are unknown a priori, impossible by definition. In such scenarios it is generally a good idea to choose a q with heavier tails than p; that is, a q such that

$$\sup_{x\in\mathcal{X}} \pi(x)/q(x) = M < \infty \; ,$$

for then

$$\mathrm{Var}_q(w(X)h(X)) = \pi[w_q h^2] - \pi[h]^2 \leq M\pi[h^2] - \pi[h]^2 = \mathrm{Var}_\pi[h] + (M-1)\pi[h^2] \; ,$$

which is bounded provided $\pi[h^2] < \infty$.

2.3.3 *Normalized Importance Sampling*

For many situations of interest, particularly when conducting Bayesian inference for the posterior of some parameters, the target distribution $\pi(x)$ is only known up to a constant of proportionality; that is, $\pi(x) = \gamma(x)/\gamma(\mathcal{X})$, where $\gamma(\mathcal{X})$ is unknown. In such cases, the *importance weights*, $\pi(x)/q(x) = \gamma(\mathcal{X})^{-1}\gamma(x)/q(x)$, are, themselves, only known up to a constant of proportionality, and, the Importance Sampling estimator, given by (12), is only known up to a constant of proportionality;

$$I_N^{\mathrm{IS}}(X_{1:N}; h) := \frac{1}{N\gamma(\mathcal{X})} \sum_{i=1}^{N} w(X_i)h(X_i) \; ,$$

where, here, $w(x) := \gamma(x)/q(x)$. The normalized Importance Sampling estimator uses the samples from the proposal to also form a MC estimator to $\gamma(\mathcal{X})$, thereby allowing the construction of a tractable estimator to $\pi[h]$ of the form

$$I_N^{\mathrm{NIS}}(X_{1:N}; h) := \sum_{i=1}^{N} \tilde{w}_i(X_{1:N}) h(X_i) \ , \tag{14}$$

where $\tilde{w}_i(X_{1:N})$ is the i-th normalized weight; that is,

$$\tilde{w}_i(X_{1:N}) := w(X_i) \Big/ \sum_{j=1}^{N} w(X_j) \ .$$

Generally, this estimator is biased. However, under appropriate conditions, the estimator satisfies the same *consistency* condition as the Importance Sampling estimator does and converges at the same rate; that is $N^{-1/2}$ (see, for example, Geweke, 1989, or Section 9.2 in Owen, 2013). The normalised importance sampling estimator is a useful estimator when conducting Bayesian inference if a good proposal distribution q can be found. A particularly useful property of the estimator is that the samples X_i and their corresponding unnormalised weights $w(X_i) = \gamma(X_i)/q(X_i)$ can be generated in parallel.

2.3.4 *Effective Sample Size*

A natural question that arises is; how efficient is the normalised importance sampling estimator? Aside from theoretical interest, measures of efficiency are useful when employing adaptive resampling schemes in particle filters (see, for instance, Del Moral, Doucet, and Jasra, 2012, Doucet and Johansen, 2011, or Section 2.4.1.2). The ubiquitous measure of efficiency is the *effective sample size* (Kong, 1992);

DEFINITION 2.3.5. *Let $X_{1:N}$ be a sequence of independent samples from the target π, and let $I_N^{\mathrm{IA}}(X_{1:N}; h)$ be the idealised estimator defined by equation (10). Moreover, let $Y_{1:N}$ be a sequence of samples from some joint proposal distribution q and let $I_N(Y_{1:N}, h)$ be any estimator of the form*

$$I_N(Y_{1:N}, h) := \sum_{i=1}^{N} \tilde{w}_i(Y_{1:N}) h(Y_i) \ ,$$

where the $\tilde{w}_{1:N}(Y_{1:N})$ are normalised weights. The effective sample size of the estimator I_N is defined, for function h, as the number of independent samples needed for the idealised estimator to have the same variance as I_N;

$$\mathrm{ESS}(I_N, h) := \frac{\mathrm{Var}_\pi[h(X)]}{\mathrm{Var}_q[I_N(Y_{1:N}, h)]} \ . \tag{15}$$

If the estimator I_N has a small bias (N is large, for example) then the effective sample size is an intuitive measure of efficiency. Unfortunately, the measure depends on the target, π, the proposal, q, and the function h which makes it an infeasible measure to consider in practice. An approximation used throughout the literature (see, for example, Kong, Liu, and Wong, 1994, Doucet et al., 2001, and Liu, 2001) is given by

$$\mathrm{ESS}(\tilde{w}_{1:N}(Y_{1:N})) := \left(\sum_{i=1}^{N} \tilde{w}_i(Y_{1:N})^2 \right)^{-1}. \tag{16}$$

This approximation of the effective sample size only depends on the normalised weights, thereby making it a useful measure of efficiency to look at in practice. This thesis will use this approximation as the measure of effiency of a weighted sequence of random variables and will, henceforth, call this approximation the *effective sample size*.

2.3.5 *Markov Chain Monte Carlo (MCMC) Algorithms*

Markov Chain Monte Carlo (MCMC) algorithms (Smith and Roberts, 1993; Tierney, 1994; Roberts and Rosenthal, 1998a; Andrieu et al., 2003; Roberts and Rosenthal, 2004) are a versatile set of procedures which extend the basic Monte Carlo approach by constructing a Markov Chain with stationary distribution π, which, utilising the concept of *reversibility* (Definition 2.3.7), is surprisingly simple (see, for instance, Section 2, Tierney, 1994, Chapter 5, Liu, 2001, Section 2.1-2.3, Roberts and Rosenthal, 2004, or Section 2.3.6 below). Although, for certain Markov Chains, *coupling from the past* ideas (see, for example, Propp and Wilson, 1998; Foss and Tweedie, 1998; Fill, 1998) allow for the exact simulation of samples from the stationary distribution, such algorithms, in general, are either not applicable or not computationally efficient. Fortunately, provided a chain has *nice* properties; namely, that it is *irreducible* (Definition 2.3.8) and *aperiodic* (Definition 2.3.9), the limiting distribution of the chain exists and is the stationary distribution of the chain (see, for instance, Theorem 1 of Section 3, Tierney, 1994, Theorem 1, Rosenthal, 2001, Theorem 4 of Section 3, Roberts and Rosenthal, 2004, or Corollary 2.3.12 below), thus allowing for the construction of ergodic, with respect to the number of particles, averages by simulating the chain for a long period of time[6] (Theorem 17.1.7, Meyn and Tweedie, 2009). Moreover, if the chain also satisfies a *drift* condition (Definition 2.3.18) back to small sets (Definition 2.3.16), the convergence of the chain to its limiting distribution is *geometrically* quick (see, for example, Theorem 1.4, Mengersen and Tweedie, 1996, Section 3.4, Roberts and Rosenthal, 2004, or Theorem 15.0.1, Meyn and Tweedie, 2009), thereby ensuring that the resulting averages satisfy a Central Limit Theorem (see, for instance, Corollary 2.1, Roberts and Rosenthal, 1997, Theorem 1, Hobert et al., 2002, or Theorem 24, Roberts and Rosenthal, 2004).

6 In fact, aperiodicity is not necessary for the construction of ergodic averages (see the discussion follwing Theorem 2.3.13).

The idea behind Markov Chain Monte Carlo (MCMC) algorithms is to construct a Markov chain, X_t, with stationary distribution π in the hope that simulating the chain from some initial state, $X_0 = x_0$, and for a long enough time $T \in \mathbb{N}$, ensures that X_T has approximately the same distribution as the stationary distribution. Recall that a Markov chain, X_t, is a Markov process $\{X_t : t \in \mathbb{N}\}$ whose stochastic dynamics are fully determined by the initial states x_0 and the distributions of the transitions; $\mathbb{P}(X_{i+1} \in A | X_i = x)$. The chains of interest will be time homogeneous so that, for any $(i, j, x) \in \mathbb{N}^2 \times \mathbb{R}^d$, and any measurable $A \subseteq \mathbb{R}^d$,

$$\mathbb{P}(X_{j+1} \in A | X_j = x) = \mathbb{P}(X_{i+1} \in A | X_i = x) \,,$$

and, therefore, the Markov chain is completely determined by its initial state $X_0 = x_0$ and the transition distributions, defined, for any $x \in \mathbb{R}^d$ and any measurable $A \subseteq \mathbb{R}^d$, by

$$P(x, A) := \mathbb{P}(X_1 \in A | X_0 = x) \,. \tag{17}$$

The n-step transition distributions will be denoted by P^n;

$$\begin{aligned} P^n(x, A) &:= \mathbb{P}(X_n \in A | X_0 = x) \\ &= \int_A P^n(x, \mathrm{d}x_n) = \int_{\mathcal{X}^{n-1} \times A} P(x, \mathrm{d}x_1) \prod_{i=1}^{n-1} P(x_i, \mathrm{d}x_{i+1}) \,. \end{aligned}$$

Intuitively, a stationary distribution, π, of a Markov Chain is one such that, if the state of the chain at some time t has distribution π, then the state of the chain at time $t + 1$, and, therefore, at any time $s \geq t$, has distribution π. Formally,

DEFINITION 2.3.6 (Stationary Distribution). *A distribution π is called a stationary distribution of a Markov chain with transition distributions $P(x, \cdot)$, if, for any measurable $A \subseteq \mathbb{R}^d$,*

$$\pi(A) = \int_{\mathbb{R}^d} \pi(\mathrm{d}x) P(x, A) \,.$$

Once a Markov Chain with stationary distribution π has been constructed, estimates to expectations of the form (8) can be formed using the generic procedure given by Algorithm 3.

Unlike the Monte Carlo estimate given by (10), the MCMC estimate given by (18) is not unbiased. The period $\{0, \ldots, s\}$ is known as the *burn-in* period and is chosen by the practitioner as the point at which it *appears* as though the chain has reached equilibrium. Chosen wisely, the bias in the estimate (18) should be small[7]. Choosing an appropriate s is an extremely important part of implementing an MCMC algorithm in practice, and there are numerous heuristics that can be followed

7 It is possible to construct MCMC algorithms whose resulting estimates are unbiased (see, for example, Section 3, Agapiou, Roberts, and Vollmer, 2018, and the more generic formulation in Jacob, O'Leary, and Atchadé, 2020).

Algorithm 3 Generic MCMC procedure .

1: Construct a Markov chain with transition densities $P(x, y)$ and stationary distribution π.
2: Initialise the chain at x_0 and choose the number of iterations $T > 0$.
3: **for** $t = 0, \ldots, T-1$ **do**
4: Simulate a realisation x_{t+1} with from $P(x_t, \cdot)$.
5: **end for**
6: Choose an $s \in \{0, \ldots, T-1\}$ and form the Monte Carlo estimate of (8) as

$$I(x_{s:T}) := \frac{1}{(T-s)} \sum_{t=s+1}^{T} h(x_t) \, . \tag{18}$$

to guide this choice (see, for example, Chapter 6, Brooks et al., 2011). However, it will not be discussed any further in this thesis. The first step in implementing an MCMC algorithm is to construct a Markov chain which has stationary distribution π (known as the target of the chain). A sufficient condition for π to be a stationary distribution of a Markov chain is that the Markov chain is reversible with respect to π (see, for example, Proposition 1, Roberts and Rosenthal, 2004) which means that initialising the chain at a random sample from π and running the chain forwards is probabilistically equivalent to ending the chain at a random sample from π and running the Markov chain backwards.

DEFINITION 2.3.7 (Reversible Chain). *Let $\pi : \mathbb{R}^d \to \mathbb{R}$ be a distribution. A Markov chain with transition distributions $P(x, \cdot)$ is reversible with respect to π if, the function $Q : \mathcal{B}(\mathbb{R}^d) \times \mathcal{B}(\mathbb{R}^d) \to \mathbb{R}$ defined, for any measurable sets $A \subseteq \mathbb{R}^d$, $B \subseteq \mathbb{R}^d$, by*

$$Q(A, B) := \int_A \pi(\mathrm{d}x) P(x, B)$$

is symmetric.

As shall be shown in Section 2.3.6, constructing Markov chains with a specified stationary distribution using this lemma is simple. However, generally, this does not guarantee that the behaviour of the chain when run for a long period of time, and, in particular, the behaviour of the estimate (18) for large T, is *good*. Specifically, it does not guarantee two properties which are of particular interest when constructing estimators to expectations; *convergence* as T tends towards infinity (i.e. a Law of Large Numbers result), and *concentration* as T tends towards infinity (i.e. a Central Limit Theorem). To determine such results it is necessary to discuss certain properties of a Markov chain. First, note that if the Markov chain has a limiting density π, then π is also a stationary density for the Markov Chain (see, for instance, (10.5), Meyn and Tweedie, 2009). Moreover, if it is assumed that the chain is ϕ-irreducible (Definition 2.3.8) and aperiodic (Definition 2.3.9), then; the chain has a unique stationary density, the limiting density of the chain exists and is unique, and, therefore, the two are equivalent (Corollary 2.3.12). Furthermore, a Law of Large Numbers result for the MCMC

estimate (18) holds (Theorem 2.3.13). A chain is irreducible if, regardless of where the chain is started, the chain can reach any set of positive measure if the chain is run for long enough;

DEFINITION 2.3.8 (Irreducible Markov Chain). *Let X_t be a Markov chain with transition distributions $P(x, \cdot)$. The chain is said to be irreducible if there exists a σ-finite measure[8], ϕ, such that, for any $x \in \mathbb{R}^d$ and any measurable set $A \in \mathcal{B}(\mathbb{R}^d)$ with $\phi(A) > 0$, there exists a number of steps $t_{x,A} \in \mathbb{N}$ such that*

$$P^{t_{x,A}}(x, A) = \mathbb{P}(X_{t_{x,A}} \in A | X_0 = x) > 0 \,.$$

A chain is aperiodic if the chain does not cycle through disjoint sets;

DEFINITION 2.3.9 (Aperiodic Markov Chain). *Let X_t be a Markov chain with transition distributions $P(x, A)$. The chain is said to be periodic if there exists an $r \geq 2$ and disjoint, measurable sets $\mathcal{X}_0, \ldots, \mathcal{X}_{r-1} \in \mathcal{B}(\mathbb{R}^d)$ such that*

$$P(x, \mathcal{X}_{(i+1) \bmod r}) = 1$$

for all $x \in \mathcal{X}_i$ and any $i = 0, \ldots, r-1$. If no such r exists then the chain is said to be aperiodic.

The importance of these two conditions in ensuring the chain has a unique limiting density can be seen by considering a chain whose state space can be decomposed into two non-empty disjoint sets A and B. Firstly, suppose the two sets are such that if $x_0 \in A$, then $\mathbb{P}(X_t \in A) = 0$ for all t, and, if $x_0 \in B$, then $\mathbb{P}(X_t \in B) = 0$ for all t. This reducible Markov chain is essentially the amalgamation of two Markov chains, one with state space A, and the other with state space B, and the behaviour of the Markov chain will be different depending on whether the chain starts- and therefore always remains- in set A or set B. Secondly, suppose the two sets are such that, for any even t, $P(X_t \in A) = 1$, and for any odd t, $\mathbb{P}(X_t \in B) = 1$ (and, therefore, $\mathbb{P}(X_t \in A) = 0$). This periodic Markov chain oscillates between two sets and, therefore, cannot have a limiting density. The chains considered in this thesis will generally have the stronger property of being irreducible on one-step transitions. It is a simple result, which stems from the discussion following the definition of irreducibility on Page 32, Roberts and Rosenthal, 2004, that this property implies both irreducibility and aperiodicity;

LEMMA 2.3.10. *Let X_t be a Markov chain with transition distributions $P(x, \cdot)$. Suppose the chain is one-step irreducible; that is, there exists a σ-finite measure ϕ, such that, for any $x \in \mathbb{R}^d$ and any measurable set $A \in \mathcal{B}(\mathbb{R}^d)$ with $\phi(A) > 0$, $P(x, A) > 0$. Then X_t is irreducible and aperiodic.*

8 A measure μ on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ is σ-finite if $\mathbb{R}^d$ is the countable union of sets in $\mathcal{B}(\mathbb{R}^d)$ each of which has finite measure; that is, $\mu(I) < \infty$.

The following theorem (see, for example, Nummelin, 1984, Tierney, 1994, Meyn and Tweedie, 2009, and Theorem 4, Roberts and Rosenthal, 2004) and corollary (see, for instance, Corollary 1, Tierney, 1994), which extends the almost everywhere condition to the whole space, illustrate that the chain being irreducible and aperiodic implies equivalence between the stationary and limiting distribution of the chain.

THEOREM 2.3.11. *Suppose X_t is an irreducible and aperiodic Markov chain with stationary distribution π. Then, for any $A \in \mathcal{B}(\mathbb{R}^d)$,*

$$\lim_{t\uparrow\infty} \|P^t(x,\cdot) - \pi\| = 0 \,, \quad \pi\text{-a.e.} \,. \tag{19}$$

COROLLARY 2.3.12. *Suppose X_t is an irreducible and aperiodic Markov chain with stationary distribution π and transition distributions $P(x,\cdot)$. Suppose further that, for any $x \in \mathbb{R}^d$, $P(x,\cdot) \ll \pi$. Then, for any $A \in \mathcal{B}(\mathbb{R}^d)$,*

$$\lim_{t\uparrow\infty} \|P^t(x,\cdot) - \pi\| = 0 \,, \quad \textit{for all } x \in \mathbb{R}^d.$$

Importantly, under these ergodicity assumptions, a Strong Law of Large Numbers result holds for the MCMC estimate (18) (see, for example, Theorem 17.0.1, Meyn and Tweedie, 2009);

THEOREM 2.3.13. *Suppose X_t is an irreducible and aperiodic Markov chain with stationary density π. Then, for any fixed $s \in \mathbb{N}$ and any π-integrable h,*

$$\operatorname*{aslim}_{T\uparrow\infty} \frac{1}{(T-s)} \sum_{t=s+1}^{T} h(X_t) = \pi[h] \,. \tag{20}$$

As shown in Corollary 6, Roberts and Rosenthal, 2004, aperiodicity is not needed to demonstrate Theorem 2.3.13. Intuitively, this follows since any irreducible and periodic chain can be decomposed into several *disjoint* sub-chains each of which satisfies Theorem 2.3.13, and the MCMC estimate of the original chain can be decomposed into a weighted sum of MCMC estimates on the sub-chains.

While such ergodicity results give an idea of *when* a Markov chain converges and demonstrate that a Strong Law of Large Numbers holds, they give no indication of how *fast* such convergence occurs and, in particular, when the MCMC estimates defined by Equation 18 satisfy a Central Limit Theorem. Two rates of convergence considered in this thesis are *uniform* and *geometric* ergodicity. The names of these forms of ergodicity are unfortunate since in both cases the rate of convergence is geometric and uniform across the state space. The difference, then, is that the constant multiplier is uniform across the state space for uniform ergodocity and dependent on where the chain is initialised for geometric ergodicity. The intuition is clear; a chain is uniformly ergodic if, wherever the chain is initialised, the difference between the n-step transition distributions and the limiting distribution can be bounded by the same quantity (that is, uniformly), and this bound decays geometrically with a rate that is independent of where the chain is initialised.

On the other hand, a chain is geometrically ergodic if the difference can be bounded by a quantity which depends on where the chain was initialised, and this bound decays geometrically with a rate that is independent of where the chain is initialised.

DEFINITION 2.3.14 (Uniform Ergodicity). *Let X_t be a Markov chain with transition distributions $P(x, \cdot)$ and stationary distribution π. The Markov chain is said to be uniformly ergodic if there exists a $\rho < 1$ and an $M < \infty$ such that, for any $t \in \mathbb{N}$ and any $x \in \mathbb{R}^d$,*

$$\|P^t(x, \cdot) - \pi\| \leq M\rho^t \ .$$

DEFINITION 2.3.15 (Geometric Ergodicity). *Let X_t be a Markov chain with transition distributions $P(x, \cdot)$ and stationary distribution π. The Markov chain is said to be geometrically ergodic if there exists a $\rho < 1$ and a function $m : \mathbb{R}^d \to [0, \infty)$, such that, for any $t \in \mathbb{N}$ and any $x \in \mathbb{R}^d$,*

$$\|P^t(x, \cdot) - \pi\| \leq m(x)\rho^t \ .$$

For statements to be made about the two forms of ergodicity, the idea of small sets and geometric drift conditions will be introduced. For a given Markov Chain with transition distributions $P(x, \cdot)$, a set C is termed small if all transitions from within C have a component of size ϵ in common[9];

DEFINITION 2.3.16. *Let X_t be a Markov chain with transition distributions $P(x, \cdot)$. A set C is said to be ϵ-small for some $\epsilon > 0$ if there exists a probability measure v, with $v(C) > 0$, such that, for any $x \in C$, v satisfies the minorization condition; $P(x, A) \geq \epsilon v(A)$, for any measurable $A \subseteq \mathbb{R}^d$. The measure v is called the minorization measure.*

If a Markov chain X_t has stationary distribution π, then, for any ϵ-small set C and minorized probability measure v, π has a component of size ϵ in common with each of the transition distributions in $\{P(x, \cdot) : x \in C\}$ since, for any measurable $A \subseteq \mathcal{X}$,

$$\pi(A) = \int_{\mathbb{R}^d} \pi(\mathrm{d}x) P(x, A) \geq \epsilon v(A) \ .$$

Thus, intuitively, if the Markov chain starts from an $x \in C$, then the difference between π and $P(x, \cdot)$ will, at most, be of size $(1 - \epsilon)$. In other words; if the chain starts from an $x \in C$ then, with probability ϵ, the distribution of the chain after one step will, at least from a probabilistic viewpoint, be π. Naturally, if the whole state space, $\mathbb{R}^d$, is small, convergence of the n-step transitions to π will occur at a rate of $(1 - \epsilon)^n$. This intuition is formalised in the following theorem (see, for example, Theorem 8, Roberts and Rosenthal, 2004, Proposition 2, Tierney, 1994, and Theorem 16.0.2, Meyn and Tweedie, 2009);

9 This definition of a small set is simpler than that typically used in the literature. Usually, the literature defines a small set with respect to general n-step transitions as opposed to the 1-step transition definition considered in this thesis.

THEOREM 2.3.17. *Let X_t be a Markov chain with stationary distribution π and transition distributions $P(x, \cdot)$. Suppose that the whole state space, $\mathbb{R}^d$, is ϵ-small. Then, for any $x \in \mathbb{R}^d$, and any $n \in \mathbb{N}$,*

$$\|P^n(x, \cdot) - \pi\| \leq (1 - \epsilon)^n \ .$$

That is, the chain is uniformly ergodic.

It is often the case for Markov chains used in practice that the whole state space is not small; one notable exception being an independence sampler with bounded weights (Theorem 2.3.31). As a result, uniform ergodicity is often too strong a condition to achieve. However, for many Markov chains used in practice, it is possible to construct, and easily define, sets C which are small. Since transitions from within C share a component of size ϵ with π, intuition suggests that the chain will converge *quickly enough* if the chain can return back to C *quickly enough*. This intuition is formalised via a *drift* condition:

Formally, a geometric drift condition.

DEFINITION 2.3.18. *Let X_t be a Markov chain with transition distributions $P(x, \cdot)$. The Markov chain is said to satisfy a drift condition if there exists a positive function f, which is finite for at least one $x \in \mathbb{R}^d$, and positive, finite constants, α, β and $\gamma < 1$, such that $C := \{x \in \mathbb{R}^d : f(x) \leq \alpha\}$ is ϵ-small for some $\epsilon > 0$, and*

$$\mathbb{E}_{P(x,\cdot)}(f(Y)) \leq f(x) + (\gamma - 1)(1 + f(x)) + \beta \mathbb{1}_C(x) \ . \tag{21}$$

As stated, this definition of a geometric *drift* is seemingly stronger than the usual definition of a geometric drift (see, for example, Meyn and Tweedie, 1994, Section 3.4, Roberts and Rosenthal, 2004, and Section 15.2.2, Meyn and Tweedie, 2009):

DEFINITION 2.3.19. *Let X_t be a Markov chain with transition distributions $P(x, \cdot)$. In the literature, the Markov chain is said to satisfy a drift condition if there exists a function $v : \mathbb{R}^d \to [1, \infty)$, which is finite for at least one $x \in \mathbb{R}^d$, an ϵ-small set C, and positive, finite constants, β and $\gamma < 1$, such that*

$$\mathbb{E}_{P(x,\cdot)}(v(Y)) \leq \gamma v(x) + \beta \mathbb{1}_C(x) \ . \tag{22}$$

However, as the following lemma shows, the two are equivalent and, therefore, can be referred to interchangeably as the *drift condition*:

LEMMA 2.3.20. *Let X_t be a Markov chain with transition distributions $P(x, \cdot)$. Then X_t satisfies a drift condition in the sense of Definition 2.3.18 if, and only if, X_t satisfies a drift condition in the sense of Definition 2.3.19.*

Proof. See A.2. □

Moreover, the former defintion is more intuitive. The former drift condition asserts that

$$\sup_{x \in C} \mathbb{E}_{P(x,\cdot)}(f(Y)) < \infty \ .$$

That is, on average, from within C, the chain moves to regions where f is bounded and, therefore, does not move *too far* away from the small set C. Moreover, for $x \notin C$, the condition asserts that

$$\mathbb{E}_{P(x,\cdot)}(f(Y)) \leq f(x) + (\gamma - 1)(1 + f(x)) < f(x) \; . \tag{23}$$

That is, on average, from outside of C, the chain drifts back towards regions where f is smaller than $f(x)$ and, therefore, back towards the small set C. Intuitively, from (23), how quickly the chain moves back to the small set depends on x and, so, any convergence bound will also depend on x. The following theorem shows that an irreducible, aperiodic Markov chain which satisfies a drift condition is geometrically ergodic (see, for instance, Theorem 9, Roberts and Rosenthal, 2004, Theorem 15.0.1, Meyn and Tweedie, 2009);

THEOREM 2.3.21. *Let X_t be an irreducible, aperiodic Markov chain with stationary distribution π, and suppose that X_t satisfies a drift condition. Then X_t is geometrically ergodic.*

Importantly, from a practical perspective, Corollary 2.1 of Roberts and Rosenthal, 1997 demonstrates that, for *reversible* chains, geometric ergodicity implies that an MCMC estimate constructed from the chain started from the stationary distribution satisfies a Central Limit Theorem for functions which are square-integrable with respect to π;

THEOREM 2.3.22. *Let X_t be an irreducible and aperiodic Markov chain with stationary distribution π, where it is assumed that $X_0 \sim \pi$. Moreover, assume that X_t is reversible with respect to π, and that the chain satisfies a drift condition. Then, for any function h with $\pi[h^2] < \infty$,*

$$\operatorname*{dlim}_{n\to\infty} \frac{1}{\sqrt{n}} \sum_{i=0}^{n} [h(X_i) - \pi[h]] = Y$$

where $Y \sim \mathrm{N}_1(0, \tau \mathrm{Var}_\pi(h(X)))$ and

$$\tau := 1 + 2 \sum_{i=1}^{\infty} \mathrm{Corr}(X_0, X_i)$$

is the integrated autocorrelation time.

This theorem shows that the limiting variance of the MCMC estimate depends on the Markov chain through the *integrated autocorrelation time*, τ, in the sense that the limiting variance is smaller if τ is smaller. As a result, chains with a smaller integrated autocorrelation time are preferred over chains with a larger integrated autocorrelation time. Heuristically, this could be used to compare Markov Chain Monte Carlo algorithms through finite sample approximations of the autocorrelations at various lags. However, see Section 2.2.2, Sherlock, Fearnhead, and Roberts, 2010 for a discussion about the drawbacks of such an approach.

Although drift conditions are extremely useful for demonstrating geometric ergodicty of certain Markov chains and, therefore, central limit theorems of the resulting MCMC estimates, there are two practical issues in using the drift condition to investigate geometric ergodicity in some settings. Firstly, drift conditions do not directly allow the comparison of two Markov chains; that is, it is not immediately obvious how to use a drift condition of one chain to derive a drift condition of another, similar chain. Secondly, due to the freedom in being able to choose the small set C and the so-called *Lyapunov* function v, it is extremely difficult to show a chain does not satisfy a drift condition and, ultimately, to demonstrate that the MCMC estimate constructed from a chain does not satisfy a central limit theorem. Fortunately, Theorem 2.3.25 below, which is an amalgamation of several different results from the literature, demonstrates that, for *non-negative* chains (see Definition 2.3.23), geometric ergodicity of a chain is equivalent to the chain having a non-zero *conductance* (see Definition 2.3.24), and both are equivalent to the MCMC estimates satisfying a central limit theorem. Before stating the theorem, the ideas of *non-negative* Markov chains and *conductance* will be introduced;

DEFINITION 2.3.23. *A Markov chain with transition distributions $P(x, \cdot)$ and stationary distribution π is said to be non-negative if, for any function $f : \mathcal{X} \to \mathbb{R}$ which is square-integrable with respect to π,*

$$\langle Pf, f \rangle_\pi := \int_{\mathcal{X}^2} \pi(\mathrm{d}x) P(x, \mathrm{d}y) f(x) f(y) \geq 0 \, .$$

DEFINITION 2.3.24. *Let X_t be a Markov chain with stationary distribution π and transition distributions $P(x, \cdot)$. The conductance of any measurable set $A \subseteq \mathcal{X}$ with $0 < \pi(A) < 1$ is the quantity*

$$\kappa(A) := \frac{1}{\pi(A)\pi(A^c)} \int_A \pi(\mathrm{d}x) P(x, A^c) \, ,$$

and the conductance of the chain is

$$\kappa := \inf_{A \in \Omega} \kappa(A) \, , \tag{24}$$

where $\Omega := \{A \subseteq \mathcal{X} : A \text{ is measurable}\}$.

Intuitively, if the conductance of a set A is small, then it is more difficult for the chain to be within A and move to outside of A than the stationary distribution suggests. Indeed, if there exists a set A such that $0 < \pi(A) < 1$ and $\kappa(A) = 0$, then, the sets A and A^c *partition* the space, and the chain, once within A, can never leave. Therefore, the chain is reducible. Moreover, if there exists a sequence of measurable sets A_i such that $0 < \pi(A_i) < 1$ for all $i \in \mathbb{N}$ and $\kappa(A_i)$ converges to 0 as $i \to \infty$ and, thus, the conductance of the chain is 0, then, the larger i is, the more difficult the chain finds it be within A_i and move outside of A_i relative to the stationary *probabilities* of A_i

and A_i^c. This intuition suggests that convergence of chains with small conductances is slower than chains with larger conductances. In fact, for reversible, non-negative Markov chains, a non-zero conductance is equivalent to geometric ergodicity, which, itself, is equivalent to the corresponding MCMC estimates satisfying a central limit theorem. The following theorem, which combines Theorems 5, 7, and 14 of Roberts and Rosenthal, 2008, with Theorem 2.5, Lawler and Sokal, 1988, proves this.

THEOREM 2.3.25. *Let X_t be a non-negative Markov chain with transition distributions $P(x, \cdot)$. Suppose that X_t is reversible with respect to π. Then, the following are equivalent;*

1. *The chain has a non-zero conductance; that is $\kappa > 0$ where κ is defined by Equation (24).*
2. *The chain is geometrically ergodic.*
3. *The MCMC estimate satisfies a Central Limit Theorem for functions which are square-integrable with respect to π; that is, suppose $X_0 \sim \pi$, then, for any function h with $\pi[h^2] < \infty$,*

$$\operatorname*{dlim}_{n\to\infty} \frac{1}{\sqrt{n}} \sum_{i=0}^{n} [h(X_i) - \pi[h]] = Y$$

where $Y \sim \mathrm{N}_1(0, \tau \mathrm{Var}_\pi(h(X)))$ and

$$\tau := 1 + 2 \sum_{i=1}^{\infty} \mathrm{Corr}(X_0, X_i)$$

is the integrated autocorrelation time.

Proof. See A.3. □

As in Sherlock, Fearnhead, and Roberts, 2010, this thesis will use the expected squared jump distance (ESJD) as a measure of efficiency of a Markov chain Monte Carlo algorithm;

DEFINITION 2.3.26. *Consider a MCMC algorithm with transition density $P(x, \cdot)$ targeting a d-dimensional distribution π. Let $X \sim \pi$ and $Y|X = x \sim P(x, \cdot)$. Then, the expected squared jump distance is defined to be $\mathbb{E}[\|Y - X\|^2]$.*

As shown by Sherlock, Fearnhead, and Roberts, 2010, maximising this measure is equivalent to minimizing a weighted sum of the lag-1 autocorrelations. In practice one will not be able to calculate the expected squared jump distance. One can, however, monitor the running mean squared jump distance of a chain as a proxy for the expected squared jump distance;

DEFINITION 2.3.27. *Consider a realisation, $x_1, x_2, \ldots, x_T$, of a Markov chain X_t. Define the mean squared jump distance of the realisation to be*

$$\frac{1}{(T-1)} \sum_{s=2}^{T} \|x_s - x_{s-1}\|^2 .$$

Given the latter is a proxy for the former we will, throughout this thesis, for consistency, refer to the mean squared jump distance as the expected squared jump distance as the context in which the phrase is used will be sufficient to determine whether we are referring to Definition 2.3.26 or Definition 2.3.27.

2.3.6 *Propose and Accept-Reject MCMC Algorithms*

Let $\pi(x) = \gamma(x)/\gamma(\mathbb{R}^d)$ be a density of interest. In practice, $\gamma(\mathbb{R}^d)$ is typically unknown and the aim is to construct Markov chains which *target* π; that is, Markov chains which have π as the stationary distribution. This thesis will concentrate on the ubiquitous propose and accept-reject Markov chains which, from a state x, propose a move to a state y from a proposal distribution $q(x, \cdot)$ with density $q(x, y)$ and accept this move with probability $\alpha(x, y)$. With a sensibly chosen acceptance probability α, that depends on π and q, such chains are reversible with respect to π (see, for example, Lemma 2.3.28). Moreover, with a sensibly chosen proposal, such chains are irreducible and aperiodic and, by Corollary 2.3.12, such chains converge to the stationary distribution wherever the chain is started. This section will introduce two commonly referenced acceptance probabilities used in the literature; namely Barker's acceptance and the Metropolis-Hastings (MH) acceptance and will show that such forms of acceptance lead to reversible chains with respect to π; see Lemma 2.3.28. Furthermore, two common proposals will be introduced; namely the independent and random walk proposals and the ergodic properties of such chains will be discussed.

The propose and accept-reject Markov chains have transition distributions $P(x, \cdot)$ of the form

$$P(x, A) = \int_A q(x, y)\alpha(x, y)\mathrm{d}y + \delta_x(A) \int_{\mathbb{R}^d} (1 - \alpha(x, y))\, \mathrm{d}y \,, \qquad (25)$$

where $q(x, y)$ is a *proposal* density, $\alpha(x, y)$ is an *acceptance* probability, and δ_x denotes the Dirac measure centred on x. A natural assumption on the proposal density, which is satisfied for many of the proposals considered in the literature, and for the proposals considered in this thesis, is that γ (and hence π) be absolutely continuous with respect to the proposal and, therefore, it is always possible to move, in one step, to any state where π is non-zero. The *weight* of a proposed move is the ratio of the density of the target at the proposed state over the density of the proposal from the current state to the proposed state; that is, $w(x, y) := \gamma(y)/q(x, y)$.

Barker's algorithm and the Metropolis-Hastings (MH) algorithm are propose and accept-reject MCMC algorithms with acceptance probabilities α_b and α_m respectively, where

$$\alpha_b(x,y) := \frac{w(x,y)}{w(y,x)+w(x,y)} , \tag{26}$$

$$\alpha_m(x,y) := 1 \wedge \frac{w(x,y)}{w(y,x)} . \tag{27}$$

Trivially, both acceptance probabilities, when weighted by the reverse transition, satisfy a certain symmetry; that is, both $w(y,x)\alpha_b(x,y)$ and $w(y,x)\alpha_m(x,y)$ are symmetric. Moreover, the Metropolis-Hastings acceptance probability dominates Barker's acceptance probability, but, only, at most, by one half;

$$\frac{1}{2}\alpha_m(x,y) \leq \alpha_b(x,y) \leq \alpha_m(x,y) . \tag{28}$$

Intuitively, therefore, the Metropolis-Hastings algorithm will *mix* more quickly than Barker's propose algorithm. This intuition can be made precise via a Peskun ordering argument (see Peskun, 1973, or, for instance, Tierney, 1998, and Mira, 2001). It is tempting, therefore, to concentrate on analysing the Metropolis-Hastings algorithm. However, Barker's acceptance probability has a number of properties which make it particularly useful in certain situations. Of particular importance to this thesis is the fact that the *multiple-proposal* extension of Barker's acceptance probability to acceptance *weights* is used throughout particle filtering schemes (see Section 2.4.1) and, therefore, throughout particle MCMC schemes. The symmetry property satisfied for both acceptance probabilities is sufficient to ensure that the Markov chains corresponding to Barker's algorithm or the Metropolis-Hastings algorithm are reversible with respect to the target distribution π and thus have π as a stationary distribution (see, for example, Tierney, 1994, or Proposition 2, Roberts and Rosenthal, 2004, for the Metropolis-Hastings specific result- the result for Barker's algorithm is a trivial extension):

LEMMA 2.3.28. *Let X_t be a propose-and-accept-reject Markov chain with either Barker's, or the Metropolis-Hastings', acceptance probability $\alpha(x,y)$. Then X_t is reversible with respect to π.*

To prove ergodicity, a continuity constraint, which typically holds for proposal densities used in practice, and holds for the proposal densities discussed in this thesis, is imposed.

ASSUMPTION 2.3.29 (Continuity of the Proposal). The proposal density $q(x,y)$ is continuous on $\mathbb{R}^d \times \mathbb{R}^d$.

This extra assumption is sufficient to demonstrate the Markov chain is one-step irreducible with respect to the target distribution π and, therefore, by Lemma 2.3.10, to demonstrate that the chain is irreducible and aperiodic. Hence, by Corollary 2.3.12, under such an assumption, the chain is ergodic; that is, the chain converges to the stationary distribution everywhere, and, by Theorem 2.3.13, the resulting MCMC

estimates satisfy a strong law of large numbers in the form of Equation (20). The following theorem formalised the one-step irreducibility claim— the assertion for the Metropolis-Hastings sampler can be seen, for instance on Page 31, Roberts and Rosenthal, 2004; the proof for Barker's random-walk sampler is a trivial extension that follows from Inequality (28);

THEOREM 2.3.30. *Let X_t be either Barker's algorithm or the Metropolis-Hastings algorithm with a proposal density $q(x, y)$ which satisfies Assumption 2.3.29. Then X_t is one-step irreducible as defined in Lemma 2.3.10.*

There are two commonly used proposals in the literature that will be discussed in this thesis; the *independent* proposal which leads to the *independence* sampler, and the *random-walk* proposal which leads to the *random-walk* sampler. The *independence* sampler proposes a new state y from a proposal $q(\cdot)$ independently of the current state x. The *random-walk* sampler, on the other hand, proposes a new state y *independently* of the *local* structure of the target π but in a way such that y is a random perturbation from x; that is, $Y|X = x \sim \mathrm{N}_d(x, \epsilon^2\mathcal{I}_d)$ for some pre-defined ϵ and, therefore,

$$q(x, y) = (2\pi\epsilon^2)^{-k/2} \exp[-(y - x)^T(y - x)/(2\epsilon^2)] \, .$$

Intuitively, if the structure of the target π is known reasonably well, and an implementable proposal q can be constructed which closely matches this structure, then the independence sampler should perform better than the *naive* random-walk sampler. However, for complex targets π where, either the structure is not known well, or, no feasibly implementable q which matches the structure of π well can be constructed, then the random-walk sampler should perform better. Indeed, if the target π is continuous then choosing ever smaller values of ϵ will lead to proposals which, although will not be far from the current state, will get accepted with probability close to one for the Metropolis-Hastings random-walk sampler, and with probability close to one-half for Barker's random-walk sampler. Balancing the trade-off between making large moves when moves are accepted and having a large, so-called, *acceptance rate* will be discussed in more detail in Section 2.3.6.3. In that section optimal-scaling results for the random-walk sampler will be highlighted.

2.3.6.1 *The Independence Sampler*

The independence sampler proposes a new state y *independently* of the current state x; that is, $q(x, y) = q(y)$. In situations where the structure of the target π is known reasonably well, and an implementable proposal q can be constructed which closely matches this structure, then the independence sampler will *mix* well. Intuitively, if one can choose a q which, *in the tails*, dominates the target π, then the weight of the current state $w(x) = \pi(x)/q(x)$ will never get so big that the chain finds it difficult to move. Therefore, on average, the chain will never

find itself moving to regions of the space with ever larger weights and, thus, to regions of the space which it finds ever difficult to leave. This intuition is formalised in Theorem 2.3.31 which demonstrates that, for both Barker's and the Metropolis-Hastings independence sampler, the MCMC estimates satisfy a central limit theorem if and only if q dominates π in the tails. This Theorem is essentially a collection of results found in the literature with a minor extension to cover Barker's independence sampler. Indeed, for the Metropolis-Hastings independence sampler one can see, for example, Liu, 1996, Corollary 4, Tierney, 1994, Theorem 1, Rosenthal, 1995, Theorem 1, Rosenthal, 2002, and Roberts and Rosenthal, 2011 for the if implication, and, for instance, Theorem 2.1, Mengersen and Tweedie, 1996, and Proposition 5.1, Roberts and Tweedie, 1996 for the result that the sampler is not geometrically ergodic if the *importance weights* are unbounded. The extension to Barker's independence sampler follows directly from Inequality 28;

THEOREM 2.3.31. *Let X_t be either Barker's or the Metropolis-Hastings independence sampler. Suppose, further that the proposal satisfies Assumption 2.3.29. Then, the chain is uniformly ergodic if*

$$\nu := \sup_{x \in \mathbb{R}^d} w(x) < \infty \,. \tag{29}$$

Furthermore, the rate of convergence is $(1-\gamma(\mathbb{R}^d)\nu^{-1}/2)^n$ for Barker's independence sampler and $(1-\gamma(\mathbb{R}^d)\nu^{-1})^n$ for the Metropolis Hastings independence sampler. Finally, the MCMC estimates corresponding to such chains satisfy central limit theorems for all functions which are square-integrable with respect to π if and only if ν is finite.

It is clear from this theorem when one independence sampler should be preferred over another. Indeed, as one would expect intuitively, an independence sampler with a smaller value of ν is preferable. Note that, for any proposal density q, $\nu \geq 1$, and this bound is achieved for $q \equiv \pi$. Hence, again, as one would expect intuitively, to maximise the rate of convergence for the independence sampler one should choose a proposal q which is easy to simulate samples from, and which matches π *closely*. Due to the independence between the proposal and the current state, the independence sampler can be efficiently extended to a *multiple-proposal* regime where, given a current state y_0, a sequence, $y_{1:N}$, of N proposals are simulated independently and identically from q. Then, a move to y_i, for any $i \in \{0, \ldots, N\}$, happens with probability $\alpha_{iN}(y_{0:N})$. The transition distribution for such a sampler is given by

$$\begin{aligned} P(x, A) = &\sum_{k=1}^{N} \int_{A} \cdots \int_{\mathbb{R}^{d\times(N-1)}} \prod_{i=1}^{N} q(y_i) \alpha_{kN}(x, y_{1:N}) \, \mathrm{d}y_k \mathrm{d}y_{-k} \\ &+ \delta_x(A) \int_{\mathbb{R}^{d\times N}} \cdots \int \prod_{i=1}^{N} q(y_i) \left(1 - \sum_{k=1}^{N} \alpha_{kN}(x, y_{1:N})\right) \mathrm{d}y_{1:N} \,, \end{aligned} \tag{30}$$

for any $x \in \mathbb{R}^d$ and any measurable $A \subseteq \mathbb{R}^d$. Such a sampler is a specific case of the more general multiple-proposal samplers introduced in Section 4.3.2.2, Lee, 2011. One can demonstate that the ergodic properties of the multiple-proposal independence sampler mirror the ergodic properties of the single-proposal independence sampler. However, in the interest of space, these results will not be stated or proved as part of this thesis.

2.3.6.2 *The Random-Walk Sampler*

The *random-walk* sampler proposes a new state, y, *independently* of the *local* structure of the target π but in a way such that y is a random perturbation from x; that is, $Y|X = x \sim \mathrm{N}_d(x, \epsilon^2\mathcal{I}_d)$ for some predefined $\epsilon > 0$ [10]. By definition,

$$q(x, y) = (2\pi\epsilon^2)^{-k/2} \exp[-(y - x)^T(y - x)/(2\epsilon^2)] \, .$$

In situations where, either the structure of the target π is not well known, or, no feasibly implementable q which matches the structure of π well can be constructed, then the random-walk sampler is preferable over the independence sampler. Note that, although the weight is defined as $w(x, y) = \gamma(y)/q(x, y)$, the proposal density q is symmetric; that is $q(x, y) = q(y, x)$ for any $(x, y) \in \mathcal{X}^2$, and, therefore, for both Barkers and the Metropolis-Hastings acceptance probability, it is the ratio

$$\frac{w(x, y)}{w(y, x)} = \frac{\gamma(y)q(y, x)}{\gamma(x)q(x, y)} = \frac{\gamma(y)}{\gamma(x)}$$

which is ultimately of interest. Hence, it is sufficient to consider the *pseudo-weight* $\tilde{w}(x) := \gamma(x)$. Intuitively, if π, and therefore γ, decays sufficiently quickly in the tails then the symmetry of the proposal about the current state of the chain will lead to a *drift* towards the parts of the space where π has most mass; that is, a drift towards *small* sets. This intuition is formally stated in Theorem 2.3.34, which consists of a slight extension of Theorem 3.2, Mengersen and Tweedie, 1996 to cover Barkers random-walk sampler and combines this with Theorem 2.3.25— which, itself, is a combination of several results in the literature; see the proof of that theorem for references— to make statements about central limit theorems. Indeed, this theorem demonstrates that the MCMC estimates corresponding to Barker's or the Metropolis-Hastings random-walk sampler satisfy a central limit theorem if π decays exponentially quickly in the tails. Firstly, the following lemma shows that random-walk chains are non-negative; the assertion about the Metropolis-Hastings random-walk sampler is demonstrated

10 This is a *naive* random-walk in the sense that it takes no account of the structure of the target. In general a random-walk can use a covariance matrix which better reflects the covariance matrix of the target. This tailored random-walk can have better mixing properties then the naive approach detailed in this thesis (see, for example, Theorem 6, Roberts and Rosenthal, 2001 and Sherlock, Fearnhead, and Roberts, 2010).

in Lemma 3.1, Baxendale, 2005, the result for Barker's random-walk sampler is a trivial extension which follows from Inequality (28):

LEMMA 2.3.32. *Let X_t be either Barker's, or, the Metropolis-Hastings random-walk sampler; that is a proposer-and-accept-reject Markov chain with proposal density*

$$q(x,y) = (2\pi\epsilon^2)^{-k/2} \exp[-(y-x)^T(y-x)/(2\epsilon^2)] \,,$$

for some $\epsilon > 0$. Then X_t is non-negative.

Secondly, the following defines what it means for π to decay exponentially quickly in the tails in one dimension;

DEFINITION 2.3.33. *In one dimension, $\pi : \mathbb{R} \to \mathbb{R}$ is said to decay exponentially quickly in the tails if there exists positive constants m_1, m_2, θ_1, and θ_2 such that, for any $(x,y) \in \mathbb{R}^2$,*

$$\begin{aligned} \pi(y)/\pi(x) &\leq \exp[-\theta_2(y-x)] \,, \quad \text{if } y \geq x \geq m_2, \\ \pi(y)/\pi(x) &\leq \exp[-\theta_1(x-y)] \,, \quad \text{if } y \leq x \leq -m_1. \end{aligned}$$

The following theorem on the existence of central limit theorems for the random-walk sampler in one-dimension, which combines Theorem 3.2, Mengersen and Tweedie, 1996, with Theorem 2.3.25 demonstrates that, in one dimension, exponential decay in the tails is sufficient to prove the existence of central limit theorems for both Barker's and the Metropolis-Hastings random-walk sampler;

THEOREM 2.3.34. *Let X_t be either Barker's, or, the Metropolis-Hastings random-walk sampler in one dimension; that is, a propose and accept-reject Markov chain with proposal density*

$$q(x,y) = \frac{1}{\sqrt{2\pi\epsilon^2}} \exp\left(-\frac{(y-x)^2}{2\epsilon^2} \right) ,$$

for some $\epsilon > 0$. Further, suppose that π is greater than zero for any $x \in \mathbb{R}$. Then, the MCMC estimates corresponding to such samplers satisfy a central limit theorem for all functions which are square-integrable with respect to π if π decays exponentially in the tails (in the sense of Definition 2.3.33).

Proof. See A.4. □

Under suitable extensions of the exponentially decaying tails condition to higher dimensions, the conclusions of Theorem 2.3.34 can be extended (see Theorem 2.1, Roberts and Tweedie, 1996). While Theorem 2.3.34 gives sufficient conditions on the target, under which, the random-walk sampler produces MCMC estimates which satisfy central limit theorems, it does not give guidance with regards to choosing a *good* step-size; that is, a step-size which results in a Markov chain with a *high* rate of *mixing*. To see the importance of the step-size on the mixing properties of the random-walk sampler consider the following example;

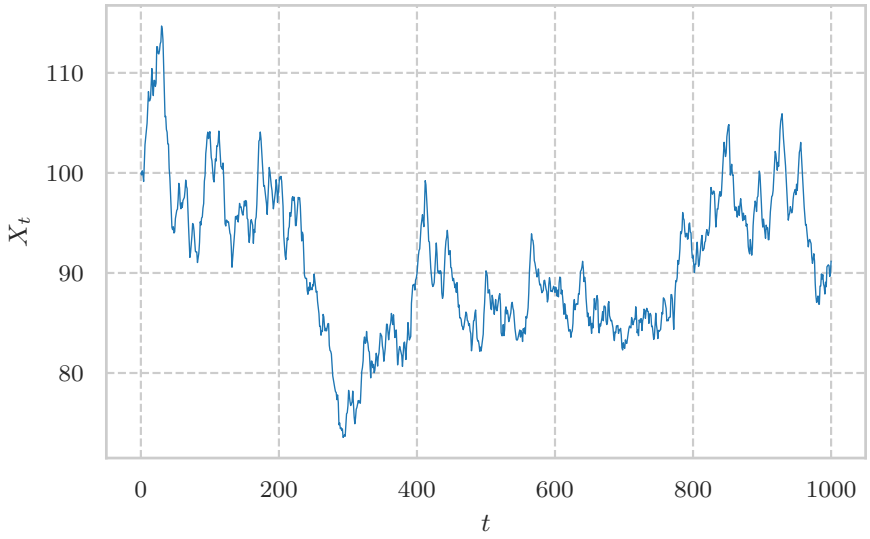

Figure 2: A sample path of the process X_t as described in Example 1 with $\alpha = 3/10$, $\sigma = 1/100$, $x_0 = 100$, $r_0 = 0.0$, and $T = 1000$.

EXAMPLE 1. *Let X_t be a growth process driven by a mean-reverting process R_t and a reversion rate α;*

$$R_{t+1}|R_t = r_t \sim \mathrm{N}(\alpha r_t, \sigma^2) \ ,$$
$$X_{t+1} = (1 + R_{t+1})X_t \ .$$

The process R_t is an autocorrelated process which reverts towards 0 provided $\alpha < 1$. The process X_t grows at the rate $R_t X_{t-1}$. A sample path of the process; that is, a simulated dataset, can be seen in Figure 2. Given the values of the process X_t, the values of the process R_t are trivially given by $X_t/X_{t-1} - 1$. Hence, for conducting Bayesian inference on (α, σ, r_0), it is sufficient to consider the posterior given $R_{1:T}$. Assuming a joint prior density f_0 for the unknown parameters, Equation (7) gives the posterior of the parameters given $R_{1:T} = r_{1:T}$;

$$g(\alpha, \sigma, r_0|r_{1:T}) \propto f_0(\alpha, \sigma, r_0)\sigma^{-T} \exp\left(-\frac{1}{2\sigma^2}\sum_{t=1}^{T}(r_t - \alpha r_{t-1})^2\right) \ , \quad (31)$$

where, as is conventional, the posterior is only given up to a constant of proportionality.

In the interest of simplicity suppose r_0 and σ are fixed and known. Moreover, suppose an improper, *flat*, prior for α; that is $f_0(\alpha) = 1$ for any $\alpha \in \mathbb{R}$ is imposed. By Equation (31), the posterior is given by

$$g(\alpha|r_{1:T}) \propto \exp\left(-\frac{1}{2\sigma^2}\sum_{t=1}^{T}(r_t - \alpha r_{t-1})^2\right) \propto \exp\left[-\frac{\bar{r}}{2\sigma^2}\left(\alpha - \frac{\bar{r}'}{\bar{r}}\right)^2\right] \ , \quad (32)$$

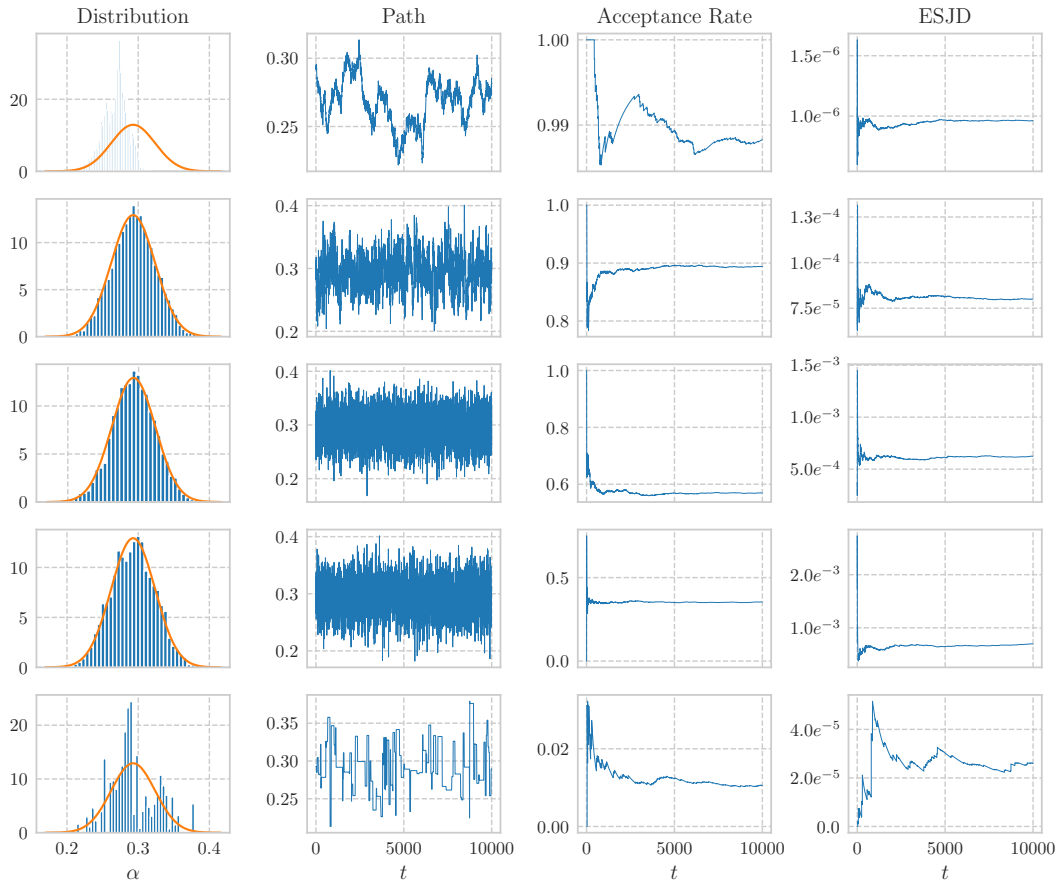

Figure 3: An illustration of the behaviour of the Metropolis Hastings random-walk sampler which targets (32) where $\sigma = 1/100$, $r_0 = 0.0$, and the realisations from the model, $r_{1:1000}$ are the same as those shown in Figure 2. The five rows, top to bottom, correspond to jump-sizes $\epsilon = 0.001$, $\epsilon = 0.01$, $\epsilon = 0.05$, $\epsilon = 0.1$, and $\epsilon = 4.0$ respectively, and the samplers were run for ten thousand iterations. The first column shows histograms of the simulated samples, with a plot of the true posterior probability density function super-imposed. The second column shows the evolution of the chain. The third and fourth columns show the evolution of the acceptance rate and expected squared jump distance respectively.

where

$$\bar{r} := \sum_{t=1}^{T} r_{t-1}^2 \ , \quad \bar{r}' := \sum_{t=1}^{T} r_{t-1} r_t \ .$$

Hence, the posterior for α is actually a one-dimensional normal distribution with mean $\bar{r}'/\bar{r}$ and variance $\sigma^2/\bar{r}$. Figures 3 and 4 show, respectively, the behaviour of the Metropolis-Hastings and Barker's random-walk sampler for five different choices of the jump-size ϵ.

It can be seen from the figures that, for both samplers, when the jump-size, ϵ, is very small— $\epsilon = 0.001$ for the top row in both figures— the acceptance rate is very close to the limiting acceptance rate, which is 1 for the Metropolis-Hastings sampler, and one-half for Barker's sampler (see the penultimate column in both figures), and the expected

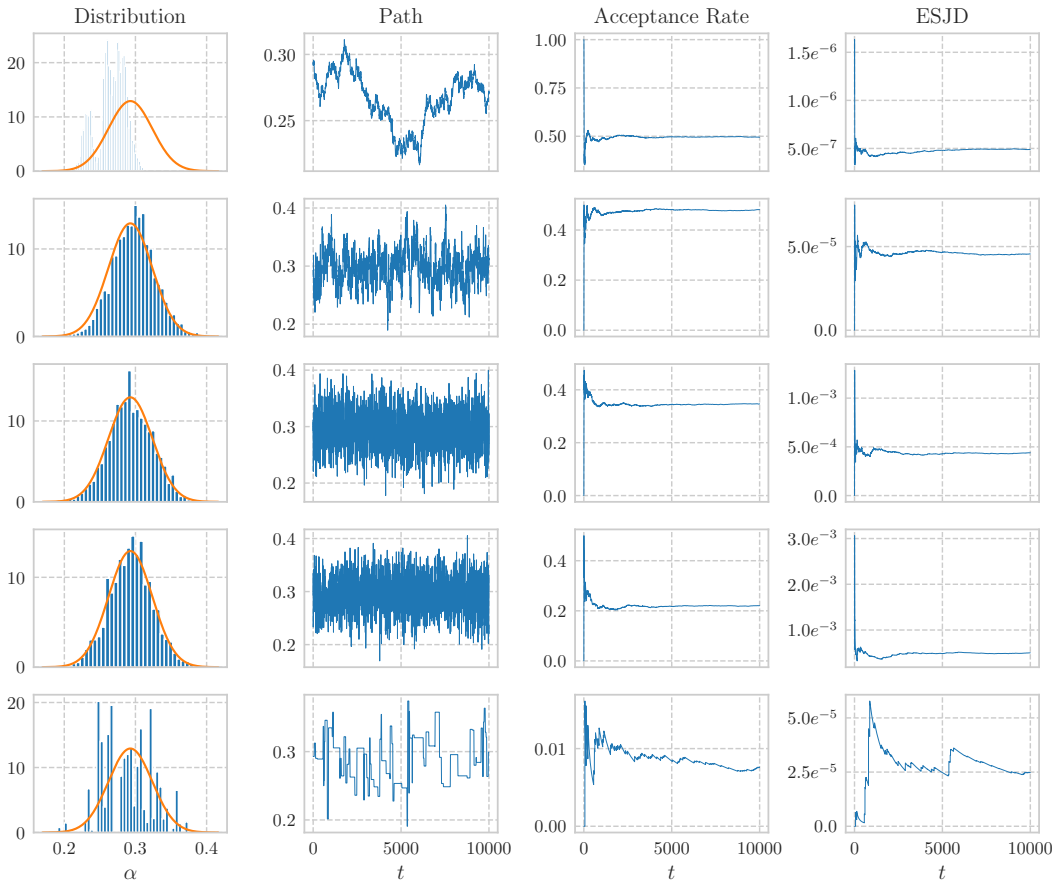

Figure 4: An illustration of the behaviour of Barker's random-walk sampler which targets (32) where $\sigma = 1/100$, $r_0 = 0.0$, and the realisations from the model, $r_{1:1000}$ are the same as those shown in Figure 2. The five rows, top to bottom, correspond to jump-sizes $\epsilon = 0.001$, $\epsilon = 0.01$, $\epsilon = 0.05$, $\epsilon = 0.1$, respectively, and the samplers were run for ten thousand iterations. The first column shows histograms of the simulated samples, with a plot of the true posterior probability density function super-imposed. The second column shows the evolution of the chain. The third and fourth columns show the evolution of the acceptance rate and expected squared jump distance respectively.

squared jump distance is very close to zero (the last column in both figures). Moreover, when the jump-size is large— $\epsilon = 4.0$ for the bottom row in both figures- the acceptance rate is close to zero and the expected squared jump distance is very close to zero. In both cases the chain does not mix well (the middle column in both figures), and the density of samples do not represent the true density particularly well (the first column in both figures). However, when ϵ is chosen to be of an appropriate size ($\epsilon \in \{0.01, 0.05, 0.1\}$ for the middle three rows), the acceptance rate is neither close to zero or one, the expected squared jump distance is relatively large, the chain mixes well, and the density of the samples represent the true density well. Furthermore, the expected squared jump distance converges to the limiting expected squared jump distance fairly quickly. This suggests that the expected squared jump distance is a useful measure to monitor when tuning the samplers.

2.3.6.3 *Optimal Scaling*

Given the observations at the end of the previous section it is natural to ask if there is an optimal choice of jump-size which *maximizes* the *rate of mixing*, thereby obtaining samples which optimally represent the true density. Moreover, from a practical viewpoint, it is natural to wonder if there a way to monitor the output from an MCMC algorithm in such a way that one can *tune* the jump-size. In general, such *optimal scaling* results are difficult to establish and depend heavily on the target π^*. However, by considering the simpler problem of a target made of d independent and identically distributed components which are *smooth*, it is possible to derive optimal scaling results in the limit as $d \to \infty$ which can then be used as general guidance for practitioners. Indeed, consider the case where the target π^* is of the form

$$\pi^*(x_{1:d}) = \prod_{i=1}^{d} \pi(x_i) \ ,$$

where each π is *smooth* (see the assumptions of Theorem 2.3.35 for a precise statement). Moreover, suppose that the chain starts in stationarity; that is, $X_0 \sim \pi^*$. Then, for the Metropolis-Hastings random-walk sampler, as $d \to \infty$, Roberts, Gelman, and Gilks, 1997 show that it is optimal to choose the jump-size ϵ to be approximately $2.38/(\psi\sqrt{d})$, where

$$\psi := \sqrt{\mathbb{E}_\pi[g'(X)^2]} \ ,$$

and $g(x) := \log \pi(x)$. Roberts, Gelman, and Gilks, 1997 also show that such a jump-size leads to an asymptotic expected acceptance rate of approximately 0.234:

THEOREM 2.3.35. *Let $\pi^* : \mathbb{R}^d \to (0, \infty)$ be a target of the form*

$$\pi^*(x_{1:d}) = \prod_{i=1}^{d} \pi(x_i) \ ,$$

where $\pi : \mathbb{R} \to (0, \infty)$ is twice continuously differentiable and the function $g : \mathbb{R} \to \mathbb{R}$, defined by $g(x) := \log \pi(x)$, satisfies

$$\mathbb{E}_\pi[g'(X)^8] < \infty \,, \quad \mathbb{E}_\pi[(g''(X) + g'(X)^2)^4] < \infty \,.$$

Furthermore, for $\lambda > 0$, let $X_0 \sim \pi$, and X_t be a Metropolis-Hastings random-walk sampler for π^* on $\mathbb{R}^d$ with jump-size $\epsilon_d := \lambda d^{-1/2}$. Finally, for any t, let Y_t^d be the first component of X_t speeded up by a factor of d; that is, $Y_t^d := X_{\lfloor dt \rfloor}$. Then, the process Y_t^d converges weakly to a diffusion process Y_t which satisfies

$$\mathrm{d}Y_t = \frac{1}{2}\bar{J}(\lambda) g'(Y_t)\, \mathrm{d}t + \sqrt{\bar{J}(\lambda)}\, \mathrm{d}W_t \,, \tag{33}$$

where

$$\bar{J}(\lambda) := 2\lambda^2 \Phi\left(-\frac{\lambda\psi}{2} \right) ,$$

Φ denotes the cumulative distribution function of a standard normal random variable, and

$$\psi := \sqrt{\mathbb{E}_\pi[g'(X)^2]} \,.$$

$\bar{J}(\lambda)$ corresponds to the speed of the diffusion process Y_t and is maximised when $\lambda = \hat{\lambda} = 2\hat{\beta}/\psi$, where $\hat{\beta}$ is the unique solution to

$$2\Phi(-\beta) = \beta\phi(\beta) \,.$$

Moreover, the limit, as d tends towards infinity, of the expected acceptance rate

$$\alpha^*(\epsilon_d) := \mathbb{E}\left(1 \wedge \frac{\pi(X + \epsilon_d Z)}{\pi(X)} \right) ,$$

where $Z \sim \mathrm{N}_d(0, 1)$ is a d-dimensional standard normal random variable, is

$$\bar{\alpha}(\lambda) := \lim_{d \uparrow \infty} \alpha^*(\epsilon_d) = 2\Phi\left(-\frac{\lambda\psi}{2} \right) . \tag{34}$$

This gives an optimal asymptotic expected acceptance rate of $\bar{\alpha}(\hat{\lambda}) = 2\Phi(-\hat{\beta})$ which is 0.234 to three decimal places.

The strength of the assumptions underpinning Theorem 2.3.35, along with the asymptotic nature of the result, suggest that such a result is only of theoretical interest. However, tuning the acceptance rate of Metropolis-Hastings random-walk samplers to 0.234 has been shown to be practically useful. Indeed, Roberts and Rosenthal, 2001 show that the *asymptotic* acceptance rate of 0.234 is approximately optimal even in dimensions as low as five. Moreover, Roberts and Rosenthal, 2001 show that the speed of the limiting diffusion (33), $\bar{J}(\lambda)$, when considered as a function of the asymptotic expected acceptance rate, $\bar{\alpha}(\lambda)$,

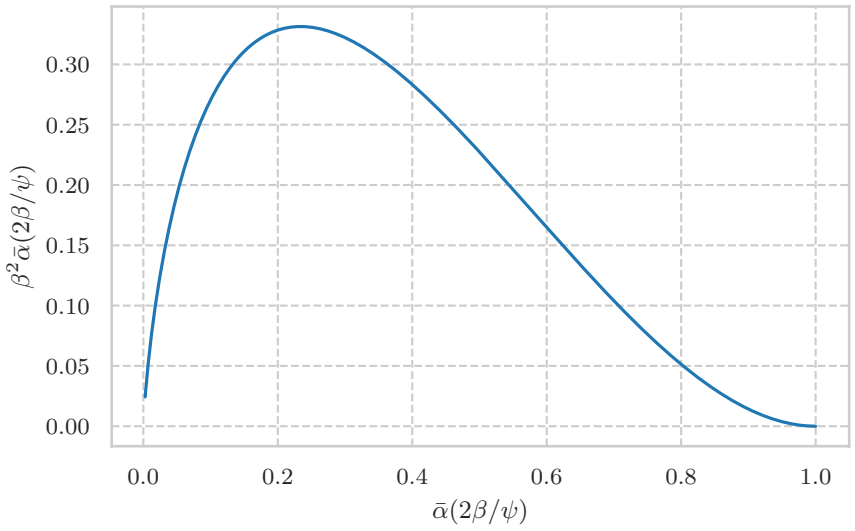

Figure 5: A plot of the asymptotic expected squared jump distance (up to a constant of proportionality) against the asymptotic acceptance rate for the Metropolis-Hastings random-walk sampler.

is relatively *flat* around the optimum, thus suggesting that only approximately tuning such samplers to 0.234 is required to achieve good mixing. This *insensitivity* to the tuning of the acceptance rate can be seen theoretically in Figure 5 and, in practice, in Figure 3. Indeed, from Figure 5, any asymptotic acceptance rate in $[0.05, 0.54]$ gives a speed of the diffusion that is above 60% of the maximum. As such, it is unnecessary to finely tune the *jump-size* to achieve the optimal acceptance rate, provided the *tuned* acceptance rate is on the same scale as the optimal asymptotic acceptance rate. This *insensitivity* is mirrored in Example 1, as is highlighted in Figure 3, where any acceptance rate in the interval $[0.25, 0.9]$ results in good mixing. Furthermore, the optimal asymptotic expected acceptance rate of 0.234 has been shown to hold under several relaxations of the independent and identically distributed assumption placed on the target (see, for instance, Roberts, 1998, Breyer and Roberts, 2000, and Roberts and Rosenthal, 2001).

Proving a diffusion limit for a MCMC algorithm, such as the one derived in Theorem 2.3.35, is, in general, a difficult task. An alternative approach, taken by Sherlock and Roberts, 2009, considers maximizing the asymptotic expected squared jump distance as the measure of *efficiency*. They demonstrate the following result for the Metropolis-Hastings random-walk sampler— the extension to Barker's random-walk sampler is trivial;

THEOREM 2.3.36. *Let $\pi^* : \mathbb{R}^d \to (0, \infty)$ be a target of the form*

$$\pi^*(x_{1:d}) = \prod_{i=1}^{d} \pi(x_i) \ ,$$

where $\pi : \mathbb{R} \to (0, \infty)$. *Furthermore, let* $X_0 \sim \pi$, *and* X_t *be either Barker's or the Metropolis-Hastings random-walk sampler for* π^* *with jump-size* $\epsilon_d := \lambda d^{-1/2}$. *Suppose that the following assumptions hold;*

(C) The marginal density, π, *is twice continuously differentiable.*

(F) The first and second derivatives of the logarithm of the marginal density, $g(x) := \log \pi(x)$, *satisfy the following;*

$$\mathbb{E}_\pi[g'(X)^2] < \infty \,, \quad \mathbb{E}_\pi[g''(X)] < \infty \,.$$

(L) The second derivative of the logarithm of the marginal density is Lipschitz continuous with Lipschitz constant a; *that is, the function* $g(x) := \log(\pi(x))$ *is such that, for any* $x_1, x_2 \in \mathbb{R}$,

$$|g''(x_2) - g''(x_1)| \leq a|x_2 - x_1| \,.$$

(VG) The gradient of the marginal density vanishes in the tails; that is,

$$\lim_{x \uparrow \infty} \pi'(x) = \lim_{x \downarrow -\infty} \pi'(x) = 0 \,.$$

Then, the expected acceptance rate,

$$\alpha^*(\epsilon) := \mathbb{E}[\alpha(X, X + \epsilon Z)] \,,$$

where $Z \sim \mathrm{N}(0, 1)$ *and* $X \sim \pi$, *is such that*

$$\lim_{d \uparrow \infty} \alpha^*(\lambda d^{-1/2}) = \bar{\alpha}(\lambda) := \mathbb{E}[\tilde{\alpha}(\exp(-\lambda^2\psi^2/2)\exp(\lambda\psi W))] \,,$$

where $W \sim \mathrm{N}(0, 1)$, $\psi := \sqrt{\mathbb{E}_\pi[g'(X)^2]}$, *and* $\tilde{\alpha}$ *is defined by* $\tilde{\alpha}(z) := z/(1 + z)$ *for Barker's random-walk sampler and by* $\tilde{\alpha}(z) := 1 \wedge z$ *for the Metropolis-Hastings random-walk sampler. Moreover, the expected squared jump distance,*

$$J(\epsilon) := \mathbb{E}[\alpha(X, X + \epsilon Z)\|\epsilon Z\|^2] \,,$$

where $Z \sim \mathrm{N}(0, 1)$ *and* $X \sim \pi$ *is such that*

$$\lim_{d \uparrow \infty} J(\lambda d^{-1/2}) = \bar{J}(\lambda) := \lambda^2 \bar{\alpha}(\lambda) \,.$$

As shown in Sherlock and Roberts, 2009, such an approach leads to the same result as Theorem 2.3.35;

COROLLARY 2.3.37. *Let the assumptions of Theorem 2.3.36 hold and let* X_t *be the Metropolis-Hastings random-walk sampler. Then,*

$$\bar{\alpha}(\lambda) = 2\Phi\left(-\frac{\lambda\psi}{2}\right) \,.$$

Moreover,

$$\bar{J}(\lambda) = 2\lambda^2\Phi\left(-\frac{\lambda\psi}{2}\right)$$

is maximised when $\lambda = \hat{\lambda} = 2\hat{\beta}/\psi$*, where* $\hat{\beta} > 0$ *is the unique solution to*

$$2\Phi(-\beta) = \beta\phi(\beta) \, .$$

$\hat{\beta} \in (1, \sqrt{2})$*; in fact,* β *is* 1.191 *to three decimal places. This gives an optimal asymptotic expected acceptance rate of* $\bar{\alpha}(\hat{\lambda}) = 2\Phi(-\hat{\beta})$*, which is* 0.234 *to three decimal places.*

Moreover, Theorem 2.3.36 allows for the optimal scaling analysis of Barker's acceptance probability— the proof of which is given in the Appendix as we failed to find the result in the literature;

COROLLARY 2.3.38. *Let the assumptions of Theorem 2.3.36 hold and let* X_t *be Barker's random-walk sampler. Then,*

$$\bar{\alpha}(\lambda) = \mathbb{E}[\{1 + \exp(\lambda^2\psi^2/2 - \lambda\psi W)\}^{-1}] \, .$$

Moreover, $\bar{J}(\lambda) = 2\lambda^2\bar{\alpha}(\lambda)$ *is maximised when* $\lambda = 2\hat{\beta}/\psi$*, where* $\hat{\beta}$ *is the global maximum of*

$$\beta^2\mathbb{E}[\{1 + \exp(2\beta^2 - 2\beta W)\}^{-1}]$$

in the positive domain. Furthermore, $\bar{J}(2\beta/\psi) < 16\beta^2\psi^{-2}\Phi(-\beta)$*, and* $\hat{\beta}$ *is* 1.228 *to three decimal places, which gives an optimal asymptotic expected acceptance rate of* 0.159 *to three decimal places.*

Proof. See A.5. □

As was the case for the Metropolis-Hastings random-walk sampler, the tuning of the acceptance rate for Barker's random-walk sampler is fairly *insensitive* around the optimal asymptotic acceptance rate. This can be seen theoretically in Figure 6 and, in practice, in Figure 4. Indeed, from Figure 6, any asymptotic acceptance rate in $[0.04, 0.3]$ gives an asymptotic expected squared jump distance that is above 60% of the maximum. As such, as was the case for the Metropolis-Hastings random-walk sampler, it is unnecessary to finely tune the *jump-size* to achieve the optimal acceptance rate, provided the *tuned* acceptance rate is on the same scale as the optimal asymptotic acceptance rate. This *insensitivity* is again mirrored in Example 1, as is highlighted in Figure 4, where any acceptance rate in the interval $[0.2, 0.4]$ results in good mixing. Note, from Figure 5, that the efficiency, in terms of the asymptotic expected squared jump distance, tends towards zero as the asymptotic acceptance-rate tends towards zero and one for the Metropolis-Hastings random-walk sampler. For Barker's random-walk sampler, as can be seen in Figure 6, the efficiency tends towards zero as the acceptance-rate tends towards zero and one half. This matches the intuition that the efficiency is worse as the *jump-size* gets very small or very large. The optimal asymptotic expected squared jump distance for the Metropolis-Hastings random-walk sampler is around 38% larger than the optimal asymptotic expected squared jump distance for Barker's random-walk sampler. This suggests that one can expect the

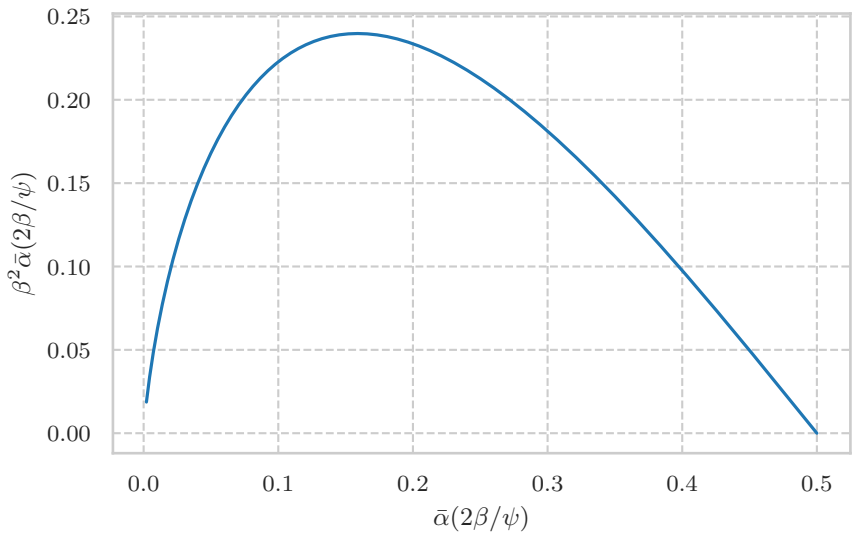

Figure 6: A plot of the asymptotic expected squared jump distance (up to a constant of proportionality) against the asymptotic acceptance rate for Barker's random-walk sampler.

Metropolis-Hastings random-walk sampler to be at least around 38% more efficient then Barker's random-walk sampler. From a practical viewpoint, the approach taken by Sherlock and Roberts, 2009, and its equivalency to the limiting diffusion approach of Roberts, Gelman, and Gilks, 1997, suggests that the expected squared jump distance can be used as a guide when tuning the Metropolis-Hastings random-walk sampler. This suggestion can be seen to work well in practice for Example 1, as can be seen in Figures 4 and 3.

2.4 THE FILTERING PROBLEM

There are two notable filtering models for which exact inference for the *filtering* distributions can be achieved. The first is when the state space is finite, in which case the forward-backward algorithm (see, for example, Rabiner, 1989; Charniak, 1996; Russell and Norvig, 2003) can be employed. The second is when the prior, transition, and observation densities satisfy certain Gaussian properties, in which case the celebrated Kalman filter shows that all of the intermediate filtering densities are Gaussian with means and variances that can be updated sequentially via the so-called *Kalman recursions* (see, for example, West and Harrison, 1999; Russell and Norvig, 2003; Grewal and Andrews, 2011). However, many interesting filtering models do not exhibit tractable distributions, and, while certain approximations, such as the extended Kalman filter (see, for example, Lefebvre, Bruyninckx, and Schutter, 2004, Rapp and Nyman, 2004, or Huang, Mourikis, and Roumeliotis, 2008) and the unscented Kalman filter (see, for instance, Julier and Uhlmann, 1997, Wan and Van Der Merwe, 2000, or Julier and Uhlmann, 2004), can be successfully applied to approximate the

filtering distributions for such models, it is difficult to *gauge* the error in these approximations and, therefore, difficult to say from a practical perspective whether the conclusions from such approximations can be *trusted* or not.

While Markov Chain Monte Carlo algorithms (Section 2.3.5) are an extremely useful set of algorithms for constructing estimators to expectations of functions with respect to some *static* distribution of interest, such as, for example, the posterior of a set of parameters, they are computationally costly and, therefore, inherently unsuited to situations where inference is needed to be performed *sequentially*. As a result, using MCMC algorithms to conduct inference for the position of a stochastic process which is evolving in real-time is infeasible. This *filtering* problem occurs frequently in many contexts, including, for example, tracking a moving object (see, for example, Gustafsson et al., 2002; Brasnett et al., 2005; Mihaylova et al., 2014), learning about the evoluation of an epidemic as it spreads (see, for instance, Yang, Karspeck, and Shaman, 2014; Del Moral and Murray, 2015; Smith, Ionides, and King, 2017), and inferring properties of stock movements in real-time (see, for example, Shephard, 1994; Barndorff-Nielsen, 1997; Kim, Shephard, and Chib, 1998). By repeatedly *mutating*, *correcting*, and *resampling* a set of *weighted particles* in such a way as to account for the sequential evolution of the target densities and for the observations, Sequential Monte Carlo (SMC) methods (see, for instance, Künsch, 2001; Cappé, Moulines, and Ryden, 2006; Doucet and Johansen, 2011; Doucet et al., 2001; Del Moral, 2012) can sequentially build ergodic Monte Carlo appoximations to expectations defined with respect to the target densities via the ratio of two unbiased estimators (see, for example, Cappé, Moulines, and Ryden, 2006; Handel, 2009; Doucet et al., 2001; Del Moral, 2012). Moreover, under suitable conditions on the *filtering* model, the mutation step, and the resampling step, these averages satisfy a Central Limit Theorem (see, for instance, Del Moral and Guionnet, 1999; Del Moral and Miclo, 2000; Chopin, 2004; Cappé, Moulines, and Ryden, 2006; Del Moral, 2012).

Consider, then, a d-dimensional Markov chain, X_t, which, for any time $t \in \mathbb{N}$, admits a *transition* density, $f_t(x_{t+1}|x_t)$; that is, for any $A \in \mathcal{B}(\mathbb{R}^d)$,

$$\mathbb{P}(X_t \in A | X_{t-1} = x_{t-1}) = \int_A f_t(x_t|x_{t-1}) \, \mathrm{d}x_t \ ,$$

and is such that $X_0 \sim f_0$ for some continuous *prior* distribution f_0. Suppose, further, that *noisy*, and potentially *partial*, observations of the Markov chain are available in the form of a sequence of k-dimensional (where $1 \leq k \leq d$) random variables, $Y_{1:\infty}$, which, given the sequence $X_{1:\infty}$, are independent, and, for any $t \in \mathbb{N}$, admit an *observation* density, $g_t(y_t|x_t)$; that is, for any $A \in \mathcal{B}(\mathbb{R}^k)$,

$$\mathbb{P}(Y_t \in A | X_t = x_t) = \int_A g_t(y_t|x_t) \, \mathrm{d}y_t \ .$$

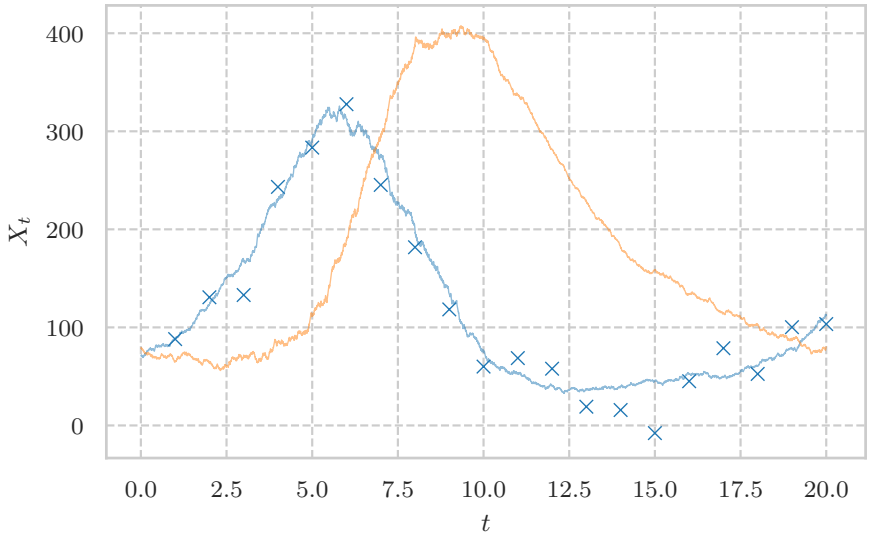

Figure 7: An illustration of the filtering problem. The solid, semi-transparent lines represent a *sample path* of a Lotka-Volterra diffusion (see, for example, Wilkinson, 2006; Boys, Wilkinson, and Kirkwood, 2008, or Section 3.1.2), where the blue line corresponds to the number of prey and the orange line corresponds to the number of predators. The crosses represent noisy observations of the prey *only*. The filtering problem concerns learning about X_t- the number of predators and prey— at time $t = 20$, given the observations.

At any time $t \geq 1$ interest lies in the posterior distribution, denoted π_t, of the Markov chain, given the observations at any time at, or prior to, t, which, via Bayes' theorem, is equal to

$$\pi_t(x_t) = \frac{g_t(y_t|x_t) \int_{\mathbb{R}^d} f_t(x_t|x_{t-1})\pi_{t-1}(x_{t-1})\, \mathrm{d}x_{t-1}}{\iint_{\mathbb{R}^{2d}} g_t(y_t|x_t) f_t(x_t|x_{t-1})\pi_{t-1}(x_{t-1})\, \mathrm{d}x_{t-1}\mathrm{d}x_t} \,. \tag{35}$$

The filtering problem consists of approximating expectations of the form

$$\pi_t[h] = \int_{\mathbb{R}^d} h(x_t)\pi_t(x_t)\, \mathrm{d}x_t \,,$$

sequentially through time as the observations, y_t, stream in. An illustration of this statistical paradigm can be seen in Figure 7.

2.4.1 *The Particle Filter*

The particle filter allows approximations to $\pi_t[h]$ to be constructed for a variety of filtering problems and these approximations satisfy several

desirable properties. In essence, the particle filtering approach to the filtering problem relies on the observation that, upon rewriting (35),

$$\pi_t(x_t) = \gamma_t(\mathbb{R}^{dt})^{-1} \int_{\mathbb{R}^{d(t-1)}} \gamma_t(x_{0:t}) \, \mathrm{d}x_{0:t-1} \ ,$$

where

$$\gamma_t(x_{0:t}) = f_0(x_0) \prod_{s=1}^{t} f_s(x_s|x_{s-1}) g_s(y_s|x_s) \ ,$$

expectations of the form $\pi_t[h]$ can be rewritten as

$$\pi_t[h] = \gamma_t(\mathbb{R}^{dt})^{-1} \int_{\mathbb{R}^{dt}} h(x_t)\gamma_t(x_{0:t}) \, \mathrm{d}x_{0:t} \ . \tag{36}$$

With this rewrite of the problem, the particle filter can be seen as a dynamic generalisation of the normalized Importance Sampling estimator of Section 2.3.2 where *particles*, $x_t^{1:N}$, along with their corresponding normalized *weights*, $\tilde{w}_t^{1:N}(x_t^{1:N})$, are updated sequentially through time to account for the evolution of the target densities π_t.

2.4.1.1 *Sequential Importance Sampling*

A naive approach to constructing an estimator of $\pi_t[h]$ would involve constructing a normalized importance sampling estimator (see Equation (14) of Section 2.3.3) of $\pi_t[h]$. Let $q_t(x_{0:t})$ be a proposal density which may or may not depend on any number of the observations $y_{1:\infty}$. For any time $t \in \mathbb{N}$, the *importance weights*, w_t, are equal to

$$\frac{\gamma_t(x_{0:t})}{q_t(x_{0:t})} = \frac{g_t(y_t|x_t) f_t(x_t|x_{t-1}) q_{t-1}(x_{0:t-1})}{q_t(x_{0:t})} w_{t-1}(x_{0:t-1}) \ . \tag{37}$$

Using such a general proposal, however, is computationally inefficient, since, in general, simulating a path, $x_{0:t}$, and calculating the *likelihood* of that path, via $q_t(x_{0:t})$, will become ever more costly as time increases. A more efficient approach is to use a *sequential* proposal;

$$q_t(x_{0:t}) = p_0(x_0) \prod_{s=1}^{t} p_s(x_s|x_{0:s-1}) \ ,$$

since, then, for any time $t \in \mathbb{N}$, the *importance weights* satisfy the following recursion;

$$w_t(x_{0:t}) = \frac{g_t(y_t|x_t) f_t(x_t|x_{t-1})}{p_t(x_t|x_{0:t-1})} w_{t-1}(x_{0:t-1}) \ , \tag{38}$$

Hence, the normalized importance weights and the normalized importance sampling estimator of $\pi_t[h]$; namely,

$$\sum_{i=1}^{N} \tilde{w}_t^{(i)}(X_{0:t}^{(1:N)}) h(X_t^{(i)}) \ , \quad \tilde{w}_t^{(i)}(X_{0:t}^{(1:N)}) = w_t(X_{0:t}^{(i)}) \Big/ \sum_{j=1}^{N} w_t(X_{0:t}^{(j)}) \ ,$$

(39)

can be efficiently updated through time. Provided, for any $t \in \mathbb{N}$, $\text{supp}(\gamma_t) \subseteq \text{supp}(q_t)$, the sequential importance sampling estimator satisfies the same ergodic properties as the normalized importance sampling estimator. Unfortunately, due to the product form of the recursion in Equation (37), the variance of the normalized weights will grow very quickly and, therefore, the estimator of the form (39) using such weights will have a large variance. This *particle degeneracy* problem makes the use of the sequential importance sampling estimator impractical for many scenarios of interest.

2.4.1.2 *Sequential Importance Resampling*

Resampling the particles at each time step according to the normalised weights, $\tilde{w}_t^{(1:N)}$, can overcome the particle degeneracy phenomenon by discarding those particles with a relatively small weight and duplicating and propagating those particles with a relatively large weight. This thesis defines a resampling procedure as follows:

DEFINITION 2.4.1 (Resampling Procedure). *Given a set of normalised weights, $\tilde{w}^{(1:N)}$, a resampling procedure consists of the following three steps;*

1. *The number of offspring assigned to each particle, denoted, $O^{(1:N)}$, is sampled from a probability mass function $\bar{\kappa}(\cdot|\tilde{w}^{(1:N)})$ such that $O^{(1)} + \ldots + O^{(N)} = N$, and;*
   a) *For each $i \in \{1, \ldots, N\}$,*
   $$\mathbb{E}(O^{(i)}) = N\tilde{w}^{(i)} , \tag{40}$$
   so that, the larger $\tilde{w}^{(i)}$ is, the more offspring particle i has on average.
   b) *For any permutation, σ, of $\{1, \ldots, N\}$,*
   $$\bar{\kappa}(o^{(1:N)}|\tilde{w}^{(1:N)}) = \bar{\kappa}(o^{(\sigma(1))}, \ldots, o^{(\sigma(N))}|\tilde{w}^{(\sigma(1))}, \ldots, \tilde{w}^{(\sigma(N))}) , \tag{41}$$
   so that, the assignment of the offspring does not depend on the order of the weights.
2. *Based on the number of offspring assigned to each particle, the ancestors of each particle, denoted $A^{(1:N)}$, are determined in such a way that, for any $i \in \{1, \ldots, N\}$,*
   $$\sum_{j=1}^{N} \mathbb{1}_{\{i\}}(A^{(j)}) = O^{(i)} . \tag{42}$$
3. *The new particles, whose ancestors are given by $A^{(1:N)}$, are then all given an equal weight.*

Therefore, broadly speaking, the resampling procedure is a mapping from a set of particles and weights $\{x_t^{(1:N)}, \tilde{w}_t^{(1:N)}(x_t^{(1:N)})\}$ to a new set of particles and weights $\{x_t^{(A_t^{(1:N)})}, 1/N\}$.

REMARK 2. *In the literature, Condition (41) is not normally imposed. Indeed, many properties of the sequential importance resampler do not require this condition to hold. Moreover, not all resampling schemes satisfy such a condition; for example, the commonly discussed stratified and systematic resampling procedures do not satisfy (41). However, any resampling scheme can be made to satisfy (41) if the weight and position pairs are relabelled randomly. This condition results in the exchangeability (see Definition 2.1.2) of the paths generated by the Sequential Importance Resampling procedure. This is not the only way to ensure that the sample paths are exchangeable. Indeed, in the Particle Markov Chain Monte Carlo literature- see, for instance, Andrieu, Doucet, and Holenstein, 2010- exchangeability of paths can be ensured by randomly permuting the ancestor variables; see the discussion following Assumptions 4.2.1 of Section 4.2. However, the condition imposed here ensures that the resampling schemes are exchangeable on their own, without any reliance on previous steps of the sequential importance resampler and, therefore, under such a condition, exchangeability of the paths is clearer and easier to demonstrate.*

For brevity, the first two steps in the resampling procedure of Definition 2.4.1 can be combined into one step which involves simulating a set of ancestors, $A^{(1:N)}$, from a probability mass function, κ, which satisfies the following assumptions;

ASSUMPTIONS 2.4.2. *Given a sequence of normalised weights, $\tilde{w}^{(1:N)}$, the resampling probability mass function, $\kappa(\cdot|\tilde{w}^{(1:N)})$, is such that;*

(U) For any $k \in \{1, \ldots, N\}$,

$$\mathbb{E}\left[\sum_{i=1}^{N} \mathbb{1}_k(A^{(i)}) \,\middle|\, \tilde{w}^{(1:N)}\right] = N\tilde{w}^{(k)} .$$

(E) For any permutation, σ, of $\{1, \ldots, N\}$,

$$\kappa(a^{(1:N)}|\tilde{w}^{(1:N)}) = \kappa(a^{(1:N)}|\tilde{w}^{(\sigma(1))}, \ldots, \tilde{w}^{(\sigma(N))}) .$$

REMARK 3. *In practice, the assignment of the ancestors is done deterministically; that is,*

$$A^{(j)} = i \,, \quad \text{if} \quad \sum_{k=1}^{i-1} O^{(k)} < j \leq \sum_{k=1}^{i} O^{(k)} .$$

As such, if (40) and (41) hold, then so do Assumptions 2.4.2.

All the resampling schemes discussed in this thesis rely on transforming a sequence $u_{1:N} \in [0,1]^N$ of sample random variables. Indeed, for

any $x_{1:N}$, and any $j \in \{0, \dots, N\}$, define $s_j(x_{1:N})$ to be the sequence of partial sums; that is,

$$s_0(x_{1:N}) := 0 \,, \quad s_j(x_{1:N}) := x_1 + \ldots + x_j \quad \text{for any } j \in \{1, \dots, N\}. \tag{43}$$

For a given permutation, σ, of $\{1, \dots, N\}$, and a sequence of normalised weights $\tilde{w}^{(1:N)}$, let $\tilde{w}_\sigma^{(1:N)} = (\tilde{w}^{(\sigma(1))}, \dots, \tilde{w}^{(\sigma(N))})$ be the relabelled normalised weights. Finally, let

$$I_j(\sigma) := (s_{j-1}(\tilde{w}_\sigma^{(1:N)}), s_j(\tilde{w}_\sigma^{(1:N)})] \,,$$

Then, given a sequence $u_{1:N} \in [0,1]^N$ of sample random variables, the number of offspring associated with each *permuted* particle $j \in \{1, \dots, N\}$ is given by

$$o_\sigma^j(u_{1:N}) := \sum_{i=1}^{N} \mathbb{1}_{I_j(\sigma)}(u_i) \,, \tag{44}$$

Given $o_\sigma^j(u_{1:N})$ for each $j \in \{1, \dots, N\}$, one can set $o^{(j)} := o_\sigma^{(\sigma^{-1}(j))}$ where σ^{-1} denotes the inverse of the permutation σ. As shown in Theorem 2.4.3 below, three well documented resampling schemes that satisfy Assumption (40) and, therefore, Assumption 2.4.2 when combined with Equation (42); multinomial, stratified, and systemic resampling, arise from (44) via different choices for $u_{1:N}$. Indeed, multinomial resampling simulates a sequence $U_{1:N}$ of independent and identically distributed standard uniform random numbers; that is, each $U_i \sim \mathrm{U}(0,1)$. Stratified resampling, on the other hand, simulates an independent, but not identically distributed, sequence $U_{1:N} \in [0,1]^N$ of random numbers by choosing, for any $i \in \{1, \dots, N\}$, $U_i \sim \mathrm{Unif}((i-1)/N, i/N)$. Finally, systematic resampling simulates a dependent and not identically distributed sequence $U_{1:N} \in [0,1]^N$ of random numbers by choosing $U_1 \sim \mathrm{Unif}(0, 1/N)$ and then, for any $i \in \{2, \dots, N\}$, deterministically setting $U_i := U_1 + (i-1)/N$. Multinomial resampling satisfies (41), whereas stratified and systematic resampling do not. Therefore, in order to satisfy (41), the weights need to be shuffled prior to applying either the stratified or systematic resampling procedure. To be precise, Algorithms 4, 5, and 6 illustrate, respectively, the multinomial, stratified, and systematic resampling procedures as defined in this thesis. Note that the stratified and systematic resampling implementations depend on a shuffle and an inverse shuffle. See Definition B.0.11 and Algorithms 24 and 25 for details.

Theorem 2.4.3 shows that all three of these resampling schemes satisfy Assumption (40). Moreover, bounds on the difference between the number of offpsring and its expectation are given and shown to be tight- as far as we are aware, a formal proof of such bounds and their tightness is novel;

THEOREM 2.4.3. *Let $\tilde{w}^{(1:N)}$ be a sequence of normalised weights and let $O^{(1:N)}$ be a sequence of offspring derived by multinomial, stratified,*

Algorithm 4 Multinomial Resampling

1: Simulate $u_{1:N}$ independently from a Unif$(0,1)$ distribution.
2: Calculate the partial sums $s_j(\tilde{w}^{(1:N)})$ for every $j = 0, \ldots, N$.
3: **for** $j = 1, \ldots, N$ **do**
4: Set

$$o^{(j)} = \sum_{i=1}^{N} \mathbb{1}_{I_j(\sigma)}(u_i) \, .$$

5: **end for**

Algorithm 5 Stratified Resampling

1: Shuffle the weights; $(\tilde{w}_\sigma^{(1:N)}, \sigma) = \text{shuffle}(\tilde{w}^{(1:N)})$.
2: **for** $i = 1, \ldots, N$ **do**
3: Sample u_i from a Unif$((i-1)/N, i/N)$ distribution.
4: **end for**
5: Calculate the partial sums $s_j(\tilde{w}_\sigma^{(1:N)})$ for every $j = 0, \ldots, N$.
6: **for** $j = 1, \ldots, N$ **do**
7: Set

$$o_\sigma^{(j)} = \sum_{i=1}^{N} \mathbb{1}_{I_j(\sigma)}(u_i) \, .$$

8: **end for**
9: Invert the shuffle on the offspring; $o^{(1:N)} = \text{inverse\_shuffle}(o_\sigma^{(1:N)}, \sigma)$.

or systematic resampling. Then, for any $j \in \{1, \ldots, N\}$, $\mathbb{E}(O^{(j)}) = N\tilde{w}^{(j)}$. *Moreover, for any permutation,* σ, *of* $\{1, \ldots, N\}$,

$$\bar{\kappa}(o^{(1:N)} | \tilde{w}^{(1:N)}) = \bar{\kappa}(o^{(\sigma(1))}, \ldots, o^{(\sigma(N))} | \tilde{w}^{(\sigma(1))}, \ldots, \tilde{w}^{(\sigma(N))}) \, . \quad (45)$$

Finally, for stratified resampling, $|O^{(j)} - N\tilde{w}^{(j)}| < 2$ *for any* $j \in \{1, \ldots, N\}$ *and this bound is tight. On the other hand, for systematic resampling,* $|O^{(j)} - N\tilde{w}^{(j)}| < 1$ *for any* $j \in \{1, \ldots, N\}$ *and this bound is tight.*

Proof. See A.6. □

Any of the aforementioned resampling schemes can, in terms of computational efficiency and/or statistical efficiency, be immediately improved by only applying them to the *residual* weights as described in Algorithm 7[11]. Theorem 2.4.4 shows that residual resampling with any resampling scheme which satisfies (40) satisfies (40) and, therefore, Assumption 2.4.2 when combined with Equation (42). Moreover, Theorem 2.4.4 shows that (41) also holds and, again, derives novel bounds on the difference between the number of offpsring and its expectation, and shows that these bounds are tight;

THEOREM 2.4.4. *Let* $\tilde{w}^{(1:N)}$ *be a sequence of normalised weights. Consider assigning* N *particles to offspring* $O^{(1:N)}$ *via residual resampling (Algorithm 7) with any resampling scheme which, given normalised*

11 In the literature residual resampling refers to multinomial resampling applied to the residuals.

Algorithm 6 Systematic Resampling

1: Shuffle the weights; $(\tilde{w}_\sigma^{(1:N)}, \sigma) = \text{shuffle}(\tilde{w}^{(1:N)})$.
2: Sample u_1 from a $\text{Unif}(0, 1/N)$.
3: **for** $i = 2, \ldots, N$ **do**
4: Set $u_i = u_1 + (i-1)/N$.
5: **end for**
6: Calculate the partial sums $s_j(\tilde{w}_\sigma^{(1:N)})$ for every $j = 0, \ldots, N$.
7: **for** $j = 1, \ldots, N$ **do**
8: Set

$$o_\sigma^{(j)} = \sum_{i=1}^{N} \mathbb{1}_{I_j(\sigma)}(u_i) \, .$$

9: **end for**
10: Invert the shuffle on the offspring; $o^{(1:N)} = \text{inverse\_shuffle}(o_\sigma^{(1:N)}, \sigma)$.

Algorithm 7 Residual Resampling

1: Initialise by setting $s = 0$.
2: **for** $j = 1, \ldots, N$ **do**
3: Set $o_b^{(j)} = \lfloor N\tilde{w}^{(j)} \rfloor$.
4: Set $s = s + o_b^{(j)}$.
5: Set $w_r^{(j)} = \tilde{w}^{(j)} - o_b^{(j)}/N$.
6: **end for**
7: Normalize the residual weights by setting, for each $j \in \{1, \ldots, N\}$,

$$\tilde{w}_r^{(j)} = w_r^{(j)} / (w_r^{(1)} + \ldots + w_r^{(N)}) \, .$$

8: Resample $N - S$ particles with weights $\tilde{w}_r^{(1:N)}$ to get offsprings $o_r^{(1:N)}$.
9: Set $o^{(j)} = o_b^{(j)} + o_r^{(j)}$ for all $j \in \{1, \ldots, N\}$.

residual weights, $\tilde{w}_r^{(1:N)}$, assigns M residual particles to offspring $O_r^{(1:N)}$ according to the mass function $\bar{\kappa}_r(\cdot|\tilde{w}_r^{(1:N)})$, which is such that

1. *For any $j \in \{1, \ldots, N\}$, $\mathbb{E}(O_r^{(j)}) = M\tilde{w}_r^{(j)}$.*
2. *For any permutation, σ, of $\{1, \ldots, N\}$,*

$$\bar{\kappa}_r(o_r^{(1:N)}|\tilde{w}_r^{(1:N)}) = \bar{\kappa}_r(o_r^{(\sigma(1))}, \ldots, o_r^{(\sigma(N))}|\tilde{w}_r^{(\sigma(1))}, \ldots, \tilde{w}_r^{(\sigma(N))}) \, .$$

Then, for any $j \in \{1, \ldots, N\}$, $\mathbb{E}(O^{(j)}) = N\tilde{w}^{(j)}$. Furtheremore, defining $\bar{\kappa}(\cdot|\tilde{w}^{(1:N)})$ to be the mass function of such a scheme,

$$\bar{\kappa}(o^{(1:N)}|\tilde{w}^{(1:N)}) = \bar{\kappa}(o^{(\sigma(1))}, \ldots, o^{(\sigma(N))}|\tilde{w}^{(\sigma(1))}, \ldots, \tilde{w}^{(\sigma(N))}) \, ,$$

for any permutation, σ, of the set $\{1, \ldots, N\}$. Moreover, for any resampling scheme used within residual resampling, $-1 < O^{(j)} - N\tilde{w}^{(j)} < N-1$. For multinomial residual resampling, this bound is tight. Furthermore, for stratified residual resampling, $-1 < O^{(j)} - N\tilde{w}^{(j)} < 2$, and this bound is tight. Finally, for systematic resampling, $|O^{(j)} - N\tilde{w}^{(j)}| < 1$, and this bound is tight.

Proof. See A.7. □

REMARK 4. *The proof of Theorem 2.4.3 demonstrates that multinomial, stratified, and systematic resampling all satisfy the assumption in Theorem 2.4.4 on the resampling scheme used in residual resampling.*

REMARK 5. *Systematic residual resampling does not improve the bounds on* $O^{(j)} - N\tilde{w}^{(j)}$ *that systematic resampling provides. However, it is more efficient in terms of computational cost.*

Any resampling leads to an immediate increase in the variance of the normalized importance sampling estimator. However, resampling also reduces the degeneracy of the particles and thereby, often, leads to a decreased variance in the estimator for larger times t. It is this reduction in degeneracy that makes the Sequantial Importance Resampling estimator an effective estimator of $\pi_t[h]$. It is natural to choose a resampling scheme which aims to minimise the increase in variance brought about by resampling; that is, one which aims to minimise the resampling variance,

$$\mathrm{Var}\left[\frac{1}{N}\sum_{i=1}^{N} O_i(\tilde{w}^{(1:N)}(\tilde{x}^{(1:N)}))h(x^{(i)})\right]. \tag{46}$$

In light of Theorem 2.4.3, it is natural to conjecture that the resampling variance for systematic resampling is no greater than the resampling variance for stratified resampling which, itself, is no greater than the resampling variance for multinomial resampling. Moreover, in light of Theorem 2.4.4, it is also natural to conjecture that the resampling variance for residual resampling, which uses a given resampling scheme, is no greater than the resampling variance of the given resampling scheme used on its own. Douc, Cappé, and Moulines, 2005 show that stratified resampling without shuffling has a smaller resampling variance than multinomial resampling, and that multinomial residual resampling has a smaller resampling variance than multinomial resampling. However, Douc, Cappé, and Moulines, 2005 also show that it is not generally true that systematic resampling without shuffling has a smaller resampling variance than multinomial resampling— however they do suggest that this might be true if one introduces a shuffle. As such, it is not immediately clear which resampling scheme one should prefer over the others. However, under certain conditions, multinomial and stratified resampling, along with their residual extensions lead to Sequential Importance Resampling Estimators which satisfy a Strong Law of Large Numbers result and a Central Limit Theorem. Moreover, their resampling variance is understood in closed form (see, for example, Del Moral and Guionnet, 1999; Del Moral and Miclo, 2000; Chopin, 2004; Cappé, Moulines, and Ryden, 2006; Del Moral, 2012). On the other hand, given the dependence between the sampled offspring that is inherent with systematic and systematic residual resampling, theoretical guarantees for these samplers is lacking. As such, this thesis will, henceforth, concentrate on the stratified residual resampling procedure.

3

SIMULATING CONDITIONED DIFFUSIONS

3.1 THE INTRODUCTION

Diffusions are a flexible class of continuous-time Markov processes whose dynamics are completely characterized by specifying an instantaneous change in mean (henceforth the drift) and an instantaneous variance (henceforth the volatility). This makes them a useful class of processes for building rich models and, as such, they are utilised in many scientific disciplines, including, but not limited to, biology (see, for example, Golightly and Wilkinson, 2011), finance (see, for instance, Aït-Sahalia and Kimmel, 2007), and engineering (see, for example, Coffey, Kalmykov, and Waldron, 2004). In biological applications, and, more generally, in applications involving reaction networks, diffusions are often used as approximate models for the evolution of the numbers of a set of species within a reaction network. In particular, the chemical Langevin diffusion is often used to approximate the chemical master equation (see, for instance, Ethier and Kurtz, 1986; Van Kampen, 1992; Wilkinson, 2006; Komorowski et al., 2009; Fearnhead, Giagos, and Sherlock, 2014).

A d-dimensional diffusion, X_t, (introduced in Section 2.1.4 of this thesis) can be defined as the solution to a stochastic differential equation (SDE)

$$\mathrm{d}X_t = \mu(X_t, t, \Theta)\,\mathrm{d}t + \sigma(X_t, t, \Theta)\,\mathrm{d}W_t\,, \quad X_0 = x_0\,, \tag{47}$$

where $t \in [0, T]$, W_t is an r-dimensional Wiener process , and $\Theta \in \mathbb{R}^m$ is a vector of unknown parameters. The drift $\mu : \mathbb{R}^d \times [0, \infty) \times \mathbb{R}^m \to \mathbb{R}^d$ corresponds to the infinitesimal change in mean, and the volatility $\zeta := \sigma\sigma^* : \mathbb{R}^d \times [0, \infty) \times \mathbb{R}^m \to \mathbb{R}^{d\times d}$ corresponds to the infinitesimal variance in the sense that

$$\mathbb{E}(X_{t+\Delta t}|X_t = x, \Theta = \theta) = x + \Delta t\mu(x, t, \theta) + o(\Delta t)\,, \tag{48}$$

$$\mathrm{Var}(X_{t+\Delta t}|X_t = x, \Theta = \theta) = \Delta t\zeta(x, t, \theta) + o(\Delta t)\,. \tag{49}$$

Note that, throughout, to avoid confusion with the inter-observation time, we use A^* to denote the transpose of a vector or matrix A. Both the drift and volatility depend on a vector of unknown parameters, Θ, which has a prior density of $p_0(\theta)$. These parameters (which drive the evolution of X_t) often relate to quantities of interest, such as the birth rate of a species, and in light of sparse, noisy, and partial observations of the diffusion, inference for these parameters, along with paths of the diffusion, can theoretically proceed in a Bayesian framework via the particle MCMC methodology of Andrieu, Doucet, and Holenstein, 2010. Such schemes rely on the construction of an unbiased approximation to the likelihood of the observations, π, which is typically obtained

through an importance-sampling and, more generally, particle-filtering approach.

Sample paths of the diffusion are infinite-dimensional and therefore, in practice, it is necessary to restrict attention to the construction of finite-dimensional skeleton paths of the diffusion. Moreover, the transition density of a large class of diffusions is intractable and exact simulation (see, for instance, Beskos, Papaspiliopoulos, and Roberts, 2006) of a skeleton path is impossible for most multivariate diffusions. Therefore, for many diffusions, it is necessary to approximate the transition density along a fine grid of skeletal points by a Gaussian density using an Euler-Maruyama (EM) step (see, for example, Kloeden and Platen, 1992 or Section 3.2).

As highlighted in Section 2.4.1.2 of this thesis, the efficiency of any sequential importance resampling scheme and, therefore, of any particle MCMC scheme depends on the variability of the importance weights. Hence, the construction of proposal densities which are consistent with respect to both the observations and the true diffusion is key to designing computationally efficient algorithms. The forward simulation (FS) proposal of Pedersen, 1995 (see Section 3.2.2) uses the EM approximation to simulate skeleton paths between consecutive observations. Such a proposal can suffer from poor performance, particularly for informative observations, since it simulates paths independently of the observations. The modified diffusion bridge (MDB) of Durham and Gallant, 2002 (see Section 3.2.3) overcomes this deficiency by using an EM approximation to the transition density between the current point of the skeleton and any subsequent point, thus leading to a tractable, Gaussian transition density between consecutive points of the skeleton given the next observation. However, such a proposal performs poorly if sample paths of the diffusion exhibit non-linear dynamics as is often the case over relatively large inter-observation times. Lindström, 2012 tackles this issue by constructing a proposal which is a mixture between the FS approach and the MDB approach (see Section 3.2.4). The downsides of such a proposal are that, firstly, it needs careful tuning, and, secondly, it is not clear how the proposal behaves as the width of the partition tends towards zero. These drawbacks also hold for the proposal of Fearnhead, 2008 which comprises of a mixture between the FS approach and an approach which simulates from the stationary distribution of the diffusion when it exists (see Section 3.2.4). Schauer, Meulen, and Zanten, 2017 (and, also, Meulen and Schauer, 2017) take a different approach and consider the form of the SDE satisfied by the diffusion conditioned on the next observation; this, in general, has the same volatility as the unconditioned diffusion and an extra term in the drift (see, for example, Chapter IV, Section 39, Rogers and Williams, 2000b) which *guides* the diffusion towards the observation. This extra term depends on the transition density of the unconditioned diffusion and thus, typically, needs to be approximated by the transition density of a tractable diffusion before forward simulation of a skeleton path (via the EM approximation) can proceed. Schauer, Meulen, and Zanten, 2017 prove that approximating the transition density of the unconditioned

diffusion with the transition density of a diffusion driven by a linear SDE leads to a diffusion which is absolutely continuous with respect to the true, conditioned diffusion. Unfortunately, implementing such an approach in a statistically efficient way can lead to a computationally expensive algorithm as shown by (Whitaker et al., 2017).

The novel proposal introduced in Section 3.3 of this thesis can be seen as a natural extension to the residual-bridge constructs of Whitaker et al., 2017 who propose improving on the MDB approach by: constructing a deterministic path which captures the non-linear dynamics of the diffusion, applying the MDB approximation to the residual process defined as the difference between the true diffusion and this path, and then adding the path back on. An appropriate choice of the deterministic path results in a residual whose dynamics are more linear and thus a proposal density which is closer to the true transition density. It is shown empirically in Whitaker et al., 2017 that, for several diffusions, this proposal, when implemented within a Metropolis-Hastings importance sampler leads to a larger empirical acceptance probability than a Metropolis-Hastings importance sampler which uses either the MDB or the construct introduced by Lindström, 2012 as a proposal distribution. Furthermore, this empirical acceptance probability is similar to the empirical acceptance probability of a Metropolis-Hastings importance sampler which uses the guided proposals of Schauer, Meulen, and Zanten, 2017 as a proposal distribution but is achieved with a considerably smaller computational cost. However, this residual-bridge approach, while accounting for the variability in the drift, does not account for the variability in the volatility and can, therefore, perform poorly in scenarios where the volatility varies substantially. This is often the case for larger inter-observation intervals, where the diffusion itself moves substantially over the state space, and, for diffusions whose volatility is time-inhomogeneous. The proposal introduced in this thesis generalizes the residual-bridge proposals of Whitaker et al., 2017 by applying the approximation used in the MDB to the difference between the true diffusion and a second, carefully chosen, approximate diffusion which is coupled with the original diffusion via the same driving Brownian motion. By attempting to account for the variability in the volatility, this new proposal can lead to greater statistical efficiency in situations where the volatility varies considerably.

To compare different approaches to simulating conditioned diffusions, we will, in this thesis, concentrate on three diffusions which commonly occur in practice; the birth-death diffusion (BD, Section 3.1.1), the Lotka-Volterra diffusion (LV, Section 3.1.2), and a diffusion corresponding to a simple model of gene expression (GE, Section 3.1.3). After introducing these diffusions in the subsequent sections, we will, in Section 3.2, describe the general framework for simulating conditioned diffusions along a discretised interval. Absolute continuity, and how it relates to this thesis, will be briefly discussed in Section 3.2.1, after which, the remainder of Section 3.2 will describe the approaches currently taken in the literature. In Section 3.3 we will introduce new bridges based on residual processes, discuss how these bridges build upon the

residual bridges proposed in Whitaker et al., 2017, and highlight, in Section 3.3.1, the extra computational cost incurred when simulating such bridges. Finally, in Section 3.3.2, we will conduct a simulation study comparing the new residual bridge constructs with the residual bridge constructs of Whitaker et al., 2017 (Section 3.2.5) and with the MDB of Durham and Gallant, 2002 (Section 3.2.3). For clarity, and to avoid duplicated results, we have avoided comparison of the new bridges with each of the bridges introduced in this thesis. Comparisons between the bridges introduced in this thesis can be found in the literature. Indeed, for a comparison between the MDB and the forward simulation approach of Pedersen, 1995, see Durham and Gallant, 2002 and Fearnhead, 2008. The latter also compares these approaches with the approach introduced in that article. Lindström, 2012 compares his approach with the forward simulation approach, the MDB, and the approach of Fearnhead, 2008. Finally, Whitaker et al., 2017 compare the MDB, the Lindström bridge, and the guided proposals of Schauer, Meulen, and Zanten, 2017 with the residual bridge constructs they introduce.

3.1.1 *The Birth-Death Diffusion*

The Birth-Death diffusion (see, for example, Whitaker et al., 2017) is an approximate model for the evolution of the number, X_t, of a species which is subject to two forces; births and deaths, with rates per species member of θ_1 and θ_2 respectively. Such a diffusion satisfies

$$\mathrm{d}X_t = (\theta_1 - \theta_2)X_t \,\mathrm{d}t + \sqrt{(\theta_1 + \theta_2)X_t}\,\mathrm{d}B_t\,, \quad X_0 = x_0$$

over the interval $[0, T]$.

3.1.2 *The Lotka-Volterra Diffusion*

The Lotka-Volterra diffusion (see, for instance, Wilkinson, 2006) is an approximate model for the evolution of the numbers, $X_t = [X_t^{(1)}, X_t^{(2)}]^*$, of two species (prey and predators respectively) which are subject to three forces; prey reproduce with rate θ_1, predators reproduce through eating prey with rate θ_2, and predators die with rate θ_3. Such a diffusion satisfies

$$\begin{aligned}\begin{bmatrix}\mathrm{d}X_t^{(1)}\\ \mathrm{d}X_t^{(2)}\end{bmatrix} &= \begin{bmatrix}\theta_1 X_t^{(1)} - \theta_2 X_t^{(1)} X_t^{(2)}\\ \theta_2 X_t^{(1)} X_t^{(2)} - \theta_3 X_t^{(2)}\end{bmatrix}\,\mathrm{d}t \\ &\quad + \begin{bmatrix}\theta_1 X_t^{(1)} + \theta_2 X_t^{(1)} X_t^{(2)} & -\theta_2 X_t^{(1)} X_t^{(2)}\\ -\theta_2 X_t^{(1)} X_t^{(2)} & \theta_2 X_t^{(1)} X_t^{(2)} + \theta_3 X_t^{(2)}\end{bmatrix}^{1/2}\,\mathrm{d}W_t\,,\end{aligned}$$

where, for a matrix A, $A^{1/2}$ denotes any matrix square-root, so that $(A^{1/2})(A^{1/2})^* = A$.

3.1.3 *A Diffusion for a Simple Gene Expression Model*

In this subsection we introduce the diffusion which approximates a simple model for gene expression (see, for instance, Komorowski et al., 2009; Golightly, Henderson, and Sherlock, 2015). This diffusion approximately describes the evolution of the numbers, $X_t = [R_t, P_t]^T$, of two biochemical species (mRNA and protein molecules respectively) which are subject to three forces; transcription with a time-inhomogeneous rate $k_R(t)$, mRNA degradation with rate γ_R, translation with rate k_P, and protein degradation with rate γ_P. As in Komorowski et al., 2009; Golightly, Henderson, and Sherlock, 2015, we take the rate $k_R(t)$ to be of the form

$$k_R(t) = b_0 \exp(-b_1(t - b_2)^2) + b_3 \ ,$$

so that the complete vector of unknown parameters is

$$\theta = (\gamma_R, \gamma_P, k_P, b_0, b_1, b_2, b_3) \ .$$

Such a diffusion satisfies

$$\begin{bmatrix} \mathrm{d}R_t \\ \mathrm{d}P_t \end{bmatrix} = \begin{bmatrix} k_R(t) - \gamma_R R_t \\ k_P R_t - \gamma_P P_t \end{bmatrix} \mathrm{d}t + \begin{bmatrix} \sqrt{k_R(t) + \gamma_R R_t} & 0 \\ 0 & \sqrt{k_P R_t + \gamma_P P_t} \end{bmatrix} \mathrm{d}W_t \ .$$

3.2 SIMULATING CONDITIONED DIFFUSIONS

Let X_t be a d-dimensional diffusion satisfying Equation (47). Consider the pre-defined sequence of times,

$$\{(t_0, \ldots, t_I) \in [0, T]^{I+1} : \ 0 =: t_0 < t_1 < \ldots < t_I := T\} \ .$$

We have noisy observations, $(y_{t_1}, \ldots, y_{t_I}) \in \mathbb{R}^{r \times I}$, of the diffusion at times $(t_1, \ldots, t_I)$ such that, for any $i \in \{1, \ldots, I\}$,

$$(Y_{t_i} | X_{t_i} = x) \sim \mathrm{N}(P_i x, V_i) \ ,$$

where $P_i \in \mathbb{R}^{r \times d}$, and $V_i \in \mathbb{R}^{r \times r}$ is symmetric and positive-definite. Denote the density of the i-th observation by $g_i(y_{t_i} | x_{t_i})$ and between any two consecutive times, t_i and t_{i+1}, define an equispaced partition, $\mathcal{P}^{(i)}_{\Delta t}$, to be the set

$$\{(t_{i[0]}, \ldots, t_{i[K_i]}) \in [t_i, t_{i+1}]^{K_i+1} : \ t_i =: t_{i[0]} < \ldots < t_{i[K_i]} := t_{i+1}\}$$

such that, for all $j \in \{0, \ldots, K_i\}$, $t_{i[j]} := t_i + j\Delta t$ with $\Delta t > 0$ and small. For convenience, denote any variable $\psi_{t_{i[j]}}$ by $\psi^{(i)}_j$ with $\psi_{t_{i[0]}}$ denoted by $\psi^{(i)}$ so that, for instance, $y_{t_{1[0]}} = y^{(1)}$ is the first observation, and $x^{(I)}_{K_I} = x_T$ is the value of the path at the final time point. Denote the transition density of the diffusion by

$$f^{s,t}_\theta(x|z) := \lim_{\epsilon \downarrow 0} \mathbb{P}(X_t \in [x, x + \epsilon) | X_s = z, \Theta = \theta)/\epsilon^d \ ,$$

where

$$[x, x+\epsilon) := \{v \in \mathbb{R}^d : x_i \leq v_i < x_i + \epsilon \text{ for all } i \in \{1, \ldots, d\}\} .$$

Interest lies in $\pi(\theta, x_{\mathcal{P}_{\Delta t}}|y^{1:I})$ which is the posterior density for Θ and the skeleton path defined at the points of $\mathcal{P}_{\Delta t} := \mathcal{P}^0_{\Delta t} \cup \ldots \cup \mathcal{P}^{I-1}_{\Delta t}$. The posterior density, π, is proportional to

$$\underbrace{\pi_0^{(\theta)}(\theta)}_{\text{Prior for } \theta} \overbrace{\pi_0^{(x^{(0)})}(x^{(0)})}^{\text{Prior for } x^{(0)}} \prod_{i=0}^{I-1} \left(\underbrace{g_{i+1}(y^{(i+1)}|x^{(i+1)})}_{\text{Observation density}} \overbrace{\prod_{k=1}^{K_i} f_\theta^{t_{i[k-1]}, t_{i[k]}}(x_k^{(i)}|x_{k-1}^{(i)})}^{\text{Density of path between obs.}} \right) .$$

The transition density for most diffusions is intractable and exact simulation techniques (Beskos, Papaspiliopoulos, and Roberts, 2006) are primarily limited to diffusions which, under a suitable transformation, have unit volatility and, therefore, are typically only applicable to one-dimensional diffusions. Hence, for small $\Delta t > 0$, it is usual to make the following Euler-Maruyama (EM) approximation; $f_\theta^{(t,t+\Delta t)}(x|z) \approx \hat{f}_\theta^{(t,t+\Delta t)}(x|z)$, where we define

$$\hat{f}_\theta^{(t,t+\Delta t)}(x|z) := \phi(x; z + \Delta t \mu(z, t, \theta), \Delta t \zeta(z, t, \theta)) ,$$

with $\phi(x; m, \Psi)$ denoting the density of a Gaussian random variable whose mean and variance are m and Ψ respectively. We consider the corresponding, approximate, posterior, $\hat{\pi}$, which is proportional to

$$\pi_0^{(\theta)}(\theta)\, \pi_0^{(x^{(0)})}(x^{(0)}) \prod_{i=0}^{I-1} g_{i+1}(y^{(i+1)}|x^{(i+1)}) \prod_{k=1}^{K_i} \hat{f}_\theta^{(t_{i_{k-1}}, t_{i_k})}(x_k^{(i)}|x_{k-1}^{(i)}) .$$

This approximation introduces a bias which decreases as Δt decreases. Therefore, a good proposal must be consistent with the diffusion for any small $\Delta t > 0$. Provided care is taken to construct a scheme which does not mix poorly, using, for example, ideas in Golightly and Wilkinson, 2008, inference for this approximate target can proceed via the particle marginal Metropolis-Hastings methodology of Andrieu, Doucet, and Holenstein, 2010. Such a scheme involves iterating over different values of θ and through the observations $y^{(1)}, \ldots, y^{(I)}$. To simplify notation, we henceforth drop θ, and to simplify exposition, and the subsequent simulation study, we fix $x^{(0)}$ and consider only one observation at time T. We emphasise that, from a statistical efficiency point of view, *nothing* is lost in making these simplifications since none of the proposals to be discussed in this thesis depend on more than the subsequent observation, hence any difference in statistical efficiency for one observation will translate into a similar or greater (due to sequential effects) difference in statistical efficiency over many observations. With these simplifications the approximate target is

$$\hat{\pi}(x_{1:K}|y) = \frac{g_1(y|x_K) \prod_{k=1}^{K} \hat{f}^{(t_{k-1}, t_k)}(x_k|x_{k-1})}{\int_{\mathbb{R}^{d \times K}} g_1(y|x_K) \prod_{k=1}^{K} \hat{f}^{(t_{k-1}, t_k)}(x_k|x_{k-1}) \, \mathrm{d}x_{1:K}} ,$$

where, for ease of exposition, we have denoted any variable $\psi_j^{(0)}$ by ψ_j, K_0 by K, t_{0_j} by t_j, and $y^{(1)}$ by y. For inexact observations, which are the focus of this chapter, any importance sampling-based approach requires the sampling of N skeleton paths, denoted by $\{x_{1:K}^{(j)}\}_{j=1}^N$, from a proposal, $q(\cdot|y)$, which is close to $\hat{\pi}$, and the calculation of the importance weights of the form

$$w_j = \hat{\pi}(x_{1:K}^{(j)}|y)/q(x_{1:K}^{(j)}|y) \; . \tag{50}$$

The optimal proposal, $q^{\mathrm{OPT}}(\cdot|y) = \hat{\pi}(\cdot|y)$, results in equal weights and, therefore, zero variance. However, for most diffusions, such a proposal cannot be implemented, thus necessitating the need to construct proposals which aim to mimic the optimal proposal.

3.2.1 *Absolute Continuity of Proposals*

In theory, any *discretized* proposal, such as the ones considered in this thesis, would have the desirable property that the proposal's limiting process, as $\Delta t \downarrow 0$, is absolutely continuous with respect to the true conditioned diffusion. Therefore, decreasing Δt will decrease the bias in the approximate inference scheme without resulting in an ever increasing variance, as measured by the variability in the weights. While it is possible to prove this absolute continuity condition for some proposals (see, for example, Delyon and Hu, 2006; Schauer, Meulen, and Zanten, 2017, and Chapter 4, Papaspiliopoulos and Roberts, 2012), it is beyond the scope of this thesis. Since we are already conducting approximate inference by discretizing the diffusion, the practical issue is not so much about whether absolute continuity holds in theory, but whether the variance of the weights is sufficiently *small* at the level of discretization the practitioner feels appropriate. Therefore, for the schemes we introduce in this thesis, we will instead demonstrate, numerically, the *robustness* of such schemes to a decreasing Δt. The robustness that we find strongly suggests that our proposals do, in fact, have limiting processes which are absolutely continuous with respect to the true conditioned diffusion.

3.2.2 *Forward Simulation*

The forward simulation (FS) approach of Pedersen, 1995 uses the proposal

$$q^{\mathrm{FS}}(x_{1:K}) = \prod_{k=1}^{K} \hat{f}^{(t_{k-1},t_k)}(x_k|x_{k-1}) \; ,$$

which leads to weights of the form $w_j = g(y|x_K^{(j)})$; that is, the weights are simply equal to the likelihood of the observation given the terminal point of the diffusion. Such a proposal produces paths which are consistent with the true diffusion but which can be inconsistent with

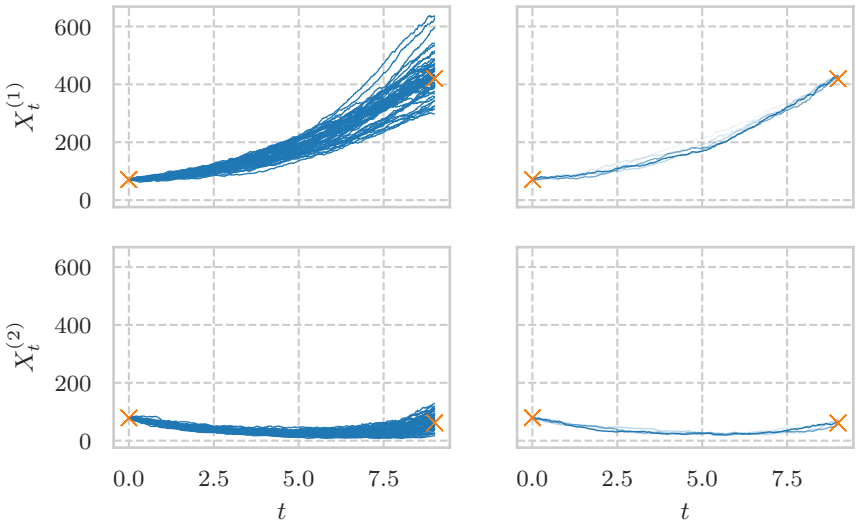

Figure 8: Two sets of plots of fifty paths simulated using the FS approach of Pedersen, 1995 on the Lotka-Volterra SDE introduced in Section 3.1.2. The plots on the left are the fifty two-dimensional simulated paths with no transparency and the plots on the right are the fifty two-dimensional paths with transparency inversely proportional to their normalised weights so that the path with the largest normalised weight has no transparency and paths with smaller normalised weights are more transparent. The two-dimensional initial condition and observation are illustrated with crosses.

the observation since x_K is simulated irrespective of the value of y. Therefore, if the noise in the observation is small, the variability of the weights is likely to be large as only a few of the simulated endpoints, $x_K^{(j)}$, will lie near the observation. This phenomena can be seen in Figure 8 where we have simulated fifty paths from the Lotka-Volterra SDE introduced in subsection 3.1.2 using the FS approach of Pedersen, 1995. For illustration purposes we have weighted each path under the assumption that the noise in the observation is small[1] and have plotted the paths twice; the paths on the left have no transparency, whereas the paths on the right have been plotted with a transparency inversely proportional to their normalised weights, $\tilde{w}_j$, so that the path with the largest normalised weight has no transparency and the paths with smaller normalised weights are more transparent. Thus, if there is large variability in the weights, the number of partially visible paths will be small, whereas, if there is small variability in the weights, the number of partially visible paths will be large. It is clear from the figure that two of the paths have the highest weight with the other paths having almost zero weight. Moreover, as one would expect, those two paths are precisely the paths whose endpoints lie closest to the observation.

1 In particular, for all the figures in this section, we have assumed that $(Y|X_K = x) \sim \mathrm{N}(x, 5I)$.

3.2.3 *The Modified Diffusion Bridge*

The modified diffusion bridge (MDB) of Durham and Gallant, 2002 overcomes the drawback of the FS approach by forming a proposal which depends on the observation y. Specifically, suppose that, at time t_k, we have simulated x_k. Conditional on this point, form the approximate diffusion, X_t^{MDB}, which satisfies, for $t \in [t_k, T]$,

$$\mathrm{d}X_t^{\mathrm{MDB}} = \mu(x_k, t_k)\,\mathrm{d}t + \sigma(x_k, t_k)\,\mathrm{d}W_t\ , \quad X_k^{\mathrm{MDB}} = x_k\ . \tag{51}$$

This approximation is equivalent to assuming that the EM approximation between the current time point and any subsequent time point is exact and leads to the following joint distribution for the approximate process, X_t^{MDB}, at the next point of the partition and at the observation time;

$$\begin{bmatrix} X_{k+1}^{\mathrm{MDB}} \\ X_K^{\mathrm{MDB}} \end{bmatrix} \Bigg| (X_k^{\mathrm{MDB}} = x_k) \sim \mathrm{N}(m_k^{\mathrm{MDB}}, \Psi_k^{\mathrm{MDB}})\ , \tag{52}$$

where

$$\begin{aligned} m_k^{\mathrm{MDB}} &:= \begin{bmatrix} x_k + \Delta t \mu(x_k, t_k) \\ x_k + (T - t_k)\mu(x_k, t_k) \end{bmatrix}\ , \\ \Psi_k^{\mathrm{MDB}} &:= \begin{bmatrix} \Delta t \zeta(x_k, t_k) & \Delta t \zeta(x_k, t_k) \\ \Delta t \zeta(x_k, t_k) & (T - t_k)\zeta(x_k, t_k) \end{bmatrix}\ . \end{aligned}$$

Consequently, the joint distribution for the approximate process at the next point of the partition and the observation, Y, is given by

$$\begin{bmatrix} X_{k+1}^{\mathrm{MDB}} \\ Y \end{bmatrix} \Bigg| (X_k^{\mathrm{MDB}} = x_k) \sim \mathrm{N}(\bar{m}_k^{\mathrm{MDB}}, \bar{\Psi}_k^{\mathrm{MDB}})\ ,$$

where

$$\begin{aligned} \bar{m}_k^{\mathrm{MDB}} &:= \begin{bmatrix} x_k + \Delta t \mu(x_k, t_k) \\ P_1 x_k + (T - t_k) P_1 \mu(x_k, t_k) \end{bmatrix}\ , \\ \bar{\Psi}_k^{\mathrm{MDB}} &:= \begin{bmatrix} \Delta t \zeta(x_k, t_k) & \Delta t \zeta(x_k, t_k) P_1^* \\ \Delta t P_1 \zeta(x_k, t_k) & (T - t_k) P_1 \zeta(x_k, t_k) P_1^* + V_1 \end{bmatrix}\ , \end{aligned}$$

Standard manipulations for the multivariate normal distribution show that

$$(X_{k+1}^{\mathrm{MDB}} | X_k^{\mathrm{MDB}} = x_k, Y = y) \sim \mathrm{N}(a_k^{\mathrm{MDB}}, C_k^{\mathrm{MDB}})\ , \tag{53}$$

where

$$\begin{aligned} a_k^{\mathrm{MDB}} &:= x_k + \Delta t \mu(x_k, t_k) \\ &\quad + \Delta t \zeta(x_k, t_k) P_1^* \Gamma_k (y - P_1 x_k - (T - t_k) P_1 \mu(x_k, t_k))\ , \end{aligned} \tag{54}$$

$$C_k^{\mathrm{MDB}} := \Delta t \zeta(x_k, t_k) - \Delta t^2 \zeta(x_k, t_k) P_1^* \Gamma_k P_1 \zeta(x_k, t_k)\ , \tag{55}$$

and

$$\Gamma_k := ((T - t_k)P_1\zeta(x_k, t_k)P_1^* + V_1)^{-1} .$$

The MDB proposal is therefore given by

$$q^{\text{MDB}}(x_{1:K}|y) = \prod_{k=1}^{K} \phi(x_k; a_{k-1}^{\text{MDB}}, C_{k-1}^{\text{MDB}}) ,$$

where ϕ denotes the density corresponding to a d-dimensional normal distribution and a_{k-1}^{MDB} and C_{k-1}^{MDB} correspond to the mean (Equation (54)) and variance matrix (Equation (55)) respectively, and, implicitly, depend on x_{k-1}, T, y, and t_{k-1}. Thus, the importance weights are given by

$$w_j = g_1(y|x_K^{(j)}) \prod_{k=1}^{K} \frac{\hat{f}^{(t_{k-1},t_k)}(x_k^{(j)}|x_{k-1}^{(j)})}{\phi(x_k^{(j)}; a_{k-1}^{\text{MDB}}, C_{k-1}^{\text{MDB}})} .$$

The approximate process, (51), is equivalent to assuming an EM approximation between the current time point and any subsequent time point, hence paths simulated using the MDB exhibit linear dynamics. Thus, even though paths simulated in this way are consistent with the observation, they are inconsistent with any non-linear dynamics of the true diffusion and, hence, can perform poorly in scenarios where the true diffusion exhibits non-linear dynamics and particularly, therefore, for relatively larger values of T. This behaviour, when compared to Figure 8, can be seen in Figure 9 where we have simulated fifty paths from the Lotka-Volterra SDE introduced in Section 3.1.2 using the MDB of Durham and Gallant, 2002. Again, for illustration purposes, paths have been plotted twice; the paths on the left have no transparency, whereas the paths on the right have transparency inversely proportional to their normalised weights. It can be seen that, even though all of the paths are consistent with the observation, none of the paths are consistent with the dynamics of the true diffusion and hence one of the paths has a much larger weight relative to the other paths.

3.2.4 *The Fearnhead and Lindström Bridges*

While paths simulated via the MDB are consistent with the observation, they are, in general, inconsistent with the dynamics of the true diffusion due to their linear transitions. On the other hand, paths simulated via FS are consistent with the dynamics of the true diffusion, but are inconsistent with the observation. This suggests that one can improve upon both bridges by combining them in such a way that, initially, the path's transitions are mostly the same as the transitions of FS, with little impact from the transitions corresponding to the MDB, but, as the paths get closer to the observation, their transitions become more similar to the transitions corresponding to the MDB. For geometrically ergodic diffusions, Fearnhead, 2008, forms a proposal based on this

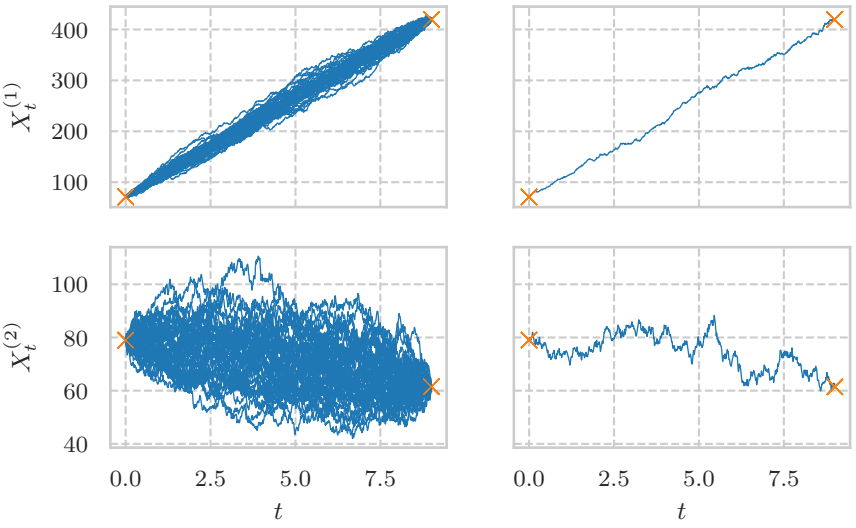

Figure 9: A plot of fifty paths simulated using the MDB of Durham and Gallant, 2002 on the Lotka-Volterra SDE introduced in Section 3.1.2. As in Figure 8, the plots on the left are the fifty two-dimensional simulated paths with no transparency and the plots on the right are the fifty two-dimensional paths with transparency inversely proportional to their normalised weights. The two-dimensional initial condition and observation are illustrated with crosses.

intuition. Indeed, let $f^{(t,T)}(\cdot|x_t)$ be the distribution corresponding to the random variable $X_T|X_t = x_t$. If the diffusion, X_t, is geometrically ergodic, then there exists a stationary distribution, π, a constant, $M > 0$, and a rate of mixing, $\rho > 0$, such that

$$\|f^{(t,T)}(\cdot|x_t) - \pi\| < M \exp[-\rho(T-t)] \, .$$

That is, $f^{(t,T)}$ converges towards π at the rate $\exp[-\rho(T-t)]$. Thus, we can approximate $f^{(t,T)}(\cdot|x_t)$ by

$$(1 - \exp[-\rho(T-t)])\pi(\cdot) + \exp[-\rho(T-t)]f^{(t,T)}(\cdot|x_t) \, . \tag{56}$$

Suppose that, at time t_k, we have simulated x_k. The density corresponding to the random variable $X_{t_{k+1}}|(X_{t_k}, Y) = (x_k, y)$ is, by Bayes' Theorem, proportional to

$$\int_{\mathbb{R}^d} g(y|x_T) f^{(t_{k+1},T)}(x_T|x_{k+1}) f^{(t_k,t_{k+1})}(x_{k+1}|x_k) \, \mathrm{d}x_T \, ,$$

which, by Equation (56), can be approximated by

$$\begin{aligned}
&(1 - \exp[-\rho(T-t_{k+1})]) f^{(t_k,t_{k+1})}(x_{k+1}|x_k) \int_{\mathbb{R}^d} g(y|x_T)\pi(x_T) \, \mathrm{d}x_T \\
&+ \exp[-\rho(T-t_{k+1})] f^{(t_k,t_{k+1})}(x_{k+1}|x_k) \int_{\mathbb{R}^d} g(y|x_T) f^{(t_{k+1},T)}(x_T, x_{k+1}) \, \mathrm{d}x_T \, .
\end{aligned}$$

The density given by

$$\frac{f^{(t_k,t_{k+1})}(x_{k+1}|x_k)\int_{\mathbb{R}^d} g(y|x_T)f^{(t_{k+1},T)}(x_T,x_{k+1})\,\mathrm{d}x_T}{\iint_{\mathbb{R}^{2d}} f^{(t_k,t_{k+1})}(x_{k+1}|x_k)g(y|x_T)f^{(t_{k+1},T)}(x_T,x_{k+1})\,\mathrm{d}x_T\mathrm{d}x_{k+1}},$$

can be approximated by the density corresponding to the MDB, and the density $f^{(t_k,t_{k+1})}(x_{k+1}|x_k)$ can be approximated by a Euler-Maruyama step, $\hat{f}^{(t_k,t_{k+1})}(x_{k+1}|x_k)$. This suggests a proposal of the form

$$\begin{aligned} q^{\text{Fearn}}(x_{k+1}|x_k,y) = (1-\exp[-\alpha(T-t_{k+1})])\beta\hat{f}^{(t_k,t_{k+1})}(x_{k+1}|x_k) \\ + \exp[-\alpha(T-t_{k+1})]q^{\text{MDB}}(x_{k+1}|x_k,x_T)\,, \end{aligned}$$

for some $\alpha > 0$ and $\beta > 0$, which leads to importance weights of the form

$$w_j = g_1(y|x_K^{(j)})\prod_{k=1}^{K}\frac{\hat{f}^{(t_{k-1},t_k)}(x_k^{(j)}|x_{k-1}^{(j)})}{q^{\text{Fearn}}(x_k^{(j)}|x_{k-1}^{(j)},y)}\,.$$

By replacing the variance of $(X_K^{\text{MDB}}|X_k^{\text{MDB}}=x_k)$ with a heuristically motivated approximation to the Mean Squared Error (MSE), Lindström, 2012 attempts to account for the bias in the Euler-Marayama step. Indeed, the MSE for the Euler-Maruyama step for a small time-step, Δt, is given by

$$\text{MSE}(X_{k+1}|X_k=x_k) = \Delta t\zeta(x_k,t_k) + \Delta t^2 D(x_k,t_k) + o(\Delta t)\,,$$

where $D(x_k,t_k)$ is an unknown matrix of size $d\times d$ (see, for example, Kloeden and Platen, 1992). Therefore, an approximation of the MSE from t_k to T is given by

$$\text{MSE}(X_K|X_k=x_k) = (T-t_k)\zeta(x_k,t_k) + (T-t_k)^2 D(x_k,t_k)\,.$$

A heuristic choice for $D(x_k,t_k)$ assumes that D is a fraction of the the variance $\zeta(x_k,t_k)$ over the time interval Δt; that is $D(x_k,t_k) = \gamma\zeta(x_k,t_k)/\Delta t$ for some $\gamma > 0$. For then, under this approximation,

$$\begin{aligned} \text{MSE}(X_K|X_k=x_k) &= (T-t_k)\zeta(x_k,t_k)\left(1+\frac{(T-t_k)\gamma}{\Delta t}\right) \\ &= (T-t_k)\zeta(x_k,t_k)(1+\gamma(K-k))\,. \end{aligned}$$

That is, the approximate MSE is simply the variance inflated, and, this inflation depends on the choice of γ. Here, $\gamma = 0$ means no inflation (that is, no bias), whereas large values of γ mean high inflation (that is, large bias), and on the the relative size of the interval $(T-t_k)$ compared to Δt; the larger the size, the greater the inflation (that is, the more bias there is), whereas, the smaller the size, the smaller the inflation (that is, the less bias there is). This makes sense intuitively. Using this approximation in Equation (52), the Lindström bridge, given

$X_k^{\text{Lind}} = x_k$, has the following joint distribution for the approximate process, X_t^{Lind}, at the next point of the partition and at the observation time;

$$\begin{bmatrix} X_{k+1}^{\text{Lind}} \\ X_K^{\text{Lind}} \end{bmatrix} \Big| (X_k^{\text{Lind}} = x_k) \sim \text{N}(m_k^{\text{Lind}}, \Psi_k^{\text{Lind}}) \ ,$$

where

$$m_k^{\text{Lind}} := \begin{bmatrix} x_k + \Delta t \mu(x_k, t_k) \\ x_k + (T - t_k)\mu(x_k, t_k) \end{bmatrix} \ ,$$

$$\Psi_k^{\text{Lind}} := \begin{bmatrix} \Delta t \zeta(x_k, t_k) & \Delta t \zeta(x_k, t_k) \\ \Delta t \zeta(x_k, t_k) & (T - t_k)\zeta(x_k, t_k)(1 + \gamma(K - k)) \end{bmatrix} \ .$$

Therefore, following the same manipulations as Section 3.2.3,

$$(X_{k+1}^{\text{Lind}} | X_k^{\text{Lind}} = x_k, Y = y) \sim \text{N}(a_k^{\text{Lind}}, C_k^{\text{Lind}}) \ , \tag{57}$$

where

$$\begin{aligned} a_k^{\text{Lind}} := x_k &+ \Delta t \mu(x_k, t_k) \\ &+ \Delta t \zeta(x_k, t_k) P_1^* \Gamma_k (y - P_1 x_k - (T - t_k) P_1 \mu(x_k, t_k)) \ , \end{aligned} \tag{58}$$

$$C_k^{\text{Lind}} := \Delta t \zeta(x_k, t_k) - \Delta t^2 \zeta(x_k, t_k) P_1^* \Gamma_k P_1 \zeta(x_k, t_k) \ , \tag{59}$$

and

$$\Gamma_k := \{P_1[(T - t_k)\zeta(x_k, t_k)(1 + \gamma(K - k))]P_1^* + V_1\}^{-1} \ .$$

The Lindström proposal is therefore given by

$$q^{\text{Lind}}(x_{1:K}|y) = \prod_{k=1}^{K} \phi(x_k; a_{k-1}^{\text{Lind}}, C_{k-1}^{\text{Lind}}) \ ,$$

where ϕ denotes the density corresponding to a d-dimensional normal distribution and a_{k-1}^{Lind} and C_{k-1}^{Lind} correspond to the mean (Equation (58)) and variance matrix (Equation (59)) respectively, and, implicitly, depend on x_{k-1}, T, y, and t_{k-1}. Thus, the importance weights are given by

$$w_j = g_1(y|x_K^{(j)}) \prod_{k=1}^{K} \frac{\hat{f}^{(t_{k-1}, t_k)}(x_k^{(j)}|x_{k-1}^{(j)})}{\phi(x_k^{(j)}; a_{k-1}^{\text{Lind}}, C_{k-1}^{\text{Lind}})} \ .$$

If $\gamma = 0$, then we have the MDB of Durham and Gallant, 2002 (Section 3.2.3). However, taking $\gamma \uparrow \infty$, gives $\Gamma_k \downarrow 0$, and

$$\begin{aligned} a_k^{\text{Lind}} &\to x_k + \Delta t \mu(x_k, t_k) \ , \\ C_k^{\text{Lind}} &\to \Delta t \zeta(x_k, t_k) \ . \end{aligned}$$

That is, we have the forward simulation approach of Pedersen, 1995 (Section 3.2.2). Therefore, like the Fearnhead bridge, the Linström bridge is a mixture between forward simulation and the Modified Diffusion Bridge.

While both the Fearnhead and Lindström bridges are appealing intuitively, and have improved performance over the forward simulation approach and the MDB, see Fearnhead, 2008, Lindström, 2012, and Whitaker et al., 2017, they both require tuning of hyperparameters. This tuning will be sensitive to; the diffusion for which the bridge is being constructed, the observation, y, being conditioned upon, and any parameters driving the diffusion. This makes them difficult to use in practice.

3.2.5 *Bridges Based on Residual Processes*

Recall that, while paths simulated via the MDB are consistent with the observation, they are, in general, inconsistent with the dynamics of the true diffusion due to their linear transitions. Whitaker et al., 2017, introduce residual-bridge proposals which deal with this issue, albeit at a greater computational cost, by constructing a deterministic path, ξ_t, which captures the non-linear dynamics of the true, conditioned diffusion, and considering the residual process, $R_t := X_t - \xi_t$, which satisfies, for $t \in [0, T]$,

$$\mathrm{d}R_t = (\mu(X_t, t) - \xi_t')\mathrm{d}t + \sigma(X_t, t)\mathrm{d}W_t \ , \quad R_0 = 0 \ .$$

If ξ_t accurately captures the non-linear dynamics of the true, conditioned diffusion, then the residual should exhibit behaviour which is more linear, hence applying the MDB to the residual and adding back ξ_t will result in a proposal which more closely resembles the optimal proposal. Suppose, then, that, at time t_k, we have simulated x_k. Applying the MDB to the residual, R_t, gives the following joint distribution for the approximate residual process, R_t^{RB}, at the next point of the partition and at the observation time;

$$\begin{bmatrix} R_{k+1}^{\mathrm{RB}} \\ R_K^{\mathrm{RB}} \end{bmatrix} \Bigg| (X_k^{\mathrm{RB}} = x_k) \sim \mathrm{N}(\gamma_k^{\mathrm{RB}}, C_k^{\mathrm{RB}}) \ ,$$

where

$$\begin{aligned} \gamma_k^{\mathrm{RB}} &:= \begin{bmatrix} (x_k - \xi_k) + \Delta t(\mu(x_k, t_k) - (\xi_{k+1} - \xi_k)/\Delta t) \\ (x_k - \xi_k) + (T - t_k)(\mu(x_k, t_k) - (\xi_{k+1} - \xi_k)/\Delta t) \end{bmatrix} \\ &= \begin{bmatrix} (x_k - \xi_{k+1}) + \Delta t\mu(x_k, t_k) \\ (x_k - \xi_k) + (T - t_k)(\mu(x_k, t_k) - (\xi_{k+1} - \xi_k)/\Delta t) \end{bmatrix} \ , \\ C_k^{\mathrm{RB}} &:= \begin{bmatrix} \Delta t\zeta(x_k, t_k) & \Delta t\zeta(x_k, t_k) \\ \Delta t\zeta(x_k, t_k) & (T - t_k)\zeta(x_k, t_k) \end{bmatrix} \ . \end{aligned}$$

Here, $X_t^{\mathrm{RB}} := R_t^{\mathrm{RB}} + \xi_t$ denotes the process which approximates the true process and, as in Whitaker et al., 2017, we have approximated ξ_k'

via the chord between (t_k, ξ_k) and (t_{k+1}, ξ_{k+1}). Adding back ξ_t leads to the following joint distribution for the approximate process, X_t^{RB}, at the next point of the partition and at the observation time;

$$\begin{bmatrix} X_{k+1}^{\mathrm{RB}} \\ X_K^{\mathrm{RB}} \end{bmatrix} \Bigg| (X_k^{\mathrm{RB}} = x_k) \sim \mathrm{N}\Bigg(\begin{bmatrix} m_{k+1}^{\mathrm{RB}} \\ m_K^{\mathrm{RB}} \end{bmatrix}, \Psi_k^{\mathrm{RB}} \Bigg) ,$$

where

$$\begin{aligned} m_{k+1}^{\mathrm{RB}} &:= x_k + \Delta t \mu(x_k, t_k) , \\ m_K^{\mathrm{RB}} &:= x_k + (\xi_K - \xi_k) + (T - t_k)(\mu(x_k, t_k) - (\xi_{k+1} - \xi_k)/\Delta t) , \\ \Psi_k^{\mathrm{RB}} &:= \begin{bmatrix} \Delta t \zeta(x_k, t_k) & \Delta t \zeta(x_k, t_k) \\ \Delta t \zeta(x_k, t_k) & (T - t_k)\zeta(x_k, t_k) \end{bmatrix} . \end{aligned}$$

Therefore, the joint distribution for the approximate process at the next point of the partition and the observation, Y, is given by

$$\begin{bmatrix} X_{k+1}^{\mathrm{RB}} \\ Y \end{bmatrix} \Bigg| (X_k^{\mathrm{RB}} = x_k) \sim \mathrm{N}\Bigg(\begin{bmatrix} \bar{m}_{k+1}^{\mathrm{RB}} \\ \bar{m}_K^{\mathrm{RB}} \end{bmatrix}, \bar{\Psi}_k^{\mathrm{RB}} \Bigg) ,$$

where

$$\begin{aligned} \bar{m}_{k+1}^{\mathrm{RB}} &:= x_k + \Delta t \mu(x_k, t_k) , \\ \bar{m}_K^{\mathrm{RB}} &:= P_1 x_k + P_1(\xi_K - \xi_k) + (T - t_k) P_1(\mu(x_k, t_k) - (\xi_{k+1} - \xi_k)/\Delta t) , \\ \bar{\Psi}_k^{\mathrm{RB}} &:= \begin{bmatrix} \Delta t \zeta(x_k, t_k) & \Delta t \zeta(x_k, t_k) P_1^* \\ \Delta t P_1 \zeta(x_k, t_k) & (T - t_k) P_1 \zeta(x_k, t_k) P_1^* + V_1 \end{bmatrix} . \end{aligned}$$

Standard manipulations for the multivariate normal distribution show that

$$(X_{k+1}^{\mathrm{RB}} | X_k^{\mathrm{RB}} = x_k, Y = y) \sim \mathrm{N}(a_k^{\mathrm{RB}}, D_k^{\mathrm{RB}}) , \tag{60}$$

where

$$a_k^{\mathrm{RB}} := x_k + \Delta t \mu(x_k, t_k) + \Delta t \zeta(x_k, t_k) P_1^* \Gamma_k E_k \tag{61}$$

$$D_k^{\mathrm{RB}} := \Delta t \zeta(x_k, t_k) - \Delta t^2 \zeta(x_k, t_k) P_1^* \Gamma_k P_1 \zeta(x_k, t_k) , \tag{62}$$

and

$$\begin{aligned} \Gamma_k &:= ((T - t_k) P_1 \zeta(x_k, t_k) P_1^* + V_1)^{-1} , \\ E_k &:= y - P_1 x_k - P_1(\xi_K - \xi_k) \\ &\quad - (T - t_k) P_1(\mu(x_k, t_k) - (\xi_{k+1} - \xi_k)/\Delta t) . \end{aligned}$$

The residual-bridge proposal is therefore given by

$$q^{\mathrm{RB}}(x_{1:K}|y) = \prod_{k=1}^{K} \phi(x_k; a_{k-1}^{\mathrm{RB}}, D_{k-1}^{\mathrm{RB}}) ,$$

where, as previously, ϕ denotes the density corresponding to a d-dimensional normal distribution and a_{k-1}^{RB} and D_{k-1}^{RB} correspond to the mean (Equation (61)) and variance matrix (Equation (62)) respectively, and, implicitly, depend on x_{k-1}, T, y, ξ_{k-1}, ξ_K, and t_{k-1}. Thus, the importance weights are given by

$$w_j = g_1(y|x_K^{(j)}) \prod_{k=1}^{K} \frac{\hat{f}^{(t_{k-1},t_k)}(x_k^{(j)}|x_{k-1}^{(j)})}{\phi(x_k^{(j)}; a_{k-1}^{\mathrm{RB}}, D_{k-1}^{\mathrm{RB}})} \ .$$

The performance of such a proposal clearly hinges on choosing a deterministic path ξ_t which has similar dynamics to the true diffusion. One natural candidate (justified, for diffusions relating to the chemical Langevin equation, by Theorem 2.1 in Chapter 11 of Ethier and Kurtz, 1986) for ξ_t is constructed by ignoring the volatility in the true diffusion. That is, if we let $\xi_t \equiv \eta_t$ be the path obtained by ignoring any stochasticity in the evolution of the diffusion, then, from (48) we have that η_t satisfies

$$\eta_{t+\Delta t} = \eta_t + \Delta t \mu(\eta_t, t) + o(\Delta t) \ ,$$

for any $[t, t+\Delta t] \subset [0, T)$. Therefore, η_t solves the ordinary differential equation (ODE)

$$\frac{\mathrm{d}\eta_t}{\mathrm{d}t} = \mu(\eta_t, t) \ , \quad \eta_0 = x_0 \ , \tag{63}$$

over $[0, T]$. We denote the residual-bridge with this choice of ξ_t by $\mathrm{RB}^{\mathrm{ODE}}$. This choice for ξ_t is independent of the observation and hence can fail to capture the true dynamics of the *conditioned* diffusion, particularly when the noise in the observation, V_1, is small, and the difference between the observation, y, and the endpoint of the deterministic path, η_K, is large. Therefore, paths simulated using this proposal can be inconsistent with the *conditioned* diffusion when the inter-observation time, T, is relatively large, since, for larger T, the stochasticity in the SDE results in dynamics which are inconsistent with η_t. As suggested by Whitaker et al., 2017, this motivates constructing a path ξ_t which is consistent with the *conditioned* diffusion by approximating the residual R_t with a tractable process, $\hat{R}_t$, and choosing

$$\xi_t = \eta_t + \mathbb{E}(\hat{R}_t | Y = y) \ .$$

One choice (justified, for diffusions relating to the chemical Langevin equation, by Theorem 2.3 in Chapter 11 of Ethier and Kurtz, 1986) for the tractable process, $\hat{R}_t$, is that given by the linear noise approximation (LNA, see, for example, Komorowski et al., 2009; Fearnhead, Giagos, and Sherlock, 2014). By Taylor expanding around η_t, defined by (63), the LNA constructs an $\hat{R}_t$ which satisfies a linear SDE, and, therefore, has Gaussian transition densities. Indeed, by taking a first-order Taylor expansion of the drift and a zeroth-order Taylor expansion of the square-root of the volatility, one arrives at an approximate process $\hat{R}_t$ which satisfies

$$\mathrm{d}\hat{R}_t = J(\eta_t, t)\hat{R}_t \, \mathrm{d}t + \sigma(\eta_t, t) \, \mathrm{d}W_t \ , \quad \hat{R}_0 = 0 \ , \tag{64}$$

over the interval $[0, T]$, where $J(\eta_t, t)$ is the $d \times d$ Jacobian matrix whose (i, j)-th entry is

$$J(\eta_t, t)_{ij} := \left. \frac{\partial \mu(x, t)_i}{\partial x_j} \right|_{x=\eta_t} .$$

Under this approximation, a tractable form for $\mathbb{E}(\hat{R}_t | Y = y)$ is available. The following lemma (see, for instance, Whitaker et al., 2017) derives a form which can be implemented in a computationally efficient manner because the ODEs that need to be solved do not involve any matrix inverses. Moreover, the ODEs only need to be solved once for a given path η_t, irrespective of the simulated path x_t.

LEMMA 3.2.1. *Let $\hat{R}_t$ be the process which satisfies (64) over the interval $[0, T]$ and let Y be such that*

$$(Y | \hat{R}_T = r) \sim N(P(r + \eta_T), V) .$$

Then

$$\mathbb{E}(\hat{R}_t | Y = y) = \phi_t G_t^{-1} G_T^* P^* (P \phi_T P^* + V)^{-1} (y - P \eta_T) ,$$

where G_t and ϕ_t satisfy, for $t \in [0, T]$, the following ODEs;

$$\frac{\mathrm{d}G_t}{\mathrm{d}t} = J(\eta_t, t) G_t , \qquad G_0 = I ,$$
$$\frac{\mathrm{d}\phi_t}{\mathrm{d}t} = J(\eta_t, t)\phi_t + \phi_t J(\eta_t, t)^* + \zeta(\eta_t, t) , \quad \phi_0 = 0 .$$

Proof. See, for example, Whitaker et al., 2017, or A.8. □

For most diffusions, G_t, and ϕ_t, will not be available analytically. However, using the Fortran subroutine `lsoda` (introduced by Petzold, 1983), both can be numerically evaluated in an accurate and efficient way at any point of the partition $\mathcal{P}^0_{\Delta t}$. We denote the residual-bridge proposal with this choice of ξ_t by $\mathrm{RB}^{\mathrm{LNA}}$. Fifty paths simulated from the Lotka-Volterra SDE, introduced in Section 3.1.2, using the $\mathrm{RB}^{\mathrm{ODE}}$ and $\mathrm{RB}^{\mathrm{LNA}}$ proposals, along with the corresponding deterministic paths, can be seen in Figures 10 and 11 respectively. As before, in both figures, the paths have been plotted twice; the paths on the left of each figure have no transparency, whereas the paths on the right of each figure have transparency inversely proportional to their normalised weights. In comparison with Figure 9, it can be seen that the paths in Figure 10 are more consistent with the true diffusion, while still being consistent with the observation. Thus, the variability in the weights is smaller. Morever, the paths in Figure 11 are more consistent with the *conditioned* diffusion, since, as can be seen by comparing with the dashed line in Figure 10, the deterministic path, ξ_t, better represents paths of the conditioned diffusion. Therefore, the variability in the weights is smaller. Although such approaches account for the non-linear dynamics of the diffusion, they still assume a constant volatility over the region of interest. This leads to poor performance for diffusions whose volatility varies greatly over this interval and, in particular, therefore, for larger inter-observation times T, or, for diffusions whose volatility is time-inhomogeneous.

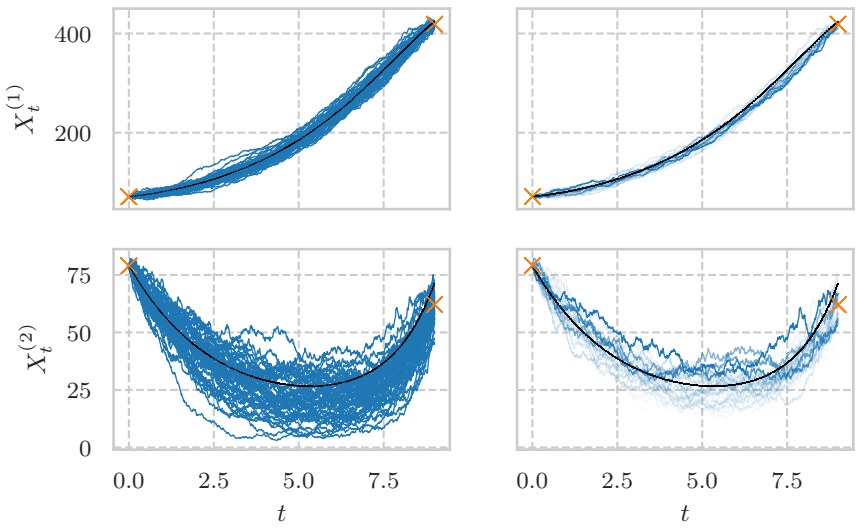

Figure 10: A plot of fifty paths simulated using the RB^{ODE} approach of Whitaker et al., 2017 from the Lotka-Volterra SDE introduced in Section 3.1.2. As with the previous figures, the plots on the left are the fifty two-dimensional simulated paths with no transparency, and the plots on the right are the fifty two-dimensional paths with transparency inversely proportional to their normalised weights. The two-dimensional initial condition and observation are illustrated with crosses, and the deterministic path, $\xi_t = \eta_t$, is plotted with a dashed line.

3.2.6 *Bridges Based on Guided Proposals*

The proposals introduced so far in this thesis, along with the novel proposal that is to be introduced in the subsequent section, are all based on approximating the unconditioned diffusion by a diffusion whose conditioned counterpart is tractable. However, other approaches considered in the literature, see, for example, Delyon and Hu, 2006, Papaspiliopoulos and Roberts, 2012, and Schauer, Meulen, and Zanten, 2017, are based upon directly approximating the true, conditioned diffusion. Indeed, consider the d-dimensional diffusion, X_t, satisfying

$$\mathrm{d}X_t = \mu(X_t, t)\,\mathrm{d}t + \sigma(X_t, t)\,\mathrm{d}W_t\,.$$

The conditioned process $\tilde{X}_t := (X_t | Y = y)$ is also a diffusion, and satisfies the SDE

$$\mathrm{d}\tilde{X}_t = [\mu(\tilde{X}_t, t) + \zeta(\tilde{X}_t, t)\nabla_{x_t} \log \pi_t(y|x_t)|_{x_t=\tilde{X}_t}]\,\mathrm{d}t + \sigma(\tilde{X}_t, t)\,\mathrm{d}W_t\,, \quad (65)$$

where $\zeta = \sigma\sigma^*$ and

$$\pi_t(y|x_t) := \int_{\mathbb{R}^d} g(y|x_T) f^{(t,T)}(x_T|x_t)\,\mathrm{d}x_T\,.$$

At any time, t, the likelihood of the observation given the current point of the process, under the dynamics of the unconditioned diffusion; that

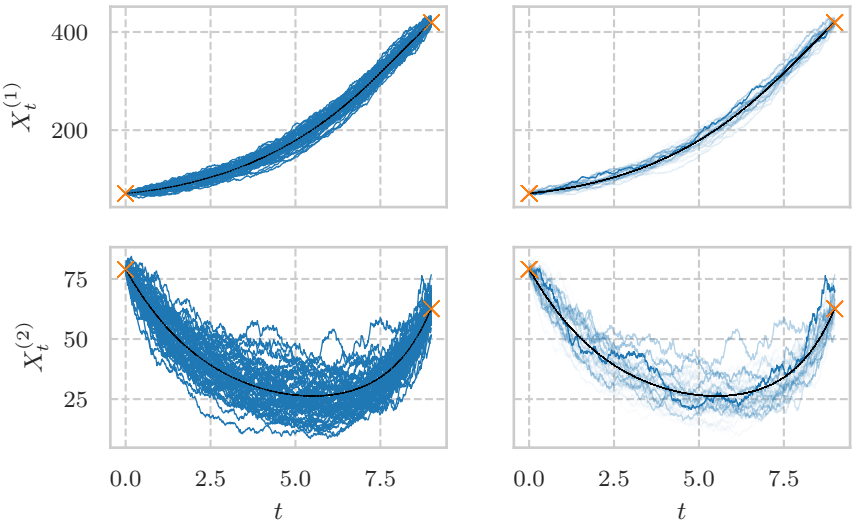

Figure 11: A plot of fifty paths simulated using the RB^{LNA} approach of Whitaker et al., 2017 from the Lotka-Volterra SDE introduced in Section 3.1.2. As with the previous figures, the plots on the left are the fifty two-dimensional simulated paths with no transparency, and the plots on the right are the fifty two-dimensional paths with transparency inversely proportional to their normalised weights. The two-dimensional initial condition and observation are illustrated with crosses, and the deterministic path, $\xi_t = \eta_t + \mathbb{E}(\hat{R}_t | Y = y)$, is plotted with a dashed line.

is, $\pi_t(y|x_t)$ is, in general, intractable. However, given a tractable approximation to this density, one can discretise (65) using, for instance, an EM-step. Given $(Y|X_T = x_T) \sim \text{N}(Px_T, V)$, a natural tractable approximation to $\pi_t(y|x_t)$, denoted $\tilde{\pi}_t(y|x_t)$, arises from approximating the true diffusion, X_s, over the interval $[0, T-t]$ and with *initial condition* $X_0 = x_t$ by a diffusion, $\tilde{X}_s$, satisfying a linear SDE:

$$\mathrm{d}\tilde{X}_s = [a(s) + b(s)\tilde{X}_s]\,\mathrm{d}s + \tilde{\sigma}(s)\,\mathrm{d}W_s\,, \quad \tilde{X}_0 = x_t \tag{66}$$

where $a : [0,\infty) \to \mathbb{R}^d$, $b : [0,\infty) \to \mathbb{R}^d$, $\tilde{\sigma} : [0,\infty) \to \mathbb{R}^{d\times r}$, and W_s is an r-dimensional Wiener process. Indeed, Lemma 3.2.2 which is a slight extension of Lemma 3.2.1, shows that, when $\tilde{X}_s$ satisfies a linear SDE, then $(Y|\tilde{X}_0 = x_t)$ is normally distributed with a mean and variance whose form is tractable:

LEMMA 3.2.2. *Let $\tilde{X}_s$ be a d-dimensional diffusion satisfying (66) over the interval $[0, T-t]$, and suppose that $(Y|\tilde{X}_{T-t} = \tilde{x}_{T-t}) \sim \text{N}(P\tilde{x}_{T-t}, V)$ for some $P \in \mathbb{R}^{r\times d}$ and some $V \in \mathbb{R}^{d\times d}$. Then*

$$(Y|\tilde{X}_0 = x_t) \sim \text{N}(P\eta_{T-t}, P\phi_{T-t}P^* + V)\,.$$

where η_s and ϕ_s satisfy, for $s \in [0, T-t]$, the following ODEs:

$$\frac{\mathrm{d}\eta_s}{\mathrm{d}s} = a(s) + b(s)\eta_s\,, \qquad \eta_0 = x_t\,,$$
$$\frac{\mathrm{d}\phi_s}{\mathrm{d}s} = b(s)\phi_s + \phi_s b(s)^* + \tilde{\sigma}(s)\tilde{\sigma}(s)^*\,, \quad \phi_0 = 0\,.$$

Proof. See, for example, Fearnhead, Giagos, and Sherlock, 2014, or A.9. □

Schauer, Meulen, and Zanten, 2017 demonstrate that the diffusion X_t^* satisfying the SDE

$$\mathrm{d}X_t^* = [\mu(X_t^*, t) + \zeta(X_t^*, t)\nabla_{x_t} \log \hat{\pi}_t(y|x_t)|_{x_t = X_t^*}] \, \mathrm{d}t + \sigma(X_t^*, t) \, \mathrm{d}W_t \, ;$$

that is, the SDE given by (65) with π_t replaced by the approximation $\hat{\pi}_t$, is absolutely continuous with respect to the true, conditioned diffusion, $\tilde{X}_t$. There are, of course, some natural choices for the linear SDE, (66), used to approximate the true SDE. For instance, one could take $b(s) \equiv 0$, $a(s) \equiv \mu(x_t, t)$, and $\sigma(s) \equiv \sigma(x_t, t)$, which, for exact observations; that is, P equal to the identity matrix and V equal to the zero matrix, leads to a proposal which is very similar to the MDB of Section 3.2.3 (see, for instance, Whitaker et al., 2017). Or, one could take the LNA as the approximating *linear* diffusion (see Whitaker et al., 2017). However, Whitaker et al., 2017 show that guided proposals which are statistically efficient (such as those which use the LNA for the approximating conditioned process) are generally too computationally expensive for their overall efficiency (statistical effiency per unit of time) to be competitively small.

3.3 NEW BRIDGES BASED ON RESIDUAL PROCESSES

We propose an extension to the approach of Whitaker et al., 2017, by constructing a *process*, U_t, which exhibits similar dynamics to the true, conditioned diffusion, and considering the residual process, $\tilde{R}_t := X_t - U_t$. We begin by constructing a deterministic path, ξ_t, which exhibits similar dynamics to the true diffusion (for instance, the path on which $\mathrm{RB}^{\mathrm{ODE}}$ or $\mathrm{RB}^{\mathrm{LNA}}$ is based). We then use this path to construct U_t which is coupled with the true diffusion through *the same driving Brownian motion* in such a way that paths of U_t exhibit similar stochastic behaviour to paths of X_t. Specifically, for an arbitrary u_0, we define U_t to be the process which satisfies

$$\mathrm{d}U_t = \xi_t' \mathrm{d}t + \sigma(\xi_t, t)\mathrm{d}B_t \, , \quad U_0 = u_0$$

over the interval $[0, T]$ and which is coupled with X_t through the same driving Brownian motion, B_t. The residual process, $\tilde{R}_t$, thus satisfies

$$\mathrm{d}\tilde{R}_t = (\mu(X_t, t) - \xi_t')\mathrm{d}t + (\sigma(X_t, t) - \sigma(\xi_t, t))\mathrm{d}B_t \, ,$$

over the interval $[0, T]$, and with initial condition $\tilde{R}_0 = x_0 - u_0$. We proceed by making the same approximation used in the MDB: suppose that we have simulated x_k at time t_k. Form an approximate process, $\tilde{R}_t^{\mathrm{MDB}}$, which satisfies

$$\mathrm{d}\tilde{R}_t^{\mathrm{MDB}} = (\mu(x_k, t_k) - \xi_{t_k}') \, \mathrm{d}t + (\sigma(x_k, t_k) - \sigma(\xi_k, t_k)) \, \mathrm{d}B_t \, ,$$

over the interval $[t_k, T]$, and has initial condition $\tilde{R}_k^{\mathrm{MDB}} = x_k - u_k$ (where, as we shall see, u_k is the superfluous value of the process U_t at

time t_k). With this approximation, we have that, conditional on having simulated x_k at time t_k, the process $X_t^{\overline{\mathrm{RB}}} := U_t + \tilde{R}_t^{\mathrm{MDB}}$ satisfies

$$\mathrm{d}X_t^{\overline{\mathrm{RB}}} = (\xi_t' + \mu(x_k, t_k) - \xi_{t_k}')\,\mathrm{d}t + (\sigma(\xi_t, t) + \sigma(x_k, t_k) - \sigma(\xi_k, t_k))\,\mathrm{d}B_t\ ,$$

over the interval $[t_k, T]$, and has initial condition $X_k^{\overline{\mathrm{RB}}} = x_k$. Approximating σ by a piecewise constant function on the partition $\mathcal{P}_{\Delta t}$,

$$\sigma(\xi_u, u) = \sum_{k=0}^{K-1} \sigma(\xi_{t_k}, t_k) \mathbb{1}_{[t_k, t_{k+1})}(u)\ ,$$

gives the following joint distribution for the approximate process, $X^{\overline{\mathrm{RB}}}$, at the next point of the partition and at the observation time:

$$\begin{bmatrix} X_{k+1}^{\overline{\mathrm{RB}}} \\ X_K^{\overline{\mathrm{RB}}} \end{bmatrix} \Bigg| (X_k^{\overline{\mathrm{RB}}} = x_k) \sim \mathrm{N}\left(\begin{bmatrix} m_{k+1}^{\overline{\mathrm{RB}}} \\ m_K^{\overline{\mathrm{RB}}} \end{bmatrix}, \Psi_k^{\overline{\mathrm{RB}}} \right)\ ,$$

where

$$\begin{aligned}
m_{k+1}^{\overline{\mathrm{RB}}} &:= x_k + \Delta t \mu(x_k, t_k)\ , \\
m_K^{\overline{\mathrm{RB}}} &:= x_k + (\xi_K - \xi_k) + (T - t_k)(\mu(x_k, t_k) - (\xi_{k+1} - \xi_k)/\Delta t)\ , \\
\Psi_k^{\overline{\mathrm{RB}}} &:= \begin{bmatrix} \Delta t \zeta(x_k, t_k) & \Delta t \zeta(x_k, t_k) \\ \Delta t \zeta(x_k, t_k) & \Phi_{k,K}^{\overline{\mathrm{RB}}} \end{bmatrix}\ , \\
\Phi_{k,K}^{\overline{\mathrm{RB}}} &:= \Delta t \zeta(x_k, t_k) + \Delta t \sum_{j=k+1}^{K-1} \varphi_{jk} \varphi_{jk}^*\ ,
\end{aligned}$$

and

$$\varphi_{jk} := \sigma(\xi_j, t_j) + \sigma(x_k, t_k) - \sigma(\xi_k, t_k)\ .$$

Thus, using (60), we see that

$$(X_{k+1}^{\overline{\mathrm{RB}}} | X_k^{\overline{\mathrm{RB}}} = x_k, Y_1 = y_1) \sim \mathrm{N}(a_k^{\overline{\mathrm{RB}}}, C_k^{\overline{\mathrm{RB}}})\ , \tag{67}$$

where

$$a_k^{\overline{\mathrm{RB}}} := x_k + \Delta t \mu(x_k, t_k) + \Delta t \zeta(x_k, t_k) P_1^* \Gamma_k D_k \tag{68}$$

$$C_k^{\overline{\mathrm{RB}}} := \Delta t \zeta(x_k, t_k) - \Delta t^2 \zeta(x_k, t_k) P_1^* \Gamma_k P_1 \zeta(x_k, t_k)\ , \tag{69}$$

and

$$\begin{aligned}
\Gamma_k &:= (P_1 \Phi_{k,K}^{\overline{\mathrm{RB}}} P_1^* + V_1)^{-1} \\
D_k &:= y_1 - P_1 x_k - P_1(\xi_K - \xi_k) \\
&\quad - (T - t_k) P_1 (\mu(x_k, t_k) - (\xi_{k+1} - \xi_k)/\Delta t)\ .
\end{aligned}$$

This extended residual-bridge proposal is therefore given by

$$q^{\mathrm{RB}}(x_{1:K} | y) = \prod_{k=1}^{K} \phi(x_k; a_{k-1}^{\overline{\mathrm{RB}}}, C_{k-1}^{\overline{\mathrm{RB}}})\ ,$$

where, as previously, ϕ denotes the density corresponding to a d-dimensional normal distribution and $a_{k-1}^{\overline{\text{RB}}}$ and $C_{k-1}^{\overline{\text{RB}}}$ correspond to the mean (Equation (68)) and variance matrix (Equation (69)) respectively, and, implicitly, depend on x_{k-1}, T, y, $\xi_{k-1:K}$, and t_{k-1}. Thus, the importance weights are given by

$$w_j = g_1(y|x_K^{(j)}) \prod_{k=1}^{K} \frac{\hat{f}^{(t_{k-1},t_k)}(x_k^{(j)}|x_{k-1}^{(j)})}{\phi(x_k^{(j)}; a_{k-1}^{\overline{\text{RB}}}, V_{k-1}^{\overline{\text{RB}}})} \ .$$

As with the residual bridge constructs of Whitaker et al., 2017, this proposal attempts to take into account the variability of the drift. However, unlike the constructs of Whitaker et al., 2017, this proposal also tries to take into account the variability of the square-root of the volatility. Therefore, such a proposal should outperform the residual-bridge constructs of Whitaker et al., 2017 in scenarios where the square-root of the volatility exhibits large variation over the interval $[0, T]$ and, thus, in particular, for relatively larger T, or, for volatilities which are time-inhomogeneous. A trade-off arises since if the square-root of the volatility varies too much then, in many cases of interest, constructing a deterministic path, ξ_t, which accurately captures the true dynamics of the diffusion will be tricky, if not impossible. To illustrate why this new residual-bridge construct might be preferred over the residual-bridge construct of Whitaker et al., 2017, consider constructing bridges to the SDE

$$\mathrm{d}X_t = \mu(t)\mathrm{d}t + \sigma(t)\mathrm{d}B_t \ , \quad X_0 = x_0 \ ,$$

over the interval $[0, T]$. It is clear that if one chooses ξ_t to be the solution of the ODE

$$\frac{\mathrm{d}\xi_t}{\mathrm{d}t} = \mu(t) \ , \quad \xi_0 = x_0 \ ,$$

then, for *any* $\sigma(t)$, this new proposal will, up to a discretisation error, simulate exact bridges of X_t, whereas, the proposal of Whitaker et al., 2017, will not. Moreover, the variability in the weights corresponding to the residual-bridge proposals of Whitaker et al., 2017, will increase the more $\sigma(t)$ varies over the region of interest.

As with the residual-bridge construct of Whitaker et al., 2017, ξ_t can be any deterministic path whose dynamics closely match those of the true conditioned diffusion. We denote this new proposal, where $\xi_t = \eta_t$ with η_t defined by (63), by $\overline{\text{RB}}^{\text{ODE}}$ and, where $\xi_t = \eta_t + \mathbb{E}(\hat{R}_t|Y_1 = y_1)$ with $\hat{R}_t$ defined by (64), by $\overline{\text{RB}}^{\text{LNA}}$. Paths simulated using this proposal look very similar to paths simulated using the residual bridge proposals of Whitaker et al., 2017, as can be seen by comparing Figures 10 and 11, with Figures 12 and 13 which show fifty paths simulated from the Lotka-Volterra SDE introduced in Section 3.1.2 using the $\overline{\text{RB}}^{\text{ODE}}$ and $\overline{\text{RB}}^{\text{LNA}}$ approaches, respectively, along with the corresponding deterministic paths, ξ_t. As throughout this chapter, in both figures, the paths have been plotted twice; the paths on the left of each figure have

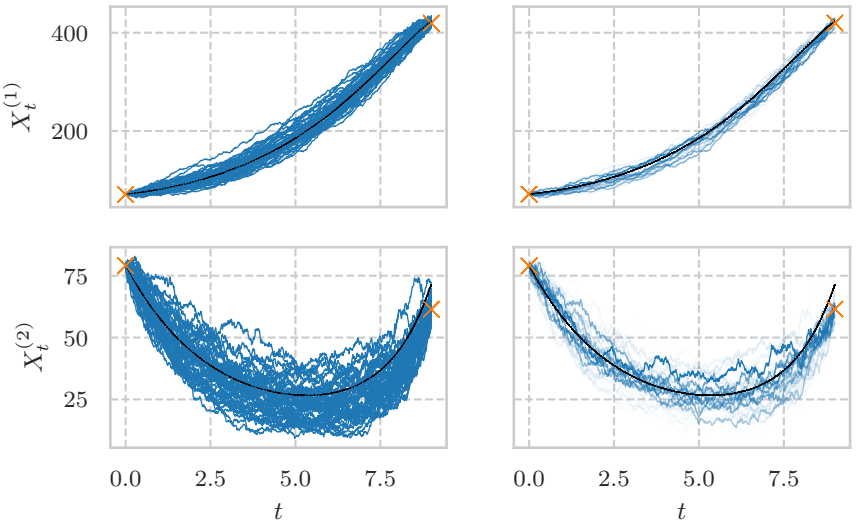

Figure 12: A plot of fifty paths simulated using the $\overline{\text{RB}}^{\text{ODE}}$ approach introduced in this chapter from the Lotka-Volterra SDE introduced in Section 3.1.2. As with the previous figures the plots on the left are the fifty two-dimensional simulated paths with no transparency and the plots on the right are the fifty two-dimensional paths with transparency inversely proportional to their normalised weights. The two-dimensional initial condition and observation are illustrated with crosses and the deterministic path, $\xi_t = \eta_t$, is plotted with a dashed line.

no transparency, whereas the paths on the right of each figure have transparency inversely proportional to their normalised weights. When compared with Figure 10 the paths in Figure 12 are more consistent with the true, *conditioned* diffusion. Thus, the variability in the weights is smaller. Similarly, when compared with Figure 11 the paths in Figure 13 are more consistent with the true, *conditioned* diffusion. Thus, once again, the variability in the weights is smaller.

3.3.1 *Computational Considerations*

Comparing the form of $\Psi_k^{\overline{\text{RB}}}$ with the form of Ψ_k^{RB}, it can be seen that the residual-bridge proposals introduced in this chapter have a larger computational cost compared to the corresponding residual-bridge proposals of Whitaker et al., 2017. Indeed, at any iteration $k \in \{0, \ldots, K-1\}$, we have $K - k - 1$ extra terms of the form

$$(\sigma(\xi_k, t_k) + \sigma(x_k, t_k) - \sigma(\xi_k, t_k))(\sigma(\xi_k, t_k) + \sigma(x_k, t_k) - \sigma(\xi_k, t_k))^* \,,$$

to calculate. Given the dependence of these terms on x_k, these terms cannot be pre-computed. We point out, however, that this difference in cost can be considerably reduced for diffusions relating to the chemical

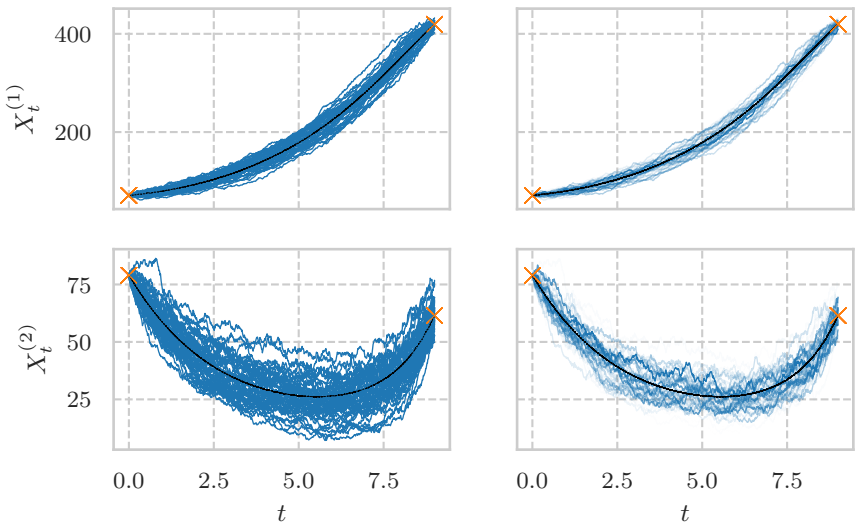

Figure 13: A plot of fifty paths simulated using the $\overline{\text{RB}}^{\text{LNA}}$ approach introduced in this chapter from the Lotka-Volterra SDE introduced in Section 3.1.2. As with the previous figures the plots on the left are the fifty two-dimensional simulated paths with no transparency and the plots on the right are the fifty two-dimensional paths with transparency inversely proportional to their normalised weights. The two-dimensional initial condition and observation are illustrated with crosses and the deterministic path, $\xi_t = \eta_t + \mathbb{E}(\hat{R}_t|Y^1 = y^1)$, is plotted with a dashed line.

Langevin diffusion (see, for example, Ethier and Kurtz, 1986), where the volatility is of the form

$$\zeta(x,t) = S\Lambda(x,t)^2 S^* \ ,$$

where $S \in \mathbb{R}^{d\times r}$ is a constant matrix, and $\Lambda \in \mathbb{R}^{r\times r}$ is a diagonal matrix. In this case, we can circumvent the calculation of partial sums of symmetric matrices of size $d\times d$ involved in the calculation of $\Psi_K^{\overline{\text{RB}}}$, and, instead, calculate partial sums of vectors of size r by letting $\sigma(x,t) = S\Lambda(x,t)$, so that

$$\Psi_K^{\overline{\text{RB}}} := \Delta t\zeta(x_k,t_k) + S\Delta t \sum_{j=k+1}^{K-1} \varphi_{jk}\varphi_{jk}^* S^* \ ,$$

where

$$\varphi_{jk} := \Lambda(\xi_j,t_j) + \Lambda(x_k,t_k) - \Lambda(\xi_k,t_k) \ .$$

Thus, if r, the number of reactions, is significantly smaller than $d^2/2$, the computational cost of calculating $\Psi_k^{\overline{\text{RB}}}$ can be significantly reduced.

3.3.2 *A Simulation Study*

In this section we compare the performance of the residual-bridge constructs introduced in this thesis to the corresponding residual-bridge

constructs of Whitaker et al., 2017, and the MDB construct of Durham and Gallant, 2002, on three diffusions; the Birth-Death (BD) diffusion (Section 3.1.1), the Lotka-Volterra (LV) diffusion (Section 3.1.2) and a diffusion corresponding to a simple model of gene expression (GE, Section 3.1.3). Due to the simplicity of the drift and volatility of the BD diffusion, the term, η_t, defined by Equation (63), along with the terms G_t and ϕ_t defined in Lemma 3.2.1 are analytically tractable with $\eta_t = x_0 \exp((\theta_1 - \theta_2)t)$, $G_t = \exp((\theta_1 - \theta_2)t)$, and

$$\phi_t = \frac{(\theta_1 + \theta_2)}{(\theta_1 - \theta_2)} \eta_t (\exp((\theta_1 - \theta_2)t) - 1) \ .$$

We use the same parameters, θ, and initial conditions, x_0, as those used in Whitaker et al., 2017, for the BD diffusion; $\theta = (\theta_1, \theta_2) = (0.1, 0.8)$, $x_0 = 50$, so that sample paths of the diffusion exhibit exponential decay. We also use the same parameters, θ, and initial conditions, x_0, as those used in Whitaker et al., 2017 for the Lotka-Volterra diffusion;

$$\theta = (\theta_1, \theta_2, \theta_3) = (0.5, 0.0025, 0.3) \ , \quad x_0 = (71, 79) \ ,$$

and we use the following parameters,

$$\theta = (\gamma_R, \gamma_P, k_P, b_0, b_1, b_2, b_3) = (0.7, 0.72, 3, 80, 0.05, 2, 50) \ ,$$

and initial condition $x_0 = (70, 70)$ for the diffusion corresponding to the simple model of gene expression. We fix Δt to be 0.01 for the BD diffusion, 0.1 for the LV diffusion, and 0.01 for the GE diffusion. We chose 10 equally-spaced values for T between; 0 and 2 for the BD diffusion, 0 and 10 for the LV diffusion, and 0 and 4 for the GE diffusion. To compare the performance of the proposals in challenging scenarios, we choose $P_1 = I$, and $\Sigma_1 = 10^{-12} I$, so that the observation, Y, is such that

$$Y | X_K = x \sim \mathrm{N}(x, 10^{-12} I) \ ,$$

and, therefore, essentially corresponds to exact observations of the diffusion[2]. For each value of T, we simulated $10,000$ values for $Y_T^{(1)}$ (where we have emphasised the dependence on T) using the EM approximation to forward simulate values of the path at points of the partition. For each collection of $10,000$ values we chose, for the BD diffusion, three terminal points for y_T, corresponding to the 2.5%, 50%, and 97.5% quantiles. For the LV and GE diffusions, we first take the logarithm of the $10,000$ simulated values of y_T. Then we choose five terminal points for $\log(y_T)$, corresponding to the mean, along with one-and-a-half standard deviations either side of the mean along the axes of the principal components. Figure 14 shows a histogram of the $10,000$ simulated observations, y_T, of the BD diffusion, where $T = 2$. The orange lines show

2 This small choice of variance in the observation is purely to generate challenging scenarios. In practice, if exact observations of the diffusion were available, the inference procedure would be slightly different (see, for example, Pedersen, 1995; Durham and Gallant, 2002) and is not considered in this thesis.

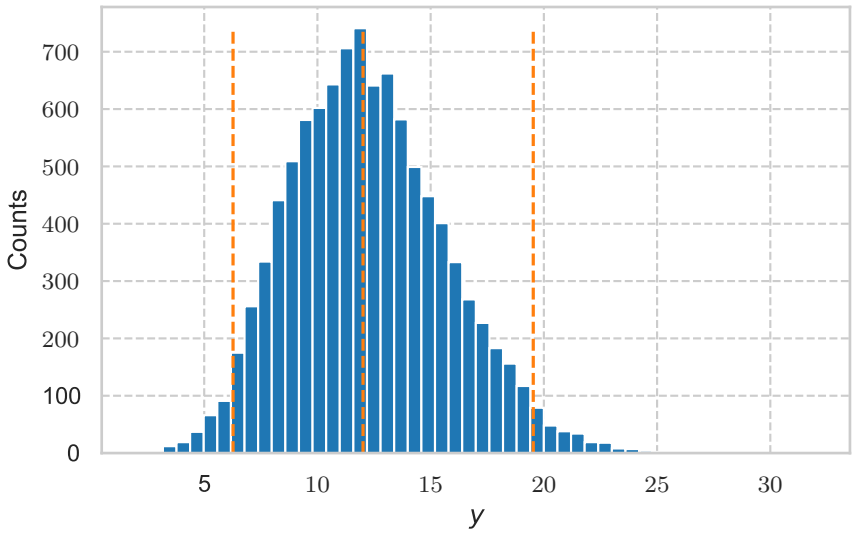

Figure 14: A histogram of the 10,000 simulated observations, y_T, of the BD diffusion, where $T = 2$. The orange lines show the locations of the 2.5%, 50%, and 97.5% quantiles.

the locations of the 2.5%, 50%, and 97.5% quantiles. Figure 15 show a scatter plot of the 10,000 simulated, two-dimensional, observations, y_T of the LV diffusion, where $T = 10$. The orange dots show the locations of the points chosen, as described above, for the simulation study. Similarly, Figure 16, shows a scatter plot of the 10,000 simulated, two-dimensional, observations, y_T of the GE diffusion, where $T = 4$. Again, the orange dots show the locations of the points chosen, as described above, for the simulation study.

For each combination of (T, y_T), we ran the MDB of Durham and Gallant, 2002, the residual-bridge construct of Whitaker et al., 2017, with the two choices for ξ_t, RB^{ODE} and RB^{LNA}, along with the residual-bridge construct introduced in this thesis (Section 3.3) with the same two choices for ξ_t, $\overline{\text{RB}}^{\text{ODE}}$ and $\overline{\text{RB}}^{\text{LNA}}$. For each of the five constructs, we simulated $N = 1,000,000$ independent skeleton paths and calculated the effective sample size per second (ESS/s) from the normalised importance weights (Section 2.3.4):

$$\text{ESS/s}\ (\tilde{w}_{1:N}) = \frac{(\tilde{w}_1^2 + \ldots + \tilde{w}_N^2)^{-1}}{\text{execution time in seconds}}\ . \tag{70}$$

For completeness we have included, in Appendix C, the average relative effective sample sizes defined by

$$\text{Rel. ESS}\ (\tilde{w}_{1:N}) = N^{-1}(\tilde{w}_1^2 + \ldots + \tilde{w}_N^2)^{-1}\ , \tag{71}$$

along with the execution times for each proposal and for each combination of (T, y_T) for the BD, LV, and GE diffusions.

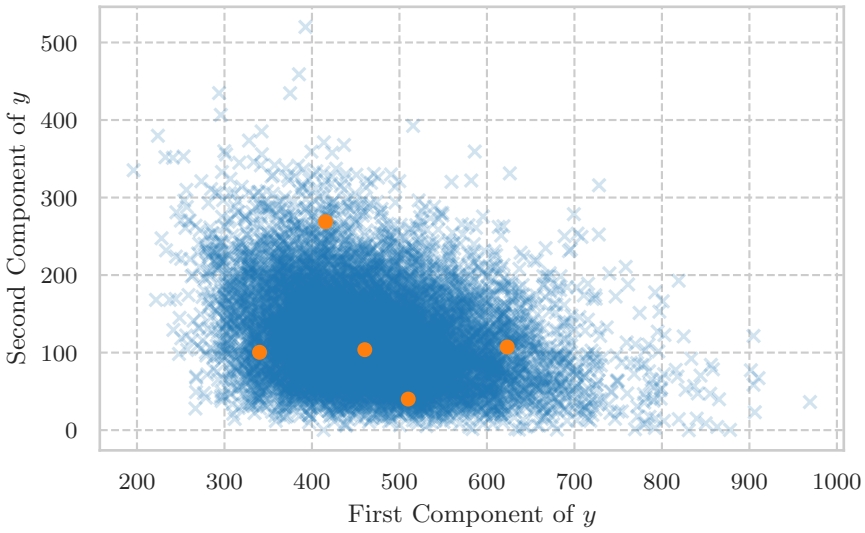

Figure 15: A scatter plot of the 10,000 simulated, two-dimensional, observations, y_T of the LV diffusion, where $T = 10$. The orange dots show the locations of the points chosen for the simulation study.

3.3.3 *Results*

To ease visualisation of comparative performance, Figures 17, 18 and 19, which illustrate the results for the BD, LV and GE diffusion respectively, plot, for four pairs of proposals, the effective sample size per second for one of the pair of proposals relative to the other for each combination of (T, y_T) for which both proposals had an effective sample size of at least *one hundred*. A small effective sample size relative to the maximum effective sample size possible (in this case one-million) implies that, for that particular combination of (T, y_T), the proposal is a poor approximation of the true conditioned diffusion, and, therefore, the resulting weights are highly variable. As such, the ESS, which, as detailed in Section 2.3.4, is really an approximation to the variance of the idealised estimator divided by the variance of the Importance Sampling estimator arising from the weights, has, as an approximation, a large variance. Thus, including relative efficiencies of proposals where one proposal has a small ESS is potentially misleading. This is why we have "dropped" such relative efficiencies from the plots, and have chosen an ESS of 100 as our threshold. The four pairs of proposals are chosen to approximate the sequential ordering in which the chapter has been presented. We emphasise that the larger the effective sample size per second, the more computationally efficient the proposal is for that particular choice of inter-observation time T and observation y_T.

The figures illustrate that across all three diffusions the statistical efficiency, in terms of the effective sample size per second, of the novel residual-bridge proposals introduced in this thesis ranges from just over half that of the efficiency of the equivalent residual-bridge proposals of Whitaker et al., 2017 to several orders of magnitude more efficient. In

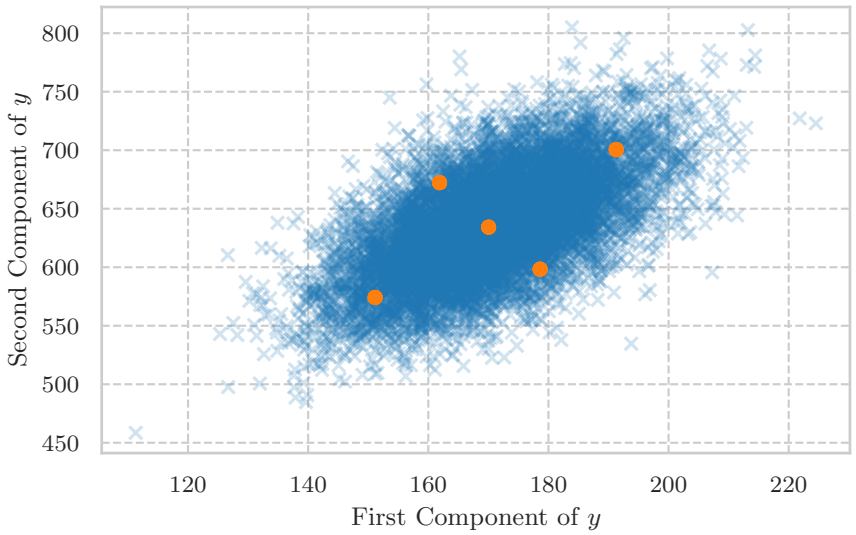

Figure 16: A scatter plot of the 10,000 simulated, two-dimensional, observations, y_T of the GE diffusion, where $T = 4$. The orange dots show the locations of the points chosen for the simulation study.

particular, our bridges tend to perform worse for the BD diffusion (Figure 17); our $\overline{\text{RB}}^{\text{ODE}}$ proposal is between 0.9 and 1.01 times as efficient as the RB^{ODE} proposal of Whitaker et al., 2017, and our $\overline{\text{RB}}^{\text{LNA}}$ proposal is between 0.55 and 0.95 times as efficient as the RB^{LNA} proposal of Whitaker et al., 2017. In contrast our bridges tend to perform the same or significantly better for the LV diffusion (Figure 18) and the GE diffusion (Figure 19). Indeed, for the LV diffusion, our $\overline{\text{RB}}^{\text{ODE}}$ proposal is between 0.86 and 119 times as efficient as the RB^{ODE} proposal of Whitaker et al., 2017, and our $\overline{\text{RB}}^{\text{LNA}}$ proposal is between 0.8 and 232 times as efficient as the RB^{LNA} proposal of Whitaker et al., 2017. Furthermore, for the GE diffusion our $\overline{\text{RB}}^{\text{ODE}}$ proposal is between 1.01 and 150 times as efficient as the RB^{ODE} proposal of Whitaker et al., 2017, and our $\overline{\text{RB}}^{\text{LNA}}$ proposal is between 0.88 and 85 times as efficient as the RB^{LNA} proposal of Whitaker et al., 2017. In all cases, the biggest differences in efficiency occur for the larger inter-observation times. For the BD diffusion, our residual-bridge proposals are least efficient, relative to the equivalent residual-bridge proposals of Whitaker et al., 2017 for the observation corresponding to the 5% quantile, which is very close to zero. Moreover, the relative efficiency of our $\overline{\text{RB}}^{\text{LNA}}$ proposal decreases monotonically with increasing inter-observation time T. On the other hand, for the LV and GE diffusions, the relative efficiency of our residual-bridge proposals is almost montonically increasing with increasing T, with the larger relative efficiencies occuring for those inter-observation times which are at least one half of the maximum inter-observation time considered. Furthermore, the improvement in efficiency, over the equivalent residual-bridge constructs of Whitaker et al., 2017, are greater for the four non-central observations suggesting

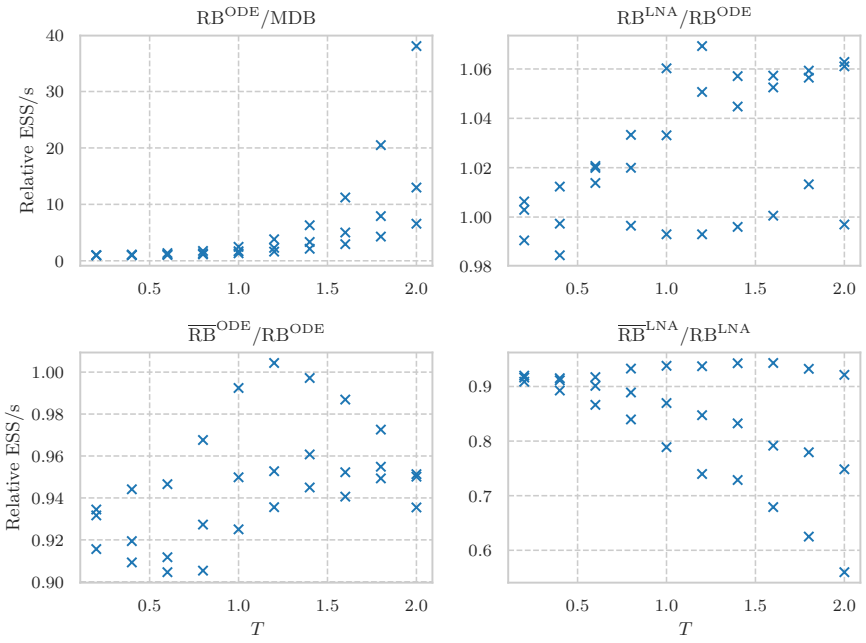

Figure 17: Plots of the comparative effective sample size per second for four pairs of proposals and for a variety of combinations of (T, y_T) corresponding to the birth-death diffusion.

that our proposals are relatively more efficient when the observations are extreme.

One drawback of the proposed residual-bridge constructs stems from the fact that, at intermediate time points, discrepancies of sample paths of the conditional diffusion from the deterministic path, ξ_t, can be relatively large. Preserving the resulting discrepancies in the drift and volatility, when for $\overline{\text{RB}}^{\text{LNA}}$ these should be 0 at time T, for example, must be sub-optimal. An interpolation scheme which is both *justifiable* and *computationally efficient*, however, eludes us. These discrepancies are particularly evident when paths of the diffusion are likely to come close to a *reflecting* boundary of the diffusion since, in this case, the approximating deterministic path produced by either the ODE or the LNA often fails to capture the true dynamics of the diffusion. This is what happens for the BD diffusion where the x-axis is a reflecting boundary and justifies why the relative efficiency of our residual-bridge proposals deteriorates the closer the observation is to the reflecting boundary. We note that, for the BD diffusion, one can transform the diffusion to a diffusion with unit volatility. Specifically, if we let

$$Z_t := 2\sqrt{\frac{X_t}{(\theta_1 + \theta_2)}}\,,$$

then Z_t satisfies

$$\mathrm{d}Z_t = \left(\frac{(\theta_1 - \theta_2)}{2} Z_t - \frac{1}{2Z_t}\right)\mathrm{d}t + \mathrm{d}B_t\,, \quad Z_0 = 2\sqrt{\frac{x_0}{(\theta_1 + \theta_2)}}\,.$$

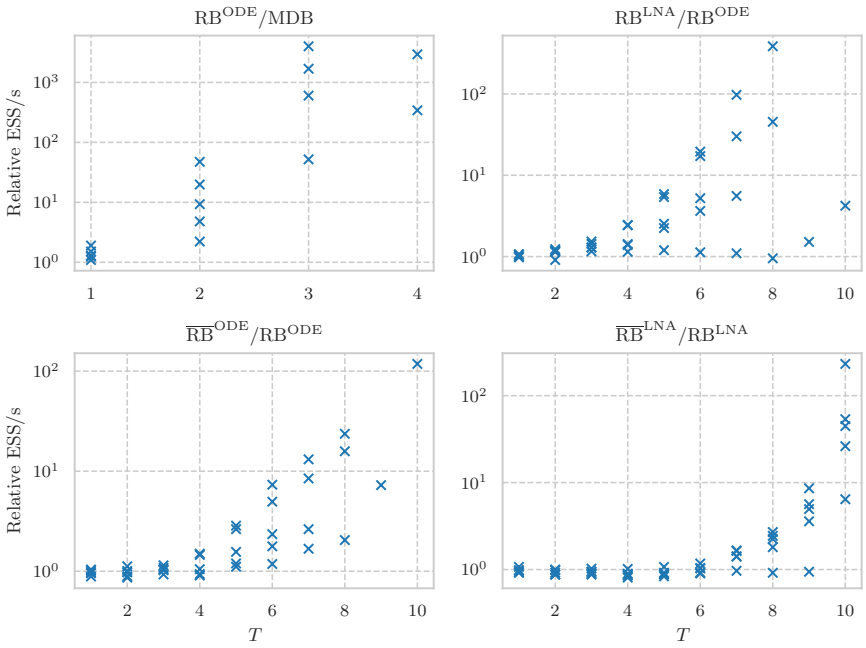

Figure 18: Plots of the comparative effective sample size per second for four pairs of proposals and for a variety of combinations of (T, y_T) corresponding to the Lotka-Volterra diffusion.

As the volatility is constant, applying the residual-bridge construct introduced in this chapter to the transformed diffusion is equivalent to applying the residual-bridge construct of Whitaker et al., 2017 to the transformed diffusion and thus the resulting effective sample sizes will be *identical* (provided, of course, the same random numbers are used). However, we emphasise that in most cases of practical interest one will not be able to transform the diffusion to one of unit volatility, and, therefore care *must* be taken when implementing the residual-bridge constructs introduced in this thesis.

3.3.4 *Absolute Continuity*

As discussed in Section 3.2.1, proving absolute continuity of the novel proposals introduced in this thesis is beyond the scope of this thesis. However, in this section, we provide numerical evidence, via a simulation study, suggesting that these residual-bridge proposals are robust to a decreasing step-size, Δt. This simulation study will partially extend the simulation study conducted in Section 3.3.2 by applying the two residual-bridge constructs introduced in this thesis, $\overline{\mathrm{RB}}^{\mathrm{ODE}}$ and $\overline{\mathrm{RB}}^{\mathrm{LNA}}$, to three diffusions; BD, LV, and GE. For consistency, we will use the same parameters and initial conditions as those used in Section 3.3.2. To test the proposals in a broad variety of scenarios, we chose three values for T; $(0.2, 1, 2)$ for the BD diffusion, $(1, 4, 7)$ for the LV diffusion, and $(0.4, 2, 3.6)$ for the GE diffusion, corresponding to a small, medium and large inter-observation interval. For each value of T, we

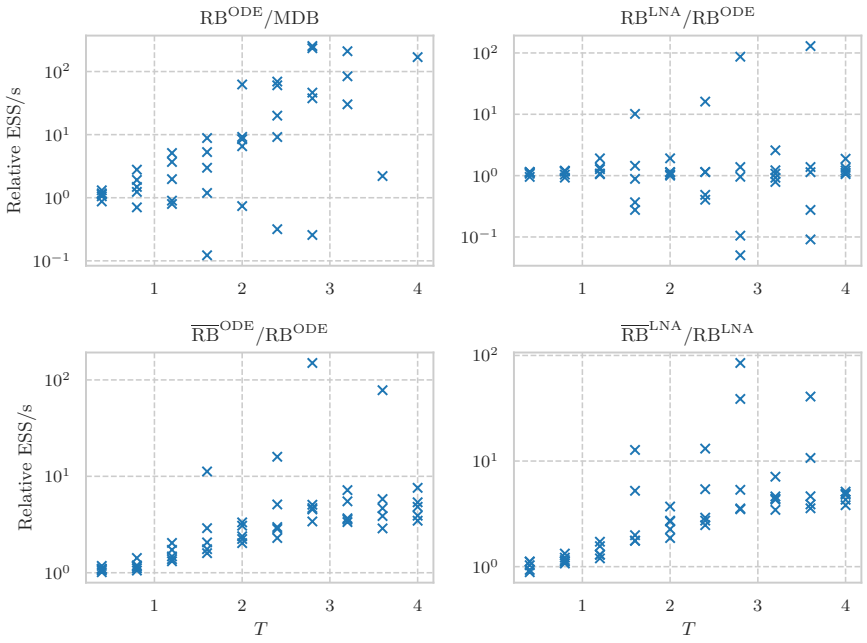

Figure 19: Plots of the comparative effective sample size per second for four pairs of proposals and for a variety of combinations of (T, y_T) corresponding to the Lotka-Volterra diffusion.

chose two observations y_T from the set of observations simulated for the simulation study detailed in Section 3.3.2 ; corresponding to the *centre* of the simulated observations and one other chosen at random. We chose five different values for Δt; $(0.01, 0.005, 0.001, 0.0005, 0.0001)$ for the BD and GE diffusions and $(0.1, 0.05, 0.01, 0.005, 0.001)$ for the LV diffusion. For each proposal, and each combination of $(T, y_T, \Delta t)$, we simulated one million independent skeleton paths and calculated the relative effective sample size (as defined by (71)) from the normalised importance weights[3].

3 For all of the models and observations, the observation variance that was used, 10^{-12}, is several orders of magnitude smaller than the eigenvalues of the variance matrix at the observation, so the empirical evidence of absolute continuity is not affected by this.

Table 1: A table showing the relative effective sample sizes for one million independent skeleton paths simulated from the two proposals; $\overline{\text{RB}}^{\text{ODE}}$ and $\overline{\text{RB}}^{\text{LNA}}$ for three diffusion models (BD, LV, and GE), three inter-observation times (small, medium, and large), a sequence of decreasing step-sizes, and for two observations (the centre, and one other chosen at random for each combination of (model, T, y_T), but fixed for the different step-sizes). The range of step-sizes are $\Delta t = 0.01, 0.005, 0.001, 0.0005, 0.0001$ for the BD and GE diffusions, and $\Delta t = 0.1, 0.05, 0.01, 0.005, 0.001$ for the LV diffusion and the results are displayed in decreasing step-size order. That is, for each group of five results, corresponding to the different values for Δt, the effective sample size corresponding to the largest and smallest value for Δt is at the top and bottom of the group respectively.

Proposal	$\overline{\text{RB}}^{\text{ODE}}$									$\overline{\text{RB}}^{\text{LNA}}$								
Diffusion Model	Birth-Death			Lotka-Volterra			Gene-Expression			Birth-Death			Lotka-Volterra			Gene-Expression		
Observation	Δt	Centre	Other	Δt	Centre	Other	Δt	Centre	Other	Δt	Centre	Other	Δt	Centre	Other	Δt	Centre	Other
Small T	0.01	0.9992	0.9990	0.1	0.9719	0.9350	0.01	0.9370	0.8407	0.01	0.9992	0.9986	0.1	0.9716	0.9621	0.01	0.9372	0.9054
	0.005	0.9995	0.9991	0.05	0.9733	0.9407	0.005	0.9408	0.8493	0.005	0.9995	0.9987	0.05	0.9731	0.9634	0.005	0.9409	0.9083
	0.001	0.9997	0.9992	0.01	0.9744	0.9449	0.001	0.9441	0.8568	0.001	0.9997	0.9987	0.01	0.9744	0.9643	0.001	0.9442	0.9107
	0.0005	0.9997	0.9992	0.005	0.9745	0.9455	0.0005	0.9444	0.8574	0.0005	0.9997	0.9987	0.005	0.9745	0.9644	0.0005	0.9445	0.9108
	0.0001	0.9997	0.9992	0.001	0.9746	0.9460	0.0001	0.9446	0.8581	0.0001	0.9997	0.9987	0.001	0.9746	0.9644	0.0001	0.9447	0.9112
Medium T	0.01	0.9926	0.9878	0.1	0.6635	0.4122	0.01	0.4289	0.2497	0.01	0.9925	0.9393	0.1	0.6574	0.6387	0.01	0.4289	0.4029
	0.005	0.9938	0.9890	0.05	0.6721	0.4396	0.005	0.4355	0.2263	0.005	0.9936	0.9408	0.05	0.6694	0.6514	0.005	0.4352	0.4118
	0.001	0.9947	0.9898	0.01	0.6767	0.4598	0.001	0.4469	0.2814	0.001	0.9944	0.9419	0.01	0.6765	0.6593	0.001	0.4468	0.4190
	0.0005	0.9948	0.9899	0.005	0.6746	0.4566	0.0005	0.4442	0.2721	0.0005	0.9945	0.9421	0.005	0.6749	0.6582	0.0005	0.4442	0.4107
	0.0001	0.9948	0.9900	0.001	0.6755	0.4487	0.0001	0.4372	0.2666	0.0001	0.9946	0.9421	0.001	0.6757	0.6573	0.0001	0.4370	0.4077
Large T	0.01	0.9367	0.9171	0.1	0.3379	0.0971	0.01	0.1404	0.0740	0.01	0.9344	0.7875	0.1	0.3350	0.3172	0.01	0.1403	0.1231
	0.005	0.9387	0.9208	0.05	0.3678	0.0839	0.005	0.1551	0.0806	0.005	0.9362	0.7926	0.05	0.3683	0.3289	0.005	0.1554	0.1401
	0.001	0.9405	0.9232	0.01	0.3688	0.0753	0.001	0.1557	0.0905	0.001	0.9378	0.7964	0.01	0.3756	0.3377	0.001	0.1556	0.1508
	0.0005	0.9406	0.9230	0.005	0.3709	0.0743	0.0005	0.1519	0.0785	0.0005	0.9378	0.7968	0.005	0.3772	0.3381	0.0005	0.1520	0.1502
	0.0001	0.9406	0.9242	0.001	0.3664	0.0716	0.0001	0.1624	0.0776	0.0001	0.9378	0.7976	0.001	0.3727	0.3375	0.0001	0.1628	0.1227

Table 1 collates the relative effective samples sizes for each diffusion as a function of the observation time, T, the observation value, y_T, and step-size, Δt. It shows that the relative effective sample size for the proposals introduced in this chapter are consistent across varying values of Δt for the scenarios considered in the simulation study. This, therefore, suggests that our residual-bridge proposals can be implemented without the need to consider the effect that decreasing the step-size, Δt, has on the resulting variability of the weights. Moreover, we stress that the smallest Δt considered here is on the border of what is computationally feasible, in the sense that any smaller Δt, with the same inter-observation interval T, will lead to an algorithm which is prohibitively costly. Therefore, it can be argued that our residual-bridge proposals are consistent for any step-size, Δt, that may be used in practice for the particular diffusions considered here.

3.4 SUMMARY

In this Chapter we introduced a new residual-bridge proposal for approximately simulating conditioned diffusions formed by applying the modified diffusion bridge approximation of Durham and Gallant, 2002 to the difference between the true diffusion and a second, approximate, diffusion driven by the same Brownian motion. By attempting to account for volatilities which are not constant, this proposal can lead to gains in efficiency over the residual-bridge constructs of Whitaker et al., 2017 in situations where the volatility varies considerably, as is often the case for larger inter-observation times and for time-inhomogeneous volatilities. We showed, via a simulation study in Section 3.3.2, how, for larger inter-observation times, this new proposal led to larger- sometimes one to two orders of magnitude larger- relative effective sample sizes per second compared to the residual-bridge constructs of Whitaker et al., 2017, for both the Lotka-Volterra diffusion (3.1.2) and a simple diffusion for gene expression (3.1.3). We highlighted that a drawback of the new proposal is that, at inter-observation time points, discrepancies of sample paths of the conditional diffusion from the deterministic path, around which the new residual-bridge construct is centered, can be relatively large. We demonstrated, in Section 3.3.2, how, for the Birth-Death diffusion, these discrepancies become evident as neither the approximating deterministic path produced by the ODE or the LNA captures the true dynamics of the diffusion as the diffusion approaches the x-axis- a reflecting boundary of the diffusion. Indeed, we showed that, for such a diffusion, these discrepancies led to lower relative effective sample sizes per second compared to the residual-bridge constructs of Whitaker et al., 2017.

4 EXCHANGEABLE PARTICLE MCMC

4.1 THE INTRODUCTION

Recently, efforts which utilise the sequential approaches introduced in Section 2.4 within Markov Chain Monte Carlo algorithms of Section 2.3.5 have resulted in powerful MCMC schemes that are tailored towards inference for densities with a certain, *sequential*, structure, such as those densities which arise when conducting inference for stochastic processes. Indeed, utilising the fact that sequential algorithms can, up to a constant of proportionality, produce unbiased approximations to expectations of functions defined on the *path space* (see, for example, Proposition 7.4.1, Del Moral, 2012, and Proposition 1, Ala-Luhtala et al., 2016), within an *auxiliary* variable, *exact-approximate*, framework, Particle Markov Chain Monte Carlo (PMCMC) methods (Andrieu, Doucet, and Holenstein, 2010; Lindsten and Schön, 2013; Lindsten, Jordan, and Schön, 2014; Chopin and Singh, 2015) are able to conduct inference for the path of a stochastic process, and any parameters of the process, *offline*. Of particular interest to this thesis is the Particle Gibbs Sampler (PGS), introduced in Andrieu, Doucet, and Holenstein, 2010, which *mimics* an *idealized* Gibbs sampler by alternating between sampling parameters given a *path*, and, sampling a path given the parameters. The latter step relies on a conditional particle filter, which, given a *reference* path, simulates N candidate paths, along with corresponding weights, using a particle filter with $N + 1$ particles which has been *conditioned* on including the reference path as one of the $N + 1$ paths simulated. The main drawback of the PGS stems from the fact that the resampling step of the particle filter can result in candidate paths of the process coalescing backwards through time. This *path degeneracy* characteristic is accentuated when the dimension of the path space is large, or, when the transition density in the mutation step is such that the weights derived in the correction step have a large variance (Pitt and Shephard, 1999; Doucet and Johansen, 2011; Lin, Chen, and Liu, 2013). In both cases, the prominence of the path degeneracy problem ultimately results in a PGS which mixes poorly (Lindsten and Schön, 2013; Lindsten, Jordan, and Schön, 2014; Chopin and Singh, 2015).

The Particle Gibbs with Ancestor Sampling (PGAS) approach of Lindsten, Jordan, and Schön, 2014 (see also Lindsten et al., 2015) attempts to overcome this path degeneracy problem by introducing an ancestor sampling step into the PGS, which, at each time step, samples a new *history* of the reference path, thus allowing the proposed paths to degenerate to a path which is different from the reference path. By allowing the *degeneration* path to differ from the reference path, the PGAS algorithm can achieve much better mixing than the PGS (Lind-

sten, Jordan, and Schön, 2014). Unfortunately, the PGAS algorithm relies on being able to calculate the *likelihood* of the reference path having a particular *history* which makes the PGAS impossible to implement in scenarios where this *likelihood* is intractable. Moreover, if the weights of the particles have large variability, or, if the model is such that the *likelihood* of the reference path having a *history* which is not the same as the current history of the reference path is relatively small, then the PGAS will offer little improvement over the PGS (Lindsten et al., 2015). This makes the PGAS particularly ill-suited to conducting inference for diffusions, and, while the rejuvenation approach of Lindsten et al., 2015 overcomes these limitations, it does so at the expense of an increase in computational cost.

In a similar spirit to the Correlated Pseudo-Marginal Markov Chain Monte Carlo (CPsMMCMC) method of Deligiannidis, Doucet, and Pitt, 2015, Dahlin et al., 2015, and Murray and Graham, 2016, this thesis introduces a set of algorithms which attempt to overcome the path degeneracy problem by making the proposed paths *closer* to the reference path being conditioned upon, and, therefore, make it more likely that the chain moves to a path which is different from the reference path. This is done by simulating particles within the Sequential Monte Carlo procedure exchangeably as opposed to independently. As shown in Algorithm 2, Section 2.1.2, exchangeable samples can, given a certain level of smoothness, be made arbitrarily close to each other, and, therefore, the probability of moving from the reference path can be made arbitrarily close to one. By using exchangeability to generalise the Independence Sampler (Section 2.3.6.1) and the Particle Gibbs Sampler (Section 4.2.4), the locality of moves in the Exchangeable Sampler (Section 4.3) and the Exchangeable Particle Gibbs Sampler (Section 4.4) can be controlled by a *scaling* parameter which can be tuned to optimise the mixing of the resulting procedure (see Sections 4.3.1 and 4.4.1). As a consequence, these samplers can lead to chains with better mixing properties and, therefore, to MCMC estimators with smaller variances. Moreover, provided one can sample particles by inverting a cumulative distribution function (Algorithm 2), then such an approach is computationally efficient and, unlike the CPsMMCMC approach, its justification does not depend on the smoothness of the likelihood with respect to all the underlying random numbers[1].

In Section 4.2 we introduce particle MCMC methods and illustrate their advantages and drawbacks. In particular, in Section 4.2.4, we introduce the Particle Gibbs Sampler and demonstrate the path degeneration phenomena. As a precursor to the Exchangeable Particle Gibbs Sampler— which attempts to overcome the path degeneracy problem by generalising the Particle Gibbs Sampler— we introduce, in Section 4.3, the Exchangeable Sampler which is a generalisation of the multiple-proposal Independence Sampler of Section 2.3.6.1. We prove, via Corollary 4.3.7, that, under certain conditions, the Exchangeable Sampler is

1 It should be noted that in order to get ϵ-close paths, in the parlance of Section 2.1.2, there is a smoothness assumption when inverting the cumulative distribution function (see the end of Section 2.1.2).

geometrically ergodic even when the *importance* weights are unbounded and, hence, in scenarios where the Independence Sampler cannot be geometrically ergodic. We investigate the assumptions underpinning the result on several simple examples; one where the importance weight is exponentially increasing in the tails, one where the importance weight is polynomially increasing in the tails, and one where the importance weight is bounded. In Section 4.3.1 we derive, through Theorem 4.3.17, an optimal scaling result which gives the asymptotic form of the acceptance rate and the asymptotic form of expected squared jump distance in the Z-space in which the underlying exchangeable Normal random variables reside. We then show, numerically, for a simple Gaussian model, how well this result holds for finite d. In Section 4.3.2 we conduct a simulation study for the Exchangeable Sampler in four scenarios; the first three corresponding to the simple examples for which we investigated the geometric ergodicity assumptions, and the fourth corresponding to a realistic example involving the simulation of a conditioned Birth-Death diffusion. In Section 4.4 we introduce the Exchangeable Particle Gibbs Sampler and show that it satisfies the same ergodicity properties as the Particle Gibbs Sampler. Then, in Section 4.4.1, we derive, through Theorem 4.4.7, a form for the asymptotic expected acceptance rate and the asymptotic expected squared jump distance for the first component of the path in the Z-space for the Exchangeable Particle Gibbs Sampler which targets a product density composed of independent and identically distributed marginals. In an effort to give guidance on how to scale the number of particles with the number of observations, we go on to analyse these forms, both numerically, and theoretically, through Corollary 4.4.9, and show that, asymptotically, one should let the number of particles scale linearly with the number of observations. Later in that section we show, numerically, for a Linear Gaussian model, how well this result holds for finite d. Finally, in Section 4.4.2, we conduct a simulation study for the Exchangeable Particle Gibbs Sampler in two scenarios; the first corresponding to a Linear Gaussian model, and the second corresponding to a Lotka-Volterra diffusion model introduced in Chapter 3.

4.2 PARTICLE MCMC ALGORITHMS

Particle MCMC methods utilise the Sequential Importance Resampling procedure of Section 2.4.1.2, which, to be consistent with the literature— in particular the seminal work of Andrieu, Doucet, and Holenstein, 2010— is really a specific case of the Sequential Monte Carlo (SMC) approach. Suppose, then, that interest lies in targeting a sequence of densities $\{\pi_t(\theta, x_{0:t}) : \ t = 0, \ldots, T\}$, where T is fixed and where, for each $t \in \{0, \ldots, T\}$, π_t is defined on the space $\mathbb{R}^p \times \mathbb{R}^{d\times(t+1)}$ and is such that

$$\pi_t(\theta, x_{0:t}) = \frac{\gamma_t(\theta, x_{0:t})}{\eta_t} = \frac{\eta_t(\theta)}{\eta_t} \frac{\gamma_t(\theta, x_{0:t})}{\eta_t(\theta)} \ ,$$

for some known $\gamma_t(\theta, x_{0:t})$, and some, typically unknown, constant

$$\eta_t = \iint_{\mathbb{R}^p \times \mathbb{R}^{d\times(t+1)}} \gamma_t(\theta, x_{0:t}) \, \mathrm{d}\theta \, \mathrm{d}x_{0:t} = \int_{\mathbb{R}^p} \eta_t(\theta) \, \mathrm{d}\theta \; .$$

Much like the Sequential Importance Resampler of Section 2.4.1.2, the Sequential Monte Carlo procedure (Algorithm 8) relies upon a sequence of proposal densities $p_0(x_0|\theta), p_1(x_1|x_0,\theta), \ldots, p_T(x_T|x_{0:T-1},\theta)$, and, given a set of normalized weights, $\tilde{w}^{(1:N)}$, an *ancestral* resampling mechanism in the form of a probability mass function $\kappa(\cdot|\tilde{w}^{(1:N)})$. The

Algorithm 8 Sequential Monte Carlo Procedure

for $i = 1, \ldots, N$ **do**
  Sample $x_0^{(i)}$ with density $p_0(x_0^{(i)}|\theta)$ and set $\tilde{x}_0^{(i)} = x_0^{(i)}$.
  Calculate the i-th weight;

$$w_0(\tilde{x}_0^{(i)};\theta) = \gamma_0(\theta, x_0^{(i)})/p_0(x_0^{(i)}|\theta) \; .$$

end for
Normalize the weights by setting, for each $i \in \{1, \ldots, N\}$,

$$\tilde{w}_0^{(i)}(\tilde{x}_0^{(1:N)};\theta) = w_0(\tilde{x}_0^{(i)};\theta)/(w_0(\tilde{x}_0^{(1)};\theta) + \ldots + w_0(\tilde{x}_0^{(N)};\theta)) \; .$$

for $t = 1, \ldots, T$ **do**
  Sample ancestors $a_{t-1}^{(1:N)}$ with mass function $\kappa(a_{t-1}^{(1:N)}|\tilde{w}_{t-1}^{(1:N)})$.
  for $i = 1, \ldots, N$ **do**
    Sample $x_t^{(i)}$ with density $p_t(x_t|\tilde{x}_{t-1}^{(a_{t-1}^{(i)})},\theta)$ and set $\tilde{x}_t^{(i)} := (\tilde{x}_{t-1}^{(a_{t-1}^{(i)})}, x_t^{(i)})$.
    Calculate the i-th weight;

$$w_t(\tilde{x}_t^{(i)};\theta) = \frac{\gamma_t(\theta, \tilde{x}_t^{(i)})}{\gamma_{t-1}(\theta, \tilde{x}_{t-1}^{(a_{t-1}^{(i)})}) p_t(x_t^{(i)}|\tilde{x}_{t-1}^{(a_{t-1}^{(i)})},\theta)} \; .$$

  end for
  Normalize the weights by setting, for each $i \in \{1, \ldots, N\}$,

$$\tilde{w}_t^{(i)}(\tilde{x}_t^{(1:N)};\theta) = w_t(\tilde{x}_t^{(i)};\theta)/(w_t(\tilde{x}_t^{(1)};\theta) + \ldots + w_t(\tilde{x}_t^{(N)};\theta)) \; .$$

end for

following assumptions are made on the proposal densities and the resampling mechanism;

ASSUMPTIONS 4.2.1.

(S) For any $\theta \in \mathbb{R}^p$,

$$\mathrm{supp}(\gamma_0(\theta,\cdot)) \subseteq \mathrm{supp}(p_0(\cdot|\theta)) \; ,$$

and, for any $t \in \{1, \ldots, T\}$, and any $(\theta, x_{0:t-1}) \in \mathbb{R}^p \times \mathbb{R}^{d\times t}$,

$$\mathrm{supp}(\gamma_t(\theta, x_{0:t-1}, \cdot)) \subseteq \mathrm{supp}(\gamma_{t-1}(\theta, x_{0:t-1}) p_t(\cdot|x_{0:t-1},\theta)) \; .$$

(U) Given a set of normalised weights, $\tilde{w}^{(1:N)}$,

$$\mathbb{E}\left[\sum_{i=1}^{N} \mathbb{1}_k(A^{(i)}) \,\middle|\, \tilde{w}^{(1:N)}\right] = N\tilde{w}^{(k)} \; ,$$

for any $k \in \{1, \ldots, N\}$.

(E) For any permutation, σ, of $\{1, \ldots, N\}$,

$$\kappa(a^{(1:N)}|\tilde{w}^{(1:N)}) = \kappa(a^{(1:N)}|\tilde{w}^{(\sigma(1))}, \ldots, \tilde{w}^{(\sigma(N))}) \ .$$

(P) For any $(k, m) \in \{1, \ldots, N\}^2$,

$$\mathbb{P}(A^{(k)} = m|\tilde{w}^{(1:N)}) = \tilde{w}^{(m)} \ .$$

The support condition, (S), ensures that, for any $(\theta, x_{0:t-1}) \in \mathbb{R}^p \times \mathbb{R}^{d\times t}$, it is possible to reach anywhere where $\gamma_t(\theta, x_{0:t-1}, \cdot)$ is non-zero. The unbiased assumption, (U), on the resampling mechanism ensures that the estimator produced by the algorithm is *unbiased* (see, for example, Proposition 1, Ala-Luhtala et al., 2016). Moreover, the exchangeable assumption on the resampling mechanism, (E), ensures that the determination of the ancestor variables does not depend on the order of the weights and, therefore, that the indices have no effect on the paths generated by the procedure. Finally, the *permutation* assumption, (P), is a *technical* condition which will make demonstrating that the Particle MH Sampler, the Particle Gibbs Sampler, and the Exchangeable Particle Gibbs Sampler correctly target the density of interest, clearer. In practice, for the Sequential Monte Carlo procedure, as highlighted in Remark 3 of Section 2.4.1.2, the ancestors are set deterministically given the number of offspring, $O^{(1:N)}$, and, therefore, (P) does not hold. However, if the ancestor variables are randomly permuted, then Assumption (P) holds given Assumption (U). Indeed, for any $j \in \{0, \ldots, N\}$, let $s_j(o^{(1:N)})$ denote the j-th partial sum of the sequence $o^{(1:N)}$; that is, $s_0(o^{(1:N)}) := 0$ and, for any $j \in \{1, \ldots, N\}$,

$$s_j(o^{(1:N)}) := \sum_{k=1}^{j} o^{(k)} \ .$$

If the ancestor variables are randomly permuted; that is, by letting, for any $k \in \{1, \ldots, N\}$, and any $s_{k-1}(O^{(1:N)}) < j \leq s_k(O^{(1:N)})$, $A^{(j)} = \sigma(k)$, then, given Assumption (U),

$$\begin{aligned}
\mathbb{P}(A^{(k)} = m|\tilde{w}^{(1:N)}) &= \sum_{n=1}^{N} \mathbb{P}(A^{(k)} = m|O^{(m)} = n)\mathbb{P}(O^{(m)} = n|\tilde{w}^{(1:N)}) \\
&= \sum_{n=1}^{N} \frac{n}{N} \mathbb{P}(O^{(m)} = n|\tilde{w}^{(1:N)}) \\
&= \tilde{w}^{(m)} \ ,
\end{aligned}$$

for any $(k, m) \in \{1, \ldots, N\}^2$, and, therefore, Assumption (P) holds.

REMARK 6. *Ultimately, interest is in paths generated by the SMC procedure and not the corresponding indices. Assuming one uses an exchangeable resampling mechanism, as we have done in this thesis— property (E) of Assumptions 4.2.1— the indices have no effect on the paths generated by the procedure, since the particles are propagated forward independently of one another and irrespective of the actual value*

of the ancestor variables, and the determination of the ancestor variables does not depend on the order of the weights. Therefore, although randomly permuting the ancestor variables ensures condition (P) of Assumptions 4.2.1 holds, thereby making some of the technical arguments clearer, it is, in practice, not necessary.

It will be useful to define the joint mass-density function corresponding to all the variables produced by the Sequential Monte Carlo procedure;

$$\Psi(x_{0:T}^{(1:N)}, a_{0:T-1}^{(1:N)}|\theta) := \\ p_0^*(x_0^{(1:N)}|\theta) \prod_{t=1}^{T} \kappa(a_{t-1}^{(1:N)}|\tilde{w}_{t-1}^{(1:N)}) p_t^*(x_t^{(1:N)}|\tilde{x}_{t-1}^{(a_{t-1}^{(1)})}, \ldots, \tilde{x}_{t-1}^{(a_{t-1}^{(N)})}, \theta) \,, \tag{72}$$

where; to ease notation, the explicit dependence of the weights on θ and $\tilde{x}_{t-1}^{(1:N)}$ has been dropped; we have recursively defined $\tilde{x}_0^{(i)} := x_0^{(i)}$ and, for any $t \in \{2, \ldots, T\}$, $\tilde{x}_t^{(i)} := (\tilde{x}_{t-1}^{(a_{t-1}^{(i)})}, x_t^{(i)})$; and

$$p_0^*(x_0^{(1:N)}|\theta) := \prod_{i=1}^{N} p_0(x_0^{(i)}|\theta) \,, \tag{73}$$

$$p_t^*(x_t^{(1:N)}|\tilde{x}_{t-1}^{(a_{t-1}^{(1)})}, \ldots, \tilde{x}_{t-1}^{(a_{t-1}^{(N)})}, \theta) := \prod_{i=1}^{N} p_t(x_t^{(i)}|\tilde{x}_{t-1}^{(a_{t-1}^{(i)})}, \theta) \,. \tag{74}$$

It will also be useful to define the *lineage* of particles back from a certain point in time as this will provide a way of tracking particle paths through time:

DEFINITION 4.2.2 (Lineage). *Let $a_{0:T-1}^{(1:N)} \in \{1, \ldots, N\}^{N \times T}$ be ancestor variables simulated via the SMC procedure of Algorithm 8. The lineage function at time t, denoted $\mathcal{L}_t : \{1, \ldots N\} \times \{0, \ldots, t\} \to \{1, \ldots, N\}$, is defined recursively by*

$$\mathcal{L}_t(k, t) := k \,, \quad \text{for any } k \in \{1, \ldots, N\}.$$
$$\mathcal{L}_t(k, s) := a_s^{(\mathcal{L}_t(k, s+1))} \,, \quad \text{for any } (k, s) \in \{1, \ldots, N\} \times \{0, \ldots, t-1\}.$$

By definition, for any $t \in \{1, \ldots, T\}$, and any $i \in \{1, \ldots, N\}$,

$$\tilde{x}_t^{(i)} = (x_0^{(\mathcal{L}_t(i,0))}, \ldots, x_{t-1}^{(\mathcal{L}_t(i,t-1))}, x_t^{(\mathcal{L}_t(i,t))}) \,.$$

4.2.1 *The Pseudo-Marginal MH Sampler*

Recall, from Algorithm 8, that the *importance* weight of the *transition* of particle j at time t is given by

$$w_t(\tilde{x}_t^{(i)}; \theta) = \frac{\gamma_t(\theta, \tilde{x}_t^{(i)})}{\gamma_{t-1}(\theta, \tilde{x}_{t-1}^{(a_{t-1}^{(i)})}) p_t(x_t^{(i)}|\tilde{x}_{t-1}^{(a_{t-1}^{(i)})}, \theta)} \,.$$

The first MCMC approach which utilises the Sequential Monte Carlo procedure relies on the fact that the Sequential Monte Carlo estimator;

$$I_T(\theta, \tilde{X}_{0:T}^{(1:N)}) := \frac{1}{N^{T+1}} \prod_{t=0}^{T} \sum_{j=1}^{N} w_t(\tilde{X}_t^{(j)}; \theta) \,, \tag{75}$$

which is a generalisation of the Sequential Importance Resampling estimator of Section 2.4.1.2, is an unbiased approximation of $\eta_T(\theta)$ (see, for example, Proposition 7.4.1, Del Moral, 2012, or Proposition 1, Ala-Luhtala et al., 2016):

THEOREM 4.2.3. *Let $\tilde{X}_{0:T}^{(1:N)}$ be the paths generated by the Sequential Monte Carlo procedure (Algorithm 8), and $\theta \in \mathbb{R}^p$. Then, for any $t \in \{0, \ldots, T\}$,*

$$\mathbb{E}_\Psi[I_T(\Theta, \tilde{X}_{0:T}^{(1:N)}) | \Theta = \theta] = \eta_T(\theta) \,,$$

where I_T is the Sequential Monte Carlo estimator given by Equation (75), and, given θ, Ψ is the density of the random variables generated by the Sequential Monte Carlo estimator, given by Equation (72).

Such an estimator, therefore, can be used in place of the exact likelihood in a Markov Chain Monte Carlo algorithm (see the Pseudo-Marginal MCMC approach in, for example, Beaumont, 2003 and Andrieu and Roberts, 2009). The intuition behind such *exact-approximate* methods is that, even though the approximation to the likelihood is not exact, even up to a constant of proportionality, and, therefore, such an MCMC scheme does not target the density of interest, it does target an extended density which is the joint density of θ and the random variables in the estimator. Moreover, given the unbiasedness property, the marginal density for θ from the extended target is equal to the density of interest. Hence, the samples of θ generated by the algorithm— obtained by ignoring the *latent* variables generated— target the density of interest. In summary, such an approach is targeting a density defined in a higher-dimensional space whose marginal is the density of interest. Formally, let

$$\pi(\theta) := \int_{\mathbb{R}^{d(T+1)}} \pi_T(\theta, x_{0:T}) \, \mathrm{d}x_{0:T}$$

be the target of interest. Consider the extended target,

$$\pi_+(\theta, x_{0:T}^{(1:N)}, a_{0:T-1}^{(1:N)}) := \frac{I_T(\theta, \tilde{x}_{0:T}^{(1:N)})}{\eta_T} \Psi(x_{0:T}^{(1:N)}, a_{0:T-1}^{(1:N)} | \theta) \,. \tag{76}$$

This target is indeed a density whose marginal density for θ is π since, by Theorem 4.2.3,

$$\mathbb{E}_\Psi\left[\frac{I_T(\Theta, \tilde{x}_{0:T}^{(1:N)})}{\eta_T} | \Theta = \theta\right] = \frac{\eta_T(\theta)}{\eta_T} \,.$$

One can construct an MCMC algorithm which targets π_+ as follows. Suppose the chain is at state $(\theta, x_{0:T}^{(1:N)}, a_{0:T-1}^{(1:N)})$. Then a new state $(\theta^*, y_{0:T}^{(1:N)}, b_{0:T-1}^{(1:N)})$ can be proposed by proposing a θ^* from some proposal density $q(\cdot|\theta)$[2] and proposing $(y_{0:T}^{(1:N)}, b_{0:T-1}^{(1:N)})$ via the Sequential Monte Carlo procedure; that is, from $\Psi(\cdot|\theta^*)$. The proposed new state can then be accepted with either Barker's or the Metropolis-Hastings acceptance probability. The Metropolis-Hastings acceptance probability is given by

$$\begin{aligned}
&1 \wedge \frac{\pi_+(\theta^*, y_{0:T}^{(1:N)}, b_{0:T-1}^{(1:N)}) q(\theta|\theta^*) \Psi(x_{0:T}^{(1:N)}, a_{0:T-1}^{(1:N)}|\theta)}{\pi_+(\theta, x_{0:T}^{(1:N)}, a_{0:T-1}^{(1:N)}) q(\theta^*|\theta) \Psi(y_{0:T}^{(1:N)}, b_{0:T-1}^{(1:N)}|\theta^*)} \\
&= 1 \wedge \frac{I_T(\theta^*, \tilde{y}_{0:T}^{(1:N)}) \Psi(y_{0:T}^{(1:N)}, b_{0:T-1}^{(1:N)}|\theta^*) q(\theta|\theta^*) \Psi(x_{0:T}^{(1:N)}, a_{0:T-1}^{(1:N)}|\theta)}{I_T(\theta, \tilde{x}_{0:T}^{(1:N)}) \Psi(x_{0:T}^{(1:N)}, a_{0:T-1}^{(1:N)}|\theta) q(\theta^*|\theta) \Psi(y_{0:T}^{(1:N)}, b_{0:T-1}^{(1:N)}|\theta^*)} \\
&= 1 \wedge \frac{I_T(\theta^*, \tilde{y}_{0:T}^{(1:N)}) q(\theta|\theta^*)}{I_T(\theta, \tilde{x}_{0:T}^{(1:N)}) q(\theta^*|\theta)} .
\end{aligned}$$

Similarly, Barker's acceptance probability is given by

$$\frac{I_T(\theta^*, \tilde{y}_{0:T}^{(1:N)}) q(\theta|\theta^*)}{I_T(\theta^*, \tilde{y}_{0:T}^{(1:N)}) q(\theta|\theta^*) + I_T(\theta, \tilde{x}_{0:T}^{(1:N)}) q(\theta^*|\theta)} .$$

As can be seen, the acceptance probability is the same as the acceptance probability in the idealised case but with the unbiased approximation, I_T/η_T, of the likelihood $\pi(\theta)$, used in place of the likelihood (a fact that was initially shown in Beaumont, 2003). Indeed, the idealised case constructs a Markov Chain which, conditional on a current state θ, proposes a new state θ^* from $q(\cdot|\theta)$ and accepts, in the random-walk Metropolis-Hastings case, with probability

$$1 \wedge \frac{\eta_T(\theta^*) q(\theta|\theta^*)}{\eta_T(\theta) q(\theta^*|\theta)} .$$

In full, the Pseudo-Marginal Metropolis-Hastings (PsMMH) sampler is given by Algorithm 9.

For detailed results relating to the ergodicity of the Pseudo-Marginal Metropolis-Hastings sampler see Andrieu and Roberts, 2009. Of course, since the sampler in the extended space is a propose and accept-reject MCMC algorithm targeting an extended density, then the results of Section 2.3.6, in particular, Theorems 2.3.30, 2.3.31, and 2.3.34, are still valid when considering the chain on the extended state space. Moreover, ergodictiy, as given by Corollary 2.3.12, of the *marginal chain*— that is, the chain that is induced by just considering how θ transitions— follows from ergodicity of the *extended chain*— that is, the chain created by the propose and accept-reject MCMC sampler on the extended space— since, intuitively, the chance of moving to a state θ^* in the marginal space is greater than the chance of moving to a state $(\theta^*, y_{0:T}^{(1:N)}, b_{0:T-1}^{(1:N)})$,

2 See, for example, the independent and random-walk proposals of Section 2.3.5.

Algorithm 9 Pseudo-Marginal Metropolis-Hastings Sampler

1: Initialise the chain at some $\theta_0 \in \mathbb{R}^p$ and choose the number of iterations, M.
2: Sample $(x_{0:T}^{(1:N)}, a_{0:T}^{(1:N)})$ with density $\Psi(\cdot|\theta_0)$ defined by (72) via the SMC procedure given by Algorithm 8.
3: Calculate $\tilde{I}_0 := I_T(\theta_0, \tilde{x}_{0:T}^{(1:N)})$ as defined by (75).
4: **for** $m = 0, \ldots, M-1$ **do**
5: Sample a θ^* with density $q(\cdot|\theta_m)$.
6: Sample $(y_{0:T}^{(1:N)}, b_{0:T}^{(1:N)})$ with density $\Psi(\cdot|\theta^*)$ defined by (72) via the SMC procedure given by Algorithm 8.
7: Calculate $\tilde{I}^* := I_T(\theta^*, \tilde{y}_{0:T}^{(1:N)})$ as defined by (75).
8: Calculate the MH acceptance probability;

$$\alpha(\theta_m, \theta^*) := 1 \wedge \frac{\tilde{I}^* q(\theta|\theta^*)}{\tilde{I}_m q(\theta^*|\theta)} .$$

9: With probability $\alpha(\theta_m, \theta^*)$ set $\theta_{m+1} = \theta^*$, $\tilde{I}_{m+1} = \tilde{I}^*$; else set $\theta_{m+1} = \theta_m$, $\tilde{I}_{m+1} = \tilde{I}_m$.
10: **end for**

for any particular $(y_{0:T}^{(1:N)}, b_{0:T-1}^{(1:N)})$, in the extended space. Furthermore, since the bounding term of the total variation distance for uniformly ergodic chains— see Definition 2.3.14— is independent of where the chain starts, then uniform ergodicity of the *marginal chain* follows from uniform ergodicity of the *extended chain*. However, given the bound on the total variation distance for geometrically ergodic chains depends on where the chain was initialised, it is not necessarily the case, without some extra assumptions, that geometric ergodicity of the *marginal chain* follows from geometric ergodicity of the *extended chain*. For formal details and a more thorough discussion, see Andrieu and Roberts, 2009.

To demonstrate the Pseudo-Marginal Metropolis-Hastings Sampler consider the following Linear Gaussian model on a one-dimensional state space:

EXAMPLE 2. *Let* $X_0 \sim \mathrm{N}(0, 1)$, *and suppose that, for any* $t \in \{1, \ldots, 100\}$, *the transition distributions are given by* $(X_t|X_{t-1} = x_{t-1}, \Theta = \theta) \sim \mathrm{N}(\theta x_{t-1}, 1)$, *and the observation distributions are given by* $Y_t|X_t = x_t \sim \mathrm{N}(x_t, 0.3)$. *For simplicity, suppose that an improper, uniform over* $\mathbb{R}$, *prior is placed on* Θ *so that* $\gamma_0(\theta, x_0) = \phi(x_0; 0, 1)$ *and, therefore,* $\eta_0(\theta) = 1$. *Moreover, suppose that, for any* $t \in \{1, \ldots, T\}$, γ_t *is defined recursively by*

$$\gamma_t(\theta, x_{0:t}) = g_t(y_t|x_t)\phi(x_t; \theta x_{t-1}, 1)\gamma_{t-1}(\theta, x_{0:t-1}) ,$$

where $\phi(\cdot; \mu, \sigma^2)$ *denotes the density of a one-dimensional normal distribution with mean* μ *and variance* σ^2.

Using the bootstrap proposal; that is, $p_0(x_0|\theta) = \phi(x_0; 0, 1)$, and, for any $t \in \{1, \ldots, T\}$, $p_t(x_t|x_{t-1}, \theta) = \phi(x_t; \theta x_{t-1}, 1)$, we ran the PsMMH for ten-thousand iterations with $N \in \{100, 1000, 10000, 100000\}$ and calculated the approximation to the log-likelihood for each simulated θ. Figure 20 shows the approximations of the log-likelihood (left column), along with a histogram of the simulated θ samples (right column)

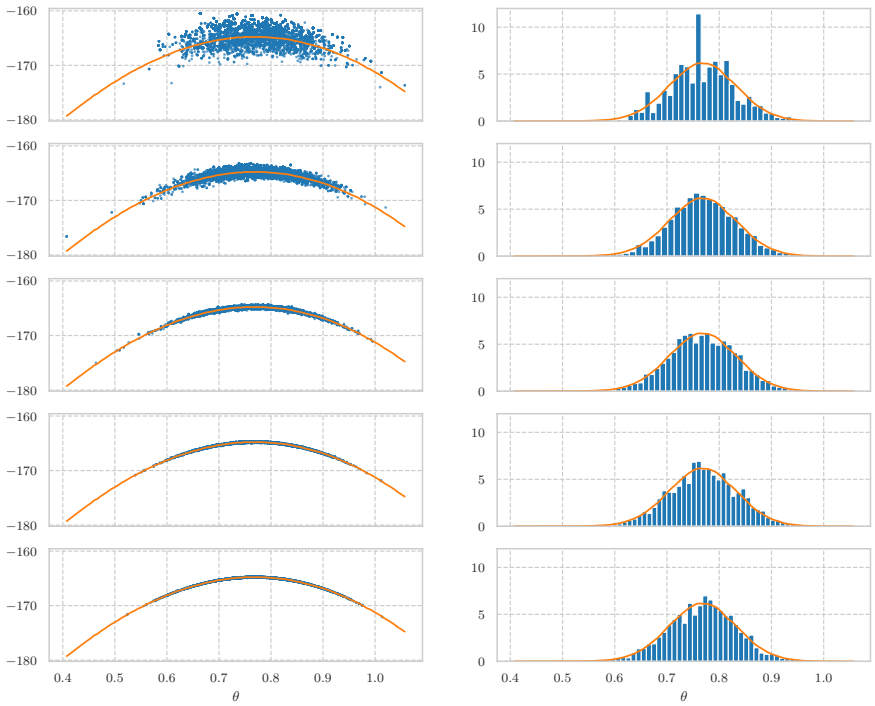

Figure 20: Plots of the approximated log-likelihood (left column) and the histogram of simulated θ samples (right column) corresponding to four simulations of the PsMMH sampler of ten-thousand iterations each, where each simulation uses a different number of particles; $N = 100$ for the top row, $N = 1000$ for the second row, $N = 10000$ for the third row, and $N = 100000$ for the bottom row. Each subplot in the left column also shows the true log-likelihood, and each subplot in the right column also shows the true posterior density; both coloured in orange.

for each $N \in \{100, 1000, 10000, 100000\}$. The top row corresponds to $N = 100$, the second row to $N = 1000$, the third row to $N = 10000$, and the bottom row to $N = 100000$. The figure also shows a plot of the true log-likelihood and the true target density. It can be seen that, even though the approximations of the log-likelihood are noisy when the number of particles is small, the sampler still targets the correct density. This is because the approximations to the likelihood are unbiased as is partially evidenced in the noise fluctuations in the log-likehood approximation around the true log-likelihood. It can also be seen that the approximations to the log-likelihood improve as the number of particles increases and this, in turn, results in an empirical density formed by the simulated samples which is closer to the true density after ten-thousand iterations.

4.2.2 *Conditional Sequential Monte Carlo*

Before proceeding to discuss other particle MCMC algorithms it will prove useful to first consider the Conditional Sequential Monte Carlo

procedure. Given a *reference* path, the Conditional Sequential Monte Carlo procedure proceeds by simulting N candidate paths, along with corresponding weights, using the Sequential Monte Carlo procedure with $N+1$ particles which has been *conditioned* on including the reference path as one of the $N+1$ paths simulated. Formally, dropping the explicit dependence on θ, recall the joint mass-density function Ψ corresponding to all the variables produced by the SMC procedure with $N+1$ particles:

$$\Psi(x_{0:T}^{(0:N)}, a_{0:T-1}^{(0:N)}|\theta) := \\ p_0^*(x_0^{(0:N)}|\theta) \prod_{t=1}^{T} \kappa(a_{t-1}^{(0:N)}|\tilde{w}_{t-1}^{(0:N)}) p_t^*(x_t^{(0:N)}|\tilde{x}_{t-1}^{(a_{t-1}^{(0)})}, \ldots, \tilde{x}_{t-1}^{(a_{t-1}^{(N)})}, \theta) ,$$

where; to ease notation, the explicit dependence of the weights on $\tilde{x}_{t-1}^{(1:N)}$ has been dropped; we have recursively defined $\tilde{x}_0^{(i)} := x_0^{(i)}$ and, for any $t \in \{2, \ldots, T\}$, $\tilde{x}_t^{(i)} := (\tilde{x}_{t-1}^{(a_{t-1}^{(i)})}, x_t^{(i)})$; and

$$p_0^*(x_0^{(0:N)}|\theta) := \prod_{i=0}^{N} p_0(x_0^{(i)}|\theta) , \tag{77}$$

$$p_t^*(x_t^{(0:N)}|\tilde{x}_{t-1}^{(a_{t-1}^{(0)})}, \ldots, \tilde{x}_{t-1}^{(a_{t-1}^{(N)})}, \theta) := \prod_{i=0}^{N} p_t(x_t^{(i)}|\tilde{x}_{t-1}^{(a_{t-1}^{(i)})}, \theta) . \tag{78}$$

The SMC procedure thus produces $N+1$ paths of the form

$$(x_0^{(\mathcal{L}_T(k,0))}, \ldots, x_T^{(\mathcal{L}_T(k,T))})$$

where $k \in \{0, \ldots, N\}$. At each step of the SMC procedure the particles are simulated forward independently of one another and thus the only dependence on the other particles comes from the resampling step. Therefore, the joint mass-density function of all the random variables produced by the SMC procedure conditional on the k-th path being fixed is given by

$$\psi(x_{0:T}^{(0:N)} \backslash \tilde{x}_T^{(k)}, a_{0:T-1}^{(0:N)} \backslash \tilde{a}_{T-1}^{(k)} | k, \tilde{x}_T^{(k)}, \tilde{a}_{T-1}^{(k)}, \theta) \\ \propto \prod_{\substack{i=0 \\ i \neq \mathcal{L}_T(k,0)}}^{N} p_0(x_0^{(i)}|\theta) \prod_{t=1}^{T} \frac{\kappa(a_{t-1}^{(0:N)}|\tilde{w}_{t-1}^{(0:N)})}{\mathbb{P}(A_{t-1}^{(\mathcal{L}_T(k,t))} = a_{t-1}^{(\mathcal{L}_T(k,t))}|\tilde{w}_{t-1}^{(0:N)})} \prod_{\substack{j=0 \\ j \neq \mathcal{L}_T(k,t)}}^{N} p_t(x_t^{(j)}|\tilde{x}_{t-1}^{(a_{t-1}^{(j)})}, \theta) . \tag{79}$$

Here, for simplicity, $\tilde{a}_T^{(k)}$ denotes the path of ancestor variables; that is,

$$\tilde{a}_T^{(k)} := a_0^{(\mathcal{L}_T(k,1))}, \ldots, a_{T-1}^{(\mathcal{L}_T(k,T))} ,$$

$x_{0:T}^{(0:N)} \backslash \tilde{x}_T^{(k)}$ denotes the sequence of variables $x_{0:T}^{(0:N)}$ with those corresponding to the k-th path removed; that is,

$$x_{0:T}^{(0:N)} \backslash \tilde{x}_T^{(k)} := x_0^{(-\mathcal{L}_T(k,0))}, \ldots, x_T^{(-\mathcal{L}_T(k,T))} ,$$

where, for any sequence $y^{(0:N)}$ and any $i \in \{0, \ldots, N\}$, $y^{(-i)}$ denotes the sequence $y^{(0:N)}$ with $y^{(i)}$ removed; that is,

$$y^{(-i)} := y^{(0)}, \ldots, y^{(i-1)}, y^{(i+1)}, \ldots, y^{(N)} ,$$

and, similarly, $a_{0:T-1}^{(0:N)} \backslash \tilde{a}_{T-1}^{(k)}$ denotes the sequence of ancestor variables $a_{0:T-1}^{(0:N)}$ with those ancestors corresponding to the k-th path removed; that is,

$$a_{0:T-1}^{(0:N)} \backslash \tilde{a}_{T-1}^{(k)} := a_0^{(-\mathcal{L}_T(k,1))}, \ldots, a_{T-1}^{(-\mathcal{L}_T(k,T))} .$$

The independence of the particles when simulating forward means that simulating particles forward in the Conditional Sequential Monte Carlo procedure is trivial. Thus, the only challenge is conditionally sampling ancestors; that is, given a sequence of weights $\tilde{w}^{(0:N)}$, simulating $a^{(-k)}$ with mass function

$$\frac{\kappa(a^{(0:N)}|\tilde{w}^{(0:N)})}{\mathbb{P}(A^{(k)} = a^{(k)}|\tilde{w}^{(0:N)})} . \tag{80}$$

Recall, from Section 2.4.1.2, that, for this thesis, a resampling procedure involves determining the number of offspring assigned to each particle; denoted, $O^{(0:N)}$, by sampling from a probability mass function, $\bar{\kappa}(\cdot|\tilde{w}^{(0:N)})$, such that $O^{(0)} + \ldots + O^{(N)} = N + 1$, and;

1. For each $i \in \{0, \ldots, N\}$,

   $$\mathbb{E}(O^{(i)}) = (N+1)\tilde{w}^{(i)} ,$$

   so that, the larger $\tilde{w}^{(i)}$ is, the more offspring particle i has on average.

2. For any permutation, σ, of $\{0, \ldots, N\}$,

   $$\bar{\kappa}(o^{(0:N)}|\tilde{w}^{(0:N)}) = \bar{\kappa}(o^{(\sigma(0))}, \ldots, o^{(\sigma(N))}|\tilde{w}^{(\sigma(0))}, \ldots, \tilde{w}^{(\sigma(N))}) ,$$

   so that, the assignment of the offspring does not depend on the order of the weights.

Thus, given a sequence of the number of offspring, $O^{(0:N)}$, assigned to each particle, where $O^{(k)} \geq 1$, one can preserve $A^{(k)} = a^{(k)}$ by simply setting $A^{(k)} = a^{(k)}$ and assigning $A^{(-k)}$ in such a way that

$$\sum_{\substack{j=0 \\ j \neq k}}^{N} \mathbb{1}_{\{k\}}(A^{(j)}) = O^{(k)} - 1 ,$$

and, for any $i \in \{0, \ldots, N\} \backslash \{k\}$,

$$\sum_{\substack{j=0 \\ j \neq k}}^{N} \mathbb{1}_{\{i\}}(A^{(j)}) = O^{(i)} .$$

Ultimately, then, the question becomes; how can one sample offspring $o^{(0:N)}$ from a probability mass function $\bar{\kappa}(o^{(0:N)}|\tilde{w}^{(0:N)})$ conditional on $o^{(k)} \geq 1$. Recall, from Section 2.4.1.2, that we will concentrate on the stratified residual resampling procedure. To this end, note, by Algorithm 5, that, for stratified resampling, the weights are initially shuffled before proceeding to determine which of the uniform samples lie in which *bucket*. Therefore, one does not need to worry about the initial ordering of the weights. Now, if $\tilde{w}^{(k)} \geq 2/(N+1)$, then, by Theorem 2.4.3, $O^{(k)} > (N+1)\tilde{w}^{(k)} - 2 \geq 0$. Thus $O^{(k)} \geq 1$. In this case, the conditional stratified resampling implementation follows exactly the same as the unconditional stratified resampling implementation given by Algorithm 5. If, on the other hand, $\tilde{w}^{(k)} < 1/(N+1)$, then, using the notation of Section 2.4.1.2, the only uniform sample that can belong to the set

$$(\tilde{w}_\sigma^{(0)} + \ldots + \tilde{w}_\sigma^{(\sigma(k)-1)}, \tilde{w}_\sigma^{(0)} + \ldots + \tilde{w}_\sigma^{(\sigma(k))}]$$

is $u_{\lceil (N+1)s_{\sigma(k)}(\tilde{w}_\sigma^{(0:N)})\rceil}$ where

$$s_{\sigma(k)}(\tilde{w}_\sigma^{(0:N)}) = \tilde{w}_\sigma^{(0)} + \ldots + \tilde{w}_\sigma^{(\sigma(k))} .$$

Therefore, in this case, the conditional stratified resampling implementation follows the unconditional stratified resampling implementation given by Algorithm 5 but where $u_{\lceil (N+1)s_{\sigma(k)}(\tilde{w}_\sigma^{(0:N)})\rceil}$ is not simulated, since, from the condition, this will have to lie in the set $(s_{\sigma(k)-1}(\tilde{w}_\sigma^{(0:N)}), s_{\sigma(k)}(\tilde{w}_\sigma^{(0:N)})]$. In the third case, where $\tilde{w}_\sigma^{(k)} \in [1/(N+1), 2/(N+1))$, the only uniform samples that can belong to the set $(s_{\sigma(k)-1}(\tilde{w}_\sigma^{(0:N)}), s_{\sigma(k)}(\tilde{w}_\sigma^{(0:N)})]$ are $u_{\lfloor (N+1)s_{\sigma(k)}(\tilde{w}_\sigma^{(0:N)})\rfloor}$ and $u_{\lceil (N+1)s_{\sigma(k)}(\tilde{w}_\sigma^{(0:N)})\rceil}$. Therefore, one chooses which of these to not simulate in a probabilistically way so that $u_{\lceil (N+1)s_{\sigma(k)}(\tilde{w}_\sigma^{(0:N)})\rceil}$ is not simulated with probability

$$\frac{(N+1)s_{\sigma(k)}(\tilde{w}_\sigma^{(0:N)}) - \lfloor (N+1)s_{\sigma(k)}(\tilde{w}_\sigma^{(0:N)})\rfloor}{(N+1)\tilde{w}_\sigma^{(\sigma(k))}} .$$

In full, Algorithm 10 gives an implementation of the conditional stratified resampling procedure where the conditioning is on $o^{(k)} \geq 1$. As in the unconditional case, the conditional stratified resampling procedure can follow a residual sampling step. The conditional stratified residual resampling procedure is given by Algorithm 11. In full, the CSMC procedure is given by Algorithm 12.

REMARK 7. *There is a slight technical issue with the conditional resampling schemes of 10 and 11 used in this thesis. Conditioning on a particle having a particular ancestor, via (80), is stronger than conditioning on that ancestor having at least one offspring. However, given that the unconditional stratified resampling scheme, and its residual extension, have very little flexibility in the number of offspring that can be assigned to each particle, as shown in Theorem 2.4.4, the distribution of offspring when using these schemes will be very similar, respectively, to*

Algorithm 10 Conditional Stratified Resampling (Conditioned on $o^{(k)} \geq 1$)

Shuffle the weights; $(\tilde{w}_\sigma^{(0:N)}, \sigma) = \text{shuffle}(\tilde{w}^{(0:N)})$.
Calculate the partial sums $s_j(\tilde{w}_\sigma^{(0:N)})$ for every $j = -1, \ldots, N$.
if $\tilde{w}^{(k)} \geq 2/(N+1)$ **then**
  for $i = 0, \ldots, N$ **do**
    Sample u_i from a $\text{Unif}(i/(N+1), (i+1)/(N+1))$ distribution.
  end for
  for $j = 0, \ldots, N$ **do**
    Set

$$o_\sigma^{(j)} = \sum_{i=0}^{N} \mathbb{1}_{I_j(\sigma)}(u_i) \, .$$

  end for
  Invert the shuffle on the offspring; $o^{(0:N)} = \text{inverse\_shuffle}(o_\sigma^{(0:N)}, \sigma)$.
else
  if $\tilde{w}^{(k)} \in [1/(N+1), 2/(N+1))$ **then**
    With probability

$$\frac{(N+1)s_{\sigma(k)}(\tilde{w}_\sigma^{(0:N)}) - \lfloor (N+1)s_{\sigma(k)}(\tilde{w}_\sigma^{(0:N)}) \rfloor}{(N+1)\tilde{w}_\sigma^{(\sigma(k))}}$$

  set $l = \lceil (N+1)s_{\sigma(k)}(\tilde{w}_\sigma^{(0:N)}) \rceil$. Else set $l = \lfloor (N+1)s_{\sigma(k)}(\tilde{w}_\sigma^{(0:N)}) \rfloor$.
  else
    Set $l = \lceil (N+1)s_{\sigma(k)}(\tilde{w}_\sigma^{(0:N)}) \rceil$.
  end if
  for $i = 0, \ldots, l-2, l, \ldots, N$ **do**
    Sample u_i from a $\text{Unif}(i/(N+1), (i+1)/(N+1))$ distribution.
  end for
  for $j = 0, \ldots, N$ **do**
    Set

$$o_\sigma^{(j)} = \sum_{\substack{i=1 \\ i \neq l}}^{N} \mathbb{1}_{I_j(\sigma)}(u_i) \, .$$

  end for
  Invert the shuffle on the offspring; $o^{(0:N)} = \text{inverse\_shuffle}(o_\sigma^{(0:N)}, \sigma)$.
  Set $o^{(k)} = o^{(k)} + 1$.
end if

Algorithm 11 Conditional Stratified Residual Resampling (Conditioned on $o^{(k)} \geq 1$)

Initialise by setting $s = 0$.
for $j = 0, \ldots, N$ **do**
  Set $o_b^{(j)} = \lfloor (N+1)\tilde{w}^{(j)} \rfloor$.
  Set $s = s + o_b^{(j)}$.
  Set $w_r^{(j)} = \tilde{w}^{(j)} - o_b^{(j)}/(N+1)$.
end for
Normalize the residual weights by setting, for each $j \in \{0, \ldots, N\}$,

$$\tilde{w}_r^{(j)} = w_r^{(j)}/(w_r^{(0)} + \ldots + w_r^{(N)}) .$$

if $o_b^{(k)} \geq 1$ **then**
  Resample, using the unconditional stratified resampling scheme, $N + 1 - S$ particles with weights $\tilde{w}_r^{(0:N)}$ to get offsprings $o_r^{(0:N)}$.
else
  Resample, using the conditional stratified resampling scheme conditioned on $o_r^{(k)} \geq 1$, $N + 1 - S$ particles with weights $\tilde{w}_r^{(0:N)}$ to get offsprings $o_r^{(0:N)}$.
end if
Set $o^{(j)} = o_b^{(j)} + o_r^{(j)}$ for all $j \in \{0, \ldots, N\}$.

Algorithm 12 Conditional Sequential Monte Carlo Procedure (Conditioned on the k-th path, $(k, \tilde{x}_T^{(k)}, \tilde{a}_{T-1}^{(k)})$)

for $i = 0, \ldots, \mathcal{L}_T(k,0) - 1, \mathcal{L}_T(k,0) + 1, \ldots, N$ **do**
  Sample $x_0^{(i)}$ with density $p_0(x_0^{(i)}|\theta)$ and set $\tilde{x}_0^{(i)} = x_0^{(i)}$.
  Calculate the i-th weight;

$$w_0(\tilde{x}_0^{(i)}; \theta) = \gamma_0(\theta, x_0^{(i)})/p_0(x_0^{(i)}|\theta) .$$

end for
Normalize the weights by setting, for each $i \in \{0, \ldots, N\}$,

$$\tilde{w}_0^{(i)}(\tilde{x}_0^{(0:N)}; \theta) = w_0(\tilde{x}_0^{(i)}; \theta)/(w_0(\tilde{x}_0^{(0)}; \theta) + \ldots + w_0(\tilde{x}_0^{(N)}; \theta)) .$$

for $t = 1, \ldots, T$ **do**
  Sample ancestors $a_{t-1}^{(-\mathcal{L}_T(k,t))}$ with mass function

$$\frac{\kappa(a_{t-1}^{(0:N)}|\tilde{w}_{t-1}^{(0:N)})}{\mathbb{P}(A_{t-1}^{(\mathcal{L}_T(k,t))} = a_{t-1}^{(\mathcal{L}_T(k,t))}|\tilde{w}_{t-1}^{(0:N)})} ,$$

  using the conditional stratified residual resampling scheme (Algorithm 11).
  for $i = 0, \ldots, \mathcal{L}_T(k,t) - 1, \mathcal{L}_T(k,t) + 1, \ldots, N$ **do**
    Sample $x_t^{(i)}$ with density $p_t(x_t|\tilde{x}_{t-1}^{(a_{t-1}^{(i)})}, \theta)$ and set $\tilde{x}_t^{(i)} := (\tilde{x}_{t-1}^{(a_{t-1}^{(i)})}, x_t^{(i)})$.
    Calculate the i-th weight;

$$w_t(\tilde{x}_t^{(i)}; \theta) = \frac{\gamma_t(\theta, \tilde{x}_t^{(i)})}{\gamma_{t-1}(\theta, \tilde{x}_{t-1}^{(a_{t-1}^{(i)})}) p_t(x_t^{(i)}|\tilde{x}_{t-1}^{(a_{t-1}^{(i)})}, \theta)} .$$

  end for
  Normalize the weights by setting, for each $i \in \{0, \ldots, N\}$,

$$\tilde{w}_t^{(i)}(\tilde{x}_t^{(0:N)}; \theta) = w_t(\tilde{x}_t^{(i)}; \theta)/(w_t(\tilde{x}_t^{(0)}; \theta) + \ldots + w_t(\tilde{x}_t^{(N)}; \theta)) .$$

end for

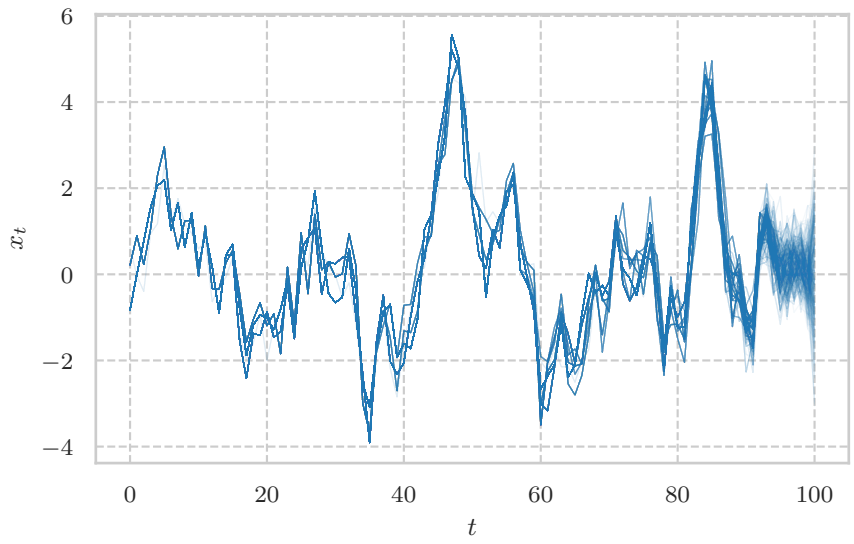

Figure 21: A plot of the five-hundred and one paths generated by the Conditional Sequential Monte Carlo procedure, which utilises the bootstrap proposal, applied to the one-dimensional Linear Gaussian model of Example 3.

the distribution of offspring when using the correctly conditioned stratified resampling and residual stratified resampling schemes.

The conditional resampling step of the CSMC procedure can result in candidate paths of the process coalescing backwards through time. To see this, consider the one-dimensional Linear Gaussian model of Example 3; that is, $X_0 \sim \mathrm{N}(0, 1)$, $\theta = 0.8$, and, for any $t \in \{1, \ldots, 100\}$, the transition distributions are given by $(X_t | X_{t-1} = x_{t-1}) \sim \mathrm{N}(\theta x_{t-1}, 1)$, and the observation distributions are given by $Y_t | X_t = x_t \sim \mathrm{N}(x_t, 0.3)$. Using the bootstrap proposal; that is, $p_0(x_0|\theta) = \phi(x_0; 0, 1)$, and, for any $t \in \{1, \ldots, T\}$, $p_t(x_t|x_{t-1}, \theta) = \phi(x_t; \theta x_{t-1}, 1)$, we ran the Conditional Sequential Monte Carlo procedure with $N = 500$ conditioned on a reference path, $x_{0:T}$ which was obtained by running the unconditional Sequential Monte Carlo procedure with ten-thousand particles and sampling one of the ten-thousand paths according to the terminal weights. Figure 21 shows the five-hundred and one paths generated by the Conditional Sequential Monte Carlo procedure. The figure illustrates the *coalescing backwards through time* behaviour of the procedure since, for the earlier observation times, the five-hundred and one paths *degenerate* into only three paths. In general, this *path degeneracy* characteristic is accentuated when the dimension of the path space is large, or, when the transition density in the mutation step is such that the weights derived in the correction step have a large variance (Pitt and Shephard, 1999; Doucet and Johansen, 2011; Lin, Chen, and Liu, 2013).

4.2.3 *Particle MH Samplers*

As shown in Section 4.2.1, the Pseudo-Marginal Metropolis-Hastings algorithm allows one to conduct inference for θ through an unbiased approximation $I_T(\theta, \tilde{X}_{0:T}^{(1:N)})$, to the likelihood $\eta_T(\theta)$. It can also be extended to permit inference for the path $x_{0:T}$. The key insight of Andrieu, Doucet, and Holenstein, 2010 is that the Sequential Importance Resampling estimator to $\pi_T[h]$ is a weighted sum over the paths simiulated by the procedure. A similar observation holds for the Sequential Monte Carlo estimator. Thus, one can extend the density given by (76) further by considering the weight of any particular path; that is, one can consider the extended density

$$\begin{aligned}&\pi_+(k, \theta, x_{0:T}^{(1:N)}, a_{0:T-1}^{(1:N)}) := \\ &\tilde{w}_T^{(k)}(\tilde{x}_T^{(1:N)}; \theta) \frac{I_T(\theta, \tilde{x}_{0:T}^{(1:N)})}{\eta_T} \Psi(x_{0:T}^{(1:N)}, a_{0:T-1}^{(1:N)} | \theta) \; . \end{aligned} \tag{81}$$

Summing over the N values of k gives the density given in (76). Hence, this is indeed a valid density. To show that such a density exhibits π_T as the marginal density for $(\theta, \tilde{x}_T^{(k)})$, it will first be useful to demonstrate the equivalence of this target with the extended target given in Andrieu, Doucet, and Holenstein, 2010:

LEMMA 4.2.4. *Under property (P) of Assumptions 4.2.1, the extended target given by (81), can be rewritten as*

$$\begin{aligned}&\pi_+(k, \theta, x_{0:T}^{(1:N)}, a_{0:T-1}^{(1:N)}) \\ &= N^{-(T+1)} \frac{\gamma_T(\theta, \tilde{x}_T^{(k)})}{\eta_T} \psi(x_{0:T}^{(1:N)} \backslash \tilde{x}_T^{(k)}, a_{0:T-1}^{(1:N)} \backslash \tilde{a}_{T-1}^{(k)} | k, \tilde{x}_T^{(k)}, \tilde{a}_{T-1}^{(k)}, \theta) \; , \end{aligned} \tag{82}$$

where ψ is the density corresponding to the Conditional Sequential Monte Carlo procedure (see Section 4.2.2).

Proof. See A.10. □

With this alternative representation for the extended density, it is trivial to see that such a density exhibits π_T as the marginal density for $(\theta, \tilde{x}_T^{(k)})$. Indeed, integrating out all the variables not involved in the path $\tilde{x}_T^{(k)}$ gives

$$\sum_{a_{0:T-1}^{(1:N)} \backslash \tilde{a}_{T-1}^{(k)}} \int_{\mathbb{R}_+} \pi_+(k, \theta, x_{0:T}^{(1:N)}, a_{0:T-1}^{(1:N)}) \, \mathrm{d}(x_{0:T}^{(1:N)} \backslash \tilde{x}_T^{(k)}) = \frac{\gamma_T(\theta, \tilde{x}_T^{(k)})}{\eta_T} = \pi_T(\theta, \tilde{x}_T^{(k)}) \; , \tag{83}$$

where, for notational simplicity, $\mathbb{R}_+ := \mathbb{R}^{d \times (T+1) \times (N-1)}$. The extended density given by (81) suggests that one can conduct inference for the path $X_{0:T}$ by running the SMC procedure to sample N *candidate*

paths and proposing one of these candidate paths with weight $\tilde{w}_T^{(k)}$. This is precisely the Particle Independent Sampler of Andrieu, Doucet, and Holenstein, 2010. Indeed, suppose the current state in the extended space is $(k, x_{0:T}^{(1:N)}, a_{0:T-1}^{(1:N)})$ and that θ is fixed. For notational simplicity, θ will be dropped from the notation that follows. Simulate a $(k^*, y_{0:T}^{(1:N)}, b_{0:T-1}^{(1:N)})$ with proposal density

$$q(k^*, y_{0:T}^{(1:N)}, b_{0:T-1}^{(1:N)}) = \tilde{w}_T^{(k^*)}(\tilde{y}_T^{(1:N)})\Psi(y_{0:T}^{(1:N)}, b_{0:T-1}^{(1:N)}) ,$$

by simulating N paths, $(y_{0:T}^{(1:N)}, b_{0:T-1}^{(1:N)})$, from Ψ using the SMC procedure (Algorithm 8), and choosing the k-th of those paths according to the normalised terminal weight $\tilde{w}_T^{(k^*)}(\tilde{y}_T^{(1:N)})$. The Metropolis-Hastings acceptance probability, which leads to the Particle Independent Metropolis-Hastings (PIMH) sampler, is given by

$$\begin{aligned}
&1 \wedge \frac{\pi_+(k^*, y_{0:T}^{(1:N)}, b_{0:T-1}^{(1:N)})q(k, x_{0:T}^{(1:N)}, a_{0:T}^{(1:N)})}{\pi_+(k, x_{0:T}^{(1:N)}, a_{0:T-1}^{(1:N)})q(k^*, y_{0:T}^{(1:N)}, b_{0:T}^{(1:N)})} \\
&= 1 \wedge \frac{\tilde{w}_T^{(k^*)}(\tilde{y}_T^{(1:N)})I_T(\theta, \tilde{y}_{0:T}^{(1:N)})\Psi(y_{0:T}^{(1:N)}, b_{0:T-1}^{(1:N)})\tilde{w}_T^{(k)}(\tilde{x}_T^{(1:N)})\Psi(x_{0:T}^{(1:N)}, a_{0:T-1}^{(1:N)})}{\tilde{w}_T^{(k)}(\tilde{x}_T^{(1:N)})I_T(\tilde{x}_{0:T}^{(1:N)})\Psi(x_{0:T}^{(1:N)}, a_{0:T-1}^{(1:N)})\tilde{w}_T^{(k^*)}(\tilde{y}_T^{(1:N)})\Psi(y_{0:T}^{(1:N)}, b_{0:T-1}^{(1:N)})} \\
&= 1 \wedge \frac{I_T(\tilde{y}_{0:T}^{(1:N)})}{I_T(\tilde{x}_{0:T}^{(1:N)})} .
\end{aligned}$$

Given an initial path $\tilde{x}_{0:T}$ from which to start the algorithm, one can use the representation of the extended target density, given by (82) in Lemma 4.2.4; and, in particular, the Conditional Sequential Monte Carlo (CSMC) procedure (see Andrieu, Doucet, and Holenstein, 2010, or Section 4.2.2), whose density is given by (79), to simulate the other $N-1$ initial, candidate, paths and calculate the initial approximation to the likelihood; $I_T(\tilde{x}_{0:T}^{(1:N)})$. In full, assuming a fixed θ, the Particle Independent Metropolis-Hastings (PIMH) Sampler is given by Algorithm 13.

It is trivial to extend such an algorithm to conduct inference for $\pi_T(\theta, x_{0:T})$ where θ is not fixed. Indeed, if one constructs a proposal $q_\Theta(\cdot|\theta)$ for θ— see, for example, the independent and random-walk proposals of Section 2.3.5— then one can target the extended density $\pi_+(k, \theta, x_{0:T}^{(1:N)}, a_{0:T-1}^{(1:N)})$ given by (81) by proposing from

$$q(k^*, \theta^*, y_{0:T}^{(1:N)}, b_{0:T-1}^{(1:N)}|\theta) = \tilde{w}_T^{(k^*)}(\tilde{y}_T^{(1:N)}; \theta^*)\Psi(y_{0:T}^{(1:N)}, b_{0:T-1}^{(1:N)}|\theta^*)q_\Theta(\theta^*|\theta) ,$$

given a current state $(k, \theta, x_{0:T}^{(1:N)}, a_{0:T-1}^{(1:N)})$, and accepting with probability

$$1 \wedge \frac{I_T(\tilde{y}_{0:T}^{(1:N)})q(\theta|\theta^*)}{I_T(\tilde{x}_{0:T}^{(1:N)})q(\theta^*|\theta)} .$$

This is the Particle Marginal Metropolis-Hastings (PMMH) Sampler of Andrieu, Doucet, and Holenstein, 2010, which, in full, is given by Algorithm 14.

Algorithm 13 Particle Independent Metropolis-Hastings Sampler

Initialise the chain at some $x_{0:T} \in \mathbb{R}^{d(T+1)}$ and choose the number of iterations, M.
Let $k = 1$, $a_t^{(1)} = 1$ for all $t \in \{0, \ldots, T-1\}$, and $x_t^{(1)} = x_t$ for all $t \in \{0, \ldots, T\}$ so that $\tilde{x}_T^{(1)} = x_{0:T}$. Define $x_0^{\text{path}} := \tilde{x}_T^{(1)}$.
Sample $(x_{0:T}^{(2:N)}, a_{0:T-1}^{(2:N)})$ with density $\psi(\cdot|1, \tilde{x}_T^{(1)}, \tilde{a}_{T-1}^{(1)})$ defined by (79) via the CSMC procedure (see Algorithm 12).
Calculate $\tilde{I}_0 := I_T(\tilde{x}_{0:T}^{(1:N)})$ as defined by (75).
for $m = 0, \ldots, M-1$ **do**
    Sample $(y_{0:T}^{(1:N)}, b_{0:T}^{(1:N)})$ with density $\Psi(\cdot)$ defined by (72) via the SMC procedure given by Algorithm 8.
    Calculate $\tilde{I}^* := I_T(\tilde{y}_{0:T}^{(1:N)})$ as defined by (75).
    Sample a $k^* \in \{1, \ldots, N\}$ with probability $\tilde{w}_T^{(k^*)}(\tilde{y}_T^{(1:N)})$.
    Calculate the MH acceptance probability;

$$\alpha(x_m^{\text{path}}, \tilde{y}_T^{(k^*)}) := 1 \wedge \frac{\tilde{I}^*}{\tilde{I}_m} .$$

    With probability $\alpha(x_m^{\text{path}}, \tilde{y}_T^{(k^*)})$ set $x_{m+1}^{\text{path}} = \tilde{y}_T^{(k^*)}$, $\tilde{I}_{m+1} = \tilde{I}^*$; else set $x_{m+1}^{\text{path}} = x_m^{\text{path}}$, $\tilde{I}_{m+1} = \tilde{I}_m$.
end for

Algorithm 14 Particle Marginal Metropolis-Hastings Sampler

Initialise the chain at some $(\theta_0, x_{0:T}) \in \mathbb{R}^p \times \mathbb{R}^{d(T+1)}$ and choose the number of iterations, M.
Let $k = 1$, $a_t^{(1)} = 1$ for all $t \in \{0, \ldots, T-1\}$, and $x_t^{(1)} = x_t$ for all $t \in \{0, \ldots, T\}$ so that $\tilde{x}_T^{(1)} = x_{0:T}$. Define $x_0^{\text{path}} := \tilde{x}_T^{(1)}$.
Sample $(x_{0:T}^{(2:N)}, a_{0:T-1}^{(2:N)})$ with density $\psi(\cdot|1, \tilde{x}_T^{(1)}, \tilde{a}_{T-1}^{(1)}, \theta_0)$ defined by (79) via the CSMC procedure (see Algorithm 12).
Calculate $\tilde{I}_0 := I_T(\theta_0, \tilde{x}_{0:T}^{(1:N)})$ as defined by (75).
for $m = 0, \ldots, M-1$ **do**
    Sample θ^* with density $q(\cdot|\theta_m)$
    Sample $(y_{0:T}^{(1:N)}, b_{0:T}^{(1:N)})$ with density $\Psi(\cdot|\theta^*)$ defined by (72) via the SMC procedure given by Algorithm 8.
    Calculate $\tilde{I}^* := I_T(\theta^*, \tilde{y}_{0:T}^{(1:N)})$ as defined by (75).
    Sample a $k^* \in \{1, \ldots, N\}$ with probability $\tilde{w}_T^{(k^*)}(\tilde{y}_T^{(1:N)}; \theta^*)$.
    Calculate the MH acceptance probability;

$$\alpha(\theta_m, x_m^{\text{path}}, \theta^*, \tilde{y}_T^{(k^*)}) := 1 \wedge \frac{\tilde{I}^* q(\theta|\theta^*)}{\tilde{I}_m q(\theta^*|\theta)} .$$

    With probability $\alpha(\theta_m, x_m^{\text{path}}, \theta^*, \tilde{y}_T^{(k^*)})$ set $\theta_{m+1} = \theta^*$, $x_{m+1}^{\text{path}} = \tilde{y}_T^{(k^*)}$, $\tilde{I}_{m+1} = \tilde{I}^*$; else set $\theta_{m+1} = \theta_m$, $x_{m+1}^{\text{path}} = x_m^{\text{path}}$, $\tilde{I}_{m+1} = \tilde{I}_m$.
end for

For formal results relating to the ergodicity of the aforementioned Particle MH Samplers see Andrieu, Doucet, and Holenstein, 2010. As was the case for the PsMMH sampler, the results of Section 2.3.6, in particular, Theorems 2.3.30, 2.3.31, and 2.3.34, are valid when considering the chain on the extended state space since the samplers in the extended space are propose and accept-reject MCMC algorithms targeting an extended density. Moreover, ergodictiy, as given by Corollary 2.3.12, of the *marginal chain*- that is, the chain that is induced by just considering how $(\theta, x^{\text{path}})$ transitions- follows from ergodicity of the *extended chain*- that is, the chain created by the propose and accept-reject MCMC sampler on the extended space- since, intuitively, the chance of moving to a state $(\theta^*, \tilde{y}_T^{(k)})$ in the marginal space is greater than the chance of moving to a state $(\theta^*, y_{0:T}^{(1:N)}, b_{0:T-1}^{(1:N)})$, for any particular $(y_{0:T}^{(1:N)} \backslash \tilde{y}_T^{(k)}, b_{0:T-1}^{(1:N)} \backslash \tilde{b}_{T-1}^{(k)})$, in the extended space. Furthermore, again, as in the PsMMH case, uniform ergodicity of the *marginal chain* follows from uniform ergodicity of the *extended chain* since the bounding term of the total variation distance for uniformly ergodic chains- see Definition 2.3.14- is independent of where the chain starts. For the PIMH sampler the chain in the extended space is uniformly ergodic if π_+/q over the extended space is uniformly bounded (Theorem 2.3.31). Therefore, with a fixed $\theta \in \mathbb{R}^p$, the PIMH Sampler is uniformly ergodic if, for fixed, finite T, and any $t \in \{0, \ldots, T\}$,

$$\sup_{x_{0:t} \in \mathbb{R}^{d(t+1)}} w_t(x_{0:t}; \theta) \leq C_t(\theta) < \infty \,, \tag{84}$$

for some $C_t(\theta)$. Moreover, for the PMMH sampler, where the proposal for θ^* is taken to be an independent proposal; that is, $q_\Theta(\cdot|\theta) = q_\Theta(\cdot)$, the chain is uniformly ergodic if, for fixed, finite T, and any $t \in \{0, \ldots, T\}$, Condition (84) holds and

$$\sup_{\theta \in \mathbb{R}^p} \frac{C_t(\theta)}{q_\Theta(\theta)} < \infty \,.$$

However, the same complexities as those described for the PsMMH Sampler hold when considering the inheritance of geometric ergodicity; that is, since the bound on the total variation distance for geometrically ergodic chains depends on where the chain was initialised, it is not necessarily the case, without some extra assumptions, that geometric ergodicity of the *marginal chain* follows from geometric ergodicity of the *extended chain*.

To demonstrate the Particle Independent Metropolis-Hastings Sampler, consider the Linear Gaussian model of Example 2 with a fixed θ:

EXAMPLE 3. *Let* $X_0 \sim \mathrm{N}(0, 1)$, $\theta = 0.8$, *and suppose that, for any* $t \in \{1, \ldots, 100\}$, *the transition distributions are given by* $(X_t | X_{t-1} = x_{t-1}) \sim \mathrm{N}(\theta x_{t-1}, 1)$, *and the observation distributions are given by* $Y_t | X_t = x_t \sim \mathrm{N}(x_t, 0.3)$. *Then,* $\gamma_0(x_0) = \phi(x_0; 0, 1)$, *and, for any* $t \in \{1, \ldots, T\}$, γ_t *is defined recursively by*

$$\gamma_t(x_{0:t}) = g_t(y_t|x_t)\phi(x_t; \theta x_{t-1}, 1)\gamma_{t-1}(x_{0:t-1}) \,,$$

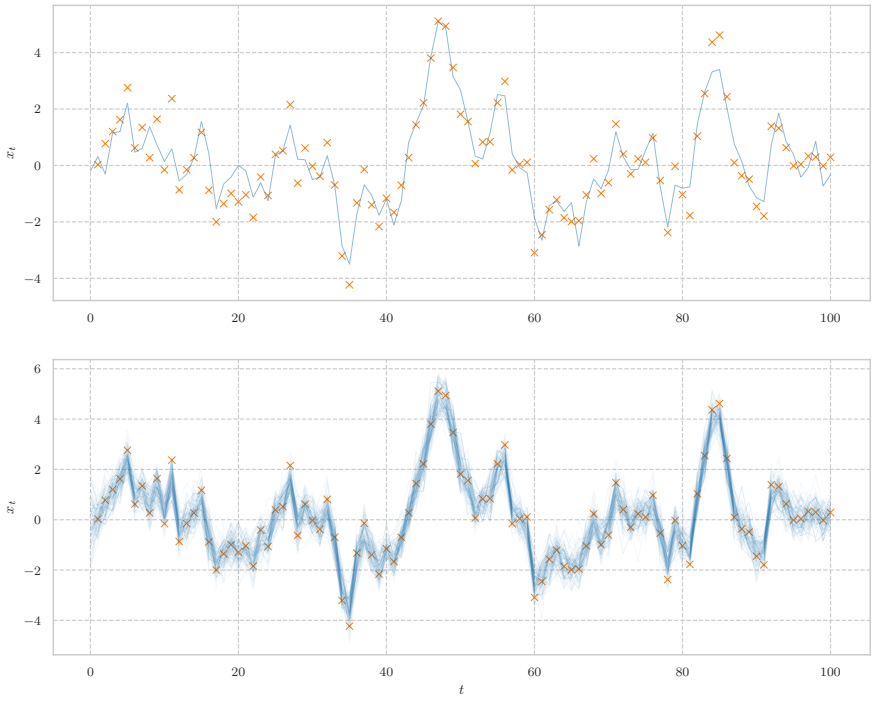

Figure 22: A plot of the path which generated the observations and the observations themselves (top row), along with a plot of one-hundred *thinned* simulated paths from the PIMH and the observations (bottom row). The simulated paths have been made semi-transparent for ease of visualisation.

where $\phi(\cdot; \mu, \sigma^2)$ *denotes the density of a one-dimensional normal distribution with mean* μ *and variance* σ^2*.*

Using the bootstrap proposal; that is, $p_0(x_0|\theta) = \phi(x_0; 0, 1)$, and, for any $t \in \{1, \ldots, T\}$, $p_t(x_t|x_{t-1}, \theta) = \phi(x_t; \theta x_{t-1}, 1)$, we ran the PIMH for ten-thousand iterations with $N = 100$. Figure 22 shows one-hundred *thinned* simulated paths from the PIMH alongside the true observations (bottom row), in comparison to the path which generated the observations (top row). Figure 23, on the other hand, shows histograms of the simulated samples for several points in time at which the observations occur ($t \in \{25, 50, 75, 100\}$), along with the true target density at these points in time. The figures demonstrate that the PIMH works as expected in this case. Indeed, Figure 22 highlights that the thinned simulated paths align with the observations and the dynamics of a *true* simulated path. Moreover, Figure 23 shows that the samples generated by the PIMH form a good empirical approximation to the true density; at least when viewed marginally at several points in time at which the observations occur.

The drawback of the Particle Marginal Metropolis-Hastings Sampler is that, at each iteration, the candidate paths, $\tilde{y}_{0:T}^{(1:N)}$, of the next state of the chain are simulated via the SMC procedure (Algorithm 8) independently of the current path of the chain; that is, independently of x_m^{path}. Therefore, the chain can get *stuck* on paths with a relatively

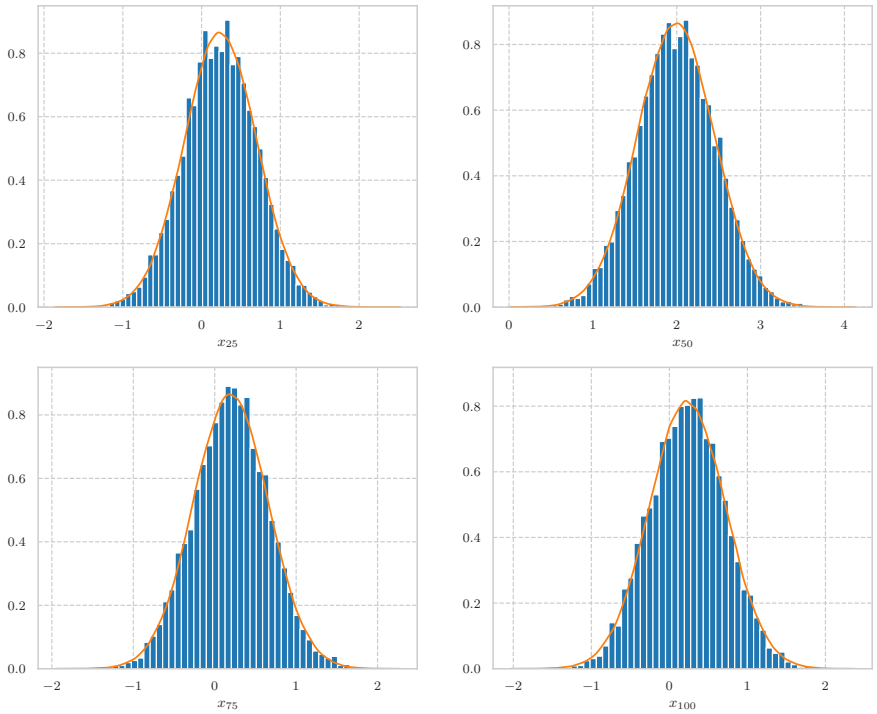

Figure 23: A plot of the histograms of the samples generated by the PIMH for several points in time at which the observations occurred ($t \in \{25, 50, 75, 100\}$), along with the true target density at these points in time.

large weight which results in poor mixing of the chain. The Correlated Pseudo-Marginal Markov Chain Monte Carlo (CPsMMCMC) approach attempts to overcome this *sticky* behaviour by correlating *all* the underlying random variables involved in the SMC procedure- including those needed during the resampling steps, thereby attempting to simulate candidate paths which are *close* to the reference path (Deligiannidis, Doucet, and Pitt, 2015; Dahlin et al., 2015; Murray and Graham, 2016). While such an approach can be effective in practice, the limiting justification relies on the *smoothness* of the SMC estimator considered as a function of *all* the underlying random variables. This condition is, in general, difficult to justify since the resampling steps in the SMC procedure are necessarily discontinuous. Moreover, while the use of a Hilbert sort procedure, as described in Deligiannidis, Doucet, and Pitt, 2015, can make the variability induced by the resampling step smaller, it still does not guarantee smoothness of the approximation and, just as importantly, is computationally costly to implement.

4.2.4 *The Particle Gibbs Sampler*

As discussed in the last Section, the Particle Marginal Metropolis-Hastings algorithm of Andrieu, Doucet, and Holenstein, 2010 can be utilised to conduct inference for the full joint target density $\pi_T(\theta, x_{0:T})$.

However, as highlighted at the end of the last section, such an algorithm can mix poorly due to the fact that the candidate paths are simulated independently of the current path of the chain. While the Correlated Pseudo-Marginal Markov Chain Monte Carlo method (Deligiannidis, Doucet, and Pitt, 2015; Dahlin et al., 2015; Murray and Graham, 2016) can work well in practice, it relies on correlating *all* the random variables involved in the Sequential Monte Carlo procedure, including those involved in the resampling steps. The discontinuity of the resampling steps means that, in general, the SMC estimator, considered as a function of the underlying random variables, is not smooth, thus invalidating the formal justification of such an approach. The Particle Gibbs Sampler (PGS) of Andrieu, Doucet, and Holenstein, 2010, on the other hand, attempts to overcome the poor mixing behaviour of the PMMH Sampler by considering the path corresponding to the current state of the chain when simulating candidate paths for the next state of the chain. Indeed, the PGS *mimics* an *idealized* Gibbs Sampler by alternating between sampling the parameters, θ, given a *path*, x_m^{path}, and, sampling a path given the parameters. The former step is achieved by sampling a θ from the joint target, $\pi_T(\theta, x_m^{\text{path}})$, conditioned on the path x_m^{path}; that is, sample a θ with density $\pi_T^{(\theta)}(\cdot|x_m^{\text{path}})$. The latter step relies on the Conditional Sequential Monte Carlo procedure (see Section 4.2.2 and Algorithm 12), which, given a *reference* path, simulates N candidate paths, along with corresponding weights, using a particle filter with $N+1$ particles which has been *conditioned* on including the reference path as one of the $N+1$ paths simulated. As was the case for the Particle Marginal Metropolis-Hastings algorithm of Andrieu, Doucet, and Holenstein, 2010, one chooses one of the $N+1$ paths with probability proportional to the final weight. The difference here is that because the current path is included in the $N+1$ candidate paths there is no acceptance step. Note that, as is the case for any Gibbs Sampler, in the situation where it is difficult to sample a θ given a path, x_m^{path}, the first step of the Particle Gibbs Sampler can be replaced by any valid MCMC step. For instance, one could replace it with a Metropolis-Hastings propose and accept-reject step. In full, the Particle Gibbs Sampler of Andrieu, Doucet, and Holenstein, 2010 is given by Algorithm 15. Formally, the sampler consists of a sequence of Gibbs steps targeting the extended density $\pi_+(k_m, \theta_m, y_{0:T}^{(0:N)}, b_{0:T-1}^{(0:N)})$ given by (81), and, therefore, has this extended density (which exhibits the target π_T as the marginal density) as the stationary density of the resulting chain. To see this, consider the two forms of the extended density, given by (81) and (82) respectively:

$$\begin{aligned}
&\pi_+(k_m, \theta_m, y_{0:T}^{(0:N)}, b_{0:T-1}^{(0:N)}) \\
&= \tilde{w}_T^{(k_m)}(\tilde{y}_T^{(0:N)}; \theta_m) \frac{I_T(\theta_m, \tilde{y}_{0:T}^{(0:N)})}{\eta_T} \Psi(y_{0:T}^{(0:N)}, b_{0:T-1}^{(0:N)}|\theta_m) \\
&= (N+1)^{-(T+1)} \frac{\gamma_T(\theta, \tilde{y}_T^{(k_m)})}{\eta_T} \psi(y_{0:T}^{(0:N)} \backslash \tilde{y}_T^{(k_m)}, b_{0:T-1}^{(0:N)} \backslash \tilde{b}_{T-1}^{(k_m)} | k_m, \tilde{y}_T^{(k_m)}, \tilde{b}_{T-1}^{(k_m)}, \theta_m) \, .
\end{aligned}$$

Algorithm 15 Particle Gibbs Sampler

Initialise the chain at some $(\theta_0, x_{0:T}) \in \mathbb{R}^p \times \mathbb{R}^{d(T+1)}$ and choose the number of iterations, M.
Let $a_t^{(0)} = 0$ for all $t \in \{0, \ldots, T-1\}$, and $x_t^{(0)} = x_t$ for all $t \in \{0, \ldots, T\}$ so that $\tilde{x}_T^{(0)} = x_{0:T}$. Define $x_0^{\text{path}} := \tilde{x}_T^{(0)}$, $a_0^{\text{path}} := \tilde{a}_T^{(0)}$, and $k_0 = 0$.
for $m = 0, \ldots, M-1$ **do**
    Set $\tilde{b}_T^{(k_m)} = a_m^{\text{path}}$ and $\tilde{y}_T^{(k_m)} = x_m^{\text{path}}$.
    Sample θ_{m+1} with density $\pi_T^{(\theta)}(\cdot|\tilde{y}_T^{(k_m)})$.
    Sample $(y_{0:T}^{(0:N)} \backslash \tilde{y}_T^{(k_m)}, b_{0:T-1}^{(0:N)} \backslash \tilde{b}_{T-1}^{(k_m)})$ with density

$$\psi(y_{0:T}^{(0:N)} \backslash \tilde{y}_T^{(k_m)}, b_{0:T-1}^{(0:N)} \backslash \tilde{b}_{T-1}^{(k_m)} | k_m, \tilde{y}_T^{(k_m)}, \tilde{b}_{T-1}^{(k_m)}, \theta_{m+1}) ,$$

    defined by (79) via the CSMC procedure (see Algorithm 12).
    Sample a $k_{m+1} \in \{0, \ldots, N\}$ with probability $\tilde{w}_T^{(k_{m+1})}(\tilde{y}_T^{(0:N)}; \theta_{m+1})$.
    Set $x_{m+1}^{\text{path}} = \tilde{y}_T^{(k_{m+1})}$ and $a_{m+1}^{\text{path}} = \tilde{b}_T^{(k_{m+1})}$.
end for

As shown in Algorithm 15, given a current state, $(k_m, \theta_m, \tilde{y}_T^{(k_m)}, \tilde{b}_{T-1}^{(k_m)})$, in the extended space, the Particle Gibbs Sampler cycles through the following steps:

1. Sample a θ_{m+1} from $\pi_T^{(\theta)}(\cdot|\tilde{y}_T^{(k_m)})$.
2. Sample a sequence $(y_{0:T}^{(0:N)} \backslash \tilde{y}_T^{(k_m)}, b_{0:T-1}^{(0:N)} \backslash \tilde{b}_{T-1}^{(k_m)})$ with density

$$\psi(y_{0:T}^{(0:N)} \backslash \tilde{y}_T^{(k_m)}, b_{0:T-1}^{(0:N)} \backslash \tilde{b}_{T-1}^{(k_m)} | k_m, \tilde{y}_T^{(k_m)}, \tilde{b}_{T-1}^{(k_m)}, \theta_{m+1}) ,$$

   defined by (79) via the CSMC procedure (see Algorithm 12).
3. Sample a $k_{m+1} \in \{0, \ldots, N\}$ with probability $\tilde{w}_T^{(k_{m+1})}(\tilde{y}_T^{(0:N)}; \theta_{m+1})$.

Note, by (83), that the extended target exhibits π_T as the marginal density for $(\theta, \tilde{y}_T^{(k_m)})$. Thus, using the terminology of Liu, 2001, Section 6.7, the first step is a *collapsed* Gibbs step. The second step is a Gibbs step on the extended space as can be seen from the second representation of the extended target:

$$\begin{aligned}
&\pi_+(k, \theta, y_{0:T}^{(0:N)}, b_{0:T-1}^{(0:N)}) \\
&= (N+1)^{-(T+1)} \frac{\gamma_T(\theta, \tilde{y}_T^{(k)})}{\eta_T} \psi(y_{0:T}^{(0:N)} \backslash \tilde{y}_T^{(k)}, b_{0:T-1}^{(0:N)} \backslash \tilde{b}_{T-1}^{(k)} | k, \tilde{y}_T^{(k)}, \tilde{b}_{T-1}^{(k)}, \theta) .
\end{aligned}$$

The third step is also a Gibbs step on the extended space as can be seen from the first representation of the extended target:

$$\begin{aligned}
&\pi_+(k, \theta, y_{0:T}^{(0:N)}, b_{0:T-1}^{(0:N)}) \\
&= \tilde{w}_T^{(k)}(\tilde{y}_T^{(0:N)}; \theta) \frac{I_T(\theta, \tilde{y}_{0:T}^{(0:N)})}{\eta_T} \Psi(y_{0:T}^{(0:N)}, b_{0:T-1}^{(0:N)} | \theta) .
\end{aligned}$$

Theorem 5 of Andrieu, Doucet, and Holenstein, 2010 demonstrates that, if each step of the Particle Gibbs Sampler is irreducible and aperiodic, then the Particle Gibbs Sampler is ergodic in the sense of Corollary

2.3.12. As such, by Theorem 2.3.13, the MCMC estimates corresponding to the samples generated by the Particle Gibbs Sampler satisfy a Strong Law of Large Numbers result. In addition, Corollary 2 of Lindsten, Douc, and Moulines, 2015, and Theorem 1 of Andrieu, Lee, and Vihola, 2018, essentially demonstrate that the sampler is uniformly ergodic (Definition 2.3.14), if, at each time $t \in \{0, \ldots, T\}$,

$$\sup_{(x_{0:t},\theta)\in\mathbb{R}^{d\times t}\times\mathbb{R}^p} w_t(x_{0:t};\theta) < \infty \,. \tag{85}$$

They do this by showing that, under such an assumption, the chain induced by the sampler satisfies a minorization condition (Definition 2.3.16) on the whole state space and, therefore, that the entire state space is small; which, by Theorem 2.3.17, is sufficient to prove uniform ergodicity of the sampler[3]. This strong result demonstrates that, under such an assumption, the difference between the chain's t-step transition distributions and the target can be bounded uniformly by a term that is independent of where the chain started, and, that decays geometrically. Indeed, Theorem 2.3.17 gives this rate in terms of the minorization constant which is analysed in depth in both Lindsten, Douc, and Moulines, 2015 and Andrieu, Lee, and Vihola, 2018. As a result, not only does such a chain produce MCMC estimates which satisfy a Central Limit Theorem— by Theorem 2.3.22— but one can choose the number of iterations to run the chain for such that the difference to stationarity is bounded by a given threshold. Furthermore, Theorem 1 of Andrieu, Lee, and Vihola, 2018 shows that condition (85) is necessary for the sampler to be geometrically ergodic; that is, the Particle Gibbs Sampler can not be geometrically ergodic if

$$\sup_{(x_{0:t},\theta)\in\mathbb{R}^{d\times t}\times\mathbb{R}^p} w_t(x_{0:t};\theta) = \infty \,.$$

These necessary and sufficient results align closely to the results derived for the Independence Sampler; in particular, the results of Theorem 2.3.31, which states that the Independence Sampler is uniformly ergodic if the independence weight, $w(x) = \pi(x)/q(x)$, is such that

$$\sup_{x\in\mathbb{R}^d} w(x) < \infty \,,$$

and that this condition is necessary for the sampler to be geometrically ergodic.

While such results give guarantees on the convergence of the Particle Gibbs Sampler, they do not give guidance on how to choose the the number of particles, N. In particular, it is practically useful to know how N should depend on T to get sufficiently good *mixing* without choosing N too large. Clearly, the larger the value of T, the more *information* the sampler has to infer, and so it is reasonable to believe that N must scale with T somehow. However, choosing N too big means wasting computational effort and increasing run-times. On the other hand, choosing N

3 Theorem 3, Chopin, 2004 provides another proof that the Particle Gibbs Sampler is uniformly erogodic under slightly stronger assumptions.

too small may potentially result in poor mixing. In the latter case the sampler would have to be run for more iterations to produce samples that *accurately* represent the target, and, this increase in the number of iterations would, again, increase the computational effort and runtimes. Propositions 4 and 5, Lindsten, Douc, and Moulines, 2015, and Theorem 3 of Andrieu, Lee, and Vihola, 2018 demonstrate that, under suitable *strong-mixing* conditions, it is sufficient to scale the number of particles, N, linearly with T in order to obtain a non-degenerate lower-bound on the minorizing constant in the limit as $T \to \infty$. In Theorem 6, Lindsten, Douc, and Moulines, 2015 show that, under weaker, *moment* conditions, the minorization constant is *probabilistically* bounded below in the limit as $T \to \infty$ provided one scales the number of particles, N, superlinearly with T— see Lindsten, Douc, and Moulines, 2015 for a detailed statement of this result. Practically speaking, this suggests, as one would intuit, that the number of particles required to achieve a sampler with good rates of mixing depends on the problem at hand, and, ultimately, on how well-behaved the SMC weights,

$$w_t(\tilde{x}_t^{(i)}; \theta) = \frac{\gamma_t(\theta, \tilde{x}_t^{(i)})}{\gamma_{t-1}(\theta, \tilde{x}_{t-1}^{(a_{t-1}^{(i)})}) p_t(x_t^{(i)} | \tilde{x}_{t-1}^{(a_{t-1}^{(i)})}, \theta)} ,$$

are.

To demonstrate the Particle Gibbs Sampler (PGS), consider, again, the one-dimensional Linear Gaussian model of Example 3; that is, $X_0 \sim \mathrm{N}(0, 1)$, $\theta = 0.8$, and, for any $t \in \{1, \ldots, 100\}$, the transition distributions are given by $(X_t | X_{t-1} = x_{t-1}) \sim \mathrm{N}(\theta x_{t-1}, 1)$, and the observation distributions are given by $Y_t | X_t = x_t \sim \mathrm{N}(x_t, 0.3)$. Using the bootstrap proposal; that is, $p_0(x_0|\theta) = \phi(x_0; 0, 1)$, and, for any $t \in \{1, \ldots, T\}$, $p_t(x_t | x_{t-1}, \theta) = \phi(x_t; \theta x_{t-1}, 1)$, we ran the PGS for ten-thousand iterations with $N = 100$. Figure 24 shows one-hundred *thinned* simulated paths from the PGS alongside the true observations (bottom row), in comparison to the path which generated the observations (top row). Figure 25, on the other hand, shows histograms of the simulated samples for several points in time at which the observations occur ($t \in \{25, 50, 75, 100\}$), along with the true target density at these points in time. The figures demonstrate that, for this example, although the PGS produces sample paths which are an okay empirical approximation to the true density— at least when viewed marginally at several points in time at which the observations occur— the samples do not represent as good an approximation as those produced by the Particle Independent Metropolis-Hastings— as can be seen by comparing with Figures 22 and 23. In particular, Figure 25, when compared with Figure 23, highlights that the empirical approximation to the true marginal density at each $t \in \{1, \ldots, 100\}$ provided by the simulated samples is worse the smaller t is. This behaviour stems from the fact that, as highlighted at the end of Section 4.2.2, the conditional resampling step of the Conditional Sequential Monte Carlo procedure can result in candidate paths of the process coalescing backwards through time; a feature that can be seen clearly in Figure 24. Although, in this

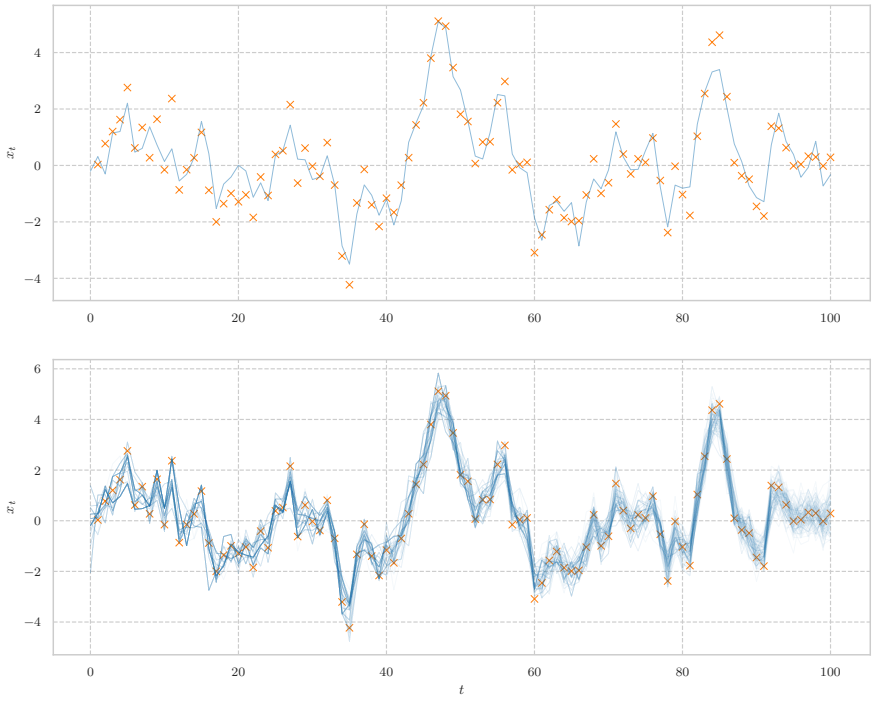

Figure 24: A plot of the path which generated the observations and the observations themselves (top row), along with a plot of one-hundred *thinned* simulated paths from the PGS and the observations (bottom row). the simulated paths have been made semi-transparent for ease of visualisation.

simple case, the sampler mixes fairly well, in general this *path degeneracy* characteristic can ultimately result in a PGS which mixes poorly (Lindsten and Schön, 2013; Lindsten, Jordan, and Schön, 2014; Chopin and Singh, 2015).

The Particle Gibbs with Ancestor Sampling (PGAS) approach of Lindsten, Jordan, and Schön, 2014 (see also Lindsten et al., 2015) attempts to overcome the path degeneracy problem of the PGS by introducing an ancestor sampling step into the PGS, which, at each time step, samples a new *history* of the reference path, thus allowing the proposed paths to degenerate to a path which is different from the reference path. By allowing the *degeneration* path to differ from the reference path, the PGAS algorithm can achieve much better mixing than the PGS (Lindsten, Jordan, and Schön, 2014). Unfortunately, the ancestral sampling step relies on being able to calculate the *likelihood* of the reference path having a particular *history* which makes the PGAS impossible to implement in scenarios where this *likelihood* is intractable. Moreover, if the weights of the particles have large variability, or, if the model is such that the *likelihood* of the reference path having a *history* which is not the same as the current history of the reference path is relatively small, then the PGAS will offer little improvement over the PGS (Lindsten et al., 2015). This makes the PGAS particularly ill-suited to conducting inference for diffusions, since; firstly, the likelihood cor-

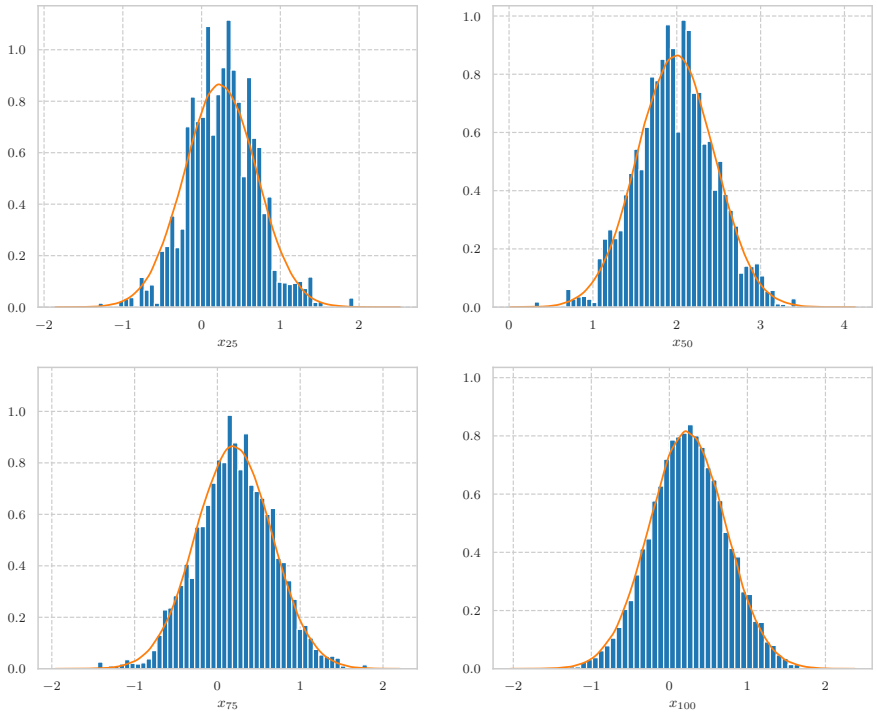

Figure 25: A plot of the histograms of the samples generated by the PGS for several points in time at which the observations occurred ($t \in \{25, 50, 75, 100\}$), along with the true target density at these points in time.

responding to a particle that was propagated according to one of the proposals introduced in Chapter 3 is intractable and computationally costly to approximate, and, secondly, as highlighted in Chapter 3, the variance of the weights corresponding to the proposals introduced in Chapter 3 can be very large. While the rejuvenation approach of Lindsten et al., 2015 overcomes these limitations, it does so at the expense of a significant increase in computational cost.

4.3 THE EXCHANGEABLE SAMPLER

In a similar spirit to the Correlated Pseudo-Marginal Markov Chain Monte Carlo method of Deligiannidis, Doucet, and Pitt, 2015, Dahlin et al., 2015, and Murray and Graham, 2016, we attempt to overcome the path degeneracy problem of the Particle Gibbs Sampler and the limitations, as highlighted at the end of the previous section, of the Particle Gibbs with Ancestor Sampling approach of Lindsten, Jordan, and Schön, 2014, and Lindsten et al., 2015 by making the proposed paths *closer* to the reference path being conditioned upon, and, therefore, make it more likely that the chain moves to a path which is different from the reference path. This is done by simulating particles within the Sequential Monte Carlo procedure exchangeably as opposed to independently. By only correlating the random variables associated

with the propagation of the particles, and not the random variables associated with the resampling steps, we avoid the practically unrealistic assumptions necessary for the limiting justification of the CPsMMCMC algorithm.

Before introducing the Exchangeable Particle Gibbs Sampler, we start, in this section, by introducing the Exchangeable Sampler which is a generalisation of the Independence Sampler of Section 2.3.6.1. As the name suggests, the Exchangeable Sampler generalisation is obtained by proposing a sample conditional on the current state of the chain in an exchangeable way (see Section 2.1.2 for an introduction to exchangeability); that is, in such a way that the joint density of the proposal and the current state is symmetric, while still emitting a proposal density q as *the* marginal density[4]. As highlighted in Section 2.1.2, and detailed in Algorithm 2 and the discussion thereafter, exchangeability allows one to propose a state which is *close* to the current state, where, as with the Random-Walk Sampler, the *closeness* of the proposal is adjustable through a tunable *jump-size*. Moreover, the user is free to construct a marginal proposal, q, which matches the structure of the target in cases where this is known reasonably well. Furthermore, as one shall see, the *weight of the transition* is independent of the current state of the chain, thus allowing one to easily extend the Exchangeable Sampler to a multiple-proposal regime, as was the case for the Independence Sampler. The Exchangeable Sampler, therefore, shares the same advantages as those the Random-Walk Sampler has, while also sharing the same advantages as those the Independence Sampler has.

Suppose, then, that one has a target density $\pi(x) = \gamma(x)/\eta$ with support on $\mathbb{R}^d$, where

$$\eta := \int_{\mathcal{X}} \gamma(x) \, \mathrm{d}x$$

is a, potentially unknown, constant. Suppose further that the current state of the chain is x. As with the Independence Sampler, let $q_0 : \mathbb{R}^d \to \mathbb{R}$ be a marginal density chosen to match the target as closely as possible. Moreover, let $q_1 : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ be a joint density which is exchangeable (see Definition 2.1.2), and which emits q_0 as the marginal density; that is,

$$q_0(y_0) = \int_{\mathbb{R}^d} q_1(y_0, y_1) \, \mathrm{d}y_1 \, , \quad q_0(y_1) = \int_{\mathbb{R}^d} q_1(y_0, y_1) \, \mathrm{d}y_0 \, .$$

Consider proposing a state y_1 with conditional density $\tilde{q}_1(y_1|y_0) = q_1(y_0, y_1)/q_0(y_0)$ where, for notational simplicity, $y_0 = x$. Set the new state, x^*, say, to be y_1 with the Metropolis-Hastings acceptance probability

$$\alpha_m(y_{0:1}) := 1 \wedge \frac{w(y_1)}{w(y_0)} \, ,$$

4 Exchangeability of the joint density ensures that there is a unique marginal density regardless of which variables are marginalised out.

where, analogously to the Independence Sampler, w denotes the *transition weight*; $w(z) := \gamma(z)/q_0(z)$, and set x^* to be y_0 otherwise. *Detailed balance* holds since

$$\begin{aligned}\pi(y_0)\tilde{q}_1(y_1|y_0)\alpha_m(y_{0:1}) &= \pi(y_0)\frac{q_1(y_{0:1})}{q_0(y_0)}\alpha_m(y_{0:1}) \\ &= \eta^{-1}w(y_0)q_1(y_{0:1})\alpha_m(y_{0:1}) \\ &= \eta^{-1}q_1(y_{0:1})[w(y_0)\wedge w(y_1)]\,,\end{aligned}$$

which, since q_1 is exchangeable, is a symmetric function of $y_{0:1}$. Therefore, the chain induced by such a procedure is reversible with respect to π (see Definition 2.3.7). Alternatively, one could use Barker's acceptance probability;

$$\alpha_b(y_{0:1}) = \frac{w(y_1)}{w(y_0)+w(y_1)}\,,$$

and the conclusions would still hold since

$$\pi(y_0)\tilde{q}_1(y_1|y_0)\alpha_b(y_{0:1}) = \eta^{-1}q_1(y_{0:1})\frac{w(y_0)w(y_1)}{w(y_0)+w(y_1)}\,,$$

is a symmetric function of $y_{0:1}$. As can be seen, the acceptance probabilities for the Exchangeable Sampler are the same as the acceptance probabilities for the Independence Sampler. Therefore, the single-proposal Exchangeable Sampler outlined above can be efficiently extended to the multiple-proposal regime. Indeed, letting $\alpha_{iN}(y_{0:N})$ be the multiple-proposal extension of either Barker's acceptance probability;

$$\alpha^b_{i,N}(y_{0:N}) = \frac{w(y_i)}{w(y_0)+\ldots+w(y_N)}\,, \tag{86}$$

or the Metropolis-Hastings' acceptance probability

$$\alpha^m_{i,N}(y_{0:N}) := \frac{w(y_i)}{w(y_0)+\ldots+w(y_N)-[w(y_k)\wedge w(y_0)]}\,, \tag{87}$$

for the i-th proposal, then Algorithm 16 gives the procedure for the multiple-proposal Exchangeable Sampler; henceforth, simply, the Exchangeable Sampler.

REMARK 8. *We note that Tjelmeland, 2004 provides a general framework for multiple-proposal samplers where the transition step involves simulating N proposals from a conditional proposal $\tilde{q}_N(y_{1:N}|y_0)$ and choosing a move to y_k with some probability. In the framework Tjelmeland, 2004 presents, one does not need to know the marginal distribution, q_0, to implement the algorithm. Moreover, the two proposals described in Tjelmeland, 2004 are such that the proposal $\tilde{q}_N$ satisfies a certain symmetry, much like the random-walk sampler does; that is, $\tilde{q}_N(y_{1:N}|y_0) = \tilde{q}_N(y_{0:i-1}, y_{i+1:N}|y_i)$ for any $i \in \{1,\ldots,N\}$. As such, for both these proposals, the transition weight, much like the transition weight for the random-walk sampler, is simply $\gamma(\cdot)$. The first proposal is very similar to the proposal we introduce in Algorithm 18 in the*

Algorithm 16 Multiple-Proposal Exchangeable Sampler

Initialise the chain at some $x_0 \in \mathbb{R}^d$ and choose the number of iterations $T > 0$.
Let $q_N(y_{0:N})$ be an exchangeable density whose marginal density is $q_0(\cdot)$.
Define, as the proposal density,

$$\tilde{q}_N(y_{1:N}|y_0) = \frac{q_N(y_{0:N})}{q_0(y_0)} .$$

for $t = 0, \ldots, T-1$ **do**
    Let $y_0 := x_t$.
    Propose a sequence $y_{1:N}$ from $\tilde{q}_N(\cdot|y_0)$.
    For each $k \in \{1, \ldots, N\}$ calculate $\alpha_{k,N}(y_{0:N})$.
    Set $\alpha_{0,N}(y_{0:N}) = 1 - (\alpha_{1,N}(y_{0:N}) + \ldots + \alpha_{N,N}(y_{0:N}))$.
    Set $x_{t+1} = y_k$ with probability $\alpha_{kN}(y_{0:N})$.
end for

case where the target and the marginal are both Normal distributions. The second mimics random-walk proposals while keeping the proposals equidistant from one another. Much like the Exchangeable Sampler introduced in this thesis, it is the symmetry that simplifies the transition weight and leads to an efficient sampler. The approaches described in Tjelmeland, 2004 are a natural alternative to the Exchangeable Sampler introduced in this thesis and so it would be interesting to compare the two approaches and, even, extend the results derived in this thesis to the methods described in Tjelmeland, 2004. Unfortunately, due to finding the results in Tjelmeland, 2004 during the write-up of this thesis, we have not done this comparison and/or extension of results.

In one-dimension one can use the conditional form of Algorithm 2, which is a multiple-sample extension of the preconditioned Crank-Nicolson proposal introduced in Cotter et al., 2013, to generate proposals in an exchangeable way while, retaining q_0 as the marginal density, and also allowing the flexibility of making the proposal as close as one wants to the current state of the chain via a tunable *jump-size*, ϵ. The full procedure is given by Algorithm 17. This procedure can be easily

Algorithm 17 Exchangeable Proposal $\tilde{q}_N(y_{1:N}|y_0)$ With Marginal q_0 and Jump-Size $\epsilon \in (0, \sqrt{2})$ for $d = 1$.

Let Φ denote the distribution function of a standard normal random variable and Q_0 denote the cumulative distribution function corresponding to the marginal density q_0.
Set $\delta := \epsilon/\sqrt{2}$.
Calculate $z_0 = \Phi^{-1}(Q_0(y_0))$.
Sample $\hat{z}_0$ from a $\mathrm{N}(0,1)$ distribution.
Set $\theta = \sqrt{1-\delta^2} z_0 + \delta \hat{z}_0$.
for $i = 1, \ldots, N$ **do**
    Sample $\hat{z}_i$ from a $\mathrm{N}(0,1)$ distribution.
    Set $z_i = \theta\sqrt{1-\delta^2} + \delta \hat{z}_i$.
    Set $y_i = Q_0^{-1}(\Phi(z_i))$.
end for

extended to a general dimension, d, provided there exists an invertible mapping, $h : \mathbb{R}^d \to \mathbb{R}^d$, such that, if $Z_{1:d} \sim \mathrm{N}_d(0, I_d)$, where I_d denotes the d-dimensional identity matrix, then $h(Z_{1:d}) \sim q_0(\cdot)$. The full procedure is given by Algorithm 18.

Algorithm 18 Exchangeable Proposal $\tilde{q}_N(y_{1:N}|y_0)$ With Marginal q_0 and Jump-Size $\epsilon \in (0, \sqrt{2})$ for General d.

1: Let $h : \mathbb{R}^d \to \mathbb{R}^d$ be an invertible mapping, such that, if $Z_{1:d} \sim \mathrm{N}_d(0, I_d)$, where I_d denotes the d-dimensional identity matrix, then $h(Z_{1:d}) \sim q_0(\cdot)$.
2: Set $\delta := \epsilon/\sqrt{2}$.
3: Calculate $z_0 = h^{-1}(y_0)$.
4: Sample $\hat{z}_0$ from a $\mathrm{N}_d(0, I_d)$ distribution.
5: Set $\theta = \sqrt{1-\delta^2} z_0 + \delta \hat{z}_0$.
6: **for** $i = 1, \ldots, N$ **do**
7: Sample $\hat{z}_i$ from a $\mathrm{N}_d(0, I_d)$ distribution.
8: Set $z_i = \theta\sqrt{1-\delta^2} + \delta \hat{z}_i$.
9: Set $y_i = h(z_i)$.
10: **end for**

REMARK 9. *In practice, it is not neccessary to know how to do the inversion since the first step of the procedure, $z_0 = h^{-1}(y_0)$, can be omitted provided one stores, in memory, the z_0 corresponding to the current state of the chain. However, the exposition of the Exchangeable Sampler and the Exchangeable Particle Gibbs Sampler of section 4.4 is clearer if the proposal is explicitly conditional on y_0.*

REMARK 10. *The implementation given by Algorithm 18 uses the same* jump-size, *ϵ, for each of the d dimensions. However, exchangeability of the proposal density and the results that follow still hold if one uses a different jump-size, ϵ_i, say, for each of the dimensions. The benefit of doing this would be to take larger jumps in the dimensions where one knew that the proposal in that dimension was closer to the target. For instance, consider simulating a conditioned diffusion as described in Chapter 3. The proposals described in that chapter of the thesis, like the Modified Diffusion Bridge (Section 3.2.3), for example, tend to be better approximations to the true conditioned diffusion as time gets closer to the time that the observation being conditioned upon occurs. As a result, it would be prudent, in such examples, to use a larger jump-size closer to the time of the observation and a smaller jump-size further away. However, to ease exposition, we will assume a fixed ϵ for each of the d dimensions.*

If q_0 is continuous, then Algorithm 18 is the combination of transformations under continuous functions, with sampling from continuous distributions. Thus, it follows that, if q_0 is continuous on $\mathbb{R}^d$, then the proposal, $\tilde{q}_N$, is continuous on $\mathbb{R}^{d\times(N+1)}$. Moreover, by construction, for any $x \in \mathbb{R}^d$ and $\epsilon > 0$, $z_i \in (-\infty, \infty)^d$ for any $i \in \{1, \ldots, N\}$. Hence, $y_i \in \{y \in \mathbb{R}^d : q_0(y) > 0\}$ for each $i \in \{1, \ldots, N\}$. Therefore,

$$\{y_{1:N} \in \mathbb{R}^{d\times N} : \tilde{q}_N(y_{1:N}|x) > 0\} \equiv \{y \in \mathbb{R}^d : q_0(y) > 0\}^N .$$

LEMMA 4.3.1. *Let q_0 be continuous on $\mathbb{R}^d$ and, for any $\epsilon \in (0, \sqrt{2})$, let $\tilde{q}_N(\cdot|\cdot)$ be the proposal density corresponding to Algorithm 18. Then $\tilde{q}_N(\cdot|\cdot)$ is continuous on $\mathbb{R}^{d\times(N+1)}$ and*

$$\{y_{1:N} \in \mathbb{R}^{d\times N} : \tilde{q}_N(y_{1:N}|x) > 0\} \equiv \{y \in \mathbb{R}^d : q_0(y) > 0\}^N . \quad (88)$$

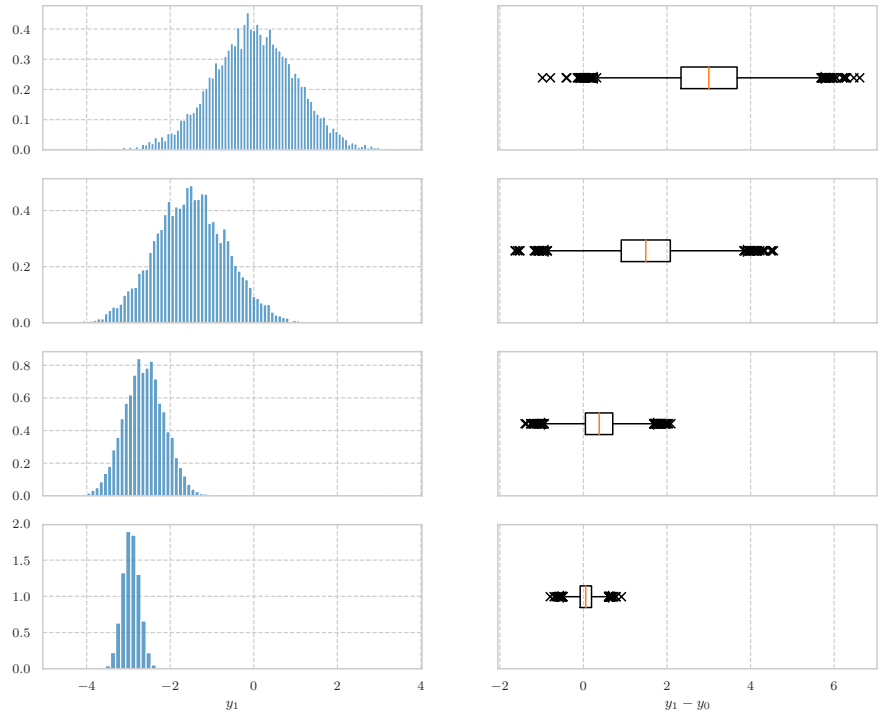

Figure 26: Plots of the histograms (left) and of ten-thousand samples simulated from $\tilde{q}_1(\cdot|-3)$ using Algorithm 18, where the marginal density, q_0, corresponds to a $\mathrm{N}(0,1)$ distribution. Each row corresponds to a different choice of the jump-size, $\epsilon \in \{0.2, 0.5, 1.0, \sqrt{2}\}$, where the top row corresponds to the largest jump-size, $\epsilon = \sqrt{2}$, and the bottom to the smallest, $\epsilon = 0.2$. The right column shows the boxplots of the corresponding differences between the simulated samples and the initial state, $y_1 - y_0$.

Before discussing the ergodic properties of the Exchangeable Sampler for a general N, it will be useful to demonstrate how, in practice, the choice of the jump-size ϵ affects the closeness of the proposals. To this end, this thesis will consider two illuminating examples. Firstly, consider the simple case where q_0 corresponds to a $\mathrm{N}(0,1)$ distribution. Suppose $y_0 = -3$ and consider simulating ten-thousand independent samples from $\tilde{q}_1(\cdot|y_0)$ using Algorithm 18. The left column of Figure 26 shows the histograms of the simulated samples for different choices of the jump-size, $\epsilon \in \{0.2, 0.5, 1.0, \sqrt{2}\}$ where the top row corresponds to the largest jump-size, $\epsilon = \sqrt{2}$, and the bottom to the smallest, $\epsilon = 0.2$. Note that the largest jump-size, $\epsilon = \sqrt{2}$, corresponds to the case of independent proposals. The right column of Figure 26 shows, via boxplots, the corresponding differences between the simulated samples and the initial state, $y_1 - y_0$. It is clear from the figure that, the smaller the jump-size, the closer the samples are to the initial state y_0. This behaviour is the same as the behaviour of samples proposed via the Random-Walk Sampler. However, it is also clear that, unlike the behaviour of samples proposed via the Random-Walk Sampler, the larger the jump-size the closer the empirical distribution function correspond-

ing to the samples resembles that of a $\mathrm{N}(0,1)$. Moreover, because the majority of the mass of a $\mathrm{N}(0,1)$ distribution lies within the range $(-3,3)$ and the initial state is -3, the distribution of the difference between the simulated proposals and the initial state, $y_1 - y_0$ is, on average, positive. Again, this is unlike the Random-Walk Sampler whose proposals have a difference from the current state which is symmetric around zero.

Secondly, consider the more complicated case where π corresponds to the conditioned Birth-Death diffusion of Section 3.1.1 and q_0 corresponds to the Modified Diffusion Bridge proposal of Section 3.2.3— see Chapter 3 for more details regarding simulating conditioned diffusions. Specifically, recall, from Section 3.2.3, that, in one-dimension, the MDB proposal of a discretised path of the diffusion, $x_{1:K}$, say, takes the form

$$q_0^{\mathrm{MDB}}(x_{1:K}|y) = \prod_{k=1}^{K} \phi(x_k; a_{k-1}^{\mathrm{MDB}}, C_{k-1}^{\mathrm{MDB}}) \,,$$

where ϕ denotes the density corresponding to a one-dimensional normal distribution and a_{k-1}^{MDB} and C_{k-1}^{MDB} correspond to the mean (Equation (54)) and variance (Equation (55)) respectively, and, implicitly, depend on x_{k-1}, T, y, and t_{k-1}. Such a proposal is equivalent to proposing K independent $\mathrm{N}(0,1)$ random variables, $Z_{1:K}$, and transforming those random variables appropriately by sequentially setting, for $k \in \{1,\ldots,K\}$, $X_k = a_{k-1}^{\mathrm{MDB}} + \sqrt{C_{k-1}^{\mathrm{MDB}}} Z_k$. Thus, simulating *exchangeable* paths corresponds to simulating sequences of K independent $\mathrm{N}(0,1)$ random variables in an exchangeable way. Given a high-weighted proposal path, y_0, simulated from the MDB, consider simulating one-hundred exchangeable samples from $\tilde{q}_{100}(\cdot|y_0)$ using Algorithm 18. Figure 27 contains plots of one-hundred samples from $\tilde{q}_{100}(\cdot|y_0)$ for different choices of the jump-size, $\epsilon \in \{0.1, 0.6, 1.0, \sqrt{2}\}$, alongside the path being conditioned upon, y_0. The top row corresponds to the largest jump-size, $\epsilon = \sqrt{2}$, and the bottom to the smallest, $\epsilon = 0.1$. The left column shows the path being conditioned upon, y_0, in orange along with the simulated exchangeable paths in blue. The right column shows the same thing but where the transparency of the simulated paths have been set to be inversely proportional to the corresponding normalised weights, and where the normalisation has occurred including the weight of the path being conditioned upon. It is clear from the figure that, the smaller the jump-size, the closer the simulated paths are to the initial path y_0. Moreover, the weights are less variable and the path being conditioned upon dominates less. Furthermore, as was the case for the previous example, the larger the jump-size, the closer the simulated paths resemble simulated paths from the MDB. In this case, the weights are more variable, and the path being conditioned upon dominates— note that, in the independent case; that is, where $\epsilon = \sqrt{2}$, the simulated paths are essentially invisible highlighting the fact that, in this case, the path being conditioned upon essentially has a normalised weight of 1.

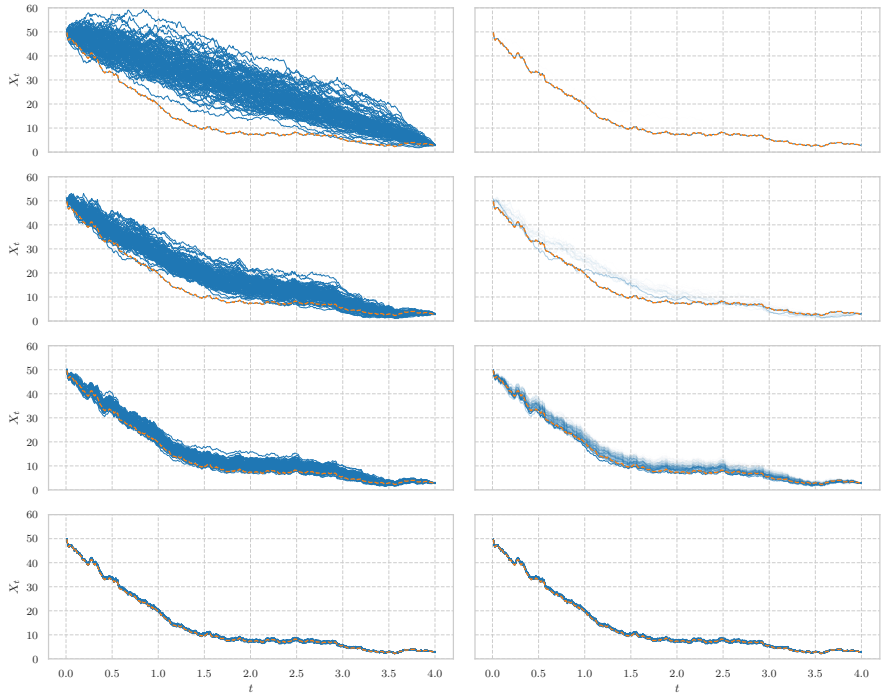

Figure 27: Plots of one-hundred samples from $\tilde{q}_{100}(\cdot|y_0)$, using Algorithm 18, where the marginal density, q_0, corresponds to the Modified Diffusion Bridge proposal of Section 3.2.3 and the target conditioned diffusion corresponds to the Birth-Death diffusion of Section 3.1.1. Here, the path being conditioned upon, y_0, is a high-weighted sample from the MDB. Each row corresponds to a different choice of the jump-size, $\epsilon \in \{0.1, 0.6, 1.0, \sqrt{2}\}$, where the top row corresponds to the largest jump-size, $\epsilon = \sqrt{2}$, and the bottom to the smallest, $\epsilon = 0.1$. The left column shows the path being conditioned upon, y_0, in orange along with the simulated exchangeable paths in blue. The right column shows the same thing but where the transparency of the simulated paths have been set to be inversely proportional to the corresponding normalised weights, and where the normalisation has occurred including the weight of the path being conditioned upon.

The Exchangeable Sampler, with proposal density given by Algorithm 18, satisfies some fundamental properties regarding reversibility, irreducibility, and non-negativity:

THEOREM 4.3.2. *Let X_t be the Markov chain corresponding to the Exchangeable Sampler with either Barker's or the Metropolis-Hastings acceptance probability. Suppose the sampler targets π using the proposal density, $\tilde{q}_N(\cdot|x)$, corresponding to Algorithm 18 which emits q_0 as the marginal density, where q_0 is continuous on $\mathbb{R}^d$, and*

$$\{y \in \mathbb{R}^d : \gamma(y) > 0\} \subseteq \{y \in \mathbb{R}^d : q_0(y) > 0\} .$$

Moreover, let $\epsilon \in (0, \sqrt{2}]$ be arbitrary. Then, the chain is;

1. *Reversible with respect to π.*

2. *One-step irreducible as defined in Lemma 2.3.10.*

3. *Non-negative.*

Proof. See A.11. □

This theorem, in conjunction with Lemma 2.3.10, Corollary 2.3.12, and Theorem 2.3.13, demonstrates that the Markov chain corresponding to the Exchangeable Sampler which targets π, has π as the limiting distribution of the chain and that the resulting MCMC estimates satisfy a Strong Law of Large Numbers result; (20). The non-negativity result will also allow us to make use of Theorem 2.3.25 to prove that, under certain conditions, the Exchangeable Sampler is geometrically ergodic for any $N \in \mathbb{N}$. First, we prove the following theorem which gives general conditions under which propose and accept-reject Markov chains satisfy a geometric drift condition;

THEOREM 4.3.3. *Let X_t be a propose and accept-reject Markov chain with state space $\mathcal{X}$, a proposal density $q(y|x)$, and an acceptance probability $\alpha(x, y)$. Suppose that there exists a function $p : \mathcal{X} \to [0, \infty)$, which is finite for at least one $x \in \mathcal{X}$, and constants, $\rho_* \in (1, \infty)$ and $\delta > 0$, such that, with $C := \{x \in \mathcal{X} : p(x) \leq \rho_*\}$, the following hold;*

(S) C is a small set.

(IM) For any $x \notin C$,

$$\mathbb{E}_{q(\cdot|x)}\left[\alpha(x, Y) \log\left(\frac{1}{\rho_*} + \frac{p(Y)}{p(x)}\right)\right] < -\delta$$

(UI) There exists a positive $\tau < \infty$ such that

$$\mu_\tau := \sup_{x \notin C} \mathbb{E}_{q(\cdot|x)}\left[\alpha(x, Y) \log\left(\frac{1}{\rho_*} + \frac{p(Y)}{p(x)}\right)^2 \mathbb{1}_{\mathcal{P}_\tau(x)}(Y)\right] < \infty ,$$

where $\mathcal{P}_\tau(x) := \{z \in \mathcal{X} : \tau p(x) \leq p(z)\}$.

(B) For $\mathcal{P}_1(x)$ defined in property (UI),

$$\xi := \sup_{x \in C} \mathbb{E}_{q(\cdot|x)}[\alpha(x, Y)(p(Y) - p(x))\mathbb{1}_{\mathcal{P}_1(x)}(Y)] < \infty .$$

Then, X_t satisfies a geometric drift condition. That is, letting $P(x, \cdot)$ denote the transition distributions of the chain, there exists a function $v : \mathbb{R}^d \to [1, \infty)$, which is finite for at least one $x \in \mathbb{R}^d$, an ϵ-small set C, and positive, finite constants, β and $\gamma < 1$, such that

$$\mathbb{E}_{P(x,\cdot)}(v(Y)) \leq \gamma v(x) + \beta \mathbb{1}_C(x) .$$

Proof. See A.12. □

Before showing how this theorem can be applied, it is worthwhile to highlight the implications of each of the assumptions. Assumption (S) of the theorem states that the set on which the function p is relatively small is a small set. Therefore, the interesting behaviour of the chain occurs when p is relatively large. Assumption (IM) asserts that, off the small set; that is, when p is relatively large, the impetus of the chain is to move to regions where p is smaller. In other words, there is a drift towards smaller p when p is large. Assumption (UI) ensures that moves from outside the small set C to regions where p is relatively much larger are uniformly *well-behaved*. Finally, assumption (B) asserts that, on average, any move from within the small set C to a region where p is relatively larger is not such that, in this region, p is *too large*. While such a theorem may seem a little contrived, its applicability can be seen immediately by using it to provide an alternative proof of the geometric ergodicity result of Theorem 2.3.34:

THEOREM 4.3.4. *Let X_t be the Metropolis-Hastings random-walk sampler in one dimension; that is, a propose-and-accept-reject Markov chain with proposal density*

$$q(x,y) = \frac{1}{\sqrt{2\pi\epsilon^2}} \exp\left(-\frac{(y-x)^2}{2\epsilon^2}\right)$$

for some $\epsilon > 0$. Further, suppose that π is greater than zero for any $x \in \mathbb{R}$, and that π decays exponentially in the tails (in the sense of Definition 2.3.33). Then, X_t is geometrically ergodic.

Proof. See A.13. □

Importantly, when combined with Theorem 4.3.2, Theorem 4.3.3 allows us to give a set of intuitive sufficient conditions, under which the Exchangeable Sampler is geometrically ergodic and has corresponding MCMC estimates which satisfy a central limit theorem for all functions which are square-integrable with respect to π. Firstly, we show that any compact set on which the transition weight is bounded above is small:

LEMMA 4.3.5. *Let X_t be the Markov chain corresponding to the Exchangeable Sampler with either Barker's or the Metropolis-Hastings acceptance probability, and with $N = 1$. Suppose, further, that the proposal density, $\tilde{q}_1(\cdot|x)$, is the density corresponding to Algorithm 18 and emits q_0 as the marginal density, where q_0 is continuous on $\mathbb{R}^d$, and*

$$\{y \in \mathbb{R}^d : \gamma(y) > 0\} \subseteq \{y \in \mathbb{R}^d : q_0(y) > 0\} .$$

Moreover, let $\epsilon \in (0, \sqrt{2}]$ be arbitrary. Then, any compact set of the form $\{x \in \mathcal{X} : w(x) \leq \bar{w}\}$, for some $\bar{w} > 0$, is small.

Proof. See A.14. □

With Lemma 4.3.5 in hand, we are now ready to state and prove two Corollaries which provide sufficient conditions under which the Exchangeable Sampler is geometrically ergodic for $N = 1$. The first corresponds to the case where the transition weights are bounded, and the second corresponds to the case where the weights are unbounded:

COROLLARY 4.3.6. *Let X_t be the Markov chain corresponding to the Exchangeable Sampler which satisfies the assumptions of Lemma 4.3.5. Suppose there exists positive, finite, constants $w_* < 1$, w^*, and δ, such that:*

(B) The transition weight, $w(x) := \gamma(x)/q_0(x)$ is bounded by w^* for any $x \in \mathcal{X}$.

(C) The set $C := \{x \in \mathcal{X} : w_* \leq w(x)\}$ is compact.

(IM) For any $x \notin C$,

$$\mathbb{E}_{\tilde{q}_1(\cdot|x)}\left[\alpha(x,Y)\log\left(w_* + \frac{w(x)}{w(Y)}\right)\right] < -\delta\,.$$

Then X_t is geometrically ergodic.

Proof. See A.15. □

COROLLARY 4.3.7. *Let X_t be the Markov chain corresponding to the Exchangeable Sampler which satisfies the assumptions of Lemma 4.3.5. Suppose there exists positive, finite, constants $w^* > 1$ and δ, such that:*

(C) The set $C := \{x \in \mathcal{X} : w(x) \leq w^*\}$ is compact.

(IM) For any $x \notin C$,

$$\mathbb{E}_{\tilde{q}_1(\cdot|x)}\left[\alpha(x,Y)\log\left(\frac{1}{w^*} + \frac{w(Y)}{w(x)}\right)\right] < -\delta\,.$$

(UI) There exists a positive $\tau < \infty$ such that

$$\mu_\tau := \sup_{x \notin C} \mathbb{E}_{q(\cdot|x)}\left[\log\left(\frac{1}{w^*} + \frac{w(Y)}{w(x)}\right)^2 \mathbb{1}_{\mathcal{P}_\tau(x)}(Y)\right] < \infty\,,$$

where $\mathcal{P}_\tau(x) := \{z \in \mathcal{X} : \tau^{-1}w(z) \geq w(x)\}$.

(B) For $\mathcal{P}_1(x)$ defined in property (UI),

$$\xi := \sup_{x \in C} \mathbb{E}_{q(\cdot|x)}[w(Y)\mathbb{1}_{\mathcal{P}_1(x)}(Y)] < \infty\,.$$

Then X_t is geometrically ergodic.

Proof. See A.16. □

In order to extend these corollaries and give sufficient conditions under which the Exchangeable Sampler is geometrically ergodic for general $N \in \mathbb{N}$, and, ultimately, under which the corresponding MCMC estimates satisfy a central limit theorem for all functions which are square-integrable with respect to π, we need a further assumption that ensures that the ratio of the weights is uniformly bounded in probability so that no one proposal leads to a weight which dominates:

ASSUMPTION 4.3.8. *For any $N \in \mathbb{N}$, $\mu_N(r) \to 1$ as $r \to \infty$, where*

$$\mu_N(r) := \inf_{x \in \mathcal{X}} \mathbb{P}_{\tilde{q}_N(\cdot|x)}\left(rw(Y_1) \geq \max_{i=2,\ldots,N} w(Y_i) \right) .$$

THEOREM 4.3.9. *Let X_t be the Markov chain corresponding to the Exchangeable Sampler which satisfies the assumptions of Lemma 4.3.5. Suppose that, either assumptions (B), (C), and (IM) of Corollary 4.3.6 hold, or that assumptions (C), (IM), (UI), and (B) of Corollary 4.3.7 hold. Finally, suppose that Assumption 4.3.8 holds. Then, X_t is geometrically ergodic and the MCMC estimates corresponding to such a sampler satisfy a central limit theorem for all functions which are square-integrable with respect to π.*

Proof. See A.17. □

In general, given a target, π, and a proposal, q_0, it is difficult to show that, either assumptions (B), (C), and (IM) of Corollary 4.3.6, or assumptions (C), (IM), (UI), and (B) of Corollary 4.3.7, hold. However, from a practical viewpoint, it is often simple to investigate such assumptions numerically, and, therefore, provide a certain level of numerical justification for using the Exchangeable Sampler in such a situation. The following illuminating examples highlight several interesting scenarios that may occur in practice:

EXAMPLE 4. *Consider the case where the target, $\pi(x)$, is a $\mathrm{N}(0,1)$ distribution, and, the marginal proposal, $q_0(x)$, is a $\mathrm{N}(0,\sigma^2)$ distribution, where $\sigma^2 := (2\tau+1)^{-1}$ for some $\tau > 0$ (i.e. $\sigma^2 < 1$). In this scenario the tails of the proposal are exponentially lighter than the tails of the target. Indeed, the weight, $w(x) = \pi(x)/q_0(x)$, is proportional to $\exp(\tau x^2)$ and, therefore, unbounded. As a result, one would expect that the sampler is not geometrically ergodic since, the further into the tails the sampler goes, the larger the weight and the increase in the weight is exponential. It is shown in Lemma 4.3.10 that, for this example, for any $\epsilon \in (0, \sqrt{2}]$, assumption (IM) of Corollary 4.3.7 is violated and, therefore, Theorem 4.3.9 does not apply.*

LEMMA 4.3.10. *Let $\pi(x)$ be the density corresponding to a $\mathrm{N}(0,1)$ distribution and $q_0(x)$ be the density corresponding to a $\mathrm{N}(0,\sigma^2)$ distribution where $\sigma^2 := (2\tau+1)^{-1}$ for some $\tau > 0$. Suppose the proposal density, $\tilde{q}_1$, is the density corresponding to Algorithm 17 and emits q_0 as the marginal density. Let $\epsilon \in (0, \sqrt{2}]$ be arbitrary. Then, for any $\delta > 0$, there exists a $x^* > 0$ such that, for any $x \geq x^*$, and any $w^* > 1$,*

$$\mathbb{E}_{\tilde{q}_1(\cdot|x)}\left[\alpha(x,Y) \log\left(\frac{1}{w^*} + \frac{w(Y)}{w(x)} \right) \right] > -\delta .$$

That is, Assumption (IM) of Corollary 4.3.7 can not hold.

Proof. See A.18. □

EXAMPLE 5. *Consider the case where the target, $\pi(x)$, is a* $\mathrm{Gamma}(\alpha_1, \beta)$ *distribution, and, the marginal proposal, $q_0(x)$, is a* $\mathrm{Gamma}(\alpha_2, \beta)$ *distribution where $\alpha_1 > \alpha_2$. In this scenario, as with the previous example, the tails of the proposal are lighter than the tails of the target, however, only polynomially so. Indeed, the weight, $w(x) = \pi(x)/q_0(x)$, is proportional to $x^{(\alpha_1-\alpha_2)}$ and, therefore, unbounded. As a result, one would expect that the sampler is geometrically ergodic since, even though the further into the tails the sampler goes, the larger the weight, the rate of increase is only polynomial. The sufficient conditions for geometric ergodicity in the case where $N = 1$, as given by Corollary 4.3.7, rely on the three quantities:*

$$\mathbb{E}_{\tilde{q}_1(\cdot|x)}\left[\alpha(x,Y)\log\left(\frac{1}{w^*}+\frac{w(Y)}{w(x)}\right)\right],$$
$$\mathbb{E}_{\tilde{q}_1(\cdot|x)}\left[\log\left(\frac{1}{w^*}+\frac{w(Y)}{w(x)}\right)^2\mathbb{1}_{\mathcal{P}_\tau(x)}(Y)\right],$$
$$\mathbb{E}_{\tilde{q}_1(\cdot|x)}[w(Y)\mathbb{1}_{\mathcal{P}_1(x)}(Y)],$$

where $\mathcal{P}_\tau(x) := \{z \in \mathcal{X} : \tau^{-1}w(z) \geq w(x)\}$. Consider the case where $\beta = 1$, $\alpha_1 = 5.5$, and $\alpha_2 = 0.5$, so that $\alpha_1 - \alpha_2 = 5$. Take $w^* = 10^5$. Then $C = [0, 10)$. Figure 28 shows plots of Monte Carlo approximations of the quantities

$$\mathbb{E}_{\tilde{q}_1(\cdot|x)}\left[\alpha(x,Y)\log\left(\frac{1}{w^*}+\frac{w(Y)}{w(x)}\right)\right], \tag{89}$$
$$\mathbb{E}_{\tilde{q}_1(\cdot|x)}\left[\log\left(\frac{1}{w^*}+\frac{w(Y)}{w(x)}\right)^2\right], \tag{90}$$
$$\mathbb{E}_{\tilde{q}_1(\cdot|x)}[w(Y)], \tag{91}$$

in the first, second, and third columns respecively. The plots in the first two columns are over the interval $x \in [10, 5000) \subset C^c$, whereas the plots in the last column are over the interval $[0, 10) = C$. Each row corresponds to a different choice of the jump-size, with $\epsilon = 1.0$ for the top row, and $\epsilon = 0.5$, $\epsilon = 0.2$, $\epsilon = 0.05$ for the second, third and fourth row respectively. The Monte Carlo approximation to the expectations have been calculated with one-hundred thousand samples. The plots show, at least empirically, that, for these choices of ϵ,

$$\mathbb{E}_{\tilde{q}_1(\cdot|x)}\left[\alpha(x,Y)\log\left(\frac{1}{w^*}+\frac{w(Y)}{w(x)}\right)\right] < 0,$$
$$\mathbb{E}_{\tilde{q}_1(\cdot|x)}\left[\log\left(\frac{1}{w^*}+\frac{w(Y)}{w(x)}\right)^2\right] < \infty,$$

for any $x \in [10, 5000) \subset C^c$, and

$$\sup_{x\in C}\mathbb{E}_{\tilde{q}_1(\cdot|x)}[w(Y)] < \infty.$$

Thus, these plots suggest that, for $\epsilon \in \{0.05, 0.2, 0.5, 1.0\}$, the conditions of Corollary 4.3.7 hold, the Exchangeable Sampler is geometrically ergodic for $N = 1$, and the MCMC estimates satisfy a central

limit theorem for all functions which are square-integrable with respect to π. Recall that, for general N, it is, by Theorem 4.3.9, sufficient to consider the quantity

$$\mu_N(r) := \inf_{x \in \mathcal{X}} \mathbb{P}_{\tilde{q}_N(\cdot|x)} \left(r w(Y_1) \geq \max_{i=2,\ldots,N} w(Y_i) \right), \tag{92}$$

given in Assumption 4.3.8. Figure 29 shows plots of Monte Carlo approximations of the quantity given by (92) over the interval $\log(r) \in [0, 50]$ for a range of values for N. Each plot corresponds to a different jump-size, with $\epsilon = 1.0$ for the top-left plot, and $\epsilon = 0.5$, $\epsilon = 0.2$, $\epsilon = 0.05$ for the top-right, bottom-left and bottom-right plot respectively. The Monte Carlo approximation to the probabilities have been calculated with one-hundred thousand samples and the infimum has been empirically approximated by taking the minimum over a range of values of x in the interval $(10^{-10}, 10^{1.5})$. These plots suggest that, for $\epsilon \in \{0.05, 0.2, 0.5, 1.0\}$, Assumption 4.3.8 holds. Thus, in conjunction with Figure 28, this suggests that, at least empirically, the conditions of Theorem 4.3.9 hold, the Exchangeable Sampler is geometrically ergodic for any $N \in \mathbb{N}$, and the MCMC estimates satisfy a central limit theorem for all functions which are square-integrable with respect to π.

EXAMPLE 6. *Consider the case where the target, $\pi(x)$, is a $\mathrm{N}(0, 1)$ distribution, and, the marginal proposal, $q_0(x)$, is a $\mathrm{T}(\nu)$ distribution. In this case the weight, $w(x) = \pi(x)/q_0(x)$, is proportional to*

$$\exp\left(-\frac{x^2}{2}\right)\left(1 + \frac{x^2}{\nu}\right)^{(1+\nu)/2}.$$

As $|x| \to \infty$, $w(x) \downarrow 0$. Therefore, the weight is bounded. As such, one would expect that, in this scenario, the Exchangeable Sampler is geometrically ergodic. The sufficient condition for geometric ergodicity in the case where $N = 1$, as given by Corollary 4.3.6, relies on the quantity:

$$\mathbb{E}_{\tilde{q}_1(\cdot|x)}\left[\alpha(x, Y) \log\left(w_* + \frac{w(x)}{w(Y)}\right)\right].$$

Consider the case where $\nu = 5$, and take $w_* = 10^{-10}$. Then, approximately, $C = [-5.81, 5.81]$. Figure 30 shows plots of Monte Carlo approximations of the quantity

$$\mathbb{E}_{\tilde{q}_1(\cdot|x)}\left[\alpha(x, Y) \log\left(w_* + \frac{w(x)}{w(Y)}\right)\right], \tag{93}$$

for x in $[-12.28, -5.81)$ and x in $(5.81, 12.28]$ in the first and second column respectively. Each row corresponds to a different choice of the jump-size, with $\epsilon = 1.2$ for the top row, and $\epsilon = 1.0$, $\epsilon = 0.8$, $\epsilon = 0.5$ for the second, third and fourth row respectively. The Monte Carlo approximations to the expectations have been calculated with one-hundred

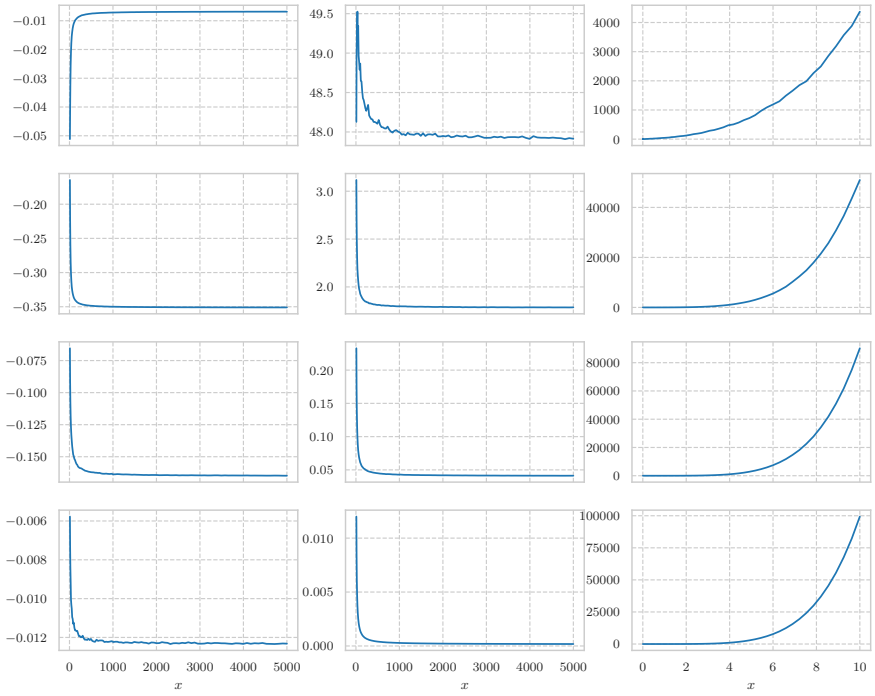

Figure 28: Plots of Monte Carlo approximations of the quantities given by (89), (90), and (91) in the first, second and third columns respectively. The plots in the first two columns are over the interval $x \in [10, 5000) \subset C^c$, whereas the plots in the last column are over the interval $[0, 10) = C$. Each row corresponds to a different choice of the jump-size, with $\epsilon = 1.0$ for the top row, and $\epsilon = 0.5$, $\epsilon = 0.2$, $\epsilon = 0.05$ for the second, third and fourth row respectively. The Monte Carlo approximation to the expectations have been calculated with one-hundred thousand samples.

thousand samples. The plots show, at least empirically, that, for these choices of ϵ*,*

$$\mathbb{E}_{\tilde{q}_1(\cdot|x)}\left[\alpha(x, Y)\log\left(w_* + \frac{w(x)}{w(Y)}\right)\right] < 0\ ,$$

for any $x \in [-12.28, -5.81) \cup (5.81, 12.28]$*. Thus, these plots suggest that, for* $\epsilon \in \{0.5, 0.8, 1.0, 1.2\}$*, the conditions of Corollary 4.3.6 hold, the Exchangeable Sampler is geometrically ergodic for* $N = 1$*, and the MCMC estimates satisfy a central limit theorem for all functions which are square-integrable with respect to* π*. As was the case for the previous example, for general* N*, it is, by Theorem 4.3.9, sufficient to consider the quantity* (92) *given in Assumption 4.3.8. Figure 31 shows plots of Monte Carlo approximations of the quantity given by* (92) *over the interval* $\log(r) \in [0, 200]$ *for a range of values for* N*. Each plot corresponds to a different jump-size, with* $\epsilon = 1.2$ *for the top-left plot, and* $\epsilon = 1.0$*,* $\epsilon = 0.8$*,* $\epsilon = 0.5$ *for the top-right, bottom-left and bottom-right plot respectively. The Monte Carlo approximations to the probabilities have been calculated with one-hundred thousand samples and the infimum has been empirically approximated by taking the minimum over a*

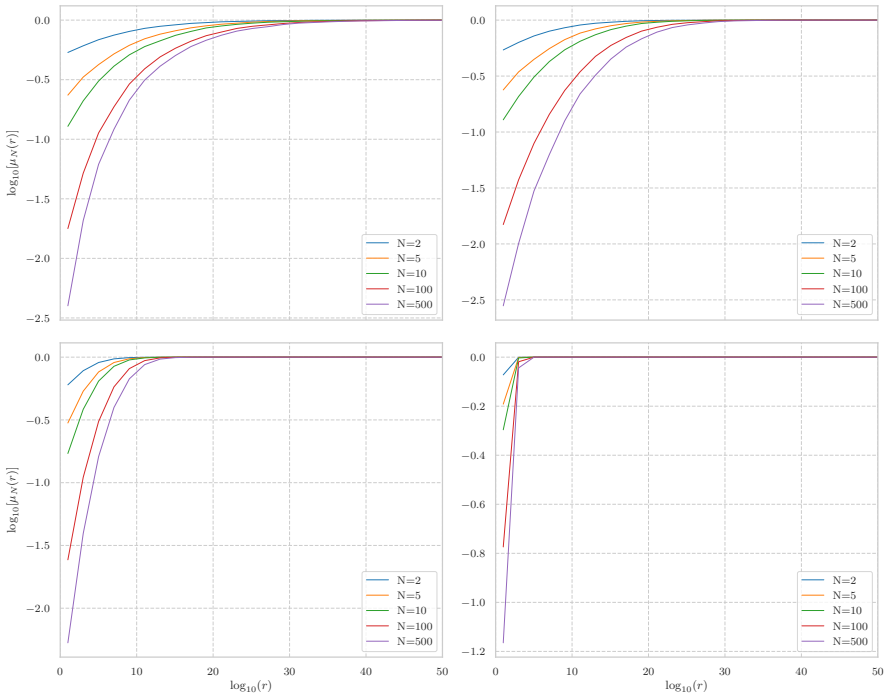

Figure 29: Plots of Monte Carlo approximations of the quantity given by (92) over the interval $\log(r) \in [0, 50]$ for a range of values for N. Each plot corresponds to a different jump-size, with $\epsilon = 1.0$ for the top-left plot, and $\epsilon = 0.5$, $\epsilon = 0.2$, $\epsilon = 0.05$ for the top-right, bottom-left and bottom-right plot respectively. The Monte Carlo approximation to the probabilities have been calculated with one-hundred thousand samples and the infimum has been empirically approximated by taking the minimum over a range of values of x in the interval $(10^{-10}, 10^{1.5})$.

range of values of x in the interval $(-1000, 1000)$. These plots suggest that, for $\epsilon \in \{0.5, 0.8\}$, Assumption 4.3.8 may not hold. Thus, for these choices of ϵ, the conditions of Theorem 4.3.9 may not hold. However, the plots do suggest that, for $\epsilon \in \{1.0, 1.2\}$, Assumption 4.3.8 does hold. Thus, in conjunction with Figure 28, this suggests that, at least empirically, for these choices of the jump-size, the conditions of Theorem 4.3.9 hold, the Exchangeable Sampler is geometrically ergodic for any $N \in \mathbb{N}$, and the MCMC estimates satisfy a central limit theorem for all functions which are square-integrable with respect to π.

As with the Random-Walk Sampler of Section 2.3.6.2, while Theorem 4.3.9 gives sufficient conditions on the target under which the Exchangeable Sampler produces MCMC estimates which satisfy central limit theorems, it does not give guidance with regards to choosing a *good* step-size; that is, a step-size which results in a Markov chain with a *high* rate of *mixing*. Moreover, in comparison with the Random-Walk Sampler, the Exchangeable Sampler has the added complexity that one is also free to choose the marginal proposal q_0 and the number of samples N to improve the mixing of the chain. To see the importance of the step-size and the number of samples on the mixing properties of

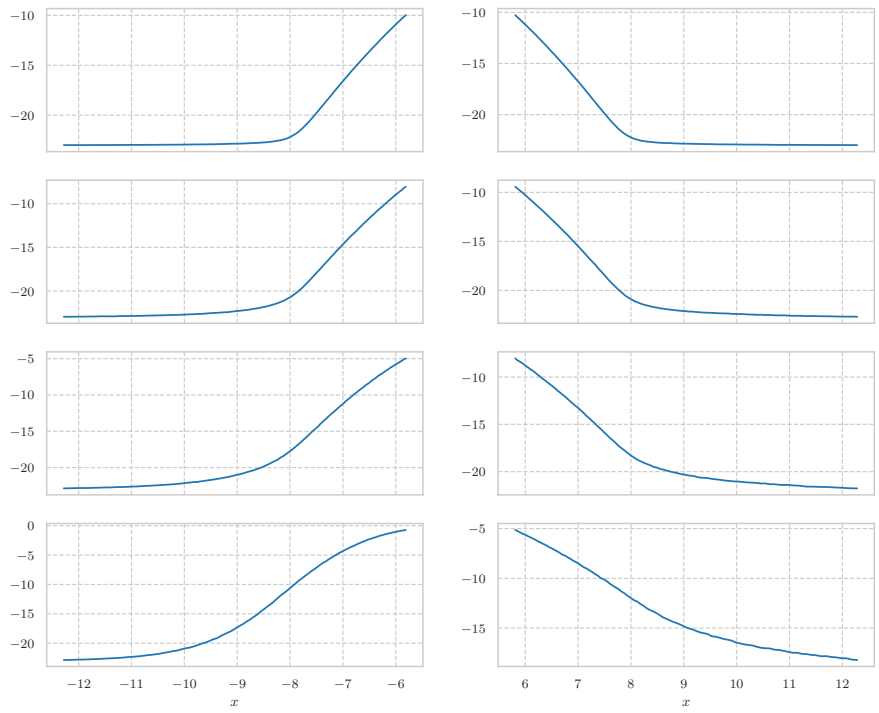

Figure 30: Plots of Monte Carlo approximations of the quantity given by (93), for x approximately in $[-12.28, -5.81)$ and x approximately in $(5.81, 12.28]$ in the first and second column respectively. Each row corresponds to a different choice of the jump-size, with $\epsilon = 1.2$ for the top row, and $\epsilon = 1.0$, $\epsilon = 0.8$, $\epsilon = 0.5$ for the second, third and fourth row respectively. The Monte Carlo approximation to the expectations have been calculated with one-hundred thousand samples.

the Exchangeable Sampler, consider, again, Example 5. Figures 32, 33, and 34 show, respectively, the behaviour of the Exchangeable sampler with $N = 1$, $N = 10$, and $N = 100$. In each figure the behaviour of the sampler has been illustrated for $\epsilon \in \{0.05, 0.2, 0.5, 1.0, \sqrt{2}\}$.

It can be seen from Figure 32 that, in the case of $N = 1$, when the jump-size, ϵ, is small, the acceptance rate- seen in the penultimate column— is very close to one and the expected squared jump distance— seen in the last column— is close to zero (as can be seen in the top row of the figure). Moreover, in the independent case; that is, when the jump-size is equal to $\sqrt{2}$, the acceptance rate is close to zero and the expected squared jump distance is close to zero (as can be seen in the bottom row of the figure). In both cases the chain does not mix well— as can be seen from the second column— and the density of samples do not represent the true density particularly well —as can be seen in the first column. However, when ϵ is chosen to be of an appropriate size ($\epsilon \in \{0.2, 0.5, 1.0\}$), the acceptance rate is neither close to zero or one, the expected squared jump distance is relatively large, the chain mixes relatively well, and the density of the samples represent the true density well (as can be seen in the middle three rows of the figure). It can be

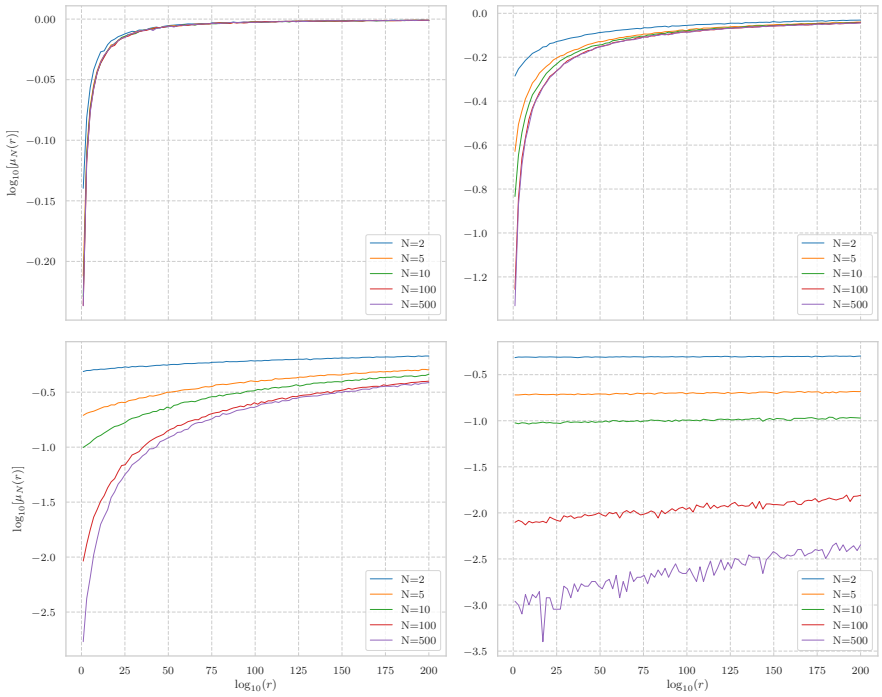

Figure 31: Plots of Monte Carlo approximations of the quantity given by (92) over the interval $\log(r) \in [0, 200]$ for a range of values for N. Each plot corresponds to a different jump-size, with $\epsilon = 1.2$ for the top-left plot, and $\epsilon = 1.0$, $\epsilon = 0.8$, $\epsilon = 0.5$ for the top-right, bottom-left and bottom-right plot respectively. The Monte Carlo approximation to the probabilities have been calculated with one-hundred thousand samples and the infimum has been empirically approximated by taking the minimum over a range of values of x in the interval $(-1000, 1000)$.

seen from Figures 33 and 34 that the larger the number of samples; that is, the larger the value for N, the larger the exepcted squared jump distance one can achieve with the Exchangeable Sampler, the larger the *optimal* acceptance rate, and the larger the *optimal* scaling. It can be seen from all three figures that the expected squared jump distance converges to the limiting expected squared jump distance fairly quickly. As was the case for the Random-Walk sampler in Section 2.3.6.2, this suggests that the expected squared jump distance is a useful measure to monitor when tuning the Exchangeable Sampler.

4.3.1 *Optimal Scaling*

Given the observations at the end of the previous section, and the comparisons to the Random-Walk sampler of Section 2.3.6.2, it is natural to ask, for the Exchangeable Sampler, if one can derive *optimal scaling* results which are in the same spirit as the optimal scaling results of Section 2.3.6.3; that is, results which practitioners can use as a general guide on how to choose the jump-size so as to maximize the rate of mixing of the chain. We follow a similar approach to that taken by Sherlock

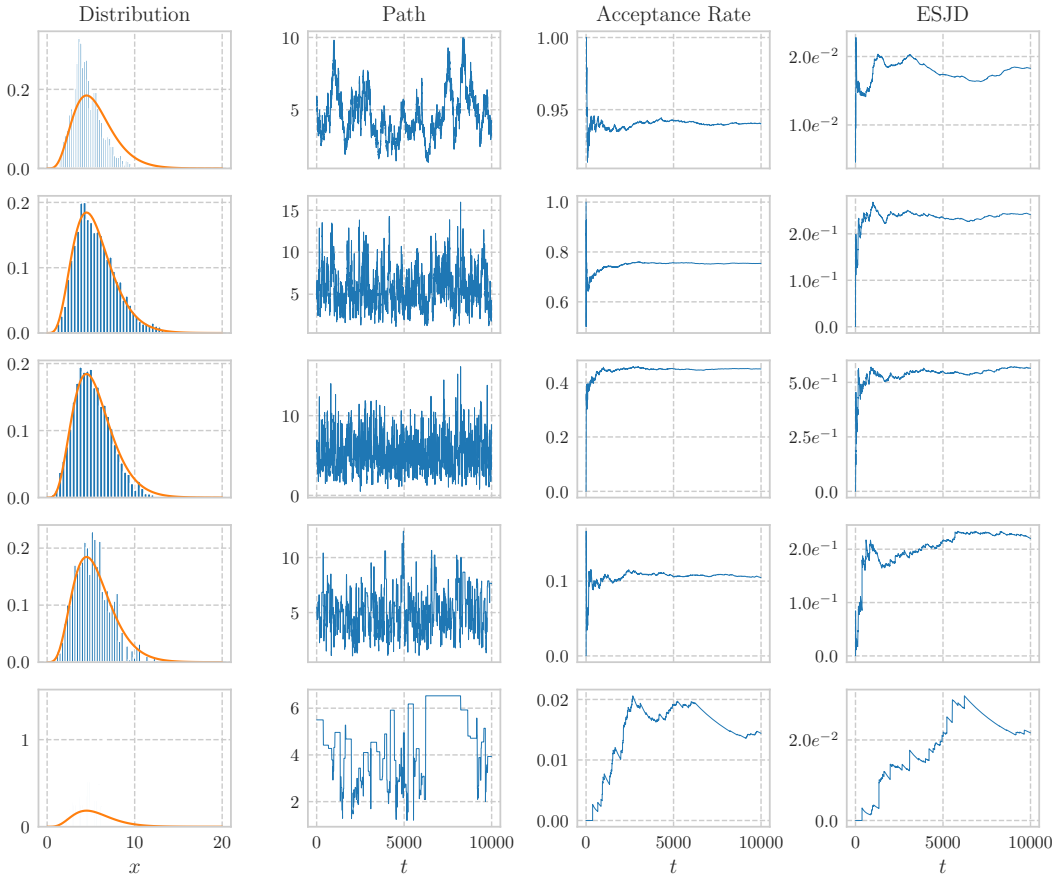

Figure 32: An illustration of the behaviour of the Exchangeable sampler, with $N = 1$, which targets a Gamma$(5.5, 1.0)$ distribution using a Gamma$(0.5, 1.0)$ distribution as the marginal proposal q_0. The five rows, top to bottom, correspond to jump-sizes $\epsilon = 0.05$, $\epsilon = 0.2$, $\epsilon = 0.5$, $\epsilon = 1.0$, and $\epsilon = \sqrt{2}$ respectively, and the samplers were run for ten-thousand iterations. The first column shows histograms of the simulated samples, with a plot of the true target density super-imposed. The second column shows the evolution of the chain. The third and fourth columns show the evolution of the acceptance rate and the expected squared jump distance respectively.

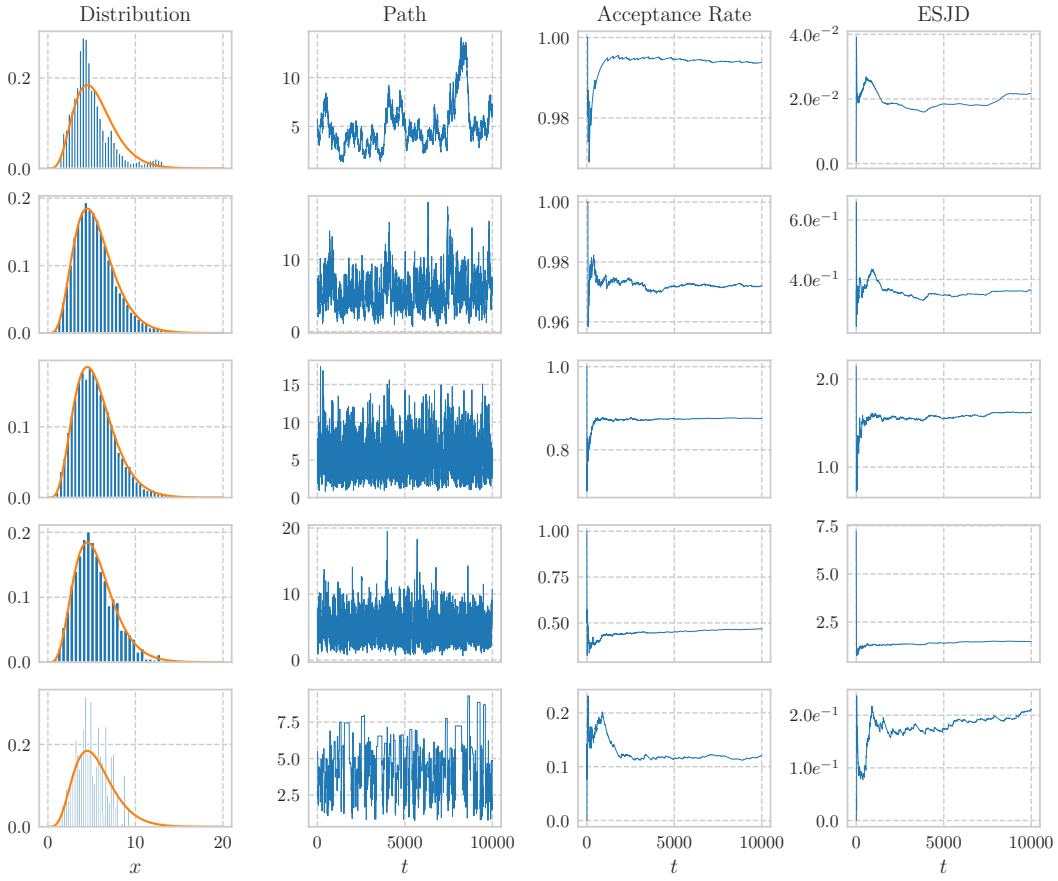

Figure 33: An illustration of the behaviour of the Exchangeable sampler, with $N = 10$, which targets a Gamma$(5.5, 1.0)$ distribution using a Gamma$(0.5, 1.0)$ distribution as the marginal proposal q_0. The five rows, top to bottom, correspond to jump-sizes $\epsilon = 0.05$, $\epsilon = 0.2$, $\epsilon = 0.5$, $\epsilon = 1.0$, and $\epsilon = \sqrt{2}$ respectively, and the samplers were run for ten thousand iterations. The first column shows histograms of the simulated samples, with a plot of the true target density super-imposed. The second column shows the evolution of the chain. The third and fourth columns show the evolution of the acceptance rate and the expected squared jump distance respectively.

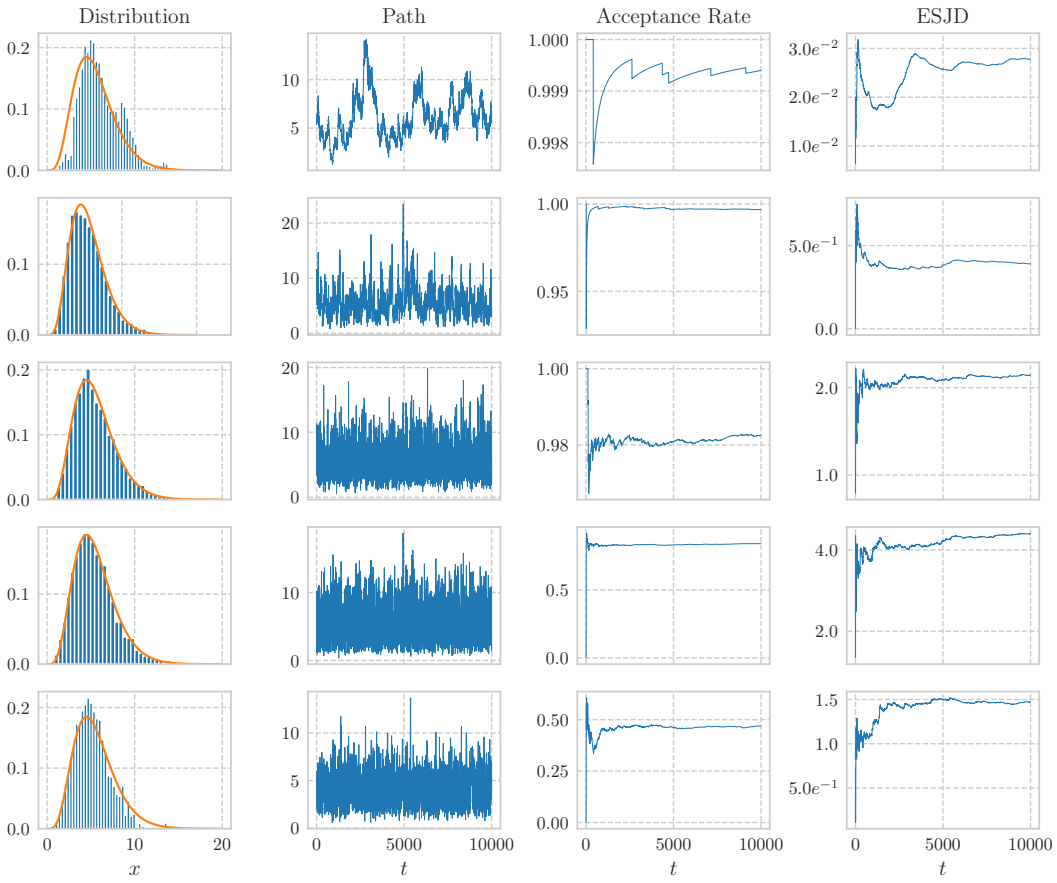

Figure 34: An illustration of the behaviour of the Exchangeable sampler, with $N = 100$, which targets a Gamma$(5.5, 1.0)$ distribution using a Gamma$(0.5, 1.0)$ distribution as the marginal proposal q_0. The five rows, top to bottom, correspond to jump-sizes $\epsilon = 0.05$, $\epsilon = 0.2$, $\epsilon = 0.5$, $\epsilon = 1.0$, and $\epsilon = \sqrt{2}$ respectively, and the samplers were run for ten thousand iterations. The first column shows histograms of the simulated samples, with a plot of the true target density super-imposed. The second column shows the evolution of the chain. The third and fourth columns show the evolution of the acceptance rate and the expected squared jump distance respectively.

and Roberts, 2009, who maximize the expected squared jump distance as the measure of efficiency. Unfortunately, the transformation, h, which transforms the variables $z_{1:N}$ to proposals $y_{1:N}$— see Algorithm 18— makes theoretically analysing the *expected squared jump distance* on the state space of the chain difficult. Instead, then, we consider the expected squared jump distance on the space in which the variables $z_{1:N}$ lie. To see that this is a reasonable metric to evaluate, consider, again, applying the Metropolis-Hastings Exchangeable Sampler to Example 5. Figure 35 shows plots of the expected squared jump distance in the original, *X-space*, against the expected squared jump distance in the underlying, *Z-space*, for a range of jump-sizes $\epsilon \in (0, \sqrt{2})$, and for a range of values of N. Each expected squared jump distance was calculated from the output of the sampler run for ten-thousand iterations. The figure suggests that, for this example, the expected squared jump distances, although on different scales, are closely correlated with one another and, so, in terms of a metric for measuring the performance of the sampler, the two can be used interchangeably. Moreover, Figure 36 illustrates plots of the acceptance rate against the the expected squared jump distances in the X-space and Z-space of the Metropolis-Hastings Exchangeable sampler with a range of values for N. The acceptance rates and expected squared jump distances were calculated for a range of jump-sizes $\epsilon \in (0, \sqrt{2})$ and the samplers were run for ten-thousand iterations. The figure suggests that the expected squared jump distance as a function of the acceptance rate has similar properties in both the X-space and Z-space. In particular, both achieve their optimum at around the same point and both are fairly insensitive to the acceptance rate around the optimum. Indeed, for $N = 1$, any acceptance rate in $[0.18, 0.66]$ achieves an expected squared jump distance in both the X-space and Z-space which is above 60% of the maximum. For $N = 10$, $N = 50$, and $N = 100$ this *insensitivity* interval becomes $[0.41, 0.9]$, $[0.54, 0.96]$, and $[0.62, 0.97]$ respectively. Therefore, for this example, monitoring either the expected squared jump distance in the X-space or the expected sqaured jump distance in the Z-space, as a function of the acceptance rate, will lead to similar conclusions and, thus, again, in terms of a metric for measuring the performance of the sampler, the two can be used interchangeably. Of course, in general, such a metric will only make sense if the transformation, h, is suitably smooth, so that *closeness* in the Z-space will result in *closeness* in the X-space. Indeed, from Section 2.1.2, Lipschitz continuity of h with a suitably small Lipschitz constant would be sufficient to justify the use of such a metric.

In order to theoretically analyse the expected squared jump distance and the expected acceptance rate in the Z-space, we consider a specific form of the Exchangeable Sampler which targets a density, π^*, of a product form by using a marginal density, q_0^*, which is also of a product form:

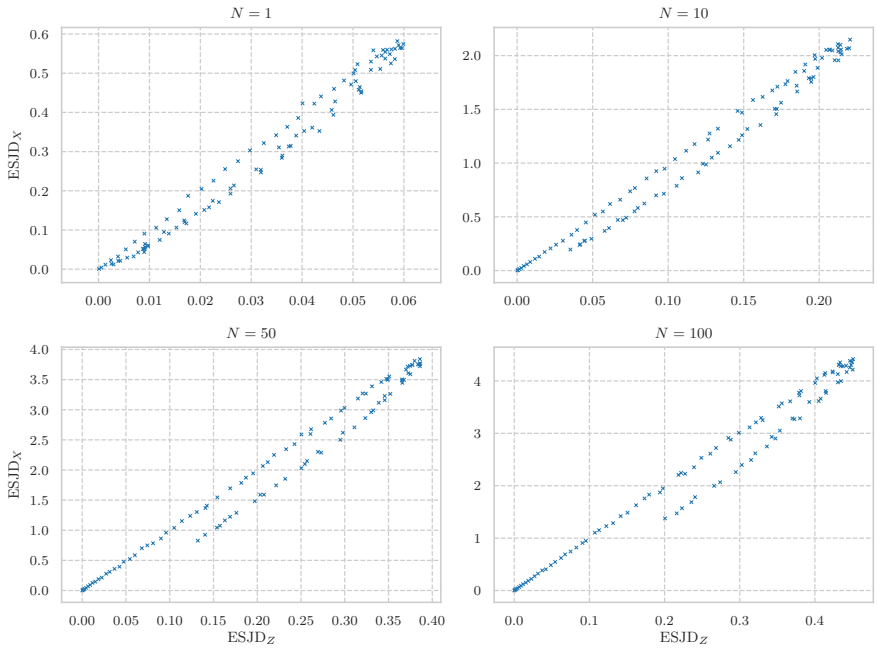

Figure 35: An illustration of the correlation between the expected squared jump distances in the X-space and Z-space of the Metropolis-Hastings Exchangeable sampler, with a range of values for N, which targets a Gamma$(5.5, 1.0)$ distribution using a Gamma$(0.5, 1.0)$ distribution as the marginal proposal q_0. The expected squared jump distances were calculated for a range of jump-sizes $\epsilon \in (0, \sqrt{2})$ and the samplers were run for ten-thousand iterations.

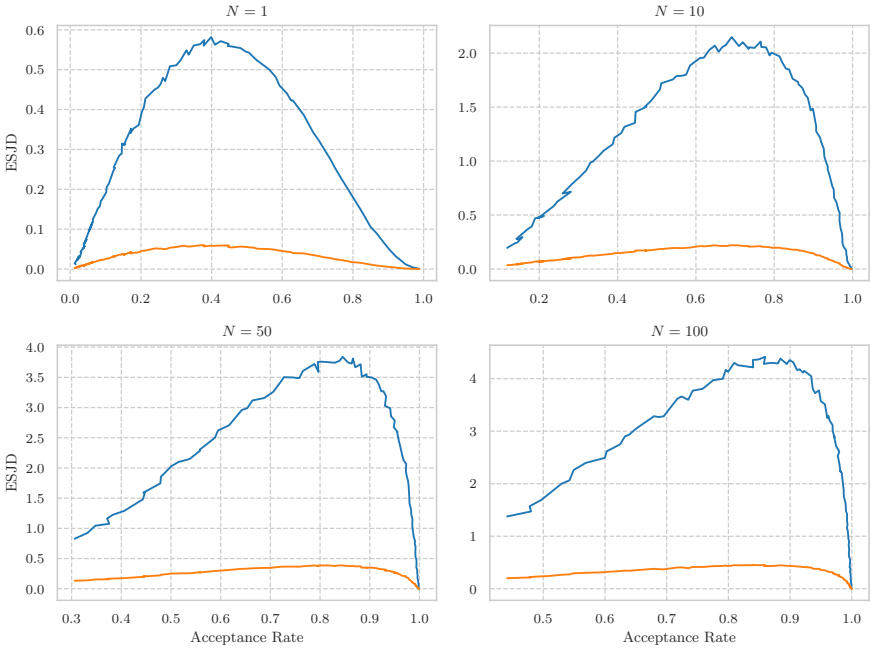

Figure 36: Plots of the acceptance rate against the the expected squared jump distances in the X-space (orange line) and Z-space (blue line) of the Metropolis-Hastings Exchangeable sampler, with a range of values for N, which targets a Gamma$(5.5, 1.0)$ distribution using a Gamma$(0.5, 1.0)$ distribution as the marginal proposal q_0. The acceptance rates and expected squared jump distances were calculated for a range of jump-sizes $\epsilon \in (0, \sqrt{2})$ and the samplers were run for ten-thousand iterations.

DEFINITION 4.3.11. *Let the number of samples, $N \in \mathbb{N}$, the dimension, d, and the jump-size, $\epsilon \in (0, \sqrt{2})$, be fixed. Consider the Exchangeable Sampler given by Algorithm 16 which targets the density*

$$\pi^*(x^{(1:d)}) := \prod_{i=1}^{d} \pi(x^{(i)}) ,$$

by using the marginal proposal density

$$q_0^*(x^{(1:d)}) := \prod_{i=1}^{d} q_0(x^{(i)}) ,$$

as part of Algorithm 18 to generate proposals. Specifically, let $X^{(1:d)}$ be an independent sequence of random variables where each $X^{(i)} \sim \pi$. Suppose q_0 is a one-dimensional density with cumulative density function Q_0. Let the transformation, h^*, be given by

$$h^*(z^{(1:d)}) := (Q_0^{-1}[\Phi(z^{(1)})], \ldots, Q_0^{-1}[\Phi(z^{(d)})]) ,$$

where Φ denotes the cumulative density function corresponding to a standard normal random variable. Then, by Theorem 2.3.2, if $Z^{(1:d)} \sim \mathrm{N}_d(0, I_d)$, then $h^*(Z^{(1:d)}) \sim q_0^*$. Now, let $h := Q_0^{-1} \circ \Phi$, so that, if $X^{(i)} \sim \pi$, then $Z^{(i)} = h^{-1}(X^{(i)})$ has density $\pi_Z(z^{(i)}) := \pi[h(z^{(i)})]|h'(z^{(i)})|$. Note that the transformation h^* satisfies the necesarry assumptions of the exchangeable proposal given by Algorithm 18. Suppose $Z_0^{(1:d)} \sim \pi_Z^*$, where

$$\pi_Z^*(z_0^{(1:d)}) := \prod_{i=1}^{d} \pi_Z(z_0^{(i)}) ,$$

and let $\hat{Z}_{0:N}^{(1:d)}$ be an independent sequence of d-dimensional random variables such that, for any $k \in \{0, \ldots, N\}$, $\hat{Z}_k^{(1:d)} \sim \mathrm{N}_d(0, I_d)$. For each $k \in \{1, \ldots, N\}$ define

$$Z_k^{(1:d)} := (1 - \delta^2) Z_0^{(1:d)} + \delta \sqrt{1 - \delta^2} \hat{Z}_0^{(1:d)} + \delta \hat{Z}_k^{(1:d)} ,$$

where $\delta := \epsilon / \sqrt{2}$. Furthermore, for any $k \in \{1, \ldots, N\}$, let $\alpha_{k,N}^*(w_{0:N}^*)$ be the multiple-proposal extension of either Barker's acceptance probability (Equation (86)) or the Metropolis-Hastings acceptance probability (Equation (87)) expressed in terms of the transition weights, which are of the form $w^* = \pi^*/q_0^*$; that is, either

$$\alpha_{k,N}^*(w_{0:N}^*) = \frac{w_k^*}{w_0^* + \ldots + w_N^*} , \quad \text{or,} \quad \alpha_{k,N}^*(w_{0:N}^*) = \frac{w_k^*}{w_0^* + \ldots + w_N^* - [w_k^* \wedge w_0^*]} .$$

Let $g^* := w^* \circ h^*$; that is,

$$g^*(z^{(1:d)}) := \prod_{i=1}^{d} w[h(z^{(i)})] ,$$

where $w = \pi/q_0$ is the marginal transition weight. Finally, let $g := w \circ h$. Then, we define the expected squared jump distance to be

$$J_N(\epsilon) := \mathbb{E}\left[\sum_{k=1}^{N} \alpha^*_{k,N}(g^*(Z_0^{(1:d)}), \ldots, g^*(Z_N^{(1:d)}))\|Z_k^{(1:d)} - Z_0^{(1:d)}\|^2\right], \tag{94}$$

and the expected acceptance rate to be

$$\alpha_N(\epsilon) := \mathbb{E}\left[\sum_{k=1}^{N} \alpha^*_{k,N}(g^*(Z_0^{(1:d)}), \ldots, g^*(Z_N^{(1:d)}))\right]. \tag{95}$$

To derive an optimal scaling result for the expected squared jump distance and expected acceptance rate, given by Equations (94) and (95) respectively, the following assumptions on the density π_Z and the transition weight expressed in terms of z; that is, h^*, are needed:

ASSUMPTIONS 4.3.12.

(B) Let $Z \sim \pi_Z$, where π_Z is as defined in Definition 4.3.11:

(B.a) For any $k \in \{1, \ldots, 4\}$, $\mathbb{E}[|Z|^k] < \infty$.

(B.b) The logarithm of the marginal transition weight expressed in terms of z; that is, $p(z) := \log[g(z)]$, where $g := w \circ h$, and w, h are defined in Definition 4.3.11, is twice differentiable and satisfies

$$\mathbb{E}[p'(Z)^2] < \infty\,, \quad \mathbb{E}[p''(Z)^2] < \infty\,.$$

(L) The second derivative of p is Lipschitz continuous with Lipschitz constant a; that is, for any $z_{0:1} \in h^{-1}(\mathcal{X}) \times h^{-1}(\mathcal{X})$,

$$|p''(z_1) - p''(z_0)| \leq a|z_1 - z_0|\,.$$

(G) The gradient of the transition weight expressed in terms of z; that is, g', is sufficiently well-behaved in the tails in the sense that

$$\lim_{z \uparrow \infty} g'(z)\phi(z) = \lim_{z \downarrow -\infty} g'(z)\phi(z) = 0\,,$$

where ϕ is the density of a standard normal random variable.

Before stating the main theorem, a few preliminary results are presented. The first derives a relationship between the first and second derivative of the logarithm of the transition weight expressed in terms of z:

LEMMA 4.3.13. *Let $p(z) := \log[g(z)]$, where $g := w \circ h$, and h, w are defined in Definition 4.3.11. Then, under Assumptions 4.3.12,*

$$\mathbb{E}[p'(Z)^2] = -\mathbb{E}[p''(Z) - Zp'(Z)]\,,$$

where $Z \sim \pi_Z$, and π_Z is as defined in Definition 4.3.11.

Proof. See A.19. □

The second Lemma derives an approximation on the difference between the logarithm of the marginal density π at a current state and a proposed state:

LEMMA 4.3.14. *Let $\hat{Z}_{0:1}, Z_0$ be an independent tuple of random variables where $\hat{Z}_i \sim \mathrm{N}(0,1)$ for $i \in \{0,1\}$, and $Z_0 \sim \pi_Z$, where π_Z is defined in Definition 4.3.11. Let*

$$Z_1 := (1-\delta^2)Z_0 + \delta\sqrt{1-\delta^2}\hat{Z}_0 + \delta\hat{Z}_1 ,$$

for some $\delta \in (0,1)$, and $p(z) := \log[g(z)]$, where $g := w \circ h$, and h, w are defined in Definition 4.3.11. Then, under Assumptions 4.3.12,

$$p(Z_1) - p(Z_0) = \delta C_1(\hat{Z}_{0:1}, Z_0) + \delta^2 C_2(\hat{Z}_{0:1}, Z_0) + \delta^3 R(\hat{Z}_{0:1}, Z_{0:1}, \delta) ,$$

where

$$C_1(\hat{Z}_{0:1}, Z_0) := p'(Z_0)(\hat{Z}_0 + \hat{Z}_1) , \quad C_2(\hat{Z}_{0:1}, Z_0) := \frac{1}{2}(\hat{Z}_0 + \hat{Z}_1)p''(Z_0) - Z_0 p'(Z_0) ,$$

and $|R(\hat{Z}_{0:1}, Z_{0:1}, \delta)| \leq R^*(\hat{Z}_{0:1}, Z_0)$ where R^* is independent of Z_1 and δ, and $\mathbb{E}[R^*(\hat{Z}_{0:1}, Z_0)] < \infty$.

Proof. See A.20. □

Using these two lemmas, a third lemma decomposes the difference between the logarithm of the weight, expressed in terms of z, at any of the N proposed states and the logarithm of the weight, expressed in terms of z, at a current state into two random variables whose limiting behaviour as d tends towards infinity, and as δ is scaled appropriately, is known:

LEMMA 4.3.15. *Let $p(z) := \log[g(z)]$, where $g := w \circ h$, and h, w are defined in Definition 4.3.11. Moreover, let $(Z_0^{(1:d)}, \hat{Z}_{0:N}^{(1:d)})$ be an independent sequence of random variables, where, for any $k \in \{0, \ldots, N\}$, $\hat{Z}_k^{(1:d)} \sim \mathrm{N}_d(0, I_d)$, and, for any $i \in \{1, \ldots, d\}$, $Z_0^{(i)} \sim \pi_Z$, where π_Z is as defined in Definition 4.3.11. For each $k \in \{1, \ldots, N\}$ define*

$$Z_k^{(1:d)} := (1-\delta_d^2)Z_0^{(1:d)} + \delta_d\sqrt{1-\delta_d^2}\hat{Z}_0^{(1:d)} + \delta_d\hat{Z}_k^{(1:d)} ,$$

where $\delta_d := \lambda d^{-1/2}/\sqrt{2}$ for some $\lambda > 0$. Define $\varphi := \mathbb{E}[p'(Z_0^{(1)})^2]$. Then, for any $k \in \{1, \ldots, N\}$, we have, under Assumptions 4.3.12,

$$\sum_{i=1}^{d}[p(Z_k^{(i)}) - p(Z_0^{(i)})] = D_k(d) + U_k(d) ,$$

where $\{D_k(d) : d \in \mathbb{N}\}$ is a collection of random variables such that

$$\operatorname*{plim}_{d\uparrow\infty} D_k(d) = -\frac{\lambda^2\varphi}{2} ,$$

and $\{U_k(d) : d \in \mathbb{N}\}$ *is a collection of random variables such that*

$$\operatorname*{dlim}_{d\uparrow\infty}(U_1(d), \ldots, U_N(d)) = (U_1, \ldots, U_N) \ ,$$

where, for each $k \in \{1, \ldots, N\}$, $U_k := A + B_k$, *and* $(A, B_{1:N})$ *is a collection of independent random variables where*

$$A \sim \mathrm{N}\left(0, \frac{\lambda^2 \varphi}{2}\right) \ ,$$

and, for any $k \in \{1, \ldots, N\}$,

$$B_k \sim \mathrm{N}\left(0, \frac{\lambda^2 \varphi}{2}\right) \ .$$

Proof. See A.21. □

A fourth lemma derives two results concerning the limit of an appropriately scaled *jump*:

LEMMA 4.3.16. *Let* $(Z_0^{(1:d)}, \hat{Z}_{0:1}^{(1:d)})$ *be an independent sequence of random variables, where, for any* $k \in \{0, 1\}$, $\hat{Z}_k^{(1:d)} \sim \mathrm{N}_d(0, I_d)$, *and, for any* $i \in \{1, \ldots, d\}$, $Z_0^{(i)} \sim \pi_Z$, *where* π_Z *is as defined in Definition 4.3.11. Define*

$$Z_1^{(1:d)} := (1 - \delta_d^2) Z_0^{(1:d)} + \delta_d \sqrt{1 - \delta_d^2} \hat{Z}_0^{(1:d)} + \delta_d \hat{Z}_1^{(1:d)} \ ,$$

where $\delta_d := \lambda d^{-1/2}/\sqrt{2}$ *for some* $\lambda > 0$. *Then, under Assumptions 4.3.12,*

$$\lim_{d\uparrow\infty} \mathbb{E}[\|Z_1^{(1:d)} - Z_0^{(1:d)}\|^2] = \lambda^2 \ , \quad \lim_{d\uparrow\infty} \mathbb{E}[(\|Z_1^{(1:d)} - Z_0^{(1:d)}\|^2 - \lambda^2)^2] = 0 \ .$$

Proof. See A.22. □

With these preliminary results in place, an optimal scaling result for the Exchangeable Sampler with either Barker's or the Metroplis-Hastings acceptance probability can be demonstrated;

THEOREM 4.3.17. *Consider the Exchangeable Sampler given in Definition 4.3.11 which targets a density,* π^*, *of a product form by using a marginal density,* q_0^*, *which is also of a product form. Suppose that Assumptions 4.3.12 hold. Then, using the multiple-proposal extension of Barker's acceptance probability (Equation (86)),*

$$\lim_{d\uparrow\infty} \alpha(\lambda d^{-1/2}) = \bar{\alpha}_b(\lambda) := 1 - \mathbb{E}\left[\left\{1 + \exp(-\xi^2)\exp(\xi W_0)\sum_{k=1}^{N}\exp(\xi W_k)\right\}^{-1}\right] \ , \tag{96}$$

where $\xi := \lambda\sqrt{\varphi}/\sqrt{2}$, φ *is defined in Lemma 4.3.15, and* $W_{0:N}$ *is an independent sequence of one-dimensional standard Normal random variables. Moreover,*

$$\lim_{d\uparrow\infty} J(\lambda d^{-1/2}) = \bar{J}_b(\lambda) := \lambda^2 \bar{\alpha}_b(\lambda) \ . \tag{97}$$

Using the multiple-proposal extension of the Metropolis-Hastings acceptance probability (Equation (87)),

$$\begin{aligned}\lim_{d\uparrow\infty} \alpha(\lambda d^{-1/2}) &= \bar{\alpha}_m(\lambda) \\ &:= \sum_{k=1}^{N} \frac{\exp(\xi W_k)}{\exp(\xi^2 - \xi W_0) + s(W_{1:N}) - [\exp(\xi^2 - \xi W_0) \wedge \exp(\xi W_k)]} ,\end{aligned} \tag{98}$$

where

$$s(w_{1:N}) := \sum_{j=1}^{N} \exp(\xi W_j) ,$$

and, as previously, $\xi := \lambda\sqrt{\varphi}/\sqrt{2}$*,* φ *is defined in Lemma 4.3.15, and* $W_{0:N}$ *is an independent sequence of one-dimensional standard Normal random variables. Moreover,*

$$\lim_{d\uparrow\infty} J(\lambda d^{-1/2}) = \bar{J}_m(\lambda) := \lambda^2 \bar{\alpha}_m(\lambda) . \tag{99}$$

Proof. See A.23. □

When $N = 1$,

$$\bar{\alpha}_b(\lambda) = \mathbb{E}\left[\frac{\exp(-\xi^2)\exp(\sqrt{2}\xi W)}{1 + \exp(-\xi^2)\exp(\sqrt{2}\xi W)}\right] ,$$

where $W \sim \mathrm{N}(0, 1)$. Moreover,

$$\begin{aligned}\bar{\alpha}_m(\lambda) &= \mathbb{E}\left[\frac{\exp(\xi W_1)}{\exp(\xi^2 - \xi W_0) + \exp(\xi W_1) - [\exp(\xi^2 - \xi W_0) \wedge \exp(\xi W_1)]}\right] \\ &= \mathbb{E}\left[\frac{\exp(-\xi^2)\exp(\sqrt{2}\xi W)}{1 + \exp(-\xi^2)\exp(\sqrt{2}\xi W) - [1 \wedge \exp(-\xi^2)\exp(\sqrt{2}\xi W)]}\right] \\ &= \mathbb{E}\left[1 \wedge \exp(-\xi^2)\exp(\sqrt{2}\xi W)\right] ,\end{aligned}$$

where $W \sim \mathrm{N}(0, 1)$. Thus, in the case of $N = 1$, the optimal scaling result for the Exchangeable Sampler, given by Theorem 4.3.17, matches the optimal scaling result for the Random-Walk Sampler, given by Theorem 2.3.36. Therefore, the results of Corollaries 2.3.37 and 2.3.38 immediately follow. Unfortunately, for general $N \in \mathbb{N}$, the asymptotic quantities given by Theorem 4.3.17 are intractable. However, Figures 37 and 38 show plots of the asymptotic expected efficiency[5]; that is, the asymptotic expected squared jump distance over the number of particles, up to a constant of proportionality, against the asymptotic acceptance rate for Barker's Exchangeable Sampler and the Metropolis-Hastings Exchangeable Sampler respectively. Each figure shows the relationship for $N \in \{1, 10, 100, 1000\}$. Note that, for both samplers, the

5 This definition of efficiency assumes that the computational complexity scales linearly with N.

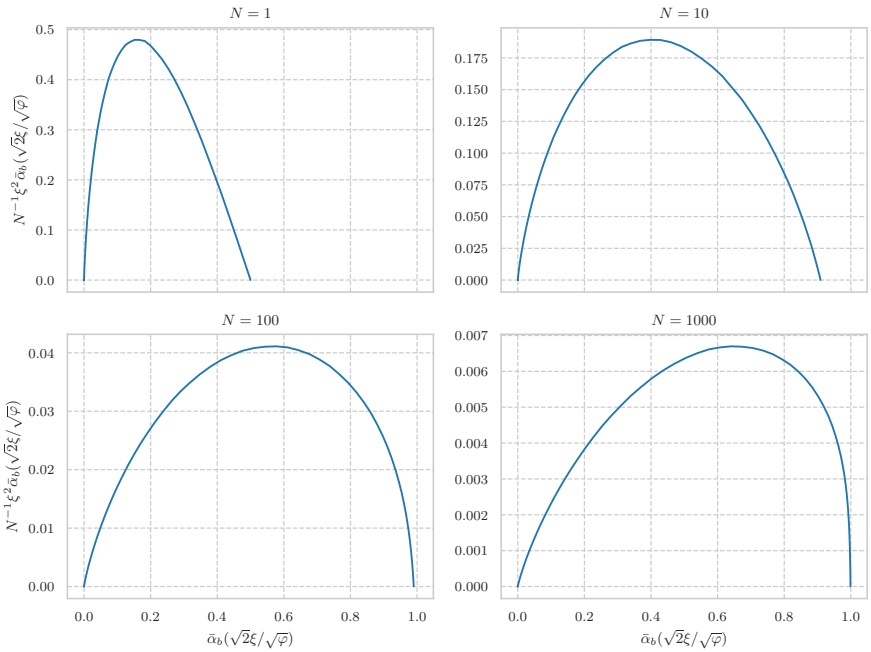

Figure 37: Plots of the asymptotic expected efficiency, up to a constant of proportionality, against the asymptotic acceptance rate for Barker's Exchangeable Sampler for a range of values for N.

optimal acceptance rate gets closer to one as the value for N increases. Also, the optimal efficiency is the largest, over the values of N considered, for $N = 1$. Moreover, for both cases, and for each value of N, the optimal asymptotic expected efficiency is fairly insensitive to choices of the asymptotic acceptance rate around the optimum. Indeed, consider Barker's Exchangeable Sampler. For $N = 1$, an asymptotic acceptance rate in the interval $[0.04, 0.33]$ leads to an asymptotic expected efficiency which is above 60% of the optimal. For $N = 10$, $N = 100$, and $N = 1000$ this interval becomes $[0.12, 0.73]$, $[0.21, 0.9]$, and $[0.23, 0.95]$ respectively. On the other hand, for the Metropolis-Hastings Sampler with $N = 1$, any asymptotic acceptance rate in the interval $[0.06, 0.53]$ leads to an asymptotic expected efficiency which is above 60% of the optimal. For $N = 10$, $N = 100$, and $N = 1000$ this interval becomes $[0.13, 0.81]$, $[0.19, 0.9]$, and $[0.23, 0.96]$ respectively. As such, as was the case for the Random-walk Sampler, it is unnecessary to finely tune the *jump-size* to achieve the optimal acceptance rate provided the *tuned* acceptance rate is on the same scale as the optimal asymptotic acceptance rate. Although these observations are purely theoretical, due to asymptotic nature of the conclusions and the strong assumptions on the target, the *insensitivity* to fine-tuning of the acceptance rate can be seen in Example 5, as was highlighted in Figures 32, 33, and 34, and discussed at the end of the last section.

Figures 37 and 38 highlight the theoretical behaviour of the asymptotic efficiency as a function of the asymptotic acceptance rate. While monitoring the *running sample efficiency* in order to tune the value of the jump-size is a sensible strategy one can use in practice to optimise

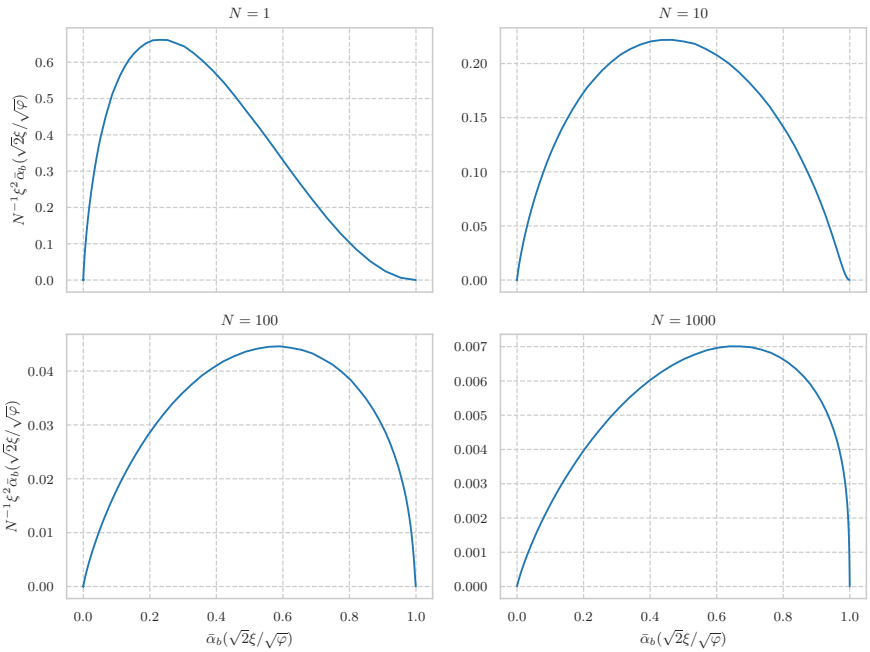

Figure 38: Plots of the asymptotic expected efficiency, up to a constant of proportionality, against the asymptotic acceptance rate for the Metropolis-Hastings Exchangeable Sampler for a range of values for N.

the mixing of the chain, it is prudent to understand to what extent such theoretical results hold for finite d. To this end, we will consider the simple task of targeting the d-dimensional posterior corresponding to the likelihood $(Y|X = x) \sim \mathrm{N}_d(x, 0.3\mathcal{I}_d)$ and a $\mathrm{N}_d([0.8, \ldots, 0.8], \mathcal{I}_d)$ prior, by using the Exchangeable Sampler (Algorithm 16) with the prior as the marginal proposal density. Here, $\mathcal{I}_d$ denotes the d-dimensional identity matrix. Note that this scenario is a d-dimensional extension of the Linear Gaussian model given by Example 2 in the case where X_0 is fixed to be 1 and $T = 1$; this observation is noteworthy as we will, in Section 4.4.1, use the same d-dimensional extension with $T > 1$ to numerically assess the theoretical optimal scaling results for the Exchangeable Particle Gibbs Sampler of Section 4.4— which we derive in that section— for finite d. The *transition* weight in this scenario is given by

$$w(x, y) \propto \exp(-10\|y - x\|^2/6) \, .$$

For each $d \in \{1, 2, 5, 10, 25, 50\}$, we set y to be the d-dimensional vector filled with 0.8; that is, $y = \mathbb{E}(Y)$, and simulated the Exchangeable Sampler for one-hundred-thousand iterations for each $N \in \{1, 10, 50, 100, 1000\}$ and for ten values of ϵ linearly spaced on the interval $[0.05, \sqrt{2}]$; that is $\epsilon \in \{0.05, 0.05 + (\sqrt{2} - 0.05)/9, 0.05 + 2(\sqrt{2} - 0.05)/9, \ldots, \sqrt{2}\}$. For each *run* of the sampler we calculated the acceptance rate and the expected squared-jump distance in the Z-space. Figures 39 and 40 show, respectively, plots of the sample efficiency against the sample acceptance rate for Barker's and the Metropolis-Hastings Exchangeable Sampler for this scenario and for $d \in \{1, 2, 5, 10, 25, 50\}$ and $N \in \{1, 10, 50, 100, 1000\}$.

The figures show that, up to a constant of proportionality, the behaviour of the sample efficiency against the sample acceptance rate for $d \in \{25, 50\}$ matches the theoretical behaviour, shown in Figures 37 and 38, exactly. However, for smaller d, this is not the case. Indeed, even though, for $d = 10$, the optimal sample acceptance rate occurs around the same place as the theoretical optimal acceptance rate, and the behaviour of the sample efficiency for higher sample acceptance rates matches the behaviour of the theoretical efficiency for higher acceptance rates, the curves do not cover the same range of acceptance rates as the theoretical curves do. Specifically, smaller sample acceptance rates are not achieved for $d = 10$. This lack of range for the sample acceptance rate, resulting in only partial matches of the *sample* curves to the theoretical curves, is exaggerated the smaller d is and the larger N is. This is because, when N is larger, one can choose a larger jump-size to get the same acceptance rate as for a smaller N, and, due to the nature of the Exchangeable Sampler, the maximum jump-size is pinned at $\epsilon = \sqrt{2}$. Indeed, for $d \in \{1, 2\}$, one should choose $\epsilon = \sqrt{2}$ in order to optimise the mixing of the chain regardless of the value of N. Note that, whatever the dimension, the plots suggest that the value of N which optimises the efficiency is always $N = 1$ which matches the advice derived from the theoretical results.

It is important to note that, in practice, the computational cost might not scale like N and so, for these scenarios, this definition of efficiency would not be the correct one. For example, one might be able to make use of parallel computations to simulate $N > 1$ particles simultaneously; that is, with approximately the same computationl cost as it would take to simulate one particle. In these scenarios, it is probable that the maximum number of particles one can simulate simultaneously will correspond to the value of N which optimises the *efficiency*.

4.3.2 *A Simulation Study*

In this section we will look at the performance of the Exchangeable Sampler in four examples. In the first example we will consider the case where the target, $\pi(x)$, is a $\mathrm{N}(0, 1)$ distribution, and, the marginal proposal, $q_0(x)$, is a $\mathrm{N}(0, 1/2)$ distribution. As highlighted in Example 4, in this case the transition weight, $w(x) = \pi(x)/q_0(x)$, is proportional to $\exp(3x^2/2)$ and, therefore, has exponentially increasing tails. As a result, one would expect that the sampler is not geometrically ergodic since, the further into the tails the sampler goes, the larger the weight, and the increase in the weight is exponential. Indeed, it was shown in Lemma 4.3.10 that, for this example, for any $\epsilon \in (0, \sqrt{2}]$, assumption (IM) of Corollary 4.3.7 is violated and, therefore, Corollary 4.3.9 does not apply. In the second example we will consider the case where the target, $\pi(x)$, is a $\mathrm{Gamma}(5.5, 1.0)$ distribution, and, the marginal proposal, $q_0(x)$, is a $\mathrm{Gamma}(0.5, 1.0)$ distribution. In this scenario, as in the first example, the tails of the proposal are lighter than the tails of the target. However, for this scenario, the tails are polynomially lighter

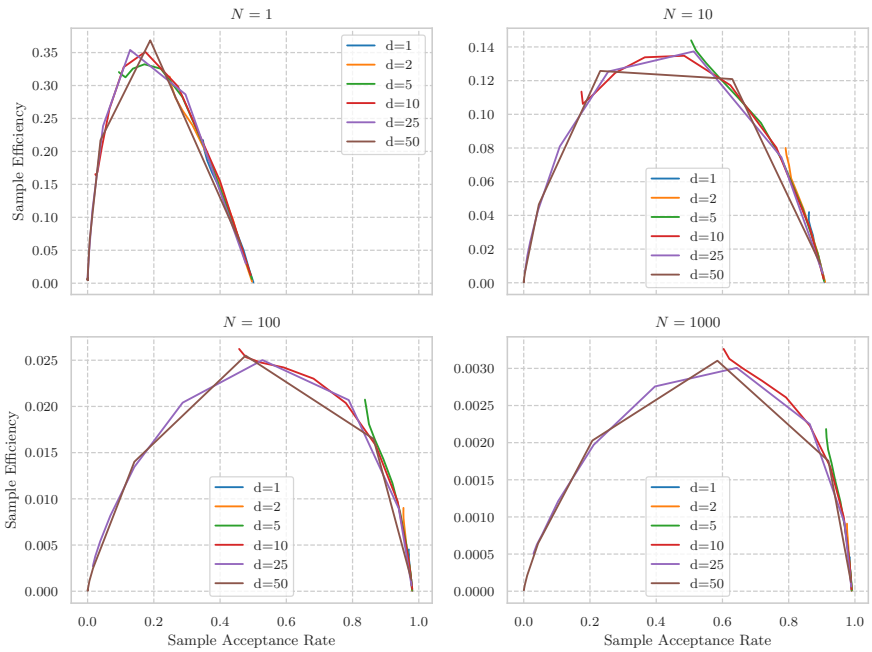

Figure 39: Plots of the sample efficiency against the sample acceptance rate for Barker's Exchangeable Sampler which targets the d-dimensional posterior corresponding to the likelihood $(Y|X = x) \sim \mathrm{N}_d(x, 0.3\mathcal{I}_d)$ and a $\mathrm{N}_d([0.8, \ldots, 0.8], \mathcal{I}_d)$ prior, by using the prior as the marginal proposal density. Each plot corresponds to a different value of $N \in \{1, 10, 50, 100, 1000\}$, and, for each N, we ran the Exchangeable Sampler for $d \in \{1, 2, 5, 10, 25, 50\}$ and for ten values of ϵ linearly spaced on the interval $[0.05, \sqrt{2}]$; that is $\epsilon \in \{0.05, 0.05 + (\sqrt{2} - 0.05)/9, 0.05 + 2(\sqrt{2} - 0.05)/9, \ldots, \sqrt{2}\}$.

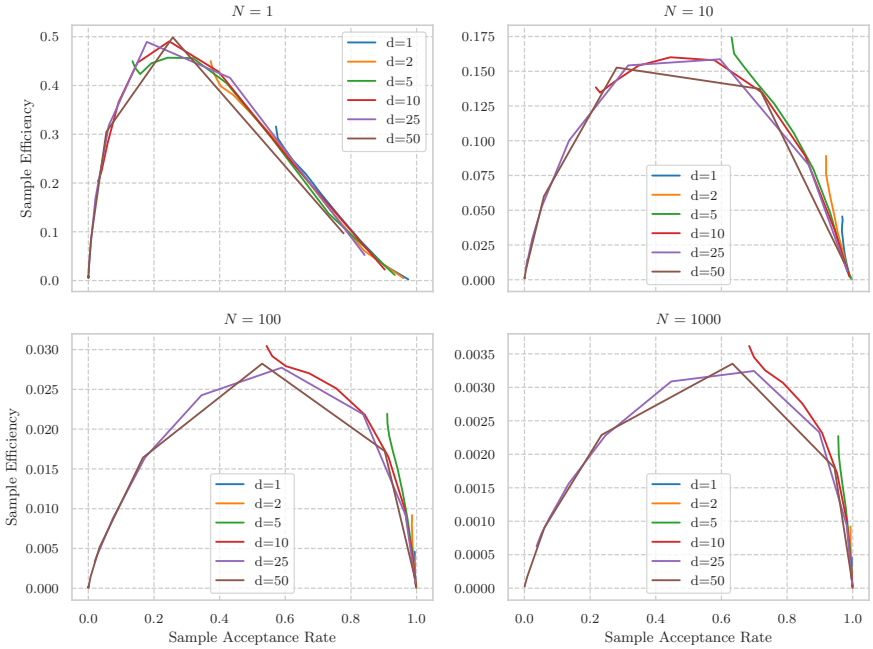

Figure 40: Plots of the sample efficiency against the sample acceptance rate for the Metropolis-Hastings Exchangeable Sampler which targets the d-dimensional posterior corresponding to the likelihood $(Y|X = x) \sim \mathrm{N}_d(x, 0.3\mathcal{I}_d)$ and a $\mathrm{N}_d([0.8, \ldots, 0.8], \mathcal{I}_d)$ prior, by using the prior as the marginal proposal density. Each plot corresponds to a different value of $N \in \{1, 10, 50, 100, 1000\}$, and, for each N, we ran the Exchangeable Sampler for $d \in \{1, 2, 5, 10, 25, 50\}$ and for ten values of ϵ linearly spaced on the interval $[0.05, \sqrt{2}]$; that is $\epsilon \in \{0.05, 0.05 + (\sqrt{2} - 0.05)/9, 0.05 + 2(\sqrt{2} - 0.05)/9, \ldots, \sqrt{2}\}$.

as opposed to exponentially lighter. Indeed, the transition weight is proportional to $\sqrt{x}$. As a result, one would expect that the sampler is geometrically ergodic since, even though the further into the tails the sampler goes the larger the weight, the rate of increase of the weight is only polynomial. The figures and discussion in Example 5 provide empirical evidence that the conditions of Theorem 4.3.9 hold for some values of ϵ and N and, therefore, that, for these values of ϵ and N, the Exchangeable Sampler is geometrically ergodic in this scenario and the MCMC estimates satisfy a central limit theorem for all functions which are square-integrable with respect to π. In the third example, we will consider the case where the target, $\pi(x)$, is a $\mathrm{N}(0,1)$ distribution, and, the marginal proposal, $q_0(x)$, is a $\mathrm{T}(5)$ distribution. In this scenario the transition weight is bounded. Indeed, the weight is proportional to

$$\exp\left(-\frac{x^2}{2}\right)\left(1+\frac{x^2}{\nu}\right)^{(1+\nu)/2},$$

which is a continuous function on $\mathbb{R}$ which tends towards zero as $|x|$ tends towards infinity. As a result, one would expect that the sampler is geometrically ergodic for any ϵ. The figures and discussion in Example 6 provide empirical evidence that the conditions of Theorem 4.3.9 hold for some values of ϵ and N but not others, and, therefore, that, for some values of ϵ and N, the Exchangeable Sampler is geometrically ergodic in this scenario and the MCMC estimates satisfy a central limit theorem for all functions which are square-integrable with respect to π. In the final example, we will consider the more realistic scenario where π corresponds to the conditioned Birth-Death diffusion of Section 3.1.1 and q_0 corresponds to the Modified Diffusion Bridge proposal of Section 3.2.3— see Chapter 3 for more details regarding simulating conditioned diffusions. Specifically, recall, from Section 3.2.3, that, in one-dimension, the MDB proposal of a discretised path of the diffusion, $x_{1:K}$, say, takes the form

$$q_0^{\mathrm{MDB}}(x_{1:K}|y) = \prod_{k=1}^{K} \phi(x_k; a_{k-1}^{\mathrm{MDB}}, C_{k-1}^{\mathrm{MDB}}),$$

where ϕ denotes the density corresponding to a one-dimensional normal distribution and a_{k-1}^{MDB} and C_{k-1}^{MDB} correspond to the mean (Equation (54)) and variance (Equation (55)) respectively, and, implicitly, depend on x_{k-1}, T, y, and t_{k-1}. Such a proposal is equivalent to proposing K independent $\mathrm{N}(0,1)$ random variables, $Z_{1:K}$, and transforming those random variables appropriately by sequentially setting, for $k \in \{1,\ldots,K\}$, $X_k = a_{k-1}^{\mathrm{MDB}} + \sqrt{C_{k-1}^{\mathrm{MDB}}} Z_k$. Thus, simulating *exchangeable* paths corresponds to simulating sequences of K independent $\mathrm{N}(0,1)$ random variables in an exchangeable way. For this example, as in Section 3.3.2, we use the same parameters, θ, and initial conditions, x_0, as those used in Whitaker et al., 2017; $\theta = (\theta_1, \theta_2) = (0.1, 0.8)$, $x_0 = 50$, so that sample paths of the diffusion exhibit exponential decay. Moreover, as

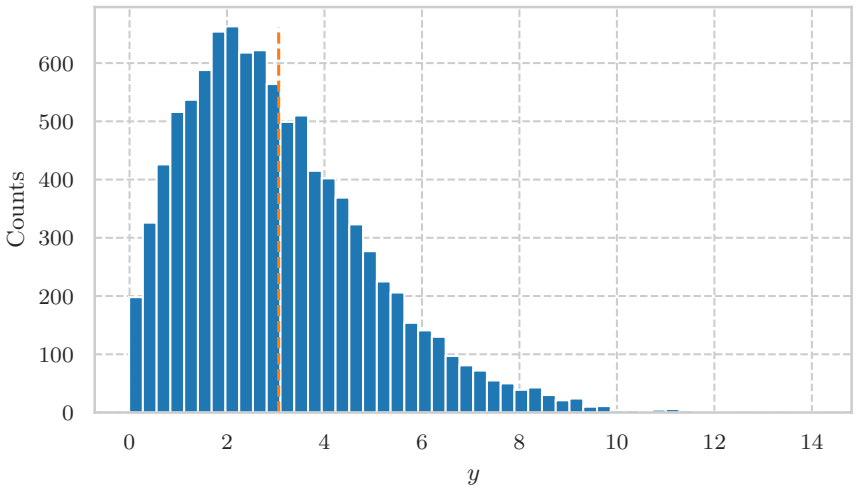

Figure 41: A histogram of the ten-thousand simulated observations, y_T, of the BD diffusion, where $T = 4$. The orange line shows the location of the mean observation.

in Section 3.3.2, we fix Δt to be 0.01, T to be 4, and choose $P_1 = I$, and $\Sigma_1 = 10^{-12}I$, so that the observation, Y, is such that

$$Y|X_K = x \sim \mathrm{N}(x, 10^{-12}I) \ ,$$

and, therefore, essentially corresponds to exact observations of the diffusion. To choose an observation to condition on, as we did in Section 3.3.2, we simulated ten-thousand values for $Y_T^{(1)}$ using the EM approximation to forward simulate values of the path at each point of the partition. We then chose the mean of the simulated terminal endpoints as the observation to condition upon. Figure 41 shows a histogram of the ten-thousand simulated observations, y_T, of the BD diffusion, where $T = 4$. The orange line shows the location of the mean observation.

For each scenario, we ran the Exchangeable Sampler for each of the forty combinations of

$$(\epsilon, N) \in \{0.01, 0.1, 0.25, 0.5, 0.75, 1.0, 1.2, \sqrt{2}\} \times \{1, 10, 50, 100, 1000\}$$

for one-hundred-thousand iterations. For all but the birth-death example, we initialised the Exchangeable Sampler at the mean of the target. For the birth-death example, we chose an initial path by simulating one-hundred paths via the MDB and choosing the path with the highest weight as our initial path. For each run of each example we calculated the sample efficiency; that is, the sample squared jump distance divided by the number of particles, in the X-space, along with the acceptance rate. For all but the Birth-Death diffusion example, we also calculate the Kolmogorov-Smirnov statistic (see Hollander and Wolfe, 1973, for example) between the simulated samples, $x_{1:M}$, and the true target, Π, as a measure of how *well* the simulated samples represent the truth:

$$\kappa(\Pi, x_{1:M}) := \sup_{x \in \mathbb{R}} \left| \Pi(x) - \frac{1}{M} \sum_{t=1}^{M} \mathbb{1}_{(-\infty, x]}(x_t) \right| . \tag{100}$$

For the Birth-Death diffusion example we simulate samples representing the *truth* by running an Independence Sampler (Section 2.3.6.1) with the residual-bridge construct of Whitaker et al., 2017, where $\xi_t = \mathbb{E}(\hat{R}_t|Y = y)$ and $\hat{R}_t$ is the process satisfying the diffusion of the Linear Noise Approxtimation; that is, satisfies the SDE 64, as the proposal. Specifically, recall, from Section 3.2.5, that, in one-dimension, the residual-bridge proposal of a discretised path of the diffusion, $x_{1:K}$, say, takes the form

$$q_0^{\mathrm{RB}}(x_{1:K}|y) = \prod_{k=1}^{K} \phi(x_k; a_{k-1}^{\mathrm{RB}}, D_{k-1}^{\mathrm{RB}}) \,,$$

where ϕ denotes the density corresponding to a one-dimensional normal distribution and a_{k-1}^{RB} and D_{k-1}^{RB} correspond to the mean (Equation (61)) and variance matrix (Equation (62)) respectively, and, implicitly, depend on x_{k-1}, T, y, ξ_{k-1}, ξ_K, and t_{k-1}. To get a good representation of the truth, we run the independence sampler for one-million iterations and, for simplicity and brevity, we focus on the two-hundreth element of each of the sample paths; x_{200}. Given the samples are *pinned* at both the start and end of the inter-observation period, it is reasonable to focus on samples at the middle of the inter-observation period as these will exhibit the most variation and, therefore, will be the hardest to represent. We calculate the two-sample Kolmogorov-Smirnov statistic (see Hollander and Wolfe, 1973, for example) between the two-hundreth element of the samples simulated via the Exchangeable Sampler, $x^{(1:M)}$, and the *true* samples, $x_*^{(1:M_*)}$, as a measure of how *well* the simulated samples represent the truth:

$$\kappa(x_*^{(1:M_*)}, x^{(1:M)}) := \sup_{x\in\mathbb{R}} \left| \frac{1}{M_*} \sum_{t=1}^{M_*} \mathbb{1}_{(-\infty,x]}(x_*^{(t)}) - \frac{1}{M} \sum_{t=1}^{M} \mathbb{1}_{(-\infty,x]}(x^{(t)}) \right| . \tag{101}$$

Here, for notational simplicity, we have dropped any explicit reference to the fact that we are considering the two-hundreth element of each of the sample paths.

4.3.3 *Results*

Starting with the scenario where the transition weight has exponentially increasing tails (Example 4 with $\sigma^2 = 1/2$); Figures 42, 43, and 44 show, respectively; a plot of the sample efficiency in the X-space against the *jump-size*, ϵ, for each value of N; a plot of the sample efficiency in the X-space against the sample acceptance rate for each value of N; and a plot of the Kolmogorov-Smirnov (KS) statistic, (100), against the jump-size for each value of N. Figure 42 shows that, for each value of N, the maximum sample efficiency is achieved with the maximal jump-size, $\epsilon = \sqrt{2}$, which corresponds to independent samples. The most *efficient* choice of N is given by $N = 1$. Figure 43 highlights that

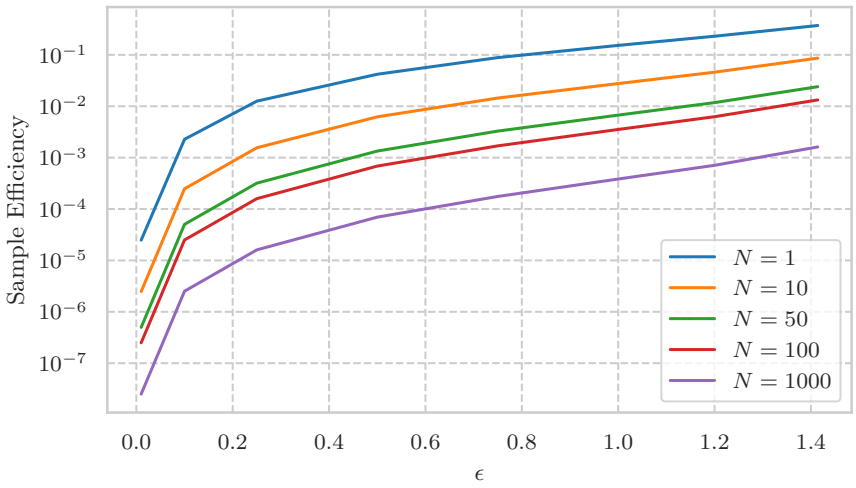

Figure 42: A plot of the sample efficiency in the X-space against the *jump-size*, ϵ, for each value of $N \in \{1, 10, 50, 100, 1000\}$ of the Exchangeable Sampler targeting a $\mathrm{N}(0, 1)$ distribution by using a $\mathrm{N}(0, 1/2)$ distribution as the marginal proposal distribution. The sampler was run for one-hundred-thousand iterations.

the optimal acceptance rate is larger the larger the value of N. On the other hand, Figure 44 illustrates that the KS statistic is minimised for values of ϵ not necessarily equal to the maximal value $\sqrt{2}$. Indeed, that figure suggests that, the smaller the value of N, the smaller the value of ϵ which corresponds to the minimal value of the KS statistic, which suggests that, even though the sample efficiency is maximised when $\epsilon = \sqrt{2}$, this might not represent the *optimal* jump-size in the sense of producing samples which most closely represent the target.

Figure 45 shows histograms of the samples simulated by the Exchangeable Sampler for $N = 1$ and for a variety of jump-sizes, $\epsilon \in \{0.01, 0.1, 0.25, 0.5, 0.75, 1.0, 1.2, \sqrt{2}\}$, where $\epsilon = 0.01$ for the top-left subplot, $\epsilon = 0.1$ for the top-right subplot, $\epsilon = 0.25$ for the subplot in the second row and the first column, and so on. With the same layout, Figure 46 shows, at each of the one-hundred-thousand iterations, the states of the Exchangeable Sampler for $N = 1$ and the same set of jump-sizes. Figure 45 highlights that, for the relatively larger jump-sizes, the simulated samples do not represent the true target well in the tails. Figure 46 shows that this is happening because the sampler gets *stuck* for periods of time when the chain goes out into the tails of the target distribution which is exactly what we expect given the exponentially increasing transition weight in the tails. Of course, if we were to start the chain far out into the tails, the chain would struggle to move for these relatively larger values of ϵ. As such, in this instance, even though larger values of the jump-size result in larger expected squared jump distances in the X-space (as shown in Figure 42), smaller values of the jump-size lead to samplers which are less prone to get stuck in the tails (as illustrated in Figure 45) and whose resulting samples more accurately

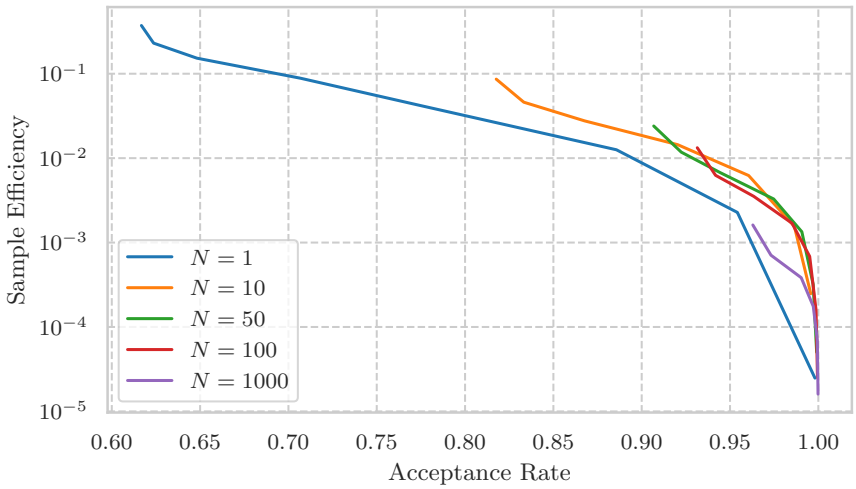

Figure 43: A plot of the sample efficiency in the X-space against the sample acceptance rate for each value of $N \in \{1, 10, 50, 100, 1000\}$ of the Exchangeable Sampler targeting a $\mathrm{N}(0, 1)$ distribution by using a $\mathrm{N}(0, 1/2)$ distribution as the marginal proposal distribution. The sampler was run for one-hundred-thousand iterations.

represent the true target (as highlighted in Figures 43 and 45), and are, therefore, preferable. This sticky behaviour is amplified for larger values of N since, the larger the value of N, the greater the chance of proposing a state in the tails which has a large weight. This can be seen in Figures 86, 87, 88, and 89 which show, for $N = 10$, $N = 50$, $N = 100$, and $N = 1000$, respectively, histograms of the samples simulated by the Exchangeable Sampler at each of the one-hundred-thousand iterations for a variety of jump-sizes, $\epsilon \in \{0.01, 0.1, 0.25, 0.5, 0.75, 1.0, 1.2, \sqrt{2}\}$, as well as in Figures 90, 91, 92, and 93 which show, for $N = 10$, $N = 50$, $N = 100$, and $N = 1000$, respectively, the states of the Exchangeable Sampler at each of the one-hundred-thousand iterations for the same set of jump-sizes.

For the second scenario, where the transition weight has polynomially increasing tails (Example 5 with $\alpha_1 = 5.5$, $\alpha_2 = 0.5$, and $\beta = 1$); Figures 47, 48, and 49 show, respectively; a plot of the sample efficiency in the X-space against the *jump-size*, ϵ, for each value of N; a plot of the sample efficiency in the X-space against the sample acceptance rate for each value of N; and a plot of the Kolmogorov-Smirnov (KS) statistic, (100), against the jump-size for each value of N. Figure 47 shows that the larger the value of N, the larger the jump-size which achieves the maximum sample efficiency; that is, the larger the value of N, the bigger the jumps you can take. Similarly, Figure 49 illustrates that the larger the value of N, the larger the jump-size which minimises the KS statistic. Figure 48 highlights that the optimal acceptance rate is larger the larger the value of N. All three figures suggest, as one would expect in this scenario, that choosing an ϵ less than $\sqrt{2}$; that is, not using independent samples, leads to a

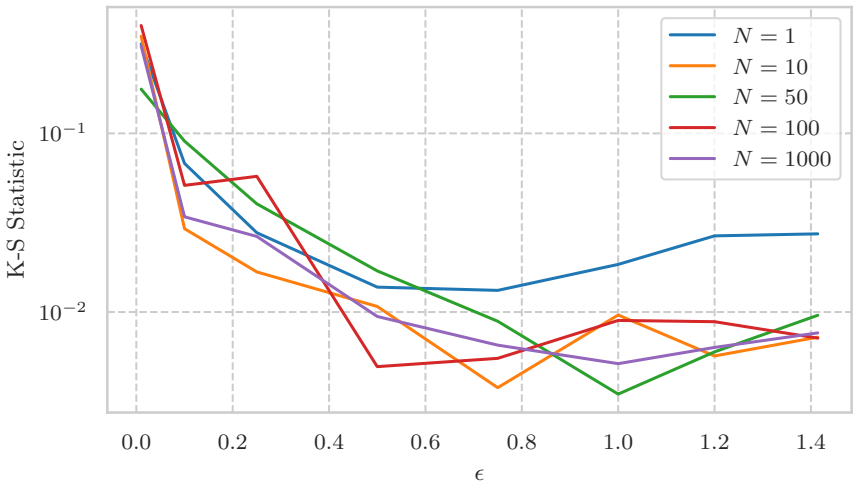

Figure 44: A plot of the Kolmogorov-Smirnov statistic, (100), against the jump-size for each value of $N \in \{1, 10, 50, 100, 1000\}$ of the Exchangeable Sampler targeting a $\mathrm{N}(0, 1)$ distribution by using a $\mathrm{N}(0, 1/2)$ distribution as the marginal proposal distribution. The sampler was run for one-hundred-thousand iterations.

better sampler. Figure 50 shows histograms of the samples simulated by the Exchangeable Sampler for $N = 1$ and for a variety of jump-sizes, $\epsilon \in \{0.01, 0.1, 0.25, 0.5, 0.75, 1.0, 1.2, \sqrt{2}\}$, where $\epsilon = 0.01$ for the top-left subplot, $\epsilon = 0.1$ for the top-right subplot, $\epsilon = 0.25$ for the subplot in the second row and the first column, and so on. With the same layout, Figure 51 shows, at each of the one-hundred-thousand iterations, the states of the Exchangeable Sampler for $N = 1$ and for a variety of jump-sizes, $\epsilon \in \{0.01, 0.1, 0.25, 0.5, 0.75, 1.0, 1.2, \sqrt{2}\}$. Figure 50 highlights that at the larger and smaller end of the jump-size scale, the simulated samples do not represent the true target well in the tails. As was the case for the first example, Figure 51 shows that this is happening because, for the relatively larger jump-sizes, the sampler gets *stuck* for periods of time when the chain goes out into the tails of the target distribution, which is exactly what we expect given the polynomially increasing transition weight in the tails. As highlighted previously, if we were to start the chain far out into the tails, the chain would struggle to move for the relatively larger values of ϵ. Moreover, for the relatively smaller jump-sizes, the sampler exhibits random-walk behaviour where, at each iteration, although the chain is moving, it is not moving very far. Unlike the first example, the sticky behaviour seen here is less pronounced for the larger values of N. This is because, even though the larger the value of N, the greater the chance of proposing a state in the tails which has a large weight, the weight increases only polynomially in the tails and thus the drift towards regions of the space where the target has more mass still occurs. This is backed up by the empirical evidence given in the discussion and figures of Examples 5 which suggest that the sampler is indeed geometrically erogodic in this

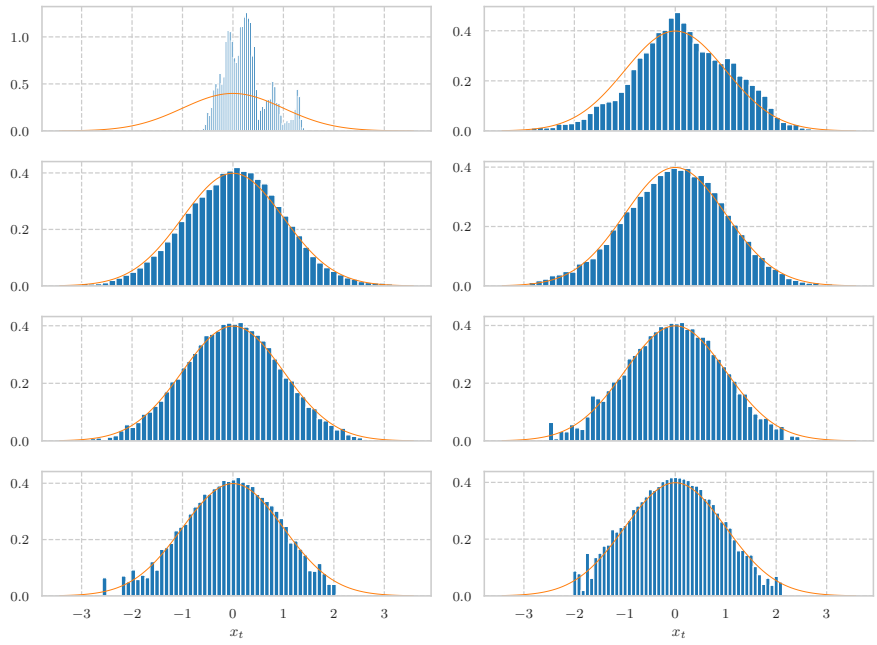

Figure 45: Histograms of the samples simulated by the Exchangeable Sampler targeting a $\mathrm{N}(0, 1)$ distribution by using a $\mathrm{N}(0, 1/2)$ distribution as the marginal proposal distribution for $N = 1$ and for a variety of jump-sizes, $\epsilon \in \{0.01, 0.1, 0.25, 0.5, 0.75, 1.0, 1.2, \sqrt{2}\}$, where $\epsilon = 0.01$ for the top-left subplot, $\epsilon = 0.1$ for the top-right subplot, $\epsilon = 0.25$ for the subplot in the second row and the first column, and so on. The orange line in each figure corresponds to the *true* density.

scenario for $\epsilon < \sqrt{2}$. The discussion and figures in Example 5 also suggest that the Exchangeable Sampler is geometrically ergodic for larger N provided $\epsilon < \sqrt{2}$. This can be seen in practice in Figures 94, 95, 96, and 97 which show, for $N = 10$, $N = 50$, $N = 100$, and $N = 1000$, respectively, histograms of the samples simulated by the Exchangeable Sampler at each of the one-hundred-thousand iterations for a variety of jump-sizes, $\epsilon \in \{0.01, 0.1, 0.25, 0.5, 0.75, 1.0, 1.2, \sqrt{2}\}$, as well as in Figures 98, 99, 100, and 101 which show, for $N = 10$, $N = 50$, $N = 100$, and $N = 1000$, respectively, the states of the Exchangeable Sampler at each of the one-hundred-thousand iterations for a variety of jump-sizes, $\epsilon \in \{0.01, 0.1, 0.25, 0.5, 0.75, 1.0, 1.2, \sqrt{2}\}$.

For the third scenario, where the transition weight is bounded (Example 6 with $\nu = 5$); Figures 52, 53, and 54 show, respectively; a plot of the sample efficiency in the X-space against the *jump-size*, ϵ, for each value of N; a plot of the sample efficiency in the X-space against the sample acceptance rate for each value of N; and a plot of the Kolmogorov-Smirnov (KS) statistic, (100), against the jump-size for each value of N. Figure 52 shows that for each value of N, the choice of the jump-size that maximises the sample efficiency is $\epsilon = \sqrt{2}$; that is, for each value of N, taking independent samples achieves the optimal sample efficiency. The figure also shows that the curves of the sample

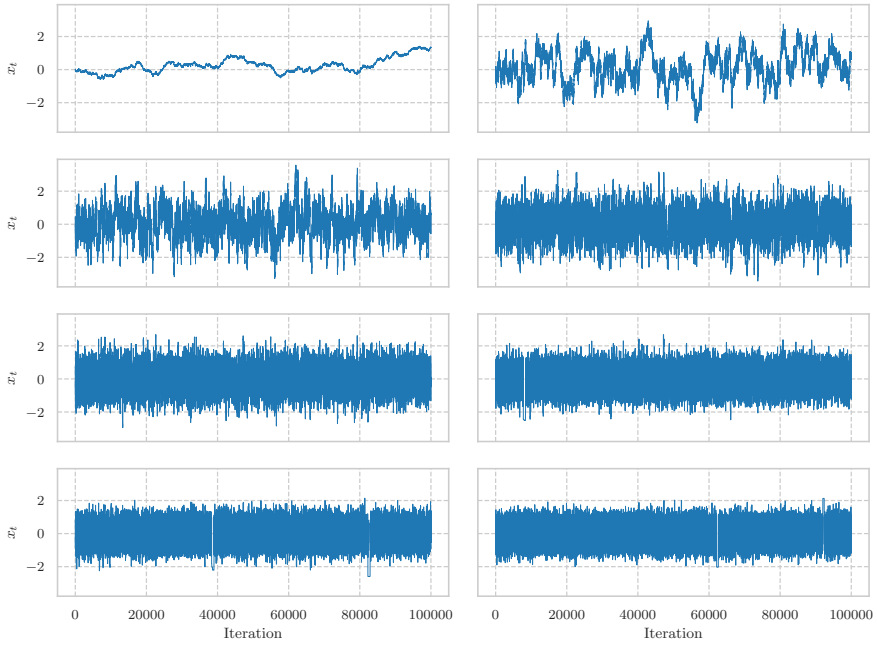

Figure 46: Plots of the states of the Exchangeable Sampler targeting a $\mathrm{N}(0, 1)$ distribution by using a $\mathrm{N}(0, 1/2)$ distribution as the marginal proposal distribution for $N = 1$ and for a variety of jump-sizes, $\epsilon \in \{0.01, 0.1, 0.25, 0.5, 0.75, 1.0, 1.2, \sqrt{2}\}$, at each of the one-hundred-thousand iterations. $\epsilon = 0.01$ in the top-left subplot, $\epsilon = 0.1$ in the top-right subplot, $\epsilon = 0.25$ in the subplot in the second row and the first column, and so on.

efficiency against the jump-size do not differ much across the values of N. Figure 53 highlights that the optimal acceptance rate, although larger the larger the value of N, also does not differ much across the values of N. Similarly, Figure 54 illustrates that the KS statistic as a function of ϵ is also invariant to the choice of N. All three figures suggest, as one would expect in this scenario where the transition weight is bounded, that the optimal choice of the jump-size is $\epsilon = \sqrt{2}$; that is, using independent samples. Figure 55 shows histograms of the samples simulated by the Exchangeable Sampler for $N = 1$ and for a variety of jump-sizes, $\epsilon \in \{0.01, 0.1, 0.25, 0.5, 0.75, 1.0, 1.2, \sqrt{2}\}$, where $\epsilon = 0.01$ for the top-left subplot, $\epsilon = 0.1$ for the top-right subplot, $\epsilon = 0.25$ for the subplot in the second row and the first column, and so on. With the same layout, Figure 56 shows, at each of the one-hundred-thousand iterations, the states of the Exchangeable Sampler for $N = 1$ and for a variety of jump-sizes, $\epsilon \in \{0.01, 0.1, 0.25, 0.5, 0.75, 1.0, 1.2, \sqrt{2}\}$. Figure 55 highlights that, for the very small jump-size, the simulated samples do not represent the true target well. Figure 56 shows that this is because the sampler exhibits random-walk behaviour where, at each iteration, although the chain is moving, it is not moving very far. Figure 56 also shows that, unlike the previous two examples, for any value of ϵ, the chain does not exhibit *sticky* behaviour. As a result, for any value of ϵ, the simulated samples provide a good representation of the target.

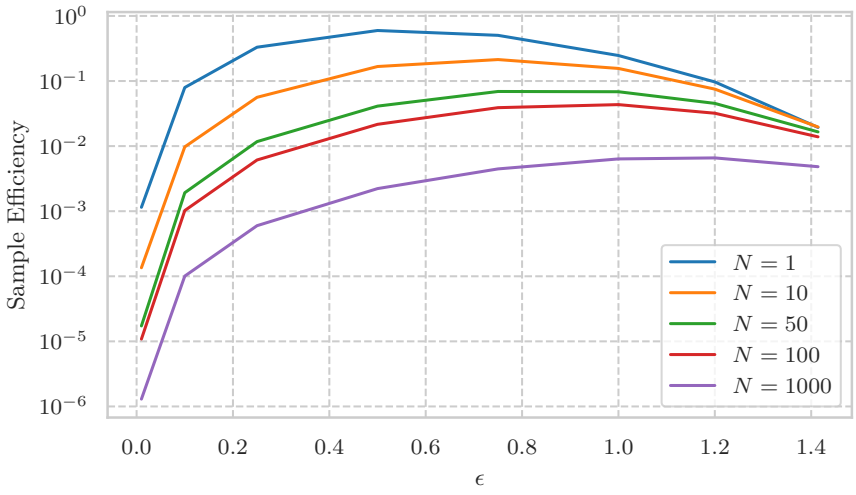

Figure 47: A plot of the sample efficiency in the X-space against the *jump-size*, ϵ, for each value of $N \in \{1, 10, 50, 100, 1000\}$ of the Exchangeable Sampler targeting a Gamma$(5.5, 1)$ distribution by using a Gamma$(0.5, 1/2)$ distribution as the marginal proposal distribution. The sampler was run for one-hundred-thousand iterations.

These figures suggest that the sampler is geometrically erogodic in this scenario for every value of ϵ, even though the discussion and figures in Example 6 could not provide evidence of this. Moreover, Figures 102, 103, 104, and 105 which show, for $N = 10$, $N = 50$, $N = 100$, and $N = 1000$, respectively, histograms of the samples simulated by the Exchangeable Sampler at each of the one-hundred-thousand iterations for a variety of jump-sizes, $\epsilon \in \{0.01, 0.1, 0.25, 0.5, 0.75, 1.0, 1.2, \sqrt{2}\}$, as well as Figures 106, 107, 108, and 109 which show, for $N = 10$, $N = 50$, $N = 100$, and $N = 1000$, respectively, the states of the Exchangeable Sampler at each of the one-hundred-thousand iterations for a variety of jump-sizes, $\epsilon \in \{0.01, 0.1, 0.25, 0.5, 0.75, 1.0, 1.2, \sqrt{2}\}$, suggest that the sampler is geometrically erogodic for all values of ϵ and all values of N.

For the fourth scenario, where the target is the conditioned Birth-Death diffusion of Section 3.1.1 and q_0 corresponds to the Modified Diffusion Bridge proposal of Section 3.2.3, Figures 57 and 58 show, respectively; a plot of the sample efficiency in the X-space against the *jump-size*, ϵ, for each value of N; and a plot of the sample efficiency in the X-space against the sample acceptance rate for each value of N. Figure 59 shows a plot of the two-sample Kolmogorov-Smirnov (KS) statistic, (101), calculated for the two-hundredth element of the sample paths, against the jump-size for each value of N. Figure 57 shows that, the larger the value of N, the larger the *optimal* jump-size which achieves the maximum sample efficiency; that is, the larger the value of N, the bigger the jumps you can take. Figure 58 highlights that the optimal acceptance rate is larger the larger the value of N. Figure 59 illustrates that, for any value of N, the choice

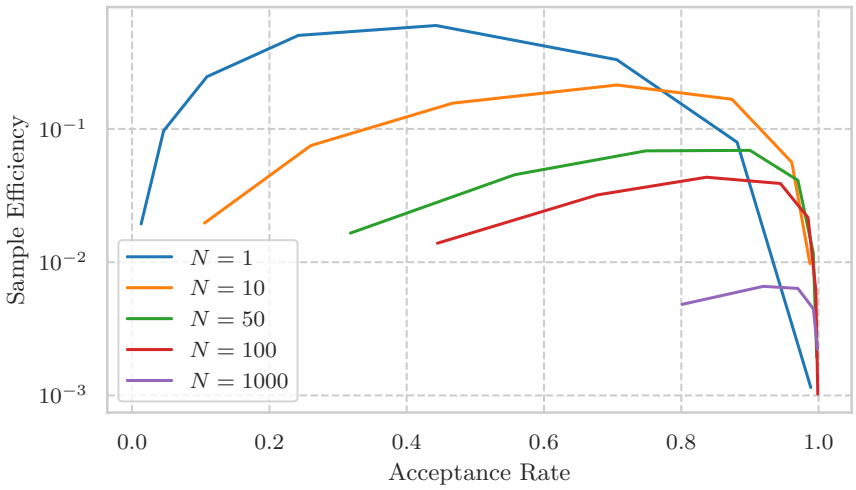

Figure 48: A plot of the sample efficiency in the X-space against the sample acceptance rate for each value of $N \in \{1, 10, 50, 100, 1000\}$ of the Exchangeable Sampler targeting a Gamma$(5.5, 1)$ distribution by using a Gamma$(0.5, 1)$ distribution as the marginal proposal distribution. The sampler was run for one-hundred-thousand iterations.

of jump-size which minimises the two-sample KS statistic is less than $\sqrt{2}$. All three figures suggest that choosing an ϵ less than $\sqrt{2}$; that is, not using independent samples, leads to a *better* sampler. Figure 60 shows histograms of the two-hundreth element of each of the sample paths simulated by the Exchangeable Sampler for $N = 1$ and for a variety of jump-sizes, $\epsilon \in \{0.01, 0.1, 0.25, 0.5, 0.75, 1.0, 1.2, \sqrt{2}\}$, where $\epsilon = 0.01$ for the top-left subplot, $\epsilon = 0.1$ for the top-right subplot, $\epsilon = 0.25$ for the subplot in the second row and the first column, and so on. With the same layout, Figure 61 shows, at each of the one-hundred-thousand iterations, the two-hundreth element of the states of the Exchangeable Sampler for $N = 1$ and for a variety of jump-sizes, $\epsilon \in \{0.01, 0.1, 0.25, 0.5, 0.75, 1.0, 1.2, \sqrt{2}\}$. Figure 60 highlights that at the larger and smaller end of the jump-size scale, the simulated samples do not represent the true target well. As has been highlighted in some of the previous scenarios, Figure 61 shows that this is happening because, for the relatively larger jump-sizes, the sampler gets *stuck* for periods of time when the chain goes out into the tails of the target distribution. As highlighted in those other examples, if we were to start the chain far out into the tails, the chain would struggle to move for the relatively larger values of ϵ. Moreover, for the relatively smaller jump-sizes, the sampler exhibits random-walk behaviour where, at each iteration, although the chain is moving, it is not moving very far. The sticky behaviour seen here suggests that the Exchangeable Sampler in this scenario is not geometrically ergodic for $\epsilon = \sqrt{2}$. However, the figures suggest that the sampler is geometrically ergodic for any $\epsilon < \sqrt{2}$. This sticky behaviour is less pronounced for the larger values of N which suggests, as was the case for the scenario where the target and

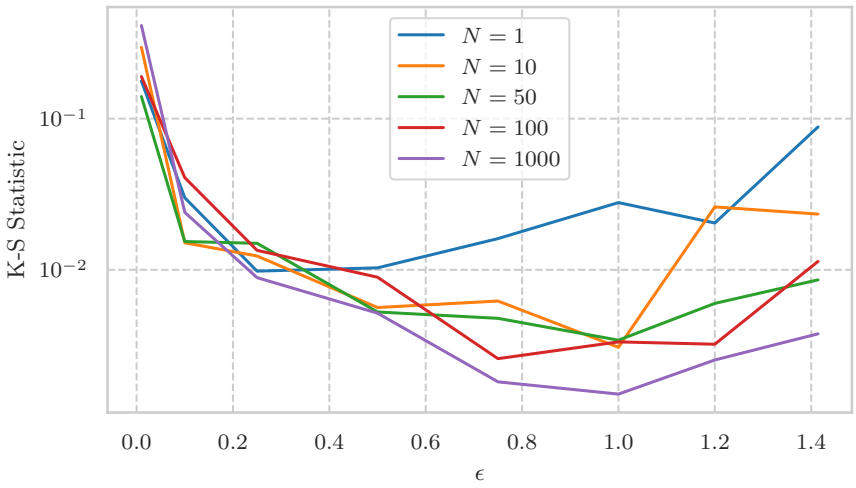

Figure 49: A plot of the Kolmogorov-Smirnov statistic, (100), against the jump-size for each value of $N \in \{1, 10, 50, 100, 1000\}$ of the Exchangeable Sampler targeting a Gamma$(5.5, 1)$ distribution by using a Gamma$(0.5, 1)$ distribution as the marginal proposal distribution. The sampler was run for one-hundred-thousand iterations.

marginal proposal distributions were both of a Gamma form, that the transition weight in this scenario is sufficienty well behaved. While this sticky behaviour is less pronounced the larger the value of N, it is still present and suggests that, no matter the value of N, the chain is not geometrically ergodic for $\epsilon = \sqrt{2}$. Indeed, one can see this by looking at Figures 110, 111, 112, and 113 which show, for $N = 10$, $N = 50$, $N = 100$, and $N = 1000$, respectively, histograms of the two-hundreth element of each of the sample paths simulated by the Exchangeable Sampler at each of the one-hundred-thousand iterations for a variety of jump-sizes, $\epsilon \in \{0.01, 0.1, 0.25, 0.5, 0.75, 1.0, 1.2, \sqrt{2}\}$, as well as Figures 114, 115, 116, and 117 which show, for $N = 10$, $N = 50$, $N = 100$, and $N = 1000$, respectively, the two-hundreth element of the states of the Exchangeable Sampler at each of the one-hundred-thousand iterations for a variety of jump-sizes, $\epsilon \in \{0.01, 0.1, 0.25, 0.5, 0.75, 1.0, 1.2, \sqrt{2}\}$. These figures also suggest, as was the case for the scenario where both the target and marginal proposal were Gamma distributions, that the the Exchangeable Sampler is geometrically ergodic for any value of N provided $\epsilon < \sqrt{2}$.

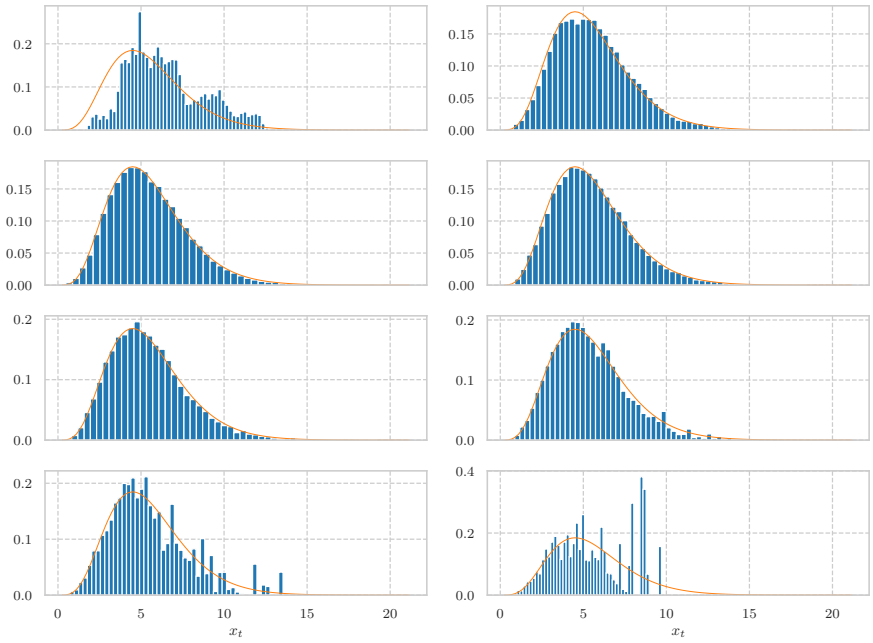

Figure 50: Histograms of the samples simulated by the Exchangeable Sampler targeting a Gamma$(5.5, 1)$ distribution by using a Gamma$(0.5, 1)$ distribution as the marginal proposal distribution for $N = 1$ and for a variety of jump-sizes, $\epsilon \in \{0.01, 0.1, 0.25, 0.5, 0.75, 1.0, 1.2, \sqrt{2}\}$ where $\epsilon = 0.01$ for the top-left subplot, $\epsilon = 0.1$ for the top-right subplot, $\epsilon = 0.25$ for the subplot in the second row and the first column, and so on. The orange line in each figure corresponds to the *true* density.

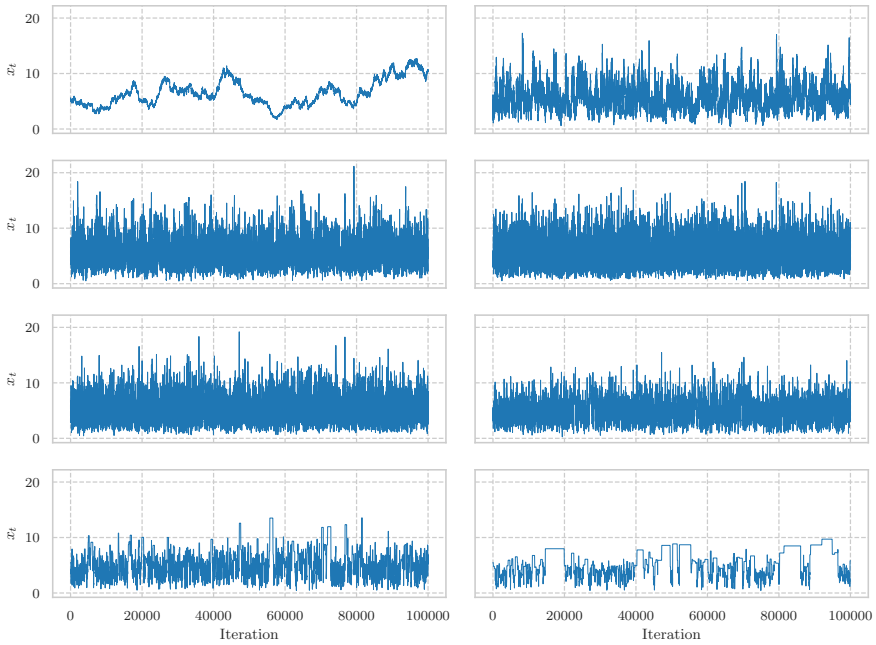

Figure 51: Plots of the states of the Exchangeable Sampler targeting a Gamma$(5.5, 1)$ distribution by using a N$(0.5, 1)$ distribution as the marginal proposal distribution for $N = 1$ and for a variety of jump-sizes, $\epsilon \in \{0.01, 0.1, 0.25, 0.5, 0.75, 1.0, 1.2, \sqrt{2}\}$, at each of the one-hundred-thousand iterations. $\epsilon = 0.01$ in the top-left subplot, $\epsilon = 0.1$ in the top-right subplot, $\epsilon = 0.25$ in the subplot in the second row and the first column, and so on.

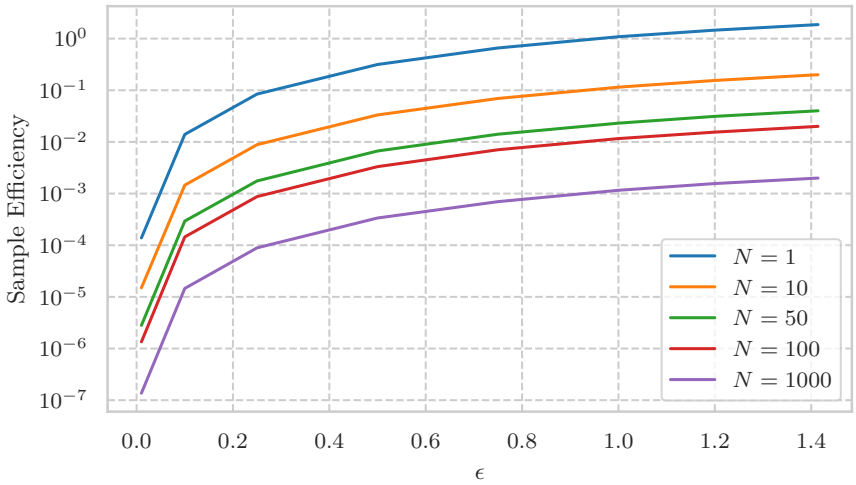

Figure 52: A plot of the sample efficiency in the X-space against the *jump-size*, ϵ, for each value of $N \in \{1, 10, 50, 100, 1000\}$ of the Exchangeable Sampler targeting a N$(0, 1)$ distribution by using a T(5) distribution as the marginal proposal distribution. The sampler was run for one-hundred-thousand iterations.

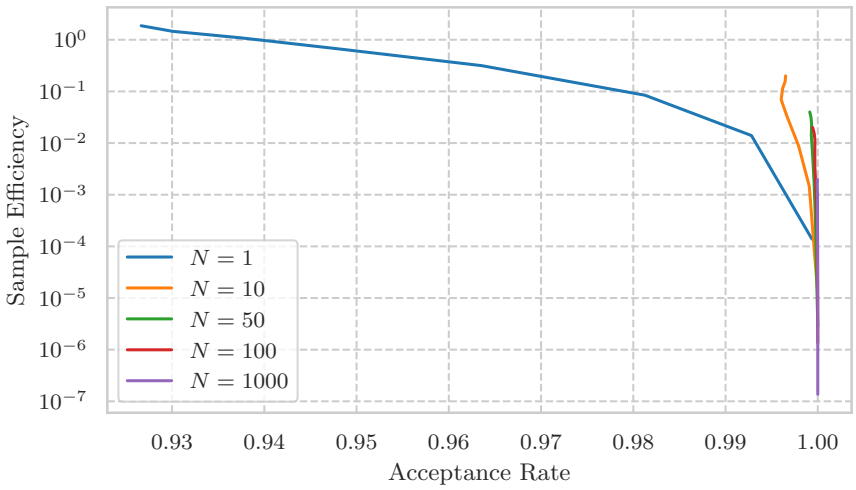

Figure 53: A plot of the sample efficiency in the X-space against the sample acceptance rate for each value of $N \in \{1, 10, 50, 100, 1000\}$ of the Exchangeable Sampler targeting a $\mathrm{N}(0, 1)$ distribution by using a $\mathrm{T}(5)$ distribution as the marginal proposal distribution. The sampler was run for one-hundred-thousand iterations.

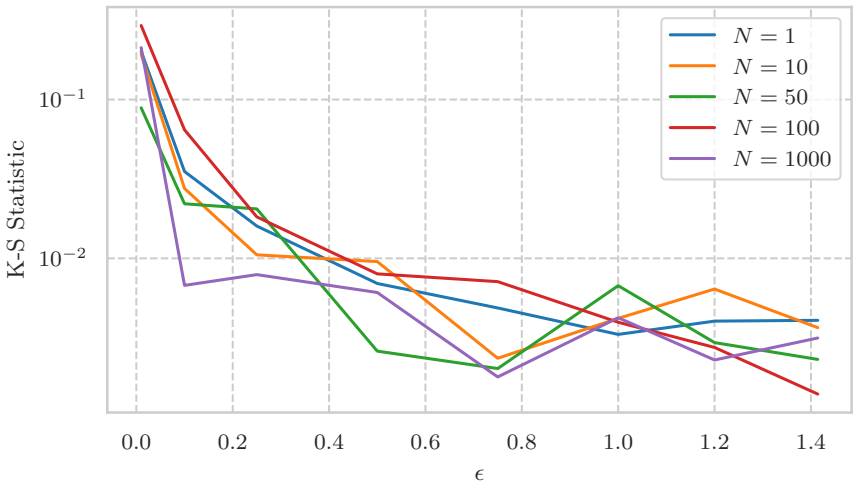

Figure 54: A plot of the Kolmogorov-Smirnov statistic, (100), against the jump-size for each value of $N \in \{1, 10, 50, 100, 1000\}$ of the Exchangeable Sampler targeting a $\mathrm{N}(0, 1)$ distribution by using a $\mathrm{T}(5)$ distribution as the marginal proposal distribution. The sampler was run for one-hundred-thousand iterations.

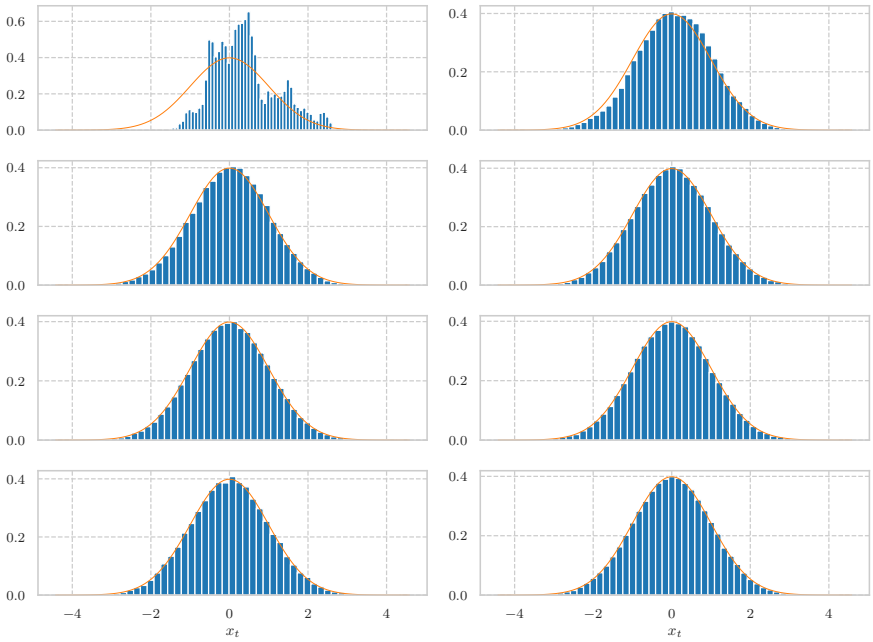

Figure 55: Histograms of the samples simulated by the Exchangeable Sampler targeting a $\mathrm{N}(0, 1)$ distribution by using a $\mathrm{T}(5)$ distribution as the marginal proposal distribution for $N = 1$ and for a variety of jump-sizes, $\epsilon \in \{0.01, 0.1, 0.25, 0.5, 0.75, 1.0, 1.2, \sqrt{2}\}$ where $\epsilon = 0.01$ for the top-left subplot, $\epsilon = 0.1$ for the top-right subplot, $\epsilon = 0.25$ for the subplot in the second row and the first column, and so on. The orange line in each figure corresponds to the *true* density.

Figure 56: Plots of the states of the Exchangeable Sampler targeting a $\mathrm{N}(0, 1)$ distribution by using a $\mathrm{T}(5)$ distribution as the marginal proposal distribution for $N = 1$ and for a variety of jump-sizes, $\epsilon \in \{0.01, 0.1, 0.25, 0.5, 0.75, 1.0, 1.2, \sqrt{2}\}$, at each of the one-hundred-thousand iterations. $\epsilon = 0.01$ in the top-left subplot, $\epsilon = 0.1$ in the top-right subplot, $\epsilon = 0.25$ in the subplot in the second row and the first column, and so on.

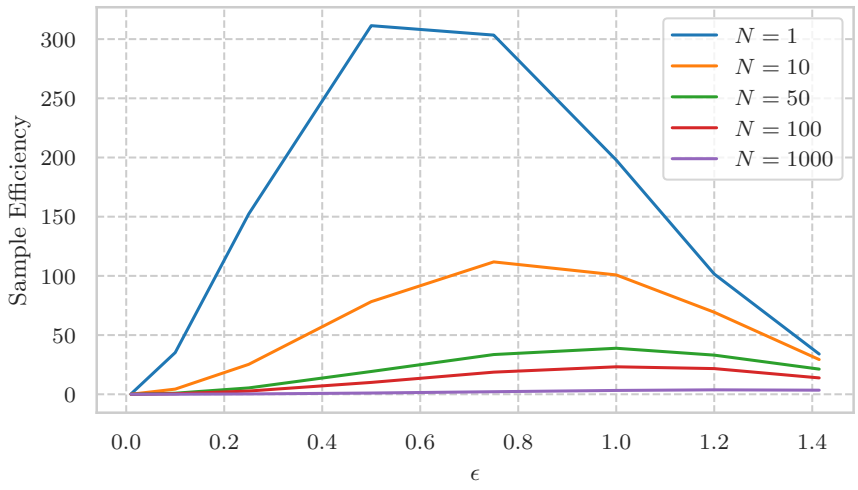

Figure 57: A plot of the sample efficiency in the X-space against the *jump-size*, ϵ, for each value of $N \in \{1, 10, 50, 100, 1000\}$ of the Exchangeable Sampler targeting a conditioned Birth-Death diffusion by using the Modified Diffusion Bridge proposal of Section 3.2.3 as the marginal proposal distribution. The sampler was run for one-hundred-thousand iterations.

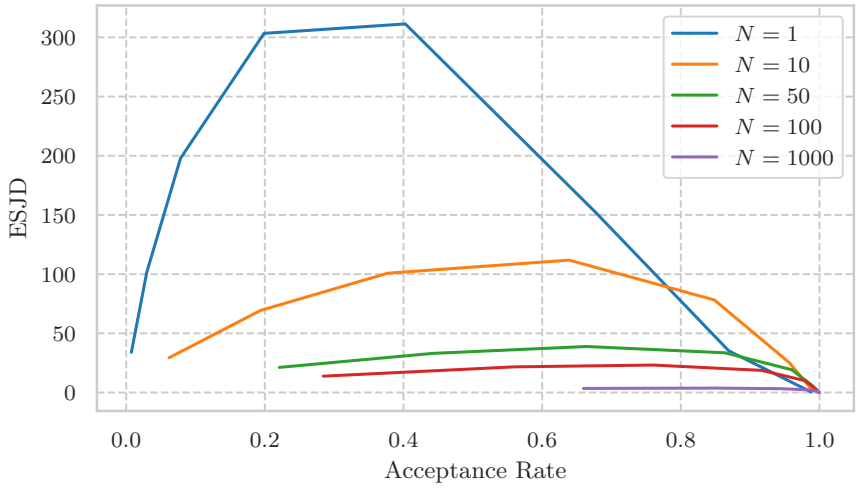

Figure 58: A plot of the sample efficiency in the X-space against the sample acceptance rate for each value of $N \in \{1, 10, 50, 100, 1000\}$ of the Exchangeable Sampler targeting a conditioned Birth-Death diffusion by using the Modified Diffusion Bridge proposal of Section 3.2.3 as the marginal proposal distribution. The sampler was run for one-hundred-thousand iterations.

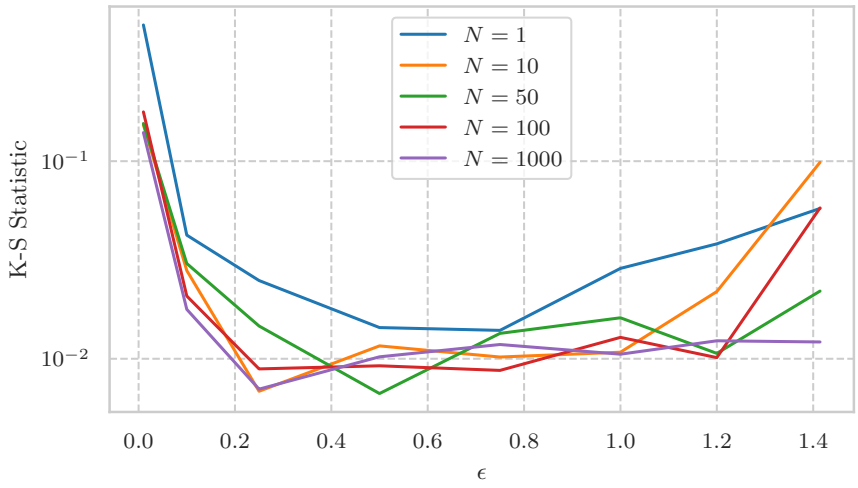

Figure 59: A plot of the two-sample Kolmogorov-Smirnov statistic calculated for the two-hundreth element of the sample paths, (101), against the jump-size for each value of $N \in \{1, 10, 50, 100, 1000\}$ of the Exchangeable Sampler targeting a conditioned Birth-Death diffusion by using the Modified Diffusion Bridge proposal of Section 3.2.3 as the marginal proposal distribution. The sampler was run for one-hundred-thousand iterations. Here, in order to get samples representing the *truth* needed to calculate the two-sample Kolmogorov-Smirnov statistic, an independence sampler which used the residual-bridge construct of Whitaker et al., 2017 was ran for one-million iterations.

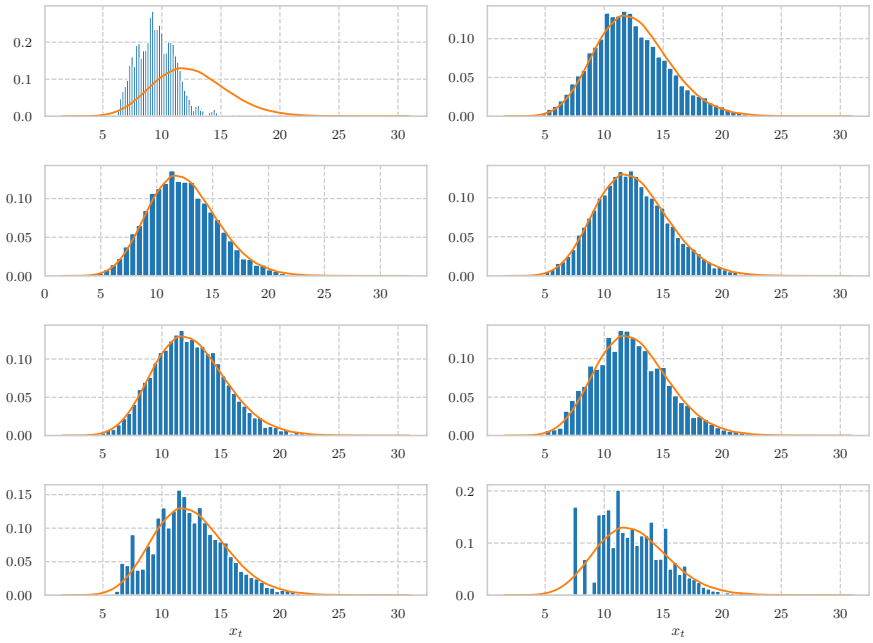

Figure 60: Histograms of the two-hundreth element of each of the sample paths simulated by the Exchangeable Sampler targeting a conditioned Birth-Death diffusion by using the Modified Diffusion Bridge proposal of Section 3.2.3 as the marginal proposal distribution. for $N = 1$ and for a variety of jump-sizes, $\epsilon \in \{0.01, 0.1, 0.25, 0.5, 0.75, 1.0, 1.2, \sqrt{2}\}$ where $\epsilon = 0.01$ for the top-left subplot, $\epsilon = 0.1$ for the top-right subplot, $\epsilon = 0.25$ for the subplot in the second row and the first column, and so on. The orange line in each figure corresponds to the *true* density.

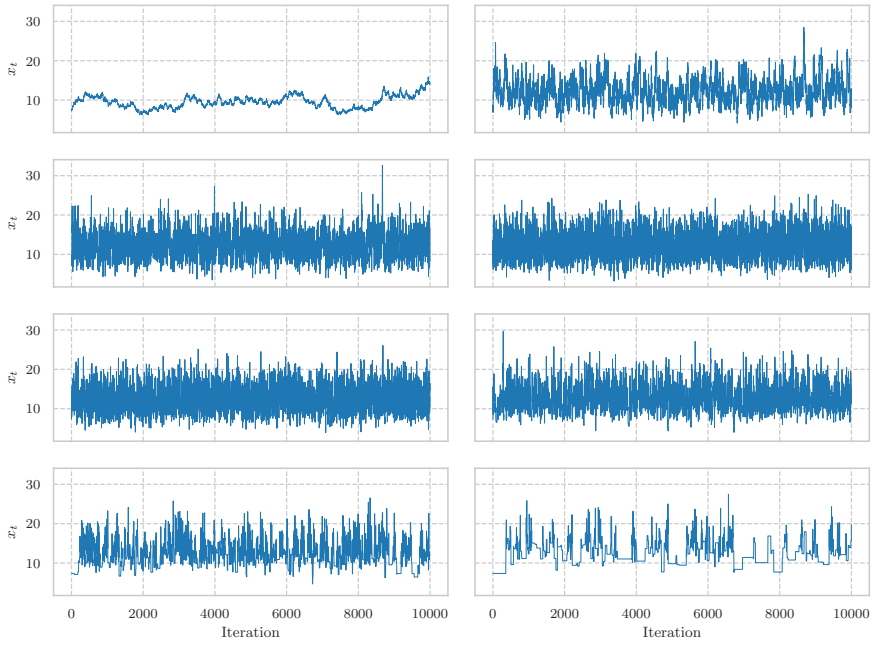

Figure 61: Plots of the two-hundreth element of the states of the Exchangeable Sampler targeting a conditioned Birth-Death diffusion by using the Modified Diffusion Bridge proposal of Section 3.2.3 as the marginal proposal distribution for $N = 1$ and for a variety of jump-sizes, $\epsilon \in \{0.01, 0.1, 0.25, 0.5, 0.75, 1.0, 1.2, \sqrt{2}\}$, at each of the one-hundred-thousand iterations. $\epsilon = 0.01$ in the top-left subplot, $\epsilon = 0.1$ in the top-right subplot, $\epsilon = 0.25$ in the subplot in the second row and the first column, and so on.

4.4 THE EXCHANGEABLE PARTICLE GIBBS SAMPLER

In the previous section, as a precursor to introducing the Exchangeable Particle Gibbs Sampler, we introduced the Exchangeable Sampler as a generalisation of the Independence Sampler. We theoretically demonstrated, through Theorem 4.3.9, that such a sampler can be geometrically ergodic and produce MCMC estimates which satisfy a central limit theorem for all functions which are square-integrable with respect to π even when the transition weight is unbounded, provided the weight is sufficiently well behaved in the tails. This is in contrast to the Independence Sampler which can not be geometrically ergodic if the transition weight is unbounded (Theorem 2.3.31). We then showed, via a simulation study in Section 4.3.2, that, in practice, tuning the *jump-size* in the Exchangeable Sampler can offer substantial improvements in the *rate of mixing* over the Independence Sampler, particularly for scenarios where the transition weight is unbounded. Much like the Particle Gibbs Sampler can be seen as an extension to Barker's (multiple-proposal) Independence Sampler, the Exchangeable Particle Gibbs Sampler (xPGS) can be seen as an extension to the Exchangeable Sampler with Barker's acceptance probabillity. Before introducing the Exchangeable Particle Gibbs Sampler, we note that there are other approaches in the literature that, at each time step of the Sequential Monte Carlo procedure, attempt to introduce some dependence between the propogated particles. While fundamentally different to the approach outlined in this thesis, they are motivated by the same considerations. In particular, Bizjajeva and Olsson, 2016 introduce a *blockwise* propogation scheme where each particle produces α, say, negatively correlated offspring simulated from some Markovian kernel. In addition, at each observation time, the Embedded Hidden Markov Model of Neal, Beal, and Roweis, 2004 uses a Markov chain to simulate dependent *pool states*, and the introduction of *sequential* dependence between *pool states* at different observation times in Shestopaloff and Neal, 2018 leads to an algorithm which has connections to particle MCMC methods where there is dependence between the propogated particles (as discussed in Section 4.3 of Shestopaloff and Neal, 2018). To introduce the Exchangeable Particle Gibbs Sampler, we start by introducing the Exchangeable Sequential Monte Carlo (xSMC) procedure, which is an extension of the Sequential Monte Carlo procedure obtained by propagating particles exchangeably as opposed to independently. Consider, then, the SMC procedure of Algorithm 8. At each step, t, of the procedure the particles, denoted $\tilde{x}_{t-1}^{(a_{t-1}^{(1:N)})}$, are propagated forward by sampling each $x_t^{(i)}$ with density $p_t(x_t|\tilde{x}_{t-1}^{(a_{t-1}^{(i)})})$ independently of one another. Recall that the proposals are freely chosen by the practitioner, provided that, for any $\theta \in \mathbb{R}^p$,

$$\text{supp}(\gamma_0(\theta, \cdot)) \subseteq \text{supp}(p_0(\cdot|\theta)) \ ,$$

and, for any $t \in \{1, \ldots, T\}$, and any $(\theta, x_{0:t-1}) \in \mathbb{R}^p \times \mathbb{R}^{d \times t}$,

$$\text{supp}(\gamma_t(\theta, x_{0:t-1}, \cdot)) \subseteq \text{supp}(\gamma_{t-1}(\theta, x_{0:t-1}) p_t(\cdot|x_{0:t-1}, \theta)) \ .$$

With the same choice of the proposal densities, p_t, the xSMC procedure, at each time $t \in \{0, \ldots, T\}$, samples $x_t^{(1:N)}$ with density $p_t^{(N)}(x_t^{(1:N)}|\tilde{x}_{t-1}^{(a_{t-1}^{(1:N)})})$, where $p_t^{(N)}$ is an exchangeable density with marginals p_t; that is, $p_t^{(N)}$ satisfies the following assumptions;

ASSUMPTIONS 4.4.1.

(X) For any permutation, σ, of $\{1, \ldots, N\}$,

$$p_t^{(N)}(y_{1:N}|x_{1:N}) = p_t^{(N)}(y_{\sigma(1)}, \ldots, y_{\sigma(N)}|x_{\sigma(1)}, \ldots, x_{\sigma(N)}) \, .$$

(M) For any $i \in \{1, \ldots, N\}$,

$$\int p_t^{(N)}(y_{1:N}|x_{1:N}) \, \mathrm{d}y_{(-i)} = p_t(y_i|x_i) \, .$$

In full, the xSMC procedure is given by Algorithm 19 and relies upon the following:

1. A sequence of marginal proposal densities,

$$p_0(x_0|\theta), p_1(x_1|x_0, \theta), \ldots, p_T(x_T|x_{0:T-1}, \theta) \, .$$

2. A sequence of proposal densities,

$$p_0^{(N)}(x_0^{(1:N)}|\theta), p_1^{(N)}(x_1^{(1:N)}|x_0^{(1:N)}, \theta), \ldots, p_T^{(N)}(x_T^{(1:N)}|x_{0:T-1}^{(1:N)}, \theta) \, ,$$

which satisfy the properties of Assumptions 4.4.1.

3. An *ancestral* resampling mechanism in the form of a probability mass function $\kappa(\cdot|\tilde{w}^{(1:N)})$, where $\tilde{w}^{(1:N)}$ is a given set of normalised weights.

The following assumptions, which are the same as Assumptions 4.2.1, are made on the *marginal* proposal densities and the resampling mechanism;

ASSUMPTIONS 4.4.2.

(S) For any $\theta \in \mathbb{R}^p$,

$$\mathrm{supp}(\gamma_0(\theta, \cdot)) \subseteq \mathrm{supp}(p_0(\cdot|\theta)) \, ,$$

and, for any $t \in \{1, \ldots, T\}$, and any $(\theta, x_{0:t-1}) \in \mathbb{R}^p \times \mathbb{R}^{d \times t}$,

$$\mathrm{supp}(\gamma_t(\theta, x_{0:t-1}, \cdot)) \subseteq \mathrm{supp}(\gamma_{t-1}(\theta, x_{0:t-1}) p_t(\cdot|x_{0:t-1}, \theta)) \, .$$

(U) Given a set of normalised weights, $\tilde{w}^{(1:N)}$,

$$\mathbb{E}\left[\sum_{i=1}^{N} \mathbb{1}_k(A^{(i)}) \,\middle|\, \tilde{w}^{(1:N)} \right] = N \tilde{w}^{(k)} \, ,$$

for any $k \in \{1, \ldots, N\}$.

(E) For any permutation, σ, of $\{1, \ldots, N\}$,

$$\kappa(a^{(1:N)}|\tilde{w}^{(1:N)}) = \kappa(a^{(1:N)}|\tilde{w}^{(\sigma(1))}, \ldots, \tilde{w}^{(\sigma(N))}) \, .$$

(P) For any $(k, m) \in \{1, \ldots, N\}^2$,

$$\mathbb{P}(A^{(k)} = m|\tilde{w}^{(1:N)}) = \tilde{w}^{(m)} \, .$$

As detailed previously in Section 4.2, the support condition, (S), ensures that, for any $(\theta, x_{0:t-1}) \in \mathbb{R}^p \times \mathbb{R}^{d \times t}$, it is possible to reach anywhere where $\gamma_t(\theta, x_{0:t-1}, \cdot)$ is non-zero. The unbiased assumption, (U), on the resampling mechanism, along with the exchangeable and marginal assumption on the joint proposal densities (Assumptions 4.4.1) ensures that the estimator produced by the algorithm is *unbiased* (see Theorem 4.4.3 below). Moreover, the exchangeable assumption on the resampling mechanism, (E), and on the joint proposal densities, (X), ensure that the determination of the ancestor variables does not depend on the order of the weights and, therefore, that the indices have no effect on the paths generated by the procedure. Finally, the *permutation* assumption, (P), is a *technical* condition which will make demonstrating that the Exchangeable Particle Gibbs Sampler correctly targets the density of interest, clearer. As highlighted in Remark 3 of Section 2.4.1.2, in practice, for the Sequential Monte Carlo procedure, the ancestors are set deterministically given the number of offspring, $O^{(1:N)}$, and, therefore, (P) does not hold. However, if the ancestor variables are randomly permuted, then Assumption (P) holds given Assumption (U).

REMARK 11. *As highlighted in Remark 6, interest is ultimately in paths generated by the SMC procedure and not the corresponding indices. The indices have no effect on the paths generated by the procedure since the particles are propagated forward in an exchangeable way by property (X) of Assumptions 4.4.1 and irrespective of the actual value of the ancestor variables; the ancestor variables are only needed to label where each particle proceeded from. Moreover, property (E) of Assumptions 4.4.2 ensures that the determination of the ancestor variables does not depend on the order of the weights. Therefore, in practice, it is not necessary to randomly permute the ancestor variables.*

As with (72), we define the joint mass-density function corresponding to all the variables produced by the Exchangeable Sequential Monte Carlo procedure;

$$\begin{aligned} &\Psi(x_{0:T}^{(1:N)}, a_{0:T-1}^{(1:N)}|\theta) := \\ &p_0^{(N)}(x_0^{(1:N)}|\theta) \prod_{t=1}^{T} \kappa(a_{t-1}^{(1:N)}|\tilde{w}_{t-1}^{(1:N)}) p_t^{(N)}(x_t^{(1:N)}|\tilde{x}_{t-1}^{(a_{t-1}^{(1)})}, \ldots, \tilde{x}_{t-1}^{(a_{t-1}^{(N)})}, \theta) \, , \end{aligned} \tag{102}$$

where; to ease notation, the explicit dependence of the weights on θ and $\tilde{x}_{t-1}^{(1:N)}$ has been dropped; and we have recursively defined $\tilde{x}_0^{(i)} := x_0^{(i)}$

Algorithm 19 Exchangeable Sequential Monte Carlo Procedure

Sample $x_0^{(1:N)}$ with density $p_0^{(N)}(x_0^{(1:N)}|\theta)$ and set $\tilde{x}_0^{(1:N)} = x_0^{(1:N)}$.
for $i = 1, \ldots, N$ **do**
    Calculate the i-th weight;

$$w_0(\tilde{x}_0^{(i)}; \theta) = \gamma_0(\theta, x_0^{(i)}) / p_0(x_0^{(i)}|\theta) \; .$$

end for
Normalize the weights by setting, for each $i \in \{1, \ldots, N\}$,

$$\tilde{w}_0^{(i)}(\tilde{x}_0^{(1:N)}; \theta) = w_0(\tilde{x}_0^{(i)}; \theta) / (w_0(\tilde{x}_0^{(1)}; \theta) + \ldots + w_0(\tilde{x}_0^{(N)}; \theta)) \; .$$

for $t = 1, \ldots, T$ **do**
    Sample ancestors $a_{t-1}^{(1:N)}$ with mass function $\kappa(a_{t-1}^{(1:N)}|\tilde{w}_{t-1}^{(1:N)})$.
    Sample $x_t^{(1:N)}$ with density $p_t^{(N)}(x_t^{(1:N)}|\tilde{x}_{t-1}^{(a_{t-1}^{(1:N)})}, \theta)$.
    for $i = 1, \ldots, N$ **do**
        Set $\tilde{x}_t^{(i)} := (\tilde{x}_{t-1}^{(a_{t-1}^{(i)})}, x_t^{(i)})$.
    end for
    for $i = 1, \ldots, N$ **do**
        Calculate the i-th weight;

$$w_t(\tilde{x}_t^{(i)}; \theta) = \frac{\gamma_t(\theta, \tilde{x}_t^{(i)})}{\gamma_{t-1}(\theta, \tilde{x}_{t-1}^{(a_{t-1}^{(i)})}) p_t(x_t^{(i)}|\tilde{x}_{t-1}^{(a_{t-1}^{(i)})}, \theta)} \; .$$

    end for
    Normalize the weights by setting, for each $i \in \{1, \ldots, N\}$,

$$\tilde{w}_t^{(i)}(\tilde{x}_t^{(1:N)}; \theta) = w_t(\tilde{x}_t^{(i)}; \theta) / (w_t(\tilde{x}_t^{(1)}; \theta) + \ldots + w_t(\tilde{x}_t^{(N)}; \theta)) \; .$$

end for

and, for any $t \in \{2, \ldots, T\}$, $\tilde{x}_t^{(i)} := (\tilde{x}_{t-1}^{(a_{t-1}^{(i)})}, x_t^{(i)})$. The xSMC procedure produces an estimator of the same form as the Sequential Monte Carlo estimator (given by (75));

$$I_T(\theta, \tilde{X}_{0:T}^{(1:N)}) := \frac{1}{N^{T+1}} \prod_{t=0}^{T} \sum_{j=1}^{N} w_t(\tilde{X}_t^{(j)}; \theta) \,, \tag{103}$$

which, as with the Sequential Monte Carlo estimator, is an unbiased approximation of $\eta_T(\theta)$:

THEOREM 4.4.3. *Let $\tilde{X}_{0:T}^{(1:N)}$ be the paths generated by the Exchangeable Sequential Monte Carlo procedure (Algorithm 19), and $\theta \in \mathbb{R}^p$. Then, for any $t \in \{0, \ldots, T\}$,*

$$\mathbb{E}_\Psi[I_T(\Theta, \tilde{X}_{0:T}^{(1:N)}) | \Theta = \theta] = \eta_T(\theta) \,,$$

where I_T is the Exchangeable Sequential Monte Carlo estimator given by Equation (103), and, given θ, Ψ is the density of the random variables generated by the Exchangeable Sequential Monte Carlo estimator, given by Equation (102).

Proof. See A.24. □

The Conditional Exchangeable Sequential Monte Carlo (CxSMC) procedure follows immediately from the Exchangeable Sequential Monte Carlo procedure, where the *conditional* propagation of particles at time t, given the known propagation of the j-th particle, is denoted

$$\begin{aligned} &\tilde{p}_t^{(N+1)}(x_t^{(0:\mathcal{L}_T(k,t)-1)}, x_t^{(\mathcal{L}_T(k,t)+1:N)} | x_t^{(\mathcal{L}_T(k,t))}, \tilde{x}_{t-1}^{(a_{t-1}^{(0:N)})}, \theta) \\ &= \frac{p_t^{(N+1)}(x_t^{(0:N)} | \tilde{x}_{t-1}^{(a_{t-1}^{(0:N)})}, \theta)}{p_t(x_t^{(\mathcal{L}_T(k,t))} | \tilde{x}_{t-1}^{(\mathcal{L}_T(k,t-1))}, \theta)} \,. \end{aligned} \tag{104}$$

As with the Conditional Sequential Monte Carlo, it will be useful to state the joint mass-density function of all the random variables produced by the xSMC procedure conditional on the k-th path being fixed:

$$\begin{aligned} &\psi(x_{0:T}^{(0:N)} \backslash \tilde{x}_T^{(k)}, a_{0:T-1}^{(0:N)} \backslash \tilde{a}_{T-1}^{(k)} | k, \tilde{x}_T^{(k)}, \tilde{a}_{T-1}^{(k)}, \theta) \\ &\propto \tilde{p}_0^{(N+1)}(x_0^{(0:\mathcal{L}_T(k,0)-1)}, x_0^{(\mathcal{L}_T(k,0)+1:N)} | x_0^{(\mathcal{L}_T(k,0))}, \theta) \\ &\quad \prod_{t=1}^{T} \frac{\kappa(a_{t-1}^{(0:N)} | \tilde{w}_{t-1}^{(0:N)})}{\mathbb{P}(A_{t-1}^{(\mathcal{L}_T(k,t))} = a_{t-1}^{(\mathcal{L}_T(k,t))} | \tilde{w}_{t-1}^{(0:N)})} \\ &\qquad \times \tilde{p}_t^{(N+1)}(x_t^{(0:\mathcal{L}_T(k,t)-1)}, x_t^{(\mathcal{L}_T(k,t)+1:N)} | x_t^{(\mathcal{L}_T(k,t))}, \tilde{x}_{t-1}^{(a_{t-1}^{(0:N)})}, \theta) \end{aligned} \tag{105}$$

To implement the Conditional Exchangeable Sequential Monte Carlo procedure one needs a sequence of proposal densities,

$$p_0^{(N+1)}(x_0^{(0:N)} | \theta), p_1^{(N+1)}(x_1^{(0:N)} | \tilde{x}_0^{(a_0^{(0:N)})}, \theta), \ldots, p_T^{(N+1)}(x_T^{(0:N)} | \tilde{x}_{T-1}^{(a_{T-1}^{(0:N)})}, \theta) \,,$$

Algorithm 20 Conditional Exchangeable Sequential Monte Carlo Procedure (Conditioned on the k-th path, $(k, \tilde{x}_T^{(k)}, \tilde{a}_{T-1}^{(k)})$)

Sample $x_0^{(0)}, \ldots, x_0^{(\mathcal{L}_T(k,0)-1)}, x_0^{(\mathcal{L}_T(k,0)+1)}, \ldots, x_0^{(N)}$ with density

$$\tilde{p}_0^{(N+1)}(x_0^{(0:\mathcal{L}_T(k,0)-1)}, x_0^{(\mathcal{L}_T(k,0)+1:N)} | x_0^{(\mathcal{L}_T(k,0))}, \theta) = \frac{p_0^{(N+1)}(x_0^{(0:N)}|\theta)}{p_0(x_0^{(\mathcal{L}_T(k,0))}|\theta)} .$$

for $i = 0, \ldots, \mathcal{L}_T(k,0) - 1, \mathcal{L}_T(k,0) + 1, \ldots, N$ **do**
    Set $\tilde{x}_0^{(i)} := x_0^{(i)}$.
end for
for $i = 1, \ldots, N$ **do**
    Calculate the i-th weight;

$$w_0(\tilde{x}_0^{(i)}; \theta) = \gamma_0(\theta, x_0^{(i)}) / p_0(x_0^{(i)}|\theta) .$$

end for
Normalize the weights by setting, for each $i \in \{0, \ldots, N\}$,

$$\tilde{w}_0^{(i)}(\tilde{x}_0^{(0:N)}; \theta) = w_0(\tilde{x}_0^{(i)}; \theta) / (w_0(\tilde{x}_0^{(0)}; \theta) + \ldots + w_0(\tilde{x}_0^{(N)}; \theta)) .$$

for $t = 1, \ldots, T$ **do**
    Sample ancestors $a_{t-1}^{(-\mathcal{L}_T(k,t))}$ with mass function

$$\frac{\kappa(a_{t-1}^{(0:N)}|\tilde{w}_{t-1}^{(0:N)})}{\mathbb{P}(A_{t-1}^{(\mathcal{L}_T(k,t))} = a_{t-1}^{(\mathcal{L}_T(k,t))}|\tilde{w}_{t-1}^{(0:N)})} ,$$

    using the conditional stratified residual resampling scheme (Algorithm 11).
    Sample $x_t^{(0)}, \ldots, x_t^{(\mathcal{L}_T(k,t)-1)}, x_t^{(\mathcal{L}_T(k,t)+1)}, \ldots, x_t^{(N)}$ with density

$$\tilde{p}_t^{(N+1)}(x_t^{(0:\mathcal{L}_T(k,t)-1)}, x_t^{(\mathcal{L}_T(k,t)+1:N)} | x_t^{(\mathcal{L}_T(k,t))}, \tilde{x}_{t-1}^{(a_{t-1}^{(0:N)})}, \theta) = \frac{p_t^{(N+1)}(x_t^{(0:N)}|\tilde{x}_{t-1}^{(a_{t-1}^{(0:N)})}, \theta)}{p_t(x_t^{(\mathcal{L}_T(k,t))}|\tilde{x}_{t-1}^{(\mathcal{L}_T(k,t-1))}, \theta)} .$$

    for $i = 0, \ldots, \mathcal{L}_T(k,t) - 1, \mathcal{L}_T(k,t) + 1, \ldots, N$ **do**
        Set $\tilde{x}_t^{(i)} := (\tilde{x}_{t-1}^{(a_{t-1}^{(i)})}, x_t^{(i)})$.
    end for
    for $i = 0, \ldots, \mathcal{L}_T(k,t) - 1, \mathcal{L}_T(k,t) + 1, \ldots, N$ **do**
        Calculate the i-th weight;

$$w_t(\tilde{x}_t^{(i)}; \theta) = \frac{\gamma_t(\theta, \tilde{x}_t^{(i)})}{\gamma_{t-1}(\theta, \tilde{x}_{t-1}^{(a_{t-1}^{(i)})}) p_t(x_t^{(i)}|\tilde{x}_{t-1}^{(a_{t-1}^{(i)})}, \theta)} .$$

    end for
    Normalize the weights by setting, for each $i \in \{0, \ldots, N\}$,

$$\tilde{w}_t^{(i)}(\tilde{x}_t^{(0:N)}; \theta) = w_t(\tilde{x}_t^{(i)}; \theta) / (w_t(\tilde{x}_t^{(0)}; \theta) + \ldots + w_t(\tilde{x}_t^{(N)}; \theta)) .$$

end for

which satisfy the properties of Assumptions 4.4.1. We also need to be able to simulate from the conditional propagation density, given by (104). As for the Exchangeable Sampler, one can, for general d, and at each time $t \in \{0, \ldots, T-1\}$, use Algorithm 21, which is a slight extension of Algorithm 18, to propagate the particles forward in an *exchangeable* way; that is, satisfying property (X) of Assumptions 4.4.1, whilst retaining $p_t(\cdot|\tilde{x}_{t-1}^{(a_{t-1}^{(i)})}, \theta)$, for $i \in \{1, \ldots, N\}$, as the marginal densities. Therefore, such a proposal density satisfies Assumptions 4.4.1, while at the same time allowing for the flexibility of making the proposals as close together as one wants via a tunable *jump-size*, ϵ_t.

Algorithm 21 Exchangeable Proposal $\tilde{p}_t^{(N+1)}(y_{1:N}|y_0, x_{0:N}, \theta)$ With Marginal p_t and Jump-Size $\epsilon_t \in (0, \sqrt{2})$ for General d.

Let $h_t : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^d$ be a mapping such that, if $Z_{1:d} \sim \mathrm{N}_d(0, I_d)$ where I_d denotes the d-dimensional identity matrix, then $h_t(Z_{1:d}, x) \sim p_t(\cdot|x, \theta)$, and that, for any $x \in \mathbb{R}^d$, the function $g_{t,x} : \mathbb{R}^d \to \mathbb{R}^d$, defined by $g_{t,x}(y) := h_t(y, x)$, is invertible.
Set $\delta_t := \epsilon_t/\sqrt{2}$.
Calculate $z_0 = g_{t,x_0}^{-1}(y_0)$.
Sample $\hat{z}_0$ from a $\mathrm{N}_d(0, I_d)$ distribution.
Set $\varphi_t = \sqrt{1-\delta_t^2} z_0 + \delta_t \hat{z}_0$.
for $i = 1, \ldots, N$ **do**
    Sample $\hat{z}_i$ from a $\mathrm{N}_d(0, I_d)$ distribution.
    Set $z_i = \varphi_t \sqrt{1-\delta_t^2} + \delta_t \hat{z}_i$.
    Set $y_i = h_t(z_i, x_i)$.
end for

REMARK 12. *Recall from the Exchangeable Sampler of Section 4.3, that, as highlighted in Remark 9, in practice, it is not neccessary to know how to do the inversion since the first step of the procedure, $z_0 = g_{t,x_0}^{-1}(y_0)$, can be omitted, provided one stores, in memory, the z_0 corresponding to the current state of the chain. However, the exposition of the Exchangeable Particle Gibbs Sampler is clearer if the proposal is explicitly conditional on y_0.*

REMARK 13. *The implementation given by Algorithm 21, much like Algorithm 18 of Section 4.3, uses the same* jump-size, *ϵ_t, for each of the d dimensions. However, as highlighted in Remark 10 for the Exchangeable Sampler, exchangeability of the proposal density and the results that follow still hold if one uses a different jump-size, $\epsilon_{t,i}$, say, for each of the dimensions. Again, as highlighted in Remark 10, the benefit of doing this would be to take larger jumps in the dimensions where one knew that the proposal, in that dimension, was closer to the target. However, as with the Exchangeable Sampler, to ease exposition, we will assume a fixed ϵ_t for each of the d dimensions.*

REMARK 14. *The implementation of the Conditional Exchangeable Sequential Monte Carlo procedure, given by Algorithm 20, which utilises the proposal given by Algorithm 21, allows for the use of a different jump-size, ϵ_t, at each time $t \in \{0, \ldots, T\}$, to propagate the particles forward. In general, because there are fewer future resampling steps the*

further in time, t, the procedure is, it would be prudent to use a smaller jump-size the smaller the value of t and a bigger jump-size the larger the value of t. While a non-constant choice for the jump-size sequence, $\epsilon_{0:t}$, allows for greater flexibility and a more effective procedure, investigating such choices is beyond the scope of this thesis.

To see how reducing the jump-sizes lessens the path degeneration problem, consider, as with the Conditional Sequential Monte Carlo procedure of Section 4.2.2, the one-dimensional Linear Gaussian model of Example 3; that is, $X_0 \sim \mathrm{N}(0,1)$, $\theta = 0.8$, and, for any $t \in \{1,\ldots,100\}$, the transition distributions are given by $(X_t|X_{t-1} = x_{t-1}) \sim \mathrm{N}(\theta x_{t-1}, 1)$, and the observation distributions are given by $Y_t|X_t = x_t \sim \mathrm{N}(x_t, 0.3)$. Using the bootstrap proposal as the marginal proposal density; that is, $p_0(x_0|\theta) = \phi(x_0; 0, 1)$, and, for any $t \in \{1,\ldots,T\}$, $p_t(x_t|x_{t-1},\theta) = \phi(x_t; \theta x_{t-1}, 1)$, we ran the Conditional Exchangeable Sequential Monte Carlo (Algorithm 20) using the proposal given by Algorithm 21 with $N = 500$ for six different choices of the sequence $\epsilon_{0:T}$. We used the same reference path, $x_{0:T}$, as that used in Section 4.2.2. Figure 62 shows plots of the five-hundred and one paths generated by the Conditional Exchangeable Sequential Monte Carlo procedure, where each subplot corresponds to a different jump-size, ϵ, which we have kept fixed for each time $t \in \{0,\ldots,T\}$; that is, for each $t \in \{0,\ldots,T\}$, $\epsilon_t = \epsilon$. From left to right and top to bottom, the subplots correspond to $\epsilon \in \{0.05, 0.2, 0.5, 0.75, 1.0, \sqrt{2}\}$. The figure illustrates that, the smaller the jump-size, the less likely the paths are to coalesce backwards through time; that is, the less likely the paths are to degenerate to the path being conditioned upon. This is because, the smaller the jump-size, the less variable the weights and, therefore, the less likely the conditioned path is to be *replicated* more than the compulsory once during resampling.

The Exchangeable Particle Gibbs Sampler is the Particle Gibbs Sampler but with the Conditional Sequential Monte Carlo procedure replaced with the Conditional Exchangeable Sequential Monte Carlo procedure (Algorithm 20). In full, the Exchangeable Particle Gibbs sampler is given by Algorithm 22.

Much like the Particle Gibbs Sampler, the Exchangeable Particle Gibbs Sampler consists of a sequence of Gibbs steps on an extended space which targets the extended density;

$$\pi_+(k,\theta,x_{0:T}^{(1:N)},a_{0:T-1}^{(1:N)}) := \\ \tilde{w}_T^{(k)}(\tilde{x}_T^{(1:N)};\theta)\frac{I_T(\theta,\tilde{x}_{0:T}^{(1:N)})}{\eta_T}\Psi(x_{0:T}^{(1:N)},a_{0:T-1}^{(1:N)}|\theta)\ , \tag{106}$$

where Ψ is the density of the random variables generated by the Exchangeable Sequential Monte Carlo procedure, given by Equation (102). Summing over the N values of k gives

$$\frac{I_T(\theta,\tilde{x}_{0:T}^{(1:N)})}{\eta_T}\Psi(x_{0:T}^{(1:N)},a_{0:T-1}^{(1:N)}|\theta)\ ,$$

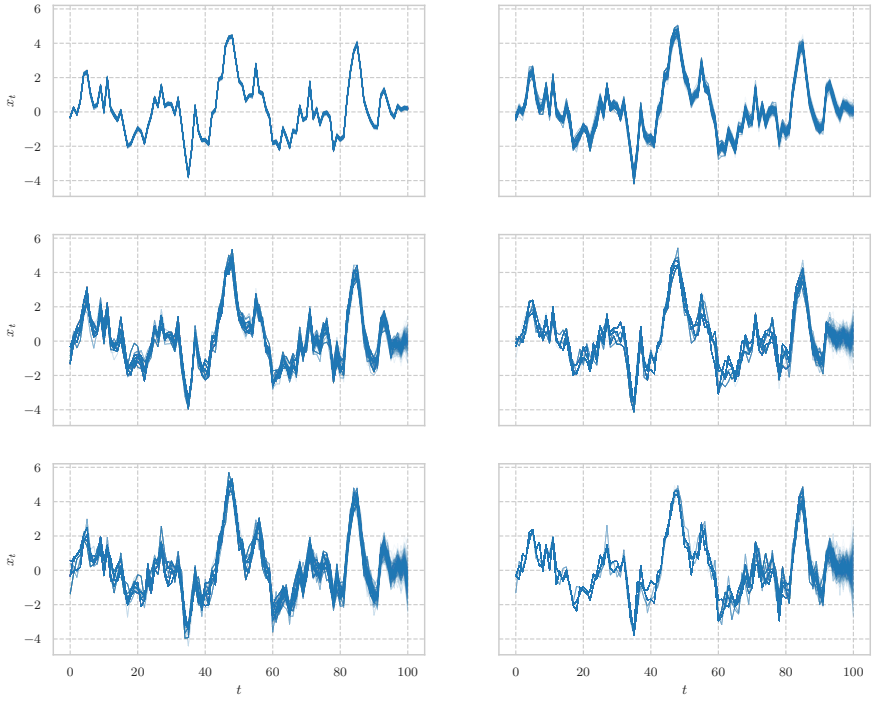

Figure 62: Plots of the five-hundred and one paths generated by the Conditional Exchangeable Sequential Monte Carlo procedure, utilising the bootstrap propsal as the marginal proposal and Algorithm 21 to propagate the particles, applied to the one-dimensional linear Gaussian model of Example 3. Each subplot corresponds to a different jump-size, ϵ, which we have kept fixed for each time $t \in \{0, \ldots, T\}$; that is, for each $t \in \{0, \ldots, T\}$, $\epsilon_t = \epsilon$. From left to right and top to bottom, the subplots correspond to $\epsilon \in \{0.05, 0.2, 0.5, 0.75, 1.0, \sqrt{2}\}$.

which, by Theorem 4.4.3, is a valid joint mass-density function. Thus, π_+ is a valid joint mass-density function. As we did for the Particle Gibbs Sampler in Lemma 4.2.4, it will be useful to rewrite the target in order to demonstrate that it exhibits π_T as the marginal density for $(\theta, \tilde{x}_T^{(k)})$;

LEMMA 4.4.4. *Under property (P) of Assumptions 4.4.2, the extended target, π_+, given by (106), can be rewritten as*

$$\begin{aligned}&\pi_+(k, \theta, x_{0:T}^{(1:N)}, a_{0:T-1}^{(1:N)}) \\ &= N^{-(T+1)} \frac{\gamma_T(\theta, \tilde{x}_T^{(k)})}{\eta_T} \psi(x_{0:T}^{(1:N)} \backslash \tilde{x}_T^{(k)}, a_{0:T-1}^{(1:N)} \backslash \tilde{a}_{T-1}^{(k)} | k, \tilde{x}_T^{(k)}, \tilde{a}_{T-1}^{(k)}, \theta) \,,\end{aligned} \tag{107}$$

where ψ is the density corresponding to the Conditional Exchangeable Sequential Monte Carlo procedure.

Proof. Given (104), the proof follows exactly the same as the proof of Lemma 4.2.4 which is given by A.10. □

Algorithm 22 Exchangeable Particle Gibbs Sampler

1: Initialise the chain at some $(\theta_0, x_{0:T}) \in \mathbb{R}^p \times \mathbb{R}^{d(T+1)}$ and choose the number of iterations, M.
2: Let $a_t^{(0)} = 0$ for all $t \in \{0, \ldots, T-1\}$, and $x_t^{(0)} = x_t$ for all $t \in \{0, \ldots, T\}$ so that $\tilde{x}_T^{(0)} = x_{0:T}$. Define $x_0^{\text{path}} := \tilde{x}_T^{(0)}$, $a_0^{\text{path}} := \tilde{a}_T^{(0)}$, and $k_0 = 0$.
3: **for** $m = 0, \ldots, M-1$ **do**
4: Set $\tilde{b}_T^{(k_m)} = a_m^{\text{path}}$ and $\tilde{y}_T^{(k_m)} = x_m^{\text{path}}$.
5: Sample θ_{m+1} with density $\pi_T^{(\theta)}(\cdot|\tilde{y}_T^{(k_m)})$.
6: Sample $(y_{0:T}^{(0:N)} \backslash \tilde{y}_T^{(k_m)}, b_{0:T-1}^{(0:N)} \backslash \tilde{b}_{T-1}^{(k_m)})$ with density

$$\psi(y_{0:T}^{(0:N)} \backslash \tilde{y}_T^{(k_m)}, b_{0:T-1}^{(0:N)} \backslash \tilde{b}_{T-1}^{(k_m)} | k_m, \tilde{y}_T^{(k_m)}, \tilde{b}_{T-1}^{(k_m)}, \theta_{m+1}) \,,$$

defined by (105) via the CxSMC procedure (see Algorithm 20).
7: Sample a $k_{m+1} \in \{0, \ldots, N\}$ with probability $\tilde{w}_T^{(k_{m+1})}(\tilde{y}_T^{(0:N)}; \theta_{m+1})$.
8: Set $x_{m+1}^{\text{path}} = \tilde{y}_T^{(k_{m+1})}$ and $a_{m+1}^{\text{path}} = \tilde{b}_T^{(k_{m+1})}$.
9: **end for**

With this alternative representation for the extended density, it is trivial to see that such a density exhibits π_T as the marginal density for $(\theta, \tilde{x}_T^{(k)})$. Indeed, integrating out all the variables not involved in the path $\tilde{x}_T^{(k)}$ gives

$$\sum_{a_{0:T-1}^{(1:N)} \backslash \tilde{a}_{T-1}^{(k)}} \int_{\mathbb{R}_+} \pi_+(k, \theta, x_{0:T}^{(1:N)}, a_{0:T-1}^{(1:N)}) \, \mathrm{d}(x_{0:T}^{(1:N)} \backslash \tilde{x}_T^{(k)}) = \frac{\gamma_T(\theta, \tilde{x}_T^{(k)})}{\eta_T} = \pi_T(\theta, \tilde{x}_T^{(k)}) \,, \tag{108}$$

where, for notational simplicity, $\mathbb{R}_+ := \mathbb{R}^{d \times (T+1) \times (N-1)}$. As with the justification of the Particle Gibbs Sampler given in Section 4.2.4, to show that the Exchangeable Particle Gibbs Sampler consists of a sequence of Gibbs steps, we consider the two forms of the extended density, given by (106) and (107) respectively:

$$\begin{aligned}
&\pi_+(k_m, \theta_m, y_{0:T}^{(0:N)}, b_{0:T-1}^{(0:N)}) \\
&= \tilde{w}_T^{(k_m)}(\tilde{y}_T^{(0:N)}; \theta_m) \frac{I_T(\theta_m, \tilde{y}_{0:T}^{(0:N)})}{\eta_T} \Psi(y_{0:T}^{(0:N)}, b_{0:T-1}^{(0:N)} | \theta_m) \\
&= (N+1)^{-(T+1)} \frac{\gamma_T(\theta, \tilde{y}_T^{(k_m)})}{\eta_T} \psi(y_{0:T}^{(0:N)} \backslash \tilde{y}_T^{(k_m)}, b_{0:T-1}^{(0:N)} \backslash \tilde{b}_{T-1}^{(k_m)} | k_m, \tilde{y}_T^{(k_m)}, \tilde{b}_{T-1}^{(k_m)}, \theta_m) \,.
\end{aligned}$$

As shown in Algorithm 22, given a current state, $(k_m, \theta_m, \tilde{y}_T^{(k_m)}, \tilde{b}_{T-1}^{(k_m)})$, in the extended space, the Exchangeable Particle Gibbs Sampler cycles through the following steps:

1. Sample a θ_{m+1} from $\pi_T^{(\theta)}(\cdot|\tilde{y}_T^{(k_m)})$.
2. Sample a sequence $(y_{0:T}^{(0:N)} \backslash \tilde{y}_T^{(k_m)}, b_{0:T-1}^{(0:N)} \backslash \tilde{b}_{T-1}^{(k_m)})$ with density

$$\psi(y_{0:T}^{(0:N)} \backslash \tilde{y}_T^{(k_m)}, b_{0:T-1}^{(0:N)} \backslash \tilde{b}_{T-1}^{(k_m)} | k_m, \tilde{y}_T^{(k_m)}, \tilde{b}_{T-1}^{(k_m)}, \theta_{m+1}) \,,$$

defined by (105) via the CxSMC procedure (see Algorithm 20).

3. Sample a $k_{m+1} \in \{0, \ldots, N\}$ with probability $\tilde{w}_T^{(k_{m+1})}(\tilde{y}_T^{(0:N)}; \theta_{m+1})$.

Note, by (108), that the extended target exhibits π_T as the marginal density for $(\theta, \tilde{y}_T^{(k_m)})$. Thus, using the terminology of Liu, 2001, Section 6.7, the first step is a *collapsed* Gibbs step. The second step is a Gibbs step on the extended space as can be seen from the second representation of the extended target:

$$\begin{aligned}&\pi_+(k, \theta, y_{0:T}^{(0:N)}, b_{0:T-1}^{(0:N)}) \\ &= (N+1)^{-(T+1)} \frac{\gamma_T(\theta, \tilde{y}_T^{(k)})}{\eta_T} \psi(y_{0:T}^{(0:N)} \backslash \tilde{y}_T^{(k)}, b_{0:T-1}^{(0:N)} \backslash \tilde{b}_{T-1}^{(k)} | k, \tilde{y}_T^{(k)}, \tilde{b}_{T-1}^{(k)}, \theta) \, .\end{aligned}$$

The third step is also a Gibbs step on the extended space as can be seen from the first representation of the extended target:

$$\begin{aligned}&\pi_+(k, \theta, y_{0:T}^{(0:N)}, b_{0:T-1}^{(0:N)}) \\ &= \tilde{w}_T^{(k)}(\tilde{y}_T^{(0:N)}; \theta) \frac{I_T(\theta, \tilde{y}_{0:T}^{(0:N)})}{\eta_T} \Psi(y_{0:T}^{(0:N)}, b_{0:T-1}^{(0:N)} | \theta) \, .\end{aligned}$$

With exactly the same proof as that of Theorem 5, Andrieu, Doucet, and Holenstein, 2010, one can, as with the Particle Gibbs Sampler, demonstrate that if each Gibbs step of the Exchangeable Particle Gibbs sampler is irreducible and aperiodic, then the Exchangeable Particle Gibbs Sampler is ergodic in the sense of Corollary 2.3.12. As such, by Theorem 2.3.13, the MCMC estimates corresponding to the samples generated by the Particle Gibbs Sampler satisfy a Strong Law of Large Numbers result. In addition, recall from Theorem 4.3.9 of Section 4.3 that the Exchangeable Sampler, which is a generalisation of the Independence Sampler, can be geometrically ergodic even if the transition weight is unbounded, provided the weight does not increase too quickly in the tails. Indeed, Section 4.3 highlights, via Example 5, a scenario where the Exchangeable Sampler is geometrically ergodic and where the transition weight is polynomially increasing in the tails. This is in contrast to the Independence Sampler which, via Theorem 2.3.31, can not be geometrically ergodic if the transition weight is unbounded. The Exchangeable Particle Gibbs Sampler, which is an extension of the Exchangeable Sampler, is a generalisation of the Particle Gibbs Sampler, which, itself, is an extension of the Independence Sampler. Recall, from Section 4.2.4, that the Particle Gibbs Sampler is uniformly ergodic if, at each time $t \in \{0, \ldots, T\}$,

$$\sup_{(x_{0:t}, \theta) \in \mathbb{R}^{d \times t} \times \mathbb{R}^p} w_t(x_{0:t}; \theta) < \infty \, ;$$

that is, if, at each time $t \in \{0, \ldots, T\}$, the transition weight at time t is uniformly bounded. Moreover, if this assumption does not hold then the Particle Gibbs Sampler can not be geometrically ergodic. Therefore, it is natural to conjecture that, under certain conditions, the Exchangeable Particle Gibbs Sampler can be geometrically ergodic even if the weights, $w_{0:T}$, are unbounded, provided the weights do not increase

too quickly in the *tails*. Even when the Exchangeable Particle Gibbs Sampler is not geometrically ergodic, the flexibility of being able to control how close the proposed paths are to the path being conditioned upon by tuning a sequence of jump-sizes, $\epsilon_{0:T}$, means that, in practice, one has the abillity to tune the algorithm to optimise the rate of mixing, and therefore provide a more efficient algorithm than the Particle Gibbs Sampler. Moreover, unlike the Particle Gibbs with Ancestor Sampling approach of Lindsten, Jordan, and Schön, 2014, the main requirement needed to be able to implement the Exchangeable Particle Gibbs Sampler is the ability, at each time $t \in \{0, \ldots, T\}$, to be able to transform a sequence of independent standard Normal random variables to random variables with a marginal density p_t and, therefore, implement Algorithm 21. While this is not always possible to do in a computationally efficient way, it is often possible using the inverse transform of Theorem 2.3.2. Of particular importance to this thesis, it is possible to implement Algorithm 21 in an efficient way when the marginal proposal density corresponds to a conditioned diffusion proposal of Chapter 3 since, in this case, the proposal density consists of a sequence of Gaussian increments. See, for example, Section 4.3 for details on how to implement Algorithm 21 in the case where the marginal proposal density corresponds to the Modified Diffusion Bridge applied to the conditioned Birth-Death diffusion.

4.4.1 *Optimal Scaling*

In the previous section we introduced the Exchangeable Particle Gibbs Sampler which extends the Particle Gibbs Sampler through the introduction of a *jump-size* which can be tuned to make the proposed paths at each step of the sampler closer to the path being conditioned upon. This flexibility allows one to improve the rate of mixing of the sampler by carefully selecting the *jump-size*. In this section, we will, under fairly stringent assumptions, derive *optimal scaling* results which are in the same spirit as the optimal scaling results of Sections 2.3.6.3 and 4.3.1; that is, results which practitioners can use as a general guide on how to choose the jump-size so as to maximize the rate of mixing of the Markov Chain induced by the sampler. As previously, we will, as in Sherlock and Roberts, 2009, use the expected squared jump distance as the measure of efficiency. As in Section 4.3.1, due to the difficulty of theoretically analysing the transformation which maps the underlying Normal random variables to the proposals, we consider the transformed space on which the Normal random variables lie. Moreover, recall that the motivation for the Exchangeable Particle Gibbs Sampler was to reduce the path degeneracy problem and, therefore, improve the mixing of the elements of the paths at times, t, closer to 0. Thus, we will take the expected squared jump distance in the first component of the path as a measure of efficiency. Heuristically, maximizing this measure of efficiency will generally reduce the degeneracy of the paths and, hence, optimize the rate of mixing across all components of the path.

In order to theoretically analyse the expected squared jump distance and the expected acceptance rate for the first component of the path in the Z-space, we consider a specific form of the Exchangeable Particle Gibbs Sampler which targets a density, π_T^*, of a product form by using a sequence of marginal densities, $p_{0:T}^*$, which, themselves, are also of a product form:

DEFINITION 4.4.5. *Let the number of samples, $N \in \mathbb{N}$, the dimension, d, the jump-size, $\epsilon \in (0, \sqrt{2})$, and T be fixed. Consider the Exchangeable Particle Gibbs Sampler (Algorithm 22) which targets the density*

$$\pi_T^*(x_{0:T}^{(1:d)}) := \prod_{t=0}^{T} \prod_{i=1}^{d} \pi(x_t^{(i)}) ,$$

by using a sequence, $p_{0:T}^*$, of marginal proposal densities of the form

$$p_t^*(x_t^{(1:d)}) = p^*(x_t^{(1:d)}) := \prod_{i=1}^{d} p(x_t^{(i)}) ,$$

as part of Algorithm 21 to generate proposal paths through the Conditional Exchangeable Sequential Monte Carlo procedure (Algorithm 20). Specifically, let $X_{0:T}^{(1:d)}$ be an independent sequence of random variables where, for any $(t, i) \in \{0, \ldots, T\} \times \{1, \ldots, d\}$, $X_t^{(i)} \sim \pi$. Suppose p is a one-dimensional density with cumulative density function P. Let the transformation, h^*, be given by

$$h^*(z^{(1:d)}) := (P^{-1}[\Phi(z^{(1)})], \ldots, P^{-1}[\Phi(z^{(d)})]) ,$$

where Φ denotes the cumulative density function corresponding to a standard normal random variable. Then, by Theorem 2.3.2, if $Z^{(1:d)} \sim \mathrm{N}_d(0, I_d)$, then $h^*(Z^{(1:d)}) \sim p^*$. Now, let $h := P^{-1} \circ \Phi$ so that, if $X_t^{(i)} \sim \pi$, then $Z_t^{(i)} = h^{-1}(X_t^{(i)})$ has density $\pi_Z(z_t^{(i)}) := \pi[h(z_t^{(i)})]|h'(z^{(i)})|$. Note that the transformation h^* satisfies the necesarry assumptions of the exchangeable proposal given by Algorithm 21. Suppose, for each $t \in \{0, \ldots, T\}$, $Z_{t,0}^{(1:d)} \sim \pi_Z^*$ where

$$\pi_Z^*(z_{t,0}^{(1:d)}) := \prod_{i=1}^{d} \pi_Z(z_{t,0}^{(i)}) ,$$

and let $\hat{Z}_{t,0:N}^{(1:d)}$ be an independent sequence of d-dimensional random variables such that, for any $k \in \{0, \ldots, N\}$, $\hat{Z}_{t,k}^{(1:d)} \sim \mathrm{N}_d(0, I_d)$. For each $k \in \{1, \ldots, N\}$ define

$$Z_{t,k}^{(1:d)} := (1 - \delta^2) Z_{t,0}^{(1:d)} + \delta \sqrt{1 - \delta^2} \hat{Z}_{t,0}^{(1:d)} + \delta \hat{Z}_{t,k}^{(1:d)} ,$$

where $\delta := \epsilon / \sqrt{2}$. Furthermore, for any $k \in \{1, \ldots, N\}$, let $\alpha_{k,N}^*(w_{0:N}^*)$ be the multiple-proposal extension Barker's acceptance probability (Equation (86)) expressed in terms of the transition weights, which are of the form $w^* = \pi^*/p^*$, where

$$\pi^*(x_t^{(1:d)}) := \prod_{i=1}^{d} \pi(x_t^{(i)}) ;$$

that is,

$$\alpha^*_{k,N}(w^*_{0:N}) = \frac{w^*_k}{w^*_0 + \ldots + w^*_N} \,.$$

Let $g^* := w^* \circ h^*$*; that is,*

$$g^*(z^{(1:d)}) := \prod_{i=1}^{d} w[h(z^{(i)})] \,,$$

where $w = \pi/p$ *is the marginal transition weight. Finally, let* $g := w \circ h$*, and* $\mathcal{L}_T$ *denote the lineage function as defined in Definition 4.2.2. Then, at each time* $t \in \{0, \ldots, t\}$*, we define the expected squared jump distance for the first component of the path in the* z*-space to be*

$$J_{t,N}(\epsilon) := \mathbb{E}\left[\sum_{k=1}^{N} \alpha^*_{k,N}(g^*(Z^{(1:d)}_{t,0}), \ldots, g^*(Z^{(1:d)}_{t,N}))\|Z^{(1:d)}_{0,\mathcal{L}_t(k,0)} - Z^{(1:d)}_{0,0}\|^2\right] , \tag{109}$$

and the expected probability that an ancestor, which is not being conditioned upon, is not equal to zero to be

$$\alpha_{t,N}(\epsilon) := \mathbb{E}\left[\sum_{k=1}^{N} \alpha^*_{k,N}(g^*(Z^{(1:d)}_{t,0}), \ldots, g^*(Z^{(1:d)}_{t,N}))\right] . \tag{110}$$

To derive an optimal scaling result for the expected squared jump distance and expected acceptance rate for the first component of the path in the Z-space, given by Equations (109) and (110) respectively, the following assumptions on the densities π_Z and the transition weight expressed in terms of z; that is, h^*, are needed:

ASSUMPTIONS 4.4.6.

(B) Let $Z \sim \pi_Z$, where π_Z is as defined in Definition 4.4.5:

(B.a) For any $k \in \{1, \ldots, 4\}$, $\mathbb{E}[|Z|^k] < \infty$.

(B.b) The logarithm of the marginal transition weight expressed in terms of z; that is, $q(z) := \log[g(z)]$, where $g := w \circ h$, and w, h are defined in Definition 4.4.5, is twice differentiable and satisfies

$$\mathbb{E}[q'(Z)^2] < \infty \,, \quad \mathbb{E}[q''(Z)^2] < \infty \,.$$

(L) The second derivative of q is Lipschitz continuous with Lipschitz constant a; that is, for any $z_{0:1} \in h^{-1}(\mathcal{X}) \times h^{-1}(\mathcal{X})$,

$$|q''(z_1) - q''(z_0)| \leq a|z_1 - z_0| \,.$$

(G) The transition weight expressed in terms of z; that is, g', is sufficiently well-behaved in the tails in the sense that

$$\lim_{z \uparrow \infty} g'(z)\phi(z) = \lim_{z \downarrow -\infty} g'(z)\phi(z) = 0 \,,$$

where ϕ is the density of a standard normal random variable.

With these assumptions in place, an optimal scaling result for the Exchangeable Particle Gibbs Sampler can be demonstrated;

THEOREM 4.4.7. *Consider the Exchangeable Particle Gibbs Sampler given in Definition 4.4.5 which targets a density, π_T^*, of a product form by using a sequence of marginal densities, $p_{0:T}^*$, which, themselves, are also of a product form. Suppose that Assumptions 4.4.6 hold. Then, for any $t \in \{0, \ldots, T\}$,*

$$\lim_{d\uparrow\infty} \alpha_{t,N}(\lambda d^{-1/2}) = \bar{\alpha}_N(\lambda) := 1 - \mathbb{E}\left[\left\{1 + \exp(-\xi^2)\exp(\xi W_0)\sum_{k=1}^{N}\exp(\xi W_k)\right\}^{-1}\right], \tag{111}$$

where $\xi := \lambda\sqrt{\varphi}/\sqrt{2}$, $\varphi := \mathbb{E}[q'(Z_0^2)]$, $Z_0 \sim \pi_Z$ where π_Z is as defined in Definition 4.4.5, and $W_{0:N}$ is an independent sequence of one-dimensional standard Normal random variables. Moreover,

$$\lim_{d\uparrow\infty} J_{T,N}(\lambda d^{-1/2}) = \bar{J}_{T,N}(\lambda) := \bar{\alpha}_N(\lambda)^T \lim_{d\uparrow\infty} J_{0,N}(\lambda d^{-1/2}) = \lambda^2 \bar{\alpha}_N(\lambda)^{(T+1)}. \tag{112}$$

Note that $\bar{\alpha}_N(\lambda)^T$ is the asymptotic expected probability of moving to a path which is not the path being conditioned upon; that is, the asymptotic acceptance rate.

Proof. See A.25. □

As with the quantities in Theorem 4.3.17, the asymptotic quantities given by Theorem 4.4.7 are intractable. For any fixed $(N, T) \in \mathbb{N}^2$, one can, as we did for the Exchangeable Sampler in Section 4.3.1, numerically analyse the asymptotic expected squared jump distance for the first component of the path in the Z-space; that is, $\bar{J}_{T,N}(\lambda)$, as a function of the asymptotic acceptance rate; that is, $\bar{\alpha}_N(\lambda)^T$ in order to derive practical guidance on how to tune the jump-size to optimise the rate of mixing. However, as with the Particle Gibbs Sampler, it is practically useful to know how N should depend on T to get sufficiently good *mixing* without choosing N too large. As highlighted in Section 4.2.4, the larger the value of T, the more *information* the sampler has to infer, and so it is reasonable to believe that N must scale with T somehow. However, choosing N too big means wasting computational effort and increasing run-times. On the other hand, choosing N too small may potentially result in poor mixing. In the latter case the sampler would have to be run for more iterations to produce samples that *accurately* represent the target, and, this increase in the number of iterations would, again, increase the computational effort and run-times. Recall, from Section 4.2.4, that, for the Particle Gibbs Sampler, Propositions 4 and 5, Lindsten, Douc, and Moulines, 2015, and Theorem 3 of Andrieu, Lee, and Vihola, 2018 demonstrate that, under suitable *strong-mixing* conditions, it is sufficient to scale the number of particles, N, linearly with T in order to obtain a non-degenerate lower-bound on the minorizing constant in the limit as $T \to \infty$. In Theorem 6, Lindsten, Douc,

and Moulines, 2015 show that, under weaker, *moment* conditions, the minorization constant is *probabilistically* bounded below in the limit as $T \to \infty$ provided one scales the number of particles, N, superlinearly with T (see Lindsten, Douc, and Moulines, 2015 for a detailed statement of the result). Given such results in the literature, it is reasonable to conjecture that, for the Exchangeable Particle Gibbs Sampler, one should, in some sense, scale the number of particles, N, at most linearly with T. To this end, we consider the asymptotic *efficiency* which we define to be the the asymptotic expected squared jump distance for the first component of the path in the z-space over the number of particles; that is,

$$e_{T,N}(\lambda) := N^{-1} \bar{J}_{T,N}(\lambda) \,, \tag{113}$$

where $\bar{J}_{T,N}$ is given by (112). Efficiency, when defined in this way, is therefore the expected squared jump distance per unit of computational cost, assuming that the computational cost of the Exchangeable Particle Gibbs Sampler scales linearly with the number of particles, N. To derive a result on how N should scale with T, we first prove the following bounds on the asymptotic expected acceptance rate, $\bar{\alpha}_N(\lambda)$;

THEOREM 4.4.8. *Let ρ_N be the asymptotic expected rejection rate defined by $\rho_N(\lambda) := 1 - \bar{\alpha}_N(\lambda)$, where $\bar{\alpha}_N$ is defined by (111). Then,*

$$(N+1)^{-1} \leq \rho_N(\lambda) \leq \mathbb{E}\{[1 + N \exp(-\xi^2 + \xi Z \sqrt{1 + 1/N})]^{-1}\} \,, \tag{114}$$

where $Z \sim \mathrm{N}(0,1)$.

Proof. See A.26. □

Using this bound we can demonstrate that, scaling N with T appropriately, one can *control* the asymptotic *efficiency*;

COROLLARY 4.4.9. *Consider the Exchangeable Particle Gibbs Sampler given in Definition 4.4.5 which targets a density, π_T^*, of a product form by using a sequence of marginal densities, $p_{0:T}^*$, which, themselves, are also of a product form. Let $e_{T,N}$ denote the asymptotic efficiency, defined by (113), and suppose that Assumptions 4.4.6 hold. Then, for any $\beta \in (0,1)$ and $c_0 > 0$, if we let $N := c_0 T^{(1-\beta)}$, then*

$$\lim_{T \uparrow \infty} N e_{T,N}(\lambda) = 0 \,.$$

Moreover, for any $c_0 > 0$, if we let $N := c_0 T$, then there exist constants $\tilde{c}_1 > 0$ and $\tilde{c}_2 > 0$ such that

$$\lambda^2 \exp(-\tilde{c}_2) \leq \lim_{T \uparrow \infty} N e_{T,N}(\lambda) \leq \lambda^2 \exp(-\tilde{c}_1) \,.$$

Finally, for any $\beta > 0$ and $c_0 > 0$, if we let $N := c_0 T^{(1+\beta)}$, then

$$\lim_{T \uparrow \infty} N e_{T,N}(\lambda) = 1 \,.$$

Proof. See A.27. □

It is clear from Corollary 4.4.9 that, if one scales N sub-linearly with T, then the asymptotic efficiency scales like 0 as T tends towards infinity. However, if N scales linearly with T, then the asymptotic efficiency scales like T^{-1} as T tends towards infinity. Finally, if N scales super-linearly with T, then the asymptotic efficiency scales like $T^{-(1+\beta)}$ as T tends towards infinity. This suggests, therefore, that, in order to optimise the efficiency of the Exchangeable Particle Gibbs Sampler, one should scale the number of particles linearly with T at least for large enough T. To see this numerically, note that $\lambda^2 = 2\xi^2/\varphi$. Therefore,

$$e_{T,N}(\lambda) = \tilde{e}_{T,N}(\xi) := \frac{2\xi^2}{N\varphi}\tilde{\alpha}_N(\xi)^{(T+1)} ,$$

where

$$\tilde{\alpha}_N(\xi) := 1 - \mathbb{E}\left[\left\{1 + \exp(-\xi^2)\exp(\xi W_0)\sum_{k=1}^{N}\exp(\xi W_k)\right\}^{-1}\right] .$$

Hence, for each T, one can numerically find the optimal $N^*(T)$ and optimal scaling, $\xi^*(T)$;

$$(N^*(T), \xi^*(T)) := \underset{(N,\xi)\in\mathbb{N}\times[0,\infty)}{\arg\min} \left[N^{-1}\xi^2\tilde{\alpha}_N(\xi)^{(T+1)}\right] .$$

Figure 63 shows that the optimal value of N scales linearly with T; roughly, $N^*(T) = 5T/2$, and that the optimal scaling and optimal asymptotic acceptance rate scale sub-linearly with T. This visualisation numerically corroborates the optimal scaling results of Corollary 4.4.9.

Now that we have given guidance on choosing N, it remains to describe how one should tune the acceptance rate in order to maximise the rate of mixing of the sampler. To this end, Figure 64 shows a plot of the true asymptotic expected efficiency (blue line) and a lower bound on the asymptotic expected efficiency provided by Theorem 4.4.8 (orange line), up to a constant of proportionality, against the asymptotic acceptance rate for the Particle Gibbs Sampler for a variety of pairs (T, N); $(T, N) \in \{(1, 2), (5, 12), (10, 25), (50, 125), (100, 250), (500, 1250)\}$. Note that we have taken $T \in \{1, 5, 10, 50, 100, 500\}$ and set $N = 5T/2$ as the resuts of Figure 63 suggest. As was the case for the Exchangeable Sampler, the optimal asymptotic expected efficiency is fairly insensitive to choices of the asymptotic acceptance rate around the optimum. This is true for every value of (T, N) considered. Indeed, for $(T, N) = (1, 2)$, an asymptotic acceptance rate in the interval $[0.2, 0.56]$ leads to an asymptotic expected efficiency which is above 60% of the optimal. This interval is similar for the other values of (T, N). As such, as was the case for the Exchangeable Sampler of Section 4.3, it is unnecessary to finely tune the *jump-size* to achieve the optimal acceptance rate, provided the *tuned* acceptance rate is on the same scale as the optimal asymptotic acceptance rate. Also, note that the value of the asymptotic acceptance rate which optimses the asymptotic efficiency is very close to the value of the asymptotic acceptance rate which optimses the lower bound on asymptotic efficiency- as given by Theorem 4.4.8. As such, for any pair

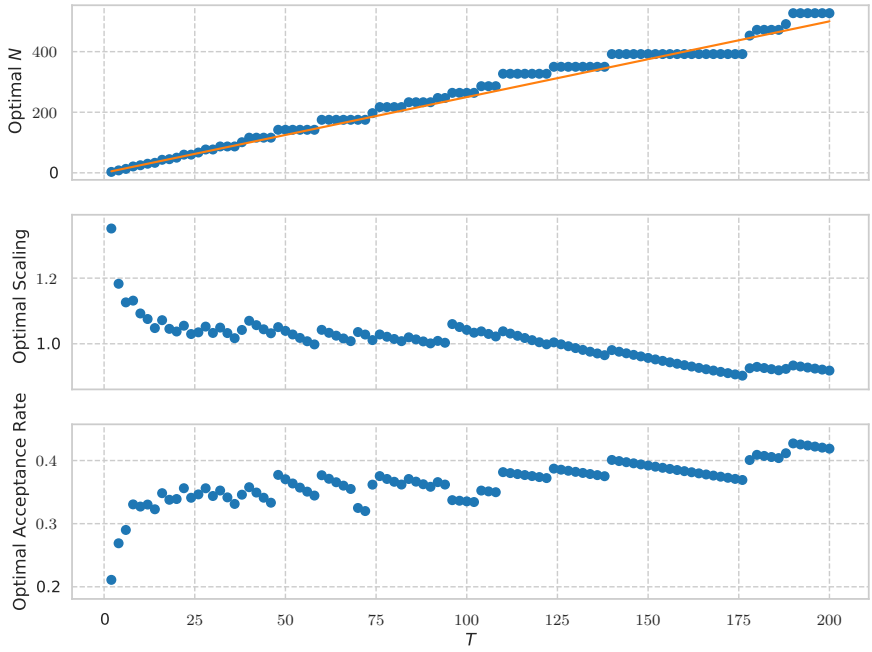

Figure 63: Plots of the optimal value of N (top subplot), optimal value of ξ (middle subplot), and optimal acceptance rate (bottom subplot) against T. The top subplot also includes the line $N = 5T/2$ in orange.

(T, N), one can find a target acceptance rate by optimising the lower bound on the asymptotic efficiency.

Figure 64 highlights the theoretical behaviour of the asymptotic efficiency as a function of the asymptotic acceptance rate, and, as was the case for the end of Section 4.3.1, while monitoring the *running sample efficiency* in order to tune the value of the jump-size is a sensible strategy one can use in practice to optimise the mixing of the chain, it is prudent to understand to what extent such theoretical results hold for finite d and for models with non-independent transition distributions. To this end, we will consider a d-dimensional extension of the Linear Gaussian model given by Example 2; that is, let $X_0 \sim \mathrm{N}_d(0, \mathcal{I}_d)$, and suppose that, for any $t \in \{1, \ldots, T\}$, we have the transition distributions given by $(X_t | X_{t-1} = x_{t-1}) \sim \mathrm{N}_d(0.8x_{t-1}, \mathcal{I}_d)$, and observation distributions given by $Y_t | X_t = x_t \sim \mathrm{N}_d(x_t, 0.3\mathcal{I}_d)$. For simplicity, suppose that an improper, uniform over $\mathbb{R}$, prior is placed on Θ so that $\gamma_0(\theta, x_0) = \phi_d(x_0; 0, \mathcal{I}_d)$ and, therefore, $\eta_0(\theta) = 1$. Moreover, suppose that, for any $t \in \{1, \ldots, T\}$, γ_t is defined recursively by

$$\gamma_t(\theta, x_{0:t}) = g_t(y_t | x_t) \phi_d(x_t; \theta x_{t-1}, \mathcal{I}_d) \gamma_{t-1}(\theta, x_{0:t-1}) \ ,$$

where $\phi_d(\cdot; \mu, \Sigma)$ denotes the density of a d-dimensional normal distribution with mean vector μ and variance-covariance matrix Σ. For each $d \in \{1, 2, 5, 10, 25, 50\}$ we simulate a sequence $y_{1:T}$ of observations from this model, and, given this sequence we simulated, for each $T \in \{5, 10, 50, 100\}$, the Exchangeable Particle Gibbs Sampler for ten-

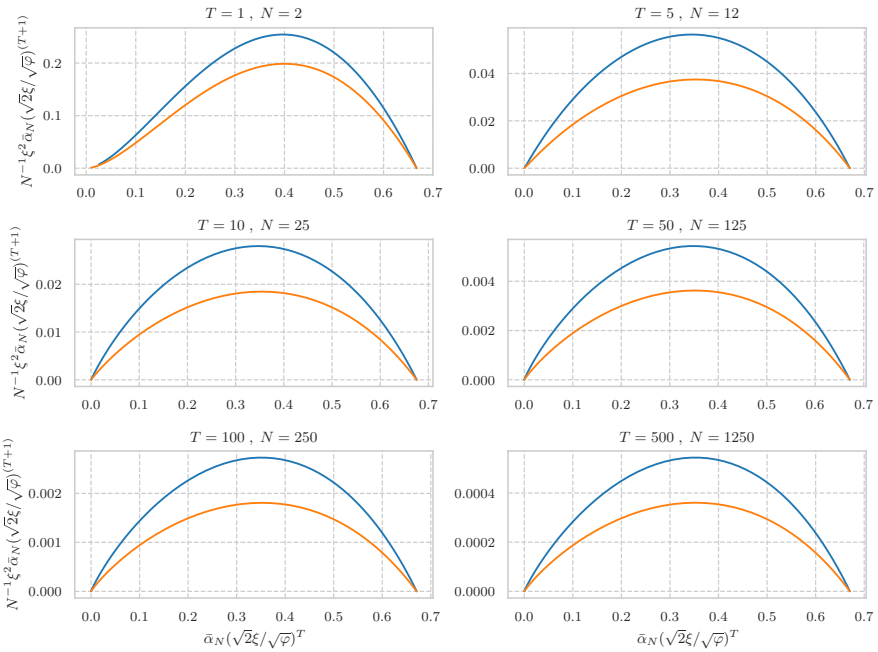

Figure 64: Plots of the true asymptotic expected efficiency (blue line) and a lower bound on the asymptotic expected efficiency provided by Theorem 4.4.8 (orange line), up to a constant of proportionality, against the asymptotic acceptance rate for the Particle Gibbs Sampler for a variety of pairs (T, N).

thousand iterations using the bootstrap proposals; that is, $p_0(x_0|\theta) = \phi_d(x_0; 0, \mathcal{I}_d)$, and, for any $t \in \{1, \ldots, T\}$, $p_t(x_t|x_{t-1}, \theta) = \phi_d(x_t; 0.8x_{t-1}, \mathcal{I}_d)$, as the marginal proposal densities for each $N \in \{1, 2, 5, 12, 25, 125, 250, 1250\}$, and for ten values of ϵ linearly spaced on the interval $[0.01, \sqrt{2}]$; that is $\epsilon \in \{0.01, 0.01 + (\sqrt{2} - 0.01)/9, 0.01 + 2(\sqrt{2} - 0.01)/9, \ldots, \sqrt{2}\}$. For each *run* of the sampler, we calculated the acceptance rate and the mean squared-jump distance for the first component of the path in the Z-space. Figure 65 shows, for each $d \in \{1, 2, 5, 10, 25, 50\}$, a plot of the value of N in the set $\{1, 2, 5, 12, 25, 125, 250, 1250\}$ that optimises the efficiency as a function of $T \in \{5, 10, 50, 100\}$ for the Exchangeable Particle Gibbs Sampler for this scenario. The figure shows that, although, for this scenario, the relationship between the value of T and the value of N which optimises the efficiency is not $N^*(T) = 5T/2$ as the theoretical behaviour shown in Figure 63 suggests, the optimal value of N does appear to scale linearly with T which is what Corollary 4.4.9 suggests. Figures 66 and 67 show, respectively, for each $d \in \{1, 2\}$, and each $d \in \{5, 10, 25, 50\}$, plots of the sample efficiency against the sample acceptance rate for the Exchangeable Particle Gibbs Sampler for this scenario and for pairs of (T, N) in $\{(1, 2), (5, 12), (10, 25), (50, 125)\}$. Figure 67 shows that, for $d \geq 5$, although the behaviour of the sample efficiency as a function of the sample acceptance rate does not match exactly the theoretical behaviour given by Figure 64, the behaviour is very similar. Indeed, the optimal acceptance rate is; around 0.4 for $(T, N) = (1, 2)$ and $(T, N) = (5, 12)$, and around 0.5 for $(T, N) = (10, 25)$ and

$(T, N) = (50, 125)$, compared to a theoretical optimal acceptance rate of around 0.4 for $(T, N) = (1, 2)$, and just below 0.4 for $(T, N) = (5, 12)$, $(T, N) = (10, 25)$ and $(T, N) = (50, 125)$. Moreover, the insensitivity of the choices of the acceptance rate around the optimum is similar to the insensitivity around the optimum in the theoretical case. In summary, for $d \geq 5$, in this particular scenario, one can get close to the optimal rate of mixing by using the theoretical optimal scaling results to tune the sampler. Figure 66, on the other hand, shows that, for $d \in \{1, 2\}$, the behaviour of the sample efficiency as a function of the sample acceptance rate is very different from the theoretical behaviour. In particular, the range of acceptance rates is much less than the theoretical range. For $d = 1$, for instance, one should choose $\epsilon = \sqrt{2}$ in order to optimise the mixing of the sampler regardless of the values considered for the pair (T, N). This is because, as was the case in Section 4.3.1, for smaller d, one can choose a larger jump-size to get the same acceptance rate for a larger d, and this is truer the smaller T is. For $d = 2$, on the other hand, one should, for each pair (T, N), choose a jump-size less than $\sqrt{2}$ to optimise the mixing of the sampler. Indeed, the optimal acceptance rate is; slightly larger than 0.2 for $(T, N) = (1, 2)$ and $(T, N) = (5, 12)$, around 0.3 for $(T, N) = (10, 25)$, and around 0.4 for $(T, N) = (50, 125)$. Moreover, for $d = 2$, the sample efficiency is fairly insensitive to choices of the acceptance rate around the optimal acceptance rate. Therefore, even for $d = 2$ for this scenario, one can use the theoretical optimal scaling results to tune the sampler.

4.4.2 *A Simulation Study*

In this section we will look at the performance of the Exchangeable Particle Gibbs Sampler in two examples. In the first example we will consider the Linear Gaussian model of Example 3; that is, $X_0 \sim \mathrm{N}(0, 1)$, $\theta = 0.8$, and, for any $t \in \{1, \ldots, 100\}$, the transition distributions are given by $(X_t | X_{t-1} = x_{t-1}) \sim \mathrm{N}(\theta x_{t-1}, 1)$, and the observation distributions are given by $Y_t | X_t = x_t \sim \mathrm{N}(x_t, 0.3)$. We will use the bootstrap proposals as the marginal proposal densities; that is, $p_0(x_0|\theta) = \phi(x_0; 0, 1)$, and, for any $t \in \{1, \ldots, T\}$, $p_t(x_t|x_{t-1}, \theta) = \phi(x_t; \theta x_{t-1}, 1)$. Note that, for this example, the *transition* weights are given by

$$w_t(\tilde{x}_t^{(i)}; \theta) = \phi(y_t; x_t, 0.3) \, ,$$

and are, therefore, bounded. Thus, by Corollary 2 of Lindsten, Douc, and Moulines, 2015, and Theorem 1 of Andrieu, Lee, and Vihola, 2018, the Particle Gibbs Sampler is uniformly ergodic as discussed in Section 4.2.4. Choosing the same observations as those given in Figure 22, we ran the Exchangeable Particle Gibbs Sampler for each of the thirty-two combinations of

$$(\epsilon, N) \in \{0.01, 0.1, 0.25, 0.5, 0.75, 1.0, 1.2, \sqrt{2}\} \times \{50, 100, 250, 1000\}$$

for one-hundred-thousand iterations. To garner an initial reference path, we ran the Sequential Monte Carlo procedure (Algorithm 8) with N

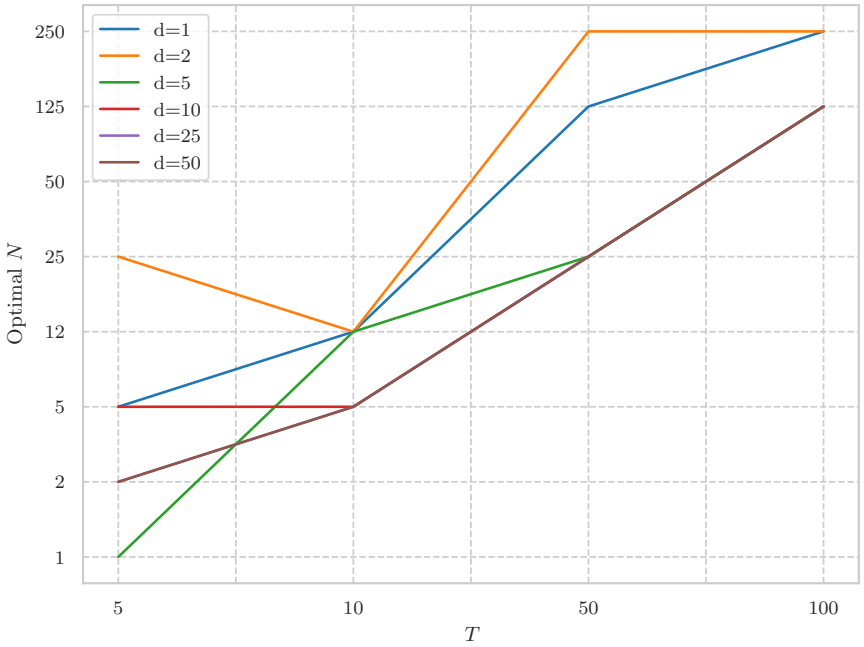

Figure 65: For each $d \in \{1, 2, 5, 10, 25, 50\}$, a plot of the value of $T \in \{5, 10, 50, 100\}$ against the value of N in the set $\{1, 2, 5, 12, 25, 125, 250, 1250\}$ which optimises the efficiency for the Exchangeable Particle Gibbs Sampler which targets the d-dimensional extension of the Linear Gaussian model given by Example 2 by using the bootstrap proposals as the marginal proposal densities.

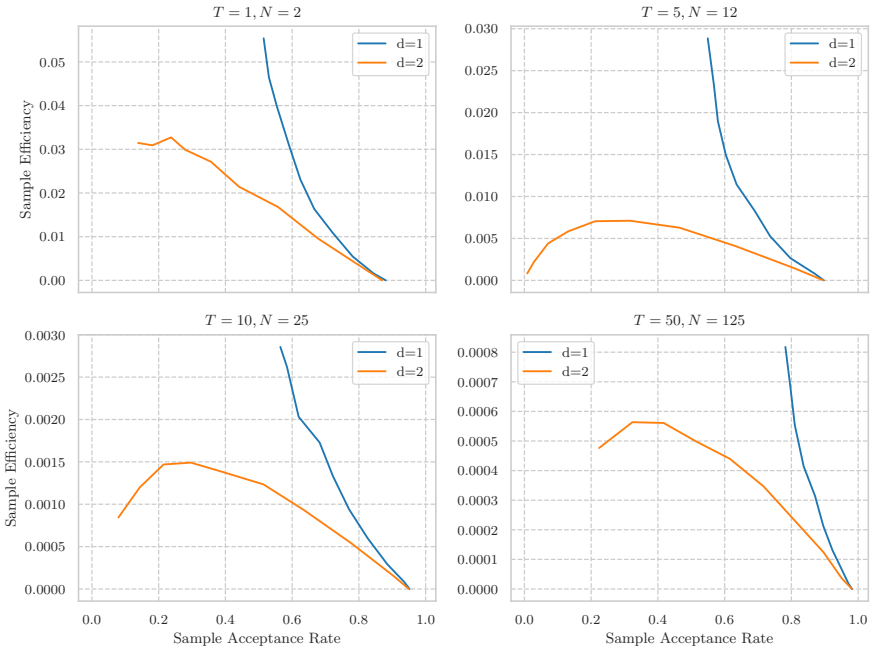

Figure 66: Plots, for each $d \in \{1, 2\}$, and each $(T, N) \in \{(1, 2), (5, 12), (10, 25), (50, 125)\}$, of the sample efficiency against the sample acceptance rate for the Exchangeable Particle Gibbs Sampler which targets the d-dimensional extension of the Linear Gaussian model given by Example 2 by using the bootstrap proposals as the marginal proposal densities.

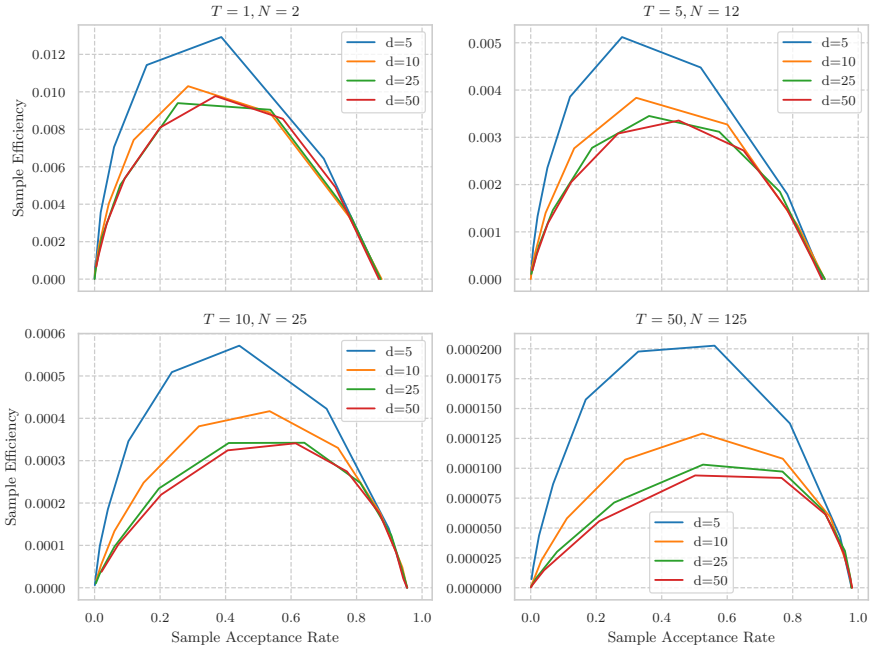

Figure 67: Plots, for each $d \in \{5, 10, 25, 50\}$, and each $(T, N) \in \{(1, 2), (5, 12), (10, 25), (50, 125)\}$, of the sample efficiency against the sample acceptance rate for the Exchangeable Particle Gibbs Sampler which targets the d-dimensional extension of the Linear Gaussian model given by Example 2 by using the bootstrap proposals as the marginal proposal densities.

particles and chose the path with the largest terminal weight, $\tilde{w}_T^{(k)}$. To generate a *good* representation of the *truth*, we ran the Particle Gibbs Sampler (Section 4.2.4) with ten-thousand particles for one-million iterations. We will compare the *performance* of the samplers on the first component of the simulated paths. As discussed when motivating the optimal scaling results of Section 4.4.1, maximizing the *performance* of the samplers for the first component of the simulated paths, will, in general, lessen the path degeneracy phenomena and, hence, optimize the rate of mixing across all components of the path. For each run of the sampler, we calculated the sample efficiency; that is, the sample squared jump distance divided by the number of particles, along with the acceptance rate for the first component of the path in the X-space. We also calculate the two-sample Kolmogorov-Smirnov statistic given by (101) between the first element of the samples simulated via the Exchangeable Particle Gibbs Sampler and the first element of the *true* samples as a measure of how *well* the simulated samples represent the truth.

In the second example we will consider a Lotka-Volterra diffusion (see Section 3.1.2) which we observe at regular intervals. In particular, we take $\theta = (\theta_1, \theta_2, \theta_3) = (0.5, 0.0025, 0.3)$ to be the parameters driving the diffusion, $x_0 = (150, 79)$ to be the initial conditions, and consider the model where, the observation distributions, for any $t \in \{2, 4, \ldots, 20\}$, are given by $Y_t | X_t = x_t \sim \mathrm{N}_2(x_t, 10^{-12}\mathcal{I}_2)$- thereby, essentially, corresponding to exact observations of the diffusion- and, for any $t \in \{2, 4, \ldots, 20\}$, the transition, $(X_t | X_{t-2} = x_{t-2})$, corresponds to the Lotka-Volterra diffusion; that is, $X_t = [X_t^{(1)}, X_t^{(2)}]^*$, where

$$\begin{bmatrix} \mathrm{d}X_t^{(1)} \\ \mathrm{d}X_t^{(2)} \end{bmatrix} = \begin{bmatrix} \theta_1 X_t^{(1)} - \theta_2 X_t^{(1)} X_t^{(2)} \\ \theta_2 X_t^{(1)} X_t^{(2)} - \theta_3 X_t^{(2)} \end{bmatrix} \mathrm{d}t + \begin{bmatrix} \theta_1 X_t^{(1)} + \theta_2 X_t^{(1)} X_t^{(2)} & -\theta_2 X_t^{(1)} X_t^{(2)} \\ -\theta_2 X_t^{(1)} X_t^{(2)} & \theta_2 X_t^{(1)} X_t^{(2)} + \theta_3 X_t^{(2)} \end{bmatrix}^{1/2} \mathrm{d}W_t \, .$$

Here, for a matrix A, $A^{1/2}$ denotes any matrix square-root, so that $(A^{1/2})(A^{1/2})^* = A$. We set $\Delta t = 0.1$ and used the EM approximation (Section 3.2.2) to forward simulate values of the diffusion, and then used the observation distribution, $Y_t | X_t = x_t \sim \mathrm{N}_2(x_t, 10^{-12}\mathcal{I}_2)$, to simulate a sequence of observations, $y_2, y_4, \ldots, y_{20}$. The observations and the paths that generated them can be seen in Figure 68. For the proposal densities for the transitions between observation times, we use the Modified Diffusion Bridge proposal of Section 3.2.3 with $\Delta t = 0.1$- see Chapter 3 for more details regarding simulating conditioned diffusions. Specifically, recall, from Section 3.2.3, that, in two-dimensions, the MDB proposal of a discretised path of the diffusion between two observation times, $x_{1:K}^*$, say, takes the form

$$q_0^{\mathrm{MDB}}(x_{1:K}^* | y^*) = \prod_{k=1}^{K} \phi(x_k^*; a_{k-1}^{\mathrm{MDB}}, C_{k-1}^{\mathrm{MDB}}) \, ,$$

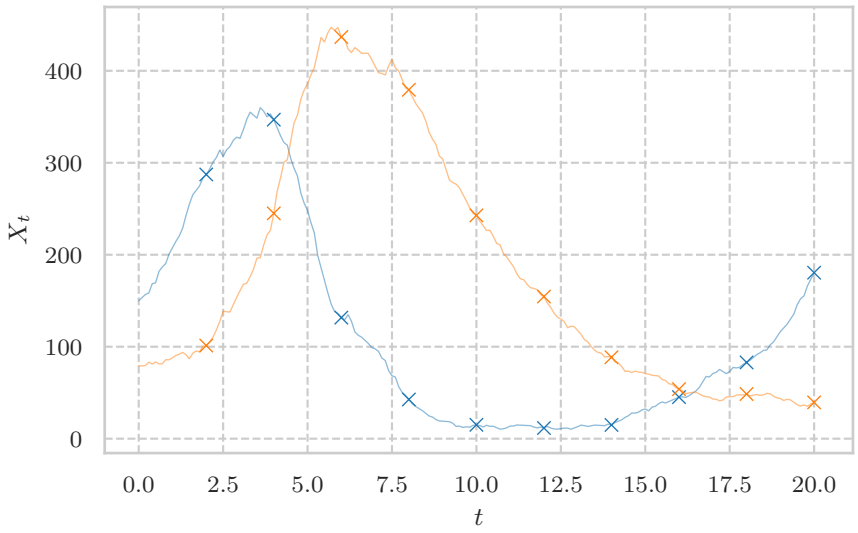

Figure 68: A plot of the observations $y_2, y_4, \ldots, y_{20}$ (orange crossed) and the path that generated them (blue solid lines) that will be used as the sequence of observations in the simulation study for the Lotka-Volterra diffusion.

where ϕ denotes the density corresponding to a two-dimensional normal distribution and a_{k-1}^{MDB} and C_{k-1}^{MDB} correspond to the mean (Equation (54)) and variance (Equation (55)) respectively, and, implicitly, depend on x_{k-1}^*, T^*, y^*, and t_{k-1}^*. Here; y^* denotes the observation which is being conditioned upon, T^* denotes the time corresponding to that observation, and t_{k-1}^* denotes the value of the path at the $(k-1)$-st inter-observation time. Such a proposal is equivalent to proposing $2K$ independent $\text{N}(0,1)$ random variables, $Z_{1:K}$, where each Z_k is formed of two independent $\text{N}(0,1)$ random variables; that is, $Z_k = (Z_k^{(1)}, Z_k^{(2)})$- one for each of the two dimensions— and transforming those random variables appropriately by sequentially setting, for $k \in \{1, \ldots, K\}$, $X_k = a_{k-1}^{\text{MDB}} + \sqrt{C_{k-1}^{\text{MDB}}} Z_k$. We ran the Exchangeable Particle Gibbs Sampler for each of the thirty-two combinations of

$$(\epsilon, N) \in \{0.01, 0.1, 0.15, 0.2, 0.3, 0.5, 0.75, \sqrt{2}\} \times \{10, 50, 100, 250\} ,$$

for one-hundred-thousand iterations. To garner an initial reference path, we ran the Sequential Monte Carlo procedure (Algorithm 8) with N particles and choose the path with the largest terminal weight, $\tilde{w}_T^{(k)}$. To generate a *good* representation of the *truth*, we ran the Particle Gibbs Sampler (Section 4.2.4) with one-thousand particles for one-million iterations using the residual-bridge construct of Whitaker et al., 2017, where $\xi_t = \mathbb{E}(\hat{R}_t | Y = y)$ and $\hat{R}_t$ is the process satisfying the diffusion of the Linear Noise Approxtimation; that is, satisfies the SDE 64, as the proposal for the transition between observations. Specifically, re-

call, from Section 3.2.5, that, in two-dimensions, the residual-bridge proposal of a discretised path of the diffusion, $x^*_{1:K}$, say, takes the form

$$q_0^{\mathrm{RB}}(x^*_{1:K}|y^*) = \prod_{k=1}^{K} \phi(x^*_k; a^{\mathrm{RB}}_{k-1}, D^{\mathrm{RB}}_{k-1}) \ ,$$

where ϕ denotes the density corresponding to a two-dimensional normal distribution and a^{RB}_{k-1} and D^{RB}_{k-1} correspond to the mean (Equation (61)) and variance matrix (Equation (62)) respectively, and, implicitly, depend on x^*_{k-1}, T^*, y^*, ξ_{k-1}, ξ_K, and t^*_{k-1}. We will compare the *performance* of the samplers on the element of the simulated paths corresponding to $t = 1$. As discussed when motivating the optimal scaling results of Section 4.4.1, maximizing the *performance* of the samplers for the section of the path corresponding to the first inter-observation time, will, in general, lessen the path degeneracy phenomena and, hence, optimize the rate of mixing across all components of the path. Moreover, given the samples are *pinned* at both the start and end of the inter-observation periods, it is reasonable to focus on samples at the middle of the *first* inter-observation period as these will exhibit the most variation and, therefore, combined with the path degeneracy phenomena, will be the hardest to represent. For each run of the sampler, we calculated the sample efficiency; that is, the sample squared jump distance divided by the number of particles, along with the acceptance rate, for the section of the path in the X-space corresponding to the first inter-observation time period. We also calculate the two-sample Kolmogorov-Smirnov statistic given by (101) between the element of the sample paths, simulated via the Exchangeable Particle Gibbs Sampler, corresponding to $t = 1$ and the same corresponding element for the *true* sample paths as a measure of how *well* the simulated samples represent the truth.

4.4.3 *Results*

Starting with the Linear Gaussian model; Figures 69 and 70 show, respectively; a plot of the sample efficiency for the first component of the sample paths in the X-space against the *jump-size*, ϵ, and against the acceptance rate, for each value of N. Figure 71 shows a plot of the two-sample Kolmogorov-Smirnov (KS) statistic, (101)— calculated for the first element of the samples simulated via the Exchangeable Particle Gibbs Sampler and the first element of the *true* samples— against the jump-size for each value of N. Figure 69 shows that the maximum sample efficiency is achieved when one takes $N = 250$ and the jump-size, ϵ, to be the maximal jump-size; that is $\epsilon = \sqrt{2}$, which corresponds to the Particle Gibbs Sampler. This optimal choice for N, in this case, lines up with the optimal scaling results of Section 4.4.1 which suggests taking $N = 5T/2$ to optimise efficiency. Simiarly, when $N = 1000$, the optimal efficiency is achieved for the maximimal jump-size. On the other hand, for $N \in \{50, 100\}$, the optimal efficiency occurs for values of the jump-size that are less than $\sqrt{2}$. Similarly, Figure 71 demonstrates that the

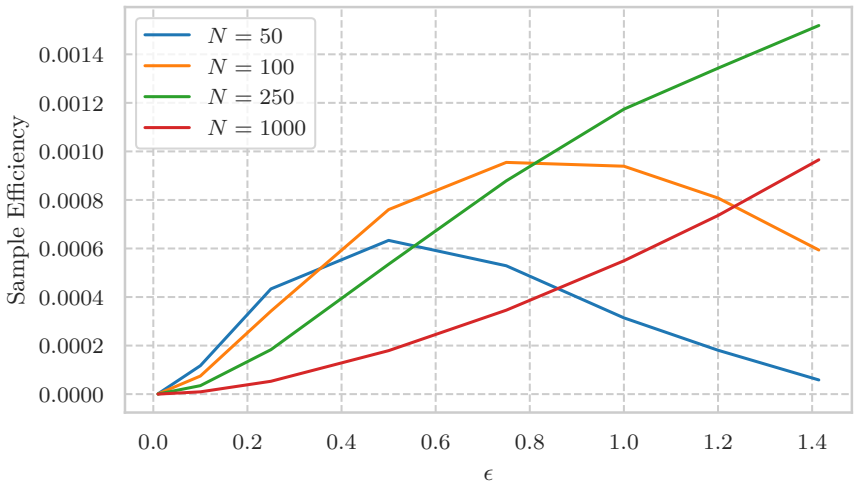

Figure 69: A plot of the sample efficiency for the first component of the sample paths in the X-space against the *jump-size*, ϵ, for each value of $N \in \{50, 100, 250, 1000\}$ of the Exchangeable Particle Gibbs Sampler applied to the Linear Gaussian model and using the bootstrap proposals as the marginal proposal densities. The sampler was run for one-hundred-thousand iterations.

two-sample KS statistic is minimised for; $\epsilon = \sqrt{2}$ when $N \in \{250, 1000\}$, and $\epsilon < \sqrt{2}$ when $N \in \{50, 100\}$. Clearly, this is because, the larger the value of N, the bigger jumps you can take while still maintaining the same level of acceptance rate, and, this increase in jump-size, results in better mixing of the chain and samples which are a better representation of the truth- at least when considering the first component of the sample paths marginally. Indeed, Figure 70 highlights that the optimal acceptance rate is larger the larger the value of N. All three figures suggest that the optimal choice of the jump-size, ϵ, in this case, depends on the choice on the number of particles N. Of course, to maximize efficiency, one would take $N = 250$ and take $\epsilon = \sqrt{2}$; that is, one would use the Particle Gibbs Sampler with $N = 250$.

Figure 72 shows histograms of the first component of the sample paths simulated by the Exchangeable Particle Gibbs Sampler for the optimal choice of N; that is, $N = 250$, and for a variety of jump-sizes,

$$\epsilon \in \{0.01, 0.1, 0.25, 0.5, 0.75, 1.0, 1.2, \sqrt{2}\} ,$$

where $\epsilon = 0.01$ for the top-left subplot, $\epsilon = 0.1$ for the top-right subplot, $\epsilon = 0.25$ for the subplot in the second row and the first column, and so on. With the same layout, Figure 73 shows, at each of the one-hundred-thousand iterations, the first component of the states of the Exchangeable Particle Gibbs Sampler for $N = 250$, and for a variety of jump-sizes,

$$\epsilon \in \{0.01, 0.1, 0.25, 0.5, 0.75, 1.0, 1.2, \sqrt{2}\} .$$

Figure 72 highlights that, for all but the smaller two jump-sizes; that is, for every $\epsilon \in \{0.25, 0.5, 0.75, 1.0, 1.2, \sqrt{2}\}$, the first components of

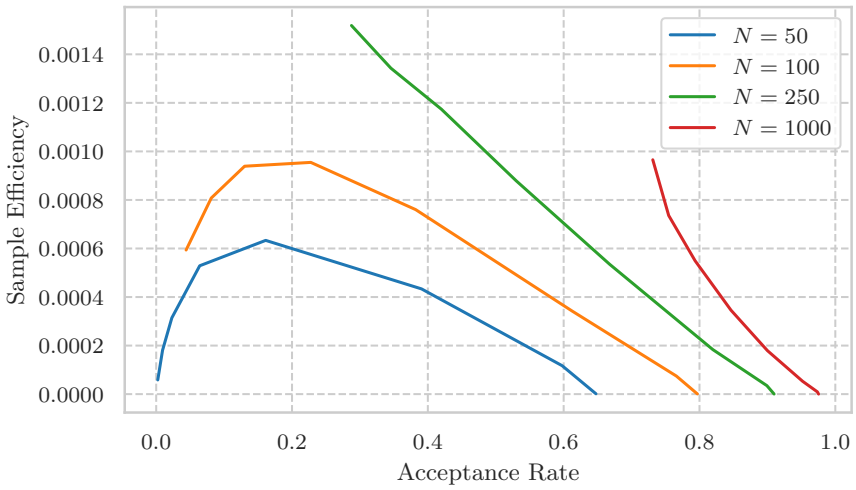

Figure 70: A plot of the sample efficiency for the first component of the sample paths in the X-space against the sample acceptance rate for the first component of the sample paths for each value of $N \in \{50, 100, 250, 1000\}$ of the Exchangeable Sampler Gibbs Sampler applied to the Linear Gaussian model and using the bootstrap proposals as the marginal proposal densities. The sampler was run for one-hundred-thousand iterations.

the simulated paths accurately represent the *truth*. Moreover, the figure also shows that, for $\epsilon = 0.1$, the first components of the simulated paths are a *good* representation of the *truth*; although, the mass in the center of the *true* distribution is slightly under-represented. The good mixing of the Exchangeable Particle Gibbs Sampler for these choices of the jump-size can be seen in Figure 73. It can also be seen, from this figure, that, when the jump-size is chosen to be very small; that is $\epsilon = 0.01$, the sampler exhibits random walk behaviour and the chain does not mix very well. This, in turn, leads to samples which do not correctly represent the *truth*- as can be seen in Figure 72. For $N = 250$, and any jump-size, the sampler does not exhibit any sticky behaviour, even as it goes out into the tails, as can be seen in Figure 73. This suggests that, for these values of N and ϵ, the Exchangeable Particle Gibbs Sampler is geometrically ergodic for this example[6]. Similar statements can be made for the case where $N = 1000$- as can be seen in Figures 120 and 123, which show, respectively; histograms of the first component of the sample paths simulated by the Exchangeable Particle Gibbs Sampler, and the evolution of the first component of the states for $N = 1000$ and for a variety of jump-sizes,

$$\epsilon \in \{0.01, 0.1, 0.25, 0.5, 0.75, 1.0, 1.2, \sqrt{2}\} .$$

Figures 118, 121, and Figures 119, 122, show the histograms and evolutions for the same set of jump-sizes, and for $N = 50$ and $N = 100$,

6 Of course, since the weights are bounded, the chain is actually uniformly ergodic by Corollary 2 of Lindsten, Douc, and Moulines, 2015, and Theorem 1 of Andrieu, Lee, and Vihola, 2018.

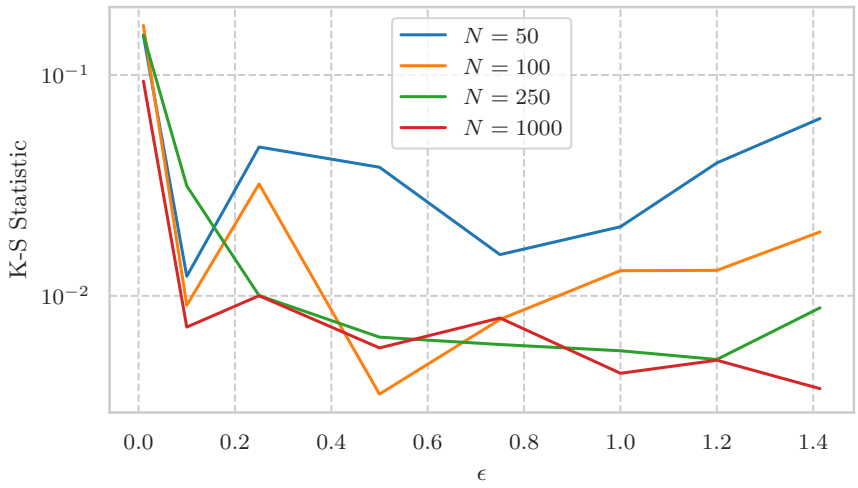

Figure 71: A plot of the two-sample Kolmogorov-Smirnov statistic, (101)- calculated for the first element of the samples simulated via the Exchangeable Particle Gibbs Sampler applied to the Linear Gaussian model and using the bootstrap proposals as the marginal proposal densities, and the first element of the *true* samples- against the jump-size for each value of $N \in \{50, 100, 250, 1000\}$. The sampler was run for one-hundred-thousand iterations.

respectively. It can be seen that, for these N- as suggested by Figures 69, 70, and 71- the largest choice of the jump-size; that is, $\epsilon = \sqrt{2}$, leads to a relatively *sticky* chain, which results in samples which represent the *truth* less well relative to smaller choices of the jump-size.

For the Lotka-Volterra diffusion model; Figures 74 and 75 show, respectively; a plot of the sample efficiency, for the section of the path in the X-space corresponding to the first inter-observation time period, against the *jump-size*, ϵ, and against the acceptance rate, for each value of N. Figure 76 shows a plot of the two-sample Kolmogorov-Smirnov (KS) statistic, (101)— calculated for the $t = 1$ element of the sample paths produced by the Exchangeable Particle Gibbs Sampler and the same element for the *true* sample paths— against the jump-size for each value of N. Figure 74 shows that the maximum sample efficiency is achieved when one takes $N = 10$ and the jump-size, ϵ, to be equal to 0.1. The figure also shows that, regardless of the value of N, the jump-size which maximises the sample efficiency is significantly less than $\sqrt{2}$. In fact, for any N, when the jump-size is chosen to be $\sqrt{2}$; that is, we use independent proposals, the sample efficiency is essentially zero. Thus, in this case, the Particle Gibbs Sampler exhibits extremely poor mixing. Simiarly, Figure 76 demonstrates that the two-sample KS statistic is minimised for values of the jump-size that are significantly less than $\sqrt{2}$, and, when $\epsilon = \sqrt{2}$, the statistic is very close to its maximal possible value, one. Therefore, in this case, the Particle Gibbs Sampler produces samples which are a very poor representation of the truth. Moreover, this figure shows that, although choosing $N = 10$ and $\epsilon = 0.1$ maximises the efficiency of the sampler-

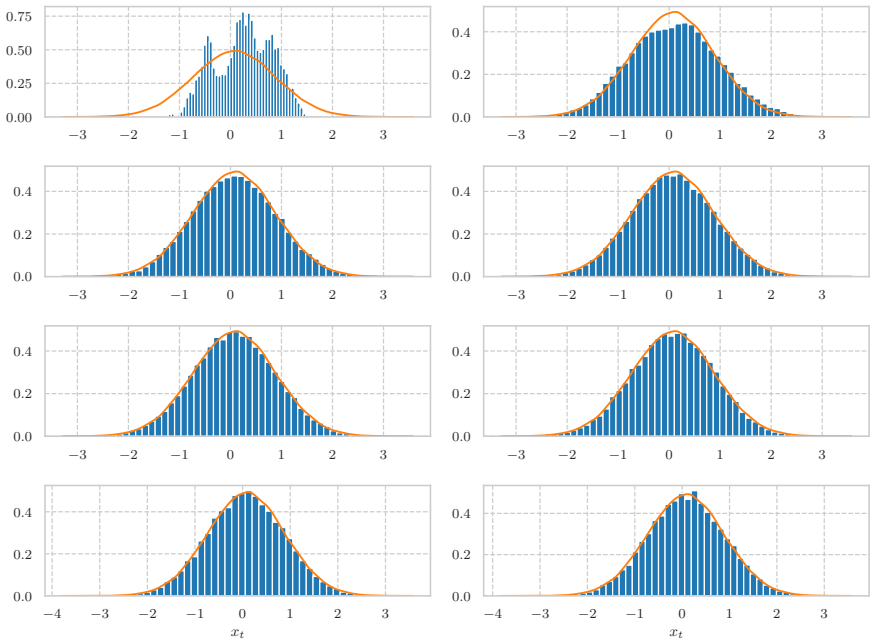

Figure 72: Histograms of the first component of the sample paths simulated by the Exchangeable Particle Gibbs Sampler- applied to the Linear Gaussian model and using the bootstrap proposals as the marginal proposal densities- with $N = 250$, and for a variety of jump-sizes, $\epsilon \in \{0.01, 0.1, 0.25, 0.5, 0.75, 1.0, 1.2, \sqrt{2}\}$, where $\epsilon = 0.01$ for the top-left subplot, $\epsilon = 0.1$ for the top-right subplot, $\epsilon = 0.25$ for the subplot in the second row and the first column, and so on. The orange line in each figure corresponds to the *true* density.

in terms of the expected squared jump distance per particle— such a choice does not lead to samples which represent the *truth*— at least in terms of the element of the path at $t = 1$ viewed marginally— the best. Indeed, for $N = 100$ or $N = 250$, and ϵ around 0.2, the two-sample KS statistic is almost an order of magnitude smaller than it is when $N = 10$ and $\epsilon = 0.1$. Of course, this is not a fair comparison, since running the sampler with $N = 100$ is about an order of magnitude slower than running the sampler with $N = 10$, so the chain, in the latter case, can be run for longer and, therefore, potentially produce samples which better represent the truth for a fixed computational cost. As has been the case for all the examples considered in this thesis, Figure 75 highlights that the optimal acceptance rate is larger the larger the value of N. All three figures suggest that, for this example, the Particle Gibbs Sampler performs very poorly.

Figure 77 shows histograms of the $t = 1$ element of the sample paths simulated via the Exchangeable Particle Gibbs Sampler for the optimal choice of N; that is, $N = 10$, and for a variety of jump-sizes,

$$\epsilon \in \{0.01, 0.1, 0.15, 0.2, 0.3, 0.5, 0.75, \sqrt{2}\} ,$$

where $\epsilon = 0.01$ for the top-left subplot, $\epsilon = 0.1$ for the top-right subplot, $\epsilon = 0.15$ for the subplot in the second row and the first column, and so

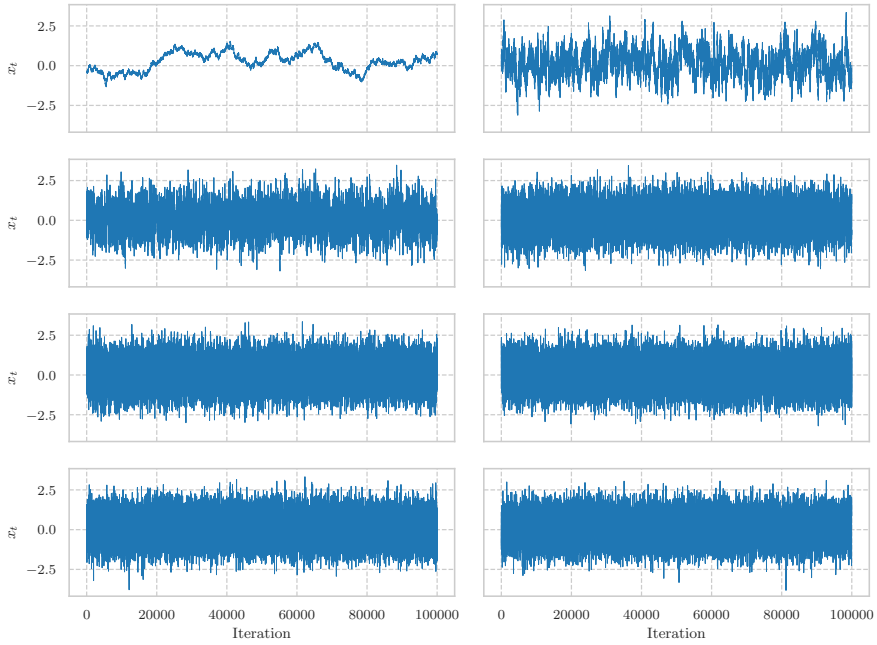

Figure 73: Plots of the first component of the states of the the Exchangeable Particle Gibbs Sampler- applied to the Linear Gaussian model and using the bootstrap proposals as the marginal proposal densities- with $N = 250$, and for a variety of jump-sizes, $\epsilon \in \{0.01, 0.1, 0.25, 0.5, 0.75, 1.0, 1.2, \sqrt{2}\}$, where $\epsilon = 0.01$ for the top-left subplot, $\epsilon = 0.1$ for the top-right subplot, $\epsilon = 0.25$ for the subplot in the second row and the first column, and so on.

on. With the same layout, Figure 78 shows, at each of the one-hundred-thousand iterations, the $t = 1$ element of the states of the Exchangeable Particle Gibbs Sampler for $N = 10$, and for a variety of jump-sizes,

$$\epsilon \in \{0.01, 0.1, 0.15, 0.2, 0.3, 0.5, 0.75, \sqrt{2}\} .$$

Figure 77 highlights that, for the smallest jump-size and for the four largest jump-sizes; that is, for every $\epsilon \in \{0.01, 0.3, 0.5, 0.75, \sqrt{2}\}$, the $t = 1$ elements of the simulated paths are a poor representation of the *truth*. Indeed, for $\epsilon \in \{0.5, 0.75, \sqrt{2}\}$, the samples do not even form a visible density. Figure 78 shows that this is because, for $\epsilon \in \{0.5, 0.75, \sqrt{2}\}$, the chain only moves a few times in the one-hundred-thousand iterations. Moreover, for $\epsilon = 0.3$, even though the chain does move, it is relatively sticky and, therefore, does not move much. For the smallest value of the jump-size; that is, $\epsilon = 0.01$, the chain exhibits random-walk behaviour. On the other hand, for $\epsilon \in \{0.1, 0.25, 0.2\}$, the chain mixes well and this leads to samples which are an okay representation of the *truth*. These figures would hence suggest that, in this case, the Exchangeable Particle Gibbs Sampler is potentially geometrically ergodic, particularly for smaller values of the jump-size. As Figure 71 suggests, for $\epsilon < \sqrt{2}$, increasing N leads to samplers whose chains are significantly less *sticky* and, therefore, whose samples are a better representation of the truth. This can be seen in Figures 124, 125, and

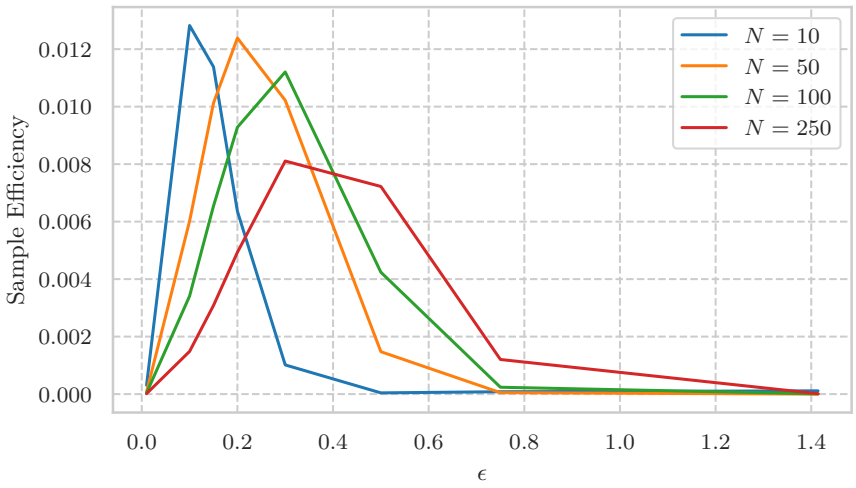

Figure 74: A plot of the sample efficiency, for the section of the path in the X-space corresponding to the first inter-observation time period, against the *jump-size*, ϵ, for each value of $N \in \{10, 50, 100, 250\}$, of the Exchangeable Particle Gibbs Sampler applied to the Lotka-Volterra diffusion model and using the Modified Diffusion Bridge proposal of Section 3.2.3 as the marginal proposal distribution. The sampler was run for one-hundred-thousand iterations.

126, which show, for $N = 50$, $N = 100$, and $N = 250$, respectively, histograms of the $t = 1$ element of the sample paths simulated via the Exchangeable Particle Gibbs Sampler for a variety of jump-sizes,

$$\epsilon \in \{0.01, 0.1, 0.15, 0.2, 0.3, 0.5, 0.75, \sqrt{2}\} ,$$

as well as Figures 127, 128, and 129, which show, for $N = 50$, $N = 100$, and $N = 250$, respectively, the $t = 1$ element of the states of the Exchangeable Particle Gibbs Sampler for a variety of jump-sizes,

$$\epsilon \in \{0.01, 0.1, 0.15, 0.2, 0.3, 0.5, 0.75, \sqrt{2}\} .$$

All these figures suggest that, for any N and $\epsilon < \sqrt{2}$, the Exchangeable Particle Gibbs Sampler is potentially geometrically ergodic. Note that, for any value of N considered, the chain corresponding to the Exchangeable Particle Gibbs Sampler with $\epsilon = \sqrt{2}$; that is, the chain corresponding to the Particle Gibbs Sampler, is extremely sticky, moving only a few times in one-hundred-thousand iterations. This suggests that, in this case, regardless of the value of N, the Particle Gibbs Sampler is not geometrically ergodic.

4.5 SUMMARY

In this Chapter we introduced two new classes of Markov Chain Monte Carlo samplers, named the Exchangeable Sampler and the Exchangeable Particle Gibbs Sampler. At each iteration, the Exchangeable Sampler use exchangeablility to simulate multiple, weighted proposals whose

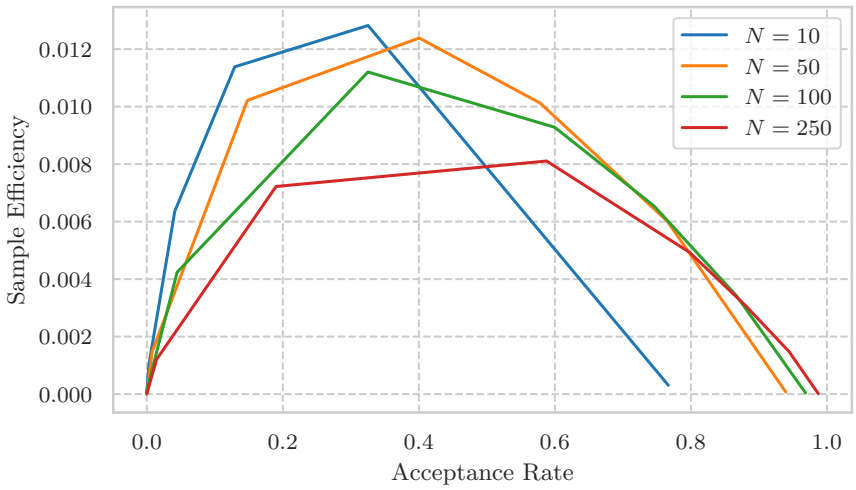

Figure 75: A plot of the sample efficiency, for the section of the path in the X-space corresponding to the first inter-observation time period, sample acceptance rate for each value of $N \in \{10, 50, 100, 250\}$, of the Exchangeable Particle Gibbs Sampler applied to the Lotka-Volterra diffusion model and using the Modified Diffusion Bridge proposal of Section 3.2.3 as the marginal proposal distribution. The sampler was run for one-hundred-thousand iterations.

weights indicate how likely the chain is to move to such a proposal, thereby generalising the multiple-proposal Independence Sampler. The Exchangeable Particle Gibbs Sampler generalises the Particle Gibbs Sampler by allowing the particles in the Conditional Sequential Monte Carlo procedure to be simulated exchangeably. By generalising the Independence Sampler and the Particle Gibbs Sampler respectively, these new samplers allow for the locality of moves to be controlled by a *scaling* parameter, which can be tuned to optimise the mixing of the resulting MCMC procedure, while still benefiting from the increase in acceptance probability that typically comes with using multiple proposals. These samplers can lead to chains with better mixing properties, and, therefore, to MCMC estimators with smaller variances than their corresponding algorithms based on independent proposals. We showed, numerically, in Section 4.3.2 for the Exchangeable Sampler, and Section 4.4.2 for the Exchangeable Particle Gibbs Sampler, how the introduction of a *tunable jump-size* can lead to significantly more efficient samplers compared to their *independent* counterparts. Of particular relevance to this thesis, we showed that both the Exchangeable Sampler and the Exchangeable Particle Gibbs Sampler were significantly more efficient than their independent counterparts when conducting inference for diffusions using the Modified Diffusion Bridge proposal of Section 3.2.3. We also provided, via Theorem 4.3.9, sufficient conditions under which the Exchangeable Sampler is geometrically ergodic. In particular, we showed that the Exchangeable Sampler can be geometrically ergodic even when the *importance* weights are unbounded and, hence, in scenarios where the Independence Sampler cannot be geometrically

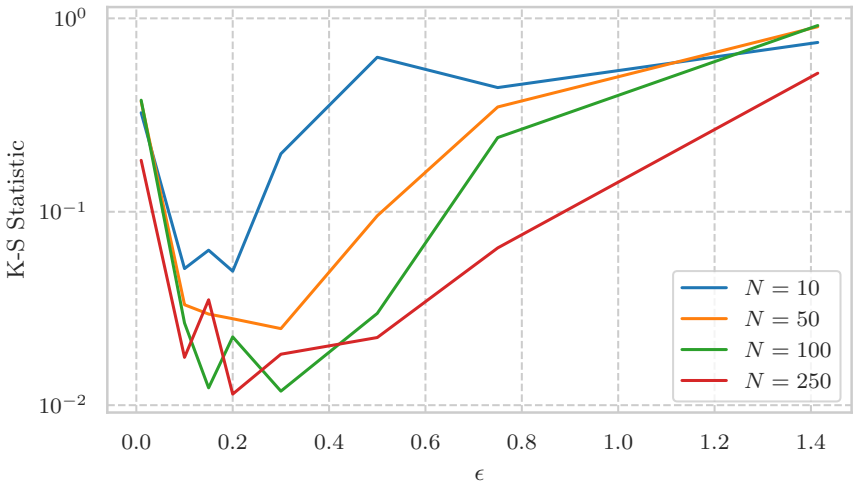

Figure 76: A plot of the two-sample Kolmogorov-Smirnov statistic, (101)- calculated for the $t = 1$ element of the sample paths simulated via the Exchangeable Particle Gibbs Sampler applied to the Lotka-Volterra diffusion model and using the Modified Diffusion Bridge proposal of Section 3.2.3 as the marginal proposal distribution, and the same element for the *true* sample paths- against the jump-size for each value of $N \in \{10, 50, 100, 250\}$. The sampler was run for one-hundred-thousand iterations.

ergodic. To provide guidance in the practical implementation of the Exchangeable Sampler and the Exchangeable Particle Gibbs Sampler, we derived, in Sections 4.3.1 and 4.4.1, asymptotic expected squared-jump distance results, and demonstrated, numerically, how the theory plays out in practice when d is finite.

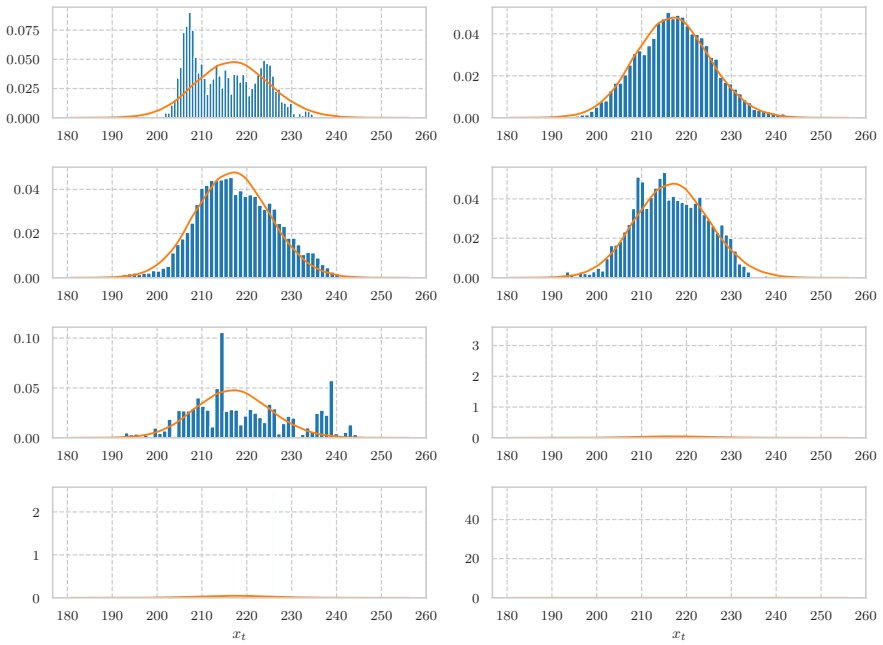

Figure 77: Histograms of the $t = 1$ element of the sample paths, simulated via the Exchangeable Particle Gibbs Sampler applied to the Lotka-Volterra diffusion model and using the Modified Diffusion Bridge proposal of Section 3.2.3 as the marginal proposal distribution, for $N = 10$, and for a variety of jump-sizes, $\epsilon \in \{0.01, 0.1, 0.15, 0.2, 0.3, 0.5, 0.75, \sqrt{2}\}$, where $\epsilon = 0.01$ for the top-left subplot, $\epsilon = 0.1$ for the top-right subplot, $\epsilon = 0.15$ for the subplot in the second row and the first column, and so on. The orange line in each figure corresponds to the *true* density.

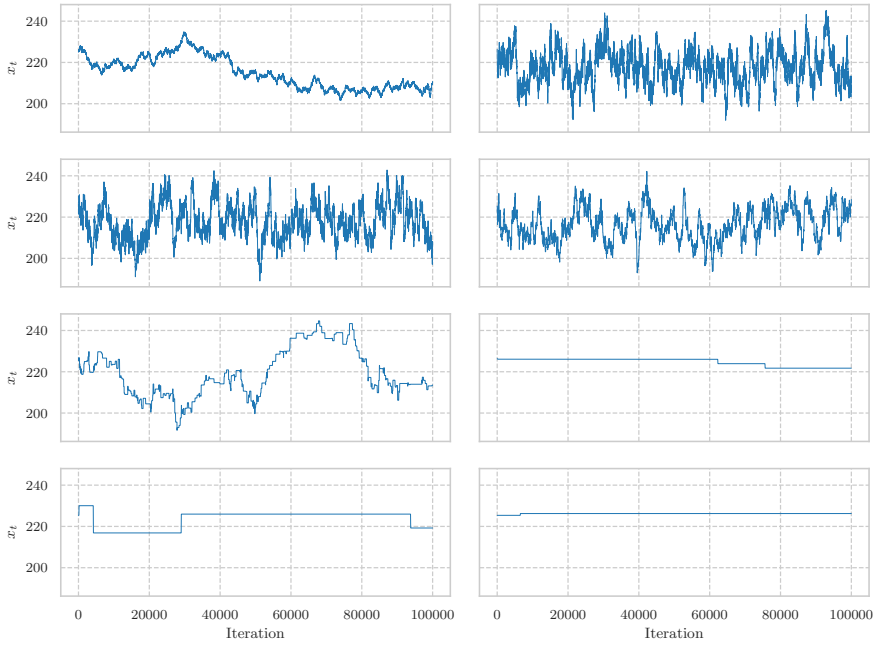

Figure 78: Plots of the $t = 1$ element of the states of the the Exchangeable Particle Gibbs Sampler- applied to the Linear Gaussian model and using the bootstrap proposals as the marginal proposal densities- for $N = 10$, and for a variety of jump-sizes, $\epsilon \in \{0.01, 0.1, 0.15, 0.2, 0.3, 0.5, 0.75, \sqrt{2}\}$, where $\epsilon = 0.01$ for the top-left subplot, $\epsilon = 0.1$ for the top-right subplot, $\epsilon = 0.15$ for the subplot in the second row and the first column, and so on.

5

CONCLUSION AND FURTHER WORK

In this thesis we built upon two strands of recent research related to conducting Bayesian inference for stochastic processes.

In our first contribution, we introduced a new residual-bridge proposal for approximately simulating conditioned diffusions formed by applying the modified diffusion bridge approximation of Durham and Gallant, 2002 to the difference between the true diffusion and a second, approximate, diffusion driven by the same Brownian motion. This new proposal attempts to account for volatilities which are not constant and can, therefore, lead to gains in efficiency over recently proposed residual-bridge constructs (Whitaker et al., 2017) in situations where the volatility varies considerably, as is often the case for larger inter-observation times and for time-inhomogeneous volatilities. We showed, in Section 3.3.2, via a simulation study, how, for larger inter-observation times, this new proposal led to larger- sometimes one to two orders of magnitude larger- relative effective sample sizes per second compared to the residual-bridge constructs of Whitaker et al., 2017, for both the Lotka-Volterra diffusion (3.1.2) and a simple diffusion for gene expression (3.1.3). We highlighted that a drawback of the new proposal is that, at inter-observation time points, discrepancies of sample paths of the conditional diffusion from the deterministic path, around which the residual-bridge is centered, can be relatively large. We demonstrated how, for the Birth-Death diffusion, these discrepancies become evident as neither the approximating deterministic path produced by the ODE or the LNA captures the true dynamics of the diffusion as the diffusion approaches the x-axis- a reflecting boundary of the diffusion. Indeed, we showed that, for the Birth-Death diffusion, these discrepancies led to lower relative effective sample sizes per second compared to the residual-bridge constructs of Whitaker et al., 2017. This is not necessarily a drawback of the residual-bridge construct itself, per se, but a drawback which stems from the deterministic path upon which the residual-bridge is constructed. Of course, using the determinstic path to approximate the volatility as well as the drift amplifies the problem and makes the bridge less robust compared to the constructs of Whitaker et al., 2017.

One natural direction for further work involves tackling the drawback of the new residual-bridge construct. In particular, developing a bridge which attempts to account for volatilities which are not constant, but which also is robust against situations where the approximating path fails to capture the true dynamics of the diffusion. As noted in this thesis, we struggled to find a *justifiable* and *computationally efficient* interpolation scheme which, instead of preserving the discrepancies in the drift and volatility over the inter-observation time, attempted to ensure these were 0 at time T. However, one avenue that was not investigated

fully involves trying to use a sample path as the path around which the new residual-bridge construct is based. Indeed, one could imagine simulating a sample path, x_t, either by the constructs of Whitaker et al., 2017 or by the constructs introduced in this thesis, and then using this sample path, that is, setting $\xi_t = x_t$, to construct the residual;

$$\mathrm{d}\tilde{R}_t = (\mu(X_t, t) - \xi_t')\mathrm{d}t + (\sigma(X_t, t) - \sigma(\xi_t, t))\mathrm{d}B_t \ .$$

By using a sample path which aims to better represent the true dynamics of the conditioned diffusion, as the path around which the residual-bridge is constructed, one can, potentially, make the novel residual-bridge construct introduced in this thesis more efficient statistically, and more robust. Of course, the performance of such a procedure will strongly depend on how well the initial sample path represents a sample from the *true* conditioned diffusion; that is, if the sample path constructed initially is not a *good* representation of the *true* conditioned diffusion, then, the discrepancies of the *true* sample paths of the conditioned diffusion from the sample path upon which the residual is based, will potentially be large. Therefore, such a proposal will potentially have the same drawbacks as the residual-bridge constructs introduced in this thesis. Moreover, having to simulate two paths to get one sample path essentially means doubling the cost per sample. Therefore, such an approach would need to significantly improve the effective sample size in order to achieve a competitive level of efficiency. One is almost always interested in simulating many of these paths as part of another procedure; like a particle filter, for instance. Therefore, one can potentially tackle both these problems by first *pre-simulating* a smaller number of sample paths and then using these sample paths- by taking their weighted mean, say- to construct a path, ξ_t, upon which to base the residual bridge. By pre-simulating, for instance, 10% of the number of sample paths one plans to simulate, one would only increase the computational cost by, at most, 10%- and, potentially, may even lead to a smaller computational cost since one would not have solve as many ODEs. Furthermore, by using a selection of conditioned sample paths, one can potentially construct a path, upon which to build the residual bridge, that accurately captures the *true* dynamics of the conditioned diffusion. Of course, where the initial starting points for the sample paths are different, such as in a particle filter where the observations are noisy, for instance, one would have to be careful about how the pre-simulated paths are constructed and combined.

In our second contribution, we introduced two new classes of Markov Chain Monte Carlo samplers, named the Exchangeable Sampler and the Exchangeable Particle Gibbs Sampler, which, at each iteration, use exchangeablility to simulate multiple, weighted proposals whose weights indicate how likely the chain is to move to such a proposal. By generalising the Independence Sampler and the Particle Gibbs Sampler respectively, these new samplers allow for the locality of moves to be controlled by a *scaling* parameter which can be tuned to optimise the mixing of the resulting MCMC procedure, while still benefiting from the increase in acceptance probability that typically comes with using mul-

tiple proposals. These samplers can lead to chains with better mixing properties, and, therefore, to MCMC estimators with smaller variances than their corresponding algorithms based on independent proposals. We showed, numerically, in Section 4.3.2 for the Exchangeable Sampler, and Section 4.4.2 for the Exchangeable Particle Gibbs Sampler, how the introduction of a *tunable jump-size* can lead to significantly more efficient samplers compared to their *independent* counterparts. In particular, in the $T = 1$ scenario, when the transition weight was unbounded, and, therefore, when the Independence Sampler could not be geometrically ergodic, the Exchangeable Sampler had better performance for values of $\epsilon < \sqrt{2}$; that is, when the samples were not simulated indepedently from one another. Moreover, both the Exchangeable Sampler and the Exchangeable Particle Gibbs Sampler were significantly more efficient than their independent counterparts when conducting inference for diffusions using the Modified Diffusion Bridge proposal of Section 3.2.3. We also provided, via Theorem 4.3.9, sufficient conditions under which the Exchangeable Sampler is geometrically ergodic. In particular, we showed that the Exchangeable Sampler can be geometrically ergodic even when the *importance* weights are unbounded and, hence, in scenarios where the Independence Sampler cannot be geometrically ergodic. We gave numerical support to this theorem by investigating the assumptions for three examples; one where the transition weight was exponentially increasing in the tails, one where the transition weight was polynomially increasing in the tails, and one where the transition weight was bounded, and seeing how the sampler performed in practice in the simulation study of Section 4.3.2. To provide guidance in the practical implementation of the Exchangeable Sampler and the Exchangeable Particle Gibbs Sampler, we derived, in Sections 4.3.1 and 4.4.1, asymptotic expected squared-jump distance results, and demonstrated, numerically, how the theory plays out in practice when d is finite.

There are several directions for further work relating to the Exchangeable Sampler and the Exchangeable Particle Gibbs Sampler. The first involves investigating further the ergodicity properties of the Exchangeable Particle Gibbs Sampler. Recall, from Theorem 2.3.31, that the Independence Sampler is uniformly ergodic if and only if the importance weight is bounded. Moreover, Corollary 2 of Lindsten, Douc, and Moulines, 2015, and Theorem 1 of Andrieu, Lee, and Vihola, 2018, demonstrate that the Particle Gibbs Sampler, which can viewed as an extension of the Independence Sampler, is uniformly ergodic if the transition weights, at each time step, are uniformly bounded. Finally, Theorem 4.3.9 provides sufficient conditions under which the Exchangeable Sampler, which, itself, is a generalisation of the Independence Sampler, is geometrically ergodic. In particular, the Exchangeable Sampler can be geometrically ergodic even when the *importance* weights are unbounded. Therefore, it is natural to conjecture that, under certain conditions, the Exchangeable Particle Gibbs Sampler, which is an extension of the Exchangeable Sampler, and a generalisation of the Particle Gibbs Sampler, can be geometrically ergodic even if the tran-

sition weights are unbounded, provided the weights do not increase too quickly in the *tails*. Investigating this conjecture is an important direction for future work.

The second direction concerns Remark 14. Recall that the Exchangeable Particle Gibbs Sampler relies on the Conditional Exchangeable Sequential Monte Carlo (CxSMC) procedure, which is given by Algorithm 20. The CxSMC procedure considered in this thesis utilises the proposal given by Algorithm 21, which allows for the use of a different jump-size, ϵ_t, at each time $t \in \{0, \ldots, T\}$, to propagate the particles forward. In this thesis, we have restricted ourselves to letting the jump-size be static; that is, $\epsilon_t = \epsilon$ for all $t \in \{0, \ldots, T\}$. However, in general, because there are fewer future resampling steps the further in time, t, the procedure is, it would be prudent to use a smaller jump-size the smaller the value of t and a bigger jump-size the larger the value of t. Intuitively, one would expect that the optimal function of ϵ_t, with respect to time t, is some concave curve between 1 and T, where $\epsilon_T = 1$. This intuition is partially confirmed by the proof, given by A.25, of the optimal scaling result of Theorem 4.4.7. Investigating this idea is another clear direction for future work.

The third direction stems from Remark 8. Tjelmeland, 2004 provides a general framework for multiple-proposal samplers. In the framework Tjelmeland, 2004 presents, one does not need a marginal distribution to implement the algorithm. Moreover, the two proposals described in Tjelmeland, 2004 are such that the resulting joint proposal satisfies a certain symmetry, much like the random-walk sampler. As such, for both those proposals, the transition weight, much like the transition weight for the random-walk sampler, is simply the target density up to a constant of proportionality. In the case where the target and the marginal are both Normal distributions, the first proposal presented in Tjelmeland, 2004 is very similar to the proposal we introduce in Algorithm 18. The second proposal, on the other hand, mimics the random-walk proposal whilst keeping the proposals equidistant from one another. Much like the Exchangeable Sampler introduced in this thesis, it is this symmetry that simplifies the transition weight and leads to an efficient sampler. The approaches described in Tjelmeland, 2004 are a natural alternative to the Exchangeable Sampler introduced in this thesis. Like the Exchangeable Sampler, they can also be extended to a $T > 1$ setting. Therefore, another interesting piece of work would involve comparing the two approaches in both the $T = 1$ and $T > 1$ settings and extending the results derived in this thesis to the methods described in Tjelmeland, 2004.

Agapiou, Sergios, Gareth O. Roberts, and Sebastian J. Vollmer (Aug. 2018). "Unbiased Monte Carlo: Posterior estimation for intractable/infinite-dimensional models". In: *Bernoulli* 24.3, pp. 1726–1786. DOI: 10.3150/16-BEJ911. URL: https://doi.org/10.3150/16-BEJ911.

Aït-Sahalia, Yacine and Robert Kimmel (2007). "Maximum likelihood estimation for stochastic volatility models". In: *Journal of Financial Economics* 83.413.

Akeret, J., A. Refregier, A. Amara, S. Seehars, and C. Hasner (Aug. 2015). "Approximate Bayesian computation for forward modeling in cosmology". In: *JCAP* 8, 043, p. 043. DOI: 10.1088/1475-7516/2015/08/043.

Ala-Luhtala, Juha, Nick Whiteley, Kari Heine, and Robert Piché (2016). "An Introduction to Twisted Particle Filters and Parameter Estimation in Non-Linear State-Space Models". In: *IEEE Transactions on Signal Processing* 64, pp. 4875–4890.

Andrieu, Christophe, Arnaud Doucet, and Roman Holenstein (2010). "Particle Markov chain Monte Carlo methods". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 72.3, pp. 269–342. ISSN: 1467-9868. DOI: 10.1111/j.1467-9868.2009.00736.x.

Andrieu, Christophe, Anthony Lee, and Matti Vihola (May 2018). "Uniform ergodicity of the iterated conditional SMC and geometric ergodicity of particle Gibbs samplers". In: *Bernoulli* 24.2, pp. 842–872. DOI: 10.3150/15-BEJ785. URL: https://doi.org/10.3150/15-BEJ785.

Andrieu, Christophe and Gareth O. Roberts (Apr. 2009). "The pseudo-marginal approach for efficient Monte Carlo computations". In: *Ann. Statist.* 37.2, pp. 697–725. DOI: 10.1214/07-AOS574. URL: https://doi.org/10.1214/07-AOS574.

Andrieu, Christophe, Nando de Freitas, Arnaud Doucet, and Michael I. Jordan (2003). "An Introduction to MCMC for Machine Learning". In: *Machine Learning* 50.1, pp. 5–43. ISSN: 1573-0565. DOI: 10.1023/A:1020281327116.

Ash, R.B. and C. Doléans-Dade (2000). *Probability and Measure Theory*. Harcourt/Academic Press. ISBN: 9780120652020.

Atchadé, Yves F. and François Perron (2007). "On the geometric ergodicity of Metropolis-Hastings algorithms". In: *Statistics* 41.1, pp. 77–84. DOI: 10.1080/10485250601033214. eprint: http://dx.doi.org/10.1080/10485250601033214.

Ball, Frank and Peter Neal (2008). "Network epidemic models with two levels of mixing". In: *Mathematical Biosciences* 212.1, pp. 69 –87. ISSN: 0025-5564. DOI: https://doi.org/10.1016/j.mbs.2008.01.001.

Barndorff-Nielsen, Ole E. (1997). "Normal Inverse Gaussian Distributions and Stochastic Volatility Modelling". In: *Scandinavian Journal of Statistics* 24.1, pp. 1–13. ISSN: 1467-9469. DOI: 10.1111/1467-9469.00045.

Baxendale, Peter H. (Feb. 2005). "Renewal theory and computable convergence rates for geometrically ergodic Markov chains". In: *Ann. Appl. Probab.* 15.1B, pp. 700–738. DOI: 10.1214/105051604000000710. URL: https://doi.org/10.1214/105051604000000710.

Beaumont, Mark A. (2003). "Estimation of Population Growth or Decline in Genetically Monitored Populations". In: *Genetics* 164.3, pp. 1139–1160. ISSN: 0016-6731. eprint: http://www.genetics.org/content/164/3/1139.full.pdf.

Berntsen, Jarle, Terje O. Espelid, and Alan Genz (Dec. 1991). "An Adaptive Algorithm for the Approximate Calculation of Multiple Integrals". In: *ACM Trans. Math. Softw.* 17.4, pp. 437–451. ISSN: 0098-3500. DOI: 10.1145/210232.210233.

Beskos, Alexandros, Omiros Papaspiliopoulos, and Gareth O. Roberts (Dec. 2006). "Retrospective exact simulation of diffusion sample paths with applications". In: *Bernoulli* 12.6, pp. 1077–1098. DOI: 10.3150/bj/1165269151.

Billingsley, P. (1995). *Probability and Measure*. Wiley Series in Probability and Statistics. Wiley. ISBN: 9780471007104.

Bizjajeva, Svetlana and Jimmy Olsson (Oct. 2016). "Antithetic sampling for sequential Monte Carlo methods with application to state-space models". English. In:

Annals of the Institute of Statistical Mathematics 68.5, pp. 1025–1053. ISSN: 0020-3157. DOI: 10.1007/s10463-015-0524-y.

Boys, R. J., D. J. Wilkinson, and T. B. L. Kirkwood (2008). "Bayesian inference for a discretely observed stochastic kinetic model". In: *Statistics and Computing* 18.2, pp. 125–135. ISSN: 1573-1375. DOI: 10.1007/s11222-007-9043-x.

Brasnett, PA, LS Mihaylova, CN Canagarajah, and DR Bull (Jan. 2005). "Particle filtering with multiple cues for object tracking in video sequences". In: *IS & T/SPIE 17th Annual Symposium Image and Video Communications Processing 2005, San Jose, CA, USA*. Vol. 5685. SPIE–The International Society for Optical Engineering, pp. 430 –441. DOI: 10.1117/12.585882.

Breyer, L.A. and G.O. Roberts (2000). "From metropolis to diffusions: Gibbs states and optimal scaling". In: *Stochastic Processes and their Applications* 90.2, pp. 181 –206. ISSN: 0304-4149. DOI: https://doi.org/10.1016/S0304-4149(00)00041-7. URL: http://www.sciencedirect.com/science/article/pii/S0304414900000417.

Brooks, S., A. Gelman, G. Jones, and X.L. Meng (2011). *Handbook of Markov Chain Monte Carlo*. Chapman & Hall/CRC Handbooks of Modern Statistical Methods. CRC Press. ISBN: 9781420079425.

Cancès, Eric, Frédéric Legoll, and Gabriel Stoltz (2007). "Theoretical and numerical comparison of some sampling methods for molecular dynamics". In: *ESAIM: M2AN* 41.2, pp. 351–389. DOI: 10.1051/m2an:2007014.

Capinski, M. and E. Kopp (2013). *Measure, Integral and Probability*. Springer Undergraduate Mathematics Series. Springer London. ISBN: 9781447136316.

Cappé, O., E. Moulines, and T. Ryden (2006). *Inference in Hidden Markov Models*. Springer Series in Statistics. Springer New York. ISBN: 9780387289823.

Charniak, E. (1996). *Statistical Language Learning*. A Bradford book. A Bradford Book. ISBN: 9780262531412.

Chopin, Nicolas (Dec. 2004). "Central limit theorem for sequential Monte Carlo methods and its application to Bayesian inference". In: *Ann. Statist.* 32.6, pp. 2385–2411. DOI: 10.1214/009053604000000698.

Chopin, Nicolas and Sumeetpal S. Singh (Aug. 2015). "On particle Gibbs sampling". In: *Bernoulli* 21.3, pp. 1855–1883. DOI: 10.3150/14-BEJ629.

Coffey, W., Y.P. Kalmykov, and J.T. Waldron (2004). *The Langevin Equation: With Applications to Stochastic Problems in Physics, Chemistry, and Electrical Engineering*. Series in contemporary chemical physics. World Scientific. ISBN: 9789812384621.

Conway, J.B. (1994). *A Course in Functional Analysis*. Graduate Texts in Mathematics. Springer New York. ISBN: 9780387972459.

Cotter, S. L., G. O. Roberts, A. M. Stuart, and D. White (2013). "MCMC Methods for Functions: Modifying Old Algorithms to Make Them Faster". In: *Statistical Science* 28.3, pp. 424–446. ISSN: 08834237, 21688745. URL: http://www.jstor.org/stable/43288425.

Dahlin, J., F. Lindsten, J. Kronander, and T. B. Schön (Nov. 2015). "Accelerating pseudo-marginal Metropolis-Hastings by correlating auxiliary variables". In: *ArXiv e-prints*. arXiv: 1511.05483 [stat.CO].

Del Moral, P. (2012). *Feynman-Kac Formulae: Genealogical and Interacting Particle Systems with Applications*. Probability and Its Applications. Springer New York. ISBN: 9781468493931.

Del Moral, P. and A. Guionnet (May 1999). "Central limit theorem for nonlinear filtering and interacting particle systems". In: *Ann. Appl. Probab.* 9.2, pp. 275–297. DOI: 10.1214/aoap/1029962742.

Del Moral, P. and L. Miclo (2000). "Branching and interacting particle systems approximations of feynman-kac formulae with applications to non-linear filtering". In: *Séminaire de Probabilités XXXIV*. Ed. by Jacques Azéma, Michel Ledoux, Michel Émery, and Marc Yor. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 1–145. ISBN: 978-3-540-46413-6. DOI: 10.1007/BFb0103798.

Del Moral, Pierre, Arnaud Doucet, and Ajay Jasra (Feb. 2012). "On adaptive resampling strategies for sequential Monte Carlo methods". In: *Bernoulli* 18.1, pp. 252–278. DOI: 10.3150/10-BEJ335. URL: https://doi.org/10.3150/10-BEJ335.

Del Moral, Pierre and Lawrence Murray (2015). "Sequential Monte Carlo with Highly Informative Observations". In: *SIAM/ASA Journal on Uncertainty Quantification* 3.1, pp. 969–997. DOI: 10.1137/15M1011214. eprint: https://doi.org/10.1137/15M1011214.

Deligiannidis, G., A. Doucet, and M. K. Pitt (Nov. 2015). "The Correlated Pseudo-Marginal Method". In: *ArXiv e-prints*. arXiv: 1511.04992 [stat.CO].

Delyon, Bernard and Ying Hu (2006). "Simulation of conditioned diffusion and application to parameter estimation". In: *Stochastic Processes and their Applications* 116.11, pp. 1660 –1675. ISSN: 0304-4149. DOI: https://doi.org/10.1016/j.spa.2006.04.004. URL: http://www.sciencedirect.com/science/article/pii/S0304414906000469.

Devroye, L. (1986). *Non-Uniform Random Variate Generation*. Springer-Verlag. ISBN: 9783540963059.

Diaconis, P. and D. Freedman (Aug. 1980). "Finite Exchangeable Sequences". In: *Ann. Probab.* 8.4, pp. 745–764. DOI: 10.1214/aop/1176994663. URL: https://doi.org/10.1214/aop/1176994663.

Donsker, M.D. (1951). *An Invariance Principle for Certain Probability Limit Theorems*. American Mathematical Society. Memoirs.

Dooren, Paul van and Luc de Ridder (1976). "An adaptive algorithm for numerical integration over an n-dimensional cube". In: *Journal of Computational and Applied Mathematics* 2.3, pp. 207 –217. ISSN: 0377-0427. DOI: http://dx.doi.org/10.1016/0771-050X(76)90005-X.

Douc, R., O. Cappé, and E. Moulines (2005). "Comparison of resampling schemes for particle filtering". In: *Proceedings of the 4th International Symposium on Image and Signal Processing and Analysis, 2005.* Vol. 1. IEEE, pp. 64–69.

Doucet, A., A. Smith, N. de Freitas, and N. Gordon (2001). *Sequential Monte Carlo Methods in Practice*. Information Science and Statistics. Springer New York. ISBN: 9781441928870.

Doucet, Arnaud and Adam M. Johansen (2011). "A Tutorial on Particle Filtering and Smoothing: Fifteen Years Later". In: *The Oxford Handbook of Nonlinear Filtering*. Ed. by D. Crisan and B. Rozovsky. Oxford University Press.

Durham, Garland B and A. Ronald Gallant (2002). "Numerical Techniques for Maximum Likelihood Estimation of Continuous-Time Diffusion Processes". In: *Journal of Business & Economic Statistics* 20.3, pp. 297–338. DOI: 10.1198/073500102288618397. eprint: http://dx.doi.org/10.1198/073500102288618397.

Durrett, R. (2010). *Probability: Theory and Examples*. Cambridge series on statistical and probabilistic mathematics. Cambridge University Press. ISBN: 9780511918445.

Durstenfeld, Richard (July 1964). "Algorithm 235: Random Permutation". In: *Commun. ACM* 7.7, pp. 420–. ISSN: 0001-0782. DOI: 10.1145/364520.364540. URL: http://doi.acm.org/10.1145/364520.364540.

Ethier, S.N. and T.G. Kurtz (1986). *Markov Processes: Characterization and Convergence*. Wiley Series in Probability and Statistics. Wiley.

Fearnhead, Paul (2008). "Computational methods for complex stochastic systems: a review of some alternatives to MCMC". In: *Statistics and Computing* 18.2, pp. 151–171. ISSN: 1573-1375. DOI: 10.1007/s11222-007-9045-8.

Fearnhead, Paul, Vasilieos Giagos, and Chris Sherlock (2014). "Inference for reaction networks using the linear noise approximation". In: *Biometrics* 70.2, pp. 457–466. ISSN: 1541-0420. DOI: 10.1111/biom.12152.

Fill, James Allen (Feb. 1998). "An interruptible algorithm for perfect sampling via Markov chains". In: *Ann. Appl. Probab.* 8.1, pp. 131–162. DOI: 10.1214/aoap/1027961037.

Foss, S. G. and R. L. Tweedie (1998). "Perfect Simulation and Backward Coupling". In: *Stochastic Models* 14.1-2, pp. 187–203.

Gelman, A., J.B. Carlin, H.S. Stern, and D.B. Rubin (2003). *Bayesian Data Analysis, Second Edition*. Chapman & Hall/CRC Texts in Statistical Science. Taylor & Francis. ISBN: 9781420057294.

Genz, Alan (1991). "An adaptive numerical integration algorithm for simplices". In: *Computing in the 90's: The First Great Lakes Computer Science Conference Kalamazoo, Michigan, USA, October 18–20, 1989 Proceedings*. Ed. by Naveed

A. Sherwani, Elise de Doncker, and John A. Kapenga. New York, NY: Springer New York, pp. 279–285. ISBN: 978-0-387-34815-5. DOI: 10.1007/BFb0038504.

Geweke, John (1989). "Bayesian Inference in Econometric Models Using Monte Carlo Integration". In: *Econometrica* 57.6, pp. 1317–1339. ISSN: 00129682, 14680262.

Glasserman, P. (2010). *Monte Carlo Methods in Financial Engineering*. Stochastic Modelling and Applied Probability. Springer New York. ISBN: 9781441918222.

Golightly, A. and D.J. Wilkinson (2008). "Bayesian inference for nonlinear multivariate diffusion models observed with error". In: *Computational Statistics & Data Analysis* 52.3, pp. 1674 –1693. ISSN: 0167-9473. DOI: http://dx.doi.org/10.1016/j.csda.2007.05.019.

Golightly, Andrew, Daniel A. Henderson, and Chris Sherlock (2015). "Delayed acceptance particle MCMC for exact inference in stochastic kinetic models". In: *Statistics and Computing* 25.5, pp. 1039–1055. ISSN: 1573-1375. DOI: 10.1007/s11222-014-9469-x. URL: http://dx.doi.org/10.1007/s11222-014-9469-x.

Golightly, Andrew and Darren J. Wilkinson (2011). "Bayesian parameter inference for stochastic biochemical network models using particle Markov chain Monte Carlo". In: *Interface Focus* 1.6, pp. 807–820. ISSN: 2042-8898. DOI: 10.1098/rsfs.2011.0047.

Grewal, M.S. and A.P. Andrews (2011). *Kalman Filtering: Theory and Practice Using MATLAB*. Wiley. ISBN: 9781118210468.

Gustafsson, F., F. Gunnarsson, N. Bergman, U. Forssell, J. Jansson, R. Karlsson, and P.-J. Nordlund (Feb. 2002). "Particle Filters for Positioning, Navigation, and Tracking". In: *Trans. Sig. Proc.* 50.2, pp. 425–437. ISSN: 1053-587X. DOI: 10.1109/78.978396.

Handel, Ramon van (2009). "Uniform time average consistency of Monte Carlo particle filters". In: *Stochastic Processes and their Applications* 119.11, pp. 3835 – 3861. ISSN: 0304-4149. DOI: http://dx.doi.org/10.1016/j.spa.2009.09.004.

Hewitt, E and L.J. Savage (Jan. 1955). "Symmetric measures on cartesian products". In: *Transactions of the American Mathematical Society* 80, pp. 470–501. DOI: 10.1090/s0002-9947-1955-0076206-8.

Hinrichs, Aicke, Erich Novak, Mario Ullrich, and Henryk Woźniakowski (2014). "The curse of dimensionality for numerical integration of smooth functions II". In: *Journal of Complexity* 30.2. Dagstuhl 2012, pp. 117 –143. ISSN: 0885-064X. DOI: http://dx.doi.org/10.1016/j.jco.2013.10.007.

Hobert, James P., Galin L. Jones, Brett Presnell, and Jeffrey S. Rosenthal (2002). "On the applicability of regenerative simulation in Markov chain Monte Carlo". In: *Biometrika* 89.4, pp. 731–743. DOI: 10.1093/biomet/89.4.731. eprint: /oup/backfile/content_public/journal/biomet/89/4/10.1093/biomet/89.4.731/2/890731.pdf.

Hollander, M. and D.A. Wolfe (1973). *Nonparametric statistical methods*. Wiley Series in Probability and Statistics - Applied Probability and Statistics Section. Wiley. ISBN: 9780471406358.

Huang, Guoquan P, Anastasios I Mourikis, and Stergios I Roumeliotis (2008). "Analysis and improvement of the consistency of extended Kalman filter based SLAM". In: *2008 IEEE International Conference on Robotics and Automation*. IEEE, pp. 473–479.

Jacob, Pierre E., John O'Leary, and Yves F. Atchadé (2020). "Unbiased Markov chain Monte Carlo methods with couplings". In: *Journal of the Royal Statistical Society Series B* 82.3, pp. 543–600. DOI: 10.1111/rssb.12336.

Jewell, C.P, M.J Keeling, and G.O Roberts (2008). "Predicting undetected infections during the 2007 foot-and-mouth disease outbreak". In: *Journal of The Royal Society Interface*. ISSN: 1742-5689. DOI: 10.1098/rsif.2008.0433.

Julier, Simon J and Jeffrey K Uhlmann (1997). "New extension of the Kalman filter to nonlinear systems". In: *Signal processing, sensor fusion, and target recognition VI*. Vol. 3068. International Society for Optics and Photonics, pp. 182–193.

Julier, Simon J and Jeffrey K Uhlmann (2004). "Unscented filtering and nonlinear estimation". In: *Proceedings of the IEEE* 92.3, pp. 401–422.

Kahn, H. and A. W. Marshall (1953). "Methods of Reducing Sample Size in Monte Carlo Computations". In: *Journal of the Operations Research Society of Amer-*

ica 1.5, pp. 263–278. ISSN: 00963984. URL: http://www.jstor.org/stable/166789.

Kallenberg, O. (1997). *Foundations of Modern Probability*. Probability and Its Applications. Springer-Verlag New York. ISBN: 9780387949574.

Kim, Sangjoon, Neil Shephard, and Siddhartha Chib (1998). "Stochastic Volatility: Likelihood Inference and Comparison with ARCH Models". In: *The Review of Economic Studies* 65.3, pp. 361–393. ISSN: 00346527, 1467937X.

Kloeden, P.E. and E. Platen (1992). *Numerical Solution of Stochastic Differential Equations*. Applications of Mathematics. Springer-Verlag. ISBN: 9783540540625.

Komorowski, Michał, Bärbel Finkenstädt, Claire V. Harper, and David A. Rand (2009). "Bayesian inference of biochemical kinetic parameters using the linear noise approximation". In: *BMC Bioinformatics* 10.1, pp. 1–10. ISSN: 1471-2105. DOI: 10.1186/1471-2105-10-343.

Kong, Augustine (1992). "A note on importance sampling using standardized weights". In: *University of Chicago, Dept. of Statistics, Tech. Rep* 348.

Kong, Augustine, Jun S. Liu, and Wing Hung Wong (1994). "Sequential Imputations and Bayesian Missing Data Problems". In: *Journal of the American Statistical Association* 89.425, pp. 278–288. ISSN: 01621459. URL: http://www.jstor.org/stable/2291224.

Kuhner, Mary K. (2006). "LAMARC 2.0: maximum likelihood and Bayesian estimation of population parameters". In: *Bioinformatics* 22.6, p. 768. DOI: 10.1093/bioinformatics/btk051.

Künsch, H.R. (2001). "State space and hidden Markov models". In: *Complex Stochastic Systems*. Ed. by O.E. Barndorff-Nielsen, D.R. Cox, and C. Klüppelberg. CRC Press, pp. 109 –173. DOI: 10.1201/9781420035988.

L'Ecuyer, Pierre (1994). "Uniform random number generation". In: *Annals of Operations Research* 53.1, pp. 77–120. ISSN: 1572-9338. DOI: 10.1007/BF02136827. URL: https://doi.org/10.1007/BF02136827.

Lawler, Gregory F. and Alan D. Sokal (1988). "Bounds on the L^2 Spectrum for Markov Chains and Markov Processes: A Generalization of Cheeger's Inequality". In: *Transactions of the American Mathematical Society* 309.2, pp. 557–580. ISSN: 00029947.

Lee, Anthony (2011). "On auxiliary variables and many-core architectures in computational statistics". PhD thesis. University of Oxford.

Lefebvre, Tine, Herman Bruyninckx, and Joris De Schutter (2004). "Kalman filters for non-linear systems: a comparison of performance". In: *International Journal of Control* 77.7, pp. 639–653. DOI: 10.1080/00207170410001704998. eprint: https://doi.org/10.1080/00207170410001704998. URL: https://doi.org/10.1080/00207170410001704998.

Lelièvre, Tony and Gabriel Stoltz (May 2016). "Partial differential equations and stochastic methods in molecular dynamics". In: *Acta Numerica* 25, pp. 681–880. DOI: 10.1017/S0962492916000039.

Lin, Ming, Rong Chen, and Jun S. Liu (2013). "Lookahead Strategies for Sequential Monte Carlo". In: *Statistical Science* 28.1, pp. 69–94. ISSN: 08834237, 21688745.

Lindsten, F. and T.B. Schön (2013). *Backward Simulation Methods for Monte Carlo Statistical Inference*. Vol. 6. Foundations and Trends® in Machine Learning 1. Now Publishers, pp. 1–143. ISBN: 9781601986986.

Lindsten, F., P. Bunch, S. S. Singh, and T. B. Schön (May 2015). "Particle ancestor sampling for near-degenerate or intractable state transition models". In: *ArXiv e-prints*. arXiv: 1505.06356 [stat.CO].

Lindsten, Fredrik, Randal Douc, and Eric Moulines (2015). "Uniform Ergodicity of the Particle Gibbs Sampler". In: *Scandinavian Journal of Statistics* 42.3, pp. 775–797. DOI: 10.1111/sjos.12136. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/sjos.12136. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/sjos.12136.

Lindsten, Fredrik, Michael I. Jordan, and Thomas B. Schön (Jan. 2014). "Particle Gibbs with Ancestor Sampling". In: *J. Mach. Learn. Res.* 15.1, pp. 2145–2184. ISSN: 1532-4435.

Lindström, Erik (2012). "A regularized bridge sampler for sparsely sampled diffusions". In: *Statistics and Computing* 22.2, pp. 615–623. ISSN: 1573-1375. DOI: 10.1007/s11222-011-9255-y.

Liu, J.S. (2001). *Monte Carlo Strategies in Scientific Computing*. Springer Series in Statistics. Springer New York. ISBN: 9780387952307.

Liu, Jun S. (1996). "Metropolized independent sampling with comparisons to rejection sampling and importance sampling". In: *Statistics and Computing* 6.2, pp. 113–119. ISSN: 1573-1375. DOI: 10.1007/BF00162521. URL: https://doi.org/10.1007/BF00162521.

Mengersen, K. L. and R. L. Tweedie (Feb. 1996). "Rates of convergence of the Hastings and Metropolis algorithms". In: *Ann. Statist.* 24.1, pp. 101–121. DOI: 10.1214/aos/1033066201.

Meulen, Frank van der and Moritz Schauer (2017). "Bayesian estimation of discretely observed multi-dimensional diffusion processes using guided proposals". In: *Electron. J. Statist.* 11.1, pp. 2358–2396. DOI: 10.1214/17-EJS1290. URL: https://doi.org/10.1214/17-EJS1290.

Meyn, Sean P. and R. L. Tweedie (Nov. 1994). "Computable Bounds for Geometric Convergence Rates of Markov Chains". In: *Ann. Appl. Probab.* 4.4, pp. 981–1011. DOI: 10.1214/aoap/1177004900. URL: https://doi.org/10.1214/aoap/1177004900.

Meyn, Sean and Richard L. Tweedie (2009). *Markov Chains and Stochastic Stability*. 2nd ed. Cambridge Mathematical Library. Cambridge University Press. DOI: 10.1017/CBO9780511626630.

Mihaylova, Lyudmila, Avishy Y. Carmi, François Septier, Amadou Gning, Sze Kim Pang, and Simon Godsill (2014). "Overview of Bayesian sequential Monte Carlo methods for group and extended object tracking". In: *Digital Signal Processing* 25, pp. 1 –16. ISSN: 1051-2004. DOI: http://dx.doi.org/10.1016/j.dsp.2013.11.006. URL: http://www.sciencedirect.com/science/article/pii/S1051200413002716.

Mira, Antonietta (Nov. 2001). "Ordering and Improving the Performance of Monte Carlo Markov Chains". In: *Statist. Sci.* 16.4, pp. 340–350. DOI: 10.1214/ss/1015346319. URL: https://doi.org/10.1214/ss/1015346319.

Murray, Iain and Matthew Graham (2016). "Pseudo-Marginal Slice Sampling". In: *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*. Ed. by Arthur Gretton and Christian C. Robert. Vol. 51. JMLR: W&CP. Cadiz, Spain, pp. 911–919.

Neal, Peter J. and Gareth O. Roberts (2004). "Statistical inference and model selection for the 1861 Hagelloch measles epidemic". In: *Biostatistics* 5.2, p. 249. DOI: 10.1093/biostatistics/5.2.249.

Neal, Radford M., Matthew J. Beal, and Sam T. Roweis (2004). "Inferring State Sequences for Non-linear Systems with Embedded Hidden Markov Models". In: *Advances in Neural Information Processing Systems 16*. Ed. by S. Thrun, L. K. Saul, and B. Schölkopf. MIT Press, pp. 401–408. URL: http://papers.nips.cc/paper/2391-inferring-state-sequences-for-non-linear-systems-with-embedded-hidden-markov-models.pdf.

Nummelin, Esa (1984). *General Irreducible Markov Chains and Non-Negative Operators*. Cambridge Tracts in Mathematics. Cambridge University Press. DOI: 10.1017/CBO9780511526237.

Oh, Man-Suk and James O. Berger (1992). "Adaptive importance sampling in Monte Carlo integration". In: *Journal of Statistical Computation and Simulation* 41.3-4, pp. 143–168. DOI: 10.1080/00949659208810398. eprint: http://dx.doi.org/10.1080/00949659208810398.

Owen, Art B. (2013). *Monte Carlo theory, methods and examples.*

Owen, Art and Yi Zhou (2000). "Safe and Effective Importance Sampling". In: *Journal of the American Statistical Association* 95.449, pp. 135–143. ISSN: 01621459.

Papaspiliopoulos, Omiros and Gareth Roberts (2012). "Importance sampling techniques for estimation of diffusion models". In: *Statistical methods for stochastic differential equations* 124, pp. 311–340.

Papoulis, A., S.U. Pillai, and S.U. Pillai (2002). *Probability, Random Variables, and Stochastic Processes*. McGraw-Hill electrical and electronic engineering series. McGraw-Hill. ISBN: 9780073660110.

Pedersen, Asger Roer (1995). "A New Approach to Maximum Likelihood Estimation for Stochastic Differential Equations Based on Discrete Observations". In: *Scandinavian Journal of Statistics* 22.1, pp. 55–71. ISSN: 03036898, 14679469.

Peskun, P. H. (1973). "Optimum Monte-Carlo Sampling Using Markov Chains". In: *Biometrika* 60.3, pp. 607–612. ISSN: 00063444. URL: http://www.jstor.org/stable/2335011.

Petzold, Linda (1983). "Automatic Selection of Methods for Solving Stiff and Nonstiff Systems of Ordinary Differential Equations". In: *SIAM Journal on Scientific and Statistical Computing* 4.1, pp. 136–148. DOI: 10.1137/0904010.

Pitt, Michael K. and Neil Shephard (1999). "Filtering via Simulation: Auxiliary Particle Filters". In: *Journal of the American Statistical Association* 94.446, pp. 590–599. ISSN: 01621459.

Propp, James and David Wilson (1998). "Coupling from the past: a user's guide". In: *Microsurveys in Discrete Probability*. Ed. by D. Aldous and J. Propp. Vol. 41. DIMACS Series in Discrete Mathematics and Theoretical Computer Science. American Mathematical Society, pp. 181–192.

Rabiner, L. R. (1989). "A tutorial on hidden Markov models and selected applications in speech recognition". In: *Proceedings of the IEEE* 77.2, pp. 257–286. ISSN: 0018-9219. DOI: 10.1109/5.18626.

Rambharat, Bhojnarine R. and Anthony E. Brockwell (Mar. 2010). "Sequential Monte Carlo pricing of American-style options under stochastic volatility models". In: *Ann. Appl. Stat.* 4.1, pp. 222–265. DOI: 10.1214/09-AOAS286.

Rapp, Knut and Per-Ole Nyman (2004). "Stability Properties of the Discrete-Time Extended Kalman Filter". In: *IFAC Proceedings Volumes* 37.13. 6th IFAC Symposium on Nonlinear Control Systems 2004 (NOLCOS 2004), Stuttgart, Germany, 1-3 September, 2004, pp. 1377 –1382. ISSN: 1474-6670. DOI: https://doi.org/10.1016/S1474-6670(17)31420-9. URL: http://www.sciencedirect.com/science/article/pii/S1474667017314209.

Richard, Jean-Francois and Wei Zhang (2007). "Efficient high-dimensional importance sampling". In: *Journal of Econometrics* 141.2, pp. 1385 –1411. ISSN: 0304-4076. DOI: http://dx.doi.org/10.1016/j.jeconom.2007.02.007.

Robert, C. and G. Casella (2004). *Monte Carlo Statistical Methods*. Springer Texts in Statistics. Springer New York. ISBN: 9781475741452.

Roberts, G. O., A. Gelman, and W. R. Gilks (Feb. 1997). "Weak convergence and optimal scaling of random walk Metropolis algorithms". In: *Ann. Appl. Probab.* 7.1, pp. 110–120. DOI: 10.1214/aoap/1034625254.

Roberts, G. O. and R. L. Tweedie (Mar. 1996). "Geometric convergence and central limit theorems for multidimensional Hastings and Metropolis algorithms". In: *Biometrika* 83.1, pp. 95–110. ISSN: 0006-3444. DOI: 10.1093/biomet/83.1.95. eprint: http://oup.prod.sis.lan/biomet/article-pdf/83/1/95/709644/83-1-95.pdf. URL: https://doi.org/10.1093/biomet/83.1.95.

Roberts, Gareth O. (1998). "Optimal Metropolis algorithms for product measures on the vertices of a hypercube". In: *Stochastics and Stochastic Reports* 62.3-4, pp. 275–283. DOI: 10.1080/17442509808834136. eprint: https://doi.org/10.1080/17442509808834136. URL: https://doi.org/10.1080/17442509808834136.

Roberts, Gareth O. and Jeffrey S. Rosenthal (1998a). "Markov-Chain Monte Carlo: Some Practical Implications of Theoretical Results". In: *The Canadian Journal of Statistics / La Revue Canadienne de Statistique* 26.1, pp. 5–20. ISSN: 03195724.

Roberts, Gareth O. and Jeffrey S. Rosenthal (1998b). "Optimal scaling of discrete approximations to Langevin diffusions". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 60.1, pp. 255–268. ISSN: 1467-9868. DOI: 10.1111/1467-9868.00123.

Roberts, Gareth O. and Jeffrey S. Rosenthal (Nov. 2001). "Optimal scaling for various Metropolis-Hastings algorithms". In: *Statist. Sci.* 16.4, pp. 351–367. DOI: 10.1214/ss/1015346320.

Roberts, Gareth O. and Jeffrey S. Rosenthal (2004). "General state space Markov chains and MCMC algorithms". In: *Probab. Surveys* 1, pp. 20–71. DOI: 10.1214/154957804100000024.

Roberts, Gareth O. and Jeffrey S. Rosenthal (June 2008). "Variance bounding Markov chains". In: *Ann. Appl. Probab.* 18.3, pp. 1201–1214. DOI: 10.1214/07-AAP486. URL: https://doi.org/10.1214/07-AAP486.

Roberts, Gareth O. and Jeffrey S. Rosenthal (2011). "Quantitative Non-Geometric Convergence Bounds for Independence Samplers". In: *Methodology and Computing in Applied Probability* 13.2, pp. 391–403. ISSN: 1573-7713. DOI: 10.1007/s11009-009-9157-z. URL: https://doi.org/10.1007/s11009-009-9157-z.

Roberts, Gareth and Jeffrey Rosenthal (1997). "Geometric Ergodicity and Hybrid Markov Chains". In: *Electron. Commun. Probab.* 2, pp. 13–25. DOI: 10.1214/ECP.v2-981.

Rogers, L.C.G. and D. Williams (2000a). *Diffusions, Markov Processes, and Martingales: Volume 1, Foundations*. Cambridge Mathematical Library. Cambridge University Press. ISBN: 9780521775946.

Rogers, L.C.G. and D. Williams (2000b). *Diffusions, Markov Processes and Martingales: Volume 2, Itô Calculus*. Cambridge Mathematical Library. Cambridge University Press. ISBN: 9780521775939.

Rosenthal, Jeffrey S. (1995). "Minorization Conditions and Convergence Rates for Markov Chain Monte Carlo". In: *Journal of the American Statistical Association* 90.430, pp. 558–566. ISSN: 01621459. URL: http://www.jstor.org/stable/2291067.

Rosenthal, Jeffrey S. (2001). "A review of asymptotic convergence for general state space Markov chains". In: *Far East Journal of Theoretical Statistics* 5.1, pp. 37–50.

Rosenthal, Jeffrey (2002). "Quantitative Convergence Rates of Markov Chains: A Simple Account". In: *Electron. Commun. Probab.* 7, pp. 123–128. DOI: 10.1214/ECP.v7-1054. URL: https://doi.org/10.1214/ECP.v7-1054.

Russell, Stuart J. and Peter Norvig (2003). *Artificial Intelligence: A Modern Approach*. 2nd ed. Pearson Education. ISBN: 0137903952.

Schauer, Moritz, Frank van der Meulen, and Harry van Zanten (Nov. 2017). "Guided proposals for simulating multi-dimensional diffusion bridges". In: *Bernoulli* 23.4A, pp. 2917–2950. DOI: 10.3150/16-BEJ833.

Sen, Deborshee, Ajay Jasra, and Yan Zhou (2017). "Some contributions to sequential Monte Carlo methods for option pricing". In: *Journal of Statistical Computation and Simulation* 87.4, pp. 733–752. DOI: 10.1080/00949655.2016.1224238.

Shephard, Neil (1994). "Local scale models". In: *Journal of Econometrics* 60.1, pp. 181 –202. ISSN: 0304-4076. DOI: http://dx.doi.org/10.1016/0304-4076(94)90043-4.

Sherlock, Chris, Paul Fearnhead, and Gareth O. Roberts (May 2010). "The Random Walk Metropolis: Linking Theory and Practice Through a Case Study". In: *Statist. Sci.* 25.2, pp. 172–190. DOI: 10.1214/10-STS327. URL: https://doi.org/10.1214/10-STS327.

Sherlock, Chris and Gareth Roberts (Aug. 2009). "Optimal scaling of the random walk Metropolis on elliptically symmetric unimodal targets". In: *Bernoulli* 15.3, pp. 774–798. DOI: 10.3150/08-BEJ176.

Shestopaloff, Alexander Y. and Radford M. Neal (Sept. 2018). "Sampling Latent States for High-Dimensional Non-Linear State Space Models with the Embedded HMM Method". In: *Bayesian Anal.* 13.3, pp. 797–822. DOI: 10.1214/17-BA1077. URL: https://doi.org/10.1214/17-BA1077.

Shreve, S.E. (2004). *Stochastic Calculus for Finance: Continuous-time models*. Springer finance. Springer. ISBN: 9780387401010.

Smith, A. F. M. and G. O. Roberts (1993). "Bayesian Computation Via the Gibbs Sampler and Related Markov Chain Monte Carlo Methods". In: *Journal of the Royal Statistical Society. Series B (Methodological)* 55.1, pp. 3–23. ISSN: 00359246.

Smith, R.A., E.L. Ionides, and A.A. King (Apr. 2017). "Infectious Disease Dynamics Inferred from Genetic Data via Sequential Monte Carlo". In: *Mol Biol Evol*. DOI: 10.1093/molbev/msx124.

Sobolev, S.L. and V.L. Vaskevich (2013). *The Theory of Cubature Formulas*. Vol. 415. Mathematics and Its Applications. Springer Netherlands. ISBN: 9789401589130.

Stroock, D.W. and S.R.S. Varadhan (1997). *Multidimensional Diffusion Processes*. Grundlehren der mathematischen Wissenschaften. Springer Berlin Heidelberg. ISBN: 9783540903536.

Süli, E. and D.F. Mayers (2003). *An Introduction to Numerical Analysis*. Cambridge University Press. ISBN: 9780521810265.

Tierney, Luke (Dec. 1994). "Markov Chains for Exploring Posterior Distributions". In: *Ann. Statist.* 22.4, pp. 1701–1728. DOI: 10.1214/aos/1176325750.

Tierney, Luke (Feb. 1998). "A note on Metropolis-Hastings kernels for general state spaces". In: *Ann. Appl. Probab.* 8.1, pp. 1–9. DOI: 10.1214/aoap/1027961031. URL: https://doi.org/10.1214/aoap/1027961031.

Tjelmeland, Håkon (2004). *Using all Metropolis-Hastings proposals to estimate mean values*. Tech. rep. Trondheim, Norway: Norwegian University of Science and Technology.

Toutenburg, H. (1971). "Fisher, R. A., and F. Yates: Statistical Tables for Biological, Agricultural and Medical Research. 6th Ed. Oliver & Boyd, Edinburgh and London 1963. X, 146 P. Preis 42 s net". In: *Biometrische Zeitschrift* 13.4, pp. 285–285. DOI: 10.1002/bimj.19710130413. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/bimj.19710130413. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/bimj.19710130413.

Van Kampen, N.G. (1992). *Stochastic Processes in Physics and Chemistry*. North-Holland Personal Library. Elsevier Science. ISBN: 9780080571386.

Wan, Eric A and Rudolph Van Der Merwe (2000). "The unscented Kalman filter for nonlinear estimation". In: *Proceedings of the IEEE 2000 Adaptive Systems for Signal Processing, Communications, and Control Symposium (Cat. No. 00EX373)*. Ieee, pp. 153–158.

West, M. and J. Harrison (1999). *Bayesian Forecasting and Dynamic Models*. Springer Series in Statistics. Springer.

Whitaker, Gavin A., Andrew Golightly, Richard J. Boys, and Chris Sherlock (2017). "Improved bridge constructs for stochastic differential equations". In: *Statistics and Computing* 27.4, pp. 885–900. ISSN: 1573-1375. DOI: 10.1007/s11222-016-9660-3.

Wiener, Norbert (1923). "Differential-Space". In: *Journal of Mathematics and Physics* 2.1-4, pp. 131–174. DOI: 10.1002/sapm192321131. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/sapm192321131. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/sapm192321131.

Wilkinson, D.J. (2006). *Stochastic Modelling for Systems Biology*. Chapman & Hall/CRC Mathematical & Computational Biology. Taylor & Francis. ISBN: 9781584885405.

Williams, D. (1991). *Probability with Martingales*. EBL-Schweitzer. Cambridge University Press. ISBN: 9781139640923.

Yang, Wan, Alicia Karspeck, and Jeffrey Shaman (Apr. 2014). "Comparison of Filtering Methods for the Modeling and Retrospective Forecasting of Influenza Epidemics". In: *PLOS Computational Biology* 10.4, pp. 1–15. DOI: 10.1371/journal.pcbi.1003583.

de Finetti, B. (1931). "Funzione caratteristica di un fenomeno aleatorio". In: *Atti della R. Accademia Nazionale dei Lincei, Ser. 6. Memorie, Classe di Scienze Fisiche, Matematiche e Naturali 4*, pp. 251–299.

A

PROOFS

A.1 PROOF OF LEMMA 2.1.3

For any $i \neq j$, and any $k \in \{1, \ldots, N\}$,

$$\mathbb{E}[\|X_i^{(k)} - X_j^{(k)}\|^2] = \mathbb{E}[c_k^2 (Z_i - Z_j)^2] \,,$$

where Z_i, $Z_j \sim \mathrm{N}(0, 1)$. Hence,

$$\mathbb{E}[\|X_i^{(k)} - X_j^{(k)}\|^2] = 2c_k^2 \,.$$

A.2 PROOF OF LEMMA 2.3.20

Firstly, note that, for a fixed set C, inequalities (21) and (22) are equivalent, since, with $v(x) := 1 + f(x)$, the former holds if, and only if,

$$\mathbb{E}_{P(x,\cdot)}(1 + f(Y)) \leq \gamma(1 + f(x)) + \beta \mathbb{1}_C(x) \,,$$

which gives the latter. As a result Definition 2.3.18 implies Definition 2.3.19. Now, suppose that X_t is a Markov chain which satisfies a drift condition in the sense of 2.3.19; that is, there exists a function $v : \mathcal{X} \to [1, \infty)$, an ϵ-small set C, and positive, finite constants, β and $\gamma < 1$, such that

$$\mathbb{E}_{P(x,\cdot)}(v(Y)) \leq \gamma v(x) + \beta \mathbb{1}_C(x) \,.$$

Let α be a finite constant such that $\alpha > \beta/(1 - \gamma) - 1$; the existence of which is guaranteed since $\beta < \infty$ and $\gamma < 1$, and define

$$C_\alpha := \{x \in C : v(x) \leq \alpha + 1\} = \{x \in C : f(x) \leq \alpha\} \,,$$

where $f(x) := v(x) - 1$, and $\gamma_\alpha := \gamma + \beta/(1 + \alpha) > \gamma$. By definition of α,

$$\gamma_\alpha < \gamma + (1 - \gamma) = 1 \,.$$

Clearly, since $C_\alpha \subseteq C$, if $x \in C_\alpha$, then

$$\mathbb{E}_{P(x,\cdot)}(v(Y)) \leq \gamma v(x) + \beta < \gamma_\alpha v(x) + \beta \,,$$

and, if $x \notin C$, then

$$\mathbb{E}_{P(x,\cdot)}(v(Y)) \leq \gamma v(x) < \gamma_\alpha v(x) \,.$$

So all that remains is to consider the case when $x \in C \backslash C_\alpha$. By definition $v(x) > \alpha + 1$, hence

$$\mathbb{E}_{P(x,\cdot)}(v(Y)) \leq \gamma v(x) + \beta < [\gamma + \beta/(1 + \alpha)] v(x) = \gamma_\alpha v(x) \,.$$

Therefore, X_t satisfies a drift condition in the sense of Definition 2.3.18.

A.3 PROOF OF THEOREM 2.3.25

By Theorem 5 of Roberts and Rosenthal, 2008, geometric ergodicity is equivalent to the chain being variance bounding (see the definition in Section 2 of that paper). Moreover, by Theorem 7 of the same paper, variance bounding is equivalent to the MCMC estimates satisfying a central limit theorem for any function h which is square-integrable with respect to the unique stationary distribution π. Furthermore, by Theorem 14 of that paper, variance bounding is equivalent to the chain have a non-zero right spectral gap, ρ, (see, for instance, Conway, 1994 for a definition). Now, Theorem 2.1 of Lawler and Sokal, 1988 provides the following bounds on the right spectral gap in terms of the conductance of the chain;

$$\kappa^2/8 \leq \rho \leq \kappa \,.$$

Hence, the chain having a non-zero right spectral gap is equivalent to the chain having a non-zero conductance, thus completing the proof.

A.4 PROOF OF THEOREM 2.3.34

Mengersen and Tweedie, 1996 show that the Metropolis-Hastings random-walk sampler is geometrically ergodic under such conditions. By Lemma 2.3.32, the Metropolis-Hastings random-walk sampler is non-negative. Hence, by Theorem 2.3.25, the chain has a non-zero conductance and, therefore, by Theorem 2.1, Lawler and Sokal, 1988, has a non-zero right spectral gap, ρ. Note that $\alpha_b(x,y) \geq \alpha_m(x,y)/2$. Therefore, letting $\mathcal{E}_P(f) := \langle f, f\rangle - \langle f, Pf\rangle$ be the *Dirichlet* form associated with a Markov chain with transition distribution P (see, for example Roberts and Rosenthal, 2008), $\mathcal{E}_{P_b}(f) \geq \mathcal{E}_{P_m}(f)/2$, where P_b and P_m denote the transition distributions corresponding to Barker's and the Metropolis-Hastings random-walk samplers respectively. Therefore (see, for example, Roberts and Rosenthal, 2008), defining ρ_b and ρ_m to be the right-spectral gaps corresponding to Barker's and the Metropolis-Hastings random-walk samplers respectively, $\rho_b \geq \rho_m/2$. Hence, Barker's random-walk sampler has a non-zero right spectral gap and, therefore, by Theorem 2.3.25, the MCMC estimates corresponding to either Barker's, or the Metropolis-Hastings random walk sampler satisfy a central limit theorem for all functions which are square-integrable with respect to π.

A.5 PROOF OF COROLLARY 2.3.38

By Theorem 2.3.36,

$$\bar{\alpha}(\lambda) = \mathbb{E}\left[\frac{\exp(-\lambda^2\psi^2/2 + \lambda\psi W)}{1+\exp(-\lambda^2\psi^2/2+\lambda\psi W)}\right] = \mathbb{E}[\{1+\exp(\lambda^2\psi^2/2 - \lambda\psi W)\}^{-1}] \,.$$

Let $\beta := \lambda\psi/2$. The aim is to maximise $\bar{J}(2\beta/\psi) = 8\beta^2\psi^{-2}\bar{\alpha}(2\beta/\psi)$. Now,

$$\bar{\alpha}(2\beta/\psi) = \int_{-\infty}^{\infty} \frac{\phi(z)}{1+\exp(2\beta^2+2\beta z)}\,\mathrm{d}z \,.$$

To derive an upper bound on this consider the two disjoint sets, $(-\infty, -\beta)$ and $[-\beta, \infty)$, which cover the real line. Firstly, $\exp(2\beta^2 + 2\beta z) > 0$ for any z and β. Hence,

$$\int_{-\infty}^{-\beta} \frac{\phi(z)}{1+\exp(2\beta^2+2\beta z)}\,\mathrm{d}z < \Phi(-\beta) \,.$$

Secondly, by Lemma B.0.7,

$$\begin{aligned}\int_{-\beta}^{\infty} \frac{\phi(z)}{1+\exp(2\beta^2+2\beta z)}\,\mathrm{d}z &< \exp(-2\beta^2)\int_{-\beta}^{\infty}\exp(-2\beta z)\phi(z)\,\mathrm{d}z \\ &= \exp(-2\beta^2)\mathbb{E}[\exp(-2\beta Z)\mathbb{1}_{[-\beta,\infty)}(Z)] \\ &= 1-\Phi(\beta) \\ &= \Phi(-\beta) \,.\end{aligned}$$

Therefore, $\bar{\alpha}(2\beta/\psi) < 2\Phi(-\beta)$. Figure 79 shows a plot of $\beta^2\bar{\alpha}(2\beta/\psi)$ against β for $\beta \in [0,2]$. There is a local maxima of $\beta^2\bar{\alpha}(2\beta/\psi)$ at $\hat{\beta} = 1.228$ to three decimal places, and $\hat{\beta}^2\bar{\alpha}(2\hat{\beta}/\psi)$ is equal to 0.240 to three decimal places. Moreover, from the proof of Corollary 2.3.37, $\beta^2\Phi(-\beta)$ is decreasing for $\beta > 2$. Hence, for $\beta > 2$,

$$\beta^2\bar{\alpha}(2\beta/\psi) < 2\beta^2\Phi(-\beta) < 8\Phi(-2) < 0.2 < 0.240 \,.$$

Combining this with Figure 79 demonstrates that the global maximum of $\beta^2\bar{\alpha}(2\beta/\psi)$ in the positive domain is 1.228 to three decimal places.

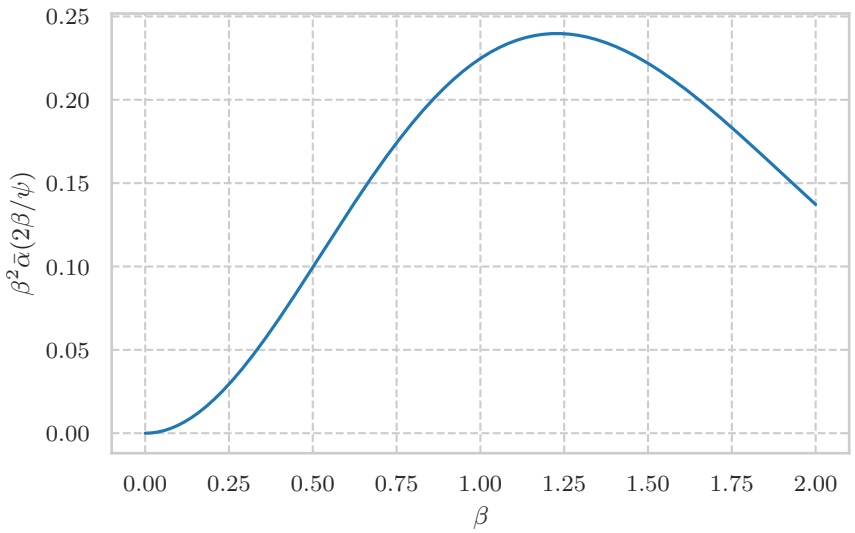

Figure 79: A plot of $\beta^2\bar{\alpha}(2\beta/\psi)$ against β for $\beta \in [0, 2]$.

A.6 PROOF OF THEOREM 2.4.3

For any $j \in \{1, \ldots, N\}$, and some permutation, σ, of $\{1, \ldots, N\}$,

$$\begin{aligned}\mathbb{E}(O^{(\sigma(j))}) &= \sum_{i=1}^{N} \mathbb{E}[\mathbb{1}_{(s_{j-1}(\tilde{w}^{(\sigma(1):\sigma(N))}), s_j(\tilde{w}^{(\sigma(1):\sigma(N))})]}(U_i)] \\ &= \sum_{i=1}^{N} \mathbb{P}(s_{j-1}(\tilde{w}^{(\sigma(1):\sigma(N))}) < U_i \leq s_j(\tilde{w}^{(\sigma(1):\sigma(N))})) \,.\end{aligned}$$

For multinomial resampling, $U_i \sim \text{Unif}(0, 1)$ for any $i \in \{1, \ldots, N\}$, and σ is the identity permutation. Therefore, for any $(i, j) \in \{1, \ldots, N\}^2$,

$$\mathbb{P}(s_{j-1}(\tilde{w}^{(\sigma(1):\sigma(N))}) < U_i \leq s_j(\tilde{w}^{(\sigma(1):\sigma(N))})) = s_j(\tilde{w}^{(1:N)}) - s_{j-1}(\tilde{w}^{(1:N)}) = \tilde{w}^{(j)} \,.$$

Hence, $\mathbb{E}(O^{(j)}) = N\tilde{w}^{(j)}$. For stratified and systematic resampling, $U_i \sim \text{Unif}((i-1)/N, i/N)$ for any $i \in \{1, \ldots, N\}$, and σ is random. Hence, for any $j \in \{1, \ldots, N\}$, if $i > \lceil N s_j(\tilde{w}^{(\sigma(1):\sigma(N))}) \rceil$, or $i \leq \lfloor N s_{j-1}(\tilde{w}^{(\sigma(1):\sigma(N))}) \rfloor$,

$$\mathbb{P}(s_{j-1}(\tilde{w}^{(\sigma(1):\sigma(N))}) < U_i \leq s_j(\tilde{w}^{(\sigma(1):\sigma(N))})) = 0 \,.$$

Moreover, for any $i \in \{\lceil N s_{j-1}(\tilde{w}^{(\sigma(1):\sigma(N))}) \rceil + 1, \ldots, \lfloor N s_j(\tilde{w}^{(\sigma(1):\sigma(N))}) \rfloor\}$,

$$\mathbb{P}(s_{j-1}(\tilde{w}^{(\sigma(1):\sigma(N))}) < U_i \leq s_j(\tilde{w}^{(\sigma(1):\sigma(N))})) = 1 \,.$$

Therefore, for brevity, dropping the partial sums' dependence on $\tilde{w}^{(\sigma(1):\sigma(N))}$,

$$O^{(\sigma(j))} = \lfloor N s_j \rfloor - \lceil N s_{j-1} \rceil + \mathbb{1}_{(\lfloor N s_j \rfloor / N, s_j]}(U_{\lceil N s_j \rceil}) + \mathbb{1}_{(s_{j-1}, \lceil N s_{j-1} \rceil / N]}(U_{\lceil N s_{j-1} \rceil}) \,. \tag{115}$$

Now,

$$\begin{aligned}\mathbb{P}(\lfloor N s_j \rfloor / N < U_{\lceil N s_j \rceil} \leq s_j) &= N s_j - \lfloor N s_j \rfloor \,, \\ \mathbb{P}(s_{j-1} < U_{\lceil N s_{j-1} \rceil} \leq \lceil N s_{j-1} \rceil / N) &= \lceil N s_{j-1} \rceil - N s_{j-1} \,.\end{aligned}$$

Hence,

$$\mathbb{E}(O^{(\sigma(j))}) = N(s_j - s_{j-1}) = N\tilde{w}^{(\sigma(j))} \,.$$

Moreover, from Equation (115) and Lemma B.0.12,

$$O^{(\sigma(j))} \geq \lfloor N s_j \rfloor - \lceil N s_{j-1} \rceil > N(s_j - s_{j-1}) - 2 = N\tilde{w}^{(\sigma(j))} - 2 \,.$$

To obtain the upper bound on $O^{(\sigma(j))} - N\tilde{w}^{(\sigma(j))}$, one can consider subtracting the indicators of the uniforms that might fall out of the interval, as opposed to adding the indicators of the uniforms that might fall inside the interval as Equation (115) does. From this viewpoint,

$$O^{(\sigma(j))} = \lceil Ns_j \rceil - \lfloor Ns_{j-1} \rfloor - \mathbb{1}_{(s_j, \lceil Ns_j \rceil / N]}(U_{\lceil Ns_j \rceil}) - \mathbb{1}_{(\lfloor Ns_{j-1} \rfloor / N, s_{j-1}]}(U_{\lceil Ns_{j-1} \rceil}) \ . \tag{116}$$

Thus, by Lemma B.0.12,

$$O^{(\sigma(j))} \leq \lceil Ns_j \rceil - \lfloor Ns_{j-1} \rfloor < N(s_j - s_{j-1}) + 2 = N\tilde{w}^{(\sigma(j))} + 2 \ .$$

To prove the tightness of the upper bound in the case of stratified resampling, consider, for any $\epsilon \in (0, 1/3)$, the normalised weight vector

$$\tilde{w}^{(1:3)} = (1/3 - \epsilon, 1/3 + 2\epsilon, 1/3 - \epsilon) \ ,$$

and suppose $u_1 \in (1/3 - \epsilon, 1/3]$, and $u_3 \in (2/3, 2/3 + \epsilon]$. Then, assuming σ is the identity permutation, $o_2 = 3$, yet $3\tilde{w}_2 = 1 + 6\epsilon$. Thus,

$$\lim_{\epsilon \downarrow 0} (o_2 - 3\tilde{w}_2) = 2 \ .$$

Similarly, to demonstrate the tightness of the lower bound, consider, for any $\epsilon \in (0, 1/2)$, the normalised weight vector

$$\tilde{w}^{(1:3)} = (\epsilon, 1 - 2\epsilon, \epsilon) \ ,$$

and suppose $u_1 \in [0, \epsilon)$, $u_3 \in (1-\epsilon, 1]$. Then, assuming σ is the identity permutation, $o_2 = 1$, yet $3\tilde{w}_2 = 3 - 6\epsilon$. Thus,

$$\lim_{\epsilon \downarrow 0} (o_2 - 3\tilde{w}_2) = -2 \ .$$

To obtain the bounds for systematic resampling, consider, again, Equation (115). Firstly, suppose that

$$U_1 + \frac{\lfloor Ns_{j-1} \rfloor}{N} = U_{\lceil Ns_{j-1} \rceil} \in \left(\frac{\lfloor Ns_{j-1} \rfloor}{N}, s_{j-1} \right] \ ,$$

and $\tilde{w}^{(\sigma(j))} \geq (\lfloor Ns_j \rfloor - \lfloor Ns_{j-1} \rfloor)/N$. Then,

$$s_{j-1} - \lfloor Ns_{j-1} \rfloor / N \leq s_j - \lfloor Ns_j \rfloor / N \ ,$$

and, therefore,

$$U_1 + \frac{\lfloor Ns_j \rfloor}{N} = U_{\lceil Ns_j \rceil} \in \left(\frac{\lfloor Ns_j \rfloor}{N}, s_j \right] \ .$$

Hence, $O^{(\sigma(j))} = \lfloor Ns_j \rfloor - \lceil Ns_{j-1} \rceil + 1$, and

$$O^{(\sigma(j))} - N\tilde{w}^{\sigma(j)} \leq \lfloor Ns_j \rfloor - \lceil Ns_{j-1} \rceil + 1 - \lfloor Ns_j \rfloor + \lfloor Ns_{j-1} \rfloor = 0 \ .$$

Suppose, on the other hand, that $\tilde{w}^{(\sigma(j))} < (\lfloor Ns_j \rfloor - \lfloor Ns_{j-1} \rfloor)/N$. Then,

$$O^{(\sigma(j))} - N\tilde{w}^{(\sigma(j))} > \lfloor Ns_j \rfloor - \lceil Ns_{j-1} \rceil - \lfloor Ns_j \rfloor + \lfloor Ns_{j-1} \rfloor = -1 \ .$$

Secondly, suppose that,

$$U_1 + \frac{\lfloor Ns_{j-1} \rfloor}{N} = U_{\lceil Ns_{j-1} \rceil} \in \left(\frac{s_{j-1}, \lceil Ns_{j-1} \rceil}{N} \right] \ ,$$

and $\tilde{w}^{(\sigma(j))} \leq (\lfloor Ns_j \rfloor - \lfloor Ns_{j-1} \rfloor)/N$. Then,

$$s_j - \lfloor Ns_j \rfloor / N \leq s_{j-1} - \lfloor Ns_{j-1} \rfloor / N \ ,$$

and, therefore,

$$U_1 + \frac{\lfloor Ns_j \rfloor}{N} = U_{\lceil Ns_j \rceil} \in \left(s_j, \frac{\lceil Ns_j \rceil}{N} \right] \ .$$

Hence, $O^{(\sigma(j))} = \lfloor Ns_j \rfloor - \lceil Ns_{j-1} \rceil + 1$, and

$$O^{(\sigma(j))} - N\tilde{w}^{(\sigma(j))} \geq \lfloor Ns_j \rfloor - \lceil Ns_{j-1} \rceil + 1 - \lfloor Ns_j \rfloor + \lfloor Ns_{j-1} \rfloor = 0 \, .$$

Suppose, on the other hand, that $\tilde{w}^{(\sigma(j))} > (\lfloor Ns_j \rfloor - \lfloor Ns_{j-1} \rfloor)/N$. Then,

$$O^{(\sigma(j))} - N\tilde{w}^{(\sigma(j))} < \lfloor Ns_j \rfloor - \lceil Ns_{j-1} \rceil + 2 - \lfloor Ns_j \rfloor + \lfloor Ns_{j-1} \rfloor = 1 \, .$$

To prove the tightness of the upper bound, consider, again, for any $\epsilon \in (0, 1/3)$, the normalised weight vector

$$\tilde{w}^{(1:3)} = (1/3 - \epsilon, 1/3 + 2\epsilon, 1/3 - \epsilon) \, .$$

Suppose that $u_1 \in (1/3 - \epsilon, 1/3]$. Then, $u_3 \in (1 - \epsilon, 1]$. Hence, assuming σ is the identity permutation, $o^{(2)} = 2$, yet $3\tilde{w}^{(2)} = 1 + 6\epsilon$. Therefore,

$$\lim_{\epsilon \downarrow 0} (o^{(2)} - 3\tilde{w}^{(2)}) = 1 \, .$$

Similarly, to demonstrate the tightness of the lower bound, consider, again, for any $\epsilon \in (0, 1/6)$, the normalised weight vector

$$\tilde{w}^{(1:3)} = (\epsilon, 1 - 2\epsilon, \epsilon) \, .$$

Suppose that $u_1 \in (0, \epsilon)$. Then, $u_3 \in (2/3, 2/3 + \epsilon)$. Note that, $2/3 + \epsilon < 2/3 + 1/6 = 5/6 = 1 - \epsilon$. Therefore, assuming σ is the identity permutation, $o^{(2)} = 2$, yet $3\tilde{w}^{(2)} = 3 - 6\epsilon$. Hence,

$$\lim_{\epsilon \downarrow 0} (o^{(2)} - 3\tilde{w}^{(2)}) = -1 \, .$$

As the weights are shuffled at the start of both the stratified and systematic resampling procedures, Equation (45) holds for these procedures. The offspring in multinomial resampling, $O^{(1:N)}$, ultimately depend on the terms

$$\mathbb{1}_{(s_{j-1}(\tilde{w}^{(1:N)}), s_j(\tilde{w}^{(1:N)})]}(U_i)$$

for any $(i, j) \in \{1, \ldots, N\}^2$. Hence $\bar{\kappa}(o^{(1:N)} | \tilde{w}^{(1:N)})$ ultimately depends on

$$\mathbb{P}(U_i \in (s_{j-1}(\tilde{w}^{(1:N)}), s_j(\tilde{w}^{(1:N)})])$$

for any $(i, j) \in \{1, \ldots, N\}^2$. This term is equal to $\mathbb{P}(U_i \in (0, \tilde{w}^{(j)}]))$ since $U_i \sim \text{Unif}(0, 1)$. Hence, the order of weights does not matter, and (45) holds.

A.7 PROOF OF THEOREM 2.4.4

By definition, $O^{(j)} = \lfloor N\tilde{w}^{(j)} \rfloor + O_r^{(j)}$, and

$$\tilde{w}_r^{(j)} = \frac{\tilde{w}^{(j)} - O_r^{(j)}/N}{\sum_{j=1}^{N} (\tilde{w}^{(j)} - O_r^{(j)}/N)} = \frac{N\tilde{w}^{(j)} - \lfloor N\tilde{w}^{(j)} \rfloor}{N - S} \, .$$

Moreover, by assumption, $\mathbb{E}(O_r^{(j)}) = (N - S)\tilde{w}_r^{(j)}$. Therefore, $\mathbb{E}(O^{(j)}) = N\tilde{w}^{(j)}$. Furthermore, $O^{(j)} \geq \lfloor N\tilde{w}^{(j)} \rfloor > N\tilde{w}^{(j)} - 1$, and,

$$|O^{(j)} - N\tilde{w}^{(j)}| = |O^{(j)} - \lfloor N\tilde{w}^{(j)} \rfloor - (N\tilde{w}^{(j)} - \lfloor N\tilde{w}^{(j)} \rfloor)| = |O_r^{(j)} - (N - S)\tilde{w}_r^{(j)}| \, .$$

Hence, the bounds obtained for systematic resampling in Theorem 2.4.3 hold for systematic residual resampling. Moreover, for stratified resampling, by Theorem 2.4.3,

$$O^{(j)} - N\tilde{w}^{(j)} < |O_r^{(j)} - (N - S)\tilde{w}_r^{(j)}| < 2 \, .$$

To obtain the general bound of $O^{(j)} - N\tilde{w}^{(j)} < N - 1$ suppose, for a contradiction, that $O^{(j)} - N\tilde{w}^{(j)} \geq N - 1$. Without loss of generality, it can be assumed that $j = 1$. If $\tilde{w}^{(1)} = 0$, then $O^{(1)} = 0$. Hence, $O^{(1)} = N$ and $\tilde{w}^{(1)} \leq 1/N$. However, if this is the case, then $\tilde{w}^{(2)} + \ldots + \tilde{w}^{(N)} = 1 - \tilde{w}^{(1)} \geq 1 - 1/N$. Therefore, there must exist at least

one $j \neq 1$ such that $\tilde{w}^{(j)} \geq 1/N$; for, if not, then $\tilde{w}^{(2)} + \ldots + \tilde{w}^{(N)} < (N-1)/N = 1 - 1/N$. Hence there exists a $j \neq 1$ such that $O^{(j)} \geq 1$ thus contradicting the assumption that $O^{(1)} = N$. Thus $O^{(j)} - N\tilde{w}^{(j)} < N - 1$. To prove the tightness of the lower bound for multinomial, stratified, and systematic residual resampling, consider, for any $\epsilon \in (0, 1/2)$, the normalised weight vector

$$\tilde{w}^{(1:2)} = (1-\epsilon, \epsilon) \, .$$

The residual normalised weight vector is thus

$$\tilde{w}_r^{(1:2)} = (1-2\epsilon, 2\epsilon) \, .$$

Suppose $u_1 \in (1-2\epsilon, 1)$. Then, assuming σ is the identity permutation, $o^{(1)} = 1$, yet $2\tilde{w}^{(1)} = 2 - 2\epsilon$. Therefore,

$$\lim_{\epsilon \downarrow 0} (o^{(1)} - 2\tilde{w}^{(1)}) = -1 \, .$$

To demonstrate the tightness of the upper bound for multinomial residual resampling, consider, for any $\epsilon \in (0, 1 - 1/N)$, the normalised weight vector

$$\tilde{w}^{(1:N)} = (1/N + \epsilon, 1/N - \epsilon/(N-1), 1/N - \epsilon/(N-1), \ldots, 1/N - \epsilon/(N-1)) \, .$$

The residual normalised weight vector is thus

$$\tilde{w}_r^{(1:N)} = (N\epsilon/(N-1), 1/(N-1) - N\epsilon/(N-1)^2, 1/(N-1) - N\epsilon/(N-1)^2, \ldots, 1/(N-1) - N\epsilon/(N-1)^2) \, .$$

Suppose $u_{1:N-1} \in (0, N\epsilon/(N-1))$. Then $o^{(1)} = N$, yet $N\tilde{w}^{(1)} = 1 + N\epsilon$. Therefore,

$$\lim_{\epsilon \downarrow 0} (o^{(1)} - N\tilde{w}^{(1)}) = N - 1 \, .$$

The prove the tightness of the upper bound for stratified residual resampling, consider, as in Theorem 2.4.3, for any $\epsilon \in (0, 1/3)$, the normalised weight vector

$$\tilde{w}^{(1:3)} = (1/3 - \epsilon, 1/3 + 2\epsilon, 1/3 - \epsilon) \, .$$

The residual normalised weight vector is thus

$$\tilde{w}_r^{(1:3)} = (1/2 - 3\epsilon/2, 3\epsilon, 1/2 - 3\epsilon/2) \, .$$

Suppose $u_1 \in (1/2 - 3\epsilon/2, 1/2]$ and $u_2 \in (1/2, 1/2 + 3\epsilon/2)$. Then, assuming σ is the identity permutation, $o^2 = 3$, yet $3\tilde{w}^{(2)} = 1 + 6\epsilon$. Therefore,

$$\lim_{\epsilon \downarrow 0} (o^{(2)} - 3\tilde{w}^{(2)}) = 2 \, .$$

The tightness of the upper bound for residual systematic resampling follows via the same argument, except, now, since $u_1 \in (1/2 - 3\epsilon/2, 1/2]$, then $u_2 \in (1 - 3\epsilon/2, 1]$ so that, assuming σ is the permutation identity, $o^{(2)} = 2$, and

$$\lim_{\epsilon \downarrow 0} (o^{(2)} - 3\tilde{w}^{(2)}) = 1 \, .$$

Equation (41) trivially holds since the only non-deterministic component of the residual resampling procedure happens during the resampling of the residuals and this is exchangeable by assumption.

A.8 PROOF OF LEMMA 3.2.1

Define the *generator*, G_t, as the solution to

$$\frac{\mathrm{d}G_t}{\mathrm{d}t} = J(\eta_t, t) G_t \, , \quad G_0 = I \, ,$$

over the interval $[0, T]$. Consider the process $G_t^{-1}\hat{R}_t$ which satisfies

$$\begin{aligned} \mathrm{d}(G_t^{-1}\hat{R}_t) &= \mathrm{d}G_t^{-1}\hat{R}_t + G_t^{-1}\mathrm{d}\hat{R}_t \\ &= -G_t^{-1}\mathrm{d}G_t G_t^{-1}\hat{R}_t + G_t^{-1}J(\eta_t, t)\hat{R}_t \mathrm{d}t + G_t^{-1}\sigma(\eta_t, t)\mathrm{d}B_t \\ &= G_t^{-1}\sigma(\eta_t, t)\mathrm{d}B_t \, . \end{aligned}$$

Therefore, for any $0 \leq s \leq t \leq T$, $G_t^{-1}\hat{R}_t$ is normally distributed with

$$\mathbb{E}(G_t^{-1}\hat{R}_t) = 0 \quad , \quad \mathrm{Cov}(G_s^{-1}\hat{R}_s, G_t^{-1}\hat{R}_t) = \int_0^s G_u^{-1}\zeta(\eta_u, u)G_u^{-*}\,\mathrm{d}u \;,$$

where G^{-*} is shorthand for $(G^{-1})^*$, and A^* denotes the transpose of the matrix A. Let ψ_t be the solution to

$$\frac{\mathrm{d}\psi_t}{\mathrm{d}t} = G_t^{-1}\zeta(\eta_t, t)G_t^{-*} \;, \quad \psi_0 = 0 \;, \tag{117}$$

over the interval $[0, T]$. Then

$$\begin{bmatrix}\hat{R}_t \\ Y\end{bmatrix} \sim \mathrm{N}\left(\begin{bmatrix}0 \\ P\eta_T\end{bmatrix}, \begin{bmatrix}G_t\psi_t G_t^* & G_t\psi_t G_T^* P^* \\ PG_T\psi_t G_t^* & PG_T\psi_T G_T^* P^* + V\end{bmatrix}\right) .$$

Therefore, by Lemma B.0.8,

$$\mathbb{E}(\hat{R}_t | Y = y) = G_t\psi_t G_T^* P^*(PG_T\psi_T G_T^* P^* + V)^{-1}(y - P\eta_T) \;.$$

To circumvent the need to calculate ψ_t, and, therefore, avoid solving the costly ODE (117) which contains inverses on the right-hand side, we let $\phi_t := G_t\psi_t G_t^*$ and note that ϕ_t solves

$$\begin{aligned}\frac{\mathrm{d}\phi_t}{\mathrm{d}t} &= \frac{\mathrm{d}G_t}{\mathrm{d}t}\psi_t G_t^* + G_t\psi_t\frac{\mathrm{d}G_t^*}{\mathrm{d}t} + G_t\frac{\mathrm{d}\psi_t}{\mathrm{d}t}G_t^* \\ &= J(\eta_t, t)G_t\psi_t G_t^* + G_t\psi_t G_t^* J(\eta_t, t)^* + \zeta(\eta_t, t) \\ &= J(\eta_t, t)\phi_t + \phi_t J(\eta_t, t)^* + \zeta(\eta_t, t) \;,\end{aligned}$$

over the interval $[0, T]$ with initial condition $\phi_0 = 0$.

A.9 PROOF OF LEMMA 3.2.2

Let η_s be defined as the solution to

$$\frac{\mathrm{d}\eta_s}{\mathrm{d}s} = a(s) + b(s)\eta_s \;, \quad \eta_0 = x_t \;,$$

over the interval $[0, T-t]$, and define $Z_s := \tilde{X}_s - \eta_s$. Then, Z_s is a diffusion satisfying the SDE

$$\mathrm{d}Z_s = b(s)Z_s\,\mathrm{d}s + \sigma(s)\,\mathrm{d}W_s \;, \quad Z_0 = 0 \;.$$

Next, let G_s be defined as the solution to

$$\frac{\mathrm{d}G_s}{\mathrm{d}s} = b(s)G_s \;, \quad G_0 = I \;,$$

over the interval $[0, T-t]$. Consider $G_s^{-1}Z_s$ which satisfies $\mathrm{d}(G_s^{-1}Z_s) = G_s^{-1}\sigma(s)\,\mathrm{d}W_s$. Thus, $G_s^{-1}Z_s$ is normally distributed for any time $s \in [0, T-t]$, with $\mathbb{E}[G_s^{-1}Z_s] = 0$ for any $s \in [0, T-t]$ and

$$\mathrm{Cov}(G_u^{-1}Z_u, G_s^{-1}Z_s) = \int_0^u G_v^{-1}\sigma(v)\sigma(v)^* G_v^{-*}\,\mathrm{d}v$$

for any $0 \leq u < s \leq T - t$. Thus, letting ψ_s be defined as the solution to

$$\frac{\mathrm{d}\psi_s}{\mathrm{d}s} = G_s^{-1}\sigma(s)\sigma(s)^* G_s^{-*} \;, \quad \psi_0 = 0 \;,$$

over the interval $[0, T-t]$,

$$(Z_{T-t} | Z_0 = 0) \sim \mathrm{N}(0, G_{T-t}\psi_{T-t}G_{T-t}^*) \;.$$

Therefore,

$$(\tilde{X}_{T-t} | \tilde{X}_0 = x_t) \sim \mathrm{N}(\eta_{T-t}, G_{T-t}\psi_{T-t}G_{T-t}^*) \;.$$

Thus, since $(Y|\tilde{X}_{T-t} = \tilde{x}_{T-t}) \sim \mathrm{N}(P\tilde{x}_{T-t}, V)$,

$$(Y|\tilde{X}_0 = x_t) \sim \mathrm{N}(P\eta_{T-t}, PG_{T-t}\psi_{T-t}G_{T-t}^*P^* + V) \,.$$

Now, as in A.8, letting $\phi_s := G_s\psi_s G_s^*$ and noting that ϕ_s solves

$$\frac{\mathrm{d}\phi_s}{\mathrm{d}s} = b(s)\phi_s + \phi_s b(s)^* + \sigma(s)\sigma(s)^* \,, \quad \phi_0 = 0$$

over the interval $[0, T-t]$,

$$(Y|\tilde{X}_0 = x_t) \sim \mathrm{N}(P\eta_{T-t}, P\phi_{T-t}P^* + V) \,.$$

A.10 PROOF OF LEMMA 4.2.4

Consider

$$\begin{aligned}
&\frac{w_T(\tilde{x}_T^{(k)};\theta)\Psi(x_{0:T}^{(1:N)}, a_{0:T-1}^{(1:N)}|\theta)}{\psi(x_{0:T}^{(1:N)}\backslash\tilde{x}_T^{(k)}, a_{0:T-1}^{(1:N)}\backslash\tilde{a}_{T-1}^{(k)}|k, \tilde{x}_T^{(k)}, \tilde{a}_{T-1}^{(k)}, \theta)} \\
&= w_T(\tilde{x}_T^{(k)};\theta)p_0(x_0^{(\mathcal{L}_T(k,0))}|\theta)\prod_{i=1}^{T}\mathbb{P}(A_{t-1}^{(\mathcal{L}_T^{(k,t)})} = a_{t-1}^{(\mathcal{L}_T(k,t))}|\tilde{w}_{t-1}^{(1:N)})p_t(x_t^{(\mathcal{L}_T(k,t))}|\tilde{x}_{t-1}^{(\mathcal{L}_T(k,t-1))}, \theta) \,.
\end{aligned}$$

By property property (P) of Assumptions 4.2.1, and the definition of the Lineage function $\mathcal{L}_T$ (Definition 4.2.2),

$$\mathbb{P}(A_{t-1}^{(\mathcal{L}_T^{(k,t)})} = a_{t-1}^{(\mathcal{L}_T(k,t))}|\tilde{w}_{t-1}^{(1:N)}) = \tilde{w}_{t-1}^{(\mathcal{L}_T(k,t-1))} \,.$$

Moreover,

$$w_0(\tilde{x}_0^{(\mathcal{L}_T(k,0))};\theta)p_0(\tilde{x}_0^{(\mathcal{L}_T(k,0))}|\theta) = \gamma_0(\theta, \tilde{x}_0^{(\mathcal{L}_T(k,0))}) \,,$$

and, for any $t \in \{1, \ldots, T\}$,

$$w_t(\tilde{x}^{(\mathcal{L}_T(k,t))};\theta)p_t(x_t^{(\mathcal{L}_T(k,t))}|\tilde{x}_{t-1}^{(\mathcal{L}_T(k,t-1))}, \theta) = \frac{\gamma_t(\theta, \tilde{x}_t^{(\mathcal{L}_T(k,t))})}{\gamma_{t-1}(\theta, \tilde{x}_{t-1}^{(\mathcal{L}_T(k,t-1))})} \,.$$

Therefore, dropping the explicit dependence of the weights on θ,

$$\begin{aligned}
&\frac{w_T(\tilde{x}_T^{(k)};\theta)\Psi(x_{0:T}^{(1:N)}, a_{0:T-1}^{(1:N)}|\theta)}{\psi(x_{0:T}^{(1:N)}\backslash\tilde{x}_T^{(k)}, a_{0:T-1}^{(1:N)}\backslash\tilde{a}_{T-1}^{(k)}|k, \tilde{x}_T^{(k)}, \tilde{a}_{T-1}^{(k)}, \theta)} \\
&= \frac{\gamma_0(\theta, \tilde{x}_0^{(\mathcal{L}_T(k,0))})}{w_0(\tilde{x}_0^{(1)}) + \ldots + w_0(\tilde{x}_0^{(N)})}\prod_{t=1}^{T-1}\frac{\gamma_t(\theta, \tilde{x}_t^{(\mathcal{L}_T(k,t))})}{\gamma_{t-1}(\theta, \tilde{x}_{t-1}^{(\mathcal{L}_T(k,t-1))})}\frac{1}{w_t(\tilde{x}_t^{(1)}) + \ldots + w_t(\tilde{x}_t^{(N)})} \\
&\quad \times \frac{\gamma_T(\theta, \tilde{x}_T^{(k)})}{\gamma_{T-1}(\theta, \tilde{x}_{T-1}^{(\mathcal{L}_T(k,T-1))})} \\
&= \gamma_T(\theta, \tilde{x}_T^{(k)})\Big(\prod_{t=0}^{T-1}\sum_{i=1}^{N} w_t(\tilde{x}_t^{(i)})\Big)^{-1} \\
&= N^{-(T+1)}\frac{\gamma_T(\theta, \tilde{x}_T^{(k)})}{I_T(\theta, \tilde{x}_{0:T}^{(1:N)})}\sum_{i=1}^{N} w_T(\tilde{x}_T^{(k)};\theta) \,.
\end{aligned}$$

Rearranging gives the required result.

A.11 PROOF OF THEOREM 4.3.2

For any $k \in \{1, \ldots, N\}$, define

$$P_k(x, B) := \int_{B\times\mathbb{R}^{d\times(N-1)}}\cdots\int \tilde{q}_N(y_{1:N}|x)\alpha_{k,N}(x, y_{1:N})\,\mathrm{d}y_k\mathrm{d}y_{-k} \,,$$

and

$$Q_k(A,B) := \int_A \pi(\mathrm{d}x) P_k(x,B) \ .$$

Q_k is symmetric since the joint proposal, $q_N(x, y_{1:N})$, is an exchangeable density and, therefore,

$$\begin{aligned}
&Q_k(A,B) \\
&= \frac{1}{\gamma(\mathbb{R}^d)} \int_{A\times B\times\mathbb{R}^{d\times(N-1)}} \cdots \int q_0(x)\tilde{q}_N(y_{1:N}|x)w(x)\alpha_{k,N}(x,y_{1:N}) \ \mathrm{d}x\mathrm{d}y_k\mathrm{d}y_{-k} \\
&= \frac{1}{\gamma(\mathbb{R}^d)} \int_{A\times B\times\mathbb{R}^{d\times(N-1)}} \cdots \int q_N(x,y_{1:N})w(y_k)\alpha_{k,N}(y_k,y_{1:k-1},x,y_{k+1:N}) \ \mathrm{d}x\mathrm{d}y_k\mathrm{d}y_{-k} \\
&= \frac{1}{\gamma(\mathbb{R}^d)} \int_{A\times B\times\mathbb{R}^{d\times(N-1)}} \cdots \int q_N(y_k,y_{1:k-1},x,y_{k+1:N})w(y_k)\alpha_{k,N}(y_k,y_{1:k-1},x,y_{k+1:N}) \ \mathrm{d}x\mathrm{d}y_k\mathrm{d}y_{-k} \\
&= \int_B \pi(\mathrm{d}y_k) P_k(y_k,A) \ .
\end{aligned}$$

Moreover, define

$$P_R(x,B) := \delta_x(B) \int_{\mathbb{R}^{d\times N}} \cdots \int \tilde{q}_N(y_{1:N}|x)\left(1 - \sum_{k=1}^N \alpha_{k,N}(x,y_{1:N})\right) \mathrm{d}y_{1:N} \ ,$$

and

$$Q_R(A,B) := \int_A \pi(\mathrm{d}x) P_R(x,B) \ .$$

Q_R is symmetric since

$$Q_R(A,B) = \int_{(A\cap B)\times\mathbb{R}^{d\times N}} \cdots \int \pi(x)\tilde{q}_N(y_{1:N}|x)\left(1 - \sum_{k=1}^N \alpha_{k,N}(x,y_{1:N})\right) \mathrm{d}x \ \mathrm{d}y_{1:N} \ .$$

Thus,

$$Q(A,B) = \int_A \pi(\mathrm{d}x) P(x,B) = \sum_{k=1}^N Q_k(A,B) + Q_R(A,B) \ ,$$

is symmetric. Hence, X_t is reversible with respect to π. Next consider, for any set $A \subseteq \mathbb{R}^d$,

$$P(x,A) \geq \sum_{k=1}^N P_k(x,A) \geq \beta \sum_{k=1}^N \int_{A^N} \cdots \int \tilde{q}_N(y_{1:N}|x)\alpha_{k,N}^m(x,y_{1:N}) \ \mathrm{d}y_{1:N} \ , \quad (118)$$

where $\beta = 1$ for the Metropolis-Hastings independence sampler, and $\beta = 1/2$ for Barker's independence sampler. Let A be such that $\pi(A) > 0$. Then, there exists an $n_A \in \mathbb{N}$ such that $\pi(A \cap B_{n_A}) > 0$ where B_n is the closed ball of radius n centred on 0. Let $C_{A,x} := A \cap B_{n_A} \backslash \{x\}$. Then

$$P(x,A) \geq \frac{1}{2} \sum_{k=1}^N \int_{C_{A,x}^N} \tilde{q}_N(y_{1:N}|x)\alpha_{k,N}^m(x,y_{1:N}) \ \mathrm{d}y_{1:N} \ .$$

For any $I \subseteq \{1,\ldots,N\}$, define

$$\Omega_I := \{y_{1:N} \in C_{A,x}^N : w(x) \geq w(y_i) \text{ for every } i \in I \text{ and } w(y_i) \geq w(x) \text{ for any } i \notin I.\} \ ,$$

and let $\Gamma := \{y \in C_{A,x} : w(y) \geq w(x)\}$. Moreover, for brevity, let $\mathcal{P}_N := \mathcal{P}(\{1, \ldots, N\})$. Then

$$P(x, A) \geq \frac{1}{2} \sum_{I \in \mathcal{P}_N} \iint_{C_{A,x}^N \cap \Omega_I} \tilde{q}_N(y_{1:N}|x) \sum_{k=1}^N \alpha_{k,N}^m(x, y_{1:N}) \, \mathrm{d}y_{1:N} \, .$$

Now, by Lemma B.0.3,

$$\sum_{k=1}^N \alpha_{k,N}^m(x, y_{1:N}) \geq \begin{cases} \dfrac{1}{|I| w(x)} \displaystyle\sum_{i \in I} w(y_i) & \text{if } I \neq \emptyset \\ 1 & \text{if } I = \emptyset \end{cases} ,$$

By assumption, q_0 is positive on $\{y \in \mathbb{R}^d : \gamma(y) > 0\}$. Hence, by (88), $\tilde{q}_N(\cdot|x)$ is positive on $\{y \in \mathbb{R}^d : \gamma(y) > 0\}^N$ for any $x \in \mathbb{R}^d$. Hence, there exists an $\eta_0 > 0$ such that

$$\inf_{y_{1:N} \in C_{A,x}^N} \tilde{q}_N(y_{1:N}|x) \geq \eta_0 > 0 \, .$$

Thus

$$\iint_{C_{A,x}^N \cap \Omega_\emptyset} \tilde{q}_N(y_{1:N}|x) \, \mathrm{d}y_{1:N} \geq \eta_0 \mathrm{Leb}(C_{A,x} \cap \Gamma)^N \, .$$

Moreover, for any $I \neq \emptyset$,

$$\begin{aligned} &\iint_{C_{A,x}^N \cap \Omega_I} \tilde{q}_N(y_{1:N}|x) \sum_{k=1}^N \alpha_{k,N}^m(x, y_{1:N}) \, \mathrm{d}y_{1:N} \\ &\geq \frac{q_0(x)}{|I| \gamma(x)} \iint_{C_{A,x}^N \cap \Omega_I} \tilde{q}_N(y_{1:N}|x) \sum_{j \in I} \frac{\pi(y_j)}{q(y_j)} \, \mathrm{d}y_{1:N} \\ &\geq \frac{1}{|I| \gamma(x)} \sum_{j \in I} \iint_{C_{A,x}^N \cap \Omega_I} \tilde{q}_N(y_{1:j-1}, x, y_{j+1:N}|y_j) \pi(y_j) \, \mathrm{d}y_{1:N} \, , \end{aligned}$$

where the last line follows since the joint density, q_N, is symmetric. By assumption, q_0 is continuous on $\mathbb{R}^d$. Therefore, by Lemma 4.3.1, $\tilde{q}_N(\cdot|\cdot)$ is continuous on $\mathbb{R}^{d \times (N+1)}$. Moreover, by assumption,

$$\{y \in \mathbb{R}^d : \gamma(y) > 0\} \subseteq \{y \in \mathbb{R}^d : q_0(y) > 0\} \, .$$

Thus, by Lemma 4.3.1, $\tilde{q}_N(y_{1:N}|x) > 0$ for any $x \in \mathbb{R}^d$ such that $\gamma(x) > 0$, and any $y_{1:N} \in A^N$. Hence, for any $j \in \{1, \ldots, N\}$, there exists an $\eta_j > 0$ such that

$$\inf_{y_{1:N} \in C_{A,x}^N} \tilde{q}_N(y_{1:j-1}, x, y_{j+1:N}|y_j) \geq \eta_j > 0 \, .$$

Therefore,

$$\begin{aligned} &\iint_{C_{A,x}^N \cap \Omega_I} \tilde{q}_N(y_{1:N}|x) \sum_{k=1}^N \alpha_{k,N}^m(x, y_{1:N}) \, \mathrm{d}y_{1:N} \\ &\geq \frac{1}{|I| \gamma(x)} \sum_{j \in I} \eta_j \pi(C_{A,x} \cap \Gamma^c) \mathrm{Leb}(C_{A,x} \cap \Gamma^c)^{|I|-1} \mathrm{Leb}(C_{A,x} \cap \Gamma)^{N-|I|} \, . \end{aligned}$$

Hence

$$\begin{aligned} P(x, A) \geq &\frac{1}{2} \eta_0 \mathrm{Leb}(C_{A,x} \cap \Gamma)^N \\ &+ \frac{1}{2|I| \gamma(x)} \sum_{I \in P_N \setminus \{\emptyset\}} \sum_{j \in I} \eta_j \pi(C_{A,x} \cap \Gamma^c) \mathrm{Leb}(C_{A,x} \cap \Gamma^c)^{|I|-1} \mathrm{Leb}(C_{A,x} \cap \Gamma)^{N-|I|} \, . \end{aligned}$$

To demonstrate that this is positive assume, for a contradiction, that $\mathrm{Leb}(C_{A,x} \cap \Gamma) = 0$, and, either $\pi(C_{A,x} \cap \Gamma^c) = 0$, or $\mathrm{Leb}(C_{A,x} \cap \Gamma^c) = 0$, where Leb denotes Lebesgue measure. π is absolutely continuous with respect to Lebesgue measure, hence this assumption implies that $\pi(C_{A,x}) = 0$, thus contradicting the definition of B_{n_A}. Hence $P(x, A) > 0$. Next, to show that the Exchangeable Sampler is non-negative, consider

$$\langle Pf, f\rangle \geq \sum_{k=1}^{N} \int_{\mathbb{R}^{d\times(N+1)}} \tilde{q}_N(y_{1:N}|x)\pi(x)\alpha_{k,N}(x, y_{1:N})f(x)f(y_k)\,\mathrm{d}x\mathrm{d}y_{1:N} .$$

For any $z_{1:N} \in \mathbb{R}^{d\times N}$, define $\xi : \mathbb{R}^{d\times N} \to \mathbb{R}$ by

$$\xi(z_{1:N}) := [w(z_1) + \ldots + w(z_N)]^{-1} .$$

By a multiple-proposal extension of Inequality (28), and Lemma B.0.1,

$$\begin{aligned}\alpha_{k,N}(x, y_{1:N}) &\geq \frac{1}{2}w(y_k)[\xi(y_{1:N}) \wedge \xi(y_{1:k-1}, x, y_{k+1:N})] \\ &= \frac{\gamma(y_k)}{2q(y_k)} \int_0^\infty \mathbb{1}_{[0,\xi(y_{1:N})]}(s)\mathbb{1}_{[0,\xi(y_{1:k-1},xy_{k+1:N})]}(s)\,\mathrm{d}s ,\end{aligned}$$

Therefore, defining

$$\psi_k(x, y_{1:N}, s) := \mathbb{1}_{[0,\xi(y_{1:N})]}(s)\mathbb{1}_{[0,\xi(y_{1:k-1},xy_{k+1:N})]}(s) ,$$

we have

$$\begin{aligned}\langle Pf, f\rangle &\geq \frac{1}{2\gamma(\mathbb{R}^d)} \sum_{k=1}^{N} \int_{\mathbb{R}^{d\times(N+1)}} \int_0^\infty \frac{\tilde{q}_N(y_{1:N}|x)}{q_0(y_k)}\gamma(x)f(x)\gamma(y_k)f(y_k)\psi_k(x, y_{1:N}, s)\,\mathrm{d}s\mathrm{d}x\mathrm{d}y_{1:N} \\ &\geq \frac{1}{2\gamma(\mathbb{R}^d)} \sum_{k=1}^{N} \int_{\mathbb{R}^{d\times(N+1)}} \int_0^\infty q_N(x, y_{1:N})w(x)f(x)w(y_k)f(y_k)\psi_k(x, y_{1:N}, s)\,\mathrm{d}s\mathrm{d}x\mathrm{d}y_{1:N} .\end{aligned}$$

Consider the joint density $q_N(x, y_{1:N})$. Let X be the random variable associated with x, Z_0 be the random variable associated with z_0, defined by Algorithm 18, Θ be the random variable associated with θ, also defined by Algorithm 18, and so on. X has density q_0. Thus $Z_0 = h^{-1}(X) \sim \mathrm{N}_d(0, I_d)$. Consider, then, the joint distribution of Θ and Z_i for any $i \in \{0, \ldots, N\}$, which is given by

$$\begin{bmatrix}\Theta \\ Z_i\end{bmatrix} \sim \mathrm{N}\left(\begin{bmatrix}0 \\ 0\end{bmatrix}, \begin{bmatrix}I_d & \sqrt{1-\delta^2}I_d \\ \sqrt{1-\delta^2}I_d & I_d\end{bmatrix}\right) .$$

Moreover, given θ, the sequence $Z_{0:N}$ is independent. Therefore, the joint density of $Z_{0:N}$, denoted $q_N^Z(z_{0:N})$, is given by

$$q_N^Z(z_{0:N}) = \int_{\mathbb{R}^d} \prod_{i=0}^{N} q_Z(z_i|\theta)q_\Theta(\theta)\,\mathrm{d}\theta ,$$

where $q_\Theta : \mathbb{R}^d \to (0, \infty)$, and $q_Z(\cdot|\theta) : \mathbb{R}^d \to (0, \infty)$ are densities defined by

$$\begin{aligned}q_\Theta(\theta) &\propto \exp(-\theta^T\theta/2) , \\ q_Z(z_i|\theta) &\propto \exp\left[-\frac{1}{2\delta^2}(z_i - \sqrt{1-\delta^2}\theta)^T(z_i - \sqrt{1-\delta^2}\theta)\right] .\end{aligned}$$

Therefore,

$$q(y_{0:N}) = \int_{\mathbb{R}^d} q_\Theta(\theta) \prod_{i=0}^{N} q_*(y_i|\theta)\,\mathrm{d}\theta ,$$

where

$$q_*(y_i|\theta) = q_Z(h^{-1}(y_i)|\theta)|\det J(y_i)| ,$$

and $\det J(y_i)$ denotes the determinant of the Jacobian matrix $J(y_i)$ which is the $d \times d$ matrix whose (j,k)-th entry is given by

$$J(y_i)_{jk} := \left.\frac{\partial y_{ij}}{\partial z_{ik}}\right|_{z_i = h^{-1}(y_i)} ,$$

where y_{ij} denotes the j-th element of y_i, and z_{ik} the k-th element of z_i. Hence,

$$q_N(y_{0:N}) = \int_{\mathbb{R}^d} \zeta_k(\theta, x, y_{-k})\zeta_k(\theta, y_k, y_{-k}) \, \mathrm{d}\theta ,$$

where

$$\zeta_k(\theta, w, y_{-k}) := q_*(w|\theta)\left[q_\Theta(\theta)\prod_{\substack{i=1\\ i\neq k}}^{N} q_*(y_i|\theta)\right]^{1/2} .$$

Thus,

$$\langle Pf, f\rangle \geq \frac{1}{2\gamma(\mathbb{R}^d)}\sum_{k=1}^{N}\int_{\mathbb{R}^{d\times(N-1)}}\int_0^\infty\int_{\mathbb{R}^d} g_k(\theta, s, y_{-k})^2 \, \mathrm{d}\theta\mathrm{d}s\mathrm{d}y_{-k} ,$$

where

$$g_k(\theta, s, y_{-k}) := \int_{\mathbb{R}^d} \zeta_k(\theta, t, y_{-k})w(t)f(t)\mathbb{1}_{[0,\xi(y_{1:k-1},t,y_{k+1:N})]}(s) \, \mathrm{d}t .$$

Therefore, $\langle Pf, f\rangle \geq 0$, and X_t is non-negative.

A.12 PROOF OF THEOREM 4.3.3

For any $\eta > 0$, define $v_\eta : \mathcal{X} \to [1, \infty)$ by $v_\eta(x) := (1 + p(x))^\eta$. Denote, by $P(x, \cdot)$, the transition distributions of the chain X_t and consider, for any $x \notin C$,

$$\begin{aligned}\frac{\mathbb{E}_{P(x,\cdot)}(v_\eta(Y))}{v_\eta(x)} &= \int_{\mathcal{X}} q(y|x)\alpha(x,y)\frac{v_\eta(y)}{v_\eta(x)} \, \mathrm{d}y + \int_{\mathcal{X}} q(y|x)(1-\alpha(x,y)) \, \mathrm{d}y \\ &= 1 - \int q(y|x)\alpha(x,y)\left(1 - \frac{v_\eta(y)}{v_\eta(x)}\right) \mathrm{d}y .\end{aligned}$$

Firstly, to demonstrate the drift condition off of the small set C, it is sufficient to show that there exists a $\zeta > 0$ and an $\eta > 0$ such that, for any $x \notin C$,

$$1 - \frac{\mathbb{E}_{P(x,\cdot)}(v_\eta(Y))}{v_\eta(x)} = \mathbb{E}\left\{\alpha(x, Y)\left[1 - \left(\frac{1 + p(Y)}{1 + p(x)}\right)^\eta\right]\right\} \geq \zeta .$$

Moreover, by assumption, for any $x \notin C$, $p(x) > \rho_*$. Hence, $(1 + p(y))/(1 + p(x)) < 1/\rho_* + p(y)/p(x)$, and it is sufficient to show that there exists a $\rho_* \in (1, \infty)$, a $\zeta > 0$, and an $\eta > 0$ such that, for all $x \notin C$,

$$\mathbb{E}\left\{\alpha(x, Y)\left[1 - \left(\frac{1}{\rho_*} + \frac{p(Y)}{p(x)}\right)^\eta\right]\right\} \geq \zeta . \tag{119}$$

Define

$$\begin{aligned} f_\eta(x, y) &:= \alpha(x, y)[1 - (1/\rho_* + p(y)/p(x))^\eta] \\ g(x, y) &:= -\alpha(x, y)\log(1/\rho_* + p(y)/p(x))\end{aligned}$$

Then, by Lemma B.0.4, for any $(x, y) \in \mathcal{X}^2$,

$$0 \geq f_\eta(x, y) - \eta g(x, y) \geq -\alpha(x, y)\frac{\eta^2}{2}\log\left(\frac{1}{\rho_*} + \frac{p(y)}{p(x)}\right)^2 . \tag{120}$$

By Assumption (IM), there exists a $\rho_* \in (1, \infty)$ and a $\delta > 0$ such that $\mathbb{E}[g(x, Y)] \geq \delta$ for any $x \notin C$. Hence, if there exists an $\eta > 0$ and a $\zeta > 0$ such that, for any $x \notin C$,

$$\mathbb{E}[f_\eta(x, Y) - \eta g(x, Y)] \geq \zeta - \eta\delta , \tag{121}$$

then

$$\mathbb{E}[f_\eta(x,Y)] \geq \zeta - \eta\delta + \eta\delta = \zeta > 0 \ ,$$

and (119) holds. Define $h_\eta(x,y) := f_\eta(x,y) - \eta g(x,y)$. Then, by Inequality (120), for any fixed $(x,y) \in \mathcal{X}^2$, $h_\eta(x,y) \uparrow 0$ as $\eta \downarrow 0$, and, although for any fixed η, the bound on h_η given by Inequality (120) could tend towards negative infinity as $p(y)/p(x)$ tends towards infinity (depending on the behaviour of α in the limit), Assumption (UI) ensures that there exists a positive $\tau < \infty$ such that this divergence is uniformly well-behaved on $\{z \in \mathcal{X} : \tau p(x) \leq p(z)\}$, thereby allowing us to demonstrate Inequality (121). Consider the disjoint sets

$$\begin{aligned}
\mathcal{X}_1(x) &:= \{y \in \mathcal{X} : 1/\rho_* + p(y)/p(x) \leq 1\} \ , \\
\mathcal{X}_2(x) &:= \{y \in \mathcal{X} : \tau p(x) \leq p(y)\} \cap \mathcal{X}_1^c(x) \ , \\
\mathcal{X}_3(x) &:= \{y \in \mathcal{X} : \tau p(x) > p(y)\} \cap \mathcal{X}_1^c(x) \ .
\end{aligned}$$

which cover $\mathcal{X}$. By definition, $1/\rho_* + p(y)/p(x) \leq 1$ for any $y \in \mathcal{X}_1(x)$. Hence, since $p(y)/p(x) > 0$,

$$\log\left(\frac{1}{\rho_*} + \frac{p(y)}{p(x)}\right)^2 \leq \log(\rho_*)^2 \ ,$$

for any $y \in \mathcal{X}_1(x)$. Therefore, by Inequality (120), and the fact that $\alpha \leq 1$,

$$\mathbb{E}[h_\eta(x,Y)\mathbb{1}_{\mathcal{X}_1(x)}(Y)] \geq -\frac{\eta^2}{2}\log(\rho_*)^2 \ .$$

Moreover, by Inequality (120), the fact that $\alpha \leq 1$, and the fact that, for any $y \in \mathcal{X}_3(x)$, $\tau p(x) > p(y)$ and $1/\rho_* + p(y)/p(x) > 1$,

$$\begin{aligned}
\mathbb{E}[h_\eta(x,Y)\mathbb{1}_{\mathcal{X}_3(x)}(Y)] &\geq -\frac{\eta^2}{2}\mathbb{E}\left[\alpha(x,Y)\log\left(\frac{1}{\rho_*} + \frac{p(Y)}{p(x)}\right)^2 \mathbb{1}_{\mathcal{X}_3(x)}(Y)\right] \\
&\geq -\frac{\eta^2}{2}\mathbb{E}[\alpha(x,Y)\log(\rho_*^{-1} + \tau)^2 \mathbb{1}_{\mathcal{X}_3(x)}(Y)] \\
&\geq -\frac{\eta^2}{2}\log(\rho_*^{-1} + \tau)^2 \ .
\end{aligned}$$

By definition, $\mathcal{X}_2(x) = \mathcal{P}_\tau(x) \cap \mathcal{X}_1^c(x)$, where $\mathcal{P}_\tau(x) = \{z \in \mathcal{X} : \tau p(x) \leq p(z)\}$. Hence, by Inequality (120), and the definition of μ_τ from Assumption (UI),

$$\begin{aligned}
\mathbb{E}[h_\eta(x,Y)\mathbb{1}_{\mathcal{X}_2(x)}(Y)] &\geq -\frac{\eta^2}{2}\mathbb{E}\left[\alpha(x,Y)\log\left(\frac{1}{\rho_*} + \frac{p(Y)}{p(x)}\right)^2 \mathbb{1}_{\mathcal{P}_\tau(x)\cap\mathcal{X}_1^c(x)}(Y)\right] \\
&\geq -\frac{\eta^2}{2}\mu_\tau \ .
\end{aligned}$$

Combining these three inequalities gives,

$$\mathbb{E}[h_\eta(x,Y)] \geq -\frac{\eta^2}{2}\left[\mu_\tau + \log(\rho_*^{-1} + \tau)^2 + \log(\rho_*)^2\right] \ , \tag{122}$$

which is true for any $x \notin C$. By Assumption (UI), $\mu_\tau < \infty$, and $\tau < \infty$, and, so, one can choose an η to be such that

$$0 < \eta < 1 \wedge 2\delta\left[\mu_\tau + \log(\rho_*^{-1} + \tau)^2 + \log(\rho_*)^2\right]^{-1} \ .$$

With this definition,

$$\frac{\eta^2}{2}\left[\mu_\tau + \log(\rho_*^{-1} + \tau)^2 + \log(\rho_*)^2\right] < \delta\eta \ . \tag{123}$$

Set

$$\zeta := \eta\delta - \frac{\eta^2}{2}\left[\mu_\tau + \log(\rho_*^{-1} + \tau)^2 + \log(\rho_*)^2\right] \ . \tag{124}$$

By (123), $\zeta > 0$, and, by (122),

$$\mathbb{E}[h_\eta(x, Y)] \geq \zeta - \eta\delta$$

for any $x \notin C$, thereby demonstrating (121), and demonstrating the drift condition off the small set. To demonstrate the drift condition on the small set, consider, for any $x \in C$,

$$\mathbb{E}_{P(x,\cdot)}(v_\eta(Y)) = \int_{\mathcal{X}} q(y|x)\alpha(x,y)(v_\eta(y) - v_\eta(x))\,\mathrm{d}y + \int_{\mathcal{X}} q(y|x)\alpha(x,y)v_\eta(x)\,\mathrm{d}y\,.$$

Since $x \in C$, then $p(x) \leq \rho_*$. Hence, since $\eta \in (0,1)$ by construction, $v_\eta(x) \leq (1+\rho_*)^\eta < (1+\rho_*)$, and the second integral is bounded by $(1+\rho_*)$. Furthermore, if $y \in \mathcal{Q}_x^c$ where $\mathcal{Q}_x := \{z \in \mathcal{X} : p(x) \leq p(z)\}$, then $v_\eta(y) - v_\eta(x) < 0$. Moreover, if $y \in \mathcal{Q}_x$, then, by Lemma B.0.5,

$$v_\eta(y) - v_\eta(x) < p(y) - p(x)\,.$$

Hence, by Assumption (B), the first integral is bounded by $\xi < \infty$, therefore demonstrating the drift condition on the small set.

A.13 PROOF OF THEOREM 4.3.4

Lemma 1.2 of Mengersen and Tweedie, 1996 demonstrates that any compact set is small. Let $p : \mathbb{R} \to [1,\infty)$ be defined by $p(x) := \exp(|x|)$. Then, $C := \{x \in \mathbb{R} : p(x) \leq \rho_*\}$, which is compact for any $\rho_* > 0$, is small for any $\rho_* > 0$. Let $x \in \mathbb{R}$ be such that $x \geq m_* > m_2 > 0$, where m_* is arbitrary. Consider the four disjoint sets

$$\begin{aligned}
\mathcal{X}_1 &:= (-\infty, -m_2]\,,\\
\mathcal{X}_2 &:= (-m_2, m_2)\,,\\
\mathcal{X}_3 &:= [m_2, x)\,,\\
\mathcal{X}_4 &:= [x, \infty)\,,
\end{aligned}$$

which cover $\mathbb{R}$. For the moment, let $\rho_* \in (1,\infty)$ be arbitrary and consider $y \in \mathcal{X}_4 = [x,\infty)$. Since π decays exponentially quickly in the tails, then $\pi(y)/\pi(x) \leq \exp(-\theta_2(y-x)) \leq 1$. Thus,

$$\begin{aligned}
&\mathbb{E}\left[\alpha(x,Y)\log\left(\frac{1}{\rho_*} + \frac{p(Y)}{p(x)}\right)\mathbb{1}_{\mathcal{X}_4}(Y)\right]\\
&\leq \mathbb{E}\{\exp[-\theta_2(Y-x)]\log(\rho_*^{-1} + \exp(Y-x))\mathbb{1}_{\mathcal{X}_4}(Y)\}\\
&= \mathbb{E}[\exp(-\theta_2\epsilon Z)\log(\rho_*^{-1} + \exp(\epsilon Z))\mathbb{1}_{[0,\infty)}(Z)]\,, \qquad (125)
\end{aligned}$$

where $Z \sim \mathrm{N}(0,1)$. Next, note that, if $y \in \mathcal{X}_3 = [m_2, x)$, then $x - y \in (0, x - m_2]$, and, since π decays exponentially in the tails, $\pi(y)/\pi(x) \geq \exp(\theta_2(x-y)) \geq 1$. Thus, $\alpha(x,y) = 1$, and

$$\begin{aligned}
&\mathbb{E}\left[\alpha(x,Y)\log\left(\frac{1}{\rho_*} + \frac{p(Y)}{p(x)}\right)\mathbb{1}_{\mathcal{X}_3}(Y)\right]\\
&= \mathbb{E}[\log(\rho_*^{-1} + \exp[-(x-Y)])\mathbb{1}_{\mathcal{X}_3}(Y)]\\
&= \mathbb{E}[\log(\rho_*^{-1} + \exp(-\epsilon Z))\mathbb{1}_{[0,\epsilon^{-1}(x-m_2))}(Z)]\,, \qquad (126)
\end{aligned}$$

where $Z \sim \mathrm{N}(0,1)$. Now, if $y \in \mathcal{X}_2 = (-m_2, m_2)$, then $-x + |y| < m_2 - m_*$. Let $\rho_*^{-1} \leq 1 - \exp(m_2 - m_*)$. If $y \in \mathcal{X}_2$, then

$$\rho_*^{-1} + \exp(-x + |y|) < \rho_*^{-1} + \exp(m_2 - m_*) \leq 1\,.$$

Thus,

$$\mathbb{E}\left[\alpha(x,Y)\log\left(\frac{1}{\rho_*} + \frac{p(Y)}{p(x)}\right)\mathbb{1}_{\mathcal{X}_2}(Y)\right] < 0\,. \qquad (127)$$

Finally, note that $\rho_*^{-1} + \exp(-x-y) \geq 1$ if and only if

$$y \leq -x + \log\left(\frac{\rho_*}{\rho_* - 1}\right) =: \varphi\,.$$

Thus,

$$\mathbb{E}\left[\alpha(x,Y)\log\left(\frac{1}{\rho_*}+\frac{p(Y)}{p(x)}\right)\mathbb{1}_{\mathcal{X}_1}(Y)\right] \leq \mathbb{E}\left[\log\left(\frac{1}{\rho_*}+\exp[-x-y]\right)\mathbb{1}_{(-\infty,\varphi\wedge -m_2)}(Y)\right].$$

Moreover, since $x \geq m_*$, then $-x-y = (x-y)-2x \leq (x-y)-2m_*$. Thus,

$$\begin{aligned}
&\mathbb{E}\left[\alpha(x,Y)\log\left(\frac{1}{\rho_*}+\frac{p(Y)}{p(x)}\right)\mathbb{1}_{\mathcal{X}_1}(Y)\right] \\
&\leq \mathbb{E}\left[\log\left(\frac{1}{\rho_*}+\exp[(x-y)-2m_*]\right)\mathbb{1}_{(-\infty,\varphi\wedge -m_2)}(Y)\right] \\
&\leq \mathbb{E}\left[\log\left(\frac{1}{\rho_*}+\exp[\epsilon Z-2m_*]\right)\mathbb{1}_{(\epsilon^{-1}\{(x-\varphi)\vee(x+m_2)\},\infty)}(Z)\right]. \qquad (128)
\end{aligned}$$

where $Z \sim \mathrm{N}(0,1)$. Combining inequalities (125), (126), (127), and (128) gives

$$\begin{aligned}
&\mathbb{E}\left[\alpha(x,Y)\log\left(\frac{1}{\rho_*}+\frac{p(Y)}{p(x)}\right)\right] \\
&< \mathbb{E}[\exp(-\theta_2\epsilon Z)\log(\rho_*^{-1}+\exp(\epsilon Z))\mathbb{1}_{[0,\infty)}(Z)] \\
&\quad + \mathbb{E}[\log(\rho_*^{-1}+\exp(-\epsilon Z))\mathbb{1}_{[0,\epsilon^{-1}(x-m_2))}(Z)] \\
&\quad + \mathbb{E}\left[\log\left(\rho_*^{-1}+\exp[\epsilon Z-2m_*]\right)\mathbb{1}_{(\epsilon^{-1}\{(x-\varphi)\vee(x+m_2)\},\infty)}(Z)\right].
\end{aligned}$$

The integrand in each expectation on the right-hand side is dominated by an exponentially decaying term. Hence, for each expectation on the right-hand side, there exists an integrable function which dominates the integrand and the assumptions for the dominated convergence theorem hold. By definition, $x \geq m_*$. Thus, as $m_* \uparrow \infty$, $x \uparrow \infty$. Taking the limit, firstly as $\rho_* \uparrow 0$, then as $m_* \uparrow \infty$, gives, for the first two terms,

$$\begin{aligned}
\lim_{m_*\uparrow\infty}\lim_{\rho_*\uparrow\infty}\mathbb{E}[\exp(-\theta_2\epsilon Z)\log(\rho_*^{-1}+\exp(\epsilon Z))\mathbb{1}_{[0,\infty)}(Z)] &= \mathbb{E}[\epsilon Z\exp(-\theta_2\epsilon Z)\mathbb{1}_{[0,\infty)}(Z)], \\
\lim_{m_*\uparrow\infty}\lim_{\rho_*\uparrow\infty}\mathbb{E}[\log(\rho_*^{-1}+\exp(-\epsilon Z))\mathbb{1}_{[0,\epsilon^{-1}(x-m_2))}(Z)] &= \mathbb{E}[-\epsilon Z\mathbb{1}_{[0,\infty)}(Z)].
\end{aligned}$$

By construction, for the final term, $\rho_*^{-1}+\exp(\epsilon z-2m_*) \geq 0$ for $z \in (\epsilon^{-1}\{(x-\varphi)\vee(x+m_2)\},\infty)$. Moreover, since $x \geq m_*$, then

$$x-\varphi = 2x-\log\left(\frac{\rho_*}{\rho_*-1}\right) \geq 2m_*-\log\left(\frac{\rho_*}{\rho_*-1}\right),$$

and, therefore,

$$\lim_{m_*\uparrow\infty}\lim_{\rho_*\uparrow\infty}(x-\varphi) = \infty.$$

Hence, taking the limit, firstly as $\rho_* \uparrow \infty$, then as $m_* \uparrow \infty$, gives, for the final term,

$$\lim_{m_*\uparrow\infty}\lim_{\rho_*\uparrow\infty}\mathbb{E}\left[\log\left(\rho_*^{-1}+\exp[\epsilon Z-2m_*]\right)\mathbb{1}_{(\epsilon^{-1}\{(x-\varphi)\vee(x+m_2)\},\infty)}(Z)\right] = 0.$$

Combining this with the limits obtained from the first two terms demonstrates that there exists a large enough $\rho_* \in (1,\infty)$ and a large enough $m_* > m_2$ such that

$$\mathbb{E}\left[\alpha(x,Y)\log\left(\frac{1}{\rho_*}+\frac{p(Y)}{p(x)}\right)\right] < \mathbb{E}[\epsilon Z(\exp(-\theta_2\epsilon Z)-1)\mathbb{1}_{[0,\infty)}(Z)],$$

where $Z \sim \mathrm{N}(0,1)$. Hence, there exists a large enough $\rho_* \in (1,\infty)$, and a $\delta > 0$ such that $C := \{x \in \mathcal{X} : p(x) \leq \rho\}$ is small, and for any $x \in C^c \cap [0,\infty)$,

$$\mathbb{E}_{q(\cdot|x)}\left[\alpha(x,Y)\log\left(\frac{1}{\rho_*}+\frac{p(Y)}{p(x)}\right)\right] < -\delta.$$

A similar argument holds if one starts by assuming that $x \leq m_* < m_1$. Thus, property (IM) of Theorem 4.3.3 is satisfied. Now, define $\mathcal{P}_r^+(x) := \mathcal{P}_r(x) \cap [0,\infty)$, $\mathcal{P}_r^-(x) := \mathcal{P}_r(x) \cap (-\infty,0)$, suppose $x > 0$, and consider

$$\mathbb{E}\left[\alpha(x,Y)\log\left(\frac{1}{\rho_*}+\frac{p(Y)}{p(x)}\right)^2\mathbb{1}_{\mathcal{P}_r^+(x)}(Y)\right] \leq \frac{1}{\sqrt{2\pi\epsilon^2}}\int_{\mathcal{P}_r^+(x)}\log\left(\rho_*^{-1}+\exp(y-x)\right)^2\exp\left(-\frac{(y-x)^2}{2\epsilon^2}\right)\mathrm{d}y.$$

Note that $y \in \mathcal{P}_r^+(x)$ if and only if $(y-x) \geq 0$. Therefore, making the substitution $s := y - x$,

$$\mathbb{E}\left[\alpha(x,Y)\log\left(\frac{1}{\rho_*}+\frac{p(Y)}{p(x)}\right)^2 \mathbb{1}_{\mathcal{P}_r^+(x)}(Y)\right] \leq \frac{1}{\sqrt{2\pi\epsilon^2}}\int_0^\infty \log\left(\rho_*^{-1}+\exp(s)\right)^2 \exp\left(-\frac{s^2}{2\epsilon^2}\right) \mathrm{d}s\ ,$$

which is bounded since the integrand is dominated by an exponentially decaying term. Next, consider

$$\begin{aligned}&\mathbb{E}\left[\alpha(x,Y)\log\left(\frac{1}{\rho_*}+\frac{p(Y)}{p(x)}\right)^2 \mathbb{1}_{\mathcal{P}_r^-(x)}(Y)\right]\\ &\leq \frac{1}{\sqrt{2\pi\epsilon^2}}\int_{\mathcal{P}_r^-(x)} \log\left(\rho_*^{-1}+\exp(-x-y)\right)^2 \exp\left(-\frac{(y-x)^2}{2\epsilon^2}\right) \mathrm{d}y\ .\end{aligned}$$

Note that $y \in \mathcal{P}_r^-(x)$ if and only if $y \leq -x$. Therefore, making the substitution $s := y - x$,

$$\mathbb{E}\left[\alpha(x,Y)\log\left(\frac{1}{\rho_*}+\frac{p(Y)}{p(x)}\right)^2 \mathbb{1}_{\mathcal{P}_r^-(x)}(Y)\right] \leq \frac{1}{\sqrt{2\pi\epsilon^2}}\int_{-\infty}^{-2x} \log\left(\rho_*^{-1}+\exp(-2x-s)\right)^2 \exp\left(-\frac{s^2}{2\epsilon^2}\right) \mathrm{d}s\ .$$

For $s \leq -2x$, $-2x - s \geq 0$. Hence,

$$\mathbb{E}\left[\alpha(x,Y)\log\left(\frac{1}{\rho_*}+\frac{p(Y)}{p(x)}\right)^2 \mathbb{1}_{\mathcal{P}_r^-(x)}(Y)\right] \leq \frac{1}{\sqrt{2\pi\epsilon^2}}\int_{-\infty}^{0} \log\left(\rho_*^{-1}+\exp(-s)\right)^2 \exp\left(-\frac{s^2}{2\epsilon^2}\right) \mathrm{d}s\ .$$

Again, this last quantity is bounded since the integrand is dominated by an exponentially decaying term. Thus,

$$\sup_{x\in C^c\cap[0,\infty)} \mathbb{E}\left[\alpha(x,Y)\log\left(\frac{1}{\rho_*}+\frac{p(Y)}{p(x)}\right)^2 \mathbb{1}_{\mathcal{P}_r(x)}(Y)\right] < \infty\ .$$

A similar argument shows that

$$\sup_{x\in C^c\cap(-\infty,0)} \mathbb{E}\left[\alpha(x,Y)\log\left(\frac{1}{\rho_*}+\frac{p(Y)}{p(x)}\right)^2 \mathbb{1}_{\mathcal{P}_r(x)}(Y)\right] < \infty\ .$$

Therefore, condition (UI) of Theorem 4.3.3 holds with $\tau = 1$. The inequality logic to demonstrate condition (B) is very similar. Indeed, suppose $x \in C \cap [0,\infty)$ and consider

$$\begin{aligned}&\mathbb{E}\left[p(x)\alpha(x,Y)\left(\frac{p(Y)}{p(x)}-1\right)\mathbb{1}_{\mathcal{P}_1^+(x)}(Y)\right]\\ &= \mathbb{E}\left[p(x)\alpha(x,Y)\Big(\exp(y-x)-1\Big)\mathbb{1}_{\mathcal{P}_1^+(x)}(Y)\right]\\ &\leq \rho_*\mathbb{E}\left[\Big(\exp(y-x)-1\Big)\mathbb{1}_{\mathcal{P}_1^+(x)}(Y)\right],\end{aligned}$$

where the last line follows since $p(x) \leq \rho_*$ for $x \in C$. Note that $y \in \mathcal{P}_1^+(x)$ if and only if $y \geq x$. Therefore, making the substitution $s := y - x$,

$$\mathbb{E}[\alpha(x,Y)(p(Y)-p(x))\mathbb{1}_{\mathcal{P}_1^+(x)}(Y)] \leq \frac{\rho_*}{\sqrt{2\pi\epsilon^2}}\int_0^\infty (\exp(s)-1)\exp\left(-\frac{s^2}{2\epsilon^2}\right) \mathrm{d}s\ .$$

Next, consider

$$\mathbb{E}\left[p(x)\alpha(x,Y)\left(\frac{p(Y)}{p(x)}-1\right)\mathbb{1}_{\mathcal{P}_1^-(x)}(Y)\right] \leq \rho_*\mathbb{E}\left[\Big(\exp(-x-y)-1\Big)\mathbb{1}_{\mathcal{P}_1^+(x)}(Y)\right].$$

Note that $y \in \mathcal{P}_1^-(x)$ if any only if $y \leq -x$. Therefore, making the substitution $s := y - x$,

$$\begin{aligned}&\mathbb{E}[\alpha(x,Y)(p(Y)-p(x))\mathbb{1}_{\mathcal{P}_1^-(x)}(Y)]\\&\leq \frac{\rho_*}{\sqrt{2\pi\epsilon^2}}\int_{-\infty}^{-2x}(\exp(-2x-s)-1)\exp\left(-\frac{s^2}{2\epsilon^2}\right)\mathrm{d}s\,.\end{aligned}$$

For $s \leq -2x$, $-2x - s \geq 0$. Hence,

$$\mathbb{E}[\alpha(x,Y)(p(Y)-p(x))\mathbb{1}_{\mathcal{P}_1^-(x)}(Y)] \leq \frac{\rho_*}{\sqrt{2\pi\epsilon^2}}\int_{-\infty}^{0}(\exp(-s)-1)\exp\left(-\frac{s^2}{2\epsilon^2}\right)\mathrm{d}s\,.$$

Combining the two results gives

$$\sup_{x\in C\cap[0,\infty)}\mathbb{E}[\alpha(x,Y)(p(Y)-p(x))\mathbb{1}_{\mathcal{P}_1(x)}(Y)] < \infty\,.$$

A similar argument shows that

$$\sup_{x\in C\cap(-\infty,0)}\mathbb{E}[\alpha(x,Y)(p(Y)-p(x))\mathbb{1}_{\mathcal{P}_1(x)}(Y)] < \infty\,.$$

Therefore, condition (B) of Theorem 4.3.3 holds. Hence, by Theorem 4.3.3, the chain is geometrically ergodic.

A.14 PROOF OF LEMMA 4.3.5

Let $\bar{w} > 0$ be arbitrarily chosen such that $C := \{x \in \mathcal{X} : w(x) \leq \bar{w}\}$ is a compact set. Let $D \subseteq \mathcal{X}$ be any compact set such that

$$\inf_{(x,y)\in C\times D}\tilde{q}_1(y|x) \geq \eta > 0\,, \quad \text{and} \quad \mathrm{Leb}(D) > 0\,.$$

The existence of such a set D is guaranteed by Lemma 4.3.1. Let $P(x,\cdot)$ be the proposal distribution corresponding to the chain and, for any $x \in C$, and any $A \subseteq \mathcal{X}$, consider

$$P(x,A) \geq \int_A \alpha(x,y)\tilde{q}_1(y|x)\,\mathrm{d}y \geq \beta\int_{A\cap D}\alpha_m(x,y)\tilde{q}_1(y|x)\,\mathrm{d}y\,,$$

where α^m corresponds to the Metropolis-Hastings acceptance probability (given by (27)), and, $\beta = 1$ for the Exchangeable Sampler with the Metropolis-Hastings acceptance probability, and $\beta = 1/2$ for the Exchangeable Sampler with Barker's acceptance probability. The weights are bounded by $\bar{w}$ on the set C, hence, for $x \in C$,

$$P(x,A) \geq \int_{A\cap D}\left(1 \wedge \frac{w(y)}{\bar{w}}\right)\tilde{q}_1(y|x)\,\mathrm{d}y \geq \eta\lambda(A)\,,$$

where

$$\lambda(A) := \int_{A\cap D}\left(1 \wedge \frac{w(y)}{\bar{w}}\right)\mathrm{d}y\,.$$

By assumption, $\mathrm{Leb}(D) > 0$, and, so, since the integrand is positive and continuous, λ is a measure. Moreover, $0 < \lambda(\mathcal{X}) \leq P(x,\mathcal{X})/\eta = 1/\eta < \infty$. Thus, $v := \lambda/\lambda(\mathcal{X})$ is a probability measure and

$$P(x,A) \geq \eta\lambda(\mathcal{X})v(A)\,.$$

Hence, C is small.

A.15 PROOF OF COROLLARY 4.3.6

Taking $p = w^{-1}$ we will show that the assumptions of Theorem 4.3.3 hold, therefore allowing us to use this theorem to deduce geometric ergodicity. Firstly, by assumptions (B) and (C), $C := \{x \in \mathcal{X} : w_* \leq w(x)\} = \{x \in \mathcal{X} : w(x)^{-1} \leq w_*^{-1}\}$ is a compact set on which the weights are bounded. Thus, by Lemma 4.3.5, C is a small set, and Assumption (S) of Theorem 4.3.3 holds. Secondly, Assumption (IM) is a rewrite of Assumption (IM) of Theorem 4.3.3 with w^{-1} in place of p and w_*^{-1} in place of ρ_*. Hence, trivially, Assumption (IM) of Theorem 4.3.3 holds. Thirdly, note that, for $\tau > 1$, if $y \in \mathcal{P}_\tau(x)$, then $w(y)/w(x) \leq \tau^{-1} < 1$. Thus, for $y \in \mathcal{P}_\tau(x)$,

$$\alpha(x,y) \log\left(w_* + \frac{w(x)}{w(y)}\right)^2 \leq \frac{w(y)}{w(x)} \log\left(w_* + \frac{w(x)}{w(y)}\right)^2 .$$

Hence, to demonstrate Assumption (UI) of Theorem 4.3.3, it is sufficient to show that

$$\lim_{z \downarrow 0} h(z) = 0 ,$$

where $h(z) := z \log(w_* + z^{-1})^2$. Rewriting h as $h(z) = z[\log(1 + w_* z) - \log(z)]^2|$ highlights the fact that it is sufficient to show that

$$\lim_{z \downarrow 0} z \log(z) = \lim_{z \downarrow 0} z \log(z)^2 = 0 . \tag{129}$$

Both these limits follow from L'Hôpital's rule. Indeed,

$$\lim_{z \downarrow 0} z \log(z) = \lim_{z \downarrow 0} \frac{\log(z)}{1/z} = -\lim_{z \downarrow 0} \frac{1/z}{1/z^2} = 0 .$$

Moreover,

$$\lim_{z \downarrow 0} z \log(z)^2 = \lim_{z \downarrow 0} \frac{\log(z)^2}{1/z} = -2 \lim_{z \downarrow 0} \frac{\log(z)/z}{1/z^2} = -2 \lim_{z \downarrow 0} z \log(z) .$$

Therefore, Assumption (UI) of Theorem 4.3.3 holds. Finally, to demonstrate Assumption (B) of Theorem 4.3.3, note that $w(x)\alpha(x,y) = w(y)\alpha(y,x)$. Therefore, since, for $x \in C$, $w(x) \geq w_*$,

$$\alpha(x,y) w(y)^{-1} = \alpha(y,x) w(x)^{-1} \leq w(x)^{-1} \leq w_*^{-1} .$$

Hence, for $x \in C$,

$$\alpha(x,y)(w(y)^{-1} - w(x)^{-1}) < \alpha(x,y) w(y)^{-1} \leq w_*^{-1} .$$

Thus, Assumption (B) of Theorem 4.3.3 holds.

A.16 PROOF OF COROLLARY 4.3.7

Taking $p = w$ we will show that the assumptions of Theorem 4.3.3 hold, therefore allowing us to use this theorem to deduce geometric ergodicity. Firstly, by assumption, $C := \{x \in \mathcal{X} : w(x) \leq w^*\}$ is a compact set on which the weights are bounded. Thus, by Lemma 4.3.5, C is a small set, and Assumption (S) of Theorem 4.3.3 holds. Secondly, Assumption (IM) is a rewrite of Assumption (IM) of Theorem 4.3.3 with w in place of p and w^* in place of ρ_*. Hence, trivially, Assumption (IM) of Theorem 4.3.3 holds. Thirdly, note that, for $\tau > 1$, if $y \in \mathcal{P}_\tau(x)$, then $w(x)/w(y) \leq \tau^{-1} < 1$. Thus, for $y \in \mathcal{P}_\tau(x)$,

$$\alpha(x,y) \log\left(\frac{1}{w^*} + \frac{w(y)}{w(x)}\right)^2 \leq \frac{w(y)}{w(x)} \log\left(\frac{1}{w^*} + \frac{w(y)}{w(x)}\right)^2 .$$

Thus, Assumption (UI) of Theorem 4.3.3 follows from Assumption (UI) of this corollary. Finally, note that, if $y \in \mathcal{P}_1(x)$, then $w(x) \leq w(y)$. Therefore, by Assumption (B),

$$\mathbb{E}_{q(\cdot|x)}[\alpha(x,Y)(w(Y) - w(x)) \mathbb{1}_{\mathcal{P}_1(x)}(Y)] \leq \mathbb{E}_{q(\cdot|x)}[w(Y) \mathbb{1}_{\mathcal{P}_1(x)}(Y)] < \infty .$$

Hence, Assumption (B) of Theorem 4.3.3 holds.

A.17 PROOF OF THEOREM 4.3.9

Let $P_N(x,\cdot)$ be the transition distributions of the chain, $\kappa_N(A)$ the conductance of any measurable set $A \subseteq \mathcal{X}$; that is,

$$\kappa_N(A) := \frac{1}{\pi(A)\pi(A^c)} \int_A \pi(\mathrm{d}x) P_N(x, A^c) \ ,$$

and κ_N the conductance of the chain; that is,

$$\kappa_N := \inf_{A \in \Omega} \kappa_N(A) \ ,$$

where $\Omega := \{A \subseteq \mathcal{X} : A \text{ is measurable}\}$. Suppose $N = 1$. Either the assumptions of Corollary 4.3.6 hold, or the assumptions of Corollary 4.3.7 hold. Hence, for $N = 1$, the sampler is geometrically ergodic. Moreover, by Theorem 4.3.2, the chain, X_t, is non-negative and reversible with respect to π. Therefore, by Theorem 2.3.25, the MCMC estimates corresponding to the chain satisfy a central limit theorem for all functions which are square-integrable with respect to π. Moreover, by Theorem 2.3.25, the chain has a non-zero conductance; that is, $\kappa_1 > 0$. To demonstrate the result for general $N \in \mathbb{N}$, we will show that the conductance of the chain, for general N; that is, κ_N, is non-zero and, again, appeal to Theorem 4.3.2 and Theorem 2.3.25. To this end, define, for any $N \in \mathbb{N}$, each $i = 1, \ldots, N$, any $r > 0$, any measurable $A \subseteq \mathcal{X}$, and any measurable $\Lambda \subseteq \mathcal{X}^N$, let

$$\begin{aligned} D_N^{(i)}(r) &:= \{y_{1:N} \in \mathcal{X}^N : w(y_k) \leq r w(y_i) \text{ for all } k \in \{1, \ldots, N\} \backslash \{i\}\} \ , \\ P_N^{(i)}(x, A) &:= \iint\limits_{A \times \mathcal{X}^{N-1}} \tilde{q}_N(y_{1:N}|x) \alpha_{i,N}(x, y_{1:N}) \, \mathrm{d}y_i \mathrm{d}y_{-i} \ , \\ I^{(i)}(x, \Lambda) &:= \int_{\Lambda} \tilde{q}_N(y_{1:N}|x) \alpha_{1,1}^m(x, y_i) \, \mathrm{d}y_i \mathrm{d}y_{-i} \ , \end{aligned}$$

where $\alpha_{i,N}^m$ denotes the the multiple-proposal Metropolis-Hastings acceptance probability (Equation (87)). By a multiple-proposal extension of Inequality (28), $\alpha_{i,N}(y_{0:N}) \geq \alpha_{i,N}^m(y_{0:N})/2$ for any $y_{0:N} \in \mathcal{X}^{N+1}$. Suppose $y_{1:N} \in D_N^{(i)}(r)$, then

$$\begin{aligned} \alpha_{i,N}^m(y_{0:N}) &= \frac{w(y_i)}{w(y_0) + \ldots + w(y_N) - [w(y_i) \wedge w(y_0)]} \\ &\geq \frac{w(y_i)}{w(y_0) + [1 + (N-1)r] w(y_i) - [w(y_i) \wedge w(y_0)]} \ . \end{aligned}$$

If $w(y_0) \leq w(y_i)$, then

$$\frac{w(y_i)}{w(y_0) + [1 + (N-1)r] w(y_i) - [w(y_i) \wedge w(y_0)]} = \frac{1}{[1 + (N-1)r]} \ .$$

On the other hand, if $w(y_0) \geq w(y_i)$, then

$$\begin{aligned} \frac{w(y_i)}{w(y_0) + [1 + (N-1)r] w(y_i) - [w(y_i) \wedge w(y_0)]} &= \frac{w(y_i)}{w(y_0) + (N-1) r w(y_i)} \\ &\geq \frac{1}{[1 + (N-1)r]} \frac{w(y_i)}{w(y_0)} \ . \end{aligned}$$

Therefore, for $y_{1:N} \in D_N^{(i)}(r)$, $\alpha_{i,N}(y_{0:N}) \geq \xi_N(r) \alpha_{1,1}^m(y_0, y_i)$ where

$$\xi_N(r) := \frac{1}{2[1 + (N-1)r]} \ .$$

Thus, for any measurable $A \subseteq \mathcal{X}$,

$$\begin{aligned} P_N^{(i)}(x, A^c) &\geq \xi_N(r) I^{(i)}[x, (A^c \times \mathcal{X}^{N-1}) \cap D_N^{(i)}(r)] \\ &= \xi_N(r) \{I^{(i)}[x, A^c \times \mathcal{X}^{N-1}] - I^{(i)}[x, (A^c \times \mathcal{X}^{N-1}) \cap D_N^{(i)}(r)^c]\} \\ &\geq \xi_N(r) \{I^{(i)}[x, A^c \times \mathcal{X}^{N-1}] - I^{(i)}[x, D_N^i(r)^c]\} \ . \end{aligned}$$

Note that, $y_{1:N} \in D_N^{(i)}(r)^c$ if and only if $rw(y_i) < w(y_k)$ for at least one $k \neq i$. Therefore,

$$D_N^{(i)}(r)^c = \left\{ y_{1:N} \in \mathcal{X}^N : \max_{k \in \{1,\dots,N\} \setminus \{i\}} w(y_k) > r w(y_i) \right\} .$$

Hence, since $\alpha_{1,1}(x,y) \leq 1$ for any $(x,y) \in \mathcal{X}^2$, then

$$\begin{aligned} I^{(i)}[x, D_N^{(i)}(r)^c] &= \int_{D_N^{(i)}(r)^c} \tilde{q}_N(y_{1:N}|x) \alpha_{1,1}(x, y_i) \, \mathrm{d}y_i \mathrm{d}y_{-i} \\ &\leq \int_{D_N^{(i)}(r)^c} \tilde{q}_N(y_{1:N}|x) \, \mathrm{d}y_i \mathrm{d}y_{-i} \\ &\leq \bar{\mu}_N^{(i)}(r) , \end{aligned}$$

where

$$\bar{\mu}_N^{(i)}(r) := \sup_{x \in \mathcal{X}} \mathbb{P}_{\tilde{q}_N(\cdot|x)} \left(\max_{k \in \{1,\dots,N\} \setminus \{i\}} \frac{w(Y_k)}{w(Y_i)} > r \right) .$$

The joint proposal, $q_N(x, y_{1:N}) = \tilde{q}_N(y_{1:N}|x) q_0(x)$, is exchangeable. Therefore, for any $i \in \{1, \dots, N\}$, $\bar{\mu}_N^{(i)}(r) = \bar{\mu}_N^{(1)}(r)$. Moreover, $\bar{\mu}_N^{(1)}(r) = 1 - \mu_N(r)$, where $\mu_N(r)$ is defined in Assumption 4.3.8. Moreover,

$$\begin{aligned} I^{(i)}[x, A^c \times \mathcal{X}^{N-1}] &= \int_{A^c \times \mathcal{X}^{N-1}} \tilde{q}_N(y_{1:N}|x) \alpha_{1,1}(x, y_i) \, \mathrm{d}y_i \mathrm{d}y_{-i} \\ &= \int_{A^c} \tilde{q}_1(y_i|x) \alpha_{1,1}(x, y_i) \, \mathrm{d}y_i = P_1^{(1)}(x, A^c) . \end{aligned}$$

Hence

$$P_N^{(i)}(x, A^c) \geq \xi_N(r) [P_1^{(1)}(x, A^c) - \bar{\mu}_N^{(i)}(r)] . \tag{130}$$

For any measurable sets $A \subseteq \mathcal{X}$ and $B \subseteq \mathcal{X}$, let

$$Q_N(A, B) := \int_A \pi(\mathrm{d}x) P_N(x, B) .$$

Then, as in the proof of Theorem 4.3.2 (see A.11),

$$Q_N(A, B) = \sum_{i=1}^N Q_N^{(i)}(A, B) + Q_N^*(A, B) ,$$

where

$$Q_N^{(i)}(A, B) := \int_A \pi(\mathrm{d}x) P_N^{(i)}(x, B) , \quad Q_N^*(A, B) := \int_A \pi(\mathrm{d}x) P_N^*(x, B) ,$$

and

$$P_N^*(x, B) := \delta_x(B) \int_{\mathbb{R}^{d \times N}} \cdots \int \tilde{q}_N(y_{1:N}|x) \left(1 - \sum_{i=1}^N \alpha_{i,N}(x, y_{1:N}) \right) \mathrm{d}y_{1:N} .$$

For any measurable $A \subseteq \mathcal{X}$, $Q_N^*(A, A^c) = 0$. Thus

$$\kappa_N(A) = \frac{1}{\pi(A)\pi(A^c)} Q_N(A, A^c) = \sum_{i=1}^N \kappa_N^{(i)}(A) ,$$

where

$$\kappa_N^{(i)}(A) := \frac{1}{\pi(A)\pi(A^c)} \int_A \pi(\mathrm{d}x) P_N^{(i)}(x, A^c) \, \mathrm{d}x .$$

By Assumption 4.3.8, $\mu_N(r) \to 1$ as $r \to \infty$. Thus, $\bar{\mu}_N^{(i)} \to 0$ as $r \to \infty$ for any $i \in \{1, \ldots, N\}$. Hence, since $\kappa_1 > 0$, there exists a $\beta \in (0,1)$ and an $r^* > 0$ such that $\mu_N^{(i)}(r^*) \leq (1-\beta)\kappa_1$ for any $i \in \{1, \ldots, N\}$. Thus, by (130),

$$\kappa_N^{(i)}(A) \geq \xi_N(r^*)[\kappa_1(A) - (1-\beta)\kappa_1] \geq \beta\xi_N(r^*) \,,$$

which is true for any $i \in \{1, \ldots, N\}$. Therefore, $\kappa_N(A) \geq N\beta\xi_N(r^*)$. This lower bound is independent of the set A, hence, $\kappa_N > 0$. By Theorem 4.3.2, the chain, X_t, is non-negative and reversible with respect to π. Therefore, by Theorem 2.3.25, the MCMC estimates corresponding to the chain satisfy a central limit theorem for all functions which are square-integrable with respect to π.

A.18 PROOF OF LEMMA 4.3.10

For brevity, we will, throughout the proof, drop the explicit dependence of the expectation on $\tilde{q}_1$. By (28), it is sufficient to show the result for the case where α corresponds to the Metropolis-Hastings acceptance probability. We will demonstrate that

$$\lim_{x \uparrow \infty} \mathbb{E}\left[\alpha(x, Y) \log\left(\frac{w(Y)}{w(x)}\right) \mathbb{1}_{\mathcal{P}(x)}(Y)\right] = 0 \,, \tag{131}$$

where $\mathcal{P}(x) := \{z \in \mathcal{X} : w(z) \leq w(x)\}$. For, if this is true, then, for any $\delta > 0$, there exists an $x^* > 0$ such that, for any $x \geq x^*$,

$$\mathbb{E}\left[\alpha(x, Y) \log\left(\frac{w(Y)}{w(x)}\right) \mathbb{1}_{\mathcal{P}(x)}(Y)\right] > -\delta \,.$$

Therefore, for any $\delta > 0$, there exists an $x^* > 0$ such that, for any $x \geq x^*$, and any $w^* > 0$,

$$\begin{aligned}
\mathbb{E}\left[\alpha(x, Y) \log\left(\frac{1}{w^*} + \frac{w(Y)}{w(x)}\right)\right] &> \mathbb{E}\left[\alpha(x, Y) \log\left(\frac{w(Y)}{w(x)}\right)\right] \\
&> \mathbb{E}\left[\alpha(x, Y) \log\left(\frac{w(Y)}{w(x)}\right) \mathbb{1}_{\mathcal{P}(x)}(Y)\right] \\
&> -\delta \,.
\end{aligned}$$

Now,

$$\begin{aligned}
\mathbb{E}\left[\alpha(x, Y) \log\left(\frac{w(Y)}{w(x)}\right) \mathbb{1}_{\mathcal{P}(x)}(Y)\right] &= \mathbb{E}\left[\frac{w(Y)}{w(x)} \log\left(\frac{w(Y)}{w(x)}\right) \mathbb{1}_{\mathcal{P}(x)}(Y)\right] \\
&= \tau \mathbb{E}[(Y^2 - x^2) \exp[\tau(Y^2 - x^2)] \mathbb{1}_{\mathcal{P}(x)}(Y)] \,,
\end{aligned}$$

where $\mathcal{P}(x) := \{z \in \mathcal{X} : z^2 \leq x^2\}$, and, since the proposal, $\tilde{q}_1$ corresponds to Algorithm 17,

$$Y = \sigma(1-\eta^2)x + \sigma\eta\sqrt{2-\eta^2}Z \,,$$

where $Z \sim \mathrm{N}(0,1)$, and $\eta := \epsilon/\sqrt{2}$. By Lemma B.0.13,

$$\mathbb{P}(Y^2 \geq x^2 - m) = \begin{cases} 1 & \text{if } x^2 \leq m \,, \\ 2 - \Phi\left(\frac{\sqrt{x^2-m} - \hat{\mu}x}{\hat{\sigma}}\right) - \Phi\left(\frac{\sqrt{x^2-M} + \hat{\mu}x}{\hat{\sigma}}\right) & \text{if } x^2 > m \,, \end{cases} \tag{132}$$

for any $m > 0$, where $\hat{\mu} := \sigma(1-\eta^2)$ and $\hat{\sigma} := \sigma\eta\sqrt{2-\eta^2}$. By assumption, $\sigma < 1$ and $(1-\delta^2) \in [0,1)$. Thus, $\hat{\mu}x < x$. Hence, $(\sqrt{x^2 - M} - \mu x)$ tends towards infinity as x tends towards infinity. Therefore, by (132),

$$\lim_{x \uparrow \infty} \mathbb{P}[(Y^2 - x^2) \geq -m] = 0 \,,$$

for any $m > 0$. Hence,

$$\lim_{x \uparrow \infty} \mathbb{P}[(Y^2 - x^2)\mathbb{1}_{\mathcal{P}(x)}(Y) \geq -m] = 0 \,,$$

for any $m > 0$. That is,

$$\operatorname*{plim}_{x\uparrow\infty}[(Y^2 - x^2)\mathbb{1}_{\mathcal{P}(x)}(Y)] = -\infty \ .$$

By the continuous mapping theorem, for any $\tau > 0$,

$$\operatorname*{plim}_{x\uparrow\infty}\left[\tau(Y^2 - x^2)\exp[\tau(Y^2 - x^2)]\mathbb{1}_{\mathcal{P}(x)}(Y)\right] = 0 \ .$$

For any $\tau > 0$, the function $f(z) := \tau z \exp(\tau z)1_{(-\infty,0]}(z)$ is bounded. Hence, by the dominated convergence theorem, the limit in (131) holds.

A.19 PROOF OF LEMMA 4.3.13

Note that $p'(z) = g'(z)/g(z)$. Thus,

$$\mathbb{E}[p'(Z)^2] = \int_{-\infty}^{\infty} p'(z)\frac{g'(z)}{g(z)}\pi_Z(z)\,\mathrm{d}z \ .$$

Now,

$$\frac{\pi_Z(z)}{g(z)} = \frac{\pi(h_*(z))|h_*'(z)|}{w[h_*(z)]} = q_0[h_*(z)]|h_*'(z)|\gamma(\mathcal{X})^{-1} \ .$$

Therefore,

$$\mathbb{E}[p'(Z)^2] = \gamma(\mathcal{X})^{-1}\int_{-\infty}^{\infty} p'(z)g'(z)q_0[h_*(z)]|h_*'(z)|\,\mathrm{d}z = \gamma(\mathcal{X})^{-1}\mathbb{E}[p'(V)g'(V)] \ ,$$

where V has density $q_0[h_*(v)]|h_*'(v)|$. By definition, if $X \sim q_0$, then $h_*^{-1}(X) \sim \mathrm{N}(0,1)$. Hence, $\phi(v) = q_0[h_*(v)]|h_*'(v)|$. Thus,

$$\begin{aligned}\mathbb{E}[p'(Z)^2] &= \gamma(\mathcal{X})^{-1}\int_{-\infty}^{\infty} p'(v)g'(v)\phi(v)\,\mathrm{d}v \\ &= \gamma(\mathcal{X})^{-1}\Big[p'(v)g(v)\phi(v)\Big]_{-\infty}^{\infty} - \gamma(\mathcal{X})^{-1}\int_{-\infty}^{\infty} g(v)[p''(v) - vp'(v)]\phi(v)\,\mathrm{d}v \ .\end{aligned}$$

Note that

$$\gamma(\mathcal{X})^{-1}g(v)\phi(v) = \gamma(\mathcal{X})^{-1}g(v)q_0[h_*(v)]|h_*'(v)| = \pi_Z(v) \ .$$

Moreover, $p'(v)g(v) = g'(v)$. Therefore, by property (G) of Assumptions 4.3.12,

$$\mathbb{E}[p'(Z)^2] = -\mathbb{E}[p''(Z) - Zp'(Z)] \ .$$

A.20 PROOF OF LEMMA 4.3.14

Let $B(\hat{Z}_{0:1}, Z_0) := (\hat{Z}_0 + \hat{Z}_1) - \delta Z_0$, and, for $i \in \{1,2,3\}$, let

$$R_i(\hat{Z}_{0:1}, Z_{0:1}, \delta) := \delta^{-(i+2)}(Z_1 - Z_0)^i - \delta^{-2}B(\hat{Z}_{0:1}, Z_0)^i \ . \tag{133}$$

Suppressing the necessary arguments, we note the following relationships;

$$R_2 = \delta^2 R_1^2 + 2R_1 B \ , \quad R_3 = \delta^4 R_1^3 + 3BR_2 - 3B^2 R_1 \ . \tag{134}$$

Let $\kappa(Z_{0:1}) := p'(Z_0)(Z_1 - Z_0) + p''(Z_0)(Z_1 - Z_0)^2/2$. By property (B.b) of Assumptions 4.3.12, p is twice differentiable. Hence, by Lemma B.0.6,

$$p(Z_1) - p(Z_0) = \kappa(Z_{0:1}) + \frac{1}{2}[p''(Z_0 + t(Z_{0:1})(Z_1 - Z_0)) - p''(Z_0)](Z_1 - Z_0)^2$$

for some $t(Z_{0:1})$ such that $|t(Z_{0:1})| \leq 1$. Note that we can rewrite κ as

$$\kappa(Z_{0:1}) = \delta C_1(\hat{Z}_{0:1}, Z_0) + \delta^2 C_2(\hat{Z}_{0:1}, Z_0) + \delta^3 p'(Z_0) R_1(\hat{Z}_{0:1}, Z_{0:1}, \delta) + \frac{1}{2}\delta^4 p''(Z_0) R_2(\hat{Z}_{0:1}, Z_{0:1}, \delta) + \frac{1}{2} p''(Z_0)\delta^3(\delta Z_0^2 - 2(\hat{Z}_0 + \hat{Z}_1)Z_0) ,$$

so that

$$p(Z_1) - p(Z_0) - \delta C_1(\hat{Z}_{0:1}, Z_0) - \delta^2 C_2(\hat{Z}_{0:1}, Z_0) = \delta^3 R(\hat{Z}_{0:1}, Z_{0:1}, \delta)$$

where

$$\begin{aligned} R(\hat{Z}_{0:1}, Z_{0:1}, \delta) := p'(Z_0) R_1(\hat{Z}_{0:1}, Z_{0:1}, \delta) &+ \frac{1}{2}\delta p''(Z_0) R_2(\hat{Z}_{0:1}, Z_{0:1}, \delta) \\ &+ \frac{1}{2} p''(Z_0)(\delta Z_0^2 - 2(\hat{Z}_0 + \hat{Z}_1)Z_0) \\ &+ \frac{1}{2}\delta^{-3}(p''(Z_0 + t(Z_{0:1})(Z_1 - Z_0)) - p''(Z_0))(Z_1 - Z_0)^2 . \end{aligned}$$

Using the fact that $\delta \in (0,1)$ we bound the modulus of each of these terms by quantities which are independent of δ and Z_1 so that, as claimed, the modulus of their sum is bounded by some $R^*(\hat{Z}_{0:1}, Z_0)$. Then, using the Cauchy-Schwartz inequality, along with Assumptions 4.3.12, we bound the expectation of this quantity. Firstly, by property (L) of Assumptions 4.3.12, we have

$$\begin{aligned} &\delta^{-3}|p''(Z_0 + t(Z_{0:1})(Z_1 - Z_0)) - p''(Z_0)||(Z_1 - Z_0)^2| \\ &\leq \delta^{-3} a |t(Z_{0:1})||Z_1 - Z_0|^3 \\ &\leq \delta^{-3} a |Z_1 - Z_0|^3 \\ &\leq a|\delta^2 R_3(\hat{Z}_{0:1}, Z_{0:1}, \delta) + B(\hat{Z}_{0:1}, Z_0)^3| \\ &\leq a(|R_3(\hat{Z}_{0:1}, Z_{0:1}, \delta)| + |B(\hat{Z}_{0:1}, Z_0)^3|) , \end{aligned}$$

where, in the second inequality, we have used the fact that $|t(Z_{1:2})| \leq 1$, in the third inequality, we have used Definition (133) and, in the fourth inequality, we have used the fact that $\delta \in (0,1)$. Secondly, since $\delta \in (0,1)$,

$$|\delta Z_0^2 - 2(\hat{Z}_0 + \hat{Z}_1)Z_0| \leq |Z_0|^2 + 2|\hat{Z}_0 + \hat{Z}_1||Z_0| .$$

Thirdly, note that

$$R_1(\hat{Z}_{0:1}, Z_{0:1}, \delta) = -\delta^{-2}(1 - \sqrt{1 - \delta^2})\hat{Z}_0 .$$

$\delta \in (0,1)$, so $\delta = \sin\theta$ for some $\theta \in (0, \pi/2)$. Therefore,

$$\delta^2 - (1 - \sqrt{1 - \delta^2}) = \sin\theta^2 - (1 - \cos\theta) = \cos\theta - (\cos\theta)^2 \geq 0 .$$

Thus $0 \leq 1 - \sqrt{1 - \delta^2} \leq \delta^2$ and

$$|R_1(\hat{Z}_{0:1}, Z_{0:1}, \delta)| \leq |\hat{Z}_0| .$$

Hence, using the relationships given by (134), along with the fact that $\delta \in (0,1)$, we have

$$\begin{aligned} &|R_2(\hat{Z}_{0:1}, Z_{0:1}, \delta)| \leq |\hat{Z}_0|^2 + 2|\hat{Z}_0||B(\hat{Z}_{0:1}, Z_0)| , \\ &|R_3(\hat{Z}_{0:1}, Z_{0:1}, \delta)| \leq |\hat{Z}_0|^3 + 3|B(\hat{Z}_{0:1}, Z_0)||\hat{Z}_0|^2 + 9|B(\hat{Z}_{0:1}, Z_0)|^2|\bar{Z}_0| . \end{aligned}$$

Moreover, we also have that

$$|B(\hat{Z}_{0:1}, Z_0)| \leq |\hat{Z}_0 + \hat{Z}_1| + |Z_0| .$$

Therefore,

$$|R(\hat{Z}_{0:1}, Z_{0:1}, \delta)| \leq R^*(\hat{Z}_{0:1}, Z_0) := |p'(Z_0)||\hat{Z}_0| + |p''(Z_0)| r_1(|\hat{Z}_0|, |\hat{Z}_1|, |Z_0|) + r_2(|\hat{Z}_0|, |\hat{Z}_1|, |Z_0|) ,$$

where r_1 and r_2 are polynomials in their arguments. The order of x_2 in the polynomials $r_1(x_{0:2})$ and $r_2(x_{0:2})$ is 2 and 3 respectively. By Cauchy-Schwartz,

$$\begin{aligned} \mathbb{E}[|R(\hat{Z}_{0:1}, Z_{0:1}, \delta)|] \leq &\sqrt{\mathbb{E}[p'(Z_0)^2]}\sqrt{\mathbb{E}[\hat{Z}_0^2]} + \sqrt{\mathbb{E}[p''(Z_0)^2]}\sqrt{\mathbb{E}[r_1(|\hat{Z}_0|, |\hat{Z}_1|, |Z_0|)^2]} \\ &+ \mathbb{E}[|r_2(|\hat{Z}_0|, |\hat{Z}_1|, |Z_0|)|] . \end{aligned}$$

Recall that $\hat{Z}_i \sim \mathrm{N}(0,1)$ for any $i \in \{0,1\}$. Therefore, for any $i \in \{0,1\}$, $\mathbb{E}[|\hat{Z}_0|^k] < \infty$ for any finite $k \in \mathbb{N}$. Thus, by property (B) of Assumptions 4.3.12, $\mathbb{E}[R^*(\hat{Z}_{0:1}, Z_0)] \leq \kappa_* < \infty$ as required.

A.21 PROOF OF LEMMA 4.3.15

Firstly, since the random variables $Z_1^{(1:d)}, \ldots, Z_N^{(1:d)}$ are identically distributed, it suffices to consider the result for $k = 1$. Let C_1, C_2 and R be as defined in Lemma 4.3.14. Via the tower property of expectations, for any $i \in \{1, \ldots, d\}$,

$$\begin{aligned}\mathbb{E}[C_1(\hat{Z}_{0:1}^{(i)}, Z_0^{(i)})^2] &= \mathbb{E}[p'(Z_0^{(i)})^2(\hat{Z}_0^{(i)} + \hat{Z}_1^{(i)})^2] \\ &= \mathbb{E}[p'(Z_0^{(i)})^2\mathbb{E}[(\hat{Z}_0^{(i)} + \hat{Z}_1^{(i)})^2]] \,.\end{aligned}$$

$\hat{Z}_0^{(i)}$ and $\hat{Z}_1^{(i)}$ are standard normal random variables. Hence, by Lemma 4.3.13,

$$\begin{aligned}\mathbb{E}[C_1(\hat{Z}_{0:1}^{(i)}, Z_0^{(i)})^2] &= 2\mathbb{E}[p'(Z_0^{(i)})^2] \\ &= -2\mathbb{E}[p''(Z_0^{(i)}) - Z_0^{(i)}p'(Z_0^{(i)})] \\ &= -2\mathbb{E}\left\{\mathbb{E}\left[\frac{1}{2}p''(Z_0^{(i)})(\hat{Z}_0^{(i)} + \hat{Z}_1^{(i)})^2 - Z_0^{(i)}p'(Z_0^{(i)})\right]\right\} \\ &= -2\mathbb{E}[C_2(\hat{Z}_{0:1}^{(i)}, Z_0^{(i)})] \,.\end{aligned} \tag{135}$$

Next, we define

$$C_4(\hat{Z}_{0:1}^{(i)}, Z_0^{(i)}) := C_1(\hat{Z}_{0:1}^{(i)}, Z_0^{(i)})^2 + 2C_2(\hat{Z}_{0:1}^{(i)}, Z_0^{(i)}) \,,$$

so that $\mathbb{E}[C_4(\hat{Z}_{0:1}^{(i)}, Z_0^{(i)})] = 0$. Let

$$\begin{aligned}D_1^{(1)}(d) &:= \delta_d^3 \sum_{i=1}^{d} R(\hat{Z}_{0:1}^{(i)}, Z_{0:1}^{(i)}, \delta_d) \,, \\ D_1^{(2)}(d) &:= \frac{\delta_d^2}{2} \sum_{i=1}^{d} C_4(\hat{Z}_{0:1}^{(i)}, Z_0^{(i)}) \,, \\ D_1^{(3)}(d) &:= -\frac{\delta_d^2}{2} \sum_{i=1}^{d} C_1(\hat{Z}_{0:1}^{(i)}, Z_0^{(i)})^2 \,.\end{aligned}$$

Define $D_1(d) := D_1^{(1)}(d) + D_1^{(2)}(d) + D_1^{(3)}(d)$. Then,

$$D_1(d) = \delta_d^2 \sum_{i=1}^{d} C_2(\hat{Z}_{0:1}^{(i)}, Z_0^{(i)}) + \delta_d^3 \sum_{i=1}^{d} R(\hat{Z}_{0:1}^{(i)}, Z_{0:1}^{(i)}, \delta_d) \,.$$

We consider the limits in probability of each of the $D_1^{(j)}$ terms. Firstly,

$$|D_1^{(1)}(d)| \leq \frac{\lambda^3}{2^{3/2}d^{1/2}} \left(\frac{1}{d}\sum_{i=1}^{d} |R(\hat{Z}_{0:1}^{(i)}, Z_{0:1}^{(i)}, \delta_d)|\right) \leq \frac{\lambda^3}{2^{3/2}d^{1/2}} \left(\frac{1}{d}\sum_{i=1}^{d} R^*(\hat{Z}_{0:1}^{(i)}, Z_0^{(i)})\right) \,,$$

where, by Lemma 4.3.14, R^* is independent of $Z_1^{(i)}$ and δ_d, and, therefore, independent of d. Moreover, by the same lemma, R^* is such that $\mathbb{E}[R^*(\hat{Z}_{0:1}^{(i)}, Z_0^{(i)})] < \infty$. Hence, by the weak law of large numbers,

$$\operatorname*{plim}_{d\to\infty} \frac{1}{d}\sum_{i=1}^{d} R^*(\hat{Z}_{0:1}^{(i)}, Z_0^{(i)}) = \mathbb{E}[R^*(\hat{Z}_{0:1}^{(i)}, Z_0^{(i)})] < \infty \,.$$

Thus,

$$\operatorname*{plim}_{d\uparrow\infty} D_1^{(1)} = 0 \,.$$

Secondly,

$$D_1^{(2)}(d) = \frac{\lambda^2}{4}\left(\frac{1}{d}\sum_{i=1}^{d} C_4(\hat{Z}_{0:1}^{(i)}, Z_0^{(i)})\right) \,.$$

Recall that $\mathbb{E}[C_4(\hat{Z}_{0:1}^{(i)}, Z_0^{(i)})] = 0$. Thus, the weak law of large numbers gives

$$\operatorname*{plim}_{d\uparrow\infty} D_1^{(2)}(d) = 0 \, .$$

Thirdly,

$$D_1^{(3)}(d) = -\frac{\lambda^2}{4}\frac{1}{d}\sum_{i=1}^{d} C_1(\hat{Z}_{0:1}^{(i)}, Z_0^{(i)})^2 \, .$$

Equation (135) gives $\mathbb{E}[C_1(\hat{Z}_{0:1}^{(i)}, Z_0^{(i)})^2] = 2\varphi$. Thus, by the weak law of large numbers,

$$\operatorname*{plim}_{d\uparrow\infty} D_1^{(3)}(d) = -\frac{\lambda^2\varphi}{2} \, .$$

Hence, by Slutsky's theorem,

$$\operatorname*{plim}_{d\uparrow\infty} D_1(d) = -\frac{\lambda^2\varphi}{2} \, .$$

Now, define, for any $k \in \{1, \ldots, N\}$,

$$U_k(d) := \sum_{i=1}^{d} [p(Z_k^{(i)}) - p(Z_0^{(i)})] - D_k(d) \, .$$

Consider $k = 1$. By Lemma 4.3.14,

$$U_1(d) = \delta_d \sum_{i=1}^{d} C_1(\hat{Z}_{0:1}^{(i)}, Z_0^{(i)}) = \delta_d \sum_{i=1}^{d} p'(Z_0^{(i)})(\hat{Z}_0^{(i)} + \hat{Z}_1^{(i)}) \, .$$

Given $Z_0^{(1:d)} = z_0^{(1:d)}$, $U_1(d)$ is a linear combination of normal random variables and so is itself a normal random variable. The same holds true for any $U_j(d)$, where $j \in \{2, \ldots, N\}$. Thus $(U_1(d), \ldots, U_N(d) | Z_0^{(1:d)} = z_0^{(1:d)})$ is an N-dimensional normal random variable. Hence, it suffices to calculate its mean vector and variance matrix. Firstly, for any $j = 1, \ldots, N$,

$$\mathbb{E}[U_j(d) | Z_0^{(1:d)} = z_0^{(1:d)}] = \delta_d \sum_{i=1}^{d} p'(z_0^{(i)}) \mathbb{E}[\hat{Z}_0^{(i)} + \hat{Z}_j^{(i)}] = 0 \, ,$$

since $\hat{Z}_0^{(i)}$ and $\hat{Z}_j^{(i)}$ are standard normal random variables. Secondly, for any $(j_1, j_2) \in \{1, \ldots, N\}^2$, we have

$$\begin{aligned}
&\operatorname{Cov}[U_{j_1}(d) \, , \, U_{j_2}(d) \mid Z_0^{(1:d)} = z_0^{(1:d)}] \\
&= \delta_d^2 \sum_{i=1}^{d} \operatorname{Cov}[C_1(\hat{Z}_0^{(i)}, \hat{Z}_{j_1}^{(i)}, Z_0^{(i)}) \, , \, C_1(\hat{Z}_0^{(i)}, \hat{Z}_{j_2}^{(i)}, Z_0^{(i)}) \mid Z_0^{(1:d)} = z_0^{(1:d)}] \\
&= \delta_d^2 \sum_{i=1}^{d} \operatorname{Cov}[p'(Z_0^{(i)})(\hat{Z}_0^{(i)} + \hat{Z}_{j_1}^{(i)}) \, , \, p'(Z_0^{(i)})(\hat{Z}_0^{(i)} + \hat{Z}_{j_2}^{(i)}) \mid Z_0^{(1:d)} = z_0^{(1:d)}] \\
&= \delta_d^2 \sum_{i=1}^{d} p'(z_0^{(i)})^2 (1 + \mathbb{1}_{\{j_1\}}(j_2)) \, ,
\end{aligned}$$

since $C_1(\hat{Z}_{0:1}^{(j_1)}, Z_0^{(j_1)})$ is independent of $C_1(\hat{Z}_{0:1}^{(j_2)}, Z_0^{(j_2)})$ when $j_1 \neq j_2$, and since $(\hat{Z}_{0:1}^{(j_1)}, \hat{Z}_{0:1}^{(j_2)})$ is a sequence of independent standard normal random variables. Thus, if we let

$$T^{(d)} := \frac{1}{d}\sum_{i=1}^{d} p'(Z_0^{(i)})^2 \, ,$$

then, for any $j \in \{1, \ldots, N\}$,

$$(U_j(d) | T^{(d)} = t) = \sqrt{t/\varphi} U_j \, ,$$

where $U_j = A + B_j$, and $(A, B_{1:N})$ is a collection of independent random variables such that $A \sim \mathrm{N}(0, \lambda^2\varphi/2)$, and $B_j \sim \mathrm{N}(0, \lambda^2\varphi/2)$. Therefore, $U_j(d) = \sqrt{T^{(d)}/\varphi}U_j$. By the weak law of large numbers,

$$\operatorname*{plim}_{d\uparrow\infty} T^{(d)} = \mathbb{E}[p'(Z_0^{(i)})^2] = \varphi \ .$$

Hence, by Slutsky's theorem,

$$\operatorname*{dlim}_{d\uparrow\infty}(U_1(d), \ldots, U_N(d)) = U_{1:N} \ ,$$

as required.

A.22 PROOF OF LEMMA 4.3.16

First note that

$$\mathbb{E}[\|Z_1^{(1:d)} - Z_0^{(1:d)}\|^2] = \mathbb{E}\left[\sum_{i=1}^{d}(Z_1^{(i)} - Z_0^{(i)})^2\right] = d\mathbb{E}[(Z_1^{(1)} - Z_0^{(1)})^2] \ .$$

By definition

$$Z_1^{(1)} - Z_0^{(1)} = -\delta_d^2 Z_0^{(1)} + \delta_d\sqrt{2 - \delta_d^2}W \ ,$$

where W is a standard normal random variable. Hence,

$$\mathbb{E}(Z_1^{(1)} - Z_0^{(1)}) = -\delta_d^2\mu_1$$

where $\mu_1 := \mathbb{E}[Z_0^{(1)}]$. Moreover,

$$\begin{aligned}\mathbb{E}[(Z_1^{(1)} - Z_0^{(1)})^2] &= \mathrm{Var}(Z_1^{(1)} - Z_0^{(1)}) + \delta_d^4\mu_1^2 \\ &= \delta_d^4\mathrm{Var}(Z_0^{(1)}) + \delta_d^2(2 - \delta_d^2) + \delta_d^4\mu_1^2 \\ &= \delta_d^4(\mu_2 - 1) + 2\delta_d^2 \ ,\end{aligned}$$

where $\mu_2 := \mathbb{E}[(Z_0^{(1)})^2]$. Thus, given property (B.a) of Assumptions 4.3.12,

$$\lim_{d\uparrow\infty} \mathbb{E}[\|Z_1^{(1:d)} - Z_0^{(1:d)}\|^2] = \lim_{d\uparrow\infty}\{d\mathbb{E}[(Z_1^{(1)} - Z_0^{(1)})^2]\} = \lambda^2 \ . \tag{136}$$

Secondly, note that, by Equation (136),

$$\begin{aligned}\lim_{d\uparrow\infty} \mathbb{E}[(\|Z_1^{(1:d)} - Z_0^{(1:d)}\|^2 - \lambda^2)^2] &= \lim_{d\uparrow\infty} \mathrm{Var}(\|Z_1^{(1:d)} - Z_0^{(1:d)}\|^2) \\ &= \lim_{d\uparrow\infty} d\mathrm{Var}[(Z_1^{(1)} - Z_0^{(1)})^2] \ .\end{aligned}$$

Now,

$$\mathrm{Var}[(Z_1^{(1)} - Z_0^{(1)})^2] = \mathrm{Var}[\delta_d^4(Z_0^{(1)})^2 + \delta_d^2(2 - \delta_d^2)W^2 - 2\delta_d^3\sqrt{2 - \delta_d^2}Z_0^{(1)}W] \ .$$

W is a standard normal random variable that is independent of $Z_0^{(1)}$. Thus, $\mathbb{E}(W^3) = \mathbb{E}(W) = 0$, and, $\mathrm{Cov}(W^2 \ , \ Z_0^{(1)}W) = 0$. Hence,

$$\begin{aligned}\mathrm{Var}[(Z_1^{(1)} - Z_0^{(1)})^2] &= \delta_d^8\mathrm{Var}[(Z_0^{(1)})^2] + \delta_d^4(2 - \delta_d^2)^2\mathrm{Var}(W^2) + 4\delta_d^6(2 - \delta_d^2)\mathrm{Var}[Z_0^{(1)}W] \\ &= \delta_d^8(\mu_4 - \mu_2^2) + \delta_d^4(2 - \delta_d^2)^2 + 4\delta_d^6(2 - \delta_d^2)\mu_2 \ ,\end{aligned}$$

where $\mu_4 := \mathbb{E}[(Z_0^{(1)})^4]$. Thus, $\mathrm{Var}[(Z_1^{(1)} - Z_0^{(1)})^2]$ is a polynomial in δ_d whose largest power of d is in the term $\delta_d^4 = \lambda^4/(4d^2)$. Therefore, by property (B.a) of Assumptions 4.3.12,

$$\lim_{d\uparrow\infty} \mathbb{E}[(\|Z_1^{(1:d)} - Z_0^{(1:d)}\|^2 - \lambda^2)^2] = 0 \ .$$

A.23 PROOF OF THEOREM 4.3.17

Consider the multiple-proposal extension of Barker's acceptance probability,

$$\alpha^*_{k,N}(w^*_{0:N}) = \frac{w^*_k}{w^*_0 + \ldots + w^*_N} .$$

Then,

$$\sum_{k=1}^{N} \alpha^*_{k,N}(w^*_{0:N}) = 1 - \frac{w^*_0}{w^*_0 + \ldots w^*_N} = 1 - \left[1 + \sum_{k=1}^{N} \frac{w^*_k}{w^*_0}\right]^{-1} .$$

By Definition 4.3.11,

$$g^*(z^{(1:d)}) = w^*[h^*(z^{(1:d)})] = \prod_{i=1}^{d} w[h(z^{(i)})] = \prod_{i=1}^{d} g(z^{(i)}) .$$

Thus,

$$\begin{aligned}\alpha^*_{k,N}(g^*(Z_0^{(1:d)}), \ldots, g^*(Z_N^{(1:d)})) &= 1 - \left[1 + \sum_{k=1}^{N} \prod_{i=1}^{d} \frac{g(Z_k^{(i)})}{g(Z_0^{(i)})}\right]^{-1} \\ &= 1 - \left\{1 + \sum_{k=1}^{N} \exp\left[\sum_{i=1}^{d} p(Z_k^{(i)}) - p(Z_0^{(i)})\right]\right\}^{-1} ,\end{aligned}$$

where $p(z) := \log[g(z)]$. By Lemma 4.3.15,

$$\alpha^*_{k,N}(g^*(Z_0^{(1:d)}), \ldots, g^*(Z_N^{(1:d)})) = 1 - \left\{1 + \sum_{k=1}^{N} \exp[D_k(d) + U_k(d)]\right\}^{-1} ,$$

where, for any $\{D_k(d) : (k,d) \in \{1,\ldots,N\} \times \mathbb{N}\}$ is a collection of random variables such that

$$\operatorname*{plim}_{d\uparrow\infty} D_k(d) = -\frac{\lambda^2\varphi}{2} ,$$

and $\{U_k(d) : (k,d) \in \{1,\ldots,N\} \times \mathbb{N}\}$ is a collection of random variables such that

$$\operatorname*{dlim}_{d\uparrow\infty}(U_1(d), \ldots, U_N(d)) = (U_1, \ldots, U_N) ,$$

where, for each $k \in \{1,\ldots,N\}$, $U_k := A + B_k$, and $(A, B_{1:N})$ is a collection of independent random variables where

$$A \sim \mathrm{N}\left(0, \frac{\lambda^2\varphi}{2}\right) ,$$

and, for any $k \in \{1,\ldots,N\}$,

$$B_k \sim \mathrm{N}\left(0, \frac{\lambda^2\varphi}{2}\right) .$$

The non-negative function $f : \mathbb{R}^{2N} \to [0,1]$ defined by

$$f(u_{1:N}, d_{1:N}) := \left[1 + \sum_{k=1}^{N} \exp(d_k + u_k)\right]^{-1}$$

is continuous and bounded above by 1, hence, by the continuous mapping theorem and the dominated convergence theorem, we have

$$\begin{aligned}\lim_{d\uparrow\infty} \alpha(\lambda d^{-1/2}) &= 1 - \lim_{d\uparrow\infty} \mathbb{E}\left[\left\{1 + \sum_{k=1}^{N} \exp(D_k(d) + U_k(d))\right\}^{-1}\right] \\ &= 1 - \mathbb{E}\left[\left\{1 + \exp(-\lambda^2\varphi/2)\exp(A)\sum_{k=1}^{N} \exp(B_k)\right\}^{-1}\right]\end{aligned}$$

Letting $\xi := \lambda\sqrt{\varphi}/\sqrt{2}$, $W_0 := A/\xi$, and, for any $k \in \{1, \ldots, N\}$, $W_k := B_k/\xi$, gives (96). Next, consider

$$\begin{aligned}
&|J(\lambda d^{-1/2}) - \lambda^2 \bar{\alpha}_b(\lambda)| \\
&= \left| \mathbb{E}\left[\sum_{k=1}^{N} \alpha^*_{k,N}(g^*(Z_0^{(1:d)}), \ldots, g^*(Z_N^{(1:d)})) \|Z_k^{(1:d)} - Z_0^{(1:d)}\|^2 \right] - \lambda^2 \bar{\alpha}_b(\lambda) \right| \\
&\leq \left| \mathbb{E}\left[\sum_{k=1}^{N} \alpha^*_{k,N}(g^*(Z_0^{(1:d)}), \ldots, g^*(Z_N^{(1:d)}))(\|Z_k^{(1:d)} - Z_0^{(1:d)}\|^2 - \lambda^2) \right] \right| + \left| \lambda^2(\alpha(\lambda d^{-1/2}) - \bar{\alpha}_b(\lambda)) \right| .
\end{aligned}$$

The second term tends towards zero as d tends towards ∞ by Lemma 4.3.16. Moreover, using the Cauchy-Schwartz inequality and the fact that $\alpha^*_{k,N}(w^*_{1:N}) \leq 1$ for any $k \in \{1, \ldots, N\}$, and any $w^*_{1:N} \in [0, \infty)^N$, the first term is equal to

$$N|\mathbb{E}[\alpha^*_{1,N}(g^*(Z_0^{(1:d)}), \ldots, g^*(Z_N^{(1:d)}))(\|Z_1^{(1:d)} - Z_0^{(1:d)}\|^2 - \lambda^2)]| \leq N\{\mathbb{E}[(\|(Z_1^{(1:d)} - Z_0^{(1:d)}\|^2 - \lambda^2)^2]|\}^{1/2} ,$$

which, via Lemma 4.3.16, tends towards zero as d tends towards ∞. Next, consider the multiple-proposal extension of the Metropolis-Hastings acceptance probability;

$$\alpha^*_{k,N}(w^*_{0:N}) = \frac{w^*_k}{w^*_0 + \ldots + w^*_N - [w^*_k \wedge w^*_0]} .$$

The proof follows from similar reasoning. First,

$$\sum_{k=1}^{N} \alpha^*_{k,N}(w^*_{0:N}) = \sum_{k=1}^{N} \frac{w^*_k/w^*_0}{1 + w^*_1/w^*_0 + \ldots w^*_N/w^*_0 - [1 \wedge w^*_k/w^*_0]} .$$

As previously, by Lemma 4.3.15, for any $k \in \{1, \ldots, N\}$,

$$\frac{w^*_k}{w^*_0} = \exp\left[\sum_{i=1}^{d} p(Z_k^{(i)}) - p(Z_0^{(i)}) \right] = \exp[D_k(d) - U_k(d)] .$$

The non-negative function $f : \mathbb{R}^{2N} \to [0, 1]$ defined by

$$f(u_{1:N}, d_{1:N}) := \sum_{k=1}^{N} \frac{\exp(d_k + u_k)}{1 + \exp(d_1 + u_1) + \ldots + \exp(d_N + u_N) - [1 \wedge \exp(d_k + u_k)]} ,$$

is continuous and bounded above by 1, hence, by the continuous mapping theorem and the dominated convergence theorem, we have

$$\begin{aligned}
\lim_{d \uparrow \infty} \alpha(\lambda d^{-1/2}) &= \lim_{d \uparrow \infty} \mathbb{E}\left[\sum_{k=1}^{N} \alpha^*_{k,N}(g^*(Z_0^{(1:d)}), \ldots, g^*(Z_N^{(1:d)})) \right] \\
&= \mathbb{E}\left[\sum_{k=1}^{N} \frac{\exp(-\xi^2)\exp(A)\exp(B_k)}{1 + \exp(-\xi^2)\exp(A)s(B_{1:N}) - [1 \wedge \exp(-\xi^2)\exp(A)\exp(B_k)]} \right] ,
\end{aligned}$$

where,

$$s(b_{1:N}) = \sum_{k=1}^{N} \exp(b_k) ,$$

and, as before, $\xi := \lambda\sqrt{\varphi}/\sqrt{2}$. Now, letting $W_0 := A/\xi$, and, for any $k \in \{1, \ldots, N\}$, $W_k := B_k/\xi$, gives (98). The proof of (99) follows in exactly the same manner as the proof of (97) and, thus, is omitted.

A.24 PROOF OF THEOREM 4.4.3

For any $t \in \{0, \ldots, T\}$ let $I_t : \mathbb{R}^p \times \mathbb{R}^{d \times N} \times \ldots \times \mathbb{R}^{d \times (t+1) \times N} \to [0, \infty)$ be defined by

$$I_t(\theta, \tilde{X}_0^{(1:N)}, \ldots, \tilde{X}_t^{(1:N)}) := \frac{1}{N^{t+1}} \prod_{s=0}^{t} \sum_{j=1}^{N} w_s(\tilde{X}_s^{(j)}) .$$

Moreover, recursively define $\mu_t : \mathbb{R}^p \times \mathbb{R}^{d\times(t+1)} \to [0,\infty)$ by

$$\begin{aligned}\mu_T(\theta, x_{0:T}) &= 1 \ , \\ \mu_t(\theta, x_{0:t}) &= \int_{\mathbb{R}^d} \frac{\gamma_{t+1}(\theta, x_{0:t+1})}{\gamma_t(\theta, x_{0:t})} \mu_{t+1}(\theta, x_{0:t+1}) \, \mathrm{d}x_{t+1} \ .\end{aligned}$$

Furthermore, for any $t \in \{0, \ldots, T-1\}$, define the filtration, $\mathcal{F}_t$, by

$$\mathcal{F}_t := (\tilde{X}_0^{(1:N)}, \ldots, \tilde{X}_t^{(1:N)}, A_0^{(1:N)}, \ldots, A_{t-1}^{(1:N)}) \ .$$

Note that

$$I_{t+1}(\theta, \tilde{X}_0^{(1:N)}, \ldots, \tilde{X}_{t+1}^{(1:N)}) \sum_{i=1}^N \tilde{w}_{t+1}^{(i)}(\tilde{X}_{t+1}^{(1:N)}; \theta) = \frac{1}{N} I_t(\theta, \tilde{X}_0^{(1:N)}, \ldots, \tilde{X}_t^{(1:N)}) \sum_{i=1}^N w_{t+1}(\tilde{X}_{t+1}^{(i)}; \theta) \ .$$

Therefore, dropping the explicit arguments of I_t and I_{t+1},

$$\begin{aligned}
&\mathbb{E}\Big[I_{t+1} \sum_{i=1}^N \tilde{w}_{t+1}^{(i)}(\tilde{X}_{t+1}^{(1:N)}; \theta) \mu_{t+1}(\theta, \tilde{X}_{t+1}^{(i)}) | \mathcal{F}_t \Big] \\
&= \frac{1}{N} I_t \mathbb{E}\Big[\sum_{i=1}^N w_{t+1}(\tilde{X}_{t+1}^{(i)}; \theta) \mu_{t+1}(\theta, \tilde{X}_{t+1}^{(i)}) | \mathcal{F}_t \Big] \\
&= \frac{1}{N} I_t \mathbb{E}\Big[\sum_{i=1}^N \int_{\mathbb{R}^{d\times N}} \frac{\gamma_{t+1}(\theta, \tilde{X}_t^{(A_t^{(i)})}, x_{t+1}^{(i)}) \mu_{t+1}(\theta, \tilde{X}_t^{(A_t^{(i)})}, x_{t+1}^{(i)})}{\gamma_t(\theta, \tilde{X}_t^{(A_t^{(i)})}) p_{t+1}(x_{t+1}^{(i)} | \tilde{X}_t^{(A_t^{(i)})}, \theta)} \\
&\qquad \times p_{t+1}^*(x_{t+1}^{(1:N)} | \tilde{X}_t^{(A_t^{(1:N)})}, \theta) \, \mathrm{d}x_{t+1}^{(1:N)} | \mathcal{F}_t \Big] \\
&= \frac{1}{N} I_t \mathbb{E}\Big[\sum_{i=1}^N \int_{\mathbb{R}^d} \frac{\gamma_{t+1}(\theta, \tilde{X}_t^{(A_t^{(i)})}, x_{t+1}^{(i)}) \mu_{t+1}(\theta, \tilde{X}_t^{(A_t^{(i)})}, x_{t+1}^{(i)})}{\gamma_t(\theta, \tilde{X}_t^{(A_t^{(i)})}) p_{t+1}(x_{t+1}^{(i)} | \tilde{X}_t^{(A_t^{(i)})}, \theta)} \\
&\qquad \times p_{t+1}(x_{t+1}^{(i)} | \tilde{X}_t^{(A_t^{(i)})}, \theta) \, \mathrm{d}x_{t+1}^{(i)} | \mathcal{F}_t \Big] \\
&= \frac{1}{N} I_t \mathbb{E}\Big[\sum_{i=1}^N \mu_t(\theta, \tilde{X}_t^{(A_t^{(i)})}) | \mathcal{F}_t \Big] \\
&= \frac{1}{N} I_t \mathbb{E}\Big[\sum_{j=1}^N \mu_t(\theta, \tilde{X}_t^{(j)}) \sum_{i=1}^N \mathbb{1}_{\{j\}}(A_t^{(i)}) | \mathcal{F}_t \Big] \\
&= I_t \sum_{j=1}^N \tilde{w}_t^{(j)}(\tilde{X}_t^{(1:N)}; \theta) \mu_t(\theta, \tilde{X}_t^{(j)}) \ ,
\end{aligned}$$

where the fourth line follows from property (M) of Assumptions 4.4.1, and the last line follows from property (U) of Assumptions 4.4.2. Thus,

$$
\begin{aligned}
&\mathbb{E}_\Psi[I_T(\Theta, \tilde{X}_0^{(1:N)}, \ldots, \tilde{X}_T^{(1:N)})|\Theta = \theta] \\
&= \mathbb{E}_\Psi\left[I_T \sum_{i=1}^N \tilde{w}_T^{(i)}(\tilde{X}_T^{(1:N)}; \theta)\mu_T(\theta, \tilde{X}_T^{(i)})\right] \\
&= \mathbb{E}_\Psi\left[I_0 \sum_{i=1}^N \tilde{w}_0^{(i)}(\tilde{X}_0^{(1:N)}; \theta)\mu_0(\theta, \tilde{X}_0^{(i)})\right] \\
&= \frac{1}{N}\mathbb{E}\left[\sum_{i=1}^N w_0(\tilde{X}_0^{(i)}; \theta)\mu_0(\theta, \tilde{X}_0^{(i)})\right] \\
&= \frac{1}{N}\sum_{i=1}^N \int_{\mathbb{R}^{d\times N}} \frac{\gamma_0(\theta, x_0^{(i)})}{p_0(x_0^{(i)}|\theta)}\mu_0(\theta, x_0^{(i)})p_0^*(x_0^{(1:N)}|\theta)\, \mathrm{d}x_0^{(1:N)} \\
&= \frac{1}{N}\sum_{i=1}^N \int_{\mathbb{R}^d} \frac{\gamma_0(\theta, x_0^{(i)})}{p_0(x_0^{(i)}|\theta)}\mu_0(\theta, x_0^{(i)})p_0(x_0^{(i)}|\theta)\, \mathrm{d}x_0^{(i)} \\
&= \int_{\mathbb{R}^d} \gamma(\theta, x_0) \int_{\mathbb{R}^d} \frac{\gamma_1(\theta, x_{0:1})}{\gamma_0(\theta, x_0)}\mu_1(\theta, x_{0:1})\, \mathrm{d}x_1 \mathrm{d}x_0 \\
&= \int_{\mathbb{R}^{d(T+1)}} \gamma_0(\theta, x_0) \prod_{t=0}^{T-1} \frac{\gamma_{t+1}(\theta, x_{0:t+1})}{\gamma_t(\theta, x_{0:t})}\, \mathrm{d}x_{0:T} \\
&= \int_{\mathbb{R}^{d(T+1)}} \gamma_T(\theta, x_{0:T})\, \mathrm{d}x_{0:T}\ ,
\end{aligned}
$$

where the sixth line follows from property (M) of Assumptions 4.4.1.

A.25 PROOF OF THEOREM 4.4.7

Assertion (111) follows immediately from (96) of Theorem 4.3.17. To demonstrate (112) we start by showing that, for any $t \in \{0, \ldots, T-1\}$,

$$J_{t+1,N}(\epsilon) = \alpha_{t+1,N}(\epsilon)J_{t,N}(\epsilon)\ .$$

To see this, note that, for any $k \in \{1, \ldots, N\}$, $\mathcal{L}_{t+1}(k, 0)$ depends only on the values $\mathcal{L}_{t+1}(k, 1), \ldots, \mathcal{L}_{t+1}(k, t)$ which, themselves, depend only on the values $A_{0:t}^{(0:N)}$. For any $s \in \{0, \ldots, t\}$, the values $A_s^{(0:N)}$ depend only on the values

$$\alpha_{0,N}^*(g^*(Z_{s,0}^{(1:d)}), \ldots, g^*(Z_{s,N}^{(1:d)})), \ldots, \alpha_{N,N}^*(g^*(Z_{s,0}^{(1:d)}), \ldots, g^*(Z_{s,N}^{(1:d)}))\ .$$

Therefore, by linearity and independence,

$$
\begin{aligned}
J_{t+1,N}(\epsilon) &= \mathbb{E}\left[\sum_{k=1}^N \alpha_{k,N}^*(g^*(Z_{t+1,0}^{(1:d)}), \ldots, g^*(Z_{t+1,N}^{(1:d)}))\|Z_{0,\mathcal{L}_{t+1}(k,0)}^{(1:d)} - Z_{0,0}^{(1:d)}\|^2\right] \\
&= \sum_{k=1}^N \mathbb{E}[\alpha_{k,N}^*(g^*(Z_{t+1,0}^{(1:d)}), \ldots, g^*(Z_{t+1,N}^{(1:d)}))]\mathbb{E}[\|Z_{0,\mathcal{L}_{t+1}(k,0)}^{(1:d)} - Z_{0,0}^{(1:d)}\|^2]\ .
\end{aligned}
$$

For any $k \in \{1, \ldots, N\}$, $\mathcal{L}_{t+1}(k, t) = j$, where $j \in \{1, \ldots, N\}$, with probability

$$\alpha_{j,N}^*(g^*(Z_{t,0}^{(1:d)}), \ldots, g^*(Z_{t,N}^{(1:d)}))\ .$$

Hence,

$$
\begin{aligned}
J_{t+1,N}(\epsilon) &= \sum_{k=1}^N \mathbb{E}[\alpha_{k,N}^*(g^*(Z_{t+1,0}^{(1:d)}), \ldots, g^*(Z_{t+1,N}^{(1:d)}))] \\
&\quad \times \mathbb{E}\left[\sum_{j=1}^N \alpha_{j,N}^*(g^*(Z_{t,0}^{(1:d)}), \ldots, g^*(Z_{t,N}^{(1:d)}))\|Z_{0,\mathcal{L}_t(j,0)}^{(1:d)} - Z_{0,0}^{(1:d)}\|^2\right] \\
&= \alpha_{t+1,N}(\epsilon)J_{t,N}(\epsilon)\ .
\end{aligned}
$$

Applying this result repeatedly gives

$$J_{T,N}(\epsilon) = \prod_{t=1}^{T} \alpha_{t,N}(\epsilon) J_{0,N}(\epsilon) \ .$$

(112) now follows from (111) and (97) of Theorem 4.3.17.

A.26 PROOF OF THEOREM 4.4.8

From (111),

$$\rho_N(\lambda) = \mathbb{E}\left[\left\{1 + \exp(-\xi^2)\exp(\xi W_0)\sum_{k=1}^{N}\exp(\xi W_k)\right\}^{-1}\right] ,$$

where $\xi := \lambda\sqrt{\varphi}/\sqrt{2}$, $\varphi := \mathbb{E}[q'(Z_0^2)]$, $Z_0 \sim \pi_Z$ where π_Z is as defined in Definition 4.4.5, and $W_{0:N}$ is an independent sequence of one-dimensional standard Normal random variables. The function $f : [0,\infty) \to (0,1]$ defined by $f(x) : 1/(1+x)$ is convex. Moreover,

$$\begin{aligned}\mathbb{E}\left[\exp(-\xi^2)\exp(\xi W_0)\sum_{k=1}^{N}\exp(\xi W_k)\right] &= N\exp(-\xi^2)\mathbb{E}[\exp(\xi W_0)]\mathbb{E}[\exp(\xi W_1)] \\ &= N\exp(-\xi^2)\exp(\xi^2/2)\exp(\xi^2/2) \\ &= N \ .\end{aligned}$$

Therefore, $\rho_N(\lambda) \geq 1/(N+1)$. Furthermore, by Jensen's inequality,

$$\sum_{k=1}^{N}\exp(\xi W_k) \geq N\exp\left(\frac{\xi}{N}\sum_{k=1}^{N} W_k\right) .$$

Hence,

$$\begin{aligned}\exp(\xi W_0)\sum_{k=1}^{N}\exp(\xi W_k) &\geq N\exp\left(\xi W_0 + \frac{\xi}{N}\sum_{k=1}^{N} W_k\right) \\ &= N\exp\left(\xi\sqrt{1+\frac{1}{N}}Z\right) ,\end{aligned}$$

where $Z \sim \mathrm{N}(0,1)$, thereby demonstrating the upper bound.

A.27 PROOF OF COROLLARY 4.4.9

Note, from Theorem 4.4.8, that

$$1 \leq \lim_{N\uparrow\infty} N\rho_N(\lambda) \leq \mathbb{E}[\exp(\xi^2 - \xi Z)] \ .$$

Hence, there exist two constants, $\tilde{c}_1$ and $\tilde{c}_2$, and an $N_0 \in \mathbb{N}$ such that, for any $N \geq N_0$,

$$\tilde{c}_1 \leq N\rho_N(\lambda) \leq \tilde{c}_2 \ .$$

Note that

$$Ne_{T,N}(\lambda) = \lambda^2(1-\rho_N(\lambda))^{(T+1)} = \lambda^2\left(1 - \frac{N\rho_N(\lambda)}{N}\right)^{(T+1)} .$$

Hence, for any $N \geq N_0$,

$$\lambda^2\left(1-\frac{\tilde{c}_2}{N}\right)^{(T+1)} \leq Ne_{T,N}(\lambda) \leq \lambda^2\left(1-\frac{\tilde{c}_1}{N}\right)^{(T+1)} .$$

The result now follows immediately.

B

TECHNICAL LEMMATA

LEMMA B.0.1. *Let $f, g : \mathcal{X} \to [0, \infty)$ be two positive functions defined on $\mathcal{X}$. Then the following relation*

$$f(x) \wedge g(y) = \int_0^\infty \mathbb{1}_{[0,f(x)]}(s)\mathbb{1}_{[0,g(y)]}(s)\, \mathrm{d}s \,,$$

where $a \wedge b$ is the minimum of a and b, holds.

Proof. Note that $h(x, y, s) := \mathbb{1}_{[0,f(x)]}(s)\mathbb{1}_{[0,g(y)]}(s)$ is equal to 1 if and only if $s \leq f(x)$ and $s \leq g(y)$, otherwise $h(x, y, s) = 0$. Hence

$$\int_0^\infty \mathbb{1}_{[0,f(x)]}(s)\mathbb{1}_{[0,g(y)]}(s)\, \mathrm{d}s = \int_0^{f(x)\wedge g(y)} \mathrm{d}s = f(x) \wedge g(y)$$

as required. □

LEMMA B.0.2. *Let α, r_1, and r_2 be non-negative numbers such that $r_2 \geq r_1$. Then*

$$\frac{r_1 + \alpha}{r_2 + \alpha} \geq \frac{r_1}{r_2} \,.$$

Proof. Note that, by assumption, $\alpha r_2 \geq \alpha r_1$. Hence $r_1 r_2 + \alpha r_2 \geq r_1 r_2 + \alpha r_1$. Dividing by $r_2(r_2 + \alpha)$ gives the result. □

LEMMA B.0.3. *Let $N \in \mathbb{N}$, and let $z_{0:N}$ be a sequence of non-negative numbers. Define $I = \{i \in \{1, \ldots, N\} : z_0 \geq z_i\}$. Then, the following inequality holds;*

$$\sum_{k=1}^N \frac{z_k}{z_0 + \ldots + z_N - [z_0 \wedge z_k]} \geq \begin{cases} \dfrac{1}{|I| z_0} \displaystyle\sum_{i \in I} z_i & \text{if } I \neq \emptyset \\ 1 & \text{if } I = \emptyset \end{cases} ,$$

where $|I|$ denotes the cardinality of the set I.

Proof. Firstly, consider the case where $I = \emptyset$. Then, by definition, $z_k > z_0$ for any $k = 1, \ldots, N$. Hence

$$\sum_{k=1}^N \frac{z_k}{z_0 + \ldots + z_N - [z_0 \wedge z_k]} = \sum_{k=1}^N \frac{z_k}{z_1 + \ldots + z_N} = 1 \,.$$

Next, consider the case where $I \neq \emptyset$. Then,

$$z_1 + \ldots + z_N \leq |I| z_0 + \sum_{i \notin I} z_i \,.$$

Moreover, if $k \in I$, then

$$z_0 + \ldots + z_N - z_k \leq |I| z_0 + \sum_{i \notin I} z_i \,.$$

Therefore, defining

$$S_J := \sum_{j \in J} z_j \,,$$

the following inequality holds;

$$\sum_{k=1}^N \frac{z_k}{z_0 + \ldots + z_N - [z_0 \wedge z_k]} \geq \sum_{k=1}^N \frac{z_k}{|I| z_0 + S_{I^c}} \,,$$

Hence, by Lemma B.0.2,

$$\sum_{k=1}^{N} \frac{z_k}{z_0 + \ldots + z_N - [z_0 \wedge z_k]} \geq \frac{S_I + S_{I^c}}{|I| z_0 + S_{I^c}} \geq \frac{1}{|I| z_0} \sum_{k \in I} z_k \ ,$$

since $S_{I^c} \geq 0$. □

LEMMA B.0.4. *Let $\eta > 0$ and $\psi > 0$. Then*

$$0 \geq 1 - \psi^\eta + \eta \log(\psi) \geq -\frac{\eta^2}{2} \log(\psi)^2 \ .$$

Proof. By Taylor's theorem there exists a ϕ with $|\phi| \leq |\eta \log(\psi)|$ such that

$$\psi^\eta = \exp(\eta \log(\psi)) = 1 + \eta \log(\psi) + \phi^2/2 \ .$$

Hence

$$1 - \psi^\eta + \eta \log(\psi) = -\phi^2/2 \ .$$

The result follows since $\eta^2 \log(\psi)^2 \geq \phi^2 \geq 0$. □

LEMMA B.0.5. *Suppose $z_2 \geq z_1 > 0$. Then, for any $\eta \in (0, 1)$,*

$$(1 + z_2)^\eta - (1 + z_1)^\eta < (z_2 - z_1) \ .$$

Proof. By the mean value theorem, there exists a $z_* \in (z_1, z_2)$ such that

$$(1 + z_2)^\eta - (1 + z_1)^\eta = \eta (1 + z_*)^{\eta - 1} (z_2 - z_1) \ .$$

The result follows since $\eta \in (0, 1)$ and $z_* > 0$. □

LEMMA B.0.6. *Let $f : [a, b] \to \mathbb{R}$ be a k-times differentiable function. Then, for any $j \leq k$, and any $(x_1, x_2) \in (a, b)^2$, there exists a $t := t(x_1, x_2) \in (-1, 1)$ such that*

$$f(x_2) - f(x_1) = \sum_{i=1}^{j} \frac{(x_2 - x_1)^i}{i!} f^{(i)}(x_1) + \frac{(x_2 - x_1)^j}{j!} [f^{(j)}(x_1 + t(x_2 - x_1)) - f^{(j)}(x_1)] \ . \tag{137}$$

Proof. This lemma is a simple application of Taylor's theorem. Indeed, there exists a ξ satisfying $|\xi - x_1| < |x_2 - x_1|$ such that

$$f(x_2) - f(x_1) = \sum_{i=1}^{j-1} \frac{(x_2 - x_1)^i}{i!} f^{(i)}(x_1) + \frac{(x_2 - x_1)^j}{j!} f^{(j)}(\xi) \ .$$

Let $t := (\xi - x_1)/(x_2 - x_1)$. Then, by assumption, $|t| < 1$ and, since

$$f^{(j)}(\xi) = f^{(j)}(x_1 + t(x_2 - x_1)) = f^{(j)}(x_1) + [f^{(j)}(x_1 + t(x_2 - x_1)) - f^{(j)}(x_1)] \ ,$$

equation (137) holds. □

LEMMA B.0.7. *Let $Z \sim \mathrm{N}(0, 1)$ be a one-dimensional standard normal random variable. Then, for any $(a, b, c, d, z_0, z_1) \in \mathbb{R}^6$,*

$$\begin{aligned} &\mathbb{E}[(aZ + b) \exp(cZ + d) \mathbb{1}_{[z_0, z_1]}(Z)] \\ &= \exp(c^2/2 + d) \{a[\phi(z_0 - c) - \phi(z_1 - c)] + (ac + b)[\Phi(z_1 - c) - \Phi(z_0 - c)]\} \ , \end{aligned}$$

where ϕ and Φ respectively denote the probability density function and the cumulative distribution function of a one-dimensional standard normal random variable.

Proof. A straightforward calculation shows that

$$\begin{aligned}
&\mathbb{E}[(aZ+b)\exp(cZ+d)\mathbb{1}_{[z_0,z_1]}(Z)] \\
&= \int_{z_0}^{z_1} (az+b)\exp(cz+d)\phi(z)\,\mathrm{d}z \\
&= \exp(c^2/2+d)\int_{z_0}^{z_1} (az+b)\phi(z-c)\,\mathrm{d}z \\
&= \exp(c^2/2+d)\int_{z_0}^{z_1} (a(z-c)+(ac+b))\phi(z-c)\,\mathrm{d}z\,,
\end{aligned}$$

from which the result follows. □

LEMMA B.0.8. *Let $X := [X_1, X_2]^T$ be an $(m+n)$-dimensional random variable with a $\mathrm{N}_{m+n}(\mu, \Sigma)$ distribution, where*

$$\mu := \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}\,, \quad \Sigma := \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{12}^T & \Sigma_{22} \end{pmatrix}\,,$$

and $\mu_1 \in \mathbb{R}^m$, $\mu_2 \in \mathbb{R}^n$, $\Sigma_{11} \in \mathbb{R}^{m\times m}$, $\Sigma_{12} \in \mathbb{R}^{m\times n}$, $\Sigma_{22} \in \mathbb{R}^{n\times n}$. Then

$$(X_1|X_2 = x_2) \sim N_m(\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2-\mu_2)\,,\ \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{12}^T)\,.$$

Proof. Let $\tilde{q}(x_1|x_2)$ be the density corresponding to the random variable $X_1|X_2 = x_2$, and let q be the density corresponding to the random variable X. Then

$$\tilde{q}(x_1|x_2) \propto q(x) \propto \exp(-(x-\mu)^T\bar{\Sigma}(x-\mu)/2)\,,$$

where

$$\bar{\Sigma} := \Sigma^{-1} = \begin{pmatrix} \bar{\Sigma}_{11} & \bar{\Sigma}_{12} \\ \bar{\Sigma}_{12}^T & \bar{\Sigma}_{22} \end{pmatrix}\,,$$

and $\bar{\Sigma}_{11} \in \mathbb{R}^{m\times m}$, $\bar{\Sigma}_{12} \in \mathbb{R}^{m\times n}$, $\bar{\Sigma}_{22} \in \mathbb{R}^{n\times n}$. Thus, a direct calculation gives

$$\tilde{q}(x_1|x_2) \propto \exp(-(x_1-\mu_1+\bar{\Sigma}_{11}^{-T}\bar{\Sigma}_{12}(x_2-\mu_2))^T\bar{\Sigma}_{11}(x_1-\mu_1+\bar{\Sigma}_{11}^{-T}\bar{\Sigma}_{12}(x_2-\mu_2))/2)\,.$$

Therefore,

$$X_1|X_2 = x_2 \sim \mathrm{N}_m(\mu_1 - \bar{\Sigma}_{11}^{-T}\bar{\Sigma}_{12}(x_2-\mu_2)\,,\ \bar{\Sigma}_{11}^{-1})\,.$$

The inverse of a block matrix is well established, and the following is easy to assert;

$$\bar{\Sigma}_{11} = (\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{12}^T)^{-1}\,, \quad \bar{\Sigma}_{12} = -\bar{\Sigma}_{11}\Sigma_{12}\Sigma_{22}^{-1}\,.$$

Thus, by symmetry of Σ_{11} and Σ_{22},

$$\bar{\Sigma}_{11}^{-1} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{12}^T\,, \quad \hat{\Sigma}_{11}^{-T}\hat{\Sigma}_{12} = -\Sigma_{12}\Sigma_{22}^{-1}\,,$$

as required. □

DEFINITION B.0.9 (Binary Search). *For a sorted sequence $x_{1:N}$; that is, a sequence such that $x_i \leq x_j$ whenever $i \neq j$, a binary search for the location of y in $x_{1:N}$; denoted* $\mathrm{binsearch}(x_{1:N}, y)$*, is an algorithm (Algorithm 23) that returns the unique integer $k \in \{0, \ldots, N+1\}$ such that $y > x_i$ for any $i \leq k$, and $y \leq x_j$ for any $j > k$.*

LEMMA B.0.10. *The binary search algorithm,* $\mathrm{binsearch}(x_{1:N}, y)$*, takes at most $1 + \lceil \log_2(2\lfloor N/2 \rfloor)\rceil$ steps to return.*

Proof. The worst possible case is when $y \leq x_1$ and, in that case, it is clear that the number of calls to binsearch needed is $1 + \lceil \log_2(2\lfloor N/2 \rfloor)\rceil$. □

Algorithm 23 Binary Search; binsearch$(x_{1:N}, y)$.

1: **if** $N = 1$ **then**
2: Set $k = 0$ if $y \leq x_1$ else set $k = 1$.
3: **else**
4: Set $i = N/2$ if N is even else set $i = (N+1)/2$.
5: Set $k = i +$ binsearch$(x_{i+1:N}, y)$ if $y > x_i$ else set $k =$ binsearch$(x_{1:i}, y)$.
6: **end if**
7: Return k

DEFINITION B.0.11 (Shuffle). *For a sequence, $x_{1:N}$, a shuffle, denoted* shuffle$(x_{1:N})$*, is an algorithm (Algorithm 24) that chooses a permutation, σ, of the set $\{1, \ldots, N\}$ uniformly at random from the $N!$ different possible permutations. This thesis implements the Fisher-Yates shuffle (see, for instance, Toutenburg, 1971) as given by Durstenfeld, 1964. For any given shuffle, σ, and a shufffled sequence $x_{1:N}$, the inverse shuffle, denoted* inverse_shuffle$(x_{1:N}, \sigma)$ *is an algorithm (Algorithm 25) that inverts the shuffle; that is, $y_{1:N} =$* inverse$_{\text{s}}$huffle$(x_{1:N}, \sigma)$ *is such that, for any $i \in \{1, \ldots, N\}$, $y_i = x_{\sigma^{-1}(i)}$.*

Algorithm 24 Shuffle; shuffle$(x_{1:N})$.

1: Initialise by setting $\sigma(i) = i$ for all $i \in \{1, \ldots, N\}$ and $x^s_{1:N} = x_{1:N}$.
2: **for** $i = 1, \ldots, N-1$ **do**
3: Sample j from a Unif$(\{i, \ldots, N\})$ distribution.
4: Swap x^s_i with x^s_j.
5: Swap $\sigma(i)$ with $\sigma(j)$.
6: **end for**
7: Return $(x^s_{1:N}, \sigma)$.

Algorithm 25 Inverse Shuffle; inverse_shuffle$(x^s_{1:N}, \sigma)$.

1: **for** $i = 1, \ldots, N$ **do**
2: Set $x_{\sigma(i)} = x^s_i$.
3: **end for**
4: Return $x_{1:N}$.

LEMMA B.0.12. *Let $(a, b) \in \mathbb{R}^2$ be such that $b \geq a$. Then*

$$\lceil b \rceil - \lfloor a \rfloor < (b-a) + 2 \,, \quad \lfloor b \rfloor - \lceil a \rceil > (b-a) - 2 \,.$$

Proof. Let $u : \mathbb{R} \to [0, 1)$ and $l : \mathbb{R} \to [0, 1)$ be the functions defined, respectively, by

$$u(x) = \lceil x \rceil - x \,, \quad l(x) = x - \lfloor x \rfloor \,.$$

Then,

$$\lceil b \rceil - \lfloor a \rfloor = (b-a) + (u(b) + l(a)) < (b-a) + 2 \,,$$

thus proving the first assertion. Moreover,

$$\lfloor b \rfloor - \lceil a \rceil = (b-a) - (u(a) + l(b)) > (b-a) - 2 \,,$$

thus proving the second assertion. □

LEMMA B.0.13. *Let $X \sim \mathrm{N}(\mu, \sigma^2)$ for some $\mu \in \mathbb{R}$ and some $\sigma \in (0, \infty)$. Then*

$$\mathbb{P}(X^2 \leq x) = \Phi\left(\frac{\sqrt{x} - \mu}{\sigma}\right) + \Phi\left(\frac{\sqrt{x} + \mu}{\sigma}\right) - 1 \,.$$

Proof. A direct calculation gives

$$\begin{aligned}\mathbb{P}(X^2 \leq x) &= \mathbb{P}(-\sqrt{x} \leq X \leq \sqrt{x}) \\ &= \mathbb{P}\bigg(-\frac{\sqrt{x}+\mu}{\sigma} \leq \frac{X-\mu}{\sigma} \leq \frac{\sqrt{x}-\mu}{\sigma}\bigg) \\ &= \mathbb{P}\bigg(\frac{X-\mu}{\sigma} \leq \frac{\sqrt{x}-\mu}{\sigma}\bigg) - \mathbb{P}\bigg(\frac{X-\mu}{\sigma} \geq \frac{\sqrt{x}+\mu}{\sigma}\bigg)\,.\end{aligned}$$

□

C

EXTRA RESULTS

C.1 RAW RESULTS FOR THE CONDITIONED DIFFUSION SIMULATION STUDY IN SECTION 3.3.2

In this appendix we include, for completeness, the raw relative effective sample sizes (as defined by (71)) and the average execution times (in seconds) for each proposal and each combination of (T, y_T) for the BD, LV, and GE diffusions detailed in Section 3.3.2. Recall that, for each combination of (T, y_T), we simulated *one million* independent skeleton paths using five different proposals; the MDB of Durham and Gallant, 2002, the residual-bridge proposal of Whitaker et al., 2017 with the two choices for ξ_t, $\mathrm{RB}^{\mathrm{ODE}}$ and $\mathrm{RB}^{\mathrm{LNA}}$, and the residual-bridge proposal introduced in this paper with the same two choices for ξ_t, $\overline{\mathrm{RB}}^{\mathrm{ODE}}$ and $\overline{\mathrm{RB}}^{\mathrm{LNA}}$. For each proposal and each combination of (T, y_T) we calculated the normalised weights for each of the one million paths according to (50) and used these to calculate the relative effective sample size (Rel. ESS) defined by (71). We also noted the average execution time (wall time) in seconds over ten identical runs for each algorithm. The relative effective sample sizes and average execution times can be seen, respectively, in Figures 80 and 83 for the BD diffusion, in Figures 81 and 84 for the LV diffusion, and Figures 82 and 85 for the GE diffusion.

C.2 EXTRA RESULTS FOR THE EXCHANGEABLE SAMPLER SIMULATION STUDY IN SECTION 4.3.2

In this appendix we include more detailed figures of the behaviour of the Exchangeable Sampler for values of $N > 1$ for the examples detailed in Section 4.3.2. For the Exchangeable Sampler which targets a $\mathrm{N}(0, 1)$ distribution by using a $\mathrm{N}(0, 1/2)$ distribution as the marginal proposal, Figures 86, 87, 88, and 89 show, for $N = 10$, $N = 50$, $N = 100$, and $N = 1000$, respectively, histograms of the samples simulated by the Exchangeable Sampler at each of the one-hundred-thousand iterations for a variety of jump-sizes, $\epsilon \in \{0.01, 0.1, 0.25, 0.5, 0.75, 1.0, 1.2, \sqrt{2}\}$. Furthermore, Figures 90, 91, 92, and 93 show, for $N = 10$, $N = 50$, $N = 100$, and $N = 1000$, respectively, the states of the Exchangeable Sampler at each of the one-hundred-thousand iterations for a variety of jump-sizes, $\epsilon \in \{0.01, 0.1, 0.25, 0.5, 0.75, 1.0, 1.2, \sqrt{2}\}$.

Similarly, for the Exchangeable Sampler which targets a $\mathrm{Gamma}(5.5, 1)$ distribution by using a $\mathrm{Gamma}(0.5, 1)$ distribution as the marginal proposal, Figures 94, 95, 96, and 97 show, for $N = 10$, $N = 50$, $N = 100$, and $N = 1000$, respectively, histograms of the samples simulated by the Exchangeable Sampler at each of the one-hundred-thousand iterations for a variety of jump-sizes, $\epsilon \in \{0.01, 0.1, 0.25, 0.5, 0.75, 1.0, 1.2, \sqrt{2}\}$. Furthermore, Figures 98, 99, 100, and 101 show, for $N = 10$, $N = 50$, $N = 100$, and $N = 1000$, respectively, the states of the Exchangeable Sampler at each of the one-hundred-thousand iterations for a variety of jump-sizes, $\epsilon \in \{0.01, 0.1, 0.25, 0.5, 0.75, 1.0, 1.2, \sqrt{2}\}$.

For the Exchangeable Sampler which targets a $\mathrm{N}(0, 1)$ distribution by using a $\mathrm{T}(5)$ distribution as the marginal proposal, Figures 102, 103, 104, and 105 show, for $N = 10$, $N = 50$, $N = 100$, and $N = 1000$, respectively, histograms of the samples simulated by the Exchangeable Sampler at each of the one-hundred-thousand iterations for a variety of jump-sizes, $\epsilon \in \{0.01, 0.1, 0.25, 0.5, 0.75, 1.0, 1.2, \sqrt{2}\}$. Furthermore, Figures 106, 107, 108, and 109 show, for $N = 10$, $N = 50$, $N = 100$, and $N = 1000$, respectively, the states of the Exchangeable Sampler at each of the one-hundred-thousand iterations for a variety of jump-sizes, $\epsilon \in \{0.01, 0.1, 0.25, 0.5, 0.75, 1.0, 1.2, \sqrt{2}\}$.

For the Exchangeable Sampler which targets a conditioned Birth-Death diffusion of Section 3.1.1 by using the Modified Diffusion Bridge proposal of Section 3.2.3 as the marginal proposal distribution, Figures 110, 111, 112, and 113 show, for $N = 10$, $N = 50$, $N = 100$, and $N = 1000$, respectively, histograms of the two-hundreth element of each of the sample paths simulated by the Exchangeable Sam-

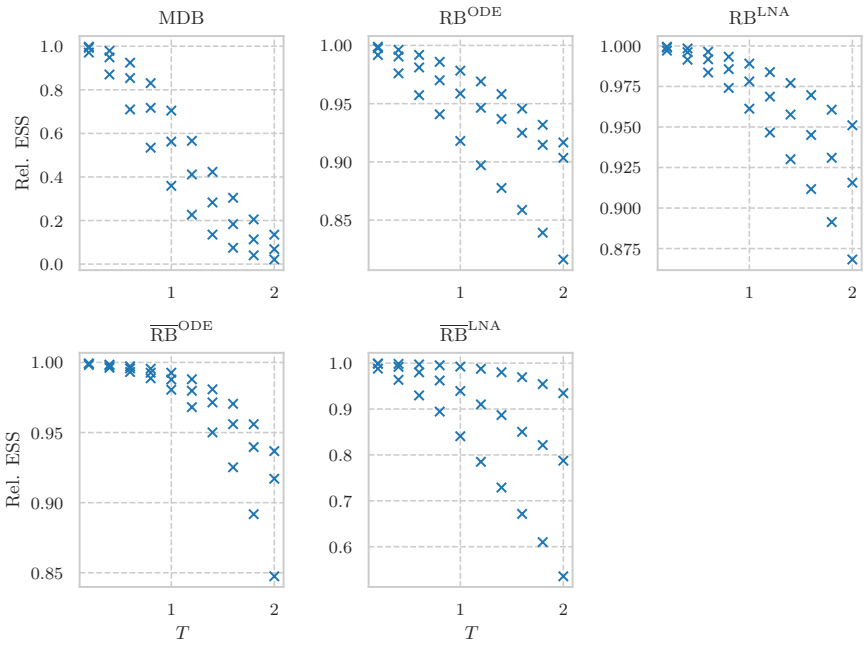

Figure 80: Plots of the relative effective sample sizes (as defined by (71)) for five proposals and for a variety of combinations of (T, y_T) corresponding to the BD diffusion.

pler at each of the one-hundred-thousand iterations for a variety of jump-sizes, $\epsilon \in \{0.01, 0.1, 0.25, 0.5, 0.75, 1.0, 1.2, \sqrt{2}\}$. Furthermore, Figures 114, 115, 116, and 117 show, for $N = 10$, $N = 50$, $N = 100$, and $N = 1000$, respectively, the two-hundreth element of the states of the Exchangeable Sampler at each of the one-hundred-thousand iterations for a variety of jump-sizes, $\epsilon \in \{0.01, 0.1, 0.25, 0.5, 0.75, 1.0, 1.2, \sqrt{2}\}$.

C.3 EXTRA RESULTS FOR THE EXCHANGEABLE PARTICLE GIBBS SAMPLER SIMULATION STUDY IN SECTION 4.4.2

In this appendix we include more detailed figures of the behaviour of the Exchangeable Particle Gibbs Sampler for the examples detailed in Section 4.4.2. For the Exchangeable Particle Gibbs Sampler applied to the Linear Gaussian model of Example 3, and using the bootstrap proposals as the marginal proposal densities, Figures 118, 119, and 120, show, for $N = 50$, $N = 100$, and $N = 1000$, respectively, histograms of the first component of the sample paths simulated by the Exchangeable Particle Gibbs Sampler at each of the one-hundred-thousand iterations for a variety of jump-sizes, $\epsilon \in \{0.01, 0.1, 0.25, 0.5, 0.75, 1.0, 1.2, \sqrt{2}\}$. Furthermore, Figures 121, 122, and 123, show, for $N = 50$, $N = 100$, and $N = 1000$, respectively, the first component of the one-hundred-thousand states of the Exchangeable Particle Gibbs Sampler for $N = 250$ and for a variety of jump-sizes,

$$\epsilon \in \{0.01, 0.1, 0.25, 0.5, 0.75, 1.0, 1.2, \sqrt{2}\} .$$

For the Exchangeable Particle Gibbs Sampler applied to the Lotka-Volterra diffusion and using the Modified Diffusion Bridge proposal of Section 3.2.3 as the marginal proposal distribution, Figures 124, 125, and 126, show, for $N = 50$, $N = 100$, and $N = 250$, respectively, histograms of the $t = 1$ element of the sample paths, simulated via the Exchangeable Particle Gibbs Sampler, for a variety of jump-sizes,

$$\epsilon \in \{0.01, 0.1, 0.15, 0.2, 0.3, 0.5, 0.75, \sqrt{2}\} .$$

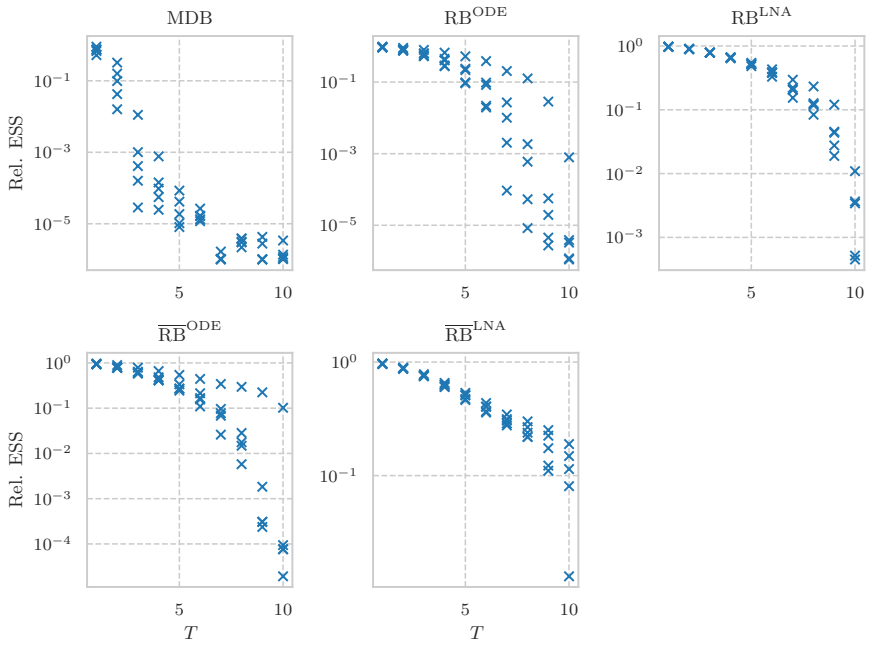

Figure 81: Plots of the relative effective sample sizes (as defined by (71)) for five proposals and for a variety of combinations of (T, y_T) corresponding to the LV diffusion.

Furthermore, Figures 127, 128, and 129, show, for $N = 50$, $N = 100$, and $N = 250$, respectively, the $t = 1$ element of the states of the Exchangeable Particle Gibbs Sampler for a variety of jump-sizes,

$$\epsilon \in \{0.01, 0.1, 0.15, 0.2, 0.3, 0.5, 0.75, \sqrt{2}\} .$$

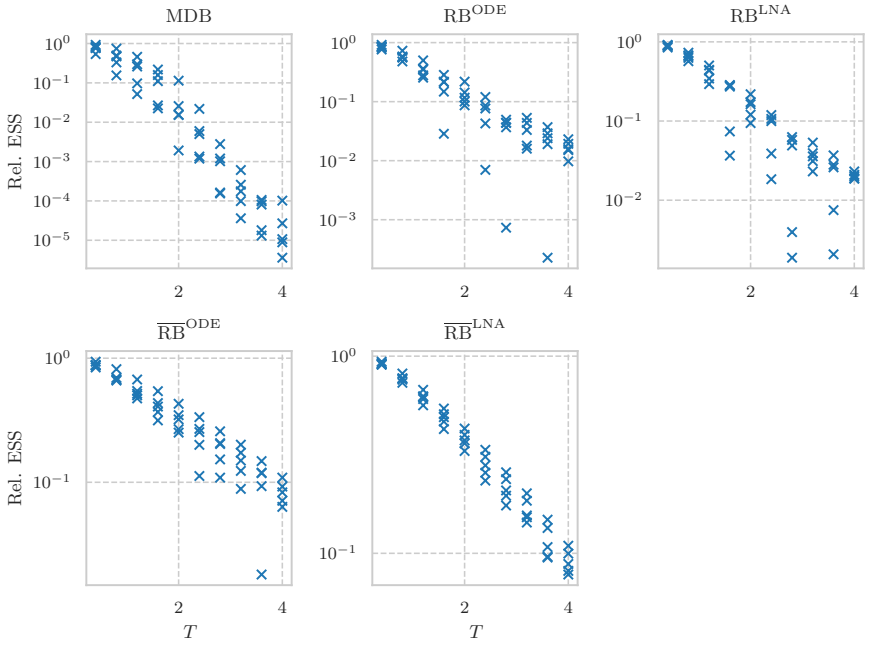

Figure 82: Plots of the relative effective sample sizes (as defined by (71)) for five proposals and for a variety of combinations of (T, y_T) corresponding to the GE diffusion.

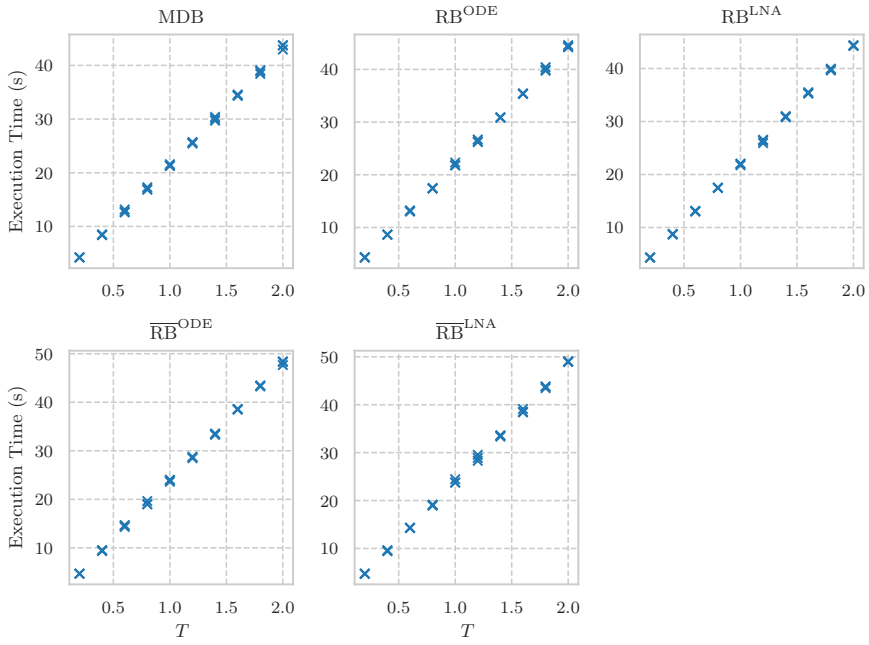

Figure 83: Plots of the average execution times for five proposals and for a variety of combinations of (T, y_T) corresponding to the BD diffusion.

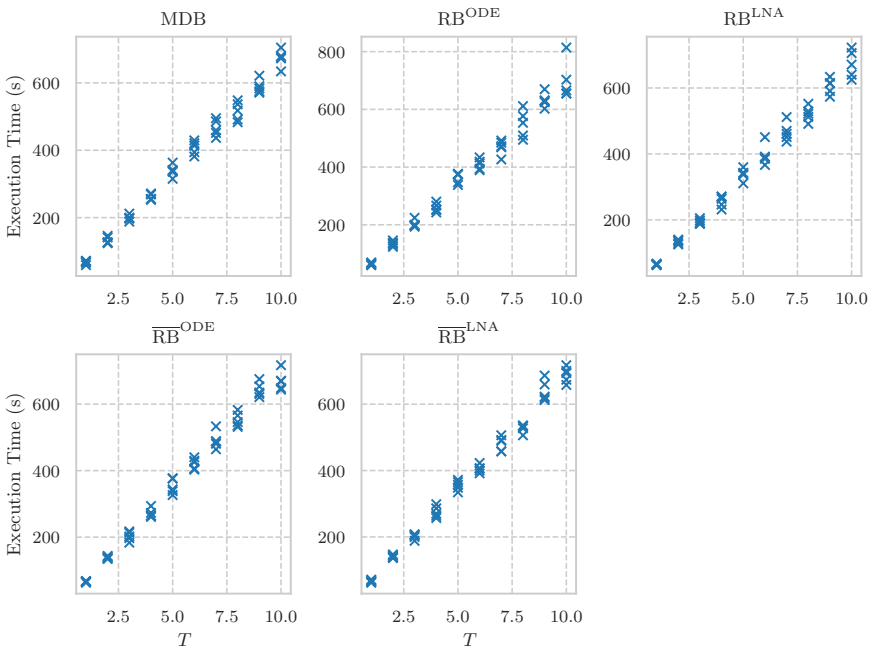

Figure 84: Plots of the average execution times for five proposals and for a variety of combinations of (T, y_T) corresponding to the LV diffusion.

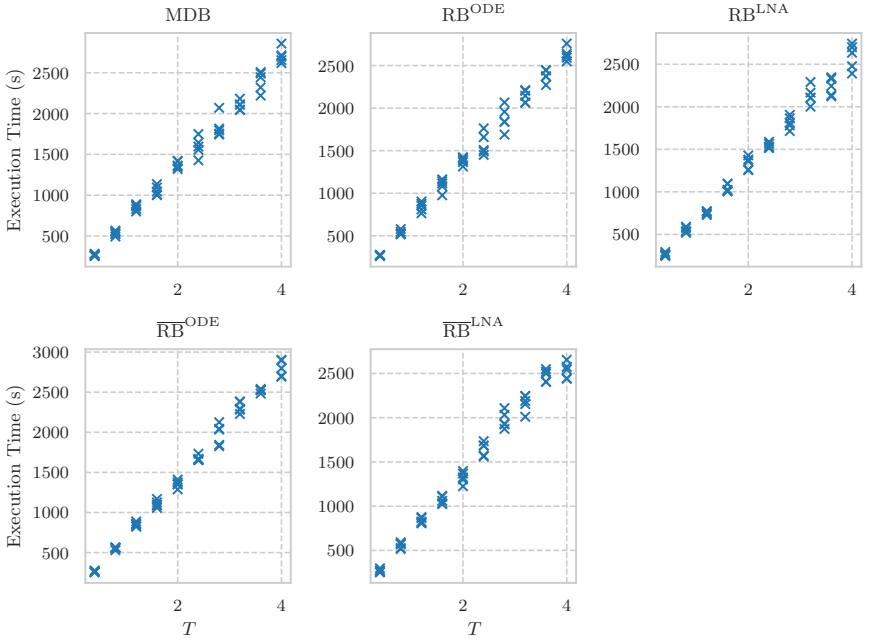

Figure 85: Plots of the average execution times for five proposals and for a variety of combinations of (T, y_T) corresponding to the GE diffusion.

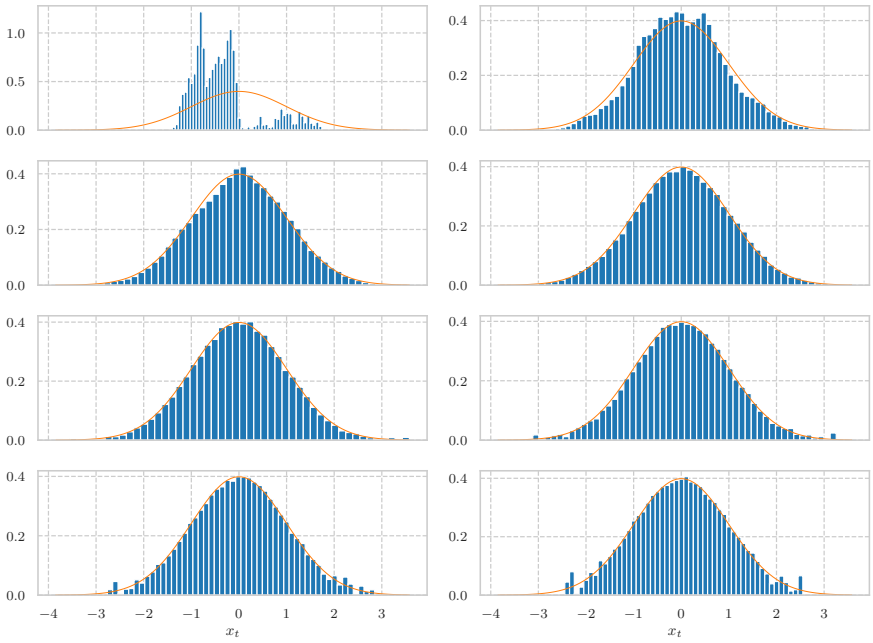

Figure 86: Histograms of the samples simulated by the Exchangeable Sampler targeting a $\mathrm{N}(0, 1)$ distribution by using a $\mathrm{N}(0, 1/2)$ distribution as the marginal proposal distribution for $N = 10$ and for a variety of jump-sizes, $\epsilon \in \{0.01, 0.1, 0.25, 0.5, 0.75, 1.0, 1.2, \sqrt{2}\}$ where $\epsilon = 0.01$ for the top-left subplot, $\epsilon = 0.1$ for the top-right subplot, $\epsilon = 0.25$ for the subplot in the second row and the first column, and so on. The orange line in each figure corresponds to the *true* density.

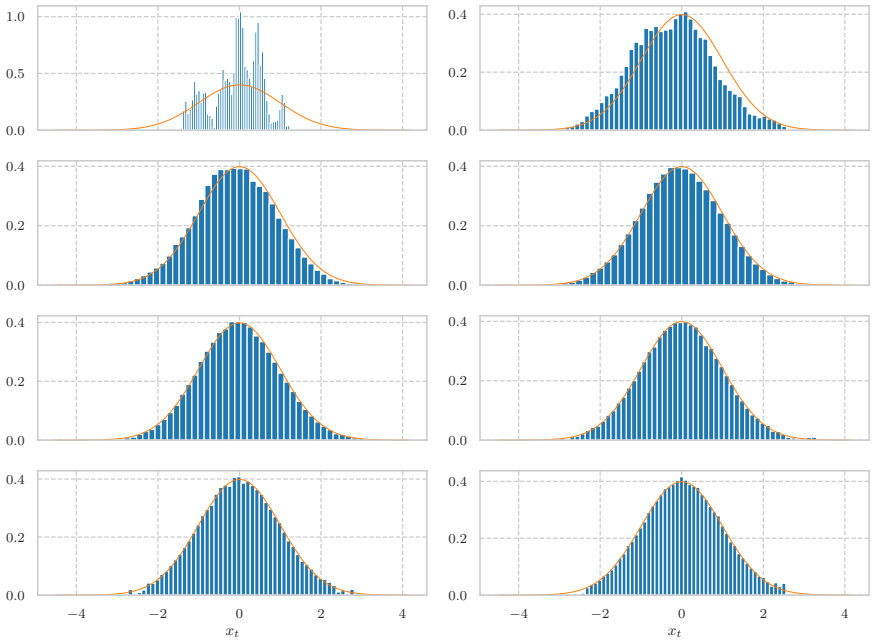

Figure 87: Histograms of the samples simulated by the Exchangeable Sampler targeting a $\mathrm{N}(0, 1)$ distribution by using a $\mathrm{N}(0, 1/2)$ distribution as the marginal proposal distribution for $N = 50$ and for a variety of jump-sizes, $\epsilon \in \{0.01, 0.1, 0.25, 0.5, 0.75, 1.0, 1.2, \sqrt{2}\}$ where $\epsilon = 0.01$ for the top-left subplot, $\epsilon = 0.1$ for the top-right subplot, $\epsilon = 0.25$ for the subplot in the second row and the first column, and so on. The orange line in each figure corresponds to the *true* density.

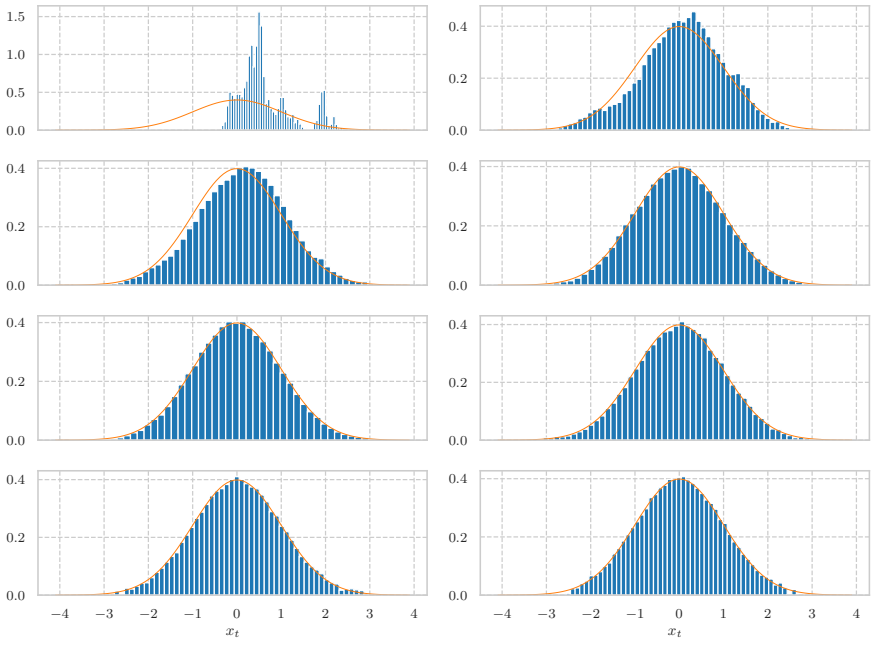

Figure 88: Histograms of the samples simulated by the Exchangeable Sampler targeting a $\mathrm{N}(0, 1)$ distribution by using a $\mathrm{N}(0, 1/2)$ distribution as the marginal proposal distribution for $N = 100$ and for a variety of jump-sizes, $\epsilon \in \{0.01, 0.1, 0.25, 0.5, 0.75, 1.0, 1.2, \sqrt{2}\}$ where $\epsilon = 0.01$ for the top-left subplot, $\epsilon = 0.1$ for the top-right subplot, $\epsilon = 0.25$ for the subplot in the second row and the first column, and so on. The orange line in each figure corresponds to the *true* density.

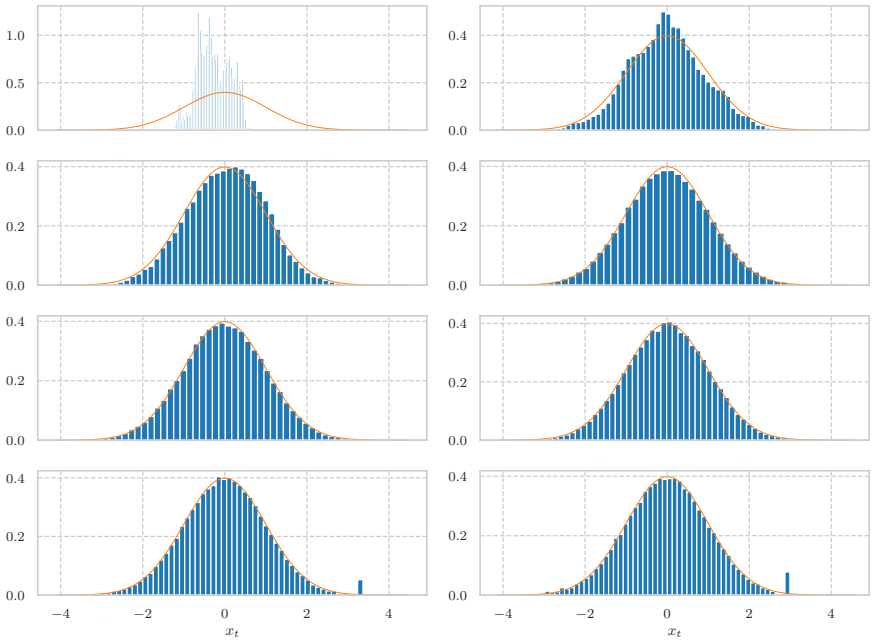

Figure 89: Histograms of the samples simulated by the Exchangeable Sampler targeting a $\mathrm{N}(0, 1)$ distribution by using a $\mathrm{N}(0, 1/2)$ distribution as the marginal proposal distribution for $N = 1000$ and for a variety of jump-sizes, $\epsilon \in \{0.01, 0.1, 0.25, 0.5, 0.75, 1.0, 1.2, \sqrt{2}\}$ where $\epsilon = 0.01$ for the top-left subplot, $\epsilon = 0.1$ for the top-right subplot, $\epsilon = 0.25$ for the subplot in the second row and the first column, and so on. The orange line in each figure corresponds to the *true* density.

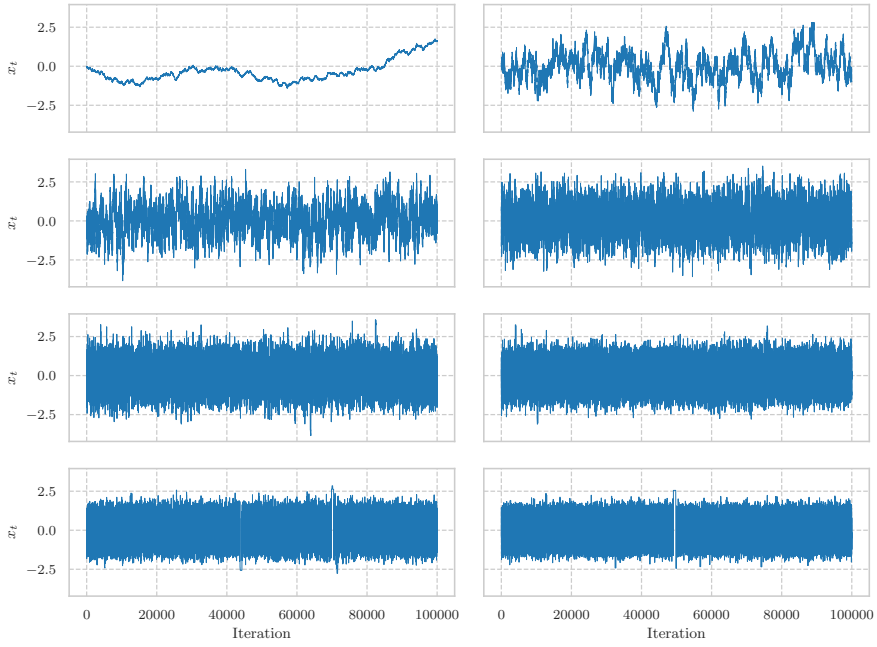

Figure 90: Plots, of the states of the Exchangeable Sampler targeting a Gamma$(0, 1)$ distribution by using a Gamma$(0, 1/2)$ distribution as the marginal proposal distribution for $N = 10$ and for a variety of jump-sizes, $\epsilon \in \{0.01, 0.1, 0.25, 0.5, 0.75, 1.0, 1.2, \sqrt{2}\}$, at each of the one-hundred-thousand iterations. $\epsilon = 0.01$ in the top-left subplot, $\epsilon = 0.1$ in the top-right subplot, $\epsilon = 0.25$ in the subplot in the second row and the first column, and so on.

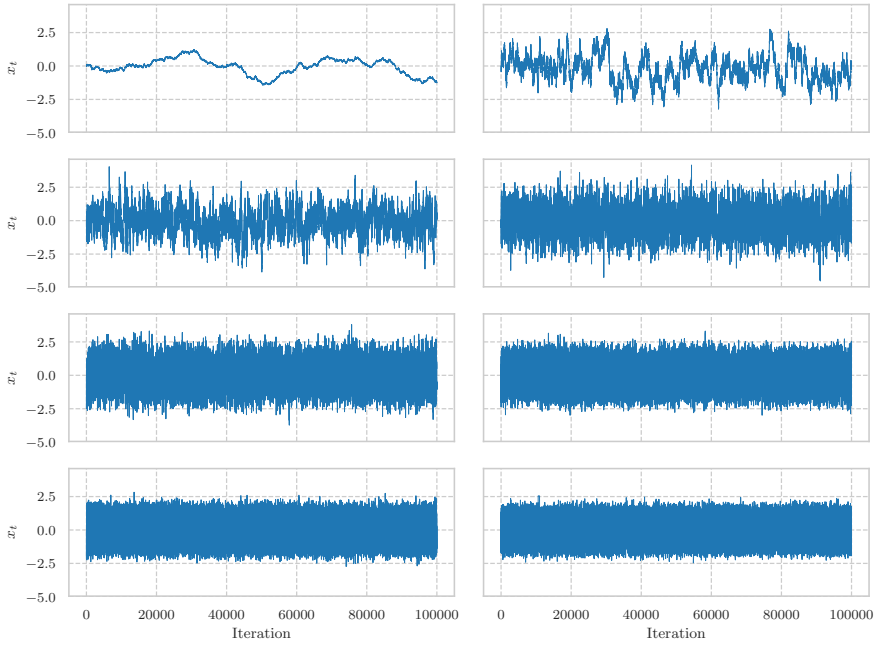

Figure 91: Plots, of the states of the Exchangeable Sampler targeting a $\mathrm{N}(0, 1)$ distribution by using a $\mathrm{N}(0, 1/2)$ distribution as the marginal proposal distribution for $N = 50$ and for a variety of jump-sizes, $\epsilon \in \{0.01, 0.1, 0.25, 0.5, 0.75, 1.0, 1.2, \sqrt{2}\}$, at each of the one-hundred-thousand iterations. $\epsilon = 0.01$ in the top-left subplot, $\epsilon = 0.1$ in the top-right subplot, $\epsilon = 0.25$ in the subplot in the second row and the first column, and so on.

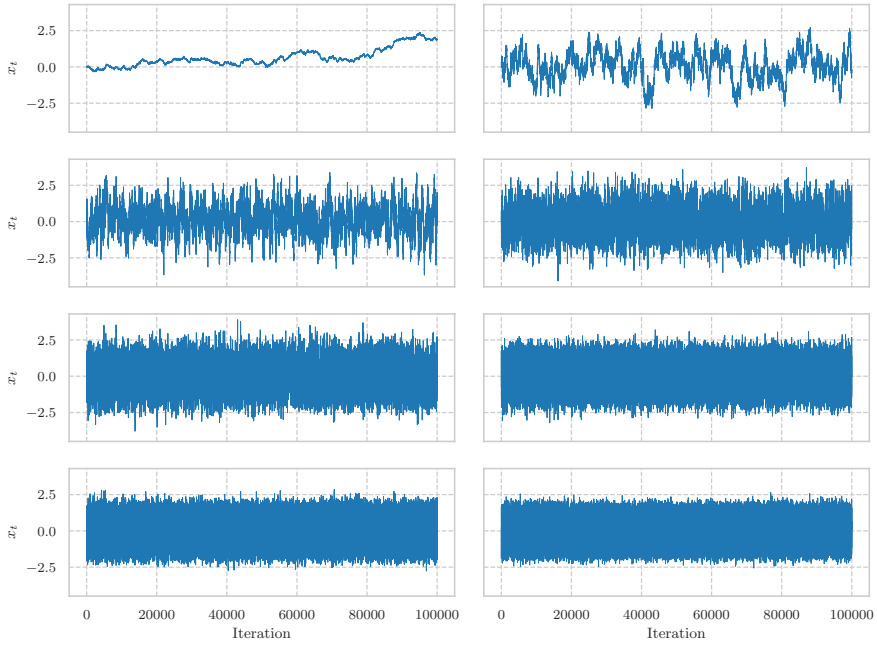

Figure 92: Plots, of the states of the Exchangeable Sampler targeting a $\mathrm{N}(0, 1)$ distribution by using a $\mathrm{N}(0, 1/2)$ distribution as the marginal proposal distribution for $N = 100$ and for a variety of jump-sizes, $\epsilon \in \{0.01, 0.1, 0.25, 0.5, 0.75, 1.0, 1.2, \sqrt{2}\}$, at each of the one-hundred-thousand iterations. $\epsilon = 0.01$ in the top-left subplot, $\epsilon = 0.1$ in the top-right subplot, $\epsilon = 0.25$ in the subplot in the second row and the first column, and so on.

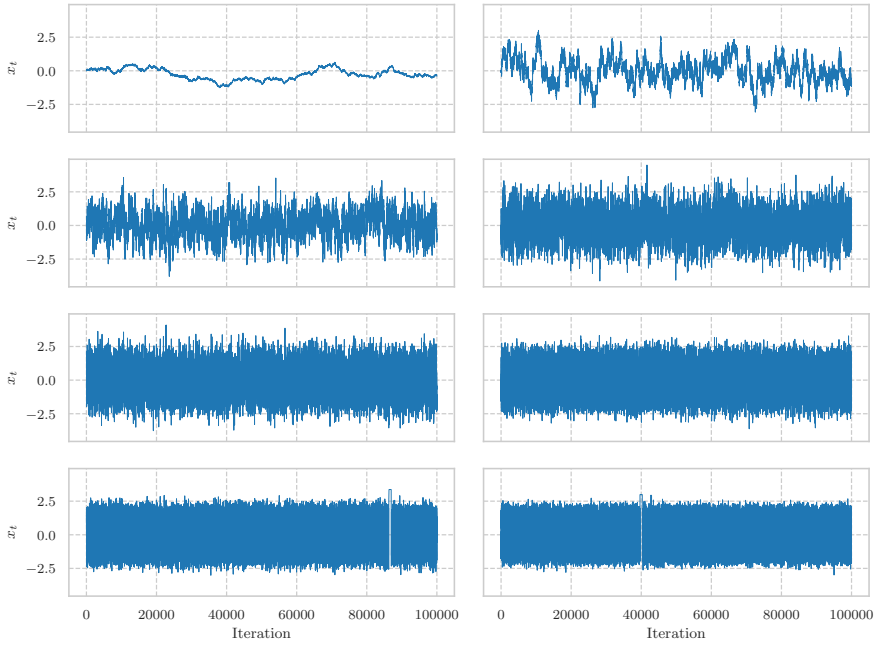

Figure 93: Plots, of the states of the Exchangeable Sampler targeting a $\mathrm{N}(0, 1)$ distribution by using a $\mathrm{N}(0, 1/2)$ distribution as the marginal proposal distribution for $N = 1000$ and for a variety of jump-sizes, $\epsilon \in \{0.01, 0.1, 0.25, 0.5, 0.75, 1.0, 1.2, \sqrt{2}\}$, at each of the one-hundred-thousand iterations. $\epsilon = 0.01$ in the top-left subplot, $\epsilon = 0.1$ in the top-right subplot, $\epsilon = 0.25$ in the subplot in the second row and the first column, and so on.

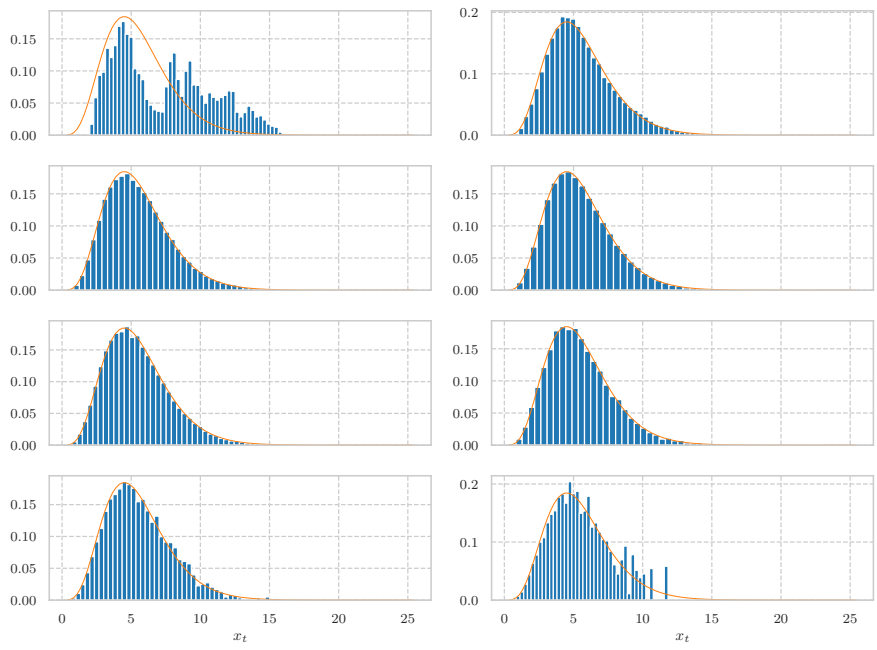

Figure 94: Histograms of the samples simulated by the Exchangeable Sampler targeting a Gamma$(5.5, 1)$ distribution by using a Gamma$(0.5, 1)$ distribution as the marginal proposal distribution for $N = 10$ and for a variety of jump-sizes, $\epsilon \in \{0.01, 0.1, 0.25, 0.5, 0.75, 1.0, 1.2, \sqrt{2}\}$ where $\epsilon = 0.01$ for the top-left subplot, $\epsilon = 0.1$ for the top-right subplot, $\epsilon = 0.25$ for the subplot in the second row and the first column, and so on. The orange line in each figure corresponds to the *true* density.

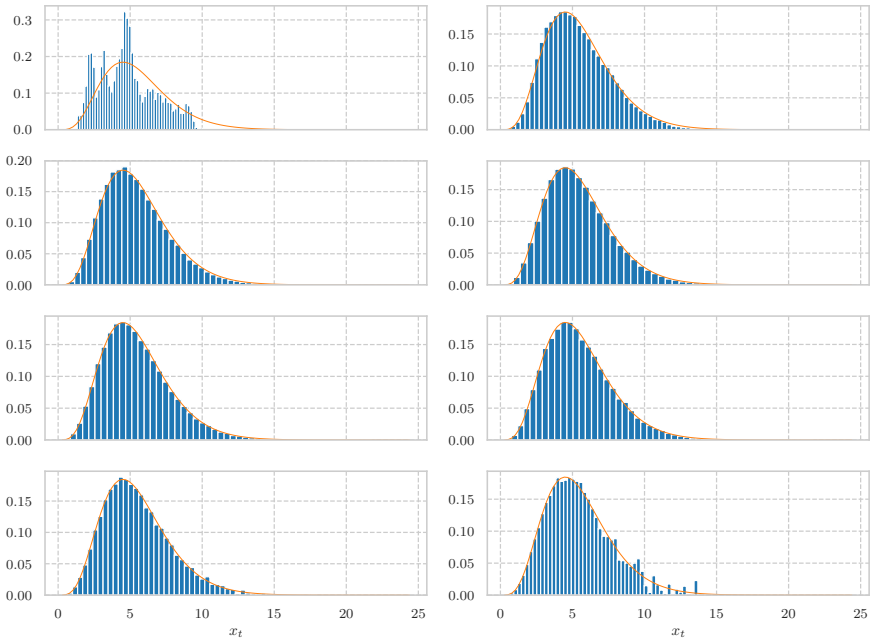

Figure 95: Histograms of the samples simulated by the Exchangeable Sampler targeting a Gamma$(5.5, 1)$ distribution by using a Gamma$(0.5, 1)$ distribution as the marginal proposal distribution for $N = 50$ and for a variety of jump-sizes, $\epsilon \in \{0.01, 0.1, 0.25, 0.5, 0.75, 1.0, 1.2, \sqrt{2}\}$ where $\epsilon = 0.01$ for the top-left subplot, $\epsilon = 0.1$ for the top-right subplot, $\epsilon = 0.25$ for the subplot in the second row and the first column, and so on. The orange line in each figure corresponds to the *true* density.

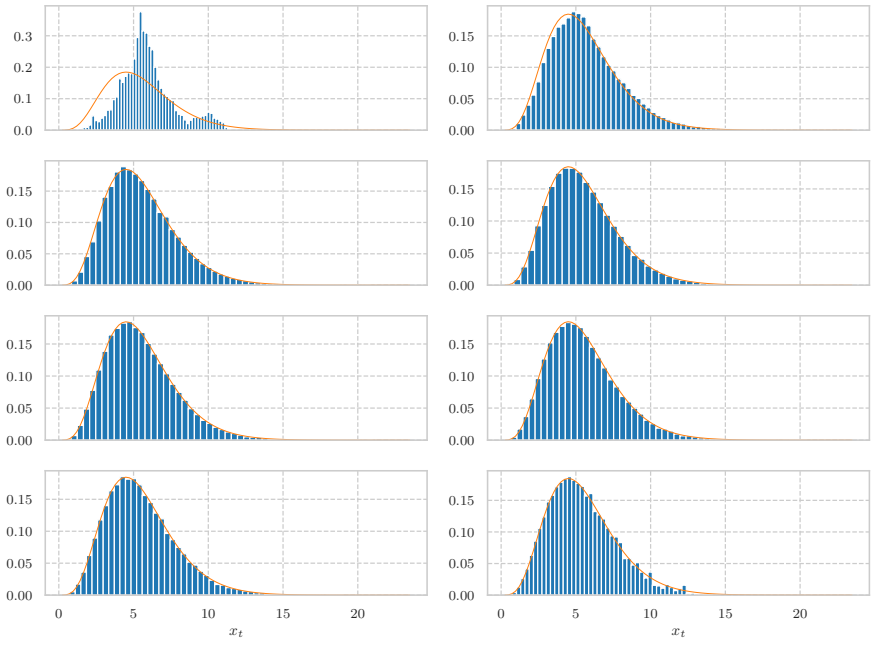

Figure 96: Histograms of the samples simulated by the Exchangeable Sampler targeting a Gamma$(5.5, 1)$ distribution by using a Gamma$(0.5, 1)$ distribution as the marginal proposal distribution for $N = 100$ and for a variety of jump-sizes, $\epsilon \in \{0.01, 0.1, 0.25, 0.5, 0.75, 1.0, 1.2, \sqrt{2}\}$ where $\epsilon = 0.01$ for the top-left subplot, $\epsilon = 0.1$ for the top-right subplot, $\epsilon = 0.25$ for the subplot in the second row and the first column, and so on. The orange line in each figure corresponds to the *true* density.

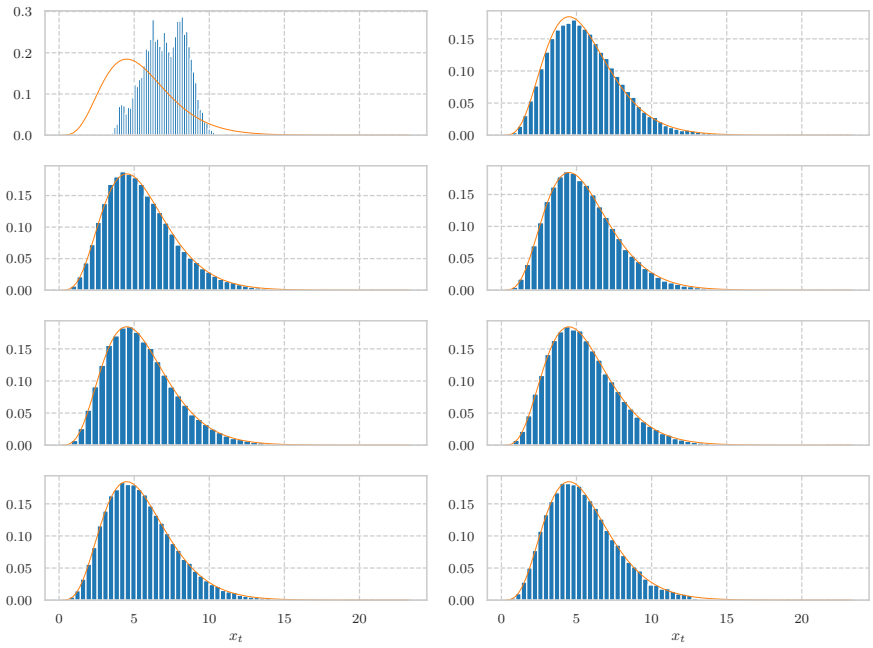

Figure 97: Histograms of the samples simulated by the Exchangeable Sampler targeting a Gamma$(5.5, 1)$ distribution by using a Gamma$(0.5, 1)$ distribution as the marginal proposal distribution for $N = 1000$ and for a variety of jump-sizes, $\epsilon \in \{0.01, 0.1, 0.25, 0.5, 0.75, 1.0, 1.2, \sqrt{2}\}$ where $\epsilon = 0.01$ for the top-left subplot, $\epsilon = 0.1$ for the top-right subplot, $\epsilon = 0.25$ for the subplot in the second row and the first column, and so on. The orange line in each figure corresponds to the *true* density.

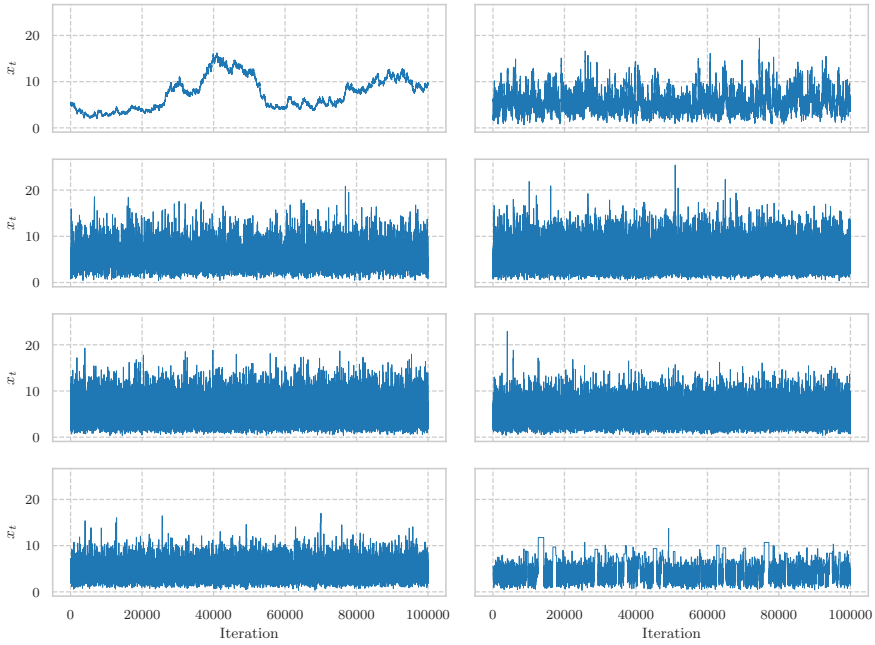

Figure 98: Plots, of the states of the Exchangeable Sampler targeting a Gamma$(5.5, 1)$ distribution by using a Gamma$(0.5, 1)$ distribution as the marginal proposal distribution for $N = 10$ and for a variety of jump-sizes, $\epsilon \in \{0.01, 0.1, 0.25, 0.5, 0.75, 1.0, 1.2, \sqrt{2}\}$, at each of the one-hundred-thousand iterations. $\epsilon = 0.01$ in the top-left subplot, $\epsilon = 0.1$ in the top-right subplot, $\epsilon = 0.25$ in the subplot in the second row and the first column, and so on.

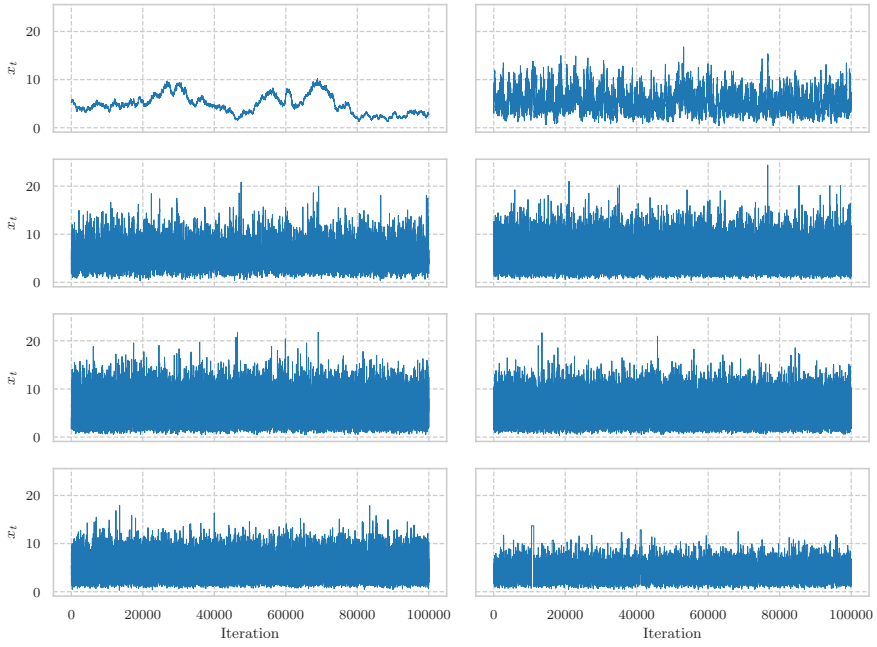

Figure 99: Plots of the states of the Exchangeable Sampler targeting a Gamma$(5.5, 1)$ distribution by using a Gamma$(0.5, 1)$ distribution as the marginal proposal distribution for $N = 50$ and for a variety of jump-sizes, $\epsilon \in \{0.01, 0.1, 0.25, 0.5, 0.75, 1.0, 1.2, \sqrt{2}\}$, at each of the one-hundred-thousand iterations. $\epsilon = 0.01$ in the top-left subplot, $\epsilon = 0.1$ in the top-right subplot, $\epsilon = 0.25$ in the subplot in the second row and the first column, and so on.

Figure 100: Plots of the states of the Exchangeable Sampler targeting a Gamma$(5.5, 1)$ distribution by using a Gamma$(0.5, 1)$ distribution as the marginal proposal distribution for $N = 100$ and for a variety of jump-sizes, $\epsilon \in \{0.01, 0.1, 0.25, 0.5, 0.75, 1.0, 1.2, \sqrt{2}\}$, at each of the one-hundred-thousand iterations. $\epsilon = 0.01$ in the top-left subplot, $\epsilon = 0.1$ in the top-right subplot, $\epsilon = 0.25$ in the subplot in the second row and the first column, and so on.

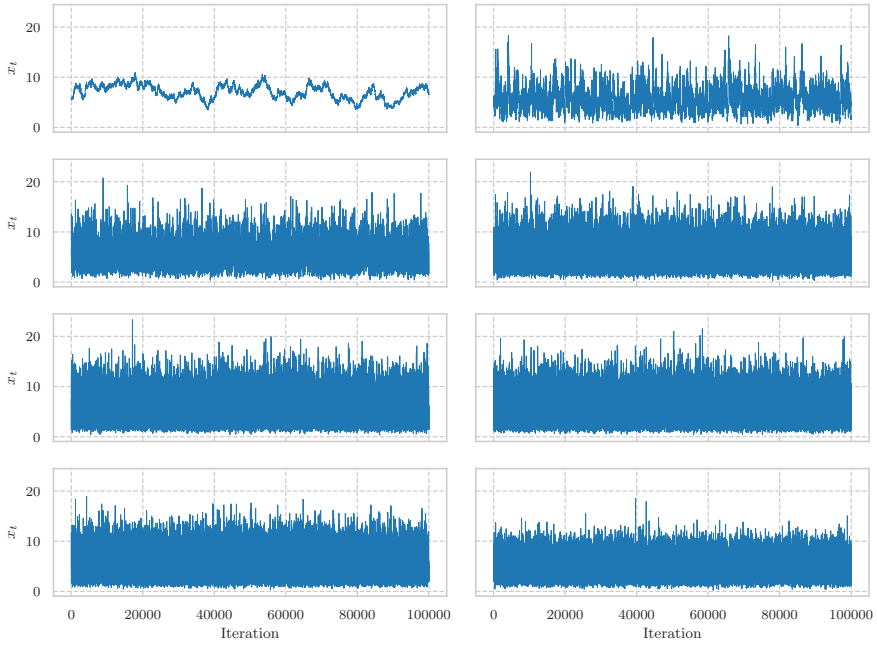

Figure 101: Plots of the states of the Exchangeable Sampler targeting a Gamma$(5.5, 1)$ distribution by using a Gamma$(0.5, 1)$ distribution as the marginal proposal distribution for $N = 1000$ and for a variety of jump-sizes, $\epsilon \in \{0.01, 0.1, 0.25, 0.5, 0.75, 1.0, 1.2, \sqrt{2}\}$, at each of the one-hundred-thousand iterations. $\epsilon = 0.01$ in the top-left subplot, $\epsilon = 0.1$ in the top-right subplot, $\epsilon = 0.25$ in the subplot in the second row and the first column, and so on.

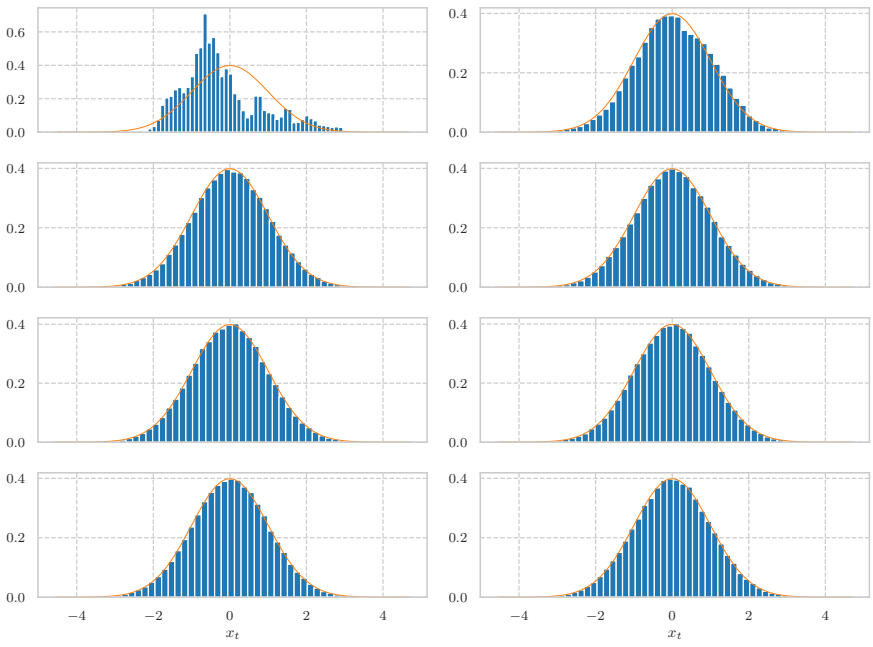

Figure 102: Histograms of the samples simulated by the Exchangeable Sampler targeting a $\mathrm{N}(0, 1)$ distribution by using a $\mathrm{T}(5)$ distribution as the marginal proposal distribution for $N = 10$ and for a variety of jump-sizes, $\epsilon \in \{0.01, 0.1, 0.25, 0.5, 0.75, 1.0, 1.2, \sqrt{2}\}$ where $\epsilon = 0.01$ for the top-left subplot, $\epsilon = 0.1$ for the top-right subplot, $\epsilon = 0.25$ for the subplot in the second row and the first column, and so on. The orange line in each figure corresponds to the *true* density.

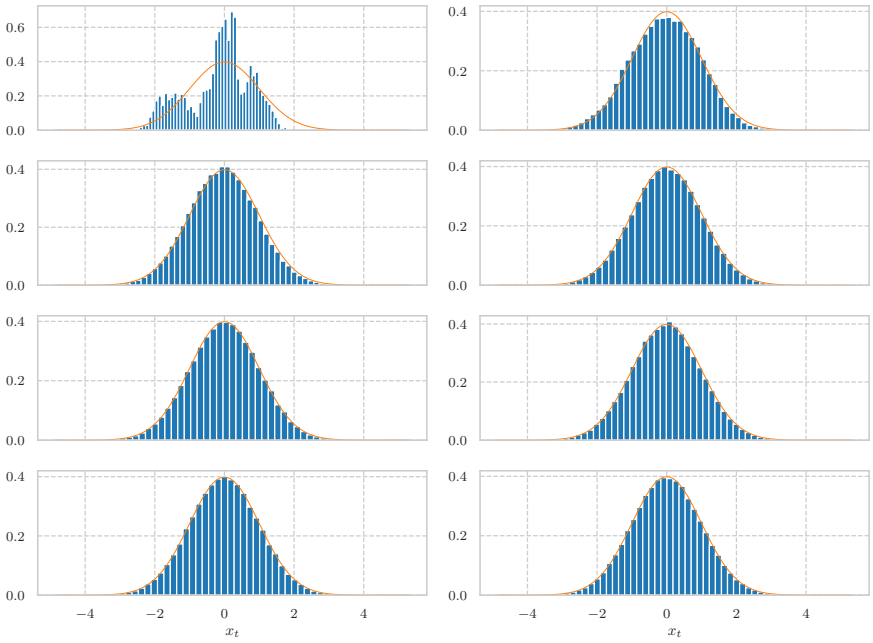

Figure 103: Histograms of the samples simulated by the Exchangeable Sampler targeting a $\mathrm{N}(0, 1)$ distribution by using a $\mathrm{T}(5)$ distribution as the marginal proposal distribution for $N = 50$ and for a variety of jump-sizes, $\epsilon \in \{0.01, 0.1, 0.25, 0.5, 0.75, 1.0, 1.2, \sqrt{2}\}$ where $\epsilon = 0.01$ for the top-left subplot, $\epsilon = 0.1$ for the top-right subplot, $\epsilon = 0.25$ for the subplot in the second row and the first column, and so on. The orange line in each figure corresponds to the *true* density.

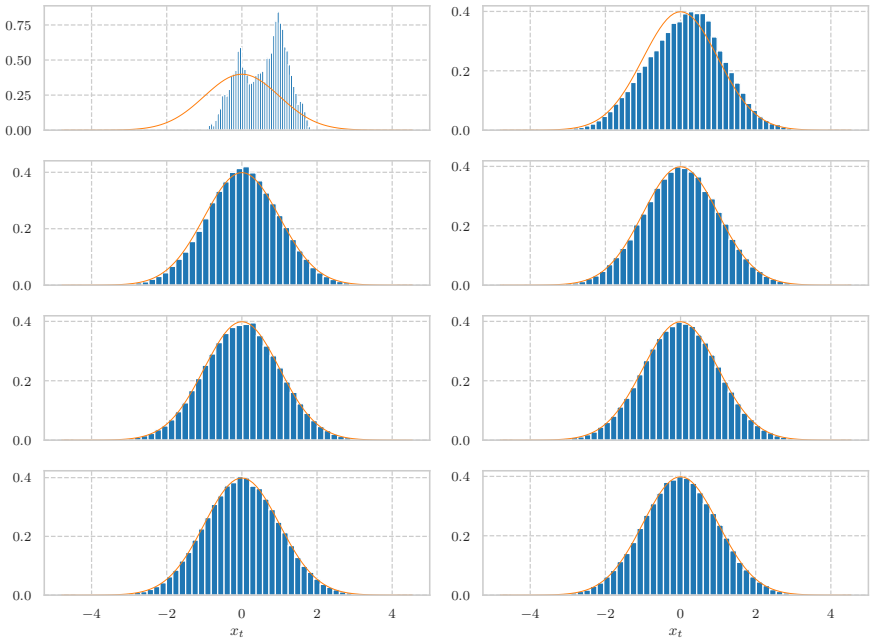

Figure 104: Histograms of the samples simulated by the Exchangeable Sampler targeting a $\mathrm{N}(0, 1)$ distribution by using a $\mathrm{T}(5)$ distribution as the marginal proposal distribution for $N = 100$ and for a variety of jump-sizes, $\epsilon \in \{0.01, 0.1, 0.25, 0.5, 0.75, 1.0, 1.2, \sqrt{2}\}$ where $\epsilon = 0.01$ for the top-left subplot, $\epsilon = 0.1$ for the top-right subplot, $\epsilon = 0.25$ for the subplot in the second row and the first column, and so on. The orange line in each figure corresponds to the *true* density.

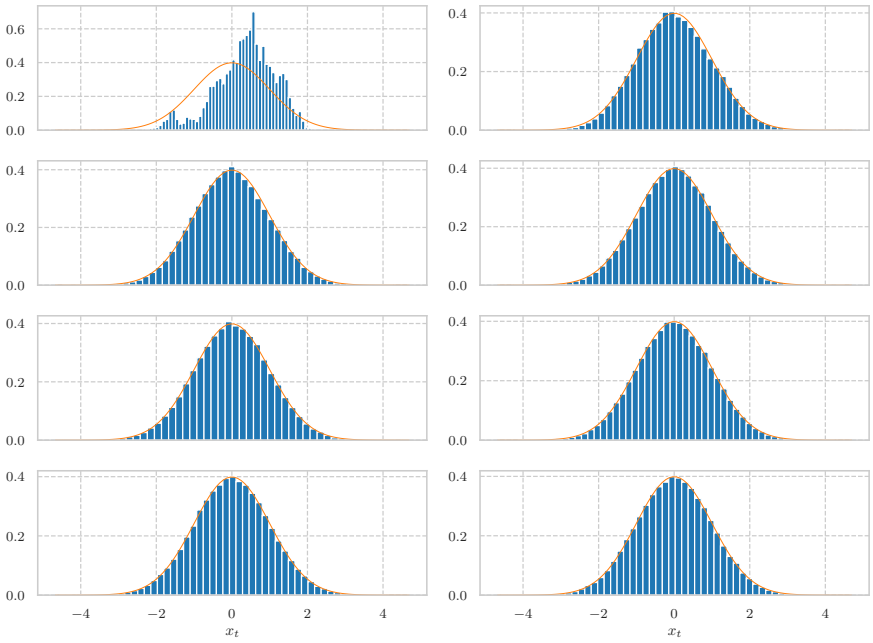

Figure 105: Histograms of the samples simulated by the Exchangeable Sampler targeting a $\mathrm{N}(0, 1)$ distribution by using a $\mathrm{T}(5)$ distribution as the marginal proposal distribution for $N = 1000$ and for a variety of jump-sizes, $\epsilon \in \{0.01, 0.1, 0.25, 0.5, 0.75, 1.0, 1.2, \sqrt{2}\}$ where $\epsilon = 0.01$ for the top-left subplot, $\epsilon = 0.1$ for the top-right subplot, $\epsilon = 0.25$ for the subplot in the second row and the first column, and so on. The orange line in each figure corresponds to the *true* density.

Figure 106: Plots, of the states of the Exchangeable Sampler targeting a $\mathrm{N}(0, 1)$ distribution by using a $\mathrm{T}(5)$ distribution as the marginal proposal distribution for $N = 10$ and for a variety of jump-sizes, $\epsilon \in \{0.01, 0.1, 0.25, 0.5, 0.75, 1.0, 1.2, \sqrt{2}\}$, at each of the one-hundred-thousand iterations. $\epsilon = 0.01$ in the top-left subplot, $\epsilon = 0.1$ in the top-right subplot, $\epsilon = 0.25$ in the subplot in the second row and the first column, and so on.

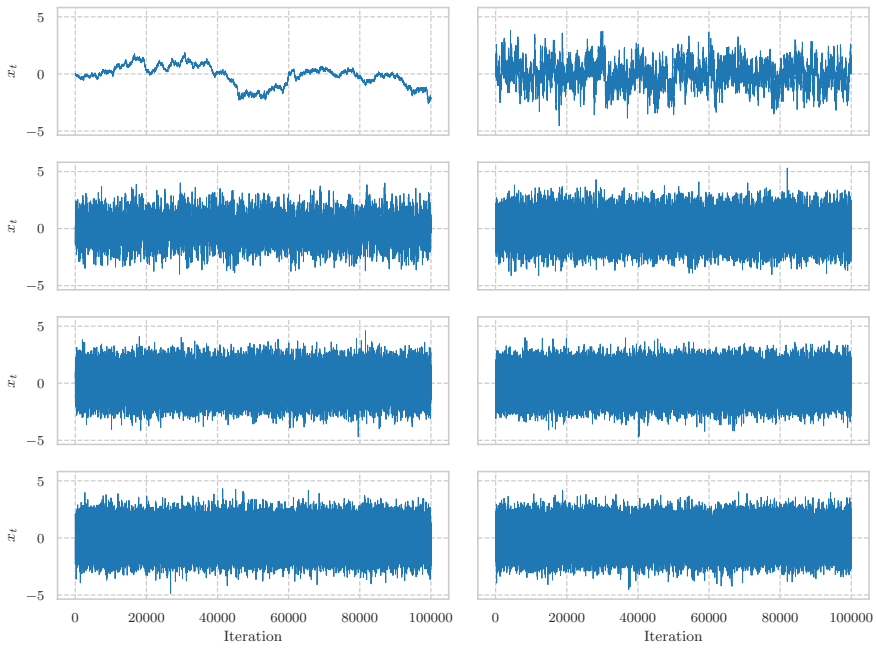

Figure 107: Plots of the states of the Exchangeable Sampler targeting a $\mathrm{N}(0, 1)$ distribution by using a $\mathrm{T}(5)$ distribution as the marginal proposal distribution for $N = 50$ and for a variety of jump-sizes, $\epsilon \in \{0.01, 0.1, 0.25, 0.5, 0.75, 1.0, 1.2, \sqrt{2}\}$, at each of the one-hundred-thousand iterations. $\epsilon = 0.01$ in the top-left subplot, $\epsilon = 0.1$ in the top-right subplot, $\epsilon = 0.25$ in the subplot in the second row and the first column, and so on.

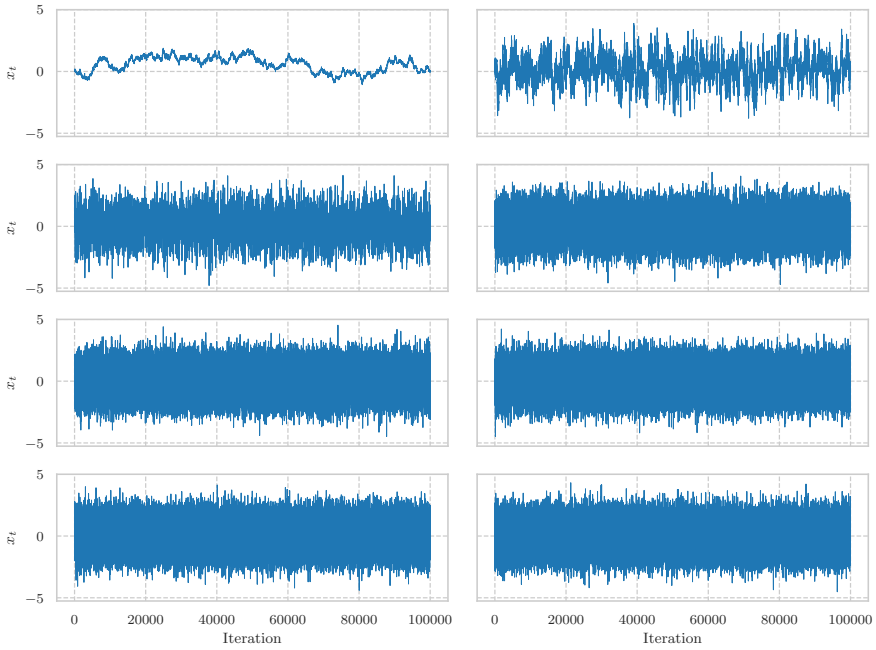

Figure 108: Plots of the states of the Exchangeable Sampler targeting a $\mathrm{N}(0, 1)$ distribution by using a $\mathrm{T}(5)$ distribution as the marginal proposal distribution for $N = 100$ and for a variety of jump-sizes, $\epsilon \in \{0.01, 0.1, 0.25, 0.5, 0.75, 1.0, 1.2, \sqrt{2}\}$, at each of the one-hundred-thousand iterations. $\epsilon = 0.01$ in the top-left subplot, $\epsilon = 0.1$ in the top-right subplot, $\epsilon = 0.25$ in the subplot in the second row and the first column, and so on.

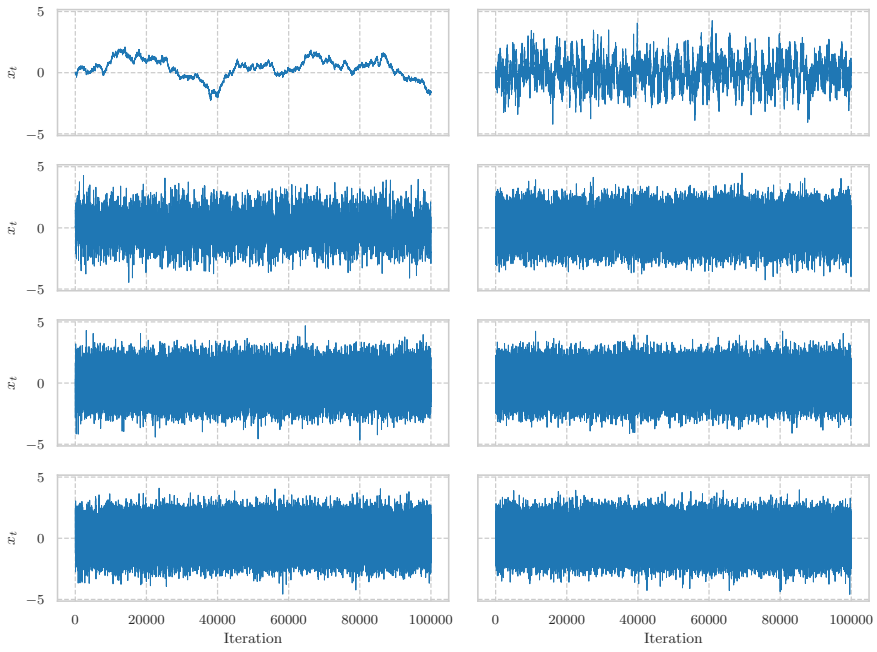

Figure 109: Plots of the states of the Exchangeable Sampler targeting a $\mathrm{N}(0, 1)$ distribution by using a $\mathrm{T}(5)$ distribution as the marginal proposal distribution for $N = 1000$ and for a variety of jump-sizes, $\epsilon \in \{0.01, 0.1, 0.25, 0.5, 0.75, 1.0, 1.2, \sqrt{2}\}$, at each of the one-hundred-thousand iterations. $\epsilon = 0.01$ in the top-left subplot, $\epsilon = 0.1$ in the top-right subplot, $\epsilon = 0.25$ in the subplot in the second row and the first column, and so on.

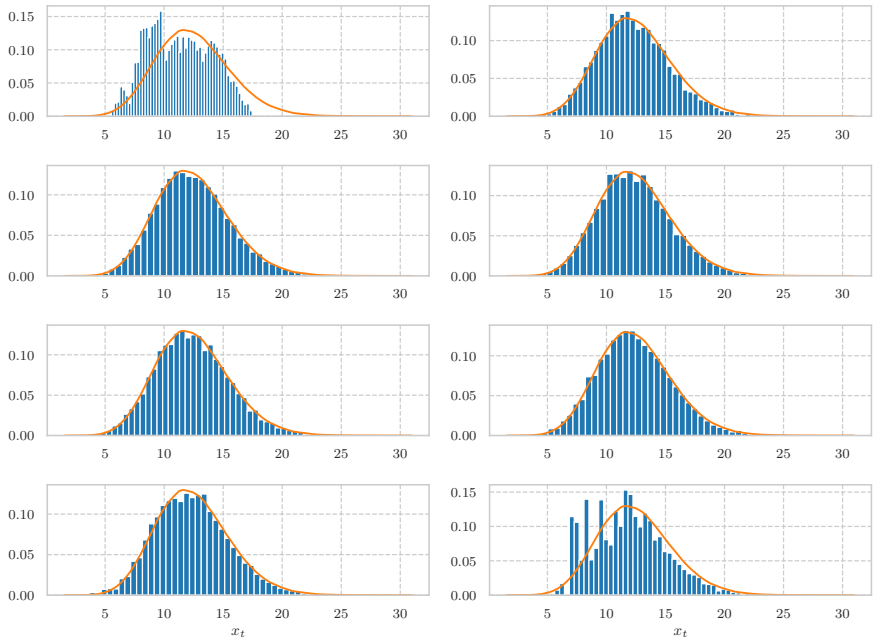

Figure 110: Histograms of the two-hundreth element of each of the sample paths simulated by the Exchangeable Sampler targeting a conditioned Birth-Death diffusion by using the Modified Diffusion Bridge proposal of Section 3.2.3 as the marginal proposal distribution. for $N = 10$ and for a variety of jump-sizes, $\epsilon \in \{0.01, 0.1, 0.25, 0.5, 0.75, 1.0, 1.2, \sqrt{2}\}$ where $\epsilon = 0.01$ for the top-left subplot, $\epsilon = 0.1$ for the top-right subplot, $\epsilon = 0.25$ for the subplot in the second row and the first column, and so on. The orange line in each figure corresponds to the *true* density.

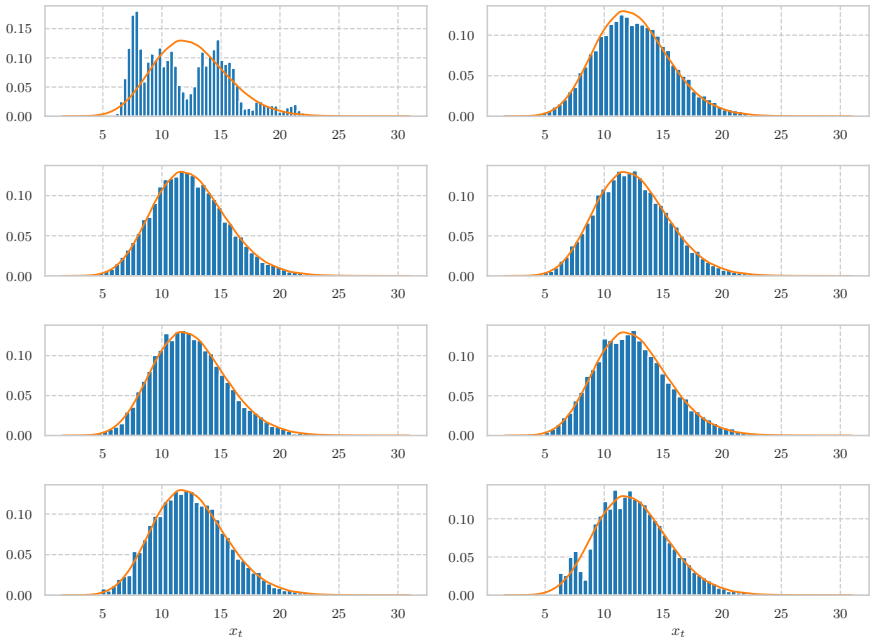

Figure 111: Histograms of the two-hundreth element of each of the sample paths simulated by the Exchangeable Sampler targeting a conditioned Birth-Death diffusion by using the Modified Diffusion Bridge proposal of Section 3.2.3 as the marginal proposal distribution. for $N = 50$ and for a variety of jump-sizes, $\epsilon \in \{0.01, 0.1, 0.25, 0.5, 0.75, 1.0, 1.2, \sqrt{2}\}$ where $\epsilon = 0.01$ for the top-left subplot, $\epsilon = 0.1$ for the top-right subplot, $\epsilon = 0.25$ for the subplot in the second row and the first column, and so on. The orange line in each figure corresponds to the *true* density.

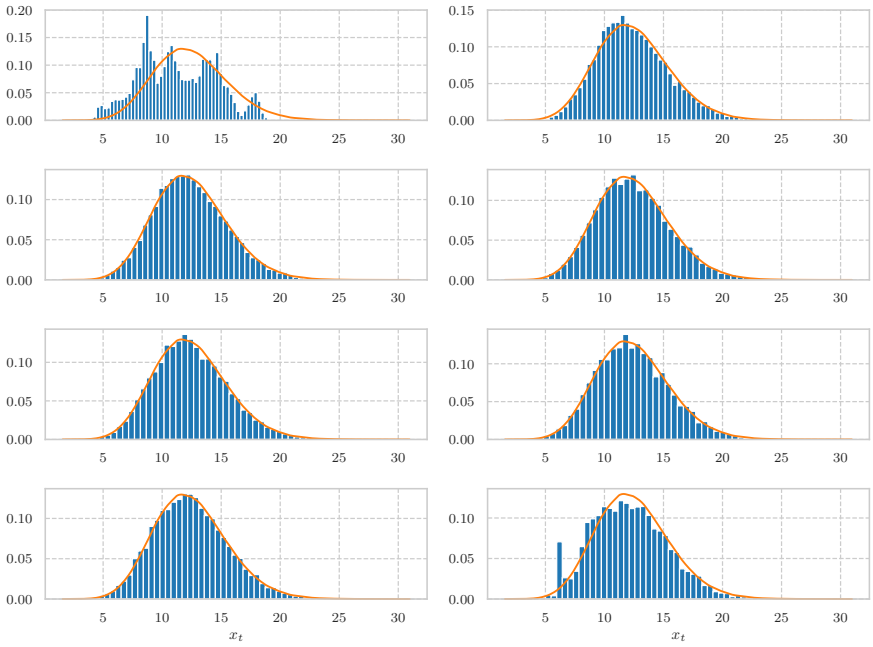

Figure 112: Histograms of the two-hundreth element of each of the sample paths simulated by the Exchangeable Sampler targeting a conditioned Birth-Death diffusion by using the Modified Diffusion Bridge proposal of Section 3.2.3 as the marginal proposal distribution. for $N = 100$ and for a variety of jump-sizes, $\epsilon \in \{0.01, 0.1, 0.25, 0.5, 0.75, 1.0, 1.2, \sqrt{2}\}$ where $\epsilon = 0.01$ for the top-left subplot, $\epsilon = 0.1$ for the top-right subplot, $\epsilon = 0.25$ for the subplot in the second row and the first column, and so on. The orange line in each figure corresponds to the *true* density.

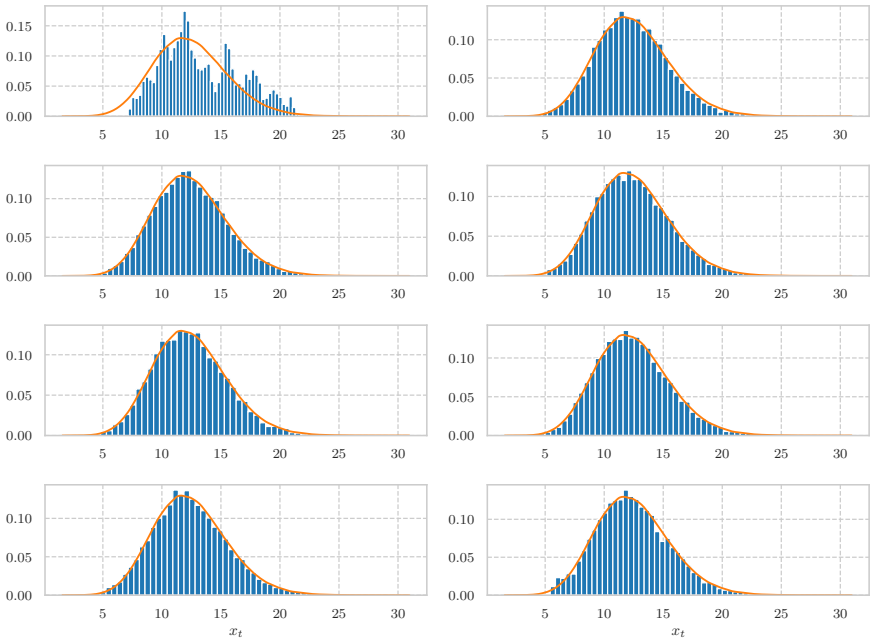

Figure 113: Histograms of the two-hundreth element of each of the sample paths simulated by the Exchangeable Sampler targeting a conditioned Birth-Death diffusion by using the Modified Diffusion Bridge proposal of Section 3.2.3 as the marginal proposal distribution. for $N = 1000$ and for a variety of jump-sizes, $\epsilon \in \{0.01, 0.1, 0.25, 0.5, 0.75, 1.0, 1.2, \sqrt{2}\}$ where $\epsilon = 0.01$ for the top-left subplot, $\epsilon = 0.1$ for the top-right subplot, $\epsilon = 0.25$ for the subplot in the second row and the first column, and so on. The orange line in each figure corresponds to the *true* density.

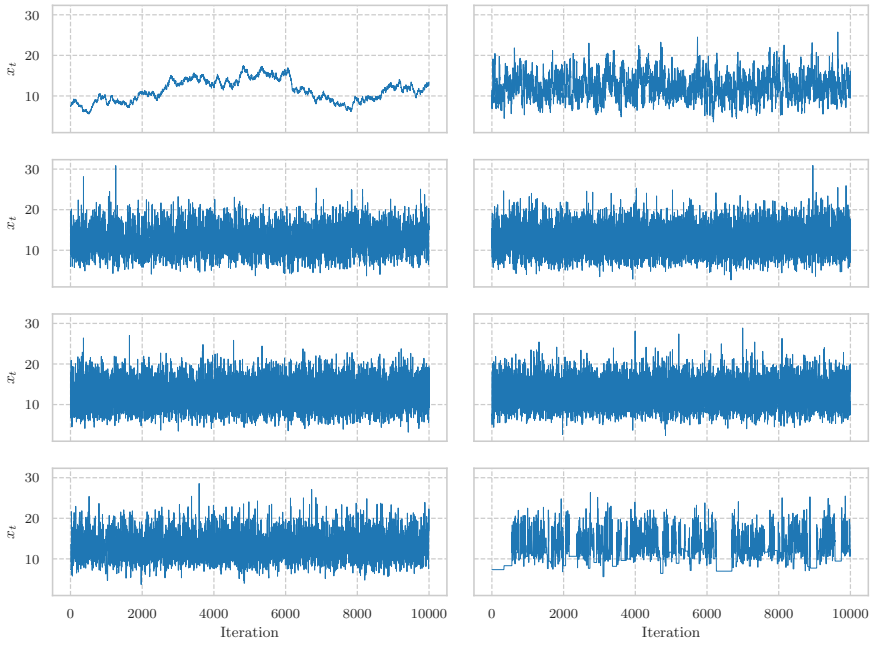

Figure 114: Plots, of the two-hundreth element of the states of the Exchangeable Sampler targeting a conditioned Birth-Death diffusion by using the Modified Diffusion Bridge proposal of Section 3.2.3 as the marginal proposal distribution for $N = 10$ and for a variety of jump-sizes, $\epsilon \in \{0.01, 0.1, 0.25, 0.5, 0.75, 1.0, 1.2, \sqrt{2}\}$, at each of the one-hundred-thousand iterations. $\epsilon = 0.01$ in the top-left subplot, $\epsilon = 0.1$ in the top-right subplot, $\epsilon = 0.25$ in the subplot in the second row and the first column, and so on.

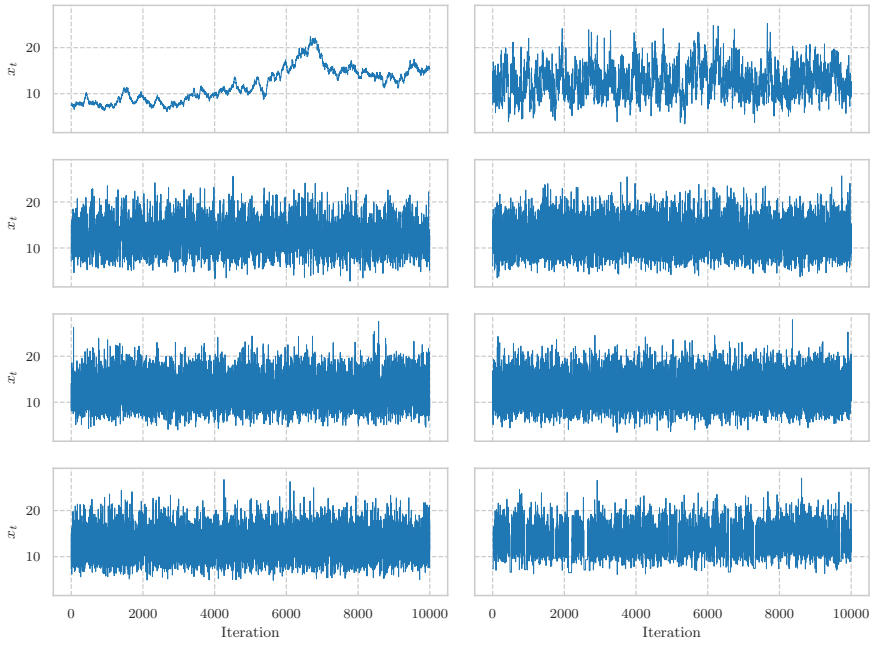

Figure 115: Plots, of the two-hundreth element of the states of the Exchangeable Sampler targeting a conditioned Birth-Death diffusion by using the Modified Diffusion Bridge proposal of Section 3.2.3 as the marginal proposal distribution for $N = 50$ and for a variety of jump-sizes, $\epsilon \in \{0.01, 0.1, 0.25, 0.5, 0.75, 1.0, 1.2, \sqrt{2}\}$, at each of the one-hundred-thousand iterations. $\epsilon = 0.01$ in the top-left subplot, $\epsilon = 0.1$ in the top-right subplot, $\epsilon = 0.25$ in the subplot in the second row and the first column, and so on.

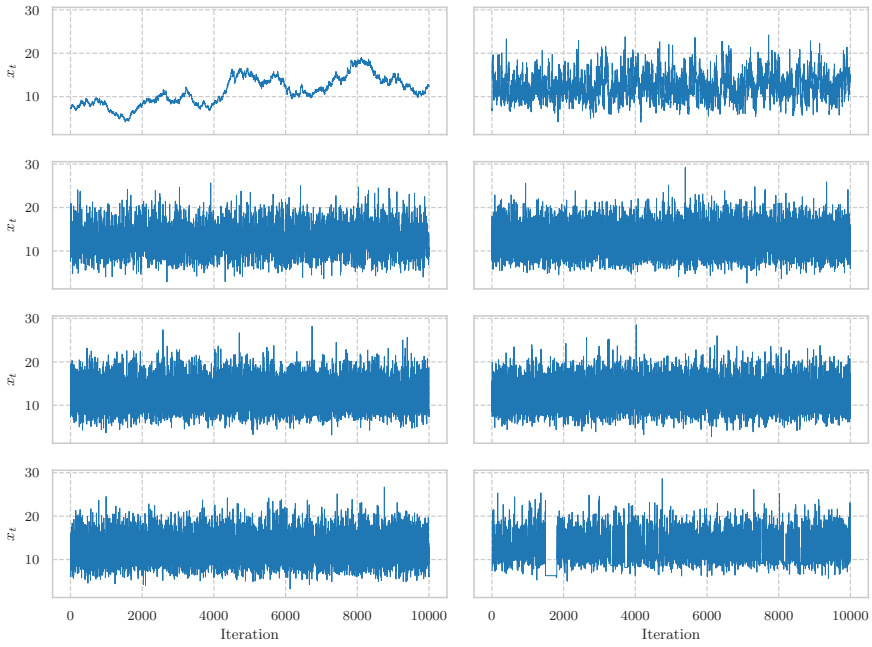

Figure 116: Plots, of the two-hundreth element of the states of the Exchangeable Sampler targeting a conditioned Birth-Death diffusion by using the Modified Diffusion Bridge proposal of Section 3.2.3 as the marginal proposal distribution for $N = 100$ and for a variety of jump-sizes, $\epsilon \in \{0.01, 0.1, 0.25, 0.5, 0.75, 1.0, 1.2, \sqrt{2}\}$, at each of the one-hundred-thousand iterations. $\epsilon = 0.01$ in the top-left subplot, $\epsilon = 0.1$ in the top-right subplot, $\epsilon = 0.25$ in the subplot in the second row and the first column, and so on.

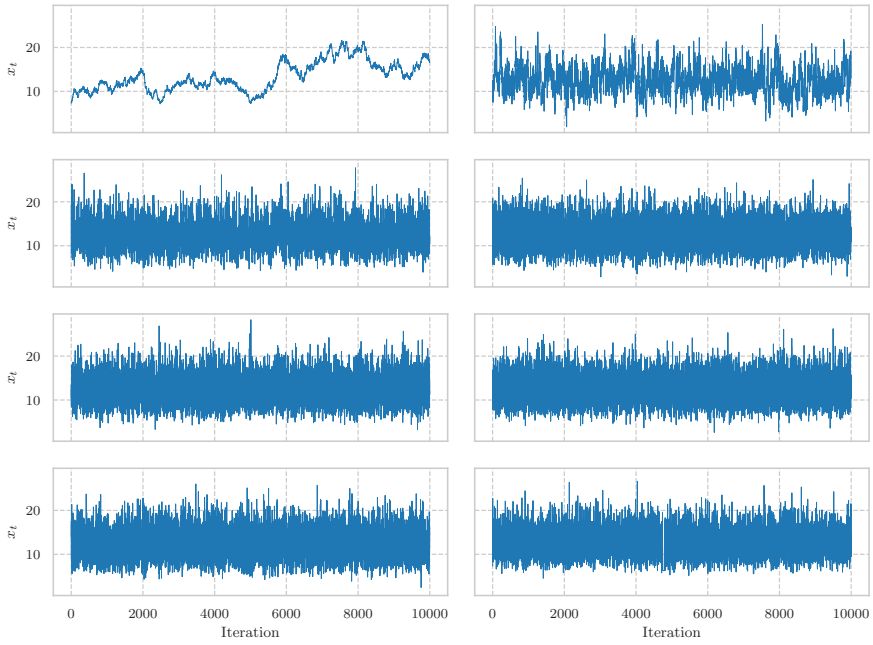

Figure 117: Plots, of the two-hundreth element of the states of the Exchangeable Sampler targeting a conditioned Birth-Death diffusion by using the Modified Diffusion Bridge proposal of Section 3.2.3 as the marginal proposal distribution for $N = 1000$ and for a variety of jump-sizes, $\epsilon \in \{0.01, 0.1, 0.25, 0.5, 0.75, 1.0, 1.2, \sqrt{2}\}$, at each of the one-hundred-thousand iterations. $\epsilon = 0.01$ in the top-left subplot, $\epsilon = 0.1$ in the top-right subplot, $\epsilon = 0.25$ in the subplot in the second row and the first column, and so on.

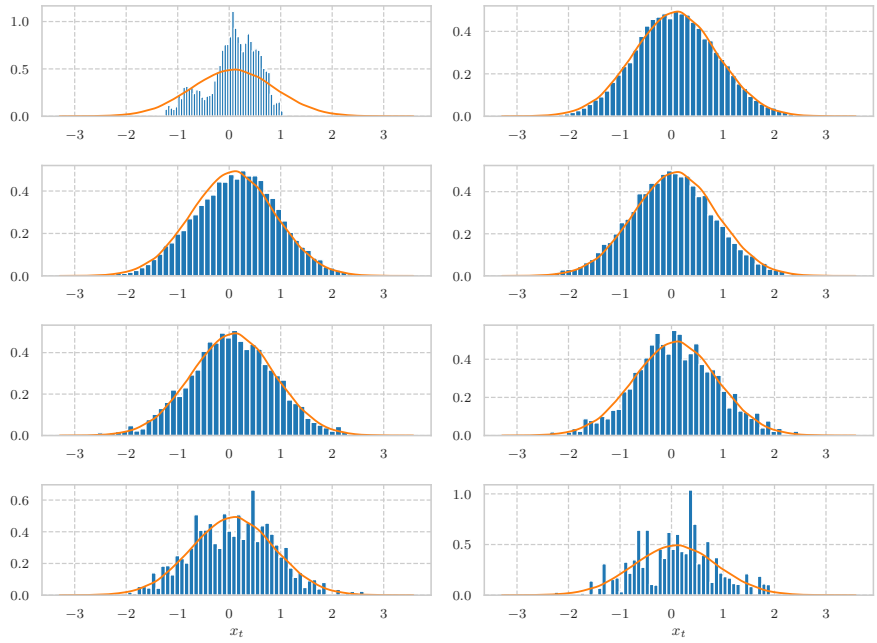

Figure 118: Histograms of the first component of the sample paths simulated by the Exchangeable Particle Gibbs Sampler- applied to the Linear Gaussian model and using the bootstrap proposals as the marginal proposal densities- with $N = 50$, and for a variety of jump-sizes, $\epsilon \in \{0.01, 0.1, 0.25, 0.5, 0.75, 1.0, 1.2, \sqrt{2}\}$, where $\epsilon = 0.01$ for the top-left subplot, $\epsilon = 0.1$ for the top-right subplot, $\epsilon = 0.25$ for the subplot in the second row and the first column, and so on. The orange line in each figure corresponds to the *true* density.

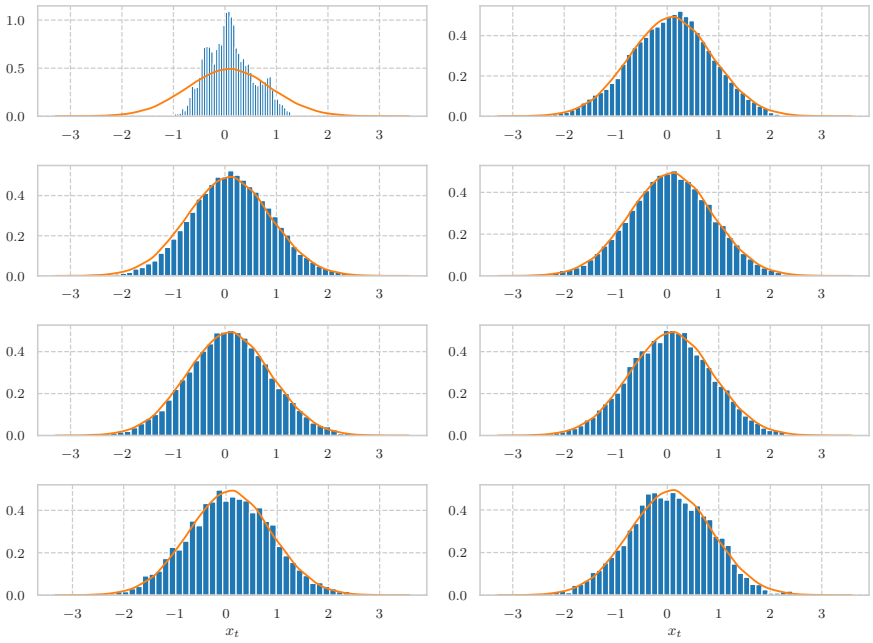

Figure 119: Histograms of the first component of the sample paths simulated by the Exchangeable Particle Gibbs Sampler- applied to the Linear Gaussian model and using the bootstrap proposals as the marginal proposal densities- with $N = 100$, and for a variety of jump-sizes, $\epsilon \in \{0.01, 0.1, 0.25, 0.5, 0.75, 1.0, 1.2, \sqrt{2}\}$, where $\epsilon = 0.01$ for the top-left subplot, $\epsilon = 0.1$ for the top-right subplot, $\epsilon = 0.25$ for the subplot in the second row and the first column, and so on. The orange line in each figure corresponds to the *true* density.

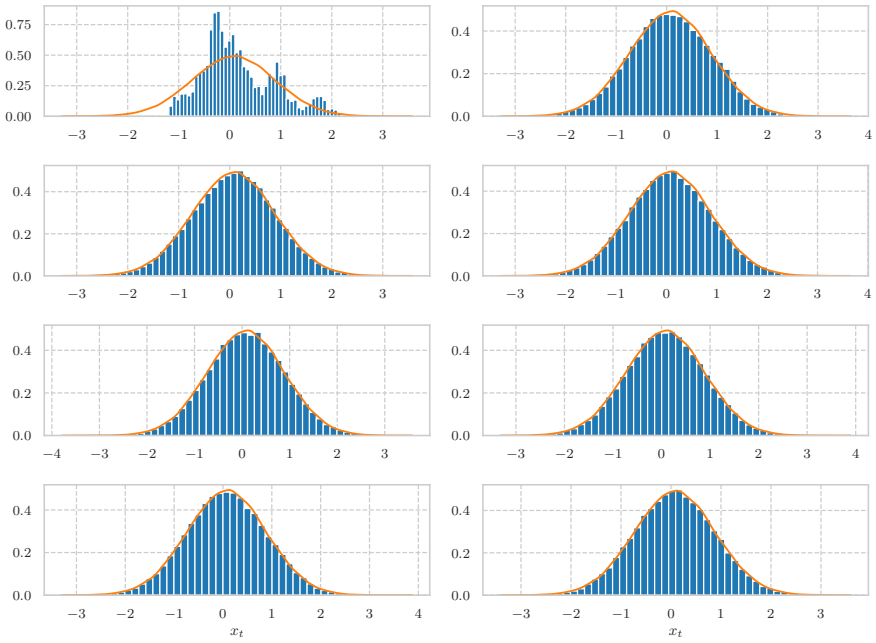

Figure 120: Histograms of the first component of the sample paths simulated by the Exchangeable Particle Gibbs Sampler- applied to the Linear Gaussian model and using the bootstrap proposals as the marginal proposal densities- with $N = 1000$, and for a variety of jump-sizes, $\epsilon \in \{0.01, 0.1, 0.25, 0.5, 0.75, 1.0, 1.2, \sqrt{2}\}$, where $\epsilon = 0.01$ for the top-left subplot, $\epsilon = 0.1$ for the top-right subplot, $\epsilon = 0.25$ for the subplot in the second row and the first column, and so on. The orange line in each figure corresponds to the *true* density.

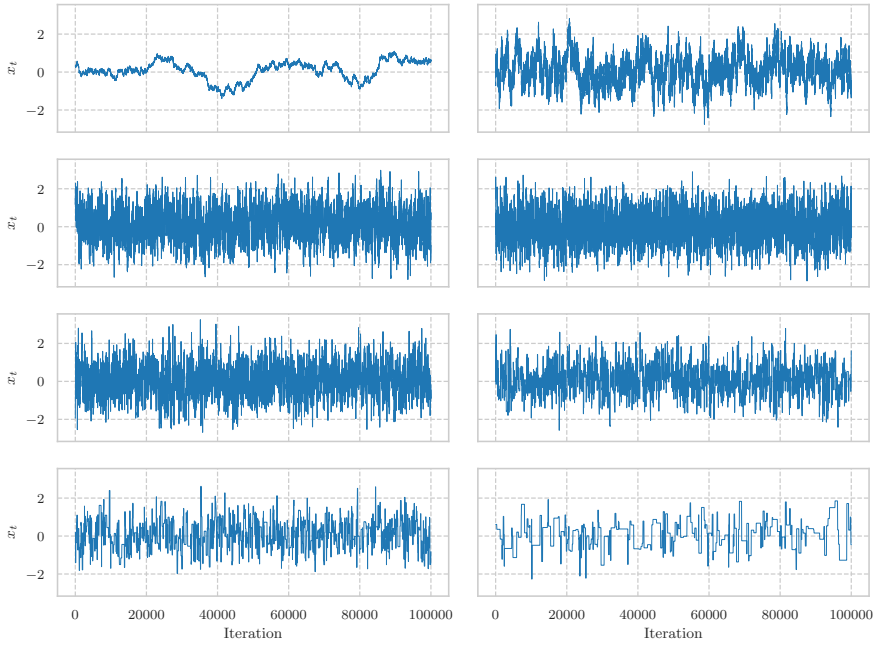

Figure 121: Plots of the first component of the states of the the Exchangeable Particle Gibbs Sampler- applied to the Linear Gaussian model and using the bootstrap proposals as the marginal proposal densities- with $N = 50$, and for a variety of jump-sizes, $\epsilon \in \{0.01, 0.1, 0.25, 0.5, 0.75, 1.0, 1.2, \sqrt{2}\}$, where $\epsilon = 0.01$ for the top-left subplot, $\epsilon = 0.1$ for the top-right subplot, $\epsilon = 0.25$ for the subplot in the second row and the first column, and so on.

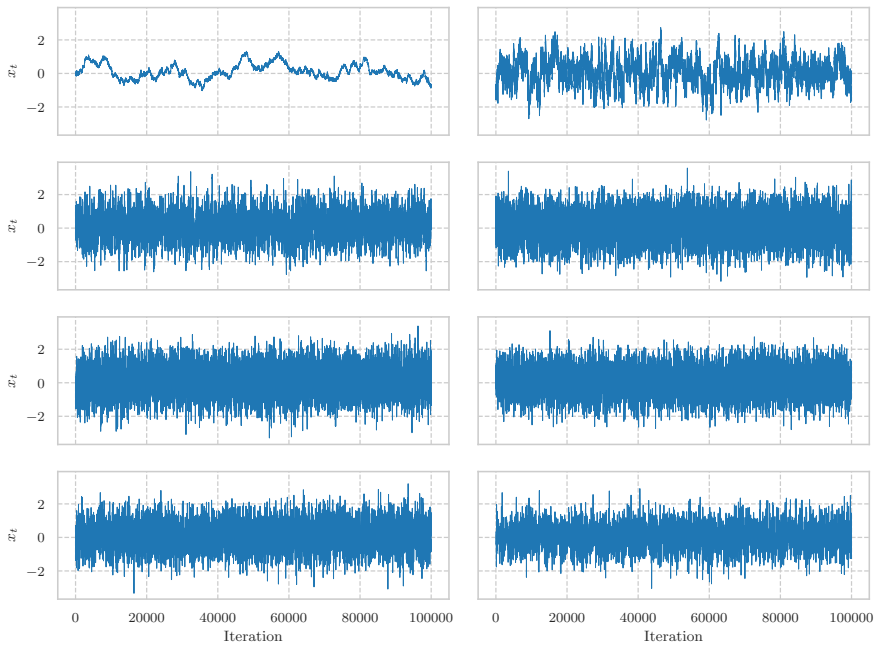

Figure 122: Plots of the first component of the states of the the Exchangeable Particle Gibbs Sampler- applied to the Linear Gaussian model and using the bootstrap proposals as the marginal proposal densities- with $N = 100$, and for a variety of jump-sizes, $\epsilon \in \{0.01, 0.1, 0.25, 0.5, 0.75, 1.0, 1.2, \sqrt{2}\}$, where $\epsilon = 0.01$ for the top-left subplot, $\epsilon = 0.1$ for the top-right subplot, $\epsilon = 0.25$ for the subplot in the second row and the first column, and so on.

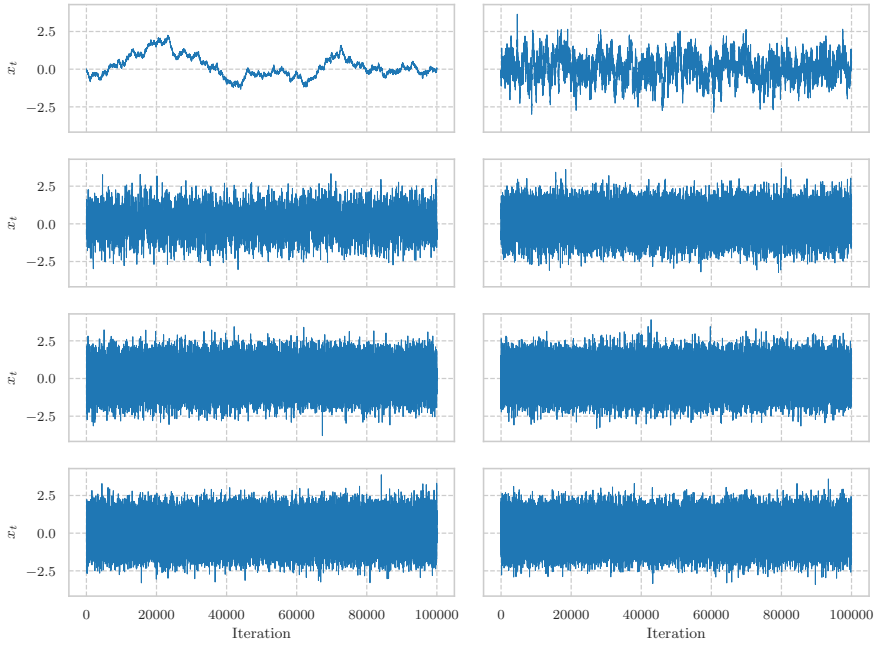

Figure 123: Plots of the first component of the states of the the Exchangeable Particle Gibbs Sampler- applied to the Linear Gaussian model and using the bootstrap proposals as the marginal proposal densities- with $N = 1000$, and for a variety of jump-sizes, $\epsilon \in \{0.01, 0.1, 0.25, 0.5, 0.75, 1.0, 1.2, \sqrt{2}\}$, where $\epsilon = 0.01$ for the top-left subplot, $\epsilon = 0.1$ for the top-right subplot, $\epsilon = 0.25$ for the subplot in the second row and the first column, and so on.

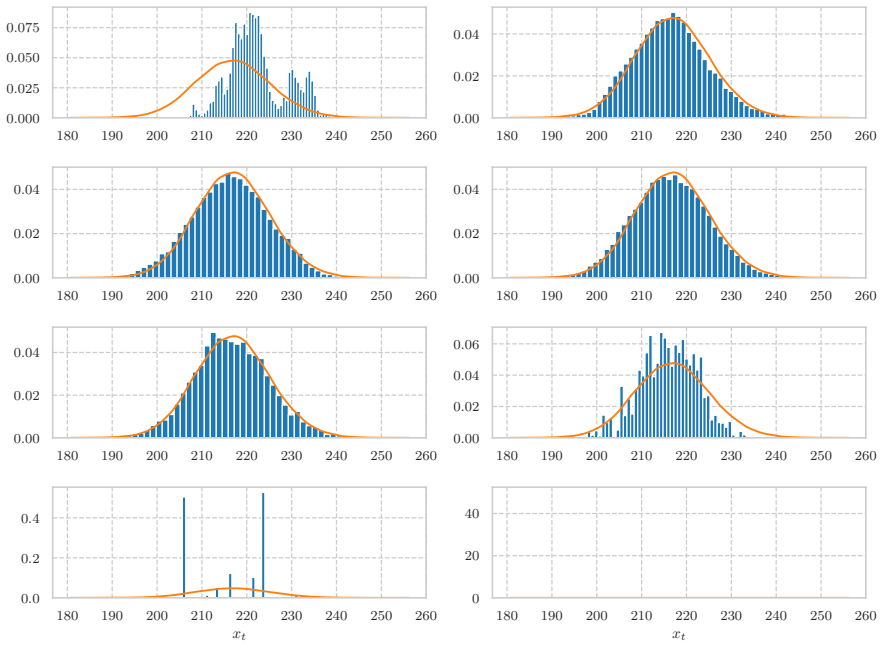

Figure 124: Histograms of the $t = 1$ element of the sample paths, simulated via the Exchangeable Particle Gibbs Sampler applied to the Lotka-Volterra diffusion model and using the Modified Diffusion Bridge proposal of Section 3.2.3 as the marginal proposal distribution, for $N = 50$, and for a variety of jump-sizes, $\epsilon \in \{0.01, 0.1, 0.15, 0.2, 0.3, 0.5, 0.75, \sqrt{2}\}$, where $\epsilon = 0.01$ for the top-left subplot, $\epsilon = 0.1$ for the top-right subplot, $\epsilon = 0.15$ for the subplot in the second row and the first column, and so on. The orange line in each figure corresponds to the *true* density.

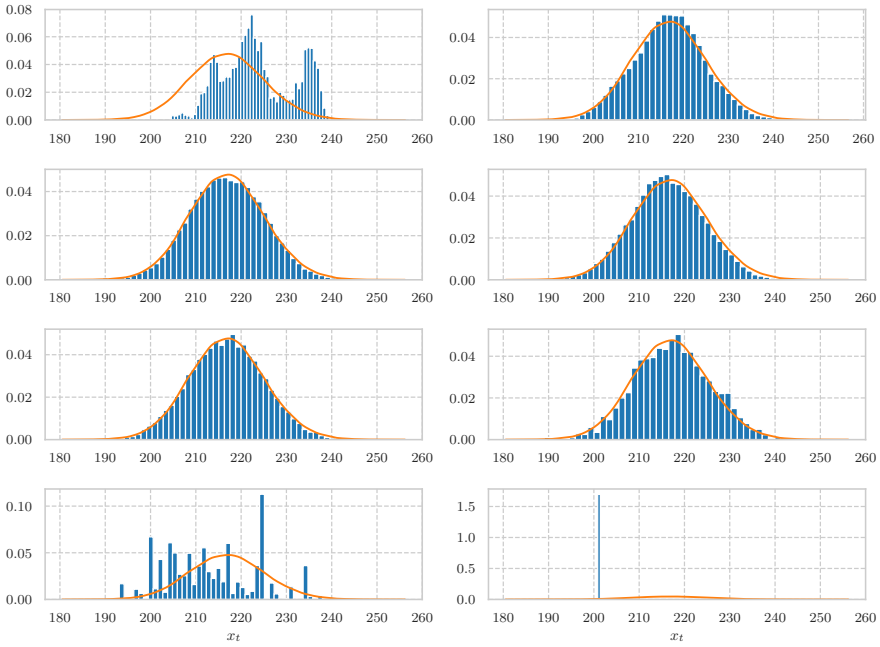

Figure 125: Histograms of the $t = 1$ element of the sample paths, simulated via the Exchangeable Particle Gibbs Sampler applied to the Lotka-Volterra diffusion model and using the Modified Diffusion Bridge proposal of Section 3.2.3 as the marginal proposal distribution, for $N = 100$, and for a variety of jump-sizes, $\epsilon \in \{0.01, 0.1, 0.15, 0.2, 0.3, 0.5, 0.75, \sqrt{2}\}$, where $\epsilon = 0.01$ for the top-left subplot, $\epsilon = 0.1$ for the top-right subplot, $\epsilon = 0.15$ for the subplot in the second row and the first column, and so on. The orange line in each figure corresponds to the *true* density.

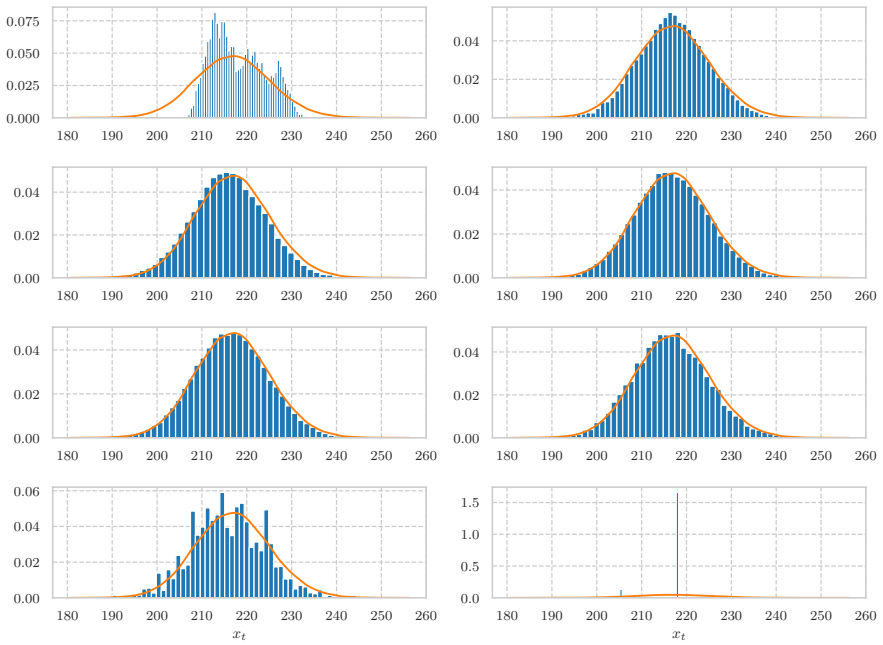

Figure 126: Histograms of the $t = 1$ element of the sample paths, simulated via the Exchangeable Particle Gibbs Sampler applied to the Lotka-Volterra diffusion model and using the Modified Diffusion Bridge proposal of Section 3.2.3 as the marginal proposal distribution, for $N = 250$, and for a variety of jump-sizes, $\epsilon \in \{0.01, 0.1, 0.15, 0.2, 0.3, 0.5, 0.75, \sqrt{2}\}$, where $\epsilon = 0.01$ for the top-left subplot, $\epsilon = 0.1$ for the top-right subplot, $\epsilon = 0.15$ for the subplot in the second row and the first column, and so on. The orange line in each figure corresponds to the *true* density.

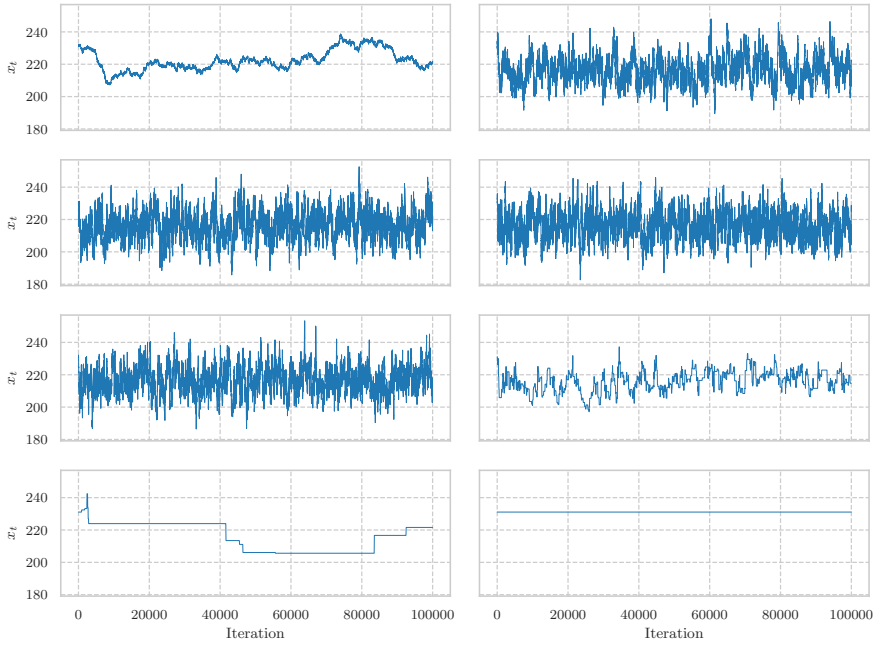

Figure 127: Plots of the $t = 1$ element of the states of the the Exchangeable Particle Gibbs Sampler- applied to the Linear Gaussian model and using the bootstrap proposals as the marginal proposal densities- for $N = 50$, and for a variety of jump-sizes, $\epsilon \in \{0.01, 0.1, 0.15, 0.2, 0.3, 0.5, 0.75, \sqrt{2}\}$, where $\epsilon = 0.01$ for the top-left subplot, $\epsilon = 0.1$ for the top-right subplot, $\epsilon = 0.15$ for the subplot in the second row and the first column, and so on.

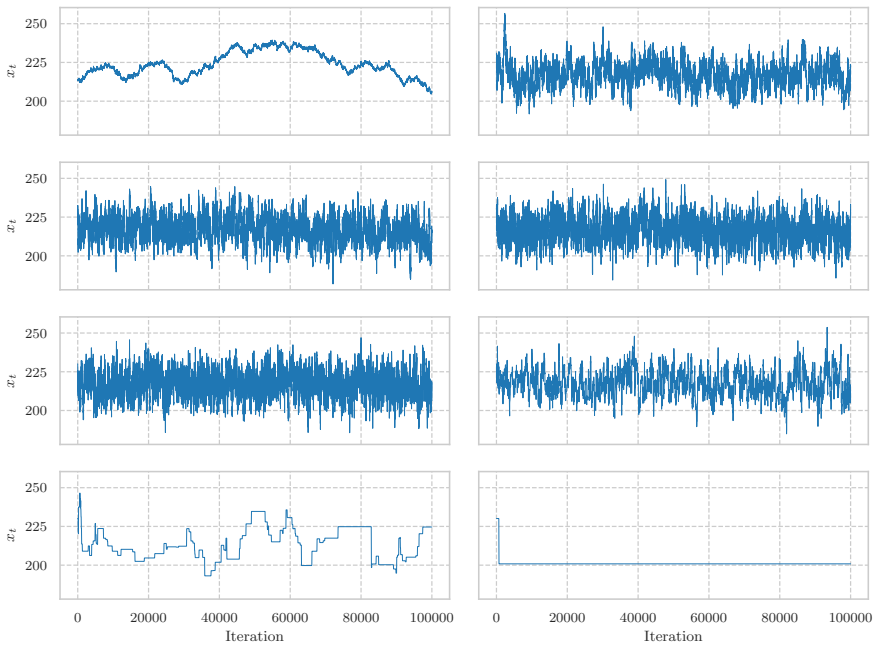

Figure 128: Plots of the $t = 1$ element of the states of the the Exchangeable Particle Gibbs Sampler- applied to the Linear Gaussian model and using the bootstrap proposals as the marginal proposal densities- for $N = 100$, and for a variety of jump-sizes, $\epsilon \in \{0.01, 0.1, 0.15, 0.2, 0.3, 0.5, 0.75, \sqrt{2}\}$, where $\epsilon = 0.01$ for the top-left subplot, $\epsilon = 0.1$ for the top-right subplot, $\epsilon = 0.15$ for the subplot in the second row and the first column, and so on.

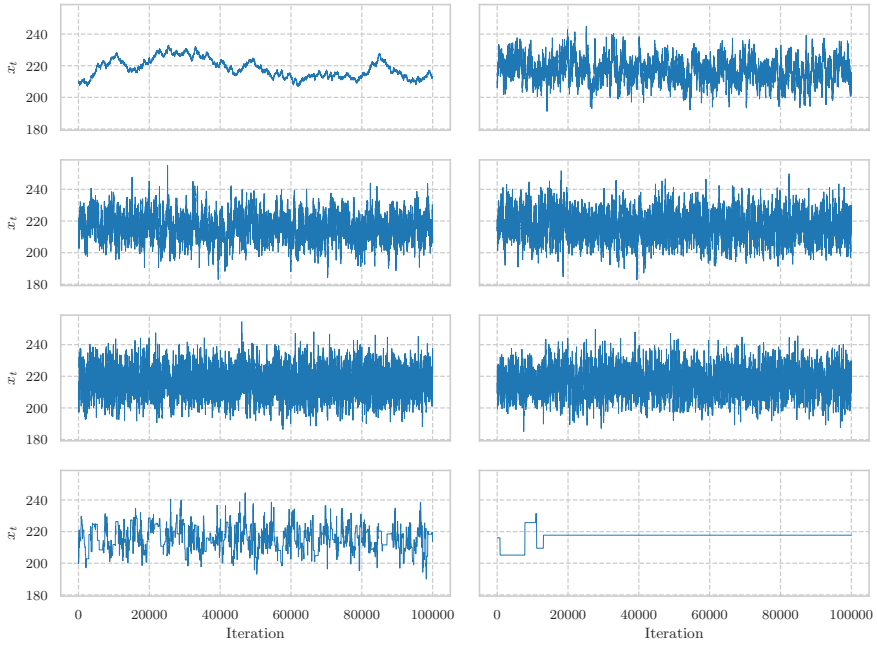

Figure 129: Plots of the $t = 1$ element of the states of the the Exchangeable Particle Gibbs Sampler- applied to the Linear Gaussian model and using the bootstrap proposals as the marginal proposal densities- for $N = 250$, and for a variety of jump-sizes, $\epsilon \in \{0.01, 0.1, 0.15, 0.2, 0.3, 0.5, 0.75, \sqrt{2}\}$, where $\epsilon = 0.01$ for the top-left subplot, $\epsilon = 0.1$ for the top-right subplot, $\epsilon = 0.15$ for the subplot in the second row and the first column, and so on.