Methodology for inference on the Markov modulated Poisson process and theory for optimal scaling of the random walk Metropolis

Christopher Sherlock, MSc.

Submitted for the degree of Doctor of Philosophy

at Lancaster University,

September 2006.

Methodology for inference on the Markov modulated Poisson process and theory for optimal scaling of the random walk Metropolis

Christopher Sherlock, MSc.

Submitted for the degree of Doctor of Philosophy

at Lancaster University, September 2006.

Abstract

Two distinct strands of research are developed: new methodology for inference on the Markov modulated Poisson process (MMPP), and new theory on optimal scaling for the random walk Metropolis (RWM).

A novel technique is presented for simulating from the exact distribution of a continuous time Markov chain over an interval given the start and end states and the infinitesimal generator. This is used to create a Gibbs sampler which samples from the exact distribution of the hidden Markov chain in an MMPP. The Gibbs sampler is compared with several Metropolis-Hastings algorithms on a variety of simulated datasets. It is found that the Gibbs sampler is more efficient than all but one of the Metropolis-Hastings algorithms, sometimes by an order of magnitude. One alternative algorithm, with reparameterisation motivated by a Taylor expansion of the MMPP log-likelihood, outperforms the Gibbs sampler when the different Poisson process intensities are similar. The Gibbs sampler is applied to modelling the occurrence of a rare DNA motif.

Two Lemmas are derived that apply to stationary Metropolis-Hastings Markov chains and simplify the analytical forms for expected acceptance rate and expected square jump distance (ESJD), a measure of efficiency. These are applied to the RWM for elliptically symmetric unimodal targets, and the existence, subject to conditions, of at least one finite optimal scaling is proved in finite dimension d. A one-to-one relationship between acceptance rate and scale parameter is also established. Asymptotic forms for ESJD and expected acceptance rate as $d \to \infty$ are then derived and conditions under which the limiting optimal acceptance rate is 0.234 are obtained. It is also shown that in a more general setting the limiting optimal acceptance rate is ≤ 0.234. Limiting efficiency results are also obtained for partial-blocking and for the exploration of elliptical targets with elliptical proposals of the same shape.

Acknowledgements

This thesis was funded by EPSRC doctoral training grant GR/P02974/01; I am grateful for the opportunity it has provided for studying such interesting topics. I would like to thank my supervisors Paul Fearnhead and Gareth Roberts for making my PhD experience so enjoyable and rewarding. Their support and guidance have been invaluable and I have very much enjoyed discussions with each on the intuition behind the work in this thesis and related ideas. I am also grateful to them both for their flexibility in allowing me the freedom to run with the ideas that have become Chapter 3, and to Gareth for his enthusiasm and encouragement as I attempted to develop the new theory. Paul showed great patience in teaching me to better structure a write up; I am very grateful and do hope it shows in this thesis!

My time at Lancaster would not have been the same without such a friendly supportive department. I would like to thank both staff and co-students for their friendship, advice, academic stimulation, and ability to distract me from work when I needed to take a break. I would particularly like to mention Emma Eastoe, Jamie Kirkham, Theo Kypraios, John Minty, Gerwyn Green, Rosemeire Fiaccone, Peter Diggle, Debbie Costain, Barry Rowlingson, Ting Li Su, Adam Butler and Mark Latham.

I would also like to thank Ludus Dance in Lancaster and my friends with All Jazzed Up in Reading for providing me with continued outlets for my love of Lindy Hop.

After 3 years of shared offices I will finish with the following, from "Shrek2":
Donkey: Are we there yet?

Declaration

This thesis is my own work and has not been submitted in substantially the same form for the award of a higher degree elsewhere. Part of the work in Chapter 2 has been accepted for publication as Fearnhead and Sherlock (2006).

The computationally intensive algorithms in Chapter 2 were coded in the C programming language and used the Gnu Scientific Library for generating random variables and manipulating matrices. All other computational work in this thesis was carried out in the R statistical environment. All computer code was my own.

Chris Sherlock

Contents

List of Figures

List of Tables

Chapter 1

Introduction and background material

1.1 Introduction

The main body of this thesis comprises two separate pieces of research broadly linked under the umbrella of Markov chain Monte Carlo (MCMC). Chapter 2 investigates the Bayesian analysis of the Markov modulated Poisson process using MCMC; it develops new methodology and applies this to a real problem in statistical genetics. Chapter 3 develops the theory of optimal scaling for a particular MCMC algorithm, the random walk Metropolis. This chapter serves as an introduction to both.

The Markov modulated Poisson process has a variety of applications in statistical modelling, which are reviewed at the start of Chapter 2. The main innovation of the chapter is an exact Gibbs sampler for analysing the Markov modulated Poisson process which samples alternately from the exact conditional distribution of

the underlying Markov chain given the parameters and the data, and then from the conditional distribution of the parameters given the underlying Markov chain and the data. The performance of the Gibbs sampler on a variety of simulated data sets is compared with several random walk Metropolis algorithms, including two new reparameterisations. The Gibbs sampler compares favourably with the random walk algorithms; it is then used to analyse occurences of a particur motif in a bacterial DNA. Part of this work (specifically the review Sections 2.1, 2.2, 2.4; Sections associated with the Gibbs sampler and its application 2.3 , 2.5.1, 2.7, and part of the discussion 2.8) has been published as Fearnhead and Sherlock (2006).

The random walk Metropolis is one of the most common MCMC algorithms employed in practice. The practising statistician must choose a scale parameter for the jump proposal distribution, yet the impact of this choice on the efficiency of an algorithm can be of many orders of magnitude. A sensible choice of scale parameter is therefore vital for obtaining accurate MCMC estimates in a reasonable time. Current theory on optimal scaling for the random walk Metropolis applies in the limit as dimension $d \to \infty$ and is reviewed at the start of Chapter 3. The chapter then develops a new theory for certain target distributions, initially deriving exact formulae in finite dimension d, before examining the limiting behaviour described by these formulae as $d \to \infty$.

Both of the main Chapters take for granted some fundamental ideas about MCMC: convergence, mixing, and a knowledge of several of the basic types of algorithm. These are reviewed in this introductory chapter, as are basic ideas about Langevin diffusions which are needed for the literature review on current optimal scaling

theory. This chapter also includes a short summary of properties of the beta function and the surface areas of hypershperes, both of which are required in Chapter 3. We start, however, with a simple introduction to the Markov modulated Poisson process.

1.2 The Markov modulated Poisson process

Chapter 2 of this thesis examines in depth the analysis of the Markov modulated Poisson process (MMPP). It reviews applications, likelihood calculations and Bayesian analysis, as well as detailing new work on an exact Gibbs sampler, on an efficient reparameterisation of the two-dimensional MMPP, and on a simulation study comparison between several MCMC approaches. This section sets the scene by introducing the MMPP itself with graphical illustrations.

Let X_t be a continuous time Markov chain on discrete state space $\{1, \ldots, d\}$ and let $\boldsymbol{\lambda}$ be a d-dimensional vector of (non-negative) intensities. The linked but stochastically independent Poisson process Y_t whose intensity is λ_{X_t} is a Markov modulated Poisson process - it is a Poisson process whose intensity is modulated my a continuous time Markov chain.

The idea is best illustrated through an example, which also serves to introduce notation that will be used throughout Chapter 2. Consider a two-dimensional Markov chain X_t with generator

$$\mathbf{Q} = \begin{bmatrix} -2 & 2 \\ 1 & -1 \end{bmatrix}$$

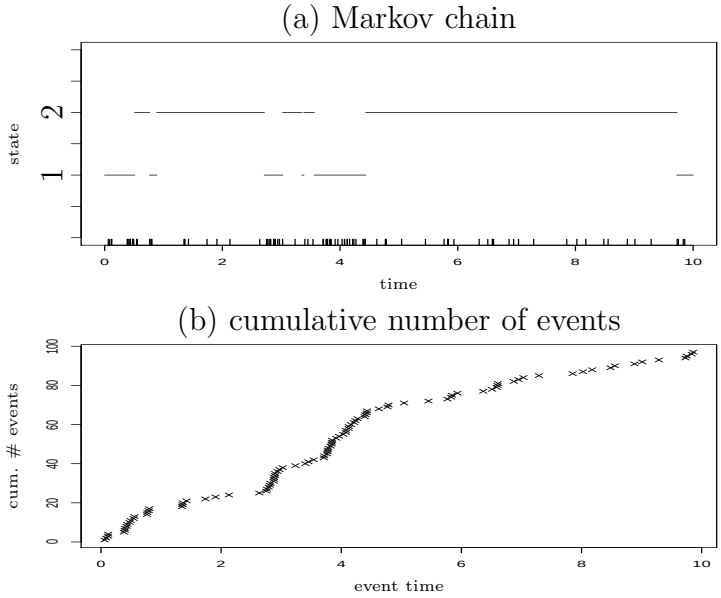

Figure 1.1: (a) A two state continuous time Markov chain simulated from generator $\mathbf{Q}$ with $q_{12} = 2$ and $q_{21} = 1$; the rug plot shows events from an MMPP simulated from this chain, with intensity vector $\boldsymbol{\lambda} = (20, 2)$. (b) Cumulative number of events of the MMPP against time.

Figure 1.1a shows a realisation from this chain over a period of 10 seconds. Notice that the chain spends approximately (in fact slightly less than) one third of its time in state 1 and two thirds of its time in state 2, as would be expected from its stationary distribution $\boldsymbol{\nu} = (1/3, 2/3)$. Now consider a Poisson process Y_t which has intensity 20 when X_t is in state 1 and intensity 2 when X_t is in state 2. This is an MMPP with intensity vector

$$\boldsymbol{\lambda} = [20, 2]$$

A realisation (obtained via the realisation of X_t) is shown as a rug plot underneath the chain. The variation in concentration of points in the process Y_t is exactly the characteristic in real processes that the MMPP as a statistical model aims to capture. This can also be seen through the variation in the slope of the cumulative plot of number of Y_t-events against time (Figure 1.1b).

1.2.1 Degenerate solutions

A given MMPP covers all lower dimensional MMPP's as special cases with certain combinations of parameter values. This degeneracy is evident in the simulation studies of Chapter 2. It is also responsible for strictly positive lower bounds on the likelihood as certain parameters approach 0 or ∞, which in turn disallows the use of improper priors for these parameters under a Bayesian analysis. This section provides an intuition into degeneracy and some consequences.

For simplicity we consider only a 2-dimensional MMPP with parameters

$$\mathbf{Q} = \begin{bmatrix} -q_{12} & q_{12} \\ q_{21} & -q_{21} \end{bmatrix}$$

and

$$\boldsymbol{\lambda} = [\lambda_1, \lambda_2]^t$$

Denote the chain's stationary distribution as

$$\boldsymbol{\nu} := [\nu_1, \nu_2]^t = \frac{1}{q_{12} + q_{21}} [q_{21}, q_{12}]^t$$

Intuitively the observed data from this approaches that of a 1-dimensional MMPP (i.e. a Poisson process) with parameter

$$\overline{\lambda} = \frac{q_{21}\lambda_1 + q_{12}\lambda_2}{q_{12} + q_{21}} = \nu_1\lambda_1 + \nu_2\lambda_2$$

as (for example)

1. $q_{21} \to \infty$ with λ_1, q_{12} fixed and $\lambda_2/q_{21} \to 0$
2. $q_{12} \to 0$ with λ_1, $q_{21} > 0$ fixed and $q_{12}\lambda_2 \to 0$
3. $q_{21} + q_{12} \to \infty$ with $\boldsymbol{\lambda}$ and $\boldsymbol{\nu}$ fixed
4. $\lambda_2 \to \lambda_1$ with all other parameters fixed.

In the first two cases the chain spends a larger and larger fraction of its time in a single state (state 1), and in the third case it oscillates so quickly between the states that (as far as the observer is concerned) it is effectively in a mean state somewhere between the two.

Now consider data from a general 2-dimensional MMPP for which n events are observed over a time window $[0, t_{obs}]$. To keep notation consistent with Chapter 2, let $\mathbf{t}'$ be the vector of event-times. Thoughout Chapter 2 it is assumed that the

underlying Markov chain starts at stationarity; given this, the likelihood for this (ordered) data is

$$L(\mathbf{Q}, \mathbf{\Lambda}|\mathbf{t}') = P(\mathbf{t}'|\mathbf{Q}, \mathbf{\Lambda}) \geq P(\mathbf{t}', \text{ chain always in state } 1|\mathbf{Q}, \mathbf{\Lambda}) = \nu_1 \, \lambda_1^n \, e^{-(\lambda_1 + q_{12}) \; t_{obs}}$$

Intuitively this scenario "chain always in state 1" corresponds to limiting cases 1 and 2 above. For a non-trivial MMPP at least one of the intensity parameters must be non-zero; without loss of generality let this be λ_1. Thus the likelihood has a strictly positive lower bound

$$L_{min}(\lambda_1) := \frac{q_{21}^{min}}{q_{12}^{max} + q_{21}^{min}} \; \lambda_1^n \; e^{-(\lambda_1 + q_{12}^{max}) \; t_{obs}}$$

on the infinite region of the state space

$$(\lambda_1, \lambda_2, q_{12}, q_{21}) \in (0, \infty) \times (0, \infty) \times (0, q_{12}^{max}) \times (q_{21}^{min}, \infty)$$

for any positive q_{21}^{min} and q_{12}^{max}. Therefore priors for λ_2, q_{12} and q_{21} which are improper over this region will lead to improper posteriors for these parameters.

1.3 Markov Chain Monte Carlo

Markov chain Monte Carlo (MCMC) algorithms provide a framework for simulating from a target random variable $\mathbf{X}$ with distribution $\pi(\cdot)$ by iteratively generating a Markov chain $\mathbf{X}_0, \mathbf{X}_1, \ldots$ with stationary distribution $\pi(\cdot)$. A Monte Carlo estimate of some function of the target random variable may then be obtained; for example

$$\hat{f}_N := \frac{1}{N} \sum_{1}^{N} f(\mathbf{X}_i) \tag{1.1}$$

Methodological development of MCMC is central to Chapter 2 of this thesis, while Chapter 3 is dedicated to one particular aspect of MCMC theory. The purpose of

this section is to introduce the various facets of MCMC that will be required in the following two Chapters. Much of the material in this section is taken from Gilks et al. (1996) and this should be assumed to be the source if no further reference is given. MCMC can be applied to discrete or continuous random variables or mixtures thereof; however only continuous random variables are investigated in Chapters 2 and 3 and so for simplicity this section is also confined to the exploration these. We first give an overview of a number of MCMC algorithms and then discuss the concepts of convergence and mixing of a Markov chain which interweave naturally with practical Monte Carlo estimation of functions of random variables and their standard error. We then detail some simple strategies for improving the efficiency of MCMC algorithms.

1.3.1 MCMC Algorithms

An incredible variety of algorithms are available under the general umbrella of MCMC. Many of these can be most easily understood in the context of the Metropolis-Hastings algorithm. We describe this first, including the ideas of stationarity and reversibility at equilibrium. We detail several sub-classes of algorithm including the random walk Metropolis before discussing the concept of partial-blocking which leads naturally to the Gibbs sampler.

1.3.1.1 The Metropolis Hastings Algorithm

The Metropolis-Hastings updating scheme provides a very general class of algorithms which proceed as follows: given current value $\mathbf{X}$, a new value $\mathbf{X}^*$ is proposed from pre-specified Lesbegue density $q(\mathbf{X}^*|\mathbf{X})$ and is then accepted or rejected according to acceptance probability

$$\alpha(\mathbf{x}, \mathbf{x}^*) = \min\left(1, \frac{\pi(\mathbf{x}^*)q\left(\mathbf{x}|\mathbf{x}^*\right)}{\pi(\mathbf{x})q\left(\mathbf{x}^*|\mathbf{x}\right)}\right) \tag{1.2}$$

If the proposed value is accepted it becomes the next current value ($\mathbf{X}' \leftarrow \mathbf{X}^*$), otherwise the current value is left unchanged ($\mathbf{X}' \leftarrow \mathbf{X}$).

Write $P(d\mathbf{x}'|\mathbf{x})$ for the conditional probability that the next value $\mathbf{X}'$ is in the hypercube with opposite corners $\mathbf{x}'$ and $\mathbf{x}' + d\mathbf{x}'$, given that $\mathbf{X} = \mathbf{x}$. We also define the joint measure of two successive realisations from the chain at stationarity as

$$A(d\mathbf{x}, d\mathbf{x}') := \pi(\mathbf{x})\ d\mathbf{x}\ P(d\mathbf{x}'|\mathbf{x})$$

it is implicit that this is valid at $(\mathbf{x}, \mathbf{x}')$.

The acceptance probability (1.2) is chosen exactly so that the chain is reversible at equilibrium with invariant distribution $\pi(\cdot)$. From the definition of $\alpha(\cdot, \cdot)$

$$\pi(\mathbf{x})q\left(\mathbf{x}^*|\mathbf{x}\right)\alpha\left(\mathbf{x}, \mathbf{x}^*\right) = \pi(\mathbf{x}^*)q\left(\mathbf{x}|\mathbf{x}^*\right)\alpha\left(\mathbf{x}^*, \mathbf{x}\right)$$

which leads directly to reversibility

$$A(d\mathbf{x}, d\mathbf{x}') = A(d\mathbf{x}', d\mathbf{x})$$

This in turn implies that $\pi(\cdot)$ is invariant since

$$\int_{\mathbf{x}\in\Re} d\mathbf{x}\ \pi(\mathbf{x})P\left(d\mathbf{x}'|\mathbf{x}\right) = \int_{\mathbf{x}\in\Re} d\mathbf{x}'\ \pi(\mathbf{x}')P\left(d\mathbf{x}|\mathbf{x}'\right) = \pi(\mathbf{x}')$$

Convergence of the chain to $\pi(\cdot)$ is discussed in Section 1.3.2.

The statistician is free to choose their proposal distribution as they like, and different types of proposal distribution lead to different classes of Metropolis-Hastings algorithm. In this thesis we are especially interested in the **random-walk Metropolis** (RWM), as applied in Metropolis et al. (1953). Here the difference between the

proposal and the current value (i.e. the proposed jump) is independent of the current value and is symmetrically distributed. For a RWM algorithm in d dimensions we therefore have

$$q\left(\mathbf{x}^*|\mathbf{x}\right) = \frac{1}{\lambda^d}\, r\left(\frac{\mathbf{x}^* - \mathbf{x}}{\lambda}\right)$$

with $r(\mathbf{y}) = r(-\mathbf{y})$ for all $\mathbf{y}$. In this case the acceptance probability simplifies to

$$\alpha(\mathbf{x}, \mathbf{x}^*) = \min\left(1, \frac{\pi(\mathbf{x}^*)}{\pi(\mathbf{x})}\right)$$

This algorithm is employed extensively in the simulation study of Chapter 2, and the process of tuning the scaling parameter λ to maximise the efficiency of the algorithm is the main subject of Chapter 3.

A related algorithm, the **multiplicative random-walk** (see for example Dellaportas and Roberts, 2003) is also employed in Chapter 2. This is simply the symmetric random-walk applied to the logarithm of each of the (non-negative) target components. Since taking logarithms shifts mass from the tails to the centre of the distribution, the multiplicative random-walk is especially efficient at exploring heavy-tailed targets. When viewed on the original target, proposed jumps are of course exponential multiples of the current value; the acceptance probability is then

$$\alpha(\mathbf{x}, \mathbf{x}^*) = \min\left(1, \frac{\prod_1^d x_i^*}{\prod_1^d x_i}\frac{\pi(\mathbf{x}^*)}{\pi(\mathbf{x})}\right)$$

In Chapter 3 we will have brief cause to mention two further Metropolis-Hastings algorithms: the **independence sampler** and the **Metropolis adjusted Langevin algorithm (MALA)**. In the former, the next value proposed is independent of the current value: $q\left(\mathbf{x}^*|\mathbf{x}\right) = r(\mathbf{x}^*)$, and the acceptance probability is therefore

$$\alpha = \min\left(1, \frac{\pi(\mathbf{x}^*)r(\mathbf{x})}{\pi(\mathbf{x})r(\mathbf{x}^*)}\right)$$

The MALA algorithm is a variant on the symmetric random-walk in which the proposed jump is biased towards the direction of increasing target density:

$$\mathbf{X}^*|\mathbf{x} \sim N\left(\mathbf{x} + \frac{\lambda^2}{2}\nabla \log \pi\left(\mathbf{x}\right), \lambda^2 \mathbf{I}_d\right)$$

The bias amount is such that the stationary distribution of the limiting process as the discretisation approaches zero is $\pi(\cdot)$.

1.3.1.2 Partial-blocking

For reasons of efficiency or analytical convenience components of the multidimensional target $\mathbf{X}$ might be grouped into k sub-blocks which are updated sequentially. In general write

$$\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_k)$$

where, in this section, and the next $\mathbf{X}_i$ is the i^{th} block of components of the current element of the chain rather than i^{th} element of the chain. It will be convenient to define the shorthand

$$\mathbf{x}_{-i} := \mathbf{x}'_1, \ldots, \mathbf{x}'_{i-1}, \mathbf{x}_{i+1}, \ldots, \mathbf{x}_k$$

One complete update of $\mathbf{X}$ then consists of k sequential updates for which the i^{th} component is proposed from $q_i\left(\mathbf{x}_i^*|\mathbf{x}_i, \mathbf{x}_{-i}\right)$ and accepted with probability

$$\alpha\left((\mathbf{x}_i, \mathbf{x}_{-i}), (\mathbf{x}_i^*, \mathbf{x}_{-i})\right) = \frac{\pi(\mathbf{x}_i^*, \mathbf{x}_{-i}) q_i\left(\mathbf{x}_i|\mathbf{x}_i^*, \mathbf{x}_{-i}\right)}{\pi(\mathbf{x}_i, \mathbf{x}_{-i}) q_i\left(\mathbf{x}_i^*|\mathbf{x}_i, \mathbf{x}_{-i}\right)} \tag{1.3}$$

Proposals for each partial update need not have the same form. For example one could be a random walk Metropolis and the next an independence sampler; this would be a **hybrid** algorithm. Acceptance probabilities are still always chosen so that each partial update is reversible at equilibrium, and the stationary distribution remains $\pi(\cdot)$.

Note that the complete update, being a sequence of k partial-updates, is not in general reversible at equilibrium. However an adaptation of the algorithm where a complete update consists of all the above partial-updates and then the same set of partial-updates in the opposite order *is* reversible. Alternatively reversibility can be achieved by performing only a single partial-update on a fixed number of components but with the components to be updated chosen at random from the full set; this scheme is often referred to as *random scan*. Algorithms which employ partial-blocking are sometimes referred to as "Metropolis-within-Gibbs"; a somewhat cryptic reference to the Gibbs sampler, which we now describe.

1.3.1.3 The Gibbs sampler

Suppose that the proposal distribution of the i^{th} partial update is actually the conditional distribution of $\mathbf{X}_i$ given $\mathbf{X}'_1, \ldots, \mathbf{X}'_{i-1}, \mathbf{X}_{i+1}, \ldots, \mathbf{X}_k$ which we denote by the shorthand $\pi_i(\mathbf{x}_i|\mathbf{x}_{-i})$. Then denoting by $\pi_{-i}(\mathbf{x}_{-i})$ the marginal distribution of $\mathbf{x}'_1, \ldots, \mathbf{x}'_{i-1}, \mathbf{x}_{i+1}, \ldots, \mathbf{x}_k$ we have that

$$q_i\left(\mathbf{x}_i^*|\mathbf{x}_i, \mathbf{x}_{-i}\right) = \pi_i(\mathbf{x}_i^*|\mathbf{x}_{-i}) = \frac{\pi(\mathbf{x}_i^*, \mathbf{x}_{-i})}{\pi_{-i}(\mathbf{x}_{-i})}$$

with a similar result for $q_i\left(\mathbf{x}_i|\mathbf{x}_i^*, \mathbf{x}_{-i}\right)$. The acceptance probability (1.3) is therefore 1. The name of this algorithm, the **Gibbs sampler**, arises from its use by Geman and Geman (1984) to analyse Gibbs distributions, but its application is far more general. One of the main innovations in Chapter 2 of this thesis is an exact Gibbs sampler for analysing the Markov modulated Poisson process.

1.3.2 Convergence and mixing of an MCMC Markov chain

Two main (and related) issues arise with regard to the efficiency of MCMC algorithms:

Convergence: The chain is unlikely to have been initialised from its stationary distribution (since if this were straightforward there would be no need for MCMC) and so a certain number of iterations are required for elements of the chain to be samples from the target distribution $\pi(\cdot)$.

Mixing: Once stationarity has been achieved the chain produces *dependent* identically distributed samples from $\pi(\cdot)$. A certain number of iterations are required to explore the target well enough to produce Monte Carlo estimates of the desired accuracy. This number of iterations is in general more than would be necessary if the elements of the chain were independent.

Note: in practice most chains never achieve perfect stationarity. In this thesis a chain is referred to as having 'reached stationarity' or 'converged' when the distribution from which an element is sampled is as close to the stationary distribution as to make no practical difference to any Monte-Carlo estimates.

For an efficient algorithm both the number of iterations required for convergence and the number of iterations needed to expore the target should be relatively small. Figure 1.2 shows so called "traceplots" of the first 1000 iterations for each of three chains exploring a standard one-dimensional Gaussian target distribution $\pi(x) = \phi(x)$ and initialised at $x = 20$. The first of these converges slowly and then mixes poorly; the second converges quickly but mixes poorly and the third converges relatively quickly and mixes well.

In Chapter 2 of this thesis we will be concerned with practical determination of a point at which a chain has converged. The method we employ is simple heuristic examination of the trace plots for the different components of the chain. Note that

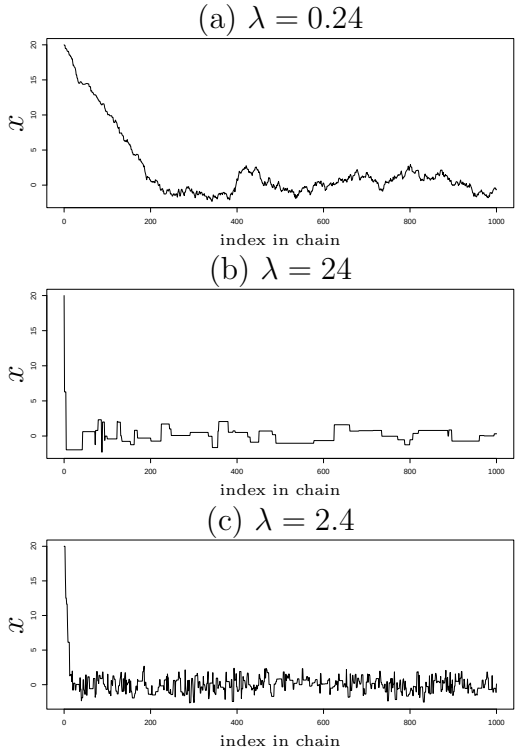

Figure 1.2: Trace plots for exploration of a standard Gaussian initialised from $x = 20$ and using the random walk Metropolis algorithm with Gaussian proposal. Proposal scale parameters for the three plots were respectively (a) 0.24, (b) 24, and (c) 2.4.

since the state space is multi-dimensional it is not sufficient to simply examine a single component. Alternative techniques are discussed in Chapter 7 of Gilks et al. (1996).

The degree to which a chain has converged can be measured by the total variational distance. For two distributions ν_1 and ν_2 on state space E with sigma algebra $\sigma(E)$, this is defined as

$$||\nu_1 - \nu_2|| := 2 \sup_{A \in \sigma(E)} |\nu_1(A) - \nu_2(A)|$$

A measure of the degree of convergence of a chain initialised at $\mathbf{x}$ and run for n iterations is therefore $||P^n(\mathbf{x}, \cdot) - \pi(\cdot)||$, where $P^i(\mathbf{x}, \cdot)$ is the distribution of a chain after i iterations from initial point at $\mathbf{x}$.

Theoretical criteria for ensuring convergence of MCMC Markov chains are examined in detail in Chapters 3 and 4 of Gilks et al. (1996) and references therein, and will not be discussed here. We do however wish to highlight the concept of geometric ergodicity. A Markov chain is geometrically ergodic with stationary distribution $\pi(\cdot)$ if

$$||P^n(\mathbf{x}, \cdot) - \pi(\cdot)|| \leq M(\mathbf{x}) r^n \tag{1.4}$$

for some positive $r < 1$ and $M(\cdot)$. Geometric convergence of the Gibbs sampler and of the RWM is discussed in Chapter 3 of Gilks et al. (1996). As well as relating to the speed of convergence, geometric ergodicity also guarantees a central limit theorem for Monte Carlo estimates such as (1.1) for functions $f(\cdot)$ such that

$$\int d\mathbf{x} \pi(\mathbf{x}) \left| f(\mathbf{x}) \right|^{2+\epsilon} < \infty \tag{1.5}$$

for some small $\epsilon > 0$. In this case

$$N^{1/2}\left(\hat{f}_N - \mathbb{E}_\pi\left[f(\mathbf{X})\right]\right) \;\Rightarrow\; N(0, \sigma_f^2) \tag{1.6}$$

where $\Rightarrow$ denotes convergence in distribution and

$$\sigma_f^2 := \mathrm{Var}_\pi\left[f(\mathbf{X})\right] < \infty$$

The central limit theorem (1.6) guarantees not only convergence of the Monte Carlo estimate (1.1) but also supplies its standard error, which decreases as $N^{-1/2}$.

Note: Total variation distance is a natural measure for defining convergence since (e.g. Meyn and Tweedie, 1993) geometric convergence as defined in (1.4) actually guarantees the given level of convergence for $f(\mathbf{X})$ for all integrable functions $f(\cdot)$. For more general functions, other distance measures may be used to define convergence, for example the f-norm and the V-norm, which is defined in terms of the f-norm (e.g. Meyn and Tweedie, 1993):

$$\begin{aligned} ||\nu_1 - \nu_2||_f &:= \sup_{g:|g|\leq f} \left| \int d\nu_1(\mathbf{x})\; g(\mathbf{x}) - \int d\nu_2(\mathbf{x})\; g(\mathbf{x}) \right| \\ |||P_1(\mathbf{x},\cdot) - P_2(\mathbf{x},\cdot)|||_V &:= \sup_x \frac{||P_1(\mathbf{x},\cdot) - P_2(\mathbf{x},\cdot)||_V}{V(x)} \end{aligned}$$

for some $f \geq 0$, some $1 \leq V < \infty$, and a chain initialised at $\mathbf{x}$. In this thesis, however our interest in the convergence of a chain is motivated by the desire for a central limit theorem such as (1.6); this theorem is used implicitly in Chapter 2. The likelihood of an MMPP with maximum and minimum Poisson intensities λ_{max} and λ_{min} and with n events observed over a time window of length t_{obs}, is bounded above by $\lambda_{max}^n e^{-\lambda_{min} t_{obs}}$. In Chapter 2 only parameters and their logarithms are considered; since exponential priors are employed the posterior then satisfies (1.5).

We therefore make no further discussion of norms.

A more accurate estimate than (1.1) is likely to be obtained by discarding the portion of the chain $\mathbf{X}_0, \ldots, \mathbf{X}_m$ up until the point at which it was deemed to have reached stationarity; iterations $1, \ldots m$ are commonly termed "burn in". Using only the remaining elements $\mathbf{X}_{m+1}, \ldots, \mathbf{X}_{m+n}$ (with $m+n=N$) our Monte Carlo estimator becomes

$$\hat{f}_n := \frac{1}{n} \sum_{m+1}^{m+n} f(\mathbf{X}_i) \tag{1.7}$$

Convergence and burn in are not discussed any further here and for the rest of this section the chain is assumed to have started at stationarity and continued for n further iterations. For a *stationary* chain, $\mathbf{X}_0$ is sampled from $\pi(\cdot)$, and so for all $k > 0$ and $i \geq 0$

$$\mathrm{Cov}\left[f(\mathbf{X}_k), f(\mathbf{X}_{k+i})\right] = \mathrm{Cov}\left[f(\mathbf{X}_0), f(\mathbf{X}_i)\right]$$

This is the *autocorrelation* at lag i. Therefore at stationarity

$$\sigma_f^2 = \lim_{n\to\infty} n\mathrm{Var}\left[\hat{f}_n\right] = \mathrm{Var}\left[f(\mathbf{X}_0)\right] + 2\sum_{1}^{\infty} \mathrm{Cov}\left[f(\mathbf{X}_0), f(\mathbf{X}_i)\right]$$

If elements of the stationary chain were independent then σ_f^2 would simply be $\mathrm{Var}\left[f(\mathbf{X}_0)\right]$ and so a measure of the inefficiency of the Monte-Carlo estimate $\hat{f}_n$ relative to the perfect i.i.d. sample is

$$\frac{\sigma_f^2}{\mathrm{Var}\left[f(\mathbf{X}_0)\right]} = 1 + 2\sum_{1}^{\infty} \mathrm{Corr}\left[f(\mathbf{X}_0), f(\mathbf{X}_i)\right] \tag{1.8}$$

This is the *integrated autocorrelation time* (ACT) and represents the effective number of dependent samples that is equivalent to a single independent sample. Alternatively $n^* = n/ACT$ may be regarded as the effective equivalent sample size if

the elements of the chain had been independent.

To estimate the ACT in practice one might examine the chain from the point at which it is deemed to have converged and estimate the lag-i autocorrelation $\mathrm{Corr}\left[f(\mathbf{X}_0), f(\mathbf{X}_i)\right]$ by

$$\hat{\gamma}_i = \frac{1}{n-i}\sum_{j=1}^{n-i}\left(f(\mathbf{X}_j) - \hat{f}_n\right)\left(f(\mathbf{X}_{j+i}) - \hat{f}_n\right) \tag{1.9}$$

Naively, substituting these into (1.8) gives an estimate of the ACT. But as noted for example in Geyer (1992) this estimate is not even consistent. For sensibly large n most of the estimated terms (1.9) consist the mean of products of two effectively independent realisations of $f(\mathbf{X}) - \hat{f}_n$, and have finite variance $O(1/(n-i))$. This is evident in Figure 1.3c which shows the estimated autocorrelation function from the last 700 iterations of the simulated chain in Figure 1.2(c). The sum of these terms consists of random noise with variance at least $O(1)$.

The simple solution employed in Chapter 2 is to visually inspect the estimated autocorrelations and then truncate the sum (1.8) at a lag l after which the autocorrelations appear to be mostly noise. This gives the estimator

$$\mathrm{ACT}_{\mathrm{est}} := 1 + 2\sum_{i=1}^{l}\hat{\gamma}_i \tag{1.10}$$

Geyer (1992) gives references for regularity conditions under which this estimator is consistent. He also discusses extensions of this window estimator and compares these with alternatives.

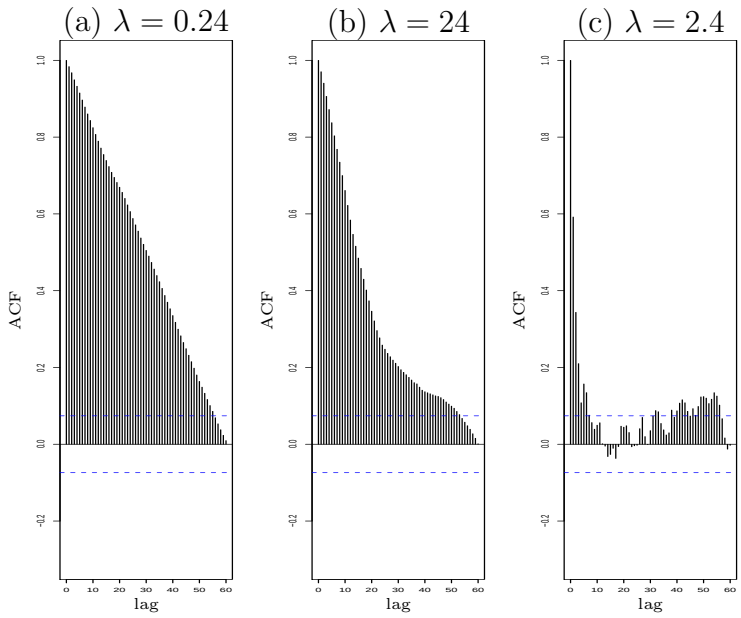

Figure 1.3: Estimated autocorrelation functions up to lag-60 for iterations 301 to 1000 of the trace plots shown in Figure 1.2. Graphs correspond to proposal scale parameters of respectively (a) 0.24, (b) 24, and (c) 2.4.

1.3.3 Improving the efficiency of MCMC algorithms

We now examine some simple strategies for improving the efficiency of MCMC algorithms.

For algorithms such as the RWM, the multiplicative random-walk, and the MALA, the scaling parameter of the jump proposal distribution may be tuned to improve the efficiency of the mixing. If proposed jumps are too small then most will be accepted but the distance moved will be small; if proposals are too large then they will rarely be accepted and again the target will be explored inefficiently. Figure 1.2 shows the exploration of a univariate standard Gaussian target via the RWM algorithm with a Gaussian proposal but three different scale parameters. Stationarity is achieved in all three runs by the 300^{th} iteration but mixing is extremely slow in the first and second runs where the scale parameter is respectively too small and too large. In the third run the scale parameter is close to its optimal value and mixing is nearly as efficient as is possible with a Gaussian proposal. The relative mixing efficiencies of the three chains may also be compared through the corresponding autocorrelation plots in Figure 1.3. These show estimated autocorrelatation functions up to lag-60 for each of the three chains from the 300^{th} iteration onwards. The area under the graphs (prior to them being dominated by random variation) gives an estimate of the ACT; this is clearly much higher for scale parameters of 0.24 and 24 than it is for a scale parameter of 2.4. Optimal scaling for the RWM algorithm is the subject of the third Chapter of this thesis and a detailed literature review is contained therein. Optimal scaling for the MALA and RWM algorithms is also reviewed in Roberts and Rosenthal (2001).

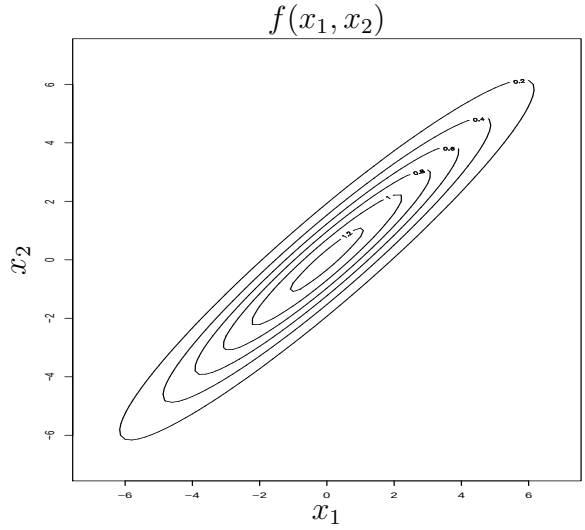

Figure 1.4: Contour plot for a two-dimensional Gaussian density with $\sigma_1^2 = \sigma_2^2 = 1$, and correlation $\rho = 0.95$.

Consider a target in which individual components are highly correlated; we will work with the example of a two component target the contours of which form a very flattened ellipse at 45 degrees to the axes, as shown in Figure 1.4. A spherical proposal distribution for a single block Metropolis-Hastings update will clearly explore the target very inefficiently: any scale parameter which allows reasonably sized jumps along the major axis of the ellipse will usually be rejected as most proposals will have a significant component along the minor axis. Sequential partial updates (for example using a Gibbs sampler or the RWM) along the x_1 and x_2 axes will also be constrained due to the narrowness of the ellipse, and exploration will again be slow.

The single block update would be made more efficient if proposals were from an

elliptical distribution with a similar shape and orientation as the target. The shape of the target is of course unknown, but might be estimated from the chain output as the chain evolves, and this is the basis behind adaptive direction sampling as considered for example in Chapter 6 of Gilks et al. (1996). Alternatively, consider the reparametrisation: $y_1 = (x_1 + x_2)/2$ and $y_2 = (x_1 - x_2)/2$; the first of these corresponds to the major axis of the ellipse and second to the minor axis. It is easy to see that a Gibbs sampler using y_1 and y_2 would proceed much more efficiently than that using x_1 and x_2. However, in general, it might be difficult to sample from the conditional distribution of Y_1 given y_2 and vice versa. A sequentially updating Metropolis-Hastings algorithm does not suffer from this problem. Scaling of the proposal along the major axis could be increased relative to that along the minor axis and the target would be explored efficiently.

Therefore one approach to increasing the mixing effiency of a partial-blocking algorithm (such as the RWM) is to make the correlation of the new parameters (with respect to the target distribution) as close as possible to zero. A similar approach has the parameters as eigenvectors of the Hessian of the target at the target's mode (if it has a single mode). Note that (at a mode) these are also the eigenvectors of the Hessian of the log of the target distribution. These eigenvectors are then approximately orthogonal in the main mass of the target. In the special case of a Gaussian target these two strategies (aiming for zero correlation or eigenvectors of the Hessian) are in fact equivalent. In Chapter 2, good mixing of the standard MMPP parameterisation in certain situations is related to approximate orthogonality at the mode.

The different curvature or scales of variation along the principal axes may be thought of as corresponding to different amounts of information about the parameters corresponding to each of the axes. Conversely the effects on the target of a unit jump from the mode along principal axes with two different curvatures will be very different. In Chapter 2, a reparameterisation of the MMPP motivated by different scales of variation in the likelihood leads to an approximate orthogonality close to or at the posterior mode in some situations. Note that the above strategies are motivated by heuristic consideration of the Gaussian-like distribution in Figure 1.4. It is easy to produce counter examples wherein, for example, under the standard parameterisation, the correlation with respect to the target is zero, and yet the chain is in fact reducible. Alternative strategies to increase mixing efficiency are considered in Chapter 6 of Gilks et al. (1996) and in Papaspiliopoulos et al. (2003).

1.3.4 Other aspects of MCMC

MCMC is most commonly used in Bayesian statistics, where it explores the posterior distribution of model parameters. Several aspects of such exploration specific to Chapter 2 are now discussed.

1.3.4.1 Extending the state space

Some problems involve hidden data models or missing data; in such cases analytical expressions for the posterior conditional distributions may only be straightforward to write down in terms of the parameters, the observed data and some other *unknown* data. In such circumstances it is common to extend the statespace of the Markov chain to include the unknown data. The chain then explores the joint

distribution of the parameters and the unknown (or hidden) data. If the hidden data is not of interest then once stationarity and good mixing have been confirmed it may be ignored in any subsequent analysis. The Gibbs sampler in Chapter 2 extends the state space in just this manner, although in its application to occurences of a DNA motif the hidden data is also analysed as it is of interest in its own right.

1.3.4.2 Label-switching

Certain likelihoods, such as those for mixture distributions and for the Markov Modulated Poisson Process, are invariant to relabelling of the states. With d states there are $d!$ identical modes and there is an innate unidentifiability of the parameters. If the joint prior distribution on the parameters is similarly invariant then so is the posterior. For inference using MCMC this is apparent in the phenomenon of *label-switching* wherein the (MCMC) chain passes from one mode to another, which is equivalent to a permutation of the states. Figure 1.5 shows trace and density plots for the first 10 000 iterations of the Gibbs sampler of Chapter 2 applied to a simulated data set with two states (replicate 1 of S4: 100 seconds of an MMPP simulated using parameters $q_{12} = q_{21} = 1$ and $\boldsymbol{\lambda} = (10, 13)^t$). Frequent label switching is apparent from the sharp jumps in the trace plots and multimodality of the kernel density estimates. Label-switching effects are not visible in the plots for q_{12} and q_{21} as the corresponding true parameter values are equal.

Even if the priors are not exchangeable they will still modulate the multimodal likelihood, and often produce a multi-modal posterior. Such a posterior is illustrated in Figure 2.8. Because of the difference between the priors, the mass in the second mode is less than a hundredth of that in the main mode. To remove the

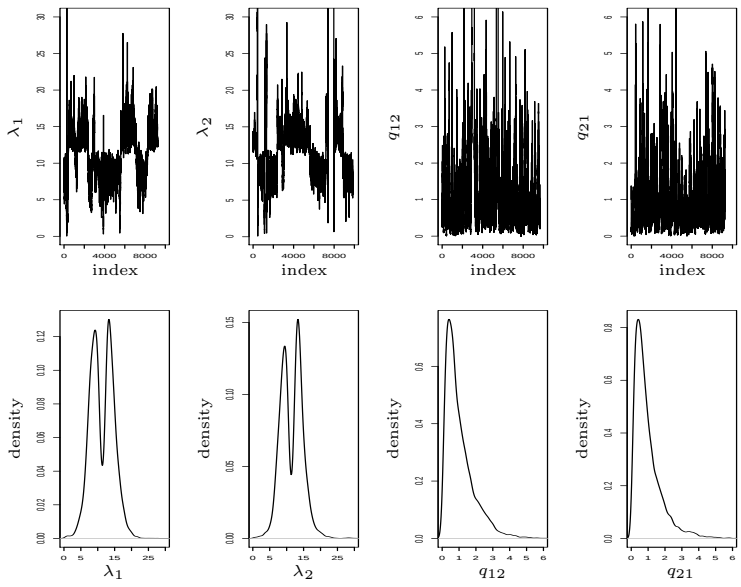

Figure 1.5: Trace and kernel density plots for the first 10 000 iterations of the Gibbs sampler of Chapter 2 on a simulated data set (replicate 1 of S4) with two states.

second mode completely it would be necessary to set a joint prior for the two states that (for example) forces an ordering on the states.

1.4 Diffusions and efficiency

1.4.1 Diffusions

Diffusion ideas are not employed in the main body of this thesis. However the summary of earlier work on optimal-scaling presented in Section 3.1.1 and the discussion of expected square jump distance in Section 3.1.4 both require a limited understanding of one-dimensional Brownian motion and stochastic differential equations (SDE's), and in particular the notion of the speed of a diffusion. A very simplistic and intuitive explanation follows; for more detail see for example Øksendal (1998).

The canonical 1-dimensional Brownian motion X_t is a stochastic process for which changes over disjoint time intervals are independent and satisfy

$$\Delta X_t \ \sim \ N(0, \Delta t)$$

where $\Delta X_t := X_{t+\Delta t} - X_t$. The relationship in the limit as $\Delta t \to 0$ corresponds to the SDE

$$dX_t = dB_t$$

The diffusion $W_t = kX_t$ corresponds to Brownian motion increments of size $dW_t = k\ dB_t$ and satisfies $\Delta X_t \ \sim \ N(0, k^2\Delta t)$. This (intuitively) establishes the connection between the coefficient of dB_t and the variance term of the normal increment.

We return to the Brownian motion X_t and add a deterministic drift $\mu(X_t)$. Increments over small time intervals now approximately satisfy

$$\Delta X_t \ \sim \ N(\mu(X_t)\Delta t, \ \Delta t)$$

and correspond to the SDE

$$dX_t = \mu(X_t) \ dt + dB_t$$

Finally consider the change in X_t over time interval $n\Delta t$ (with $n\Delta t$ small and n an integer). Small increments are still approximately independent, so for small Δt

$$\Delta X_t \ \sim \ N(n\mu(X_t) \ \Delta t, \ n\Delta t)$$

Conceptually this is the same as would have been observed over interval Δt if the diffusion had been speeded up by a factor n. In general therefore for a diffusion with constant speed h we have (approximately)

$$\Delta X_t \ \sim \ N(h\mu(X_t) \ \Delta t, \ h\Delta t) \tag{1.11}$$

with equality in the limit as $\Delta t \to 0$. This corresponds to the stochastic differential equation

$$dX_t = h \ \mu(X_t) \ dt + h^{1/2} dB_t \tag{1.12}$$

If the drift term satisfies

$$\mu(\cdot) = -\frac{1}{2}\nabla \log \pi(\cdot)$$

then the diffusion has stationary distribution $\pi(\cdot)$ and is known as a *Langevin* diffusion.

1.4.2 Autocorrelation for diffusions

The autocorrelation at time t_0 and lag-t for some function $f(\cdot)$ of a Langevin diffusion with speed h is defined as

$$\rho_{h,f}(t;t_0) := \mathrm{Corr}\left[f\left(X_{t_0}\right), f\left(X_{t_0+t}\right)\right]$$

We will assume X_t is stationary so that the correlation is independent of t_0, and drop t_0 from the notation. By the same argument as at the end of the previous section we may define a new diffusion $\tilde{X}_t$ which is stochastically identical to X_t but with speed 1: $\tilde{X}_{ht} := X_t$. Therefore

$$\rho_{h,f}(t) = \rho_{1,f}(ht)$$

For continuous process X_t the integrated autocorrelation time is

$$I_h(t) = \int_0^\infty dt\ \rho_{h,f}(t) = \int_0^\infty dt\ \rho_{1,f}(ht) = \frac{1}{h}\int_0^\infty dt\ \rho_{1,f}(t) \tag{1.13}$$

Thus minimising the integrated ACT is equivalent to maximising the speed of the diffusion.

1.5 Beta functions and hyperspheres

In Chapter 3 we are concerned with spherically symmetric random variables of general dimension d, for which marginal one-dimensional and marginal radial distributions will turn out to be related through beta random variables. This section gives an overview of basic results on the beta function and beta random variables that will be required. It also introduces the formula for the (d-1)-dimensional 'surface area' of a hypersphere in $\Re^d$, which is used in Chapter 3.

1.5.1 Beta functions and Beta random variates

For $a > 0$ and $b > 0$, the (complete) Beta function is

$$\begin{aligned} B(a,b) &= \int_0^1 dz\ z^{a-1}(1-z)^{b-1} \\ \text{(equivalently)} &= 2\int_0^{\pi/2} d\psi\ \sin^{2a-1}\psi\cos^{2b-1}\psi \\ &= \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)} \end{aligned}$$

In Chapter 3 we will require an asymptotic appromation to $B\left(\frac{1}{2}, \frac{d-1}{2}\right)$, which we obtain now via Stirling's approximation:

$$\frac{\Gamma(b+1)}{(2\pi)^{1/2}b^{b+1/2}e^{-b}} \to 1$$

From this

$$d^{1/2}\frac{\Gamma\left(\frac{d-1}{2}\right)}{\Gamma\left(\frac{d}{2}\right)} \to 2^{1/2}$$

and therefore

$$d^{1/2}B\left(\frac{1}{2}, \frac{d-1}{2}\right) \to (2\pi)^{1/2} \tag{1.14}$$

A $Beta(a,b)$ random variate has density function $f(x) = \frac{1}{B(a,b)}x^{a-1}(1-x)^{b-1}$

1.5.2 The surface area of a hypershpere

The (d-1)-dimesnional 'area' of a hypershpere of radius r in $\Re^d$ is $a_d r^{d-1}$ where $a_d = \frac{2\pi^{d/2}}{\Gamma(\frac{d}{2})}$. The first few terms are: $a_1 = 2, a_2 = 2\pi, a_3 = 4\pi$

The formula arises from the equality

$$(2\pi)^{d/2} = \int_{\Re^d} d\mathbf{x}\ e^{-\frac{1}{2}\left(x_1^2+\cdots+x_d^2\right)} = \int_0^\infty dr\ a_d r^{d-1} e^{-\frac{1}{2}r^2} = a_d \Gamma(d/2) 2^{d/2-1}$$

We also note that

$$a_d/a_{d-1} = \frac{\pi^{1/2}\Gamma(\frac{d-1}{2})}{\Gamma(\frac{d}{2})} = B\left(\frac{1}{2}, \frac{d-1}{2}\right) \tag{1.15}$$

Chapter 2

Bayesian analysis of the Markov modulated Poisson process

2.1 Introduction

A Markov Modulated Poisson Process (MMPP) is a Poisson process whose intensity depends on the current state of an independently evolving continuous time Markov chain. Points from the MMPP are often referred to as *the observed data* and the underlying Markov chain as *the hidden data*.

MMPP's are used in modelling a variety of phenomena, for example:

- The arrivals of photons from single molecule fluorescence experiments (Burzykowski et al., 2003; Kou et al., 2005). Here the arrival rate of photons at a receptor is modulated by the state of a molecule which (in the simplest model formulation) alternates between its ground state and an excited state.
- Frequency of bank transactions where a customer's bank details have been

obtained by a criminal and there are intermittent "contamination" periods where both the customer and the criminal are accessing the account (Scott, 1999).

- Wet deposition of a radionuclide emitted from a point source (Davison and Ramesh, 1996).
- Requests for web pages from users of the World Wide Web (Scott and Smyth, 2003); these show bursts of activity followed by periods of almost no activity.

Fischer and Meier-Hellstern (1992) note many further examples of the use of MMPP's for modelling overflow in telecommunications networks and in modelling packetised voice and data streams.

In some applications the exact timings of all observed events are known and in others data are accumulated over fixed intervals. In the latter situation the observed data often appear as either a count of the number of events in each interval or a binary indication for each interval as to whether there were no events or at least one event.

MMPP parameters can be fitted to data by matching certain theoretical moments to those observed (see Fischer and Meier-Hellstern (1992) and references therein). However, it is possible to calculate the likelihood of arrival data for an MMPP (for example Asmussen (2000); see also Section 2.4). Ryden (1996) summarises several likelihood approaches.

Here we consider Bayesian analysis and focus on exploring the posterior distribution via Markov chain Monte Carlo (MCMC). Metropolis-Hastings algorithms

(e.g. Gilks et al., 1996) provide a standard mechanism for Bayesian inference about parameters when the likelihood is computable. This approach is employed for example by Kou et al. (2005). Alternatively an approximate Gibbs sampler has been developed in Scott (1999) and Scott and Smyth (2003). This Gibbs sampler is only applicable to event-time data, and restricts the possible transitions of the underlying Markov chain. The approximation is based on requiring that certain transitions of the underlying chain only occur at event-times (see Section 2.5 for further details). For the examples considered in Scott (1999) and Scott and Smyth (2003) this approximate Gibbs sampler is very efficient.

The key presentation of this chapter is an exact Gibbs sampler which alternately samples from the *true* conditional distribution of the hidden chain given the parameters and the data and then the conditional distribution of the parameters given the hidden chain and the data (Section 2.5). As background to this (Section 2.2) we detail the forward-backward algorithm (Baum et al., 1970). We then exhibit (Section 2.3) a generic algorithm for sampling from the exact distribution of a continuous time Markov chain over a known interval given the start and end states. This is an extension of the technique developed in Fearnhead and Meligkotsidou (2004) and is key to the construction of our Gibbs sampler.

Section 2.4 reviews the derivation of the likelihood function for MMPP's for the data-formats mentioned above (exact timings, and either interval counts or binary indicators). The likelihood function is necessary for Metropolis-Hastings inference, and the extended state spaces introduced in its derivation are fundamental to the construction of our Gibbs sampler.

Section 2.5 details the Gibbs sampler itself for all three data formats. Priors are discussed, as is the use of the Gibbs sampler for model choice. Finally, limiting forms of the observed information matrix are derived and used to assess when the Gibbs sampler is likely to be most efficient and to suggest possible reparameterisations for Metropolis-Hastings schemes. There then follows (Section 2.6) a detailed comparison between the Gibbs sampler and various Metropolis-Hastings random walk algorithms for simulated event time data on two-dimensional MMPP's. Here we find that the Gibbs sampler is more efficient than many of the Metropolis-Hastings algorithms tested, sometimes by an order of magnitude. In Section 2.7 we apply the Gibbs sampler to choose between one, two, and three dimensional models to explain variations in the frequency of occurence of the Chi site in E.coli DNA. One advantage of the the Gibbs sampler is that it allows us to sample from the exact conditional distribution of the underlying Markov chain given the data, and this allows us to identify regions of high and low Chi site intensity. The chapter concludes in Section 2.8 with a discussion.

2.2 The forward-backward algorithm

The forward-backward algorithm (Baum et al., 1970) applies to any discretely observed Hidden Markov Model (HMM) and allows sampling of the state of the hidden chain at the observation times given the states at the start and end of the observation window. The algorithm is easily extended to the case where there is a prior distribution on the initial state and no knowledge of the end state of the chain.

We first describe a general HMM. Let an unobserved (discrete or continuous time)

Markov chain evolve over a d-dimensional state space. We observe a second process over a window $[0, t_{obs}]$ at specific times $t'_1, \ldots, t'_n$. Suppose that the value of the observed process at time t'_k is d_k, and define $\mathbf{d} := (d_1, \ldots, d_n)^t$. For notational convenience define $t'_0 = 0$, $t'_{n+1} = t_{obs}$ and $\mathbf{t}' = (t'_0, \ldots, t'_{n+1})$. Also write s_k for the state of the unobserved Markov chain at time t'_k. The likelihood of the observed process depends on the state of the hidden process via a likelihood vector $\mathbf{l}^{(k)}$ with $k = 1, \ldots, n$ where $l_i^{(k)} := P(d_k | S_k = i)$. From this define a likelihood matrix $\mathbf{L}^{(k)} = \text{diag}(\mathbf{l}^{(k)})$.

Let $\mathbf{T}^{(k)}$ be the k^{th} transition matrix of the Markov chain (i.e. $T_{ij}^{(k)}$ is the probability that the unobserved process is in state j just before t'_k given that it is in state i at t'_{k-1}).

We define probability matrices

$$\begin{aligned} A_{s,s_{n+1}}^{(n+1)} &= P(s_{n+1} | s_n = s) \\ A_{s,s_{n+1}}^{(k)} &= P(d_k, \ldots, d_n, s_{n+1} | s_{k-1} = s) \quad (0 < k \leq n) \end{aligned}$$

And note that

$$P(d_k, \ldots, d_n, s_{n+1} | s_{k-1}) = \sum_{s_k=1}^{d} P(s_k | s_{k-1}) P(d_k | s_k) P(d_{k+1}, \ldots, d_n, s_{n+1} | s_k)$$

Therefore the matrices may be calculated via a backwards recursion

$$\begin{aligned} \mathbf{A}^{(n+1)} &= \mathbf{T}^{(n+1)} \\ \mathbf{A}^{(k)} &= \mathbf{T}^{(k)} \mathbf{L}^{(k)} \mathbf{A}^{(k+1)} \quad (0 < k \leq n) \end{aligned}$$

These matrices accumulate information about the chain through the data. The final accumulation step creates $\mathbf{A}^{(0)}$, where $A_{s_0,s_{n+1}}^{(0)} = P(\mathbf{d}, s_{n+1} | s_0)$ is proportional

to the likelihood of the observed data given the start and end states.

Using the Markov property we therefore have

$$\begin{aligned} P(S_k = s|\ \mathbf{d}, s_{k-1}, s_{n+1}) &= P(S_k = s|\ d_k, \ldots, d_n, s_{k-1}, s_{n+1}) \\ &= \frac{T^{(k)}_{s_{k-1},s}\ l^{(k)}_s A^{(k+1)}_{s,s_{n+1}}}{A^{(k)}_{s_{k-1},s_{n+1}}} \end{aligned} \tag{2.1}$$

Using (2.1) we may proceed forwards through the observation times $t'_1, \ldots t'_n$, simulating the state at each observation point in turn. This algorithm is often presented in the equivalent formulation of a forwards accumulation of information and a backwards simulation step through the observation times.

If the start and end states of the chain are unknown, but a prior distribution $\boldsymbol{\mu}$ on the state of the hidden process is provided then with a slight adjustment to the algorithm we may simulate the states at the start and end times of the chain as well as at the observation times.

The start state is simulated from

$$P(S_0 = s|\mathbf{d}) = \frac{\mu_s\ [\mathbf{A}^{(1)}\mathbf{1}]_s}{\boldsymbol{\mu}^t\ \mathbf{A}^{(1)}\mathbf{1}} \tag{2.2}$$

where $\mathbf{1}$ is the d-dimensional vector of ones.

The state s_k $(k > 0)$ is then simulated from

$$P(S_k = s|\mathbf{d}, s_{k-1}) = \frac{T^{(k)}_{s_{k-1},s}\ l^{(k)}_s\ [\mathbf{A}^{(k+1)}\mathbf{1}]_s}{[\mathbf{A}^{(k)}\mathbf{1}]_{s_{k-1}}} \tag{2.3}$$

The observation times in a Markov Modulated Poisson Process correspond to actual events from the observed Poisson process. Therefore not only do the observations

contain information about the state of the hidden chain, but so do the intervals between observations, since these contain no events. In Section 2.4 we derive likelihoods through accumulation steps modified to take this into account. In a similar way we can use the forward-backward algorithm to simulate the state of the hidden chain at observation times for the first stage of our Gibbs sampler. The second stage of the Gibbs sampler simulates a realisation from the exact distribution of the full underlying Markov chain conditional on the data. This is more challenging and relies on a technique for simulating a realisation from a continuous time Markov chain over an interval given the start and end states.

2.3 Simulating a continuous time Markov chain over an interval given the start and end states

Let continuous time Markov chain W_t have generator matrix $\mathbf{G}$, and let it start the interval $[0, t]$ in state s_0 and finish in state s_t. We describe a method for simulating from the exact conditional distribution of the chain given the start and end states.

The behaviour of W_t on entering state i until leaving that state can be thought of in terms of a Poisson process of rate $\rho_i := -g_{ii}$ and a set of transition probabilities

$$\begin{aligned} p_{ij} &= g_{ij}/\rho_i \quad (i \neq j) \\ &= 0 \qquad\quad (i = j) \end{aligned}$$

The Poisson process is started as soon as the chain enters state i; at the first event from the process the chain changes to a state j determined at random using the transition probabilities for state i. A new Poisson process is then initiated with

intensity corresponding to the new state.

An alternative formulation uses events from a single dominating Poisson process U_t to determine when transitions may occur; crucially the intensity ρ of the Poisson process is independent of the chain state. We call events in this dominating Poisson process "U-events". Probabilities for the various state changes are presented in the form of a transition matrix $\mathbf{M}$.

The intensity of the dominating process must necessarily be at least as high as the highest (in modulus) diagonal element of $\mathbf{G}$. With $\rho = \max \rho_i$ the stochastic transition matrix for the discrete time sequence of states at "U-events" is

$$\mathbf{M} := \frac{1}{\rho}\mathbf{G} + \mathbf{I}$$

For any state i with $\rho_i < \rho$, $\mathbf{M}$ specifies a non-zero probability of no change in the state, so that the rate of events that change the state is ρ_i. Considering an interval of length t straightforward expansion of the transition matrix for the interval gives

$$e^{\mathbf{G}t} = e^{-\rho \mathbf{I} t} e^{\rho \mathbf{M} t} = \sum_{r=0}^{\infty} e^{-\rho t} \frac{(\rho t)^r}{r!} \mathbf{M}^r \tag{2.4}$$

The $(i,j)^{th}$ element of the left hand side is $P(W_t = j | W_0 = i)$. If we define $N_U(t)$ as the number of U-events over the interval of length t then the $(i,j)^{th}$ element on right hand side can be interpreted as

$$\sum_{r=0}^{\infty} P(N_U(t) = r)\ P(W_t = j | W_0 = i, N_U(t) = r)$$

Thus conditional on start and end states s_0 and s_t, the distribution of the number of dominating U-events is given by

$$P(N_U(t) = r) = \frac{e^{-\rho t} \frac{(\rho t)^r}{r!} [\mathbf{M}^r]_{s_0, s_t}}{[e^{\mathbf{G}t}]_{s_0, s_t}} \tag{2.5}$$

We have used a single dominating Poisson process with fixed intensity independent of the chain state. Therefore conditional on the number of dominating events, the positions of these events and the state changes that occur at the events are independent of each other and may be simulated separately. Furthermore, since U_t is a simple Poisson process the U-events are distributed uniformly over the interval $[0, t]$.

Suppose that r dominating U-events are simulated at times $t_1^*, \ldots, t_r^*$, and let these correspond to (possible) changes of state of W_t to $s_1^*, \ldots, s_r^*$. For convenience we define $t_0^* := 0$ and $s_0^* := s_0$.

Now

$$P(W_t = s_t | W_0 = s_0) = [\mathbf{M}^r]_{s_0, s_t}$$

The start and end state for each interval are assumed known, and so we employ the forward-backward algorithm of Section 2.2 with $\mathbf{L}^{(k)} = \mathbf{I}$ and $\mathbf{T}^{(k)} = \mathbf{M}$ to simulate the state change at each U-event

$$P(W_{t_j^*} = s | W_{t_{j-1}^*} = s_{j-1}^*, W_t = s_t) = \frac{[\mathbf{M}]_{s_{j-1}^*, s} [\mathbf{M}^{r-j}]_{s, s_t}}{[\mathbf{M}^{r-j+1}]_{s_{j-1}^*, s_t}} \quad (j = 1, \ldots, r) \qquad (2.6)$$

Our algorithm then becomes

(i) Simulate the number of dominating events using (2.5).

(ii) Simulate the position of each dominating event from a uniform distribution over the interval [0,t].

(iii) Simulate the state changes at the dominating events using (2.6).

2.4 Likelihood for MMPP's

We now focus exclusively on MMPP's. Let a (hidden) continuous-time Markov chain X_t on state space $\{1, \ldots, d\}$ have generator matrix $\mathbf{Q}$ and stationary distribution $\boldsymbol{\nu}$.

An MMPP is a Poisson process Y_t whose intensity is λ_i when $X_t = i$, but in all other ways is evolving independently of X_t. We write $\boldsymbol{\lambda} := (\lambda_1, \ldots, \lambda_d)^t$ and $\boldsymbol{\Lambda} := \text{diag}(\boldsymbol{\lambda})$.

We are interested in Bayesian inference about $\boldsymbol{\lambda}, \mathbf{Q}$, and X_t. Here we review derivations of likelihood for the three different data types mentioned in the introduction. Likelihoods are required for inference about $\boldsymbol{\lambda}$ and $\mathbf{Q}$ using Metropolis-Hastings schemes (see Section 2.5). The accumulation steps and extended state spaces used here are essential also for our Gibbs sampler which allows inference for $\boldsymbol{\lambda}, \mathbf{Q}$, and X_t is detailed in section 2.5.1.

The Y process is (fully or partially) observed over an interval $[0, t_{obs}]$ with t_{obs} known and fixed in advance. We employ the symbol $\mathbf{1}$ for the matrix or (horizontal or vertical) vector all of whose elements are one, and similarly $\mathbf{0}$ is a matrix or vector all of whose elements are zero.

We are interested in inference for three commonly encountered data formats

D1 Exact times are recorded for each of the n observed events (see Kou et al. (2005), and Scott and Smyth (2003) for example uses of this data format).

D2 A fixed series of $n+1$ contiguous accumulation intervals of length t_i is used, and associated with the i^{th} interval is a binary indicator b_i which is zero if there are no Y-events over the interval and one otherwise (see for example Davison and Ramesh, 1996).

D3 A fixed series of $n+1$ contiguous accumulation intervals of length t_i is used, and associated with the i^{th} interval is a count c_i of the number of Y-events over the interval (see for example Burzykowski et al., 2003).

In each case it is possible to derive the likelihood function. We summarise the three derivations; for more details see Asmussen (2000), Davison and Ramesh (1996), and Burzykowski et al. (2003) respectively.

2.4.1 Likelihood for event-time data

We first consider the data format D1. We write $N_Y(t)$ for the number of Y-events in the interval $[0,t]$, so that $N_Y(0) = 0$ and $N_Y(t_{obs}) = n$, the total number of events. For notational convenience we set $t'_0 = 0$, $t'_{n+1} = t_{obs}$ and let $t'_1, \ldots, t'_n$ be the event times for the n events. Define $t_k = t'_k - t'_{k-1}, k = 1, \ldots, n+1$; these are respectively the time from the start of the observation period to the first event, the inter-event times, and the time from the last event until the end of the observation period. We define $\mathbf{t} := (t_1, \ldots, t_{n+1})^t$.

We first derive a form for

$$P_{ij}^{(0)}(t) := P(\text{there are no } Y \text{ events in } (0,t) \text{ and } X_t = j \mid X_0 = i)$$

We define a meta-Markov process W_t on an extended state space $\{1, \ldots, d, 1^*\}$, and let W_t combine X_t and Y_t as follows: W_t matches X_t exactly up until just

before the first Y event. At the first such event W moves to the absorbing state 1^*. So if the first Y-event occurs at time t'

$$\begin{aligned} \text{for } t < t',\ W_t &= X_t \\ \text{for } t \geq t',\ W_t &= 1^* \end{aligned}$$

The generator matrix for W_t is

$$\mathbf{G}_w = \begin{bmatrix} \mathbf{Q} - \mathbf{\Lambda} & \boldsymbol{\lambda} \\ \mathbf{0} & 0 \end{bmatrix} \tag{2.7}$$

So the transition matrix at time t is

$$e^{\mathbf{G}_w t} = \begin{bmatrix} e^{(\mathbf{Q}-\mathbf{\Lambda})t} & (\mathbf{Q}-\mathbf{\Lambda})^{-1}(e^{(\mathbf{Q}-\mathbf{\Lambda})t} - \mathbf{I})\boldsymbol{\lambda} \\ \mathbf{0} & 1 \end{bmatrix} \tag{2.8}$$

From the definition of W we see that

$$P_{ij}^{(0)}(t) = [e^{(\mathbf{Q}-\mathbf{\Lambda})t}]_{ij} \tag{2.9}$$

So the likelihood of the observed data, and that the chain ends in state j given that it starts in state i is the $(i,j)^{th}$ element of

$$e^{(\mathbf{Q}-\mathbf{\Lambda})t_1}\mathbf{\Lambda}e^{(\mathbf{Q}-\mathbf{\Lambda})t_2}\mathbf{\Lambda}\ldots\mathbf{\Lambda}e^{(\mathbf{Q}-\mathbf{\Lambda})t_{n+1}}$$

This is the $\mathbf{A}^{(0)}$ matrix of the forward-backward algorithm as described in Section 2.2. Assuming that the chain starts in its stationary distribution, the likelihood of the observed data is therefore

$$L(\mathbf{Q}, \mathbf{\Lambda}, \mathbf{t}) = \boldsymbol{\nu}^t e^{(\mathbf{Q}-\mathbf{\Lambda})t_1}\mathbf{\Lambda}\ldots e^{(\mathbf{Q}-\mathbf{\Lambda})t_n}\mathbf{\Lambda}e^{(\mathbf{Q}-\mathbf{\Lambda})t_{n+1}}\mathbf{1} \tag{2.10}$$

2.4.2 Likelihood for accumulation interval formats

We now consider data formats D2 and D3 and for simplicity assume all the interval lengths to be equal ($t_i = t^*\ , i = 1, \ldots, n+1$). Extension to the more general case is straightforward.

Define

$$P_{ij}^{(s)} = P(\text{there are } s \text{ } Y\text{-events over } (0, t^*) \text{ and } X_{t^*} = j | X_0 = i)]$$

and

$$\overline{P}_{ij} = P(\text{there is at least one } Y\text{-event over } (0, t^*) \text{ and } X_{t^*} = j | X_0 = i)]$$

With b_i as the binary indicator for at least one event in the i^{th} interval, the likelihood for D2 is therefore

$$\boldsymbol{\nu}^t \left(\prod_{i=1}^{n+1} \mathbf{P}^{(0)^{1-b_i}} \, \overline{\mathbf{P}}^{\, b_i} \right) \mathbf{1}$$

and with count c_i of the number of events for each interval the likelihood for D3 is

$$\boldsymbol{\nu}^t \left(\prod_{i=1}^{n+1} \mathbf{P}^{(c_i)} \right) \mathbf{1}$$

$\mathbf{P}^{(0)}$ is given by (2.9) and so it remains to calculate the matrices $\mathbf{P}^{(c)}$ $(c > 0)$, and $\overline{\mathbf{P}}$.

Since the probability of finishing interval $(0, t^*)$ in state j given starting state i is the $(i, j)^{th}$ element of $e^{\mathbf{Q}t^*}$, we see that

$$\overline{\mathbf{P}} = e^{\mathbf{Q}t^*} - e^{(\mathbf{Q} - \boldsymbol{\Lambda})t^*}$$

For format D3 define $c_{max} = \max\ c_i$ and create a new meta-process V_t on state space $S = (1^{(0)}, \ldots, d^{(0)}, 1^{(1)}, \ldots d^{(1)}, \ldots, 1^{(c_{max})}, \ldots, d^{(c_{max})}, 1^*)$. If the number of

Y-events observed up until time t in the accumulation interval containing t is $N_Y^*(t)$, then for $N_Y^*(t) \leq c_{max}$ $V_t = X_t^{(N_Y^*(t))}$ and otherwise $V_t = 1^*$. For example if at time t, the hidden process is in state 3 and there have been 7 events so far in the accumulation interval containing t, then the meta-process V_t is in state $3^{(7)}$

The generator matrix for V_t is

$$\mathbf{G_v} = \begin{bmatrix} \mathbf{Q}-\mathbf{\Lambda} & \mathbf{\Lambda} & \mathbf{0} & \dots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{Q}-\mathbf{\Lambda} & \mathbf{\Lambda} & \dots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{Q}-\mathbf{\Lambda} & \dots & \mathbf{0} & \mathbf{0} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \dots & \mathbf{Q}-\mathbf{\Lambda} & \boldsymbol{\lambda} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} & 0 \end{bmatrix} \tag{2.11}$$

and the block of square matrices comprising of the top d rows of $e^{\mathbf{G}_v t^*}$ gives the $(d \times d)$ conditional transition matrices $\mathbf{P}^{(r)}$.

2.5 Bayesian approach

We are interested in Bayesian analysis of MMPP's. In the Bayesian framework, beliefs about the parameters before examining the data are collected into a prior distribution on the parameter vector. This is then modified by the data to produce a posterior parameter distribution.

Metropolis-Hasting algorithms (see Section 1.3.1.1) provide one possible Bayesian approach to inference for parameters of an MMPP, and this approach is adopted by Kou et al. (2005) for example. The prior is multiplied by the likelihood for

a specific parameter vector to obtain the posterior distribution up to a constant of proportionality. However all Metropolis-Hastings algorithms need to be tuned, and this can be time consuming.

Scott (1999) describes an approximate Gibbs sampler for event-time observed data on two-state MMPP's. This was mentioned briefly in Section 2.1, and is now expanded upon. By assuming that transitions from state-1 to state-2 occur only at event times the possible behaviours of the chain over each closed inter-event interval are divided into 5 classes. These classes are specified in terms of the start and end states of the chain and whether or not there is a state change over the interval. The forward-backward algorithm is then applied to sample from the exact distribution of these classes for each interval in turn.

For each interval, given the chain-class, both the state at the end of each event interval and the amount of time spent in state 1 over the interval are conditionally independent of each other and of the values for previous inter-event intervals. These are sampled, together with the initial state, for the entire chain. With conjugate Gamma priors, the sampled data provide sufficient statistics for draws of a new set of parameter values.

Scott and Smyth (2003) generalise this process to a d-dimensional Markov Poisson cascade. Here a Markov chain modulates an ordered superposition of Poisson processes. Each Poisson process may only be active if all "lower" processes are also active. In a similar manner to the previous paper it is assumed that each activation and deactivation step is associated with an event from the "higher" process.

Several deactivations may occur simultaneously, corresponding to just one event, thus the chain may transition from a given state to any "lower" state; however the chain may only transition to a higher state via all intermediate states. The formulation again allows a draw from an underlying Markov chain, and at the same time stops any activation and subsequent deactivation with no associated event.

We aim to demonstrate a Gibbs sampler that samples from the underlying hidden chain X_t and then from the parameter vector given the underlying chain. Unlike the algorithms of Scott (1999) and Scott and Smyth (2003) our solution contains no constraint forcing certain state changes to occur at observed event times, and it allows all possible transitions between states. Our Gibbs sampler may also be applied to all three of the data formats described in Section 2.4

As a shorthand we write the state of the chain at event-times and at the start and end of the observation period as $S_i = X_{t'_i}$. The distribution of the new parameter vector depends on the underlying chain through the starting state (ν_{s_0}) and three further sufficient statistics, which we now define.

We write $\tilde{t}_i$ for the total time spent in state i by the hidden chain, r_{ij} for the number of times the chain transitions from state i to state j ($r_{ii} = 0 \; \forall i$), and n_i for the number of Y-events that occur while the chain is in state i. We correspondingly define $\tilde{\mathbf{t}} = (\tilde{t}_1, \ldots, \tilde{t}_d)^t$, $\mathbf{n} = (n_1, \ldots, n_d)^t$, and $\mathbf{R}$ as the matrix with elements r_{ij} .

2.5.1 Gibbs sampler

Our Gibbs sampler acts on augmented state-space $\{\boldsymbol{\lambda}, \mathbf{Q}, X_t\}$, and each iteration has 3 distinct stages:

1. Given the parameter values $(\boldsymbol{\lambda}, \mathbf{Q})$ use the second form of the the forward-backward algorithm, specified by (2.2) and (2.3) in Section 2.2), to simulate the state of the hidden chain X_t at the start and end of the observation interval ($t'_0 = 0$ and $t_{obs} = t'_{n+1}$) and at a set of time points $t'_1, \ldots, t'_n$. For data format D1 $t'_1, \ldots, t'_n$ correspond to event times; for formats D2 and D3 $t'_1, \ldots, t'_{n+1}$ are the end-points of accumulation intervals.

2. Given the parameter values and the finite set of states produced in stage 1, apply the technique of Section 2.3 to each interval in turn to simulate the full underlying hidden chain X_t from it's exact conditional distribution.

3. Simulate a new set of parameter values.

We now describe how each of the stages may be implemented for each of the three data formats.

Data format D1

For *stage 1* we apply the forward-backward algorithm of section 2.2 modified to take account of the fact that observation times $t'_1, \ldots, t'_n$ correspond exactly to events of the observed process and that therefore there are no Y-events between observation times. For the k^{th} interval, which has width $t_k = t'_k - t'_{k-1}$, the transition matrix is $\mathbf{T}^{(k)} = e^{(\mathbf{Q}-\boldsymbol{\Lambda})t_k}$, and the likelihood vector for the k^{th} observation

point is $\mathbf{l}^{(k)} = \boldsymbol{\lambda}$.

This process is exactly equivalent to straightforward application of the second form of the forward-backward algorithm to the meta-process W_t of section 2.4.1 on the extended state space $\{1, \ldots, d, 1^*\}$, but replacing the d-dimensional vector $\mathbf{1}$ with the $d+1$-dimensional vector $(1, \ldots, 1, 0)^t$. For the k^{th} interval, the transition matrix is now $\mathbf{T}^{(k)} = e^{\mathbf{G}_w t_k}$, where $\mathbf{G}_w$ is defined in (2.7) and $e^{\mathbf{G}_w t}$ is given explicitly in (2.8). The likelihood vector is $\mathbf{l}^{(k)} = (\boldsymbol{\lambda}, 0)^t$.

Stage 2 applies the technique of Section 2.3 directly to extended state space $\{1, \ldots, d, 1^*\}$ with generator matrix $\mathbf{G}_w$.

Figure 2.1 shows the first two stages for data format **D1**.

Stage 3 is especially simple if conjugate gamma priors are used for the parameters since the likelihood for the full data (observed data and complete underlying Markov chain) is

$$L(x_t, \mathbf{t}|\mathbf{Q}, \boldsymbol{\lambda}) \propto \nu_{s_0} \times \prod_{i=1}^{d} \prod_{j \neq i} \left(q_{ij}^{r_{ij}} e^{-q_{ij}\tilde{t}_i} \right) \times \prod_{i=1}^{d} \lambda_{s_i}^{n_i} e^{-\lambda_i \tilde{t}_i} \tag{2.12}$$

Thus independent priors $\lambda_i \sim Gam(\alpha_i, \beta_i)$ produce independent posteriors

$$\lambda_i \sim Gam(\alpha_i + n_i, \beta_i + \tilde{t}_i) \tag{2.13}$$

Were it not for the factor ν_{s_0}, which is itself a function of $\mathbf{Q}$, choosing independent priors $q_{ij} \sim Gam(\gamma_{ij}, \delta_{ij})$ $(j \neq i)$ would lead to independent posteriors

$$q_{ij} \sim Gam(\gamma_{ij} + r_{ij}, \delta_{ij} + \tilde{t}_i) \tag{2.14}$$

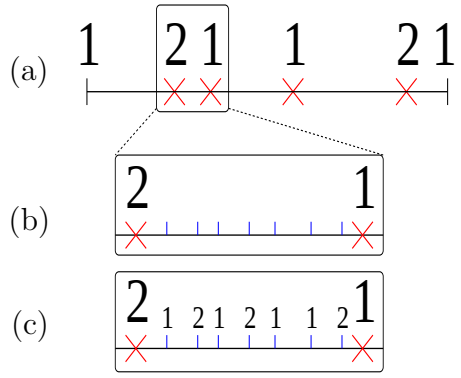

Figure 2.1: The Gibbs sampler (a) first simulates the chain state at observation times and the start and end time; for each interval it then simulates (b) the number of dominating events and their positions, and finally (c) the state changes that may or may not occur at these dominating events. The figure applies to a two-state chain with $\lambda_2 + q_{21} > \lambda_1 + q_{12}$.

However since ν_{s_0} is bounded between 0 and 1 we may employ rejection sampling, simulating $\mathbf{Q}$ from (2.14) and accepting with probability $\nu_{s_0}(\mathbf{Q})$.

Data format D2

For *stage 1* we apply the second form of the forward-backward algorithm with likelihood vector $\mathbf{l}^{(k)} = \mathbf{1}$ and transition matrix dependent on the binary indicator (b_k) for the interval

$$\mathbf{T}^{(k)} = \mathbf{P}^{(0)\,1-b_k}\ \overline{\mathbf{P}}^{\,b_k}$$

For *stage 2* we first consider the meta-process $\overline{W}_t$ on state space $\{1, \ldots, d, 1^*, \ldots, d^*\}$ with $\overline{W}_t = X_t$ when $Y_t = 0$ and $\overline{W}_t = X_t^*$ otherwise.

This has generator matrix

$$\mathbf{G}_{\overline{\mathbf{w}}} = \begin{bmatrix} \mathbf{Q} - \mathbf{\Lambda} & \mathbf{\Lambda} \\ \mathbf{0}^* & \mathbf{Q} \end{bmatrix}$$

For a given interval suppose that we have simulated X_t starting in state s_0 and ending state s_1. On the extended state space this corresponds to starting in state s_0 and finishing in state s_1 if there have been no events over the interval, otherwise finishing in s_1^* . We simulate the underlying chain from the algorithm of section 2.3. This also supplies the time of the first event in the interval, which we use for simulating the new parameters in stage 3.

In *stage 3*, for accumulation interval i define t_{ij}^* as the amount of time that the hidden chain spends in state j between the start of the interval and either the time of the first event (if there is a first event) or the end of the interval; write $t_{\cdot j}^* = \sum_i t_{ij}^*$. Let n_j^* be the number of intervals for which the chain is in state j at the first event of the interval. Then the likelihood is

$$L(x_t, \mathbf{t}|\mathbf{Q}, \boldsymbol{\lambda}) \propto \nu_{s_0} \times \prod_{i=1}^{d} \prod_{j \neq i} \left(q_{ij}^{r_{ij}} e^{-q_{ij}\tilde{t}_i} \right) \times \prod_{j=1}^{d} \lambda_j^{n_j^*} e^{-\lambda_j t_{\cdot j}^*}$$

We then proceed as with data format D1.

Data format D3

For this data format we consider the meta-process V_t on extended state space $\{1^{(0)}, \ldots, d^{(0)}, 1^{(1)}, \ldots d^{(1)}, \ldots, 1^{(c_{max})}, \ldots, d^{(c_{max})}, 1^*\}$ as defined in section 2.4.2.

For the application of the forward-backward algorithm in *stage 1*, the transition matrices are $\mathbf{T}^{(k)} = \mathbf{P}^{(c_k)}$ and the likelihood vectors are $\mathbf{l}^{(k)} = \mathbf{1}$. For *stage 2*, in

simulating from the exact distribution of the underlying chain for an interval where the start state is s_0, the end state is s_1 and there are c_k events observed we use the generator matrix $\mathbf{G}_v$ as defined in (2.11) with start state s_0 but end state $s_1^{(c_k)}$.

The algorithm also simulates from the exact distribution of the times at which each of the c_k events occurs over the interval, therefore we may perform *stage 3* exactly as for data format D1.

2.5.2 Choice of prior

As discussed in Section 1.2.1, the likelihood of an MMPP is bounded below as certain combinations of parameters approach 0 or ∞ and care must be taken in the choice of priors. That improper priors are inappropriate may also be seen from the behaviour of our Gibbs sampler. For example, there is a non-zero probability that the Gibbs sampler will simulate a chain that never enters a particular state s. The conditional posterior distribution of λ_s is then identical to the prior.

In Section 2.5.1 we saw that the Gibbs sampler is simplest to implement when parameters have independent Gamma priors. With little prior knowledge we might be tempted to apply priors with a low ratio of mean to standard-deviation, and therefore a shape parameter less than one. However such a density function approaches infinity as the argument approaches zero, and since the likelihood is bounded below the posterior distribution for each parameter contains a (probably unintended) infinite mode at zero. The vaguest "safe" prior is therefore exponentially distributed. Note also that Gamma priors combined with a likelihood that is bounded below as some parameters tend to infinity (with others fixed) will produce posterior tails

that are heavier than Gaussian.

To place priors in context, a $Gam(\alpha_i, \beta_i)$ prior for λ_i is equivalent to having previously observed the chain in state i for β_i seconds and noted α_i Y-events. Similarly a $Gam(\gamma_{i,j}, \delta_{i,j})$ prior on $q_{i,j}$ is equivalent to having previously observed the chain in state i for $\delta_{i,j}$ seconds (in total) and noted $\gamma_{i,j}$ jumps from state i to state j.

It might seem odd to have somehow managed to observed the chain in state i for different time periods depending on the states jumped to. If, as seems more intuitive, $\delta_{i,j} = \delta_i \ \ \forall i, j$ then we may consider priors for the $q_{i,j}$ in terms of independent gamma priors for the modulus of each of the diagonal elements $\rho_i = -q_{i,i} = \sum_{j \neq i} q_{i,j}$ and a Dirichlet prior for $\mathbf{f}_i := (q_{i,1}/\rho_i, \ldots, q_{i,i-1}/\rho_i, q_{i,i+1}/\rho_i, \ldots, q_{i,d}/\rho_i)$.

$$\begin{aligned} \rho_i &\sim Gam\left(\sum_{j \neq i} \gamma_{ij} \ , \ \delta_i\right) \\ \mathbf{f}_i &\sim Dir(\gamma_{i,1}, \ldots, \gamma_{i,i-1}, \gamma_{i,i+1}, \ldots, \gamma_{i,d}) \end{aligned}$$

The parameter ρ_i governs the (exponentially distributed) time we expect X_t to remain in state i after arriving, and about which we might have a more intuitive feel. We would (for example) hope that the residence time is likely to be less than t_{obs} since if the chain remains in a single state for the entire observation period there is little justification in using an MMPP to model the data. Similarly we might expect ρ_i not to be several orders of magnitude larger than λ_i since we are then either in limiting case (3) of Section 1.2.1, or the state has no impact on the observed data and is redundant. From these restrictions it seems more intuitive to think of the prior for $\mathbf{Q}$ in terms of gamma/Dirichlet combinations.

We finally note that truncating the priors for ρ_i or λ_i does not affect their conjugacy with the full likelihood. Flat priors truncated *within set limits* such as extremes from the previous paragraph may therefore be used.

2.5.3 Model choice

2.5.3.1 Theory for model choice

We wish to compare models using posterior model probabilities

$$P_i = \frac{P(M_i|\mathbf{t})}{\sum_{j=1}^{r} P(M_j|\mathbf{t})}$$

where $M_1, \ldots, M_r$ are the models under consideration. We will assume uniform prior model probabilities, which leads to

$$P_i = \frac{P(\mathbf{t}|M_i)}{\sum_{j=1}^{r} P(\mathbf{t}|M_j)}$$

and we must therefore ascertain the probability of the observed data for each model.

For a given model we denote the prior and posterior distributions of the parameters by $\pi(\cdot)$ and $P(\cdot)$ respectively. With likelihood $f(\cdot)$

$$P(\mathbf{t}) = \frac{f(\mathbf{t}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{P(\boldsymbol{\theta}|\mathbf{t})}$$

Chib (1995) considers estimating the posterior probability of the data from the output of a Gibbs sampler which samples from the conditional distribution $p(\mathbf{x}|\mathbf{t}, \boldsymbol{\theta})$ of hidden data $\mathbf{x}$ and then from the posterior distribution of the parameters given the full data $P(\boldsymbol{\theta}|\mathbf{t}, \mathbf{x})$. With Gibbs sampler output $\{\mathbf{x}^{(g)}\}$ for $g = 1, \ldots, G$ he notes that an appropriate Monte Carlo estimator of $P(\boldsymbol{\theta}|\mathbf{t})$ at $\boldsymbol{\theta}^*$ is

$$\hat{P}(\boldsymbol{\theta}^*|\mathbf{t}) = \frac{1}{G}\sum_{g=1}^{G} P(\boldsymbol{\theta}^*|\mathbf{t}, \mathbf{x}^{(g)})$$

He therefore suggests estimating the marginal likelihood from

$$\log \hat{P}(\mathbf{t}) = \log f(\mathbf{t}|\boldsymbol{\theta}^*) + \log \pi(\boldsymbol{\theta}^*) - \log \left(\frac{1}{G} \sum_{g=1}^{G} P(\boldsymbol{\theta}^*|\mathbf{t}, \mathbf{x}^{(g)}) \right)$$

The arbitrary value $\boldsymbol{\theta}^*$ should be chosen close to the posterior mode to reduce the variance of the estimator. By running the Gibbs sampler for several competing models we may estimate the marginal probabilities for the data given each model and hence the posterior model probabilities.

2.5.3.2 Implementation of model choice

For each realisation of the hidden chain X_t that is produced by the Gibbs sampler we must calculate the posterior probability of $\boldsymbol{\theta}^*$ given the observed data and the hidden chain.

From the expansion of the full-data likelihood (2.12), with independent gamma priors, the posterior λ_i's follow independent gamma posteriors (2.13). However, when multiplied by independent gamma priors (2.12) gives the posterior distribution of $\mathbf{Q}$ as

$$P(\mathbf{Q}|x_t) \propto \nu_{s_0}(\mathbf{Q}) \times \prod_{i=1}^{d} \prod_{j \neq i} \mathrm{Gam}(q_{ij}; \gamma_{ij} + r_{ij}, \delta_{ij} + \tilde{t}_i) \tag{2.15}$$

where $\mathrm{Gam}(x; a, b)$ is the density function of a gamma random variable with shape parameter a and rate parameter b evaluated at x. The normalisation constant for this density is not known and so we estimate it by Monte-Carlo integration. For each realisation of the hidden chain we repeatedly sample $\mathbf{Q}$ from the distribution of independent gammas that results from ignoring the factor ν_{s_0} in (2.15). For each $\mathbf{Q}$ we calculate the stationary distribution $\boldsymbol{\nu}$ and we find the mean (over the samples) of the s_0^{th} components of this vector. This gives the inverse of the nor-

malisation constant.

2.5.4 The information matrix

The observed Fisher information matrix at a particular point in parameter space is the local curvature of the log-likelihood function at that point. Specifically, for parameter $\boldsymbol{\theta}$ and log-likelihood $l(\boldsymbol{\theta})$

$$V_{ij} = -\frac{\partial^2 l}{\partial\theta_i \partial\theta_j}$$

The information matrix may suggest reparameterisations for inference via Metropolis-Hastings algorithms and it also impacts on the efficiency of the Gibbs sampler. For analytical convenience, throughout this section we consider the observed information rather than its expectation. We will however be discussing observed information matrices calculated from two different likelihoods:

V* the information matrix calculated from the likelihood for the (observed) event time data.

V the information matrix calculated from the likelihood for the (observed) event time data *and* the (hidden) underlying Markov chain

In this section these are referred to respectively as the observed information and the complete (or full) information. We first obtain an intuition into variations in the efficiency of our Gibbs sampler.

2.5.4.1 Efficiency of our Gibbs sampler algorithm

Louis (1982) showed that for hidden data problems, the observed and complete

information are related as follows

$$V_{ij}^{*} = E_R[V_{ij}] - (E_R[U_i U_j] - E_R[U_i]E_R[U_j])$$

where $\mathbf{U}$ is the score function for the full data, and $E_R[.]$ denotes expectation taken over the distribution of the full data given the observed data.

Defining

$$Cov_R[U_i, U_j] := E_R[U_i U_j] - E_R[U_i]E_R[U_j]$$

We write the equation more simply as

$$V_{ij}^{*} = E_R[V_{ij}] - Cov_R[U_i, U_j] \tag{2.16}$$

Therefore, as would be expected, the more information the observed data contains about the hidden chain, the smaller the second term and the closer observed information is to the complete information.

Sahu and Roberts (1999) investigated the geometric rate of convergence to stationarity of a Gibbs sampler on a Gaussian approximation to the joint posterior distribution of the observed and missing data. They show that the geometric rate of convergence of the Gibbs sampler is equal to the maximum eigenvalue of the matrix

$$\mathbf{I} - \mathbf{V}^{*}\mathbf{V}^{-1}$$

Thus the closer the information from the observed data is to the full data information, the more efficient the algorithm. Although the data from an MMPP is not in general Gaussian, this relationship suggests that the efficiency of a Gibbs sampler increases with the amount of information about the underlying chain extractable

from the observed event data.

2.5.4.2 Reparameterisations and limiting forms for the information matrix

As discussed in Section 1.3.3, a random walk Metropolis algorithm with partial updates and parameters close to the eigenvectors of the Hessian at the log-posterior mode is likely to be tuneable so that it is more efficient than a similar algorithm with a different parameterisation. This strategy arose through consideration of possibly very different scales along each principal axis of the target near the mode. In situations where the data are much stronger than the prior we may reasonably approximate the log-posterior curvature by the information matrix; the different sizes of the principal axes then correspond to different amounts of information in the data about the parameters that correspond to these axes.

If an MMPP with stationary distribution $\boldsymbol{\nu}$ is observed over a reasonably long time window compared to the convergence time of the hidden Markov chain then components of $\boldsymbol{\nu}$ give the approximate fractions of time that the chain spends in each state. The overall average intensity of the MMPP is therefore approximately $\overline{\lambda} := \boldsymbol{\nu}^t\boldsymbol{\lambda}$. Since the overall number of events is observed, intuitively $\overline{\lambda}$ should be reasonably well determined by the data compared to any other parameter $\lambda^{\perp} := \boldsymbol{\xi}^t\boldsymbol{\lambda}$ for any $\boldsymbol{\xi}$ with $\boldsymbol{\xi}^t\boldsymbol{\nu} = 0$. This reparameterisation motivates algorithm M4 in Section 2.6.

We now describe two limiting forms for the information matrix, justifying the

"default" parameterisation $(\boldsymbol{\lambda}, \mathbf{Q})$ and another possible reparameterisation, corresponding to algorithm M5 in Section 2.6. The limiting approximations are relatively easy to derive and are suitable when the data contain either almost complete or very little information about the hidden chain.

1. **The limit of complete knowledge of the underlying chain**

   From (2.12) the log-likelihood of the full (observed and hidden) data is

   $$l(\mathbf{t},\ \text{chain} \mid \boldsymbol{\lambda}, \mathbf{Q}) = \log \nu_{s_0} + \sum_{i=1}^{d} \sum_{j \neq i} \left(r_{ij} \log q_{ij} - q_{ij} \tilde{t}_i \right) + \sum_{i=1}^{d} \left(n_i \log \lambda_i - \lambda_{s_i} \tilde{t}_i \right)$$

   Therefore the score for the $\boldsymbol{\lambda}$ parameters is

   $$u_i := \frac{\partial l}{\partial \lambda_i} = \frac{n_i}{\lambda_i} - \tilde{t}_i \tag{2.17}$$

   and the observed Fisher information matrix for the $\boldsymbol{\lambda}$ parameters for the full data is

   $$V_{ij} := -\frac{\partial^2 l}{\partial \lambda_i \partial \lambda_j} = \frac{n_i}{\lambda_i^2} \delta_{ij} \tag{2.18}$$

   Similarly the score for the $\mathbf{Q}$ parameters is

   $$\frac{\partial l}{\partial q_{ij}} = \frac{1}{\nu_{s_0}} \frac{\partial \nu_{s_0}}{\partial q_{ij}} + \frac{r_{ij}}{q_{ij}} - \tilde{t}_i \tag{2.19}$$

   and the observed information is

   $$-\frac{\partial^2 l}{\partial q_{ij} \partial q_{kl}} = -\frac{1}{\nu_{s_0}} \frac{\partial^2 \nu_{s_0}}{\partial q_{ij} \partial q_{kl}} + \frac{1}{\nu_{s_0}^2} \frac{\partial \nu_{s_0}}{\partial q_{ij}} \frac{\partial \nu_{s_0}}{\partial q_{kl}} + \frac{r_{ij}}{q_{ij}^2} \delta_{ik} \delta_{jl} \tag{2.20}$$

   Second derivatives involving components of both $\boldsymbol{\lambda}$ and $\mathbf{Q}$ vanish.

Equation 2.16 combined with (2.17)-(2.20) would give a form for the information matrix for the observed data. For simplicity we just write down the portion of the information matrix corresponding to components of $\boldsymbol{\lambda}$. For all $i = 1, \ldots, d$ and $j = 1, \ldots, d$:

$$V_{ij}^{*} = E_R\left[\frac{N_i}{\lambda_i^2}\right]\delta_{ij} - Cov_R\left[\frac{N_i}{\lambda_i} - \tilde{T}_i, \frac{N_j}{\lambda_j} - \tilde{T}_j\right] \tag{2.21}$$

If the observed data contains complete information about the chain then

$$Cov_R\left[\frac{N_i}{\lambda_i} - \tilde{T}_i, \frac{N_j}{\lambda_j} - \tilde{T}_j\right] = 0$$

and so the $\boldsymbol{\Lambda}$ portion of the information matrix is diagonal

$$V_{ij}^{*} = E_R\left[\frac{N_i}{\lambda_i^2}\right]\delta_{ij} = \frac{n_i}{\lambda_i^2}\delta_{ij} \tag{2.22}$$

Non-diagonal terms in the $\mathbf{Q}$ portion of the information matrix that are not associated with shrinking covariances consist of derivatives of the stationary components. These do not increase as the observation window increases but the sufficient statistics do, therefore for large enough observation windows the full data information matrix is approximately diagonal; with the limit corresponding to the simple form of the full data likelihood. This suggests that with a large enough observation window, a tuned random walk Metropolis algorithm with partial updates using the standard parameterisation will be reasonably efficient.

2. **The limit of no knowledge of the underlying chain**

We now investigate the two-dimensional MMPP with $\lambda_1 \approx \lambda_2$. First reparameterise, setting

$$\overline{\lambda} := \nu_1\lambda_1 + \nu_2\lambda_2 \quad \text{and} \quad q := q_{12} + q_{21}$$

where $\boldsymbol{\nu}$ is the stationary distribution of the chain. Now Taylor expand the likelihood in the relative intensity difference

$$\delta := (\lambda_2 - \lambda_1)/\overline{\lambda}$$

After some algebra (see Appendix A) we obtain

$$l(\overline{\lambda}, q, \delta, \nu_1) = n\log\overline{\lambda} - \overline{\lambda}t_{obs} + 2\delta^2\nu_1\nu_2 f(\overline{\lambda}\mathbf{t}, q\mathbf{t}) + \delta^3\nu_1\nu_2(\nu_2-\nu_1)g(\overline{\lambda}\mathbf{t}, q\mathbf{t}) + O(\delta^4) \tag{2.23}$$

for some $f(\cdot,\cdot)$ and $g(\cdot,\cdot)$. At points where $\delta = 0$ (which implies that the observed data contains no information about the chain) the information matrix $\mathbf{V}^*(\overline{\lambda}, q, \delta, \nu_1)$ has a particularly simple form, with all components zero apart from

$$V^*_{\overline{\lambda},\overline{\lambda}} = \frac{n}{\overline{\lambda}^2} \quad \text{and} \quad V^*_{\delta,\delta} = 4\nu_1\nu_2 f(\overline{\lambda}\mathbf{t}, q\mathbf{t})$$

The ν_1 and q elements are zero since when $\lambda_1 = \lambda_2$ the two states are indistinguishable and there can be no information on the chain in the observed data. Further the MMPP has degenerated into a simple Poisson process, which is reflected in the information on $\overline{\lambda}$. The information matrix suggests that when there is very little indication as to the behaviour of the underlying chain a reparameterisation to $\overline{\lambda}$ and δ may be preferable. However, in (2.23) all variations of $O(\delta^2)$ are captured by the three parameters $\overline{\lambda}, q$ and

$$\alpha := 2\delta(\nu_1\nu_2)^{1/2}$$

Thus for fixed $\overline{\lambda}$ and q and small enough δ, the parameter α should have a bigger impact on the likelihood than

$$\beta := \delta(\nu_2 - \nu_1)$$

Hence there might be more information about α than about β since with this reparameterisation the log-likelihood is

$$l(\overline{\lambda}, q, \delta, \nu_1) = n \log \overline{\lambda} - \overline{\lambda} t_{obs} + \frac{1}{2}\alpha^2 f(\overline{\lambda}\mathbf{t}, q\mathbf{t}) + \frac{1}{4}\alpha^2\beta g(\overline{\lambda}\mathbf{t}, q\mathbf{t}) + O(\delta^4)$$

with both α and β being $O(\delta)$. This motivates algorithm M5 in Section 2.6.

We note that Davison and Ramesh (1996) reparameterise to $\overline{\lambda},\ q,\ \alpha^2/4,$ and $\alpha^2/4\delta^2$ for their simulation study of information loss through collecting binary response data over accumulation intervals rather than exact event times. However they cite the only reason for this reparameterisation as its invariance to label-switching.

2.5.5 Implementation of the Gibbs sampler

Gibbs sampler code used in Sections 2.6 and 2.7 was written in C, as was the code for all the Metropolis-Hastings algorithms used for comparison in Section 2.6. Matrix exponentials were calculated by truncating (2.4). The truncation was set so that the error in each element of the matrix exponential was less than a pre-determined tolerance (this was efficient as errors decay faster than geometrically, and accurate as it involves summing only positive values). The sum can be evaluated efficiently for all interval lengths by calculating and storing the required powers of $\mathbf{M}$ once for each iteration. The powers of $\mathbf{M}$ are also then used when simulating the underlying chain.

2.6 Simulation studies

We aim to test how the performance of our Gibbs sampler depends on features of the data, and to compare it with a number of random walk Metropolis-Hastings algorithms. For simplicity and because this contains the most information, we simulate data in event-time format.

Also for simplicity we confine ourselves to 2-dimensional MMPP's, and mostly to runs of $t_{obs} = 100$ seconds with $q_{12} = q_{21} = 1$. We also examine longer data sets, and data produced using asymmetric generators. For the core comparisons we perform runs on several replicate data sets.

We choose all $q_{ij} >> 1/t_{obs}$ so that the hidden chain changes state many times over the observation period, thus allowing a chance of inferring the values of the $\mathbf{Q}$ parameters (provided the visits to the states are discernible). Similarly we choose each $\lambda_i >> q_{ij}$ so that most visits to a given state will contain several observed events, making it easier to identify the separate states as well as infer $\boldsymbol{\lambda}$. Thus in our simulated data, the relative difference $\delta := (\lambda_2 - \lambda_1)/\overline{\lambda}$ provides the main guage as to ease of inference. The larger this difference the greater our knowledge of the underlying chain, the easier it is to distinguish the parameters, and consequently too, the lighter the posterior tails. Table 2.1 lists the core simulated data sets; additional simulated data sets are detailed in Appendix B.

The many options for random walk algorithms include choices between

1. Partial and complete blocking of parameters. If the former, then sequential updating, random scan, or some other updating method could be chosen.

Dataset	λ_1	λ_2	q_{12}	q_{21}	t	replicates
S1	10	90	1	1	100	3
S2	10	30	1	1	100	3
S3	10	17	1	1	100	3
S4	10	13	1	1	100	3

Table 2.1: Parameter values for the core simulated data sets.

2. Performing the random walk on the posterior of the random variable $\mathbf{Y}$, or on the posterior of $(\log Y_1, \ldots, \log Y_d)$. As discussed in Section 1.3.1.1, these are respectively known as additive and multiplicative random walks.

3. Different re-parameterisations of the components of $\boldsymbol{\lambda}$ and $\mathbf{Q}$

4. Different distributions for the proposed jump.

It is found in Section 3.3.4.2 that the choice between Gaussian proposals and proposals that decay exponentially does not signifiantly alter the optimum efficiency on the targets examined. For this reason, and for simplicity we use Gaussian jump proposals in all of our random walk simulations.

For random walk Metropolis updates the variances of the jump proposal distribution(s) must be tuned to achieve optimal mixing of the (MCMC) chain (see Section 1.3.3 and the whole of Chapter 3). In various situations in the limit as the number of dimensions $d \to \infty$ the optimum acceptance rate is approximately 0.234 (see Sections 3.1.1 3.3.1.7, 3.3.1.8, and 3.3.2.2). Our block updates have $d = 4$ and sequential updates $d = 1$, and for these dimensions the optimal acceptance rate depends on both target and proposal. We find in Sections 3.3.4.1 and 3.3.4.2 that

for a Gaussian jump proposal on a Gaussian target the optimal acceptance rates for 1-dimensional and 4-dimensional updates are approximately 0.44 and 0.30 respectively; but for a Gaussian target with an exponentially decaying proposal the corresponding values are 0.31 and 0.24. In each case however, efficiency as a function of acceptance rate has a relatively flat mode, indicating that the exact choice of acceptance rate is relatively unimportant provided that it is not too close to zero or one . We choose to tune our random walk algorithms (approximately) to the acceptance rate for the Gaussian, since although posteriors are not Gaussian, close to a mode they may be approximated as such.

The chain is run for a reasonable number of iterations, until it appears to have forgotten its starting position. Subsequent acceptance rates are examined and for sequential updates we simply alter each parameter to give an acceptance rate hopefully closer to 0.44 and re-iterate the procedure. For block updates it is possible to achieve an acceptance rate of 0.30 with jumps tuned approximately correctly for only one parameter, and all other proposed jumps having variances smaller than the optimal. In this case we must carefully increase the parameters re-iterating many times until all have achieved their maximal value such that the acceptance rate is still about 0.30.

Our main comparison is between the Gibbs sampler and a multiplicative sequential random walk (M1). We choose this latter firstly because we believe (Section 2.5.2) that posterior distributions contain heavy tails, and multiplicative random walks are generally better at exploring heavy tails than additive random walks (see Section 1.3.1.1). Tuning can be very time consuming and so we choose sequential

updates since parameters may then be tuned individually.

We undertake comparisons with other random walk algorithms, the full list being

M1 The multiplicative sequential random walk, with all parameters updated individually.

M2 The additive sequential random walk, with all parameters updated individually.

M3 The block multiplicative random walk, with all parameters updated at once.

M4 An additive random walk with reparameterisation suggested in Section 2.5.4.2.

M5 A mixed multiplicative/additive random walk with reparameterisation suggested in Section 2.5.4.2.

For algorithm M4, given $\mathbf{Q}$ (and hence $\boldsymbol{\nu}$) we perform a sequential update on $\overline{\lambda} := \boldsymbol{\nu}^t\boldsymbol{\lambda}$. We then pick a random direction $\boldsymbol{\xi}$ perpendicular to $\boldsymbol{\nu}$ and perform a sequential update on $\boldsymbol{\xi}^t\boldsymbol{\lambda}$; for a d-dimensional MMPP this is equivalent to a block update on the remaining $d-1$ degrees of freedom in $\boldsymbol{\lambda}$. Components of $\mathbf{Q}$ are then updated sequentially. All updates are additive.

For algorithm M5 we use the reparameterisation $(\overline{\lambda}, q, \alpha, \beta)$ as defined in Section 2.5.4.2. Parameters are updated sequentially, with multiplicative random walks on $\overline{\lambda}, q$ and α and an additive random walk on β, since it may take negative values. The parameter α as originally defined may also take negative values and so might not be suitable for a simple multiplicative random walk scheme. However $\alpha > 0$ corresponds to $\lambda_2 > \lambda_1$ which, as described later in this section, is exactly

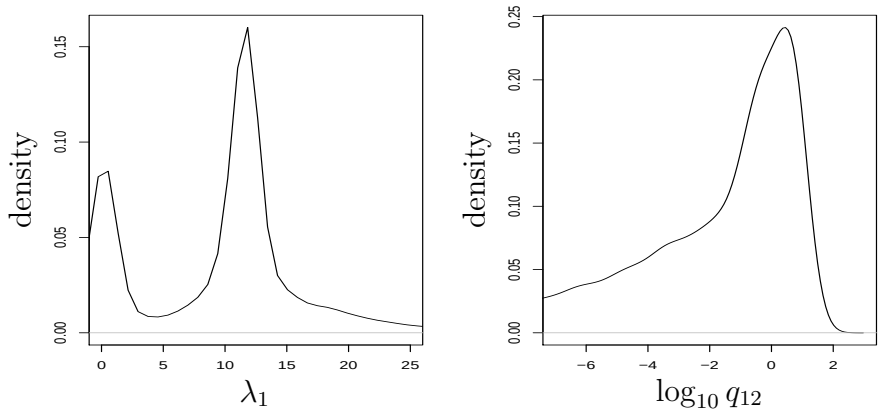

Figure 2.2: Density plots for λ_1 and $log_{10}(q_{12})$ for 20 000 iterations of the Gibbs sampler on replicate 1 of S4. Plots for the other two parameters are very similar due to frequent label-switching.

the ordering we perform on the output from the standard parameterisation before calculating ACT's. We therefore simply choose a positive initial value for α.

We suspect (Section 2.5.2) that vague gamma priors with shape parameter less than 1, might cause spurious modes close to the origin. To illustrate this we perform single runs of 20 000 iterations with the Gibbs sampler on S3 and S4 with sensible prior means but shape parameter 0.1 . Figure 2.2 shows posterior densities for (unordered) λ_1 and $\log_{10}(q_{12})$ from the Gibbs sampler run on the first replicate of S4 with gamma prior means of $\lambda_1 = \lambda_2 = n/t_{obs}$ and $q_{12} = q_{21} = 1$, but shape parameter 0.1.

As suspected there is a clear second mode for λ_1 close to zero; the heavy tails for $\log_{10}(q_{12})$ indicate that q_{12} itself also has a large second mode very close to zero.

For S3 results are similar but the second modes are less pronounced as the likelihood tails off that much more quickly away from its two modes.

In our main comparisons for the Gibbs sampler and M1 we perform 100 000 iterations on the first replicate of each data set, and 10 000 iterations on the second and third replicates of S1-S4. For M2-M4 we perform 10 000 iterations on the first and second replicates of S1-S4. Algorithm M5 is expected to perform well only when $\lambda_1 \approx \lambda_2$; we therefore treat this as a subsidiary investigation and simply perform 10 000 iterations on replicate 1 of S1-S4.

In the absence of further information it seems sensible to use the same prior distributions for each state. In the above runs the prior for each parameter is exponential with the mean for each q component always set to 1. The mean for each λ component is n/t_{obs} where n is the number of simulated events in the data set and t_{obs} is the observation period.

Since our priors are exchangeable and the likelihood of an MMPP is invariant under permutation of states, so too is the joint posterior. The (MCMC) chain may therefore be subject to label-switching (see Section 1.3.4.2). For label-switching to occur on the posterior of a two-dimensional MMPP which has the two (symmetric) modes reasonably well separated, the MCMC chain must pass through the region between the two modes, which generally corresponds to a low posterior density. The frequency of label-switching can therefore provide a heuristic indication of the ability of the chain to explore areas away from the main posterior mass (e.g. Celeux et al., 2000). Hence we place no restriction on our MCMC algorithms and allow

label-switching to occur.

With frequent label-switching, estimated marginal posterior distributions for each component will be very similar since they are sampled from the same (by symmetry) overall distribution. For example in a two-dimensional MMPP, the estimated posterior mean and variance for λ_1 will be approximately the same as those for λ_2; similarly for q_{12} and q_{21}. We wish to use the integrated autocorrelation time (ACT) for each component as our main measure of the degree of mixing of that component of an MCMC chain, but this too can be misleading in the presence of label-switching. If a component mixes well within each posterior mode, and occasionally switches between them, then the overall ACT may be higher for this chain than for one that mixes less well and is confined to a single mode.

We overcome these two label-switching problems by (if necessary) permuting the states in each parameter vector of the MCMC output so that $\lambda_1 \leq \lambda_2$. This is equivalent to using a joint prior distribution on λ_1 and λ_2 with this constraint built in. An argument can be made (e.g. Celeux et al., 2000) for a more discerning discrimination mechanism which takes into account the shape of the posterior distribution. We first note that the particular technique used only makes a strong difference in cases where there is significant mass between the symmetric modes. Secondly, our main interest is in the ACT's of the parameters as a measure of the efficiency of a given algorithm and these will be compared for different algorithms on the same posterior. Moreover different functionals of parameters will give different ACT's so that there is no absolutely correct measure of efficiency in any case. Ordering output according to λ_1 and λ_2 is a simple approach that nonethe-

less lends the desired properties to our ACT's. The ordering has been performed in all ACT's and qq plots shown for the standard parameter vector, but not for trace plots or density plots.

Initial values for all of the above runs are the true parameter values, which are known since the data is simulated. To investigate the tail behaviour of the Gibbs sampler and each of algorithms M1-M4 we identify two tail excursions from the main runs and perform three runs with each algorithm starting in each of these two tails.

2.6.1 Accuracy

We wish to compare the efficiencies of the different algorithms using their ACT's, however (for example) an algorithm that poorly explores a heavy tail may have a lower ACT than one that better explores the whole posterior. We therefore first assess the accuracy of the sampling distributions of the different algorithms. Using qq plots we visually compare the sampling distributions of the 100 000 iteration runs of the Gibbs sampler and M1. We then compare the Gibbs sampler and algorithms M1-M4 with a best estimate of the true posterior.

For runs S1, S2, there is no discernable difference between the sampling distributions of M1 and M2 (this is also the case for additional comparisons using datasets simulated from the same $\boldsymbol{\lambda}$ but where $(q_{12}, q_{21}) = (0.5, 2.0)$ or $(2.0, 0.5)$). With such a strong difference between the λ's it is relatively easy to discern the two states of the hidden chain and so the posterior has relatively light tails, and both algorithms can well explore the main posterior mass.

For S3 and S4, there are two places where the distributions differ noticeably: S3 for large λ_2 and S4 for small λ_1. Figure 2.3 shows the plots for S3, along with plots comparing the first 10 000 iterations of each run with iterations 11 000 to 100 000 of the exact same run. Since the ACT's for either algorithm are all less than 100 these plots are effectively comparing two independent runs of each algorithm. If an algorithm explores the posterior well, the two sampling distributions should both be close to the true posterior and so the qq plots should be a straight line. In general over all the S3 and S4 runs as well as over the additional runs with different $\mathbf{Q}$ or longer t_{obs} the Gibbs sampler has a better self-similarity than M1. In particular for the two cases where the Gibbs and M1 qq plots differ strongly, it is M1 that exhibits the poorer self-similarity.

We wish to compare the 10 000 iteration runs of all the algorithms against a "true posterior". In most cases it matters little whether we use output from the Gibbs sampler or M1 to represent this true posterior, but (from the previous paragraph) in the two cases where they disagree we are inclined to trust the output from the 100 000 iteration Gibbs sampler rather than from the M1 runs.

Figure 2.4 shows a comparison for the S4 data set between iterations 11 000 to 100 000 of the Gibbs sampler run with iterations 1-10 000 of the Gibbs sampler and each of algorithms M1-M4; qq plots for algorithm M5 are given in Figure B.1. Each algorithm (including M5) performs worse for this data set than for any other in replicate 1. The plots illustrate two points that recur through all the comparisons:

1. Quantile positions compare reasonably with their confidence limits so there

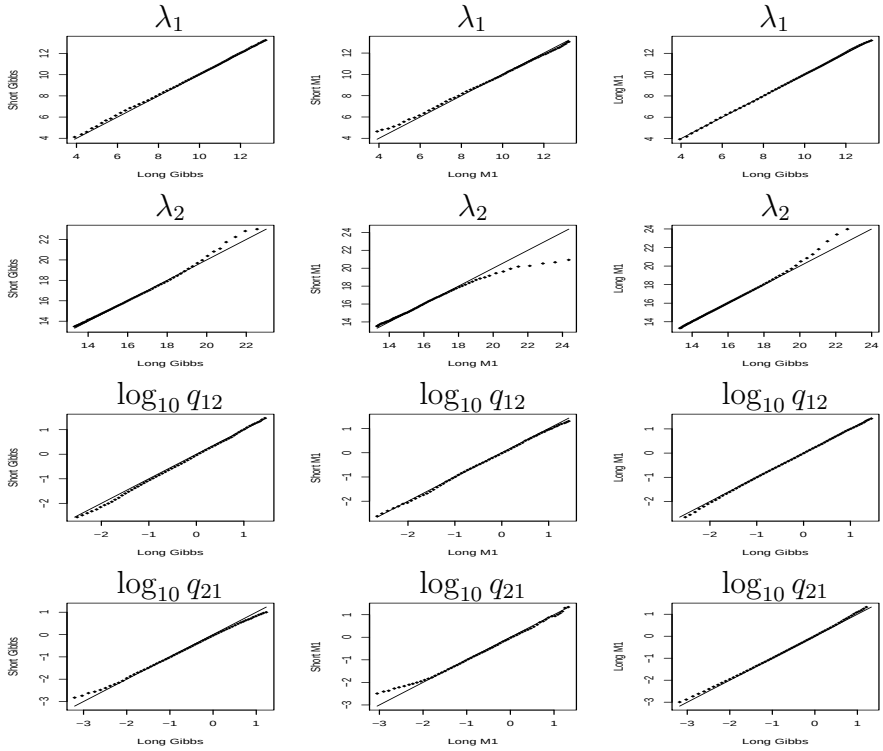

Figure 2.3: qq plots for runs of the Gibbs sampler and M1 on replicate 1 of S3. For each parameter, plots compare the the first 10 000 iterations of the Gibbs sampler against iterations 11 000 to 100 000, then the first 10 000 iterations of M1 against iterations 11 000 to 100 000, and finally all 100 000 iterations of M1 against the full 100 000 iterations of the Gibbs sampler.

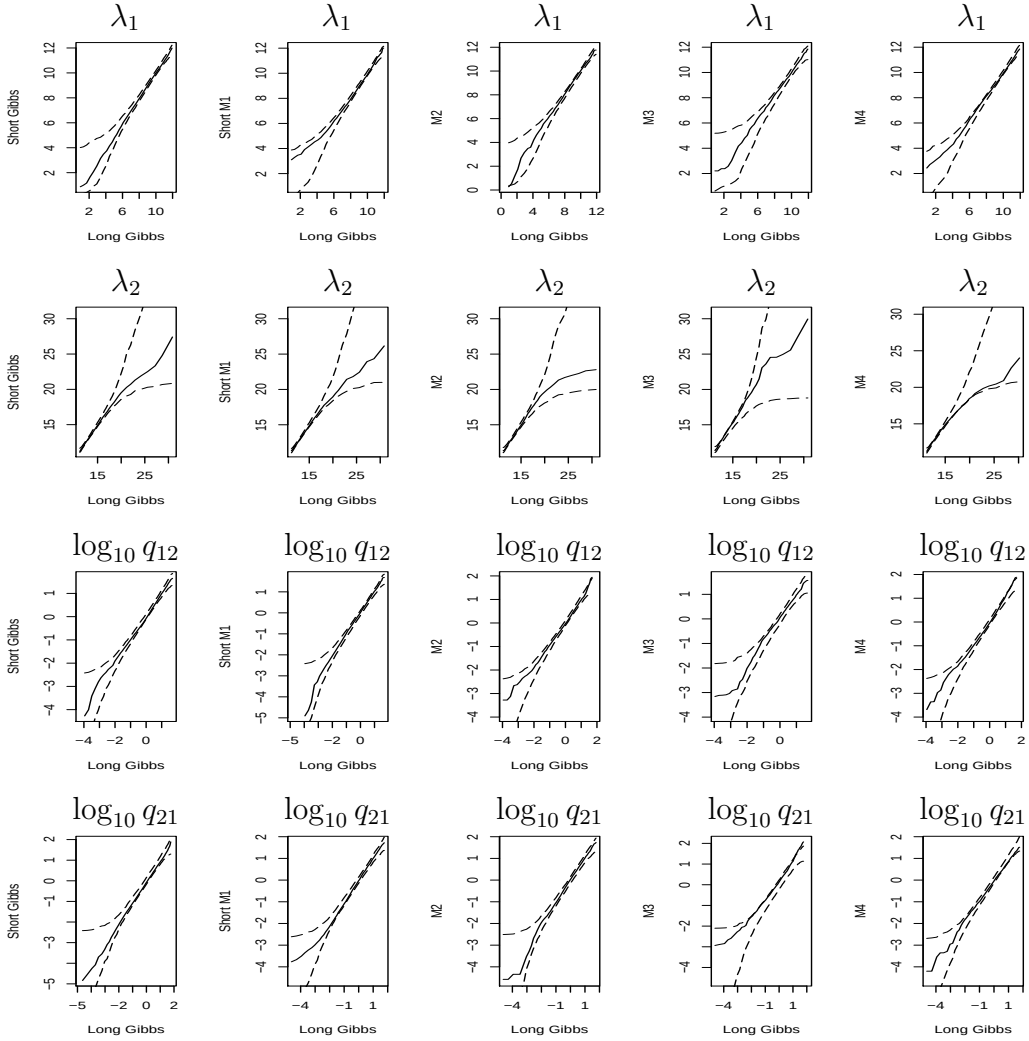

Figure 2.4: qq plots for replicate 1 of S4, comparing the first 10 000 iterations of the Gibbs sampler and of algorithm M1-M4 against iterations 11 000 - 100 000 of the Gibbs sampler. Dashed lines are approximate 95% confidence limits obtained by repeated sampling from iterations 11 000 to 100 000 of the Long Gibbs data; sample sizes were 10 000/ACT, which is the effective sample size of the data being compared to the Long Gibbs run.

is no reason to doubt the overall sampling accuracy of any algorithm

2. Both M2 and M4 appear to stick sometimes at relatively low q_{ij} values.

The slight overlap outside of the confidence limits for high λ_2 and low q_{12} in algorithm M5 is due to an excursion of about 40 iterations to $\lambda_2 \approx 40$ to 50, $\log_{10} q_{12} \approx -5$ to $-4 << \log_{10} q_{21}$; this also adversely affects the ACT's, as discussed in Section 2.6.2.1.

The second point is confirmed by a cursory examination of the trace plots (not shown, but see Figure 2.5 for similar behaviour) where over the 10 000 iterations there are many instances of one or other of the $\mathbf{Q}$ parameters remaining unchanged at a relatively low value for 20 or more iterations. This is because the chain is close to a degenerate state similar to Case 2 in Section 1.2.1 of Chapter 1, where q_{ij} is small, q_{ji} is O(1), $\lambda_i \approx n/t_{obs}$, but λ_j is far from the overall mean value. Additive updates to q_{ij} will most of the time propose making it O(1), taking $\overline{\lambda}$ far from n/t_{obs}. The likelihood for such a state is low, and so the move is usually rejected.

We wish to examine the behaviour of the algorithms in the tails of the posterior distribution, and therefore start the Gibbs sampler and each of algorithms M1-M4 at some low-probability point in parameter space. (As a subsidiary algorithm M5 has not been examined in the same detail as the other algorithms, however the short excursion to high λ_2 and low q_{12}, mentioned earlier in this section, suggests reasonable tail behaviour). It is unsatisfactory to simply pick a random point far from the main mass as this might be of such low density that its vicinity is in practice never visited by any of the algorithms, and the comparison would be meaningless.

Examining the output of our simulation study we identify two excursions where at least one parameter is far from its modal value: the M4 run on replicate 2 of S3, and the M2 run on replicate 1 of the additional data set with $\overline{\lambda} = (10, 13),\ q_{12} = 2.0,\ q_{21} = 0.5$. We pick specific parameter vectors from these excursions as starting points for our runs, and shall refer to these runs respectively as Excursions 1 and 2.

Runs for Excursion 1 start with $\lambda_1 = 13.9,\ \lambda_2 = 4.4,\ q_{12} = 1.4,\ q_{21} = 9.2$. The log-likelihood at this point is 2158.4, compared with a modal log-likelihood of 2167.3.

Runs for Excursion 2 start with $\lambda_1 = 12.9,\ \lambda_2 = 56.2,\ q_{12} = 0.00157,\ q_{21} = 0.1878$. The log-likelihood at this point is 1909.4, compared with a modal log-likelihood of 1912.2.

For Excursion 1 the log-likelihood gives a clear delineation between the tail and the main mass, and so we use a cut-off value of 2163.3 (about a 50^{th} of the modal value), above which the algorithm is deemed to have joined the main posterior mass. Excursion 2 is actually close to a region where the log-likelihood is approximately 1911 but which is not part of the main posterior mass and is not in the neighbourhood of the main posterior mode. The log-likelihood in this area appears to stay always below 1911.7, whereas in the main mass it often exceeds this value. Therefore we use 1911.7 as the cut-off value for Excursion 2.

For each excursion we perform 3 runs of each algorithm. Table 2.2 shows the mean number of iterations taken to reach the main posterior mass. We note the

Ex	Gibbs	M1	M2	M3	M4
1	28	21	84	143	113
2	4	53	279	132	38

Table 2.2: Mean number of iterations to find the main posterior mass for the 3 runs of each algorithm in each of the two excursions.

consistently good performance of the Gibbs sampler, and that the multiplicative sequential random walk (M1) appears to perform better than the additive sequential random walk (M2). Taking CPU times into account, with the Gibbs sampler and the block multiplicative random walks respectively about 3 and 4 times faster than the sequential random walks (see Section 2.6.2.1), the performance of the Gibbs sampler is even more impressive, and those of the sequential and block multiplicative random walks are comparable.

The additive random walk performs so poorly in Excursion 2 as it takes a long time for λ_2 to reach sensible values, and (for reasons already noted) the algorithm spends long periods with none of its proposed updates to q_{12} being accepted (Figure 2.5 shows the first 500 iterations of the first run for M2).

Due to our method of choosing the two excursions, the starting point for each is necessarily a point where some algorithm performs poorly. It so happens that both of the algorithms in question (M2 and M4) are additive random walks, and we have therefore biassed our testing against additive random walks. A more complete set of tail tests would find alternative tail starting points where the Gibbs sampler or the multiplicative random walks perform poorly. Such points, however, are not

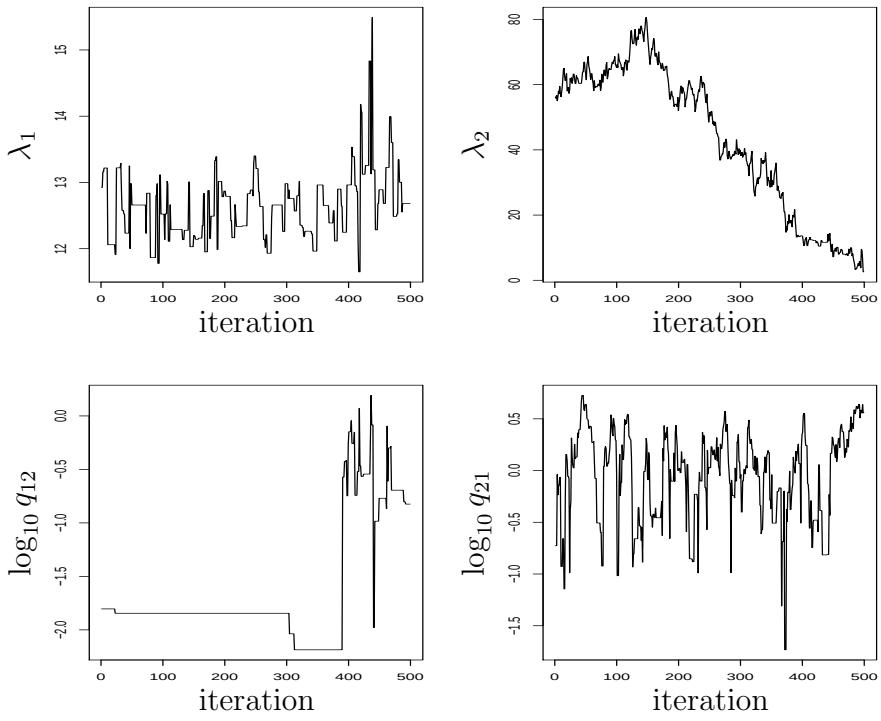

Figure 2.5: Trace plots for the first 500 iterations of the first run of M2 at Excursion2.

obvious from our results.

2.6.2 Efficiency

We now compare efficiencies of the algorithms, mainly through estimated autocorrelation times, but also, through label-switching.

2.6.2.1 Integrated autocorrelation time

The integrated autocorrelation time of a parameter gives the effective number of consecutive realisations equivalent to one independent realisation of that parameter. Thus it provides a measure for direct comparison of efficiencies between algorithms. However different algorithms take different amounts of time to produce a fresh realisation of the parameter vector. A more practical measure of efficiency is therefore the time to produce a single (in effect) independent realisation of a parameter. The exact timing and CPU used should not be important so we use timings relative to those of the Gibbs sampler. Our measure of efficiency is therefore

$$ACT_{rel} = ACT \times \frac{\text{time per iteration (algorithm)}}{\text{time per iteration (Gibbs)}}$$

Lower values correspond to greater efficiency.

Timings relative to the Gibbs sampler are consistent at about 3.2 for (M1), 3.1 (M2), 0.8 (M3), 3.0 (M4), and 3.5 (M5) (see Table B.6). The most CPU intensive operation is the forward-backward accumulation step used to calculate the likelihood in (M1-M5) and to begin sampling the states at event times in the Gibbs sampler. The Gibbs sampler and M3 apply this once per iteration, whereas M1, M2, M4 and M5 apply it once per parameter.

Different functions of a particular parameter will posses different ACT's. Intuition reinforced by the trace plots and qq plots of section 2.6.1 leads us to believe that some of the algorithms behave poorly at low values of q_{ij} and this would not be picked up if we simply examined the ACT's of the parameters. We therefore use the ACT's of $\log(q_{12})$ and $\log(q_{21})$ instead.

We compare estimated relative autocorrelation times for the parameters λ_1, λ_2, $\log q_{12}$ and $\log q_{21}$ with states ordered so that $\lambda_1 < \lambda_2$ for the first 10 000 iterations in every run performed. Results for the core runs are shown in Table 2.3. Results for additional runs are given in Appendix B.

There is a great deal of variability across parameters and replicates due to random variation, as well as (inevitably) imperfect tuning of the random walk algorithms. Further, different replicates are different datasets, and while simulated from the same parameter values these will have different properties, especially when $\delta := (\lambda_2 - \lambda_1)/\overline{\lambda}$ is small.

Despite all the variability it is clear that the Gibbs sampler is more efficient than the Metropolis-Hastings algorithms M1-M4 and more efficient than M5 on runs S1 and S2. The contrast in efficiency is most striking for S1 where the Gibbs sampler is consistently at least an order of magnitude more efficient than any of the other algorithms; for most of the other datasets the Gibbs sampler is at least twice as efficient as random-walk algorithms M1-M4. However M5 performs simlarly to the Gibbs sampler on S3 and arguably outperforms the Gibbs sampler on S4, the sce-

Data	Alg	λ_1	λ_2	q_{12}	q_{21}
S1	Gibbs	1.2	1.2	1.4	1.4
	M1	14.0	14.1	13.3	14.9
	M2	13.5	12.9	15.3	12.8
	M3	11.8	10.5	10.5	9.8
	M4	7.9	82.4	14.3	14.8
	M5	9.5	14.2	105.0	161.5
S2	Gibbs	4.2	3.3	5.8	6.0
	M1	22.0	19.4	28.2	28.6
	M2	19.1	20.6	29.6	27.3
	M3	10.5	15.2	16.6	18.5
	M4	12.5	25.7	27.4	25.1
	M5	13.2	14.6	27.2	29.9
S3	Gibbs	24.2	18.3	33.3	23.6
	M1	63.7	49.5	76.4	68.5
	M2	101.5	84.8	104.6	73.3
	M3	38.2	24.9	34.3	32.0
	M4	74.8	55.2	85.2	72.1
	M5	17.7	22.0	23.2	18.5
S4	Gibbs	32.2	28.9	35.5	53.0
	M1	90.7	91.0	106.1	127.7
	M2	99.6	130.0	110.9	155.4
	M3	73.9	62.1	86.9	85.1
	M4	76.1	94.2	109.3	109.9
	M5	23.6	70.6	10.6	8.2

Table 2.3: ACT_{rel} for replicate 1 of simulated data sets S1-S4.

nario closest to that for which it was designed. A qualitatively similar pattern is observed in all additional replicates (see Tables B.2, B.3, and B.4). In general the Gibbs sampler appears to be the most efficient algorithm on most of the simulated datasets we analysed. The only exceptions occur for datasets where the Poisson intensities corresponding to the two different states of the chain are similar; in these cases sometimes M5 is the most efficient of all the algorithms.

For S1 (and additional datasets with very different λ_1 and λ_2 - see Appendix B) the reparameterisation M5 is very inefficient due to the poor mixing of the q parameters. However for all other datasets the reparameterisation is at least as efficient as any of the other sequential random-walks, and is often much more efficient than these. Further investigation of the exploration of S4 by algorithm M5 showed that the high ACT_{rel} for λ_2 is due to the short excursion to high λ_2 and low q_{12} already mentioned in Section 2.6.1.

The multiplicative sequential random walk (M1) does not appear to be any more efficient than its additive counterpart (M2). We must recall though that the ACT measures the efficiency of an algorithm at exploring the portion of the parameter space that it actually explores. It does not penalise an algorithm that completely misses a heavy tail for example, and we have already remarked that M1 appears to explore heavy tails better than M2.

To assess the success of the reparameterisation M4 we compare the two additive random walks. For data set S4, for 7 of the 8 ACT's across the 2 replicates the M4 reparameterisation performs better than M2, perhaps justifying the reparam-

eterisation in this case. For parameter λ_2 in both replicates of S1, M4 behaves noticeably worse than M2 and in all other cases there is no clear difference between the performance of the two algorithms. The reparameterisation does not produce the hoped for improvement in S3, nor in the additional replicates (see Appendix B) similar to S3 except with $t_{obs} = 400$ instead of 100. An explanation for these results is presented in 2.6.3.

It is also noticeable that the ACT's for all algorithms tend to increase with decreasing $(\lambda_2 - \lambda_1)/\overline{\lambda}$, and therefore heavier posterior tails, which are naturally more slowly explored.

2.6.2.2 Label-switching

Label-switching can provide another indication of an algorithm's ability to explore areas of low mass. In many of the runs either the region between the modes is of such low density that switching does not occur for any algorithm, or the modes are so close together that all algorithms are effectively switching nearly every iteration. However for replicates 1 and 2 of S3 and replicate 1 of S4 the spacing of the modes allows a meaningful comparison of frequency of label-switching. Table 2.4 shows the mean number of label switches per 10 000 iterations (recall that for replicate 1, both the Gibbs sampler and M1 were run for 100 000 iterations).

Overall it appears that M4 is best able to switch between the modes. Firstly we note that $\overline{\lambda}$ is invariant to label-switches; secondly that for S3 and S4 we will often find $\nu_1 \approx \nu_2$ and so parameter $\lambda^{\perp} := \nu_2\lambda_1 - \nu_1\lambda_2$ (for which large jumps are performed) is approximately parallel to the line (in λ-space) between

Data	Gibbs	M1	M2	M3	M4
S3 rep. 1	20	12	1	2	9
S3 rep. 2	1	0	0	0	2
S4 rep. 1	20	15	13	9	97

Table 2.4: Mean number of label-switches per 10 000 iterations for replicates 1 and 2 of S3 and replicate 1 of S4.

the two modes. The Gibbs sampler label-switches more frequently than any of the Metropolis-Hastings random walks with the standard parameterisation.

2.6.3 Information matrices and efficiency

Some of the results in Section 2.6.2.1 may be explained through the properties of the observed information matrices. As discussed in Section 2.5.4.2, the closer this matrix is to diagonal with respect to a particular parameterisation, the more efficient this parameterisation is likely to be under a (tuned) sequentially updating MCMC scheme. An approximate measure of the closeness of a (symmetric) matrix to diagonality is obtained by examining the eigenvectors, normalised to be of length 1. If the matrix is in fact diagonal then each of these will have a single non-zero component, which will be of length 1. We examine the largest component of each normalised eigenvector; the closeness of these to 1 gives a measure of the closeness of the matrix to diagonal.

The information matrix for the observed data ($\mathbf{V}^*$) was estimated at the posterior mode by numerical differentiation of the log-likelihood for the observed data. Table 2.5 shows the observed-data information matrices for replicate 1 of S1 and S4

Data	Information Matrix
S1	$\begin{bmatrix} 5.33 & -0.06 & 0.50 & 0.17 \\ -0.06 & 0.47 & -0.13 & -0.21 \\ 0.50 & 0.013 & 51.62 & -7.00 \\ 0.17 & -0.21 & -7.00 & 37.55 \end{bmatrix}$
S4	$\begin{bmatrix} 1.71 & 0.20 & 4.19 & -4.78 \\ 0.20 & 2.67 & 3.21 & -7.35 \\ 4.19 & 3.22 & 16.27 & -20.76 \\ -4.78 & -7.34 & -20.76 & 34.64 \end{bmatrix}$

Table 2.5: Information matrices for S1 and S4 at the MLE, estimated by numerical differentiation.

with repsect to the standard parameterisation $(\lambda_1, \lambda_2, q_{12}, q_{21})$. Table B.5 shows the same matrices for replicate 1 of S2 and S3.

The information matrix for S1 is very close to diagonal, as predicted by (2.22), with all eigenvectors having one component of at least 0.92. By contrast, the maximum component for the eigenvectors of the information matrix for S4 range between 0.57 and 0.82. Information matrices for S2 and S3 lie between these extremes.

Information matrices were transformed to the following reparameterisations (see Section 2.5.4.2 for notation): $(\lambda_1, \lambda_2, \nu_1, q)$, $(\overline{\lambda}, \delta, \nu_1, q)$, $(\overline{\lambda}, \lambda^{\perp}, q_{12}, q_{21})$ (corresponding to algorithm M4), and $(\overline{\lambda}, q, \alpha, \beta)$ (corresponding to M5). In each case we looked at the maximum component of each (normalised) eigenvector. For S1 all parameterisations except M4 performed well; for M4 the maximum components were

between 0.76 and 0.78, but for the eigenvectors of the other matrices all maximum components were at least 0.92. The best parameterisation for S1 was $(\lambda_1, \lambda_2, \nu_1, q)$, with all maximum components > 0.999. For S2, S3, and S4 the best parameterisation was $(\overline{\lambda}, q, \alpha, \beta)$ with all maximum components at least 0.98.

The success of the $(\overline{\lambda}, q, \alpha, \beta)$ reparameterisation for S2 (and even for S1) is perhaps surprising as it was suggested by an expansion valid for $\lambda_1 \approx \lambda_2$. The results indicate that this might be a good reparameterisation over a broad range of data information on the hidden chain. $\mathbf{Q}$ parameters might be explored more efficiently if the multiplicatice random walk could somehow be worked into updates for β as well as α.

The approximately diagonal nature of the $\boldsymbol{\lambda}$-components of the information matrices of S1 and S2 also provides insight into the mixing properties of M4 for these data sets. The reparameterisation of M4 has $\overline{\lambda} = \nu_1\lambda_1 + \nu_2\lambda_2$ and $\lambda^{\perp} = \nu_2\lambda_1 - \nu_1\lambda_2$. However for S1 and S2 we have $q_{21} \approx q_{12}$ and so $\nu_1 \approx \nu_2 \approx 1/2$; the new parameters are at approximately 45 degrees to the optimum and the best random walk will propose roughly equal scaling for both parameters, constrained by the parameter for which there is most information: the smaller parameter, λ_1. Therefore λ_2 will be explored less efficiently and the ACT's will be higher, as found in Section 2.6.2.1.

2.7 Analysis Chi site data for *E.coli*

2.7.1 Background and the *E.coli* data

In recent years there has been an explosion in the amount of data describing both the genomes of different organisms, and the biological processes that effect the evolution of these genomes. There is much current interest in understanding the function of different features of the genome and what affects the biological processes such as mutation and recombination. One approach to learning about these is via genome segmentation (e.g. Li et al., 2002): partitioning a genome into regions that are homogeneous in terms of some characteristic (e.g GC content), and then looking for correlations between this characteristic and either another characteristic, or a biological process of interest. For example regions with high recombination rates are known to correlate with regions of high GC content (Kong et al., 2002) which has led to various possible explanations of how recombination hotspots may have evolved Eyrie-Walker and Hurst (2001); Marais (2003); Galtier et al. (2001).

Here we consider segmentation of a bacterial genome based on the rate of occurence of a particular DNA motif - called the Chi site. The Chi site is a motif of 8 base pairs: GCTGGTGG. It is of interest because it stimulates DNA repair by homologous recombination (Gruss and Michel, 2001), so the occurence of Chi sites has been conjectured to be related to recombination hotspots.

Our data is for *E.coli* DNA and consists of the position (in bases) of Chi sites along the genome. Figure 2.6 shows a schematic of the circular double stranded DNA genome of *E.coli*, with the two strands represented by the inner and outer

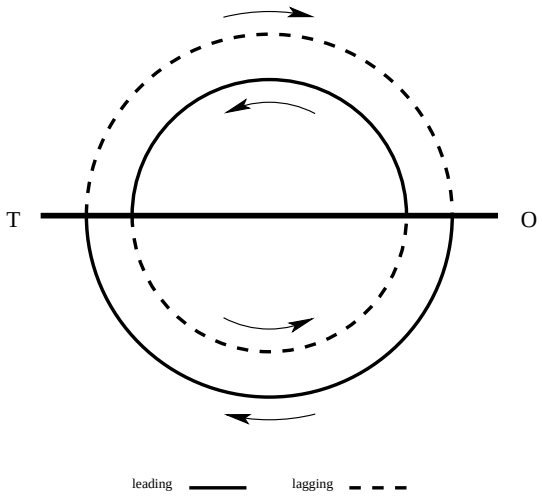

Figure 2.6: Schematic of the leading and lagging strands on the inner and outer rings of the *E.coli* genome split by the replication origin (O) and terminus (T), together with the direction relevant for Chi site identification.

rings. There is a 1-1 mapping of bases between the outer and inner strands (C $\leftrightarrow$ G and A $\leftrightarrow$ T) so that each uniquely determines the other. The figure also indicates a directionality associated with different halves of each strand as split by the replication origin (O) and terminus (T). The molecular mechanisms of DNA replication differ between the two half-strands and they are termed *leading* and *lagging*, as indicated in the figure.

The 1-1 mapping between base pairs together with the reversing of directionality between inner and outer strands implies that searching for the Chi site in the outer strand is equivalent to searching for CCACCAGC in the inner strand. This sequence is different enough from the sequence for the Chi site in the inner strand that occurences of the Chi site in the inner and outer strands are effectively independent. Occurence of Chi sites in leading and lagging halves are also independent

since these are separate parts of the genome. Thus our data consists of four independent sets of positions of Chi sites - along leading and lagging halves of both inner and outer strands. Figure 2.7 shows the cumulative number of events along the genome for each of these data sets.

The replication and repair mechanisms for leading strands are different to those for lagging strands so in general we might expect them to have different compositional properties (densities of nucleotides and oligonucleotides). A bias in the frequency of Chi sites favouring leading strands has been noted in several genomes, including *E.coli* (e.g. Karoui et al., 1999) and is evident from the figure. A more open question is whether there is variation within the leading and/or lagging strands, rather than just between the leading and lagging strands.

Our aim is to first determine whether Chi sites appear to occur uniformly at random within each of the leading and lagging strands, or whether there is evidence of the intensity of the occurence of Chi sites varying across either strand. Secondly, if there is variation then we would like to infer the regions with strong evidence for either a high or low intensity of Chi sites.

The *E.coli* genome (defined as single strand length) is 4 639 675 bases long so each of the individual halves are 2319.838 kilobases (kb) long. Henceforth we use units of kb.

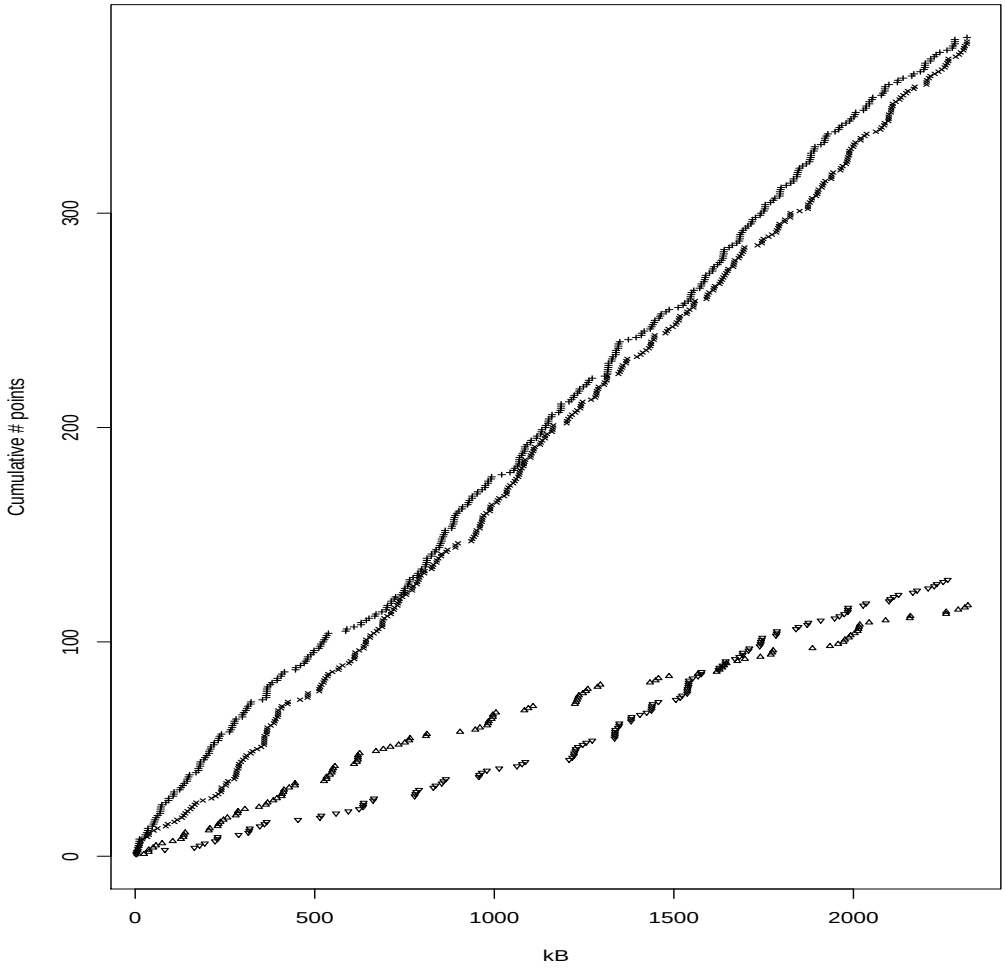

Figure 2.7: Cumulative number of occurences of the Chi site along the genome for leading (+) and lagging ($\triangle$) halves of the outer strand and leading ($\times$) and lagging (∇) halves of the inner strand.

2.7.2 Model and prior

We analyse the positions of occurences of the Chi site along first leading then lagging strands using our Gibbs sampler. These positions are discrete bases and our Gibbs sampler applies to continuous data, however each of the four strands is over 2319kb long and contains less than 400 occurences of the 8-base Chi site, so it is reasonable to model this discrete process as continuous. Furthermore, a straightforward approach to discrete modelling would involve applying the forward-backward algorithm across the entire genome, which would be computationally prohibitive.

One of our aims is to perform model choice, and the choice of model will depend on the priors for each model; in particular we cannot use uninformative priors (e.g. Bernardo and Smith, 1995, Chapter 6). For the results presented here, exponential priors are used for each λ_i and for each total intensity with which the underlying Markov chain leaves state i, ρ_i's; uniform priors are employed for each vector of transition probabilities.

We first analyse the inner leading and lagging strands and use the results from these to inform priors for analyses of the outer leading and lagging strands, which we use to perform model choice. We also tested robustness of our results to variation in the priors.

We analyse the inner strands using exponential priors, the means of which are chosen empirically from the data for each strand. The mean for all λ parameters is set to n/t_{obs}, where n and t_{obs} are respectively the number of Chi sites and the total length in kb of the strand. The mean for all q parameters needs to

be somewhere between $1/t_{obs}$ and n/t_{obs} for an analysis to be feasible so we set it to $\sqrt{n}/t_{obs}$. These latter choices are rather arbitrary, but the resulting posteriors are only used to inform the (weak) priors for the analyses of the outer strands.

Since states for the analyses of the inner strand are exchangeable, we order the results such that $\lambda_1 \leq \lambda_2$ and use the posterior means as means for the exponential priors in the analysis of the outer strands. Since the runs for the outer strands have non-exchangeable priors, we may not order the output and must treat it exactly as it appears.

For each strand we analyse the 1-d case analytically and the 2-d and 3-d cases using 100 000 iterations of our Gibbs sampler.

2.7.3 Results

Figure 2.8 shows trace plots for the first 20 000 iterations and ACF's over the first 10 000 iterations for the 2-d run on the lagging strand of the outer ring. The trace plot for λ_1 shows one of only 6 mode-switch-and-return's (all brief), indicating that the different priors fix quite firmly the ordering of the states. These brief switches do however exert a strong (and spurious for our purposes) influence on the ACF's, and so we show ACF's for a period in which there is no mode-switching; the mixing appears to be satisfactory.

Posterior model probabilities for the leading and lagging strands are calculated according to the method described in Section 2.5.3 and are given Table 2.6. They

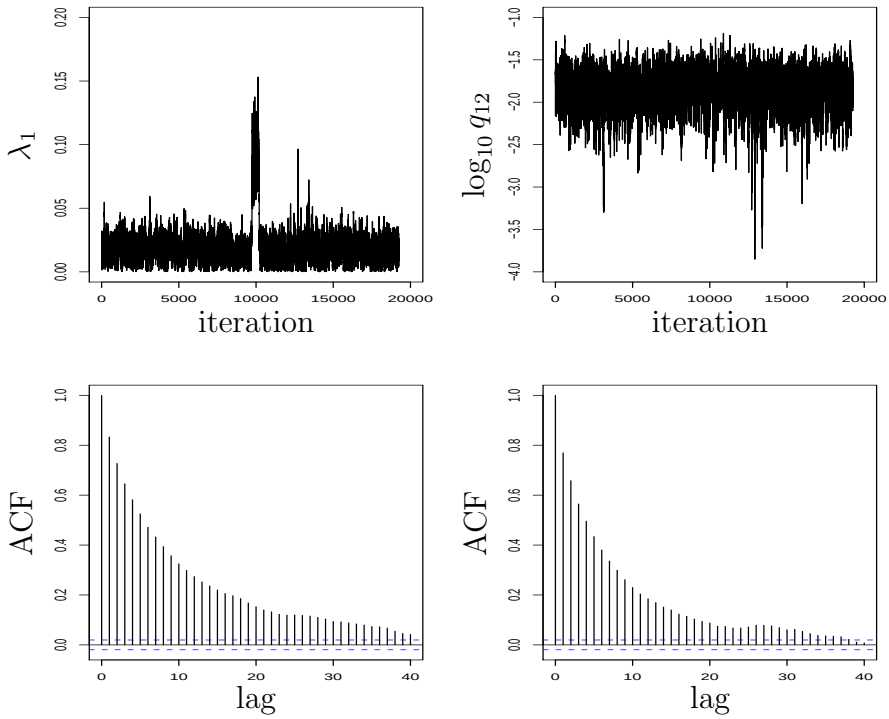

Figure 2.8: Trace plots for the first 20 000 iterations and and ACF's for the first 10 000 iterations of the Gibbs sampler for the lagging strand of the outer ring with non-exchangeable priors derived from the run for the lagging strand of the inner ring.

Dataset	1-D	2-D	3-D
lagging (outer)	<0.01	0.83	0.17
leading (outer)	0.30	0.44	0.26

Table 2.6: Posterior model probabilities for leading and lagging halves of the outer strand.

indicate a clear choice of a two-dimensional model over a one-dimensional model for the lagging strand. There is also substantial evidence for a two-dimensional model in preference to a three-dimensional model. From the model probabilities alone there is nothing to choose between one, two, and three dimensional models for leading strands. An alternative view on the evidence is found through contour plots of the posterior for $\boldsymbol{\lambda}$ in the 2-component models (Figure 2.9). For the lagging plot the two components of the modal $\boldsymbol{\lambda}$ differ by nearly an order of magnitude, whereas for the leading plot the ratio between the larger and smaller components is less than two. Moreover the lagging plot shows only a single mode as the second mode, corresponding to the mode switch noted in the trace plots, has approximately $1/20^{th}$ the mass of the main mode. By contrast the leading plot shows two modes with a great deal of mass between them, including the neighbourhood of some of the points (λ_*, λ_*) which correspond to a simple Poisson process. The two plots indicate a clear choice of a model with (at least) two different λ values for lagging strands and uncertainty between one and two λ values for the two-dimensional model.

For the two-dimensional model for lagging strands the posterior mean parameter values correspond to intensities of 20.8 and 92.1 Chi sites per megabase (Mb), and an intensity of 16.0 transfers per Mb from the lower state to the higher state and 21.1 transfers per Mb from the higher state to the lower state. The one-dimensional model for leading strands has posterior mean intensity of 164.7 Chi sites per Mb.

Evaluation of posterior model probabilities is subject to Monte-Carlo error which depends on the special parameter value $\boldsymbol{\theta}^*$. However experimentation showed the

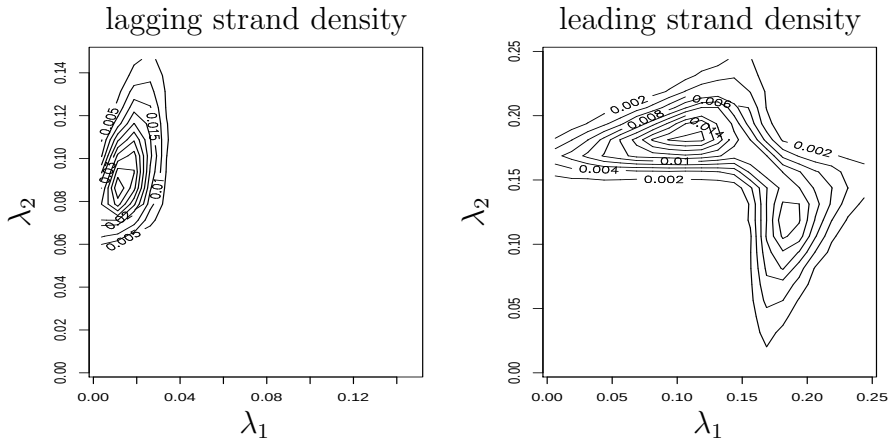

Figure 2.9: Contour plots of λ_1 vs. λ_2 from all 100 000 iterations for the lagging and leading strands

first decimal place of $\log_{10}(P(\mathbf{t}|\text{ model}))$ to be extremely robust to even quite large variations of $\boldsymbol{\theta}^*$ from the posterior mean value, only altering when there was virtually no posterior mass at the parameter value. Posterior model probabilities may also be sensitive to the exact prior used, and since the data contains less information about the $\mathbf{Q}$ parameters than the $\boldsymbol{\lambda}$ parameters, the $\mathbf{Q}$ priors may be particularly influential. Further analyses of the outer and inner rings were performed with exchangeable exponential priors for $\boldsymbol{\lambda}$ and with exchangeable exponential, (approximately) normal, and truncated exponential priors for each ρ_i. There was little change in the posterior means for the ordered $\boldsymbol{\lambda}$ vector, but a great deal of variability in $\mathbf{Q}$ as expected. However the posterior model probabilities always indicated at least a two-state model for lagging strands and little to choose between one and two state models for leading strands.

A possible biological explanation for our results is given by how replication differs

on leading and lagging strands. Leading DNA strands are replicated continuously whereas lagging strands are replicated in fragments. It may be the fragmentary nature of replication that is causing the hetrogeneity in rate of occurrence of Chi sites.

We can use the output of the Gibbs sampler to perform segmentation of the lagging strands based on the intensity of occurence of Chi sites. Figure 2.10 plots the mean (over 1000 chains sampled every 100 iterations) intensity against position along the genome. This gives a "smoothed signal" of Chi site intensity which could be used to evaluate correlations with (say) recombination rates across the genome. An alternative segmentation might be based on the posterior probabilities that a given point along the genome is in each of the possible states - for this segementation, at each point the chain is simply set to the state with the highest posterior probability.

2.8 Discussion

We have presented a novel approach to simulating directly from the conditional distribution of a continuous time Markov process and shown how this can be used to implement a Gibbs sampler for analysing MMPPs. The Gibbs sampler can analyse data where the event times are directly observed, and also data where the number of events or even only the presence/absence of events is known for a sequence of time intervals.

The Gibbs sampler has a number of advantages over standard Metropolis-Hastings samplers. Firstly, the Gibbs sampler requires no tuning; tuning for Metropolis-Hastings algorithms can be time consuming - especially for long datasets where

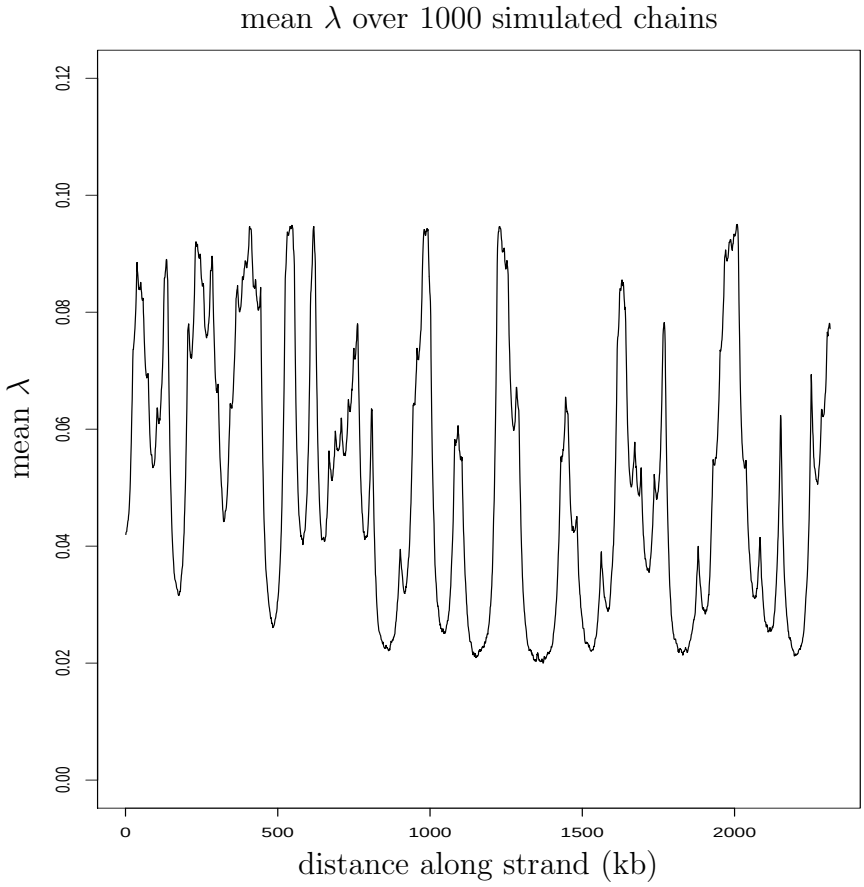

Figure 2.10: Mean λ value at each point in the lagging strand, derived from 1000 effectively independent simulations of the parameter vector and the underlying chain.

the algorithm takes longer to run and for algorithms involving blocking of parameters. Further such tuning is valid for the area of the posterior being explored whilst the tuning takes place (hopefully the mode); there is no guarantee that it will be appropriate for as yet unseen tail areas that the algorithm should eventually explore.

Secondly, our simulation results suggest that the Gibbs sampler is more efficient than many Metropolis-Hastings random walks for fixed CPU time. For a wide range of data sets that we analysed the relative ACT (taking CPU time into account) of the Gibbs sampler was always smaller than the relative ACT for any of the standard Metropolis-Hastings algorithms tested, sometimes by an order of magnitude. In general, the more information about the hidden underlying Markov chain contained in the data, the more efficient the Gibbs sampler. Finally, a by-product of the Gibbs sampler is that we can investigate the posterior distribution of the underlying chain.

There has been previous work on developing a Gibbs sampler for MMPP's. Scott (1999) and Scott and Smyth (2003) present an approximate Gibbs sampler that can be applied to certain MMPP's, assuming the event times are directly observed. Their approximation is to assume that certain state changes coincide precisely with observed events. In many situations this approximation will be negligible; Scott (1999) models times at which a bank account is accessed, where a criminal may or may not have obtained the bank details; it is argued that it is sensible to *define* the arrival of a criminal as the time at which he/she first accesses the account. Further Scott and Smyth (2003) argue that forcing state changes to start and

end at event times "eliminates the possibility of pathological bursts containing no events". However their Gibbs sampler also places restrictions on the allowable state changes: all transitions to states with lower intensities than the current state are permitted, but out of all the (ordered) states with higher intensity than the current state, transitions are only permitted to the state immediately adjacent to the current one. Also the approximation of restricting state changes to event times will become less accurate as the rates of the generator for the hidden chain increase towards the same order of magnitude as the intensities of the observed process. Our Gibbs sampler avoids these issues and there is little extra cost in implementing it.

Blackwell (2003) and Bladt and Sorensen (2005) use rejection sampling to sample from the exact distribution of a discretely observed continuous-time Markov process. A chain is simulated forward from a given observed state, and if the simulated state at the next observation time does not match the corresponding observed state then the chain is rejected and the process repeated until a match is achieved. A similar technique could replace stage 2 of our Gibbs sampler, where we simulate from the hidden chain and the observed event process and accept the hidden chain if the chain finishes in the correct state and there are no observed events. This is efficent only when the number of rejected chains is small. It is straightforward to calculate the expected number of simulations until acceptance for an interval of known length given the start and end states. We calculated this for the simulated states at event times at every iteration of our Gibbs sampler for every one of the 1164 intervals in the S4 data set. On average for about 700 of the intervals 3 or fewer chain simulations were expected to be required. However the

distribution of the expected number of simulations had a very heavy right hand tail, with about 200 intervals requiring at least 10 simulations and about 20 requiring more than 100 simulations, so that the mean expected number of simulations per interval was around 20. This number is likely to increase as the number of hidden states increases. In practice stage 2 of our Gibbs sampler takes a very small proportion of the CPU time and this would be likely to remain small if rejection sampling were to be used instead, unless the number of rejections was large.

Both our simulation study and consideration of the form of the information matrices for MMPPs gives insight into how to implement Metropolis-Hastings schemes. Multiplicative random walks are preferable to additive random walks since they mix better on the more heavy tailed posteriors (for example the additive random walks sometimes stick at low q values). The efficiency of the standard parameterisation $(\mathbf{\Lambda}, \mathbf{Q})$ increases with increasing information in the observed data about the the hidden underlying chain, since at the extreme of complete knowledge the information matrix is approximately diagonal with respect to this parameterisation. By examining the information matrices for simulated datasets at the modal values we checked a variety of alternative parameterisations for the two-dimensional MMPP's when there is less knowledge about the underlying chain. The most promising, $(\overline{\lambda}, q, \alpha, \beta)$ is based on the form of a cubic Taylor expansion of the log-likelihood for small $(\lambda_2 - \lambda_1)/\overline{\lambda}$. An implementation of this parameterisation was found to outperform all the other Metropolis-Hastings algorithms on datasets containing relatively little information about the hidden chain, and on such datasets its performance was comparable with that of the Gibbs sampler.

The posterior distribution for MMPPs can be highly complicated; with multiple modes (due to the invariance of the likelihood to label-switching), and heavy tails (for example, due to degeneracy into a lower dimensional MMPP). In particular, it is not possible to use improper priors for MMPPs, as these lead to improper posteriors. Since we are mainly interested in comparing algorithms using ACT's rather than in the posterior distribution itself we use symmetric priors and (as post-processing) order the states according to components of the intensity vector. Stephens (2000) and Celeux et al. (2000) discuss possible better ways of labelling states if the main interest is in summarising the posterior distribution.

We considered the application of MMPPs to modelling the occurence of a specific DNA motif in *E.coli*. We found evidence for heterogeneity in the occurrence of this DNA motif, the Chi site, in the lagging strand; which may have a biological explanation in terms of the replication process on this strand. The output of our Gibbs sampler also enables us to segment the lagging strand into regions of high and low intensity of these Chi sites. Ideally we would like to use this segmentation to test for correlation of high Chi site intensity with regions of high recombination rates, but unfortunately data is not currently available on the variation in recombination rate in *E.coli*.

Chapter 3

Optimal scaling of the random walk Metropolis

3.1 Introduction

The random walk Metropolis algorithm (RWM) is introduced in Section 1.3.1.1. Consider the behaviour of this algorithm as a function of some overall parameter for the scale of proposed jumps. If most proposed jumps are small compared with some measure of the scale of variability of the target distribution then, although these jumps will often be accepted, the chain will move slowly and exploration of the target distribution will be relatively inefficient. If the jumps proposed are relatively large compared with the target distribution's scale, then many will not be accepted, the chain will rarely move and will again explore the target distribution inefficiently. This suggests that given a particular target and form for the jump proposal distribution, there may exist a finite scale parameter for the proposal such that the algorithm will explore the target as efficiently as possible. This chapter is concerned with the definition and existence of an optimal-scaling, its asymptotic

properties, and the process of finding it. We start with a review of current literature on the topic, which is concerned with asymptotic properties.

3.1.1 Existing results for optimal scaling of the RWM

Previous theoretical investigations into optimal scaling of the RWM reviewed in this section have all taken the same general approach. The chain is "speeded up" by a factor of some positive power of dimension d. Provided the scale parameter of the proposal distribution decreases with d just quickly enough to compensate for the speed up, a single component of the new chain is shown to approach a Langevin diffusion, as defined in Section 1.4.1, as $d \to \infty$. The speed of this diffusion is then optimised in terms of the d-independent constant of proportionality in the scale parameter.

Roberts et al. (1997) investigate optimal-scaling on target densities of the form

$$\pi(\mathbf{x}) = \prod_{1}^{d} f(x_i) \tag{3.1}$$

using Gaussian jump proposals

$$\mathbf{Y}^{(d)} \sim N(\mathbf{0}, \sigma_d^2 \mathbf{I}_d) \tag{3.2}$$

For a random walk on a d dimensional target they define a speeded up (discrete) process which at time t is $Z_t^{(d)} = X_{1\ [td]}^{(d)}$, the first component of the chain after iteration $[td]$ (here $[x]$ denotes the largest integer less than or equal to x). It is shown that, subject to conditions on the first two derivatives of $f(\cdot)$, if the proposal standard deviation is chosen to be $\sigma_d = l/d^{1/2}$ (or $l/(d-1)^{1/2}$) for some $l > 0$, then

$Z^{(d)} \xrightarrow{D} Z$ where Z satisfies an SDE of the form

$$dZ_t = \frac{1}{2}h(l)\ \left(\log f\left(Z_t\right)\right)'\ dt + h(l)^{1/2}\ dB_t \tag{3.3}$$

with speed

$$h(l) := 2l^2\Phi\left(-\frac{1}{2}lI^{1/2}\right) \tag{3.4}$$

Here

$$I := \mathbb{E}\left[\left((\log f)'\right)^2\right] \tag{3.5}$$

is a measure of the roughness of the target and

$$2\Phi\left(-\frac{1}{2}lI^{1/2}\right) \tag{3.6}$$

corresponds to the acceptance rate.

Maximising the speed of this diffusion leads to setting $l = \frac{2.38}{I^{1/2}}$, so that the standard deviation is

$$\sigma_d = \frac{2.38}{I^{1/2}d^{1/2}}$$

This leads to an optimal acceptance rate of approximately 0.234; a value that is independent of $f(\cdot)$.

Roberts and Rosenthal (2001) also examine optimal-scaling, this time on "stretched" target distributions of the form

$$\prod_1^d C_i f(C_i x_i) \tag{3.7}$$

with the C_i sampled from some fixed distribution with $\mathbb{E}\left[C_i\right] = 1$, and using Gaussian jump proposals as in (3.2). In Theorem 5 of the paper $\sigma_d = l/d^{1/2}$ as in Roberts et al. (1997), and $W_t^{(d)} := C_1 X_{1[td]}^{(d)}$; this is the (scaleless) first component

of the Markov chain after iteration $[td]$. It is then shown that $W_t^{(d)}$ converges to a limiting diffusion process W_t satisfying an SDE of the form (3.3) with speed

$$h^*(l) = \frac{C_1^2}{b} \times h(lb^{1/2}) \tag{3.8}$$

with $h(\cdot)$ as defined in (3.4), I the same measure of the roughness of $f(\cdot)$ as in (3.5), and

$$b := \mathbb{E}\left[C_i^2\right]$$

Analogously with the result for i.i.d. components, the speed of the diffusion is maximised when $lb^{1/2} = 2.38/I^{1/2}$ and this still corresponds to an optimal acceptance rate of 0.234. Further, the overall speed at this optimum is the optimum speed for the i.i.d. target but multiplied by the factor C_1^2/b.

In the paper this is then compared with the diffusion that would arise if the target were of the form

$$\pi\left(\mathbf{x}\right) = \prod_1^d C\ f(Cx_i) \tag{3.9}$$

The relative efficiency between the "stretched" target and this target, if $C_1 = 1$ is given as $\mathbb{E}\left[C_i^2\right]/\mathbb{E}\left[C_i\right]^2$. However some confusion arises from this discussion because it is unclear exactly what is being compared, and because the target is referred to as having the form (3.1), i.e. setting $C = 1$. (Note that there is further confusion with the ensuing Theorem 6 since in this Theorem the $C_i's$ for an elliptical proposal are taken to be squared scale parameters rather than the inverse scale parameters that they denoted in Theorem 5). The following is intended to clarify the intent of the paper, addressing several general points before focusing on the above. In Section 3.4 some of these ideas will be tied in with new results derived in Section 3.3.2. Throughout this discussion the C_i's are inverse scale parameters.

We first note that the speed $(C_1^2/b^2) \times h(lb^{1/2})$ and efficiency factor C_1^2/b are those of the scaleless ("transformed") diffusion W_t. The stochastically identical diffusion (W_t/C_1) exploring the first component of the original target's space has speed $(1/b^2) \times h(lb^{1/2})$ and efficiency factor $1/b$ compared to the exploration of a target with i.i.d. components with unit scale parameters. If the expectation of the inverse scale parameters C_i were not constrained to 1 (for example each C_i was multiplied by a factor $\mathbb{E}[C_i]$) then the optimum speed of the diffusion W_t/C_1 exploring the original target would be multiplied by the factor $1/\mathbb{E}[C_i]^2$ but the speed of the scaleless diffusion W_t would be unchanged. **Therefore (3.8) holds even if** $\mathbb{E}[C_i] \neq 1$. Further, multiplying each C_i by a factor k will multiply $\mathbb{E}[C_i^2]$ by k^2, so the reduction factor b needs no adjustment to specifically account for changes in $\mathbb{E}[C_i]$. Hence the speed of the diffusion (W/C_1) exploring the original target remains $(1/b^2) \times h(lb^{1/2})$ even if $\mathbb{E}[C_i] \neq 1$.

Compare this with the special case (3.9) where $C_i = C \ \forall \ i$ and therefore $b = C^2$. Naturally the speed of the scaleless diffusion W_t reduces to (3.4), the same as that for i.i.d. target components (3.1). However the speed of the diffusion W_t/C exploring the original space is now

$$h^{**}(l) = \frac{1}{C^2} h(lC^{1/2})$$

Therefore the limiting ratio of optimum efficiencies *in the original space* for exploration of the first component of a "stretched" target with axes C_i and a target with all axes identically scaled by $\mathbb{E}[C_i]$ is

$$\text{rel.eff}_{\text{orig}} := \frac{\text{eff}_{\text{orig}}(\text{stretched target and spherical proposal})}{\text{eff}_{\text{orig}}(\text{i.i.d. target and spherical proposal})} = \frac{\mathbb{E}[C_i]^2}{\mathbb{E}[C_i^2]} \tag{3.10}$$

In the transformed (scaleless) space this ratio is given by (3.8)

$$\text{rel.eff}_{\text{trans}} := \frac{\text{eff}_{\text{trans}}\big(\text{stretched target and spherical proposal}\big)}{\text{eff}_{\text{trans}}\big(\text{i.i.d. target and spherical proposal}\big)} = \frac{C_1^2}{\mathbb{E}\left[C_i^2\right]} \tag{3.11}$$

If a "typical" component is considered, where $C_1 = \mathbb{E}\left[C_i\right]$ then the ratio (3.11) becomes the same as (3.10).

It is then noted that exploration of a stretched target using a similarly stretched Gaussian proposal is equivalent to exploring a non-stretched target (3.9) using a spherical Gaussian proposal. Therefore (3.11) still holds in this case.

Bedard (2006c) also considers targets of the form (3.7) but allows each $C_i(d)$ to be dependent on dimension such that the squared scale parameter of the i^{th} component $C_i(d)^{-2} = K_i/d^{\gamma_i}$. Note for this review, notation and the groupings of the components have been significantly altered and simplified from the original paper to be more consistent with notation in the rest of this Chapter. Let the number of the squared scale parameters proportional to $d^{-\gamma_i}$ be denoted $n_i(d)$, and let the number of distinct powers of d be $m < \infty$. The set of constants $\{K_j\}$ that correspond to components varying according to any one particular power $d^{-\gamma_j}$ are assumed to arise from a distribution satisfying $\mathbb{E}\left[K^{-2}\right] < \infty$. Further it is assumed (without loss of generality) that for each of these distributions, $\mathbb{E}\left[K^{-1/2}\right] = 1$; $\mathbb{E}\left[K^{-1}\right]$ is denoted b_i. Denote by α the smallest power such that as $d \to \infty$

$$\frac{n_i(d)\ d^{\gamma_i}}{d^{\alpha}} < \infty \quad \forall\ i$$

The sequence of transformed chains is considered

$$\mathbf{Z}_t^{(d)} := \left[X_{1\ [d^{\alpha}t]}^{(d)}, \ldots, X_{d\ [d^{\alpha}t]}^{(d)}\right]$$

using a spherical Gaussian proposal as in (3.2) with $\sigma_d = l/d^{\alpha/2}$ for some fixed l. It is shown that $\mathbf{Z}^{(d)}$ approaches a limiting diffusion $\mathbf{Z}_t$ whose i^{th} component Z_i satisfies an SDE of the form (3.3) with speed

$$h(l) = 2l^2\Phi\left(-\frac{1}{2}lE^{1/2}\right)$$

where the weighted roughness

$$E = I \times \lim_{d\to\infty} \sum_1^m \frac{b_i\ n_i(d)\ d^{\gamma_i}}{d^\alpha}$$

is proportional to I, the roughness measure defined in (3.5). However this diffusion limit is shown to hold if and only if

$$\lim_{d\to\infty} \frac{d^\lambda}{\sum_1^d d^{\gamma_i}} = 0 \tag{3.12}$$

where λ is the largest power of d that is repeated only finitely often.

It is noted that the simple hierarchical model $X_1 \sim N(0,1),\ X_i \sim N(X_1,1),\ i = 2,\ldots,d$ can be transformed to a Gaussian target with independent components and variances of $O(d)$, $O(1/d)$, and of $O(1)$ with multiplicity $d-2$. Such a target fails to satisfy the necessary and sufficient condition (3.12) for the theorem, and so 0.234 may not be the optimal limiting acceptance rate in this case. Bedard (2006b) investigates more general targets of the form (3.7) with $C_i(d)$ dependent on dimension and which fail to satisfy (3.12). It is found that the optimal scaling must be $\sigma^2(d) = l^2/d^\lambda$ where, as above, λ is the largest power of d repeated only finitely often. A limiting process is found that is a Langevin diffusion on components corresponding to the smaller powers γ (i.e. those with relatively stretched axes), and a discrete Metropolis-Hastings accept reject step on each of the remaining components. In these cases it is found that the asymptotically optimal acceptance rate

is no longer necessarily 0.234. Bedard (2006a) investigates the application of these results in several standard settings such as the Normal hierarchichal model and a variance components model.

Neal and Roberts (2006) again consider target densities of the form (3.1) and spherical Gaussian proposals with standard deviation $\sigma_d = l_c/d^{1/2}$, but use the random walk Metropolis algorithm with partial blocking (also known as "random walk Metropolis within Gibbs"; see Section 1.3.1.2). Components to update are chosen by random scan rather than sequentially. At each iteration the proposed jump is along only a subset of all the d components. This subset is of size dc_d (with $c_d \leq 1$ and $c_d \to c$) and is chosen freshly at random each iteration. It is shown (again subject to differentiability conditions on $f(\cdot)$) that the process $U_t^{(d)} := X_{1\ [dt]}^{(d)}$ converges to a diffusion U_t again satisfying an SDE of the form (3.3), with speed

$$h_c(l) = 2cl_c^2\Phi\left(-\frac{1}{2}l_c(cI)^{1/2}\right) \tag{3.13}$$

This has the same form as (3.4) but with $c\ l_c^2$ replacing l^2. Thus the optimal scaling is

$$\sigma_d = \frac{2.38}{c^{1/2}I^{1/2}d^{1/2}}$$

and this again corresponds to an optimal acceptance rate of 0.234. Most importantly however, it is clear from (3.13) that the optimal speed does not depend on c. Thus there is no advantage in using large block updates and possibly some slight disadvantage since these are generally computationally more expensive. This is contrary to the generally held intuition that "block-updating improves MCMC mixing". It is accepted that partial updates on a target with independent components might have special properties not shared by partial updates on more general

targets so an alternative target is considered.

$$\mathbf{X}^{(d)} \sim N(\mathbf{0}, \mathbf{\Sigma}^{(d)}(\rho))$$

where $\Sigma_{ii}^{(d)}(\rho) = 1$ and $\Sigma_{ij}^{(d)}(\rho) = \rho(j \neq i)$ with $0 < \rho < 1$. A limiting diffusion is found for which the optimal scaling (and hence the optimal speed) is reduced by a factor $(1-\rho)^{1/2}$

$$\sigma_d = \frac{2.38(1-\rho)^{1/2}}{c^{1/2} I^{1/2} d^{1/2}}$$

However the limiting optimal acceptance rate is once more 0.234 and once again the optimal speed is independent of c.

To test empirically for even more generality, a simulation study is detailed on three different targets: the $N\left(\mathbf{0}, \mathbf{\Sigma}^{(d)}(\rho)\right)$ distribution for which the theoretical results have been shown to hold, as well as $t_{50}\left(\mathbf{0}, \mathbf{\Sigma}^{(d)}(\rho)\right)$ and a target with independent components that follow a double-exponential $(\exp(-|x_i|))$. Tests are carried out for different values of ρ, c and d, and efficiency is measured by square jumping distance along the first component of the chain (multiplied by the normalising factor $d/(1-\rho)$). It is found to be remarkably consistent for each target across all the c, d and ρ values studied.

Through this same general approach an optimal acceptance rate of 0.234 has also been shown to hold for the RWM as applied to

- a spatially homogeneous Gibbs distribution (Markov random field) where the correlations decay at least exponentially quickly with distance (Breyer and Roberts, 2000)
- a discrete target with i.i.d. components each having mass p at the origin

and $1-p$ at 1. Here a fixed fraction f of the components are updated each iteration, with the components themselves chosen at random each time, and with f acting in place of a scale parameter (Roberts, 1998).

These two results are also summarised in Roberts and Rosenthal (2001).

3.1.2 Motivation for this chapter

The above results are all asymptotic and apply to exploration of a single component of certain specific classes of target distributions. Several questions arise immediately:

1. Real problems are finite dimensional. In such finite dimensional problems is there always an optimal scale parameter?

2. There is clearly a one-to-one mapping between the scale parameter and the form (3.6) corresponding to the acceptance rate that arises from the limiting diffusion. This justifies the use of one as a proxy for the other. In (real) finite dimensional problems is there always a one-to-one mapping between acceptance rate and scale parameter?

3. Are there further classes of distributions for which the limiting optimal acceptance rate is 0.234 ? If so, how does the optimal scale parameter behave in these cases and how does the use of partial blocking affect efficiency?

4. Are there classes of distributions for which the limiting optimal acceptance rate is not 0.234? If so, is it possible to characterise them?

5. The results reviewed in Section 3.1.1 optimise the scaling parameter for the limiting process along a single component. Does taking the infinite dimen-

sional limit of the optimal scale parameters for a sequence of finite dimensional processes lead to the same results? Does considering all components at once affect this?

The theory and ideas presented in this chapter go some way towards addressing each of the above questions.

3.1.3 A new approach

We wish to analyse expectations of acceptance rate and of some measure of efficiency of a RWM algorithm when the Markov chain is stationary. As will be discussed in more detail in Section 3.1.4 we take the expected square jumping distance (ESJD) as our measure of efficiency. In general these expectations (ESJD and acceptance rate) are sums of integrals across four different regions of the product space ($\Re^d \times \Re^d$) where $\Re^d$ is the state-space of the Markov chain. Through two "exchangeability lemmas" we reduce these sums to a single term (see Section 3.2).

We then investigate ESJD and expected acceptance rate for the random walk Metropolis algorithm on spherical (Section 3.3.1) and then elliptical (Section 3.3.2) target distributions with Lesbegue density monotonically decreasing from the origin. Our key successes are

(i) Exact forms for the acceptance rate and ESJD in finite dimension d, in terms of simple expectations that may be evaluated numerically, and help to address Questions 1 and 2 above.

(ii) Limiting results as dimension $d \to \infty$ that address Questions 3, 4 and 5 above, extending the class of target distributions for which the asymptotically

optimal acceptance rate is known to be 0.234 and providing classes for which different limiting optimal acceptance rates apply. Conditions on the limit theorems give valuable insight into when and why the different results apply.

We also examine the effects of partial blocking on spherical and elliptical targets (Section 3.3.3) in the limit as $d \to \infty$. Then in Section 3.3.4, through exact forms for expected acceptance rate and ESJD we examine specific combinations of target and proposal in finite dimensions, and compare behaviour with both finite dimensional and limiting theoretical results. A final simulation study points out fundamental difficulties with the idea of a single optimal scaling for targets which vary on at least two radically different scales.

3.1.4 Elliptically symmetric distributions and expected square jump distance as a measure of efficiency

The most general target distributions that we shall examine in this Chapter possess elliptical symmetry. If a d-dimensional target distribution has elliptical contours then there is a simple invertible linear transformation $\mathbf{T} : \Re^d \to \Re^d$, consisting of stretching along orthogonal principal components, which produces a spherically symmetric target. Of course $\mathbf{T}$ is not unique; to fix it (up to an arbitrary rotation) we define $\mathbf{T}$ to be the transformation that produces a spherically symmetric target with unit scale parameter. Here the exact meaning of "unit scale parameter" may be decided arbitrarily or by convention. The scale parameter β_i along the i^{th} principal axis of the ellipse is the i^{th} eigenvalue of $\mathbf{T}^{-1}$ since $\mathbf{T}^{-1}$ maps the spherically symmetric unit target to the elliptical target under consideration.

Let $\mathbf{X}$ and $\mathbf{X}'$ be consecutive elements of a stationary chain exploring a d-dimensional

target distribution. We wish to express the efficiency of the chain in terms of some amalgamation of the expected distances that it will move along each principal axis. A naive measure would be

$$S^2_{d,\text{ naive}} := \mathbb{E}\left[|\mathbf{X}' - \mathbf{X}|^2\right] \tag{3.14}$$

where expectation is with respect to the joint law of consecutive elements $\mathbf{X}$ and $\mathbf{X}'$ at stationarity.

We argue that this is not the most natural measure of efficiency for an elliptical target. Consider for simplicity a two dimensional ellipse with the target distribution "stretched" along x_1 and "squashed" along x_2. An efficient scheme would optimise exploration of the whole target and so require larger jumps along the x_1 axis and smaller jumps along the x_2 axis. In general we would like the relative sizes of the jumps along i^{th} principle axis to be proportional to the relative spacing of the contours along that axis, which is in turn proportional to the scale parameter β_i. An alternative perspective is attained by considering the transformed target $\mathbf{T}(\mathbf{X})$; this *is* spherically symmetric and so in the transformed space we do wish to give equal weight to equal size jumps along any axis.

Either of the above consideration leads to the following definition of the expected square jump distance (ESJD) for an elliptical target:

$$S^2_d := \mathbb{E}\left[||\mathbf{X}' - \mathbf{X}||^2_\beta\right] := \mathbb{E}\left[\sum_1^d \frac{1}{\beta_i^2}\left(X'_i - X_i\right)^2\right] \tag{3.15}$$

where X'_i and X_i are the components of $\mathbf{X}'$ and $\mathbf{X}$ along the i^{th} principal axis. For a spherical target $\beta_i = \beta \;\forall\; i$ and the ESJD is proportional to the naive definition (3.14). Later, when considering spherically symmetric targets, we therefore simply

optimise this naive definition, or equivalently set $\beta = 1$. It is possible to define a more general ESJD, provided the target has a finite covariance $\mathbf{\Sigma}$,

$$S_d^2 := \mathbb{E}\left[(\mathbf{X}' - \mathbf{X})^t \, \mathbf{\Sigma}^{-1} \, (\mathbf{X}' - \mathbf{X})\right] \tag{3.16}$$

For an elliptical target with finite covariance matrix, definitions (3.15) and (3.16) differ only by a constant of proportionality since with respect to the principal axes $\mathbf{\Sigma} \propto diag(\beta_1^2, \ldots, \beta_d^2)$.

We shall be concerned with maximising the ESJD rather than examining a single component and maximising the speed of a limiting diffusion. This shift of emphasis compared to the current literature is driven by what is feasibly achievable through the new theory that we develop. That the ESJD takes into account the efficiencies along all components has both advantages and disadvantages. However ESJD is often a reasonable measure of overall efficiency as we now discuss.

We first provide a simple relationship between ESJD along a single component, and lag-1 autocorrelation at stationarity (when the variance is finite). Define $\sigma_i^2 := \mathrm{Var}\left[X_i\right] = \mathrm{Var}\left[X_i'\right]$, and note that $\mathbb{E}\left[X_i' - X_i\right] = 0$, so

$$\mathbb{E}\left[(X_i' - X_i)^2\right] = \mathrm{Var}\left[X_i' - X_i\right] = 2\sigma_i^2\left(1 - \mathrm{Corr}\left[X_i, X_i'\right]\right) = 2\sigma^2\beta_i^2\left(1 - \mathrm{Corr}\left[X_i, X_i'\right]\right)$$

where σ^2 is the variance along any component of the target with unit scale parameter. Thus maximising ESJD along any component is in fact equivalent to minimising the lag-1 autocorrelation of that component. Similarly for an elliptical target the full ESJD is

$$S_d^2 = \sigma^2\left(d - \sum_1^d \mathrm{Corr}\left[X_i, X_i'\right]\right)$$

so maximising the ESJD is equivalent to minimising the sum of the correlations of individual principal components.

A common practical measure of the efficiency of an MCMC run is the integrated auto-correlation time along one or more components (see Section 1.3.2). This is determined not only by the lag-1 autocorrelation but by the expected sum of all the lagged autocorrelations and relates directly to the variance of Monte-Carlo estimates. Even along a single component, maximising ESJD is not necessarily equivalent to minimising the ACT. However, there are several problems with the integrated ACT in theory and practice.

- The ACT of X and the ACT of some function $f(X)$ may behave very differently and be optimised by quite different scaling parameters.
- An integrated ACT estimated from a real chain is subject to noise that increases as the number of lags included in the sum increases.
- Even at low lags the theoretical expected ACT and the observed ACT may differ radically. Consider for example an (almost) irreducible one-dimensional algorithm run on the uniform distribution over $[-2,-1] \cup [1,2]$. In practice only one half of the space will be explored and for an efficient chain the auto-correlations will decrease quickly with lag. However theoretical auto-correlations at stationarity will be large and positive even for very high lags.

ESJD does not suffer from the second and third problems in the list. The first problem (which also affects ESJD) arises from the fact that different functions $f(\cdot)$ lead to different sequences $f(\mathbf{X}_i)$, so that the accuracy of a Monte Carlo estimate of $\mathbb{E}[f(\mathbf{X})]$ depends on $f(\cdot)$. The ACT for any function of the chain can be bounded

above using the geometric rate of convergence of the chain (e.g. Gilks et al., 1996, Chapter 3). But this rate is rarely known and in any case only supplies an upper bound: there is no easy to obtain best measure of efficiency for the chain. However, a further justification for the use of ESJD is expanded upon below: that (at least for certain targets) in the limit as $d \to \infty$ the scale parameter that maximises the ESJD also maximises the integrated ACT and does not depend on the function $f(\cdot)$ (provided $f(\cdot)$ is differentiable).

As discussed in Section 3.1.1, for several forms of target with independent components, which are identical up to a scale parameter, the limit of a single component of the speeded-up Markov chain becomes a Langevin diffusion with speed h. As was discussed in Section 1.4.2, minimising the integrated ACT is equivalent to maximising the speed h. However as the limiting diffusion is approached, the ESJD along any component is more and more closely approximated by a small increment in the diffusion along that component. From (1.11)

$$\mathbb{E}\left[|\Delta X_t|^2\right] \approx h^2\mu^2(\Delta t)^2 + h\Delta t \approx h\Delta t$$

where both approximations become exact as $\Delta t \to 0$. Therefore maximising the ESJD along a single component is also equivalent to maximising the speed of the diffusion, h. Similarly maximising the total ESJD is equivalent to maximising the sum of the speeds of the diffusions over all components.

Investigations in this chapter will concern spherically and elliptically symmetric targets, and the only target with such symmetries covered in Roberts and Rosenthal (2001) is the Gaussian. Nevertheless it seems plausible that the general principle of speeding up time to produce limiting diffusion processes will hold. Further evidence

for this intuition arises from the description in Section 3.4.1 and is discussed briefly in the context of further work in Section 3.4.3.

3.2 Region exchangeability and some consequences

In this section we present two main Exchangeability Lemmas with extensions. The lemmas are applicable to any "sensible" Metropolis-Hastings algorithm at stationarity on (almost) any target with Lesbegue density. They apply to expectations of certain functionals of the Metropolis-Hastings Markov chain, including expected square jumping distance and acceptance rate, and lead to a simplification of their closed form.

3.2.1 Definitions and assumptions

The general form of the Metropolis-Hastings algorithm was described in Section 1.3.1.1. As in that introductory section we will only be concerned with an element of the Metropolis-Hastings Markov chain at stationarity and the element immediately following it. We therefore define the current instance of the chain $\mathbf{X} := \mathbf{X}_m$, the proposed next instance of the chain $\mathbf{X}^* := \mathbf{X}^*_{m+1}$, and the actual next instance of the chain $\mathbf{X}' := \mathbf{X}_{m+1}$. We also define the proposed jump $\mathbf{Y}^* := \mathbf{X}^*_{m+1} - \mathbf{X}_m$ and the actual jump $\mathbf{Y} := \mathbf{X}_{m+1} - \mathbf{X}_m$. The acceptance rate $\alpha(\mathbf{x}, \mathbf{x}^*)$ has the form given in (1.2).

The target distribution is assumed to posses a Lesbegue density $\pi(\cdot)$ and the proposal to posses Lesbegue density $q(\mathbf{x}^*|\mathbf{x})$. It is assumed that the chain has reached stationarity, so that the marginal distributions of both $\mathbf{X}$ and $\mathbf{X}'$ are $\pi(\cdot)$. The

law for $\mathbf{X}'$ given $\mathbf{X} = \mathbf{x}$ is denoted $P(d\mathbf{x}'|\mathbf{x})$ and the joint law of two successive elements is $\pi(\mathbf{x})d\mathbf{x}\ P(d\mathbf{x}'|\mathbf{x})$ which we denote

$$A(d\mathbf{x}, d\mathbf{x}') := \pi(\mathbf{x})d\mathbf{x}\ q(\mathbf{x}'|\mathbf{x})\,\alpha(\mathbf{x},\mathbf{x}')\ \mathbf{1}_{\{\mathbf{x}'\neq\mathbf{x}\}}\ d\mathbf{x}' + \pi(\mathbf{x})d\mathbf{x} \int d\mathbf{x}^*\ q(\mathbf{x}^*|\mathbf{x})\,(1-\alpha(\mathbf{x},\mathbf{x}^*))\ \mathbf{1}_{\{\mathbf{x}'=\mathbf{x}\}} \quad (3.17)$$

We assume that the space of possible values for element $\mathbf{x}$ of a d-dimensional chain is $\Re^d$, and partition the space of possible values for $\mathbf{x}^*$ (and so for $\mathbf{x}'$) given $\mathbf{x}$ into the following four disjoint regions:

the identity region $R_{id}(\mathbf{x}) := \{\mathbf{x}\}$; this is a null set under $q(\cdot|\mathbf{x})$, but is in general not null under $P(\cdot|\mathbf{x})$.

the equality region $R_{eq}(\mathbf{x}) := \{\mathbf{x}' \in \Re^d : \mathbf{x}' \notin R_{id}(\mathbf{x}), \frac{\pi(\mathbf{x}')q(\mathbf{x}|\mathbf{x}')}{\pi(\mathbf{x})q(\mathbf{x}'|\mathbf{x})} = 1\}$; we assume initally that $R_{eq}(\mathbf{x})$ is null under $q(\cdot|\mathbf{x})$ (and therefore under $P(\cdot|\mathbf{x})$) for all $\mathbf{x}$, but later relax this assumption for certain $q(\cdot|\cdot)$.

the acceptance region $R_a(\mathbf{x}) := \{\mathbf{x}' \in \Re^d : \alpha(\mathbf{x},\mathbf{x}') = 1,\ \mathbf{x}' \notin \{R_{eq}(\mathbf{x}) \cup R_{id}(\mathbf{x})\}\}$; this is the remainder of the region where we are guaranteed to accept the proposal.

the rejection region $R_r(\mathbf{x}) := \{\mathbf{x}' \in \Re^d : \alpha(\mathbf{x},\mathbf{x}') < 1\}$; this is the region where there is a positive probability that we will reject the proposal.

For vectors $(\mathbf{x}, \mathbf{x}')$ in $\Re^d \times \Re^d$ we employ the shorthand

$$\begin{aligned} R_{ID}(\mathbf{x},\mathbf{x}') &:= \{(\mathbf{x},\mathbf{x}') : \mathbf{x} \in \Re^d, \mathbf{x}' \in R_{id}(\mathbf{x})\} \\ R_{EQ}(\mathbf{x},\mathbf{x}') &:= \{(\mathbf{x},\mathbf{x}') : \mathbf{x} \in \Re^d, \mathbf{x}' \in R_{eq}(\mathbf{x})\} \\ R_{A}(\mathbf{x},\mathbf{x}') &:= \{(\mathbf{x},\mathbf{x}') : \mathbf{x} \in \Re^d, \mathbf{x}' \in R_{a}(\mathbf{x})\} \\ R_{R}(\mathbf{x},\mathbf{x}') &:= \{(\mathbf{x},\mathbf{x}') : \mathbf{x} \in \Re^d, \mathbf{x}' \in R_{r}(\mathbf{x})\} \end{aligned}$$

3.2.2 Exchangeability, ESJD and expected acceptance rate

An exchangeability between the regions $R_a(\cdot)$ and $R_r(\cdot)$ follows directly from their definitions

$$\mathbf{x}' \in R_a(\mathbf{x}) \Leftrightarrow \mathbf{x} \in R_r(\mathbf{x}')$$

so that

$$R_A(\mathbf{x}, \mathbf{x}') = R_R(\mathbf{x}', \mathbf{x})$$

and vice versa. For example in the context of the RWM, if a proposed jump from $\mathbf{x}$ to $\mathbf{x}+\mathbf{y}^*$ sees a reduction in the stationary density, and therefore a probability of rejection, the reverse jump would lead to an increase in the density and therefore guaranteed acceptance.

This leads to the main results of this section: two Exchangeability Lemmas (Lemmas 1 and 2), which apply to any Metropolis-Hastings algorithm that uses a single block update, subject to minor conditions on the target distribution. These conditions may be relaxed in the case of symmetric proposals and this is examined in Section 3.2.3. Equivalent lemmas apply when components are updated in several blocks rather than all at once; these extensions are derived in Section 3.2.4. The Exchangeability Lemmas provide a simpler form for the ESJD and expected acceptance rate than might be naively obtained though simply plugging in the joint law (3.17); they form the basis of all our subsequent work. In all of the lemmas, $\mathbf{X}$ and $\mathbf{X}'$ are consecutive elements of *any* chain produced from a Metropolis-Hastings algorithm which has reached stationarity, subject to conditions on the target and proposal distributions to be specified.

Some of the lemmas apply to expectations of acceptance rates; these expectations

are with respect to the joint law of current position and proposed move:

$$A^*(d\mathbf{x}, d\mathbf{x}^*) := \pi(\mathbf{x})q\left(\mathbf{x}^*|\mathbf{x}\right)\ d\mathbf{x}\ d\mathbf{x}^* \tag{3.18}$$

Others apply to classes of functions $h(\mathbf{X}, \mathbf{X}')$ which satisfy the symmetry condition

$$h(\mathbf{x}, \mathbf{x}') = c \times h(\mathbf{x}', \mathbf{x}) \quad \forall\ \mathbf{x}, \mathbf{x}' \ \ (with\ c = \pm 1) \tag{3.19}$$

and (potentially) the further condition

$$h(\mathbf{x}, \mathbf{x}) = 0 \quad \forall\ \mathbf{x} \tag{3.20}$$

The equality region is also required to be null

$$\int_{R_{eq}(\mathbf{x})} d\mathbf{x}'\ q(\mathbf{x}'|\mathbf{x}) = 0 \quad \forall\ \mathbf{x} \tag{3.21}$$

Note that (3.21) implies both

$$\int_{R_{EQ}} d\mathbf{x}\ d\mathbf{x}'\ \pi(\mathbf{x})q(\mathbf{x}'|\mathbf{x}) = 0$$

and

$$\int_{R_{EQ}} d\mathbf{x}\ d\mathbf{x}'\ \pi(\mathbf{x})q(\mathbf{x}'|\mathbf{x})h(\mathbf{x}, \mathbf{x}') = 0$$

In calculating the expectation of functions $h(\mathbf{X}, \mathbf{X}')$ we must use the joint law (3.17). The first lemma allows us, to reduce such expectations to a single simple integral over the acceptance region, the second allows the same simplification for the expected acceptance rate.

Lemma 1 *Consider two consecutive elements,* $\mathbf{X}$ *and* $\mathbf{X}'$*, of some Metropolis-Hastings Markov chain with stationary Lesbegue density* $\pi(\cdot)$*, proposal Lesbegue density* $q\left(\mathbf{x}'|\mathbf{x}\right)$ *and joint law* $A(d\mathbf{x}, d\mathbf{x}')$*. At stationarity, for any function* $h(\cdot, \cdot)$ *satisfying (3.19)*

$$\int_{(\mathbf{x},\mathbf{x}')\in R_A} A(d\mathbf{x}, d\mathbf{x}')\ h(\mathbf{x}, \mathbf{x}') = c \times \int_{(\mathbf{x},\mathbf{x}')\in R_R} A(d\mathbf{x}, d\mathbf{x}')\ h(\mathbf{x}, \mathbf{x}')$$

If in addition (3.20) and (3.21) hold then

$$\mathbb{E}\left[h(\mathbf{X},\mathbf{X}')\right] = (1+c) \times \int_{(\mathbf{x},\mathbf{x}')\in R_A} d\mathbf{x}\; d\mathbf{x}'\; \pi(\mathbf{x})\; q\left(\mathbf{x}'|\mathbf{x}\right)\; h(\mathbf{x},\mathbf{x}')$$

Proof: *The first result is a consequence of region exchangeability, reversibility, and the symmetry of $h(\cdot,\cdot)$ which we apply consecutively below, and then relabel:*

$$\begin{aligned}
\int_{(\mathbf{x},\mathbf{x}')\in R_A} A(d\mathbf{x},d\mathbf{x}')h(\mathbf{x},\mathbf{x}') &= \int_{(\mathbf{x}',\mathbf{x})\in R_R} A(d\mathbf{x},d\mathbf{x}')h(\mathbf{x},\mathbf{x}') \\
&= \int_{(\mathbf{x}',\mathbf{x})\in R_R} A(d\mathbf{x}',d\mathbf{x})h(\mathbf{x},\mathbf{x}') \\
&= c \times \int_{(\mathbf{x}',\mathbf{x})\in R_R} A(d\mathbf{x}',d\mathbf{x})h(\mathbf{x}',\mathbf{x}) \\
&= c \times \int_{(\mathbf{x},\mathbf{x}')\in R_R} A(d\mathbf{x},d\mathbf{x}')h(\mathbf{x},\mathbf{x}')
\end{aligned}$$

Equation (3.20) states that $h(\mathbf{x},\mathbf{x}') = 0$ in R_{ID}. Further, $\alpha(\mathbf{x},\mathbf{x}^*) = 1\ \forall\ (\mathbf{x},\mathbf{x}^*) \in R_{EQ}(\mathbf{x},\mathbf{x}^*)$, and (3.21) holds. Therefore

$$\int_{(\mathbf{x},\mathbf{x}')\in R_{ID}\cup R_{EQ}} A(d\mathbf{x},d\mathbf{x}')h(\mathbf{x},\mathbf{x}') = 0$$

and R_{ID} and R_{EQ} contribute nothing to the overall expectation of $h(\mathbf{X},\mathbf{X}')$. Since $\alpha(\mathbf{x},\mathbf{x}^*) = 1\ \forall\ (\mathbf{x},\mathbf{x}^*) \in R_A(\mathbf{x},\mathbf{x}^*)$ the second result follows from the first.

Setting $h(\mathbf{x},\mathbf{x}') = ||\mathbf{x}'-\mathbf{x}||_\beta^2 = \sum_1^d \frac{1}{\beta_i^2}\left(x_i'-x_i\right)^2$ in Lemma 1 leads to the following:

Corollary 1 *Consider any stationary Markov chain that has been produced by a Metropolis-Hastings algorithm with both target and proposal being Lesbegue densities, with the target in fact elliptical, and with equality region satisfying (3.21). The expected square jumping distance for the chain at stationarity is the same over the acceptance and rejection regions and equal to half the complete expectation, and therefore*

$$S_d^2 = 2\int_{(\mathbf{x},\mathbf{x}')\in R_A} \pi(\mathbf{x}) q\left(\mathbf{x}'|\mathbf{x}\right) ||\mathbf{x}'-\mathbf{x}||_\beta^2 \tag{3.22}$$

Clearly the Lemma also applies to the naive ESJD, $|\mathbf{x}' - \mathbf{x}|^2$. We now examine the acceptance rate $\alpha(\mathbf{x}, \mathbf{x}^*) = \min\left(1, \frac{\pi(\mathbf{x}^*)q(\mathbf{x}|\mathbf{x}^*)}{\pi(\mathbf{x})q(\mathbf{x}^*|\mathbf{x})}\right)$, using joint law (3.18) and again simplifying its expectation to a single term.

Lemma 2 *Let* $\mathbf{X}$ *be an element from some Metropolis-Hastings Markov chain with stationary Lesbegue density* $\pi(\cdot)$*, proposal Lesbegue density* $q(\mathbf{x}^*|\mathbf{x})$ *and acceptance rate* $\alpha(\mathbf{x}, \mathbf{x}^*)$*. At stationarity*

$$\int_{(\mathbf{x},\mathbf{x}^*)\in R_A} A^*(d\mathbf{x}, d\mathbf{x}^*)\ \alpha(\mathbf{x}, \mathbf{x}^*) = \int_{(\mathbf{x},\mathbf{x}^*)\in R_R} A^*(d\mathbf{x}, d\mathbf{x}^*)\ \alpha(\mathbf{x}, \mathbf{x}^*)$$

If in addition (3.21) holds then the overall expected acceptance rate is given by

$$\mathbb{E}\left[\alpha(\mathbf{X}, \mathbf{X}^*)\right] = 2\int_{(\mathbf{x},\mathbf{x}^*)\in R_A} d\mathbf{x}\ d\mathbf{x}^*\ \pi(\mathbf{x})q\left(\mathbf{x}^*|\mathbf{x}\right)$$

Proof: *After simplifying the expression we apply region exchangeability, relabel, and note that* $\alpha(\mathbf{x}, \mathbf{x}^*) = 1\ \ \forall\ (\mathbf{x}, \mathbf{x}^*) \in R_A$*.*

$$\begin{aligned}
\int_{(\mathbf{x},\mathbf{x}^*)\in R_R} A^*(d\mathbf{x}, d\mathbf{x}^*)\ \alpha(\mathbf{x}, \mathbf{x}^*) &= \int_{(\mathbf{x},\mathbf{x}^*)\in R_R} d\mathbf{x}d\mathbf{x}^*\pi(\mathbf{x})q\left(\mathbf{x}^*|\mathbf{x}\right) \times \frac{\pi(\mathbf{x}^*)q\left(\mathbf{x}|\mathbf{x}^*\right)}{\pi(\mathbf{x})q\left(\mathbf{x}^*|\mathbf{x}\right)} \\
&= \int_{(\mathbf{x},\mathbf{x}^*)\in R_R} d\mathbf{x}d\mathbf{x}^*\pi(\mathbf{x}^*)q\left(\mathbf{x}|\mathbf{x}^*\right) \\
&= \int_{(\mathbf{x}^*,\mathbf{x})\in R_A} d\mathbf{x}d\mathbf{x}^*\pi(\mathbf{x}^*)q\left(\mathbf{x}|\mathbf{x}^*\right) \\
&= \int_{(\mathbf{x},\mathbf{x}^*)\in R_A} d\mathbf{x}d\mathbf{x}^*\pi(\mathbf{x})q\left(\mathbf{x}^*|\mathbf{x}\right) \\
&= \int_{(\mathbf{x},\mathbf{x}^*)\in R_A} A^*(d\mathbf{x}, d\mathbf{x}^*)\ \alpha(\mathbf{x}, \mathbf{x}^*)
\end{aligned}$$

which proves the first part of the lemma. Since $q(\cdot,\cdot)$ *is a density,* $R_{id}(\mathbf{x})$ *is null with respect to* $q(\cdot|\mathbf{x})$*. Further, (3.21) holds so both* R_{EQ} *and* R_{ID} *are null with respect to* $\pi(\mathbf{x})q\left(\mathbf{x}^*|d\mathbf{x}\right)$*. The second part of the lemma then follows immediately as the acceptance rate in* R_A *is* 1*.*

Lemmas 1 and 2 hold for stationary chains only and thus suggest a number of possible tests for stationarity. These are discussed briefly in Section 3.4.3 as possibilities for further work but are not the main focus of this Chapter.

3.2.3 Extension for symmetric proposals

If the proposal distribution is symmetric (i.e. $q(\mathbf{x}^*|\mathbf{x}) = q(\mathbf{x}|\mathbf{x}^*)$) then we may extend Lemmas 1 and 2 to deal with cases where R_{EQ} is not null.

With a symmetric proposal, the acceptance probability becomes

$$\alpha(\mathbf{x}, \mathbf{x}^*) = \min(1, \pi(\mathbf{x}^*)/\pi(\mathbf{x}))$$

and $R_{EQ}(\mathbf{x}, \mathbf{x}^*) = \{(\mathbf{x}, \mathbf{x}^*) : \pi(\mathbf{x}) = \pi(\mathbf{x}^*)\}$.

With each $\mathbf{x} \in \Re^d$ we associate an equivalence class $C^*(\mathbf{x}) = \{\mathbf{x}^* : \pi(\mathbf{x}) = \pi(\mathbf{x}^*)\}$, and a portion of the equality region R_{EQ}:

$$C(\mathbf{x}) = (C^*(\mathbf{x}) \times C^*(\mathbf{x})) \cap (\Re^d \times \Re^d \setminus R_{ID})$$

Consider those disjoint sets defined by $C(\cdot)$ that posses non-zero Lesbegue measure. At most a finite number of these can share any given non-zero measure since the total measure is less than or equal to 1. Also as each has non-zero measure the differing measures may be ordered. Combining these two ideas we see that there are a countable number of such classes, which we denote C_i. We then partition R_{EQ} into the union of a null set and

$$C_1 \cup C_2 \cup \ldots$$

Each of the product spaces C_i is then partitioned into product spaces $C_i^{(1)}$ and $C_i^{(2)}$ via the following rule: for any distinct unordered pair $(\mathbf{x}, \mathbf{x}^*)$ arbitrarily assign one ordered couplet $(\mathbf{x}, \mathbf{x}^*)$ to $C_i^{(1)}$; assign the other ordered couplet $(\mathbf{x}^*, \mathbf{x})$ to $C_i^{(2)}$. These product spaces are then exchangeable in the same sense as the acceptance and rejection regions are exchangeable:

$$(\mathbf{x}, \mathbf{x}^*) \in C_i^{(1)} \Leftrightarrow (\mathbf{x}^*, \mathbf{x}) \in C_i^{(2)}$$

The following may then be proved in a similar manner to Lemmas 1 and 2:

Lemma 3 *Define $C_i^{(b)}$ as above and let $\mathbf{X}$ and $\mathbf{X}'$ be two consecutive elements of any stationary Metropolis-Hastings Markov chain on Lesbegue target density $\pi(\cdot)$ using **symmetric** Lesbegue proposal density $q(\mathbf{x}'|\mathbf{x})$. For any scalar function $h(\mathbf{x}, \mathbf{x}')$ satisfying (3.19) and (3.20)*

$$\mathbb{E}\left[h(\mathbf{X}, \mathbf{X}')\right] = (1+c) \times \int_{(\mathbf{x},\mathbf{x}') \in R_A \cup C_1^{(1)} \cup C_2^{(1)} \cup \ldots} d\mathbf{x} d\mathbf{x}' \ \pi(\mathbf{x}) q\left(\mathbf{x}'|\mathbf{x}\right) \ h(\mathbf{x}, \mathbf{x}')$$

and the expected acceptance rate of proposals $\mathbf{X}^*$ satisfies

$$\mathbb{E}\left[\alpha(\mathbf{X}, \mathbf{X}^*)\right] = 2 \times \int_{(\mathbf{x},\mathbf{x}^*) \in R_A \cup C_1^{(1)} \cup C_2^{(1)} \cup \ldots} d\mathbf{x} d\mathbf{x}' \ \pi(\mathbf{x}) q\left(\mathbf{x}'|\mathbf{x}\right)$$

3.2.4 Extension for partial blocking

We now consider the effect of updating components separately using several sub-blocks rather than a single block. Partition the complete space into k sub-spaces: $\Re^d = \mathcal{E}_1 \oplus \cdots \oplus \mathcal{E}_k$ and update $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_k)$ to $\mathbf{x}' = (\mathbf{x}'_1, \ldots, \mathbf{x}'_k)$ via k sub-blocks using k separate proposal Lesbegue densities with an accept/reject stage after each. As in Section 1.3.1.2 we define

$$\mathbf{x}_{-i} := \mathbf{x}'_1, \ldots, \mathbf{x}'_{i-1}, \mathbf{x}_{i+1}, \ldots, \mathbf{x}_k \tag{3.23}$$

so that the proposal for the i^{th} sub-block is

$$q_i\left(\mathbf{x}_i^*|\mathbf{x}_i, \mathbf{x}_{-i}\right) \tag{3.24}$$

Also define

$$\Re_{-i}^d = \mathcal{E}_1 \oplus \cdots \oplus \mathcal{E}_{i-1} \oplus \mathcal{E}_{i+1} \cdots \oplus \mathcal{E}_k$$

Denote the acceptance, rejection, identity and equality regions for the i^{th} update as in Section 3.2.1 but with the subscript i. The joint law $A(d\mathbf{x}, d\mathbf{x}')$ may be decomposed into the product of the conditional joint laws $A_i(d\mathbf{x}_i, d\mathbf{x}_i'|x_{-i})$ within each block:

$$\begin{aligned} A(d\mathbf{x}, d\mathbf{x}') &= A((d\mathbf{x}_1, \ldots, d\mathbf{x}_k), (d\mathbf{x}_1', \ldots, d\mathbf{x}_k')) \\ &= A_1(d\mathbf{x}_1, d\mathbf{x}_1'|d\mathbf{x}_{-1}) \ \ldots \ A_k(d\mathbf{x}_k, d\mathbf{x}_k'|d\mathbf{x}_{-k}) \\ &= \prod_1^k A_i(d\mathbf{x}_i, d\mathbf{x}_i'|d\mathbf{x}_{-i}) \end{aligned} \tag{3.25}$$

where the last line merely introduces a convenient shorthand for the individual conditional laws and does not indicate independence. We will also require the marginal law for each block

$$A_{-i}(d\mathbf{x}_i, d\mathbf{x}_i') := \int_{\Re_{-i}^d \times \Re_{-i}^d} d\mathbf{x}_1 d\mathbf{x}_1' \ldots, d\mathbf{x}_{i-1} d\mathbf{x}_{i-1}' d\mathbf{x}_{i+1} d\mathbf{x}_{i+1}' \ldots d\mathbf{x}_k d\mathbf{x}_k' \ A(d\mathbf{x}, d\mathbf{x}'))$$

Consider only functions $h(\cdot, \cdot)$ of the form

$$h(\mathbf{x}, \mathbf{x}') = \sum_1^k h_i(\mathbf{x}_i, \mathbf{x}_i') \tag{3.26}$$

This includes for example $h(\mathbf{x}, \mathbf{x}') = ||\mathbf{x} - \mathbf{x}'||_\beta^2$, provided each subspace $\mathcal{E}_i$ is the span of some subset of the principal axes. In these circumstances Lemma 1 may be extended:

Lemma 4 *Consider a Metropolis-Hastings Markov chain with stationary Lesbegue density that is updated using k partial blocks on orthogonal subspaces as defined in (3.24). Let $h(\cdot,\cdot)$ be a scalar function which decomposes according to (3.26), with each $h_i(\cdot,\cdot)$ satisfying (3.19) with the same c. Let the chain be stationary and let* $\mathbf{X}$ *and* $\mathbf{X'}$ *be two consecutive elements (after a complete set of partial updates) of the chain, with joint law $A(d\mathbf{x}, d\mathbf{x}')$ and marginal laws $A_{-i}(d\mathbf{x}_i, d\mathbf{x}'_i)$ for each of the partial blocks. Then*

$$\int_{(\mathbf{x}_i,\mathbf{x}'_i)\in R_{A_i}} A_{-i}(d\mathbf{x}, d\mathbf{x}')\ h_i(\mathbf{x}_i,\mathbf{x}'_i) = c \times \int_{(\mathbf{x}_i,\mathbf{x}'_i)\in R_{R_i}} A_{-i}(d\mathbf{x}, d\mathbf{x}')\ h_i(\mathbf{x}_i,\mathbf{x}'_i)$$

If in addition $h_i(\mathbf{x},\mathbf{x}) = 0 \ \ \forall i$ and furthermore each R_{eq_i} and corresponding $q_i(\cdot|\cdot)$ satisfy (3.21), then

$$\mathbb{E}\left[h(\mathbf{X},\mathbf{X}')\right] = (1+c) \times \sum_1^k \int_{(\mathbf{x}_i,\mathbf{x}'_i)\in R_{A_i}} A(d\mathbf{x}, d\mathbf{x}')\ h_i(\mathbf{x}_i,\mathbf{x}'_i)$$

Proof: *Once the chain has reached stationarity,* $\mathbf{X}$ *is still a draw from the stationary distribution after any of the partial updates. Also the acceptance probabilities for partial updates are chosen exactly so that the chain is reversible at stationarity across each partial update. Thus the conditional law for the i^{th} block is symmetric and hence so is the marginal law*

$$A_{-i}(d\mathbf{x}_i, d\mathbf{x}'_i) = A_{-i}(d\mathbf{x}'_i, d\mathbf{x}_i)$$

Acceptance and rejection regions for each partial update are chosen to be exchangeable and the first result then follows by applying the first part of Lemma 1 to each partial update. Next, combining (3.25) and (3.26) we obtain

$$\begin{aligned}\int_{\Re^d\times\Re^d} A(d\mathbf{x}, d\mathbf{x}')h(\mathbf{x},\mathbf{x}') &= \sum_{j=1}^k \int_{\Re^d\times\Re^d} \prod_{i=1}^k A_i(d\mathbf{x}_i, d\mathbf{x}'_i|\mathbf{x}_{-i})\ h_j(\mathbf{x}_j,\mathbf{x}'_j) \\ &= \sum_{j=1}^k \int_{\mathcal{E}_j\times\mathcal{E}_j} A_{-j}(d\mathbf{x}_j, d\mathbf{x}'_j)\ h_j(\mathbf{x}_j,\mathbf{x}'_j)\end{aligned}$$

The second result then follows by applying the second part of Lemma 1 to each partial update.

Setting $h(\mathbf{x}, \mathbf{x}') = ||\mathbf{x} - \mathbf{x}'||_\beta^2$ provides the following:

Corollary 2 *Consider any stationary Markov chain that has been produced by a Metropolis-Hastings algorithm on an elliptical target density with partial updates on spaces $\mathcal{E}_i$, each spanned by some subset of the principal axes, and where (3.21) holds for R_{eq_i} and corresponding proposal Lesbegue density $q_i(\cdot|\cdot)$. In this situation the integrated square jumping distance is the same over the acceptance and rejection regions and equal to half the complete expectation, and therefore*

$$\mathbb{E}\left[||\mathbf{X}' - \mathbf{X}||_\beta^2\right] = 2\sum_1^k \int_{(\mathbf{x}_i, \mathbf{x}_i') \in R_{A_i}} \pi(\mathbf{x}_i) q\left(\mathbf{x}_i'|\mathbf{x}_i\right) ||\mathbf{x}_i' - \mathbf{x}_i||_\beta^2 \tag{3.27}$$

Since each partial update is reversible at equilibrium and the acceptance and rejection regions are exchangeable Lemma 2 may also be applied to the acceptance probability for each partial update, leading to the following:

Lemma 5 *Consider a Metropolis-Hastings Markov chain that is updated using k partial blocks on orthogonal subspaces as defined in (3.24). Let $\mathbf{X}$ be any element from the chain at equilibrium; let $\mathbf{X}_i \quad (i = 1, \ldots, k)$ be the component along the i^{th} partial block after $i - 1$ further partial updates and let $\mathbf{X}_i^*$ be the proposed next partial update. Write the marginal law for each proposed partial update as $A_{-i}(d\mathbf{x}_i, d\mathbf{x}_i^*)$. At stationarity*

$$\int_{(\mathbf{x}_i, \mathbf{x}_i*) \in R_{A_i}} A_{-i}(d\mathbf{x}, d\mathbf{x}^*)\ \alpha_i(\mathbf{x}_i, \mathbf{x}_i^*) = \int_{(\mathbf{x}_i, \mathbf{x}_i^*) \in R_{R_i}} A_{-i}(d\mathbf{x}, d\mathbf{x}^*)\ \alpha_i(\mathbf{x}_i, \mathbf{x}_i^*)$$

If in addition (3.21) holds for each R_{eq_i} and its corresponding $q_i(\cdot|\cdot) \quad \forall i$ then

$$\mathbb{E}\left[\alpha_i(\mathbf{X}, \mathbf{X}^*)\right] = 2 \times \int_{(\mathbf{x}_i, \mathbf{x}_i^*) \in R_{A_i}} A(d\mathbf{x}, d\mathbf{x}^*)\ \alpha_i(\mathbf{x}_i, \mathbf{x}_i^*)$$

3.3 The random walk Metropolis

We will start by defining the random walk Metropolis algorithm and restating the Exchangeability Lemmas and the definition of ESJD in terms specific to this algorithm. Next we consider optimal scaling of the random walk Metropolis on spherically symmetric targets in both finite dimensions and then in the limit as $d \to \infty$ (Section 3.3.1). We then generalise some of the results to elliptically symmetric targets (Section 3.3.2) and examine the effects of partial blocking on efficiency as dimension $d \to \infty$ (Section 3.3.3). In Section 3.3.4 exact analytical and computational results for the variation of expected acceptance rate and ESJD with scale parameter are compared with theory from the previous sections. A simulation study is also conducted on unimodal targets each of which varies on two radically different scales.

First consider a d-dimensional random walk Metropolis algorithm where the jump proposal distribution has an overall scale parameter λ. As in Section 1.3.1.1 we write

$$q\left(\mathbf{x}^*|\mathbf{x}\right) = \frac{1}{\lambda^d} r((\mathbf{x}^* - \mathbf{x})/\lambda) = \frac{1}{\lambda^d} r(\mathbf{y}^*/\lambda)$$

Here $r(\cdot)$ is the Lesbegue density function for a jump proposal with unit scale parameter. The exact meaning of "unit scale parameter" is arbitrary and may (for example) be taken from the conventional parametrisation if one exists. For the symmetric random walk we also specify that $r(\mathbf{y}) = r(-\mathbf{y})$.

Since $q\left(\mathbf{x}^*|\mathbf{x}\right) = q\left(\mathbf{x}|\mathbf{x}^*\right)$ the acceptance probability simplifies to

$$\alpha\left(\mathbf{x}, \mathbf{x}^*\right) = \min\left(1, \frac{\pi(\mathbf{x}^*)}{\pi(\mathbf{x})}\right) = \min\left(1, \frac{\pi(\mathbf{x} + \mathbf{y}^*)}{\pi(\mathbf{x})}\right)$$

For a d-dimensional random walk we also define the expected acceptance rate as

$$\overline{\alpha}_d := E[\alpha(\mathbf{X}, \mathbf{X}^*)]$$

where the expectation is with respect to the joint law for the current value $\mathbf{X}$ and the proposed value $\mathbf{X}^*$. We consider only spherical and elliptically symmetric targets for which the ESJD is as defined in (3.15), or equivalently

$$S_d^2 = \sum_{i=1}^{d} \frac{1}{\beta_i^2} E[Y_i^2]$$

Expectation here is with respect to the law for the realised jump $\mathbf{Y}$. If both $\pi(\cdot)$ and $r(\cdot)$ are spherically symmetric then $S_d^2/d = E[(Y_i)^2]/\beta^2$ is the ESJD over any single component, and maximising this is equivalent to minimising the expected lag-1 autocorrelation over that component. As noted in Section 3.1.4 for spherically symmetric targets maximising the ESJD is equivalent to maximising the naive ESJD (3.14). Since the latter involves no floating constant of proportionality (or simply sets the target scale parameter to 1) we will consider this as our definition of ESJD throughout our examination of spherically symmetric random variables in Section 3.3.1.

As before, we denote the target Lesbegue density function as $\pi(\mathbf{x})$. In the region R_A, where acceptance is guaranteed, we have $\mathbf{x}' = \mathbf{x}^*$ and $\mathbf{y} = \mathbf{y}^*$ so that for integrals over R_A we need not distinguish between proposed and accepted values.

Therefore from Corollary 1 and Lemma 2

$$\overline{\alpha}_d(\lambda) = \frac{2}{\lambda^d} \int_{R_A} d\mathbf{x}\, d\mathbf{y}\, \pi(\mathbf{x})\, r(\mathbf{y}/\lambda) \tag{3.28}$$

$$S_d^2(\lambda) = \frac{2}{\lambda^d} \int_{R_A} d\mathbf{x}\, d\mathbf{y}\, |\mathbf{y}|^2\, \pi(\mathbf{x})\, r(\mathbf{y}/\lambda) \tag{3.29}$$

3.3.1 Spherically symmetric unimodal target distributions

In Section 3.3.1.1 for isotropic (spherically symmetric) unimodal targets we derive analytical forms valid at any dimension d for expected acceptance rate and ESJD in terms of simple expectations of the target's marginal distribution function along a single axis. Consequences for optimal-scaling in finite dimensions are then explored in Section 3.3.1.2, while Section 3.3.1.3 rewrites the exact forms from Section 3.3.1.1 in terms of the more intuitive marginal radial distribution and density functions.

Progressing to asymptotic optimal-scaling behaviour, we start (Sections 3.3.1.4 and 3.3.1.5) with some limit theory for marginal radial and marginal one-dimensional distribution functions of spherically symmetric random variables. Limiting results for optimal scaling are split into three sections: we first obtain simple limiting forms for expected acceptance rate and ESJD in terms of the limiting marginal radial distribution (Section 3.3.1.6). We then explore the existence (Section 3.3.1.7) and the properties (Section 3.3.1.8) of an optimal scaling in the limit as $d \to \infty$ and examine the consequences for the limiting acceptance rate.

3.3.1.1 Expected acceptance rate and ESJD for a finite dimensional target in terms of its marginal one-dimensional distribution function

Let the target density of spherically symmetric d-dimensional random variable $\mathbf{X}^{(d)}$ be $\pi_d(\mathbf{x}) = f_d(|\mathbf{x}|)$ with $f_d(x)$ a strictly monotonically decreasing function in non-negative x. For clarity of exposition we sometimes drop the subscript or superscript d and refer simply to $\mathbf{X}$, $\pi(\mathbf{x})$, and $f(|\mathbf{x}|)$. We will relax the strictness

of the monotonicity later in this section.

Now consider $\mathbf{x}$ and $\mathbf{x}+\mathbf{y}$ as two points in the same space $\Re^d$ rather than a single point in the product space $\Re^d \times \Re^d$. Thus R_A corresponds to the region where $\pi(\mathbf{x}+\mathbf{y}) > \pi(\mathbf{x})$, but

$$\begin{aligned} \pi(\mathbf{x}+\mathbf{y}) > \pi(\mathbf{x}) &\Leftrightarrow |\mathbf{x}+\mathbf{y}|^2 < |\mathbf{x}|^2 \\ &\Leftrightarrow 2\mathbf{x}.\mathbf{y} + \mathbf{y}.\mathbf{y} < 0 \\ &\Leftrightarrow \mathbf{x}.\hat{\mathbf{y}} < -\frac{1}{2}|\mathbf{y}| \end{aligned} \tag{3.30}$$

where $\hat{\mathbf{y}}$ is the unit vector in the direction of $\mathbf{y}$. So

$$(\mathbf{x}, \mathbf{x}+\mathbf{y}) \in R_A \Leftrightarrow \mathbf{y} \in \Re^d \text{ and } \mathbf{x}\cdot\hat{\mathbf{y}} < -\frac{1}{2}|\mathbf{y}|$$

Figure 3.1 shows the geometric intuition behind the equivalence between $|\mathbf{x}+\mathbf{y}|^2 < |\mathbf{x}|^2$ and $\mathbf{x}\cdot\hat{\mathbf{y}} < -\frac{1}{2}|\mathbf{y}|$. Here the component of $\mathbf{x}$ in the $\hat{\mathbf{y}}$ direction is denoted x_1 and the vector component of $\mathbf{x}$ perpendicular to $\mathbf{y}$ is $\mathbf{x}^-$. The contribution of $|\mathbf{x}^-|^2$ to $|\mathbf{x}|^2$ and $|\mathbf{x}+\mathbf{y}|^2$ is the same and so the only quantities relevant to the comparisons between the two magnitudes are x_1^2 and $(x_1+|\mathbf{y}|)^2$. The latter is clearly the smaller if and only if $|\mathbf{y}| < -2x_1$, as occurs in the figure.

Denote the one-dimensional marginal distribution function of the target $\mathbf{X}^{(d)}$ along unit vector $\hat{\mathbf{y}}$ as $F_{1|d}(x)$. Since $\mathbf{X}^{(d)}$ is spherically symmetric, this is independent of $\hat{\mathbf{y}}$, and we simply refer to it as *the* one-dimensional marginal distribution function of $\mathbf{X}^{(d)}$.

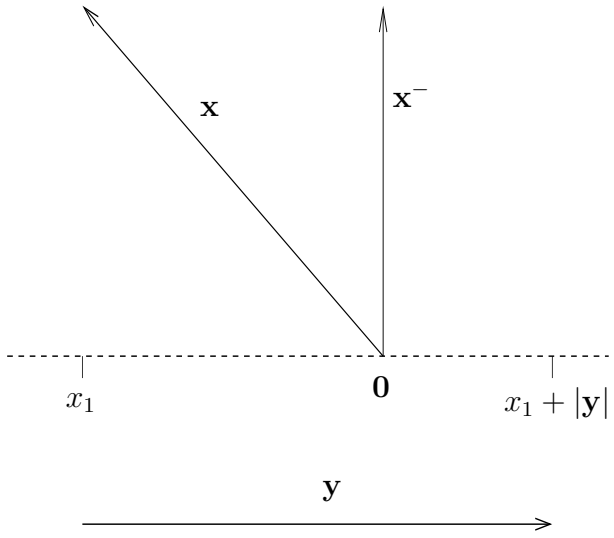

Figure 3.1: For proposed jump $\mathbf{y}$, current position $\mathbf{x}$ is decomposed into x_1, the component parallel to $\mathbf{y}$, and $\mathbf{x}^-$, the vector component perpendicular to $\mathbf{y}$.

Thus (3.28) and (3.29) become

$$\begin{aligned}\overline{\alpha}_d(\lambda) &= \frac{2}{\lambda^d}\int_{\Re^d} d\mathbf{y}\; r\left(\mathbf{y}/\lambda\right) F_{1|d}\left(-\frac{1}{2}|\mathbf{y}|\right)\\ S_d^2(\lambda) &= \frac{2}{\lambda^d}\int_{\Re^d} d\mathbf{y}\; |\mathbf{y}|^2 r\left(\mathbf{y}/\lambda\right) F_{1|d}\left(-\frac{1}{2}|\mathbf{y}|\right)\end{aligned}$$

Substituting $\mathbf{y}' = \frac{1}{\lambda}\mathbf{y}$ and relabelling we obtain

$$\begin{aligned}\overline{\alpha}_d(\lambda) &= 2\int_{\Re^d} d\mathbf{y}\; r\left(\mathbf{y}\right) F_{1|d}\left(-\frac{1}{2}\lambda|\mathbf{y}|\right)\\ S_d^2(\lambda) &= 2\lambda^2\int_{\Re^d} d\mathbf{y}\; |\mathbf{y}|^2\; r\left(\mathbf{y}\right) F_{1|d}\left(-\frac{1}{2}\lambda|\mathbf{y}|\right)\end{aligned}$$

Or equivalently

$$\overline{\alpha}_d(\lambda) = 2\mathbb{E}\left[F_{1|d}\left(-\frac{1}{2}\lambda\,|\mathbf{Y}|\right)\right] \tag{3.31}$$

$$S_d^2(\lambda) = 2\lambda^2\,\mathbb{E}\left[|\mathbf{Y}|^2\,F_{1|d}\left(-\frac{1}{2}\lambda\,|\mathbf{Y}|\right)\right] \tag{3.32}$$

where expectation is with respect to measure $r(\cdot)$ with *unit scale parameter*.

We now relax the "strict monotonicity" condition to "monotonicity". First consider the simple example of a uniform ball:

$$\begin{aligned}\pi(\mathbf{x}) &= c \quad \text{if} \quad |\mathbf{x}| \leq 1\\ &= 0 \quad \text{if} \quad |\mathbf{x}| > 1\end{aligned}$$

for some constant c. The rejection region $\{(\mathbf{x},\mathbf{x}') : |\mathbf{x}| \leq 1 \leq |\mathbf{x}'|\}$ is null with respect to the stationary distribution and, provided we start the algorithm in the support of $\pi(\cdot)$ the acceptance region $\{(\mathbf{x},\mathbf{x}') : |\mathbf{x}'| \leq 1 \leq |\mathbf{x}|\}$ is also null. However one part of the equality region $R^*_{EQ} := \{(\mathbf{x},\mathbf{x}') : |\mathbf{x}'| \leq 1, |\mathbf{x}| \leq 1, \mathbf{x} \neq \mathbf{x}'\}$ has Lesbegue measure 1.

We partition R^*_{EQ} into

$$\begin{aligned}C^{(1)} &:= \{(\mathbf{x},\mathbf{x}') : |\mathbf{x}'| < |\mathbf{x}| \leq 1\}\\ C^{(2)} &:= \{(\mathbf{x},\mathbf{x}') : |\mathbf{x}| < |\mathbf{x}'| \leq 1\}\\ C^{(null)} &:= \{(\mathbf{x},\mathbf{x}') : |\mathbf{x}| = |\mathbf{x}'| \leq 1, \mathbf{x} \neq \mathbf{x}'\}\end{aligned}$$

and apply Lemma 3. The third set is null and the first two sets correspond exactly to the acceptance and rejection region in our standard problem with strict monotonicity (consider for example altering the density slightly so that it has a small

downwards slope away from the origin when $|\mathbf{x}| \leq 1$ and is 0 thereafter).

Now consider a general spherically symmetric density function with

$$|\mathbf{x}_1| < |\mathbf{x}_2| \Rightarrow \pi(\mathbf{x}_1) \geq \pi(\mathbf{x}_2)$$

For brevity we will sometimes refer to such functions as *unimodal* and *isotropic*. These functions differ from the strictly monotonic by containing a (possibly countably infinite) series of plateaus. For target $\pi(\cdot)$, write $f(|\mathbf{x}|) = \pi(\mathbf{x})$ and let the i^{th} plateau be over the region $a_i < |\mathbf{x}| \leq b_i$ with $b_i \leq a_{i+1}$. Compare this function with the strictly monotonic function $g(|\mathbf{x}|)$ defined with a linear interpolation replacing each plateau:

$$\begin{aligned} g(x) &= f_i^l - \frac{x - (a_i - \epsilon)}{(b_i + \epsilon) - (a_i - \epsilon)} \times \left(f_i^l - f_i^h\right) \quad (\text{for } x \in [a_i - \epsilon, b_i + \epsilon], \text{ any } i) \\ &= f(x) \quad \text{elsewhere.} \end{aligned}$$

for some small $\epsilon > 0$, and where

$$f_i^l = \frac{f(a_i - \epsilon) + f(a_i)}{2} \quad \text{and} \quad f_i^h = \frac{f(b_i + \epsilon) + f(b_i)}{2}$$

If plateaux i and $i + 1$ are adjacent or nearly adjacent, simply define $g(\cdot)$ in the overlap region of size at most 2ϵ as the maximum of the two possibilities. For target $g(|\mathbf{x}|)$, regions R_{EQ}, R_A, and R_R only differ from those of $f(|\mathbf{x}|)$ when both $\mathbf{x}$ and $\mathbf{x}'$ occupy the same plateau (extended at each end by ϵ). But the entire portion of any space with $\mathbf{x}$ and $\mathbf{x}'$ on the same plateau (and $\mathbf{x} \neq \mathbf{x}'$) is an equality region for $f(\cdot)$ and so may be treated as for the uniform ball and partitioned into a null set and two non-null regions *which correspond exactly to the acceptance and rejection regions of target* $g(|\mathbf{x}|)$. Let $\epsilon \to 0$ to see that equations (3.31) and (3.32) apply for any unimodal isotropic target density.

The marginal distribution function $F_{1|d}(-\lambda |\mathbf{Y}|/2)$ is bounded and decreasing in λ. Also $\lim_{x\to\infty} F_{1|d}(-x) = 0$ and by symmetry, provided the marginal distribution function is continuous at the origin, $\lim_{x\to 0} F_{1|d}(-x) = 0.5$. Applying the bounded convergence theorem to (3.31) we therefore obtain the following, true whatever the dimension of the random walk:

Corollary 3 *Let λ be the scaling parameter for any RWM algorithm on a unimodal isotropic target Lesbegue density. In this situation the expected acceptance rate at stationarity $\overline{\alpha}_d(\lambda)$ decreases with increasing λ with $\lim_{\lambda\to 0} \overline{\alpha}_d(\lambda) = 1$ and $\lim_{\lambda\to\infty} \overline{\alpha}_d(\lambda) = 0$.*

In fact this result holds true for all unimodal elliptically symmetric targets, as will be discussed in Section 3.3.2.

Now suppose $\mathbf{X}^{(d)} \sim N(\mathbf{0}, \lambda_t^2 \mathbf{I}_d)$. Here $F_{1|d}(x) = \Phi(x/\lambda_t)$, where $\Phi(\cdot)$ is the distribution function of a standard Gaussian, and we have exactly that

$$\overline{\alpha}_d(\lambda) = 2\mathbb{E}\left[\Phi\left(-\frac{1}{2}\frac{\lambda}{\lambda_t}|\mathbf{Y}|\right)\right] \tag{3.33}$$

$$S_d^2(\lambda) = 2\lambda^2\,\mathbb{E}\left[|\mathbf{Y}|^2\,\Phi\left(-\frac{1}{2}\frac{\lambda}{\lambda_t}|\mathbf{Y}|\right)\right] \tag{3.34}$$

where expectation is with respect to proposal density $r(\cdot)$ with unit scale parameter. No other distributions are both spherically symmetric and have independent components and hence such a simple one-dimensional marginal distribution independent of the axis. Nevertheless (3.31) and (3.32) will prove extremely useful when considering both finite dimensional behaviour and the limit as $d \to \infty$.

For many spherically symmetric target distributions it is more intuitive to think in terms of the marginal radial distribution, which is easily derived from the general density function in $\Re^d$. In this case analytical results in terms of simple expectations of marginal radial densities and standard functions are possible (see Section 3.3.1.3). We first examine the validity of the principal of optimal-scaling in finite dimensions.

3.3.1.2 Optimal scaling for spherically symmetric unimodal targets in finite dimensions

In this Section we use (3.32) to explore behaviour of the ESJD for finite dimensional unimodal spherically symmetric targets. We prove the existence of at least one (finite) optimal scaling, subject to conditions on the moments of the target and proposal. Results from this section will later be shown to apply to more general unimodal elliptically symmetric distributions (see Section 3.3.2). We introduce the notation $\overline{r}_d(y)$ for the density of $|\mathbf{Y}|$ for general proposal $\mathbf{Y}$.

Lemma 6 *Consider a spherically symmetric unimodal d-dimensional target Lesbegue density $\pi(\mathbf{x})$ with 1-dimensional marginal distribution function $F_{1|d}(x)$. Let $\pi(\cdot)$ be explored through a RWM algorithm with proposal Lesbegue density $\frac{1}{\lambda^d} r(\mathbf{y}/\lambda)$. Consider the expected square jump distance along any component of the Markov chain at stationarity, $S_d^2(\lambda)$. If $\mathbb{E}_\pi\left[|\mathbf{X}|^2\right] < \infty$ and $\mathbb{E}_r\left[|\mathbf{Y}|^2\right] < \infty$ then*

$$\lim_{\lambda \to 0} S_d^2(\lambda) = 0 \tag{3.35}$$

$$\lim_{\lambda \to \infty} S_d^2(\lambda) = 0 \tag{3.36}$$

$$S_d^2(\lambda) > 0 \quad \textit{for all} \quad \lambda \in (0, c) \tag{3.37}$$

for some c with $0 < c \leq \infty$.

Proof: *Since* $\mathbb{E}\left[|\mathbf{Y}|^2\right]$ *is finite we may apply the dominated convergence theorem to prove (3.35).*

$$\begin{aligned}
\lim_{\lambda\to 0} S_d^2(\lambda) &= \lim_{\lambda\to 0} 2\lambda^2\ \mathbb{E}_{\mathbf{Y}}\left[|\mathbf{Y}|^2 F_{1|d}\left(-\frac{1}{2}\lambda|\mathbf{Y}|\right)\right] \\
&= \lim_{\lambda\to 0}\left(2\lambda^2\right)\times\mathbb{E}_{\mathbf{Y}}\left[|\mathbf{Y}|^2 \lim_{\lambda\to 0} F_{1|d}\left(-\frac{1}{2}\lambda|\mathbf{Y}|\right)\right] \\
&= \lim_{\lambda\to 0}\left(2\lambda^2\right)\times\mathbb{E}_{\mathbf{Y}}\left[\frac{1}{2}|\mathbf{Y}|^2\right] \\
&= 0
\end{aligned}$$

To show (3.36) we first construct an upper bound for $S_d^2(\lambda)$.

$$\begin{aligned}
S_d^2(\lambda) &= 2\lambda^2\mathbb{E}_{\mathbf{Y}}\left[|\mathbf{Y}|^2 F_{1|d}\left(-\frac{1}{2}\lambda|\mathbf{Y}|\right)\right] \\
&= 2\lambda^2\int_0^\infty dy\ \overline{r}_d(y)\, y^2\int_{-\infty}^{-\frac{\lambda}{2}y} dx\ f_{1|d}(x) \\
&= 2\lambda^2\int_0^\infty dy\ \overline{r}_d(y)\, y^2\int_{\frac{\lambda}{2}y}^{\infty} dx\ f_{1|d}(x) \\
&= 2\lambda^2\int_0^\infty dx\ f_{1|d}(x)\int_0^{\frac{2x}{\lambda}} dy\ \overline{r}_d(y)\, y^2 \\
&\leq 2\lambda^2\int_0^\infty dx\ f_{1|d}(x)\int_0^{\frac{2x}{\lambda}} dy\ \overline{r}_d(y)\left(\frac{2x}{\lambda}\right)^2 \\
&= 8\mathbb{E}_{f_{1|d}}\left[X^2\ \int_0^{\frac{2x}{\lambda}} dy\ \overline{r}_d(y)\right]
\end{aligned}$$

Now $\mathbb{E}_{f_{1|d}}[X^2] \leq \mathbb{E}_f\left[|\mathbf{X}|^2\right] < \infty$ *so by the dominated convergence theorem*

$$\lim_{\lambda\to\infty} S_d^2(\lambda) \leq 8\mathbb{E}_{f_{1|d}}\left[X^2\ \lim_{\lambda\to\infty}\int_0^{\frac{2X}{\lambda}} dy\ \overline{r}_d(y)\right] = 0$$

since $\overline{r}_d(y)$ *is a Lesbegue density.*

Finally, since $\mathbb{E}_{\overline{r}_d}[Y^2] < \infty$ *there is an* a *for which* $0 < \int_0^b dy\ \overline{r}_d(y)\ y^2 < \infty$ *for all* $b > a$ *and so*

$$S_d^2(\lambda) \geq 2\lambda^2 \int_0^b dy\ \overline{r}_d(y)\, y^2 F_{1|d}\left(-\frac{1}{2}\lambda y\right) \geq 2\lambda^2 F_{1|d}\left(-\frac{1}{2}\lambda b\right) \int_0^b dy\ \overline{r}_d(y)\, y^2$$

But $F_{1|d}(x)$ *is the distribution function of a symmetric Lesbegue density and so is strictly positive for* x *greater than some* $-\epsilon$*. Thus* S_d^2 *is certainly strictly positive for* $0 < \lambda < 2\epsilon/b$*, proving (3.37).*

Note that the ESJD need not be stricly positive for all $\lambda \in (0, \infty)$. Consider for example a proposal which has density zero inside the sphere of radius λ and a target which has mass only inside the unit sphere. The acceptance probability will be zero everywhere for $\lambda > 2$.

The next corollary follows immediately from Lemma 6, and validates our search for optimal scaling(s) in finite dimensions for unimodal isotropic targets.

Corollary 4 *Consider a spherically symmetric unimodal* d*-dimensional target Lesbegue density* $\pi(\mathbf{x})$*. Let* $\pi(\cdot)$ *be explored via a RWM algorithm with proposal Lesbegue density* $\frac{1}{\lambda^d} r(\mathbf{y}/\lambda)$*. If* $\mathbb{E}_\pi\left[|\mathbf{X}|^2\right] < \infty$ *and* $\mathbb{E}_r\left[|\mathbf{Y}|^2\right] < \infty$ *then the ESJD of the Markov chain at stationarity attains its maximum at a finite non-zero value (or values) of* λ*.*

3.3.1.3 Expected acceptance rate and ESJD for a finite dimensional target in terms of its marginal radial density

We seek expressions for the expected acceptance rate and the ESJD in terms of marginal radial densities or distribution functions. We start by deriving the one-dimensional marginal distribution function of a spherically symmetric random

variable in terms of its marginal radial distribution function. The derivation is straightforward and is given below; alternatively the marginal one-dimensional distribution function may be deduced from standard results on the partitioning of components of spherically symmetric distributions (e.g. Fang et al., 1990, Section 2.3).

We introduce some further notation; write $\overline{F}_d(r)$ and $\overline{f}_d(r)$ for the marginal radial distribution and density functions of d-dimensional spherically symmetric target $\mathbf{X}^{(d)}$; these are the distribution and density functions of $\left|\mathbf{X}^{(d)}\right|$. We continue to denote the marginal distribution of $\mathbf{X}^{(d)}$ along any particular axis by $F_{1|d}(x)$.

Lemma 7 *For any d-dimensional spherically symmetric random variable with marginal radial distribution function $\overline{F}_d(r)$ the 1-dimensional marginal distribution function along any axis is*

$$F_{1|d}(x_1) = \frac{1}{2}\left(1 + sign(x_1)\ \mathbb{E}\left[\overline{F}_d\left(\frac{|x_1|}{U_d^{1/2}}\right)\right]\right) \tag{3.38}$$

where $sign(x) = 1$ for $x \geq 0$ and $sign(x) = -1$ for $x < 0$. Here

$$\begin{aligned} U_1 &= 1 \\ U_d &\sim Beta\left(\frac{1}{2}, \frac{d-1}{2}\right) \quad (d > 1) \end{aligned}$$

Proof: *To allow for a possible point mass at the origin we define*

$$p_d := \overline{F}_d(0)$$

Clearly

$$F_{1|d}(0) = P(X_1 < 0) + P(X_1 = 0) = \frac{1}{2}(1 - p_d) + p_d = \frac{1}{2}(1 + p_d) = \frac{1}{2}\left(1 + \overline{F}_d(0)\right)$$

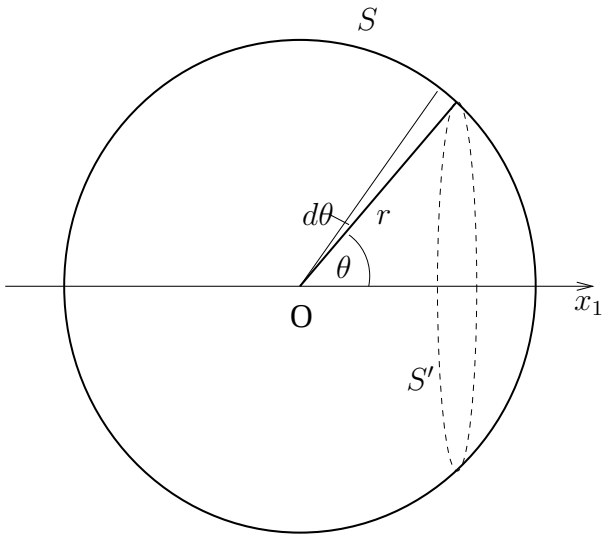

Figure 3.2: A d-dimensional spherical shell S at distance r from the origin, and a $(d-1)$-dimensional spherical shell $S' \subset S$ at angle θ to the x_1 axis.

and the result therefore holds for $x_1 = 0$. *For* $x_1 > 0$, *by symmetry we have* $F_{1|d}(-x_1) = 1 - F_{1|d}(x_1)$ *and so the result for* $-x_1$ *would follow from a proof for* $+x_1$. *Also for* $d = 1$ *and for* $x_1 > 0$, $F_{1|1}(x_1) = (1-p_1)/2+p_1+1/2\times\big(\overline{F}_1(x_1) - p_1\big)$. *We need therefore only consider* $d \geq 2$ *and* $x_1 > 0$.

Following the notation of Section 1.5.2, the probability mass per unit "area" of a d-dimensional hyperspherical shell at radius $r > 0$ *is*

$$\frac{d\overline{F}_d(r)}{a_d\, r^{d-1}}$$

Consider the d-1 dimensional hyperspherical shell consisting of that part of the original shell at angle $[\theta, \theta + d\theta)$ *to the* x_1 *axis (see Figure 3.2). The total mass in*

this shell is

$$\frac{d\overline{F}_d(r)}{a_d r^{d-1}} \times r\ d\theta\ a_{d-1}(r\sin\theta)^{d-2} = \frac{a_{d-1}}{a_d}\sin^{d-2}\theta\ d\overline{F}_d(r)\ d\theta$$

Since $F_{1|d}(0) = (1+p_d)/2$, *for non-negative* x_1

$$\begin{aligned} F_{1|d}(x_1) &= \frac{1+p_d}{2} + \frac{a_{d-1}}{a_d}\int_0^{\pi/2} d\theta\ \sin^{d-2}\theta \int_{0^+}^{x_1/\cos\theta} d\overline{F}_d(r) \\ &= \frac{1+p_d}{2} + \frac{a_{d-1}}{a_d}\int_0^{\pi/2} d\theta\ \sin^{d-2}\theta\ \left(\overline{F}_d(x_1/\cos\theta) - p_d\right) \end{aligned}$$

Now substitute $u^{1/2} = \cos\theta$ *so that* $-2\sin\theta\ d\theta = u^{-1/2}du$ *and recall (1.15) for the ratio* a_d/a_{d-1} *to obtain*

$$\begin{aligned} F_{1|d}(x_1) &= \frac{1}{2}\left(1 + p_d + \int_0^1 du\ g_d(u)\left(\overline{F}_d\left(x_1/u^{1/2}\right) - p_d\right)\right) \\ &= \frac{1}{2}\left(1 + \int_0^1 du\ g_d(u)\ \overline{F}_d\left(x_1/u^{1/2}\right)\right) \end{aligned}$$

Here

$$\begin{aligned} g_d(u) &:= \frac{1}{B\left(\frac{1}{2},\frac{d-1}{2}\right)} u^{-1/2}(1-u)^{(d-3)/2} \quad (0 \le u \le 1) \\ &\qquad 0 \qquad\qquad (u < 0\ or\ u > 1) \end{aligned}$$

which is the density function of a Beta $\left(\frac{1}{2},\frac{d-1}{2}\right)$ *random variable.*

Lemma 7 applies whatever the form of the radial distribution function $\overline{F}_d(\cdot)$. In our investigations in to the random walk Metropolis algorithm we are concerned only with targets that posses a density with respect to the Lesbegue measure in $\Re^d$. In this case both the marginal one-dimensional and radial density functions $f_{1|d}(\cdot)$ and $\overline{f}_d(\cdot)$ exist trivially, and consequently the corresponding distribution functions are continuous. Further $\overline{F}_d(0) = 0$ as there can be no point mass at the

origin (or anywhere else), and hence

$$\begin{aligned}\mathbb{E}\left[\overline{F}_d\left(\frac{x_1}{U_d^{1/2}}\right)\right] &= \int_0^\infty dG_d(u)\ \int_0^{x_1/u^{1/2}} d\overline{F}_d\,(x) \\ &= \int_0^\infty d\overline{F}_d\,(x)\ \int_0^{x_1^2/x^2} dG_d(u) \\ &= \mathbb{E}\left[G_d\left(\left(\frac{x_1}{X^{(d)}}\right)^2\right)\right]\end{aligned}$$

where $X^{(d)} = \left|\mathbf{X}^{(d)}\right|$ and $G_d(u_1) = \int_0^{u_1} du\ g_d(u)$ is the distribution function of a $Beta(1/2,(d-1)/2)$ random variable with $G_d(u_1) = 1$ for $u_1 \geq 1$. We may therefore re-write (3.38) as

$$F_{1|d}(x_1) = \frac{1}{2}\left(1 + sign(x_1)\ \mathbb{E}_{X^{(d)}}\left[G_d\left(\left(\frac{x_1}{X^{(d)}}\right)^2\right)\right]\right) \quad (d \geq 2) \tag{3.39}$$

Substituting (3.38) into (3.31) and (3.32) we obtain

$$\overline{\alpha}_d(\lambda) = 1 - \mathbb{E}_{\mathbf{Y},U}\left[\overline{F}_d\left(\frac{\lambda\,|\mathbf{Y}|}{2U^{1/2}}\right)\right] \tag{3.40}$$

$$S_d^2(\lambda) = \lambda^2\mathbb{E}_{\mathbf{Y},U}\left[|\mathbf{Y}|^2\left(1 - \overline{F}_d\left(\frac{\lambda\,|\mathbf{Y}|}{2U^{1/2}}\right)\right)\right] \tag{3.41}$$

where $U \sim Beta(1/2,(d-1)/2)$. The above forms are interesting because they express both the expected acceptance rate and ESJD in terms expectations; further, these expectations are over quantities whose distributions are known to the statistician and from which he or she could simulate. However the marginal radial distribution function of any specific target density may not be easy to obtain. Alternative forms of more practical use when examining the behaviour of specific combinations of target and proposal are obtained by substituting (3.39) into (3.31) and (3.32). To simplify the notation we first define for non-negative u

$$K_d(u) := 1 - G_d(u^2)$$

so that $K_d(0) = 1$ and $K_d(u) = 0$ for $u \geq 1$; also $K_1(u) = 1$ for $0 \leq u < 1$. Then

$$\overline{\alpha}_d(\lambda) = \mathbb{E}_{\mathbf{Y},X^{(d)}}\left[K_d\left(\frac{\lambda\left|\mathbf{Y}\right|}{2X^{(d)}}\right)\right] \tag{3.42}$$

$$S_d^2(\lambda) = \lambda^2\mathbb{E}_{\mathbf{Y},X^{(d)}}\left[\left|\mathbf{Y}\right|^2\left(K_d\left(\frac{\lambda\left|\mathbf{Y}\right|}{2X^{(d)}}\right)\right)\right] \tag{3.43}$$

These expectations depend on $\mathbf{Y}$ only through $|\mathbf{Y}|$, so writing $\overline{r}_d(y)$ for the marginal radial density of $|\mathbf{Y}|$ we obtain the expected acceptance rate and ESJD in terms of straightforward double integrals.

$$\overline{\alpha}_d(\lambda) = \int_0^\infty dy \int_{\frac{1}{2}\lambda y}^\infty dx\ \overline{r}_d(y)\,\overline{f}_d(x)\,K_d\left(\frac{\lambda y}{2x}\right) \tag{3.44}$$

$$S_d^2(\lambda) = \lambda^2\int_0^\infty dy \int_{\frac{1}{2}\lambda y}^\infty dx\ \overline{r}_d(y)\,\overline{f}_d(x)\,y^2K_d\left(\frac{\lambda y}{2x}\right) \tag{3.45}$$

Obtaining the marginal radial density of the proposed jump may itself require a multi-dimensional integral, but in the event that $\mathbf{Y}^{(d)}$ is also spherically symmetrical then $\overline{r}_d(|\mathbf{y}|) = a_d\left|\mathbf{y}\right|^{d-1}r_d(\mathbf{y})$. For certain simple combinations of target and proposal the double integral may be reduced to a single integral or even removed completely (see Section 3.3.4.1), but even if this is not possible, a simple **R** routine will quickly evaluate the integrals numerically for various values of λ and d and the behaviour of the optimal scaling and acceptance rates can be ascertained (Section 3.3.4.2).

In the event that $r(\mathbf{y})$ is not spherically symmetric but is easy to simulate from, then one or both of the expectations in (3.40) and (3.41) or (3.42) and (3.43) may be evaluated by Monte Carlo approximation.

3.3.1.4 Limit theorems for spherically symmetric distributions

This section is dedicated to limit theorems for the marginal one-dimensional distribution function of a spherically symmetric distribution. We will show that if

the modulii of a (suitably rescaled) sequence of spherically symmetric random variables converges in probability to a non-zero constant then the sequence of (rescaled) one-dimensional marginal distributions converges to a standard Gaussian. More generally, even when rescaled convergence in probability is not achieved, provided the sequence of (suitably rescaled) marginal radial distribution functions converges weakly then the sequence of (rescaled) marginal one-dimensional distribution functions converges weakly to a scaled mixture of normals. The first result clearly follows as a special case of the second, however we prove it separately using a much simpler argument that gives an intuition into the underlying reason for the mixture form. The mixture result itself follows from a theorem in Fang et al. (1990); we provide an alternative proof from first principles in Appendix C.

We first define some notation for the convergence of random variables that will be used throughout the rest of this chapter. Weak convergence, also known as convergence in distribution, is denoted by $\xrightarrow{D}$; convergence in probability is denoted $\xrightarrow{p}$ and convergence in mean square by $\xrightarrow{m.s.}$. We now recall some properties of convergence in probability, which will subsequently be used without further comment or reference. For any random variable Z with $P(Z = \infty) = 0$ and any two sequences of random variables $X_n \xrightarrow{p} a$, $Y_n \xrightarrow{p} b \neq 0$

$$X_n Y_n \xrightarrow{p} ab\ , \quad Y_n^k \xrightarrow{p} b^k \text{ for any } k, \quad \text{and} \quad X_n Z \xrightarrow{p} aZ$$

As well as being essential to the proof of Theorem 1, the following simple generalisation of the weak law of large numbers for triangular sequences of chi-square random variables will also play a part in the discussion of convergence in probability for elliptically symmetric random variables in Section 3.3.2.1. A proof which mirrors exactly that for the weak law of large numbers, is given for the sake of

completeness in Appendix C.

Lemma 8 *Let* $\{\mathbf{Z}^{(n)} : \mathbf{Z}^{(n)} \sim N(\mathbf{0}, \mathbf{I}_n)\}$ *be a sequence of independent random variables in* $\Re^n$.

$$\frac{\left|\mathbf{Z}^{(n)}\right|}{n^{1/2}} \xrightarrow{p} 1$$

We may simulate a spherically symmetric random variable $\mathbf{X}^{(d)}$ on $\Re^d$ by simulating a random direction and independently simulating a random length $R^{(d)}$. This may be taken as defining a spherically symmetric random variable; for equivalence with other possible definitions see Theorem 2.3 of Fang et al. (1990). *Fix* a p-dimensional space V_p ($p \leq d$) which is $\Re^p$, possibly rotated within $\Re^d$. Let $\mathbf{Z}^{(d)} \sim N(\mathbf{0}, \mathbf{I}_d)$ be independent of $R^{(d)}$, and let $\mathbf{Z}^{(p|d)} \sim N(\mathbf{0}, \mathbf{I}_p)$ be the p-component marginal distribution of $\mathbf{Z}^{(d)}$ over $V_p \leq \Re^d$ (in the context of vector spaces, "$\leq$" denotes "is a subspace of"). Then with $\mathbf{X}^{(p|d)}$ denoting the p-dimensional random variable consisting of the components of $\mathbf{X}^{(d)}$ in V_p, we may write

$$\mathbf{X}^{(d)} = \frac{\mathbf{Z}^{(d)}}{|\mathbf{Z}^{(d)}|} \times R^{(d)} \quad \text{and} \quad \mathbf{X}^{(p|d)} = \frac{\mathbf{Z}^{(p|d)}}{|\mathbf{Z}^{(d)}|} \times R^{(d)}$$

We may of course relate $\mathbf{X}^{(d)}$ to any other spherically symmetric random variable in this manner, but the Gaussian is the only spherically symmetric random variable with independent components and therefore a simple form for the distribution of any p-component marginal.

Theorem 1 *Let* $\mathbf{X}^{(d)}, \mathbf{X}^{(p|d)}, \mathbf{Z}^{(d)}, \mathbf{Z}^{(p|d)}$ *and* V_p *be defined as above. If there exist* k_d *such that the marginal radius,* $R^{(d)} := |\mathbf{X}^{(d)}|$ *satisfies* $R^{(d)}/k_d \xrightarrow{p} 1$ *then* $\frac{d^{1/2}}{k_d}\mathbf{X}^{(p|d)} \xrightarrow{p} \mathbf{Z}^{(p|d)} \sim N(\mathbf{0}, \mathbf{I}_p)$.

Proof: $|\mathbf{Z}^{(d)}|^2 \sim \chi^2_d$ *, so from Lemma 8* $|\mathbf{Z}^{(d)}|/d^{1/2} \xrightarrow{p} 1$*. Also* $R^{(d)}/k_d \xrightarrow{p} 1$*, and so*

$$\frac{d^{1/2}}{k_d}\mathbf{X}^{(p|d)} = \mathbf{Z}^{(p|d)} \times \frac{R^{(d)}/k_d}{|\mathbf{Z}^{(d)}|/d^{1/2}} \xrightarrow{p} \mathbf{Z}^{(p|d)}$$

In actual fact our subspace V_p will be a one-dimensional subspace of $\Re^d$ but it will *not be fixed* as $d \to \infty$. However since $\mathbf{X}^{(d)}$ is spherically symmetrical, the marginal distribution along any given one-dimensional subspace of $\Re^d$ is the same as that along any other one-dimensional subspace, at least one of which converges in probability to $\mathbf{Z}^{(1|d)} \sim N(0,1)$. Convergence in probability implies convergence in distribution and it follows that *any* 1-component projection of $\mathbf{X}^{(d)}$ (suitably rescaled) converges in distribution to a standard Gaussian.

From real analysis we have the following result (a proof of which is given for completeness in Appendix C):

Lemma 9 *Let $G_d(x)$ be a sequence of monotonic functions, with identical finite upper and lower bounds. If the sequence converges to a continuous limit $G(x)$ then it does so uniformly in x.*

Thus for the one-dimensional marginal distribution function we obtain:

Corollary 5 *Let $\{\mathbf{X}^{(d)}\}$ be a sequence of d-dimensional spherically symmetric random variables with one-dimensional marginal distribution functions $F_{1|d}(\cdot)$, and let there be a k_d such that $\left|\mathbf{X}^{(d)}\right|/k_d \xrightarrow{p} 1$. Then*

$$F_{1|d}\left(\frac{k_d}{d^{1/2}}x_1\right) \to \Phi(x_1) \quad \textit{uniformly}$$

A general sequence of d-dimensional isotropic random variables may not satisfy the criterion of convergence in probability of the rescaled modulus. However, provided

the (suitably rescaled) sequence of marginal distribution functions of the modulii of the random variables converges to some distribution function $\overline{\Theta}(r)$ then the (suitably rescaled) one-dimensional marginal distribution function converges to that of a scaled mixture of normals; the intuition behind this result is clear from Theorem 1.

Convergence of the sequence of characteristic functions of a sequence of d-dimensional isotropic random variables to that of a mixture of normals is proved as Theorem 2.21 of Fang et al. (1990). Clearly the marginal distribution along any given axis is a mixture of one-dimensional standard normals and so a single component may be written as $X_1 = RZ$ with Z a standard Gaussian and R the mixing distribution.

We require a result for the limiting marginal one-dimensional distribution function to which we apply Lemma 9 to give uniform convergence. We therefore obtain the following:

Theorem 2 *Let $\{\mathbf{X}^{(d)}\}$ be a sequence of d-dimensional spherically symmetric random variables. If there exist k_d such that $\left|\mathbf{X}^{(d)}\right|/k_d \xrightarrow{D} R$ where R has distribution function $\overline{\Theta}(r)$ with $\overline{\Theta}(0) = 0$. then the sequence of marginal one-dimensional distributions of $\mathbf{X}^{(d)}$ satisfies*

$$F_{1|d}\left(\frac{k_d}{d^{1/2}}x_1\right) \to \Theta(x_1) := \mathbb{E}_R\left[\Phi\left(\frac{x_1}{R}\right)\right] \quad \textit{uniformly in } x_1 \tag{3.46}$$

where $\Phi(\cdot)$ is the standard Gaussian distribution function.

For an alternative proof of this theorem (from first principles) see Appendix C.

$\left|\mathbf{X}^{(d)}\right|$ posseses a Lesbegue density and therefore no point mass at the origin; however the rescaled limit R may posses such a point mass. We now examine the

consequences of this possibility, in particular for the continuity of $\Theta(\cdot)$. It will be shown in Lemma 10 that if $\overline{\Theta}(0) = p > 0$ then $\Theta(x)$ is discontinuous at the origin. In such circumstances Theorem 2 continues to hold except that Lemma 9 no longer applies and convergence is therefore no longer uniform in x_1. It would also be necessary to specify $\Phi(x_1/0)$ to be 0 if $x_1 < 0$ and 1 if $x_1 \geq 0$ (since distribution functions are right-continuous). This specification is unnecessary in the theorem as it stands since $\Phi(\cdot)$ is bounded and therefore with no point mass at the origin the term $\Phi(x_1/0)$ makes zero contribution to the expectation.

Insistence on $\overline{\Theta}(0) = 0$ ensures continuity of $\Theta(x_1)$ for all $x_1 \in \Re$ since by the bounded convergence theorem

$$\lim_{x_1 \to k} \Theta(x_1) = \lim_{x_1 \to k} \mathbb{E}_R \left[\Phi\left(x_1/R\right)\right] = \mathbb{E}_R \left[\lim_{x_1 \to k} \Phi\left(x_1/R\right)\right]$$

The above holds whether the limit is approached from above or below and possible discrepancies between the two at $x_1 = 0$ due to the extended definition for $R = 0$ are avoided.

If $\overline{\Theta}(0) = p > 0$ then consideration of R as the mixture

$$\begin{aligned} R &= 0 \quad \text{with probability} \quad p \\ &= R^* \quad \text{with probability} \quad 1 - p \end{aligned}$$

where R^* has no mass at the origin, leads directly to the following:

Lemma 10 *Let $\mathbf{X}^{(d)}$ be a sequence of spherically symmetric random variables with an associated scaling k_d such that $\left|\mathbf{X}^{(d)}\right| / k_d \xrightarrow{D} R$ where R has distribution function $\overline{\Theta}(r)$. Consider the limiting marginal one-dimensional distribution function*

$\Theta(x)$ of a spherically symmetric random variable; this is continuous for $x \neq 0$ and

$$\lim_{x \uparrow 0} \Theta(x) = \frac{1}{2}(1-p) \quad \textit{and} \quad \Theta(0) = \frac{1}{2}(1+p)$$

where $p := \overline{\Theta}(0)$.

Note that $\overline{\Theta}(x)$ (and hence also $\Theta(x)$) is only unique up to a constant: if scaling k_d produces $\overline{\Theta}(x)$ then scaling ak_d will produce $\overline{\Theta}(ax)$.

The condition of convergence of the rescaled modulus to 1 or to random variable R will turn out to be the key determinant in the behaviour of the optimal acceptance rate as $d \to \infty$; we now examine the limiting behaviour in more detail.

3.3.1.5 Limiting forms for the rescaled modulus of the target

For many of the standard sequences of density functions (e.g. $\pi_d(\mathbf{x}) \propto |\mathbf{x}|^a e^{-|\mathbf{x}|^c}$) there is a k_d such that $\left|\mathbf{X}^{(d)}\right| / k_d \xrightarrow{p} 1$. Clearly this is equivalent to $\left|\mathbf{X}^{(d)}\right| / k_d \xrightarrow{p} c$ for any $c \in (0, \infty)$, with each k_d divided by c. However choosing the sequence k_d to have values too small by a factor which tends to infinity produces a point mass of 1 at ∞, and similarly choosing the values too large by this factor results in a point mass of 1 at 0, neither of which are informative.

It is easy to construct d-dependent density forms for which there is no rescaling sequence k_d that gives convergence in probability to 1. For example if $\pi_d(\mathbf{x}) \propto |\mathbf{x}|^{-d+1} e^{-\frac{1}{2}|\mathbf{x}|^2}$ then the marginal radial distribution function is always a standard Gaussian. It is also easy to construct (highly artificial) sequences $\{\mathbf{X}^{(d)}\}$ such that there is no scaling to give any convergence, except to 0 or ∞ (for example $\pi_d(\mathbf{x}) \propto |\mathbf{x}|^{-d+1} e^{-\frac{\sin^2 d}{2}|\mathbf{x}|^2}$).

Also of interest are sequences of random variables with the same functional form for density in terms of the radial distance. However even in such cases as these, where the only explicit dependence on d is through the normalisation constant, there are spherically symmetric random variables for which there is no re-scaling such that $\left|\mathbf{X}^{(d)}\right|/k_d \xrightarrow{p} 1$. Consider for example $\pi_d(\mathbf{x}) = f_d(|\mathbf{x}|)$ where

$$f_d(x) = \frac{1}{(2\pi e^d)^{1/2}} x^{-\frac{1}{2}\log x} = \frac{1}{(2\pi e^d)^{1/2}} e^{-\frac{1}{2}(\log x)^2} \tag{3.47}$$

Then

$$\begin{aligned} \overline{f}_{d+1}(x) &\propto x^d e^{-\frac{1}{2}(\log x)^2} \\ \log \overline{f}_{d+1}(x) &= const + d\log x - \frac{1}{2}(\log x)^2 \\ \left(\log \overline{f}_{d+1}(x)\right)' &= \frac{d}{x} - \frac{\log x}{x} \end{aligned}$$

Thus $\overline{f}_{d+1}(x)$ is maximised at $x = e^d$ and this therefore must be the re-scaling factor k_d that could potentially lead to some limiting distribution. We substitue $u = x/e^d$ and find its density function $h_{d+1}(u)$

$$\begin{aligned} h_{d+1}(u) &\propto u^d e^{-\frac{1}{2}(\log u + d)^2} \\ &\propto u^d e^{-\frac{1}{2}(\log u)^2} e^{-d\log u} \\ &= e^{-\frac{1}{2}(\log u)^2} \end{aligned}$$

The rescaled density is therefore independent of d and in fact exhibits the same form as the original density as a function of the radius

$$h_d(u) = \frac{1}{(2\pi e)^{1/2}} e^{-\frac{1}{2}(\log u)^2} \tag{3.48}$$

This is because $U = \log X$ has Gaussian density with mean d, and dividing X by e^{d-1} is equivalent to subtracting $d-1$ from U. The result is a Gaussian with mean

1, which is the density of the transformed X at $d = 1$.

We therefore seek a sufficient condition for there to exist a sequence k_d such that $\left|\mathbf{X}^{(d)}\right| / k_d \xrightarrow{p} 1$ when the density of spherically symmetric $\mathbf{X}^{(d)}$ only depends on dimension through the normalisation constant. We will find that polynomial tail behaviour of the log density guarantees the required convergence.

Define

$$g(|\mathbf{x}|) := -\log f(|\mathbf{x}|) \tag{3.49}$$

We assume throughout that $g(r)$ is twice differentiable everywhere on the interval $[0, \infty)$ and define

$$h(r) := r(rg'(r))' \tag{3.50}$$

Also write the d-dimensional marginal radial density as

$$f_d^*(r) := c_d r^{d-1} \exp(-g(r)) \tag{3.51}$$

where $c_d^{-1} = \int_0^\infty dr\ r^{d-1} \exp(-g(r))$. Note that $g(r)$ being twice differentiable implies that both $g'(r)$ and $g(r)$ are bounded on any compact interval $[0, s]$ for $s < \infty$ and that therefore $f_d^*(r)$ has support $[0, \infty)$.

Finally let r_d be any maximum of the marginal radial density and define the modulus of the rescaled target as $U^{(d)} := \left|\mathbf{X}^{(d)}\right| / r_d$. Then the density of $U^{(d)}$ is

$$f_d^{**}(u) \propto u^{d-1} \exp(-g(ur_d)) \tag{3.52}$$

The following Lemma provides intuition behind our use of the function $h(r)$.

Lemma 11 *Let r_d be any maximum of the marginal radial density function $f_d^*(\cdot)$ and let $g(\cdot), h(\cdot)$, and $f_d^{**}(\cdot)$ be defined as in (3.49), (3.50), and (3.52). Then*

$$(\log f_d^{**}(1))'' = -h(r_d)$$

Proof: *The derivatives of the log-density of the rescaled modulus are*

$$\begin{aligned}(\log f_d^{**}(u))' &= \frac{d-1}{u} - r_d g'(ur_d) \\ (\log f_d^{**}(u))'' &= -\left(\frac{d-1}{u^2} - r_d^2 g''(ur_d)\right)\end{aligned}$$

At the maximum, $u = 1$, and $d - 1 = r_d\ g'(r_d)$, from which

$$\begin{aligned}(\log f_d^{**}(u))'' &= -\left(r_d g'(r_d) + r_d^2 g''(r_d)\right) \\ &= -h(r_d)\end{aligned}$$

Thus $h(r)$ is the (negative) curvature of the log-density of the rescaled modulus at its maximum.

We will now show that for large enough d, and subject to a simple condition on $h(\cdot)$ the maxima r_d are unique and provide an increasing sequence tending to infinity. We will then show that using the r_d as a rescaling sequence produces the desired convergence in probability to 1.

Lemma 12 *If $h(r) > 0$ for all $r > r^*$ then for all $d > d_0$ (for some $d_0 < \infty$) the marginal radial density function has a single maximum r_d, with $r_d \to \infty$ monotonically as $d \to \infty$.*

Proof: *First note that*

$$\log f_d^*(r) = \log c_d + (d-1)\log r - g(r)$$

so that

$$(\log f_d^*(r))' = \frac{d-1}{r} - g'(r) \tag{3.53}$$

and any local maximum r_d *must satisfy*

$$r_d \; g'(r_d) = d - 1 \tag{3.54}$$

Differentiability of $g(r)$ *implies differentiability of* $f_d^*(r)$ *and since* $g(r)$ *is bounded for finite* r, $f_d^*(r)$ *has support thoughout* $[0,\infty)$. *Further,* $f_d^*(0) = 0$ *for* $d > 1$. *Any differentiable density function* $f_d^*(r)$ *with support throughout* $[0,\infty)$ *and with* $f_d^*(0) = 0$ *must have at least one local maximum (otherwise it could not integrate to* 1*).*

Set $d_0 = 1 + \sup_{r\in[0,r^*]} rg'(r)$ *so that by (3.54) for* $d > d_0$, *any maximum must occur at* $r_d > r^*$ *Now*

$$(rg'(r))' = h(r)/r > 0 \quad for \quad r > r^*$$

so for $r > r^*$, $rg'(r)$ *is always increasing. Thus for* $d > d_0$ *there can be at most one solution to* $rg'(r) = d-1$, *and* $f_d^*(r)$ *posseses exactly one local maximum.*

Further as $rg'(r)$ *is increasing and since* $r_d g'(r_d) = d-1$ *then* r_d *must increase monotonically with* d. *Suppose that it approaches some limit and so* $r_d \le r_0$ *for all* d. *Since* $g'(r)$ *is bounded for finite* r *it has a finite upper bound* b *on* $[0, r_0]$, *so* $d - 1 = r_d g'(r_d) \le r_0 b$ *which is a contradiction for all* $d > r_0 b + 1$.

It is actually possible to show far more for the rescaled modulus than simple convergence in probability to 1; the following is proved in Appendix C.

Lemma 13 *If there exist positive* k *and* a *such that*

$$h(r)/r^a \to k \quad as \quad r \to \infty \tag{3.55}$$

then there is a sequence $\{r_d\}$ *with* $r_d \to \infty$ *as* $d \to \infty$*, such that*

$$\frac{1}{a}\left(kr_d^a\right)^{1/2}\left(\left(\frac{\left|\mathbf{X}^{(d)}\right|}{r_d}\right)^a - 1\right) \xrightarrow{D} N(0,1)$$

The desired convergence in probability result follows immediately from Lemma 13 since $a > 0$ and $r_d \to \infty$ as $d \to \infty$.

Corollary 6 *If there exist positive* k *and* a *such that*

$$h(r)/r^a \to k \quad as \quad r \to \infty$$

then there is a sequence r_d *with* $r_d \to \infty$ *as* $d \to \infty$ *such that*

$$\frac{\left|\mathbf{X}^{(d)}\right|}{r_d} \xrightarrow{p} 1$$

The condition (3.55) implies that $h(r_d)$ is increasing polynomially in r_d, which itself is increasing in d, without bound. Thus the (negative) curvature of the log density at the maximum is increasing in d without bound, which intuitively corresponds to convergence in probability. Lemma 13 shows limiting normality and so is far stronger than is required for there to be a rescaling which produces convergence in probability to 1 of the modulii. It therefore seems likely that a weaker condition than (3.55) may be sufficient.

3.3.1.6 Limit theorems for expected acceptance rate and ESJD

We now return to the RWM and consider ESJD and expected acceptance rate on a unimodal spherically symmetric target as $d \to \infty$.

We have found that the scaled one-dimensional marginal distribution function of the target approaches some continuous limiting distribution function $\Theta(\cdot)$ as $d \to \infty$ whenever the rescaled radial distribution function approaches some limiting distribution function $\overline{\Theta}(\cdot)$ with $\overline{\Theta}(0) = 0$. In all further discussions the relation $\overline{\Theta}(0) = 0$ is assumed to hold. The limiting marginal distribution function is in general a scaled mixture of normal distribution functions, but in the special case that $\overline{\Theta}(r)$ has a single discontinuous step of height 1 at $r = 1$ (or at any other finite value) then the scaled mixture of normals reduces to the standard Gaussian cumulative distribution function $\Phi(\cdot)$.

Consider a sequence of jump proposal random variables $\{\mathbf{Y}^{(d)}\}$ with unit scale parameter. If there exist $k_y^{(d)}$ such that $\left|\mathbf{Y}^{(d)}\right| / k_y^{(d)}$ converges to unity (in a sense to be defined) then simple limit results are possible. Since the working statistician is free to choose $\mathbf{Y}^{(d)}$ this convergence condition can be ensured in advance. If the standard parametrisation of $\mathbf{Y}^{(d)}$, gives scaling $k_y^{(d)}$ we define $\tilde{\mathbf{Y}}^{(d)} := \mathbf{Y}^{(d)} / k_y^{(d)}$ so that $\tilde{\mathbf{Y}}^{(d)}$ converges (in a sense to be defined) to 1.

We also define a rescaled scaling parameter

$$\mu(d) := \frac{1}{2} \frac{d^{1/2} k_y^{(d)}}{k_x^{(d)}} \lambda \tag{3.56}$$

Applying Lemma 9 gives

$$F_{1|d}\left(-\frac{1}{2}\frac{k_x^{(d)}}{d^{1/2}}\lambda\left|\mathbf{Y}\right|\right)\rightarrow\Theta\left(-\frac{1}{2}\lambda\left|\mathbf{Y}\right|\right)\quad\text{uniformly}$$

for some continuous distribution function $\Theta(\cdot)$ representing either a Gaussian (Corollary 5) or a scaled mixture of Gaussians (Theorem 2).

The key result of this section (Theorem 3) requires two further Lemmas. The former is a standard result (see for example Grimmett and Stirzaker (2001) section 7.2), and the latter is proved in Appendix C.

Lemma 14 *For all bounded continuous functions* $g(\cdot)$

$$X_n\xrightarrow{D}\ X\Leftrightarrow\mathbb{E}\left[g(X_n)\right]\rightarrow\mathbb{E}\left[g(X)\right]$$

Lemma 15 *Let* $\{U_d\}$ *be a sequence of random variables such that*

$$U_d\xrightarrow{m.s.}\ 1$$

and let $\{G_d(\cdot)\}$ *be a sequence of functions with* $0\leq G_d(u)\leq 1$*. Then*

$$\mathbb{E}\left[G_d\left(U_d\right)\right]\rightarrow c\Rightarrow\mathbb{E}\left[U_d^2G_d\left(U_d\right)\right]\rightarrow c \tag{3.57}$$

Theorem 3 *Let* $\{\mathbf{X}^{(d)}\}$ *be a sequence of d-dimensional unimodal spherically symmetric targets and let* $\{\mathbf{Y}^{(d)}\}$ *be a corresponding sequence of jump proposals. If there exist* $k_x^{(d)}$ *such that* $\{\overline{F}_d\left(\cdot\right)\}$*, the sequence of marginal radial distribution functions of* $\{\left|\mathbf{X}^{(d)}\right|\}$*, satisfies* $\overline{F}_d\left(k_x^{(d)}\ x\right)\rightarrow\overline{\Theta}\left(x\right)$ *with* $\overline{\Theta}\left(0\right)=0$*, then for fixed* μ*,*

(i) If there exist $k_y^{(d)}$ *such that* $|\mathbf{Y}^{(d)}|/k_y^{(d)}\xrightarrow{p}\ 1$ *then*

$$\overline{\alpha}_d(\mu)\rightarrow 2\Theta\left(-\mu\right) \tag{3.58}$$

(ii) If in fact $|\mathbf{Y}^{(d)}|/k_y^{(d)} \xrightarrow{m.s.} 1$ *then*

$$\frac{d}{4{k_x^{(d)}}^2} S_d^2(\mu) \to 2\mu^2 \Theta(-\mu) \tag{3.59}$$

where $\Theta(x)$ *is the marginal one-dimensional distribution function corresponding to* $\overline{\Theta}(x)$.

Proof: *First note that* $-\frac{1}{2}\lambda \left|\mathbf{Y}^{(d)}\right| = -\mu \left|\tilde{\mathbf{Y}}^{(d)}\right| \times k_x^{(d)}/d^{1/2}$ *whence (3.31) and (3.32) become*

$$\begin{aligned} \overline{\alpha}_d(\mu) &= 2\mathbb{E}\left[F_{1|d}\left(-\mu\left|\tilde{\mathbf{Y}}^{(d)}\right|\frac{k_x^{(d)}}{d^{1/2}}\right)\right] \\ S_d^2(\mu) &= \frac{8\mu^2{k_x^{(d)}}^2}{d}\mathbb{E}\left[\left|\tilde{\mathbf{Y}}^{(d)}\right|^2 F_{1|d}\left(-\mu\left|\tilde{\mathbf{Y}}^{(d)}\right|\frac{k_x^{(d)}}{d^{1/2}}\right)\right] \end{aligned} \tag{3.60}$$

(i) $\Theta(-\mu x)$ *is bounded, and continuous by Lemma 10 and we are given that* $\left|\tilde{\mathbf{Y}}^{(d)}\right| \xrightarrow{D} 1$, *so by Lemma 14*

$$\mathbb{E}\left[\Theta\left(-\mu\left|\tilde{\mathbf{Y}}^{(d)}\right|\right)\right] \to \mathbb{E}\left[\Theta(-\mu)\right] = \Theta(-\mu)$$

and given $\epsilon > 0$ *we may find a* d_1 *such that for all* $d > d_1$

$$\left|\mathbb{E}\left[\Theta\left(-\mu\left|\tilde{\mathbf{Y}}^{(d)}\right|\right)\right] - \Theta(-\mu)\right| < \frac{\epsilon}{2}$$

Also from Theorem 2 $\exists\ d_2$ *such that for all* $d > d_2$ *and for all* $\left|\tilde{\mathbf{Y}}^{(d)}\right|$

$$\left|F_{1|d}\left(-\mu\left|\tilde{\mathbf{Y}}^{(d)}\right|\frac{k_x^{(d)}}{d^{1/2}}\right) - \Theta\left(-\mu\left|\tilde{\mathbf{Y}}^{(d)}\right|\right)\right| < \frac{\epsilon}{2}$$

Thus

$$\mathbb{E}\left[\left|F_{1|d}\left(-\mu\left|\tilde{\mathbf{Y}}^{(d)}\right|\frac{k_x^{(d)}}{d^{1/2}}\right) - \Theta\left(-\mu\left|\tilde{\mathbf{Y}}^{(d)}\right|\right)\right|\right] < \frac{\epsilon}{2}$$

Writing

$$F_{1|d}\left(-\mu|\mathbf{y}|\frac{k_x^{(d)}}{d^{1/2}}\right)-\Theta\left(-\mu\right)=\left(F_{1|d}\left(-\mu|\mathbf{y}|\frac{k_x^{(d)}}{d^{1/2}}\right)-\Theta\left(-\mu|\mathbf{y}|\right)\right)$$
$$+\left(\Theta\left(-\mu\left|\mathbf{y}\right|\right)-\Theta\left(-\mu\right)\right) \quad (3.61)$$

we separate the two expectations and apply the triangle inequality for $d > max(d_1, d_2)$ *to find*

$$\left|\mathbb{E}\left[F_{1|d}\left(-\mu\left|\tilde{\mathbf{Y}}^{(d)}\right|\frac{k_x^{(d)}}{d^{1/2}}\right)-\Theta\left(-\mu\right)\right]\right|\leq\epsilon$$

(ii) Since convergence in mean square implies convergence in probability the first part of the Theorem continues to hold. The second half then follows directly from (3.60) and Lemma 15, by substituting $U_d=\left|\tilde{\mathbf{Y}}^{(d)}\right|$, $c=2\Theta(-\mu)$, *and* $G_d=2F_{1|d}\left(-\mu\left|\tilde{\mathbf{Y}}^{(d)}\right|\frac{k_x^{(d)}}{d^{1/2}}\right)$.

The special case where $\left|\mathbf{X}^{(d)}\right|/k_x^{(d)}\xrightarrow{p}1$ merits explicit note.

Corollary 7 *Let* $\mathbf{X}^{(d)}$ *be a d-dimensional unimodal spherically symmetric target distribution and let* $\mathbf{Y}^{(d)}$ *be the jump proposal distribution. If there exist* $k_x^{(d)}$ *such that* $\left|\mathbf{X}^{(d)}\right|/k_x^{(d)}\xrightarrow{p}1$ *then for fixed* μ,

(i) If there exist $k_y^{(d)}$ *such that* $|\mathbf{Y}^{(d)}|/k_y^{(d)}\xrightarrow{p}1$ *then*

$$\overline{\alpha}_d(\mu)\to 2\Phi\left(-\mu\right) \quad (3.62)$$

(ii) If in fact $|\mathbf{Y}^{(d)}|/k_y^{(d)}\xrightarrow{m.s.}1$ *then*

$$\frac{d}{4{k_x^{(d)}}^2}S_d^2(\mu)\to 2\mu^2\Phi\left(-\mu\right) \quad (3.63)$$

where $\Phi(x)$ *is the cumulative distribution function of a standard Gaussian.*

With these asymptotic forms for expected acceptance rate and ESJD we are finally equipped to examine the issue of optimal-scaling in the limit as $d\to\infty$.

3.3.1.7 The existence of an asymptotically optimal scaling

It was shown in section 3.3.1.2 that there is at least one finite optimal scaling for any spherically symmetric unimodal finite-dimensional target with finite second moment provided the second moment of the proposal is also finite. We now investigate the existence of a finite optimal scaling as $d \to \infty$. We assume throughout that there is a sequence $\{k_y^{(d)}\}$ such that the rescaled proposal satisfies $\left|\mathbf{Y}^{(d)}\right| / k_y^{(d)} \xrightarrow{m.s.} 1$.

We first consider the special case where there is a sequence $k_x^{(d)}$ such that the rescaled target satisfies $\left|\mathbf{X}^{(d)}\right| / k_x^{(d)} \xrightarrow{p} 1$. Differentiating (3.63) we see that in the limit, as $d \to \infty$

$$\frac{1}{\mu}\frac{dS_d^2}{d\mu} \propto D_p(\mu) := 2\Phi(-\mu) - \mu\phi(-\mu)$$

This is plotted in Figure 3.4(a) and is zero for a maximum in the ESJD at

$$\hat{\mu}_p :\approx 1.19 \tag{3.64}$$

Substituting into (3.62) provides the expected acceptance rate at this optimal scaling

$$\hat{\alpha}_p :\approx 0.234 \tag{3.65}$$

More generally we have $\left|\mathbf{X}^{(d)}\right| / k_x^{(d)} \xrightarrow{D} R$. First consider the equivalent cases of classes of target distributions for which limiting rescaled radius has a point mass at zero or infinity or both. A point mass p at infinity for R is not strictly forbidden in Theorem 3. However such a point mass does imply (using Theorem 2) that $\Theta(x_1) \to p/2$ as $x_1 \to -\infty$. Therefore the rescaled ESJD given in (3.59) tends to infinity as $\mu \to \infty$ and there is no finite optimal scaling $\hat{\mu}$ corresponding to $k_x^{(d)}$. Of course, if the limiting marginal radial distribution contains a point mass

at infinity then there may be a possible alternative rescaling $k_x^{*(d)}$ that shifts this mass to $(0, \infty)$; however any mass that was in $(0, \infty)$ with the first scaling $k_x^{(d)}$ would then move to the origin; and this *is* strictly forbidden in Theorem 3. In general, targets that vary on at least two very different scales are not amenable to the current approach. Indeed the very idea that there is a single optimal-scaling is highly debatable. A simulation study is conducted to this end in Section 3.3.4.3 and the concept of an optimal scaling in such cases is discussed. For the remaining theory in this Chapter however we only consider targets for which the rescaled limiting radius has no point mass at either the origin or infinity.

Differentiate (3.59) to find that in the limit as $d \to \infty$

$$\frac{1}{\mu}\frac{dS_d^2}{d\mu} \propto D(\mu) := 2\Theta(-\mu) - \mu\theta(-\mu) = 2\mathbb{E}_{\overline{\theta}(r)}\left[\Phi\left(-\frac{\mu}{R}\right)\right] - \mathbb{E}_{\overline{\theta}(r)}\left[\frac{\mu}{R}\phi\left(\frac{\mu}{R}\right)\right]$$

We seek zeros of $D(\cdot)$ that correspond to maxima. Substituting $V = R/\mu$ gives

$$D(\mu) = \mathbb{E}_{\mu\overline{\theta}(\mu v)}\left[2\Phi\left(-\frac{1}{V}\right) - \frac{1}{V}\phi\left(\frac{1}{V}\right)\right]$$

Figure 3.3 shows the graph of $2\Phi\left(-\frac{1}{v}\right) - \frac{1}{v}\phi\left(\frac{1}{v}\right)$, which has a zero at $v^* \approx 1/1.19$, is positive for all $v > v^*$ and asymptotes to 1. By making μ sufficiently small, as much of the mass as we like of Lesbegue density $\mu\overline{\theta}(\mu v)$ can be made to reside in the interval $v \in (v^*, \infty)$. Therefore $D(\mu)$ is always positive for small enough μ (in other words the ESJD always increases initially as the scaling parameter is increased from zero). For an optimal scaling to exist for a continuously differentiable ESJD we require $D(\mu) < 0$ for some finite μ. As we shall discover, this condition will not always hold.

Here we consider three further examples; the respective derivative functions $D(\mu)$ are plotted in Figure 3.4.

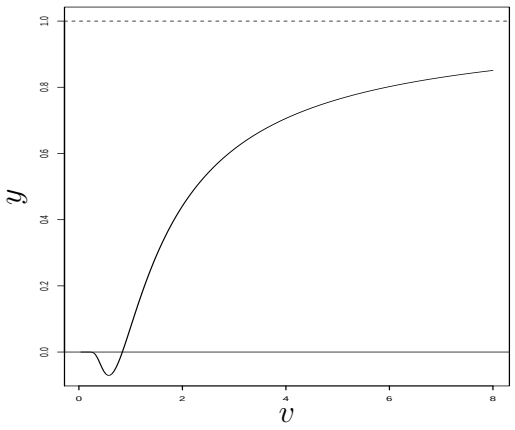

Figure 3.3: Graph of $y = 2\Phi\left(-\frac{1}{v}\right) - \frac{1}{v}\phi\left(\frac{1}{v}\right)$

- If the marginal radial distribution is exponential, $\mu\overline{\theta}(\mu r) = \mu e^{-\mu r}$, and the derivative function $D(\mu) < 0$ for all $\mu > 2.86$ (Figure 3.4(b)).

- Consider a slight alteration of the example density (3.47) from Section 3.3.1.5: $f_d(x) \propto \mathbf{1}_{\{x \leq 1\}} + \exp\left(-\frac{1}{2}(\log x)^2\right)\mathbf{1}_{\{x>1\}}$. This is unimodal (in the extended sense of Section 3.2.3) and has the same limiting marginal radial distribution (3.48). Its tail decays more slowly than the exponential example above and there is no finite optimal μ (Figure 3.4(c)).

- Consider R as the polynomially tailed t_3 distribution, which has finite second moment. For this also there is no finite optimal μ (Figure 3.4(d)).

We now briefly examine the relationship between the limit of the maxima of the finite dimensional expected square jump distances $\{S_d^2\}$ and the maximum (or lack thereof) of the limiting expected square jump distance. A proof of the following result is given in the Appendix C.

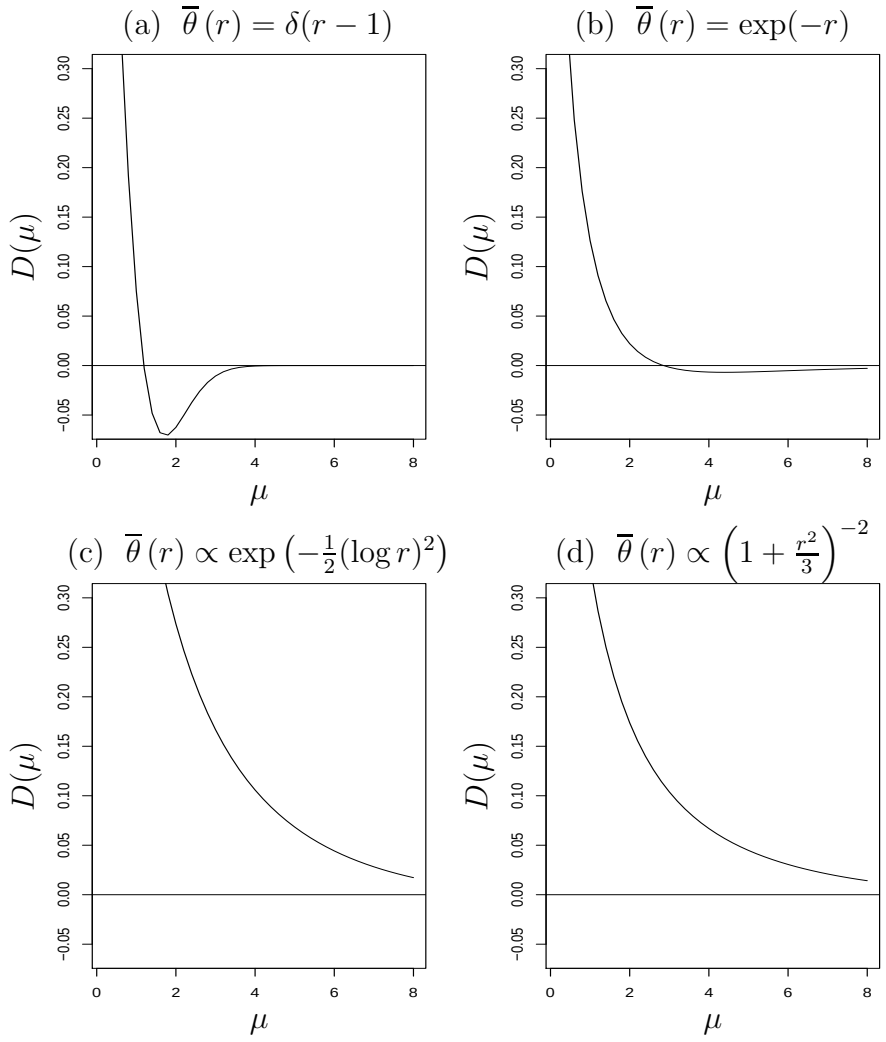

Figure 3.4: The derivative function $D(\mu)$ of the expected squared jump distance when the limiting radial distribution is a) a point mass at 1; b) the unit exponential; c) the heavy tailed limit (3.48); d) Student's t with 3 degrees of freedom.

Lemma 16 *Let $\{h_d(x)\}$ be a sequence of functions defined on $[0, \infty)$, Let the pointwise limit of the sequence, $h(x)$, be continuously differentiable and have a finite number of local maxima.*

(i) If $h(x)$ has a local maximum at $x^* < \infty$ then x^* is a limit point of local maxima of $h_d(x)$.

(ii) If $h(x)$ is strictly monotonically increasing in x then denoting by x_d^* the smallest x at which $h_d(x)$ achieves its global maximum: $\lim_{d\to\infty} x_d^* = \infty$

We will be working with a form for the limiting ESJD. The scaling μ which maximises this is usually the same as the limit of the scalings that optimise the sequence of ESJD's. However there are sequences of distributions where this is not the case, such as the mixture of normals in (3.99). Statements made through the remainder of this chapter concern optimising the limit function. To clarify this we define the **asymptotically optimal rescaled scaling** (AORS) to be the value $\hat{\mu}$ that optimises the limiting ESJD, simiarly the **asymptotically optimal scaling** (AOS) is defined to be $\hat{\lambda}_d = (2k_x^{(d)}\hat{\mu})/(d^{1/2}k_y^{(d)})$. Finally the **asymptotically optimal acceptance rate** (AOA) is the limiting expected acceptance rate that results from using the AORS.

Consider a sequence of random variables with limiting (rescaled) marginal radial distribution corresponding to Figure 3.4(a) or 3.4(b). If the ESJD of each element in the sequence has a single maximum, then by Lemma 16 the limit of these maxima is the maximum of the limit function, subject to the scaling $k_x^{(d)}$.

Alternatively for sequences of random variables with limiting (rescaled) marginal radial distribution functions as in Figures 3.4(c) or 3.4(d) the limiting μ^* is ∞,

and the limiting optimal acceptance rate is therefore zero. Note that if in an attempt to counteract this effect, bigger rescalings $k_x^{(d)}$ were used, the limiting radial distribution would simply be a point mass at zero.

3.3.1.8 Asymptotically optimal scaling and acceptance rate

For μ to be optimal we require $D(\mu) = 0$, or

$$2\Theta(-\mu) = \mu\theta(-\mu) \tag{3.66}$$

There may not always be a solution for μ (see Section 3.3.1.7) but when there is, denote this value as $\hat{\mu}$. Optimal scaling is therefore achieved by setting $\mu = \hat{\mu}$, so that re-arranging (3.56) we obtain the following corollary to Theorem 3:

Corollary 8 *Let $\{\mathbf{X}^{(d)}\}$ be a sequence of d-dimensional spherically symmetric unimodal target distributions and let $\{\mathbf{Y}^{(d)}\}$ be a sequence of jump proposal distributions. If there exist $k_x^{(d)}$ and $k_y^{(d)}$ such that the marginal radial distribution function of $\mathbf{X}^{(d)}$ satisfies $\left|\mathbf{X}^{(d)}\right|/k_d \xrightarrow{D} R$ where R has distribution function $\overline{\Theta}(r)$ with $\overline{\Theta}(0) = 0$, and $|\mathbf{Y}^{(d)}|/k_y^{(d)} \xrightarrow{m.s.} 1$ and provided there is a solution $\hat{\mu}$ to (3.66) then the asymptotically optimal scaling is*

$$\hat{\lambda}_d = 2\hat{\mu}\,\frac{k_x^{(d)}}{d^{1/2}k_y^{(d)}}$$

This has some especially interesting consequences in certain specific situations, which are investigated further in Section 3.3.4.2:

If we have chosen to propose jumps from the same distribution as the target (up to a scaling constant), or more generally if $k_x^{(d)} = k_y^{(d)}$ then AOS is $\hat{\lambda}_d = 2\hat{\mu}/d^{1/2}$

If $k_x^{(d)}/k_y^{(d)} \propto d^{1/2}$ (e.g. exponential target and Gaussian proposal) then the AOS $\hat{\lambda}_d$ is a fixed non-zero constant.

As discussed in Section 3.3.1.7, if $\left|\mathbf{X}^{(d)}\right| / k_x^{(d)} \xrightarrow{p} 1$ we obtain a AORS of $\hat{\mu}_p \approx 1.1906$ and a corresponding acceptance rate of $\hat{\alpha}_p \approx 0.234$.

Let us now explore the asymptotically optimal acceptance rate, if it exists. We will show that the AOA is at most $\hat{\alpha}_p$ with the upper bound achieved if and only if the marginal radius of the rescaled target converges in probability to 1.

Define

$$\overline{\alpha}_\infty(\mu) := \lim_{d\to\infty} \overline{\alpha}_d(\mu)$$

From Theorems 2 and 3

$$\overline{\alpha}_\infty(\mu) = 2\Theta(-\mu) = 2\mathbb{E}_R\left[\Phi\left(\frac{-\mu}{R}\right)\right]$$

where R is the marginal radius of the limit of the sequence of scaled targets and $P(R \leq r) = \overline{\Theta}(r)$. In the theorem that follows the rescaled marginal radial distribution must have no point mass at 0 - this is an explicit condition of Theorem 3, on which this theorem relies. Following the discussion in Section 3.3.1.7 the Theorem can only apply to distributions for which the limiting marginal radius has no point mass at infinity. It is not necessary to state this however since it is implied by the condition that there be an optimal value $\hat{\mu}$.

Theorem 4 *Let $\{\mathbf{X}^{(d)}\}$ be a sequence of d-dimensional spherically symmetric unimodal targets and let $\{\mathbf{Y}^{(d)}\}$ be a sequence of jump proposals. Let there exist $k_x^{(d)}$ and $k_y^{(d)}$ such that $|\mathbf{Y}^{(d)}|/k_y^{(d)} \xrightarrow{m.s.} 1$ and $\left|\mathbf{X}^{(d)}\right| / k_x^{(d)} \xrightarrow{D} R$ for some R with no point mass at 0. If there is an asymptotically optimal acceptance rate it is*

$$\overline{\alpha}_\infty(\mu) \leq \hat{\alpha}_p \approx 0.234$$

Equality is achieved if and only if there exist $k_x^{(d)}$ *such that* $\left|\mathbf{X}^{(d)}\right| / k_x^{(d)} \xrightarrow{p} 1$.

Proof: *Observe that*

$$\theta(-\mu) = \mathbb{E}\left[\frac{1}{R}\phi\left(-\frac{\mu}{R}\right)\right]$$

So (3.66) becomes

$$2\mathbb{E}\left[\Phi\left(-\frac{\mu}{R}\right)\right] = \mathbb{E}\left[\frac{\mu}{R}\phi\left(-\frac{\mu}{R}\right)\right] \tag{3.67}$$

For a given distribution of R*, this has solution* $\hat{\mu}$*, from which the optimal acceptance rate is*

$$\hat{\alpha} := \overline{\alpha}_\infty(\hat{\mu}) = 2\mathbb{E}\left[\Phi\left(-\frac{\hat{\mu}}{R}\right)\right]$$

Now substitute $V := \Phi\left(-\frac{\hat{\mu}}{R}\right)$ *so that for* $\mu \geq 0$ *and* $R \geq 0$ *we have* $v \in [0, 0.5]$*. Also define*

$$h(v) := -\Phi^{-1}(v)\ \phi(\Phi^{-1}(v))$$

The optimal acceptance rate is therefore

$$\hat{\alpha} = 2\ \mathbb{E}[V]$$

and (3.67) is satisfied, becoming

$$2\ \mathbb{E}[V] = \mathbb{E}[h(V)] \tag{3.68}$$

But

$$\frac{d^2h}{dv^2} = 2\ \frac{\Phi^{-1}(v)}{\phi(\Phi^{-1}(v))} \leq 0 \quad \text{for } v \in [0, 0.5]$$

the inequality being strict for $v \in (0, .5)$*, which corresponds to* $r \in (0, \infty)$*. Therefore by Jensen's inequality*

$$\mathbb{E}[h(V)] \leq h(\mathbb{E}[V]) \tag{3.69}$$

Since the second derivative of $h(\cdot)$ *is strictly negative except at the (finite) end points, equality is achieved if and only if all the mass in* V *is concentrated in one*

place v_0; this corresponds to all the mass in R being concentrated at $-\hat{\mu}/\Phi^{-1}(v_0)$. This is exactly the situation $\left|\mathbf{X}^{(d)}\right|/k_x^{(d)} \xrightarrow{p} 1$.

Substitute $m := -\Phi^{-1}(\mathbb{E}[V])$, so that (3.68) and (3.69) combine to give

$$2\Phi(-m) \leq m\ \phi(m)$$

When there is equality the single solution to this equation is $\hat{m} = \hat{\mu}_p$. Figure 3.4 (a) shows the graph of $y = 2\Phi(-v) - v\ \phi(-v)$ from which we observe that the inequality is strict if and only if $m > \hat{\mu}_p$ and hence $2\Phi(-m) \leq 2\Phi(-\hat{\mu}_p)$.

Therefore the optimal acceptance rate is

$$\hat{\alpha} = \mathbb{E}[V] = 2\Phi(-m) \leq 2\phi(-\hat{\mu}_p) \approx 0.234$$

with equality achieved if and only if $\left|\mathbf{X}^{(d)}\right|/k_x^{(d)} \xrightarrow{p} 1$.

We considered limiting forms for $\left|\mathbf{X}^{(d)}\right|/k_x^{(d)}$ in Section 3.3.1.5; in Section 3.3.4.2. we will examine acceptance rate for specific examples where convergence in probability to 1 is not achieved.

3.3.2 Elliptically symmetric distributions

As discussed in Section 3.1.4 an elliptically symmetric target $\mathbf{X}$ may be defined in terms of an associated orthogonal linear map $\mathbf{T}$ such that $\mathbf{T}(\mathbf{X})$ is spherically symmetric with unit scale parameter. We write $\mathbf{X}_* = \mathbf{T}(\mathbf{X})$ and $\mathbf{Y}_*$ for the corresponding jump proposal. Since $\mathbf{T}(\cdot)$ is linear the jump proposal satisfies

$$\mathbf{Y}_* = \mathbf{T}(\mathbf{X} + \mathbf{Y}) - \mathbf{T}(\mathbf{X}) = \mathbf{T}(\mathbf{Y})$$

The expected square jump distance (3.15) is preserved under the transformation since (choosing axes corresponding to the principal axes)

$$\mathbb{E}\left[\sum \frac{1}{\beta_i^2} Y_i^2\right] = \mathbb{E}\left[\sum Y_{*i}^2\right]$$

Naive ESJD (3.14) is clearly not preserved. Suppose now that the target $\mathbf{X}$ has elliptical contours monotonically decreasing from the origin. Since $\mathbf{X}_*$ is spherically symmetric, the acceptance region in the transformed space is the region where $|\mathbf{X}_* + \mathbf{Y}_*| < |\mathbf{X}_*|$. Further, the quantity we wish to optimise is $|\mathbf{Y}_*|^2$. We may therefore apply (3.31) and (3.32) in the transformed space. Write $F_{1|d}^*(\cdot)$ for the one-dimensional marginal density of spherically symmetric $\mathbf{X}_* = \mathbf{T}(\mathbf{X})$, and recall that $\mathbf{Y}_* = \mathbf{T}(\mathbf{Y})$, where $\mathbf{Y}$ has unit scale parameter. We wish to optimise the ESJD

$$S_d^2(\lambda) := 2\lambda^2 \mathbb{E}\left[|\mathbf{Y}_*|^2 F_{1|d}^*\left(-\frac{1}{2}\lambda|\mathbf{Y}_*|\right)\right] \tag{3.70}$$

Here expectation is with respect to Lesbegue measure $r_*(\cdot)$ of $\mathbf{Y}_*$. The acceptance rate in the original space is the same as that in the transformed space and is therefore given by

$$\overline{\alpha}_d(\lambda) = 2\mathbb{E}\left[F_{1|d}^*\left(-\frac{1}{2}\lambda\,|\mathbf{Y}_*|\right)\right] \tag{3.71}$$

Similar explicit formulae apply for calculation of acceptance rate and ESJD in terms of double integrals in the transformed space (analogues of (3.44) and (3.45)). Furthermore Lemma 6 and Corollaries 3 and 4 are now seen to hold for all unimodal elliptically symmetric targets.

Our goal is now to find conditions on the scale parameters β_i such that Corollary 7 applies in the transformed space. When dealing with each target random variable

with elliptical contours we will convert it to a spherically symmetric random variable by a simple invertible linear map. For the limit results to continue to apply we will need the condition of convergence in probability to 1 of the modulii of the rescaled targets and the condition of mean square convergence of the modulii of the rescaled jump proposals to carry through the transformation. Section 3.3.2.1 investigates convergence implications for sequences of orthogonal linear maps on sequences of spherically symmetric distributions in general and Section 3.3.2.2 relates these results to the random walk Metropolis.

3.3.2.1 Orthogonal linear maps on spherically symmetric distributions

In this section we show that, subject to conditions on the eigenvalues of a sequence of orthogonal linear maps between spherically symmetric and elliptically symmetric random variables, both convergence in probability and convergence in mean square of the modulus do carry through the mapping. Throughout this section and the next we will employ the following shorthand for the arithmetic mean of the squares of a set of d-dependent scalar values $\alpha_1(d), \ldots, \alpha_d(d)$

$$\overline{\alpha^2}(d) := \frac{1}{d}\sum_{1}^{d} \alpha_i(d)^2$$

We employ a similar shorthand for the corresponding harmonic mean

$$\tilde{\alpha^2}(d) := \left(\frac{1}{d}\sum_{1}^{d} \frac{1}{\alpha_i(d)^2}\right)^{-1}$$

In Lemma 18 we will relate convergence in probability of the modulus of linear mappings on a general set of spherical distribution with increasing dimension to convergence in probability of the sequence modulii of isotropic Gaussians. As prelude we first offer a condition for convergence in probability of a sequence of linear combinations of Chi-squared random variables.

Lemma 17 *Let $V_1, V_2, \ldots$ be a sequence of independent identically distributed χ^2_1 random variables and let $a_1(d), a_2(d), \ldots$ be a sequence of sequences of positive coefficients. Define $a_{max}(d) := \max_{i=1,\ldots d} a_i(d)$. Then*

$$\frac{\sum_1^d a_i(d)V_i}{\sum_1^d a_i(d)} \xrightarrow{p} 1 \qquad \text{if and only if} \qquad \frac{a_{max}(d)}{\sum_1^d a_i(d)} \to 0$$

Proof: *We first prove implication from left to right. If the right hand side fails then there is some $\delta > 0$ such that there exists a sequence $d_1, d_2, \ldots$ tending to infinity for which*

$$\frac{a_{max}(d_k)}{\sum_1^{d_k} a_i(d_k)} = \delta_k \geq \delta$$

Since all the V_i are non-negative

$$\begin{aligned} V_{j_{max}} \geq \frac{1+\epsilon}{\delta} &\Rightarrow \delta_k V_{j_{max}} \geq 1+\epsilon \\ &\Rightarrow \delta_k V_{j_{max}} + \frac{\sum_{-j_{max}} a_i(d_k)V_i}{\sum_1^{d_k} a_i(d_k)} \geq 1+\epsilon \\ &\Rightarrow \frac{\sum_1^{d_k} a_i(d_k)V_i}{\sum_1^{d_k} a_i(d_k)} - 1 \geq \epsilon \\ &\Rightarrow \left| \frac{\sum_1^{d_k} a_i(d_k)V_i}{\sum_1^{d_k} a_i(d_k)} - 1 \right| \geq \epsilon \end{aligned}$$

where $j_{max}(k)$ is the subscript of the largest eigenvalue $a_{max}(d_k)$, and $\sum_{-j_{max}}$ indicates a sum over all indices from $1 \ldots d_k$ excluding j_{max}. Therefore

$$P\left(\left| \frac{\sum_1^{d_j} a_i(d_j)V_i}{\sum_1^{d_j} a_i(d_j)} - 1 \right| \geq \epsilon \right) \geq P\left(V_{j_{max}} \geq \frac{1+\epsilon}{\delta} \right) = P\left(\chi^2_1 > \frac{1+\epsilon}{\delta} \right) > 0$$

We now prove the implication from right to left. Define $X_d := \sum_1^d a_i(d)V_i$ and note that $\mathbb{E}[X_d] = \sum_1^d a_i(d)$, and since the V_i are independent, $Var[X_d] = 2\sum_1^d a_i(d)^2$.

Therefore

$$\frac{Var[X_d]}{(\sum_1^d a_i(d))^2} = \frac{2\sum_1^d a_i(d)^2}{\left(\sum_1^d a_i(d)\right)^2} \leq \frac{2a_{max}\sum_1^d a_i(d)}{\left(\sum_1^d a_i(d)\right)^2} = \frac{2a_{max}}{\sum_1^d a_i(d)}$$

Combining this with Chebyshev's inequality we obtain

$$\begin{aligned} P\left(\left|\frac{X_d}{\sum_1^d a_i(d)} - 1\right| \geq \epsilon\right) &= P\left(\left|\frac{X_d - \sum_1^d a_i(d)}{\sum_1^d a_i(d)}\right| \geq \epsilon\right) \\ &\leq \frac{Var[X_d]}{\epsilon^2 \left(\sum_1^d a_i(d)\right)^2} \\ &\leq \frac{2a_{max}}{\epsilon^2 \sum_1^d a_i(d)} \end{aligned}$$

For a given $\epsilon > 0$ this can be made as small as we like.

Consider orthogonal linear map $\mathbf{S}$ with eigenvalues α_i applied to $\mathbf{Z} \sim N(\mathbf{0}, \mathbf{I}_d)$. Since $\mathbf{Z}$ is spherical we may without loss of generality choose as axes the principal axes of the linear map. Thus $|\mathbf{S}(\mathbf{Z})|^2 = \sum \alpha_i^2 Z_i^2$ which is the form described by Lemma 17. This translates to the following corollary:

Corollary 9 *Let $\mathbf{Z} \sim N(\mathbf{0}, \mathbf{I}_d)$ and orthogonal linear transformation $\mathbf{S}^{(d)} : \Re^d \to \Re^d$ have eigenvalues $\alpha_1(d), \ldots, \alpha_d(d)$. Define $\alpha_{max}(d) := \max_{i=1,\ldots d} \alpha_i(d)$. Then*

$$\frac{\left|\mathbf{S}^{(d)}(\mathbf{Z})\right|}{\left(\sum_1^d \alpha_i(d)^2\right)^{1/2}} \xrightarrow{p} 1 \tag{3.72}$$

if and only if

$$\frac{\alpha_{max}(d)^2}{\sum_1^d \alpha_i(d)^2} \to 0 \tag{3.73}$$

We now relate our general $\mathbf{U}$ to a Gaussian $\mathbf{Z}$, first tackling the continued convergence in probability to 1 of a transformed target and then examining convergence in mean square.

Lemma 18 *Let $\mathbf{S}^{(d)}$ be a sequence of orthogonal linear maps on $\Re^d$ with eigenvalues $\alpha_1(d), \ldots, \alpha_d(d)$. Write $\overline{\alpha^2}(d) := \frac{1}{d}\sum_1^d \alpha_i(d)^2$ and let $\mathbf{U}^{(d)}$ be a sequence of spherically symmetric random variables in $\Re^d$. Then*

$$\frac{\left|\mathbf{S}^{(d)}\left(\mathbf{U}^{(d)}\right)\right|}{\left(\overline{\alpha^2}(d)\right)^{1/2}} \xrightarrow{p} 1 \Leftrightarrow \left|\mathbf{U}^{(d)}\right| \xrightarrow{p} 1$$

provided the eigenvalues of $\mathbf{S}^{(d)}$ *satisfy (3.73).*

Proof: *We decompose* $\mathbf{U}^{(d)}$ *into a Gaussian related to direction and a scalar length as in the introduction to Theorem 1, so*

$$\left|\mathbf{S}^{(d)}\left(\mathbf{U}^{(d)}\right)\right| = \left|\mathbf{S}^{(d)}\left(\frac{\mathbf{Z}^{(d)}}{\left|\mathbf{Z}^{(d)}\right|}R^{(d)}\right)\right| = \frac{R^{(d)}}{\left|\mathbf{Z}^{(d)}\right|}\left|\mathbf{S}^{(d)}\left(\mathbf{Z}^{(d)}\right)\right| \tag{3.74}$$

From Lemma 8, $\left|\mathbf{Z}^{(d)}\right|/d^{1/2} \xrightarrow{p} 1$ *so that*

$$\frac{\left|\mathbf{S}^{(d)}\left(\mathbf{U}^{(d)}\right)\right|}{\left(\overline{\alpha^2}(d)\right)^{1/2}} \xrightarrow{p} 1 \Leftrightarrow \frac{1}{d^{1/2}}\frac{\left|\mathbf{S}^{(d)}\left(\mathbf{Z}^{(d)}\right)\right|}{\left(\overline{\alpha^2}(d)\right)^{1/2}}R^{(d)} \xrightarrow{p} 1$$

Since $R^{(d)} = \left|\mathbf{U}^{(d)}\right|$*, applying Corollary 9 gives the desired result immediately.*

Proof of a corresponding result for convergence in mean square requires the following simple scaling relation between the second moments.

Lemma 19 *Let* $\mathbf{S}^{(d)}$ *be a sequence of orthogonal linear maps on* $\Re^d$ *with eigenvalues* $\alpha_1(d), \ldots, \alpha_d(d)$ *and let* $\mathbf{U}^{(d)}$ *be a sequence of spherically symmetric random variables in* $\Re^d$ *then*

$$\mathbb{E}\left[\left|\mathbf{U}^{(d)}\right|^2\right] = \frac{1}{\overline{\alpha^2}(d)}\mathbb{E}\left[\left|\mathbf{S}^{(d)}\left(\mathbf{U}^{(d)}\right)\right|^2\right]$$

Proof: *Since* $\mathbf{U}^{(d)}$ *is spherically symmetric we may without loss of generality consider it with axes along the principal components of* $\mathbf{S}^{(d)}$*. Then*

$$\mathbb{E}\left[\left|\mathbf{S}^{(d)}\left(\mathbf{U}^{(d)}\right)\right|^2\right] = \mathbb{E}\left[\sum_1^d \alpha_i(d)^2\left(U_i^{(d)}\right)^2\right] = \sum_1^d \alpha_i(d)^2\ \mathbb{E}\left[\left(U_i^{(d)}\right)^2\right]$$

But $\mathbf{U}^{(d)}$ *is spherically symmetric so this is*

$$\sum_1^d \alpha_i(d)^2\mathbb{E}\left[\left(U_1^{(d)}\right)^2\right] = \frac{1}{d}\sum_1^d \alpha_i(d)^2\sum_1^d \mathbb{E}\left[\left(U_i^{(d)}\right)^2\right] = \overline{\alpha^2}\ \mathbb{E}\left[\left|\mathbf{U}^{(d)}\right|^2\right]$$

We will also require Scheffe's Lemma (e.g. Williams, 1991) which relates convergence of the second moment and convergence in probability (to a constant) with convergence in expectation.

Lemma 20 *(**Scheffe's Lemma**): If* $\mathbb{E}\left[\left(Y^{(n)}\right)^2\right] \to 1$ *and* $Y^{(n)} \xrightarrow{p} 1$ *then* $\mathbb{E}\left[Y^{(n)}\right] \to 1$.

We are now in a position to prove the analogue of Lemma 18 for convergence in mean square.

Lemma 21 *Let* $\mathbf{S}^{(d)}$ *be a sequence of orthogonal linear maps on* $\Re^d$ *with eigenvalues* $\alpha_1(d), \ldots, \alpha_d(d)$ *and let* $\mathbf{U}^{(d)}$ *be a sequence of spherically symmetric random variables in* $\Re^d$. *Then*

$$\left|\mathbf{U}^{(d)}\right| \xrightarrow{m.s.} 1 \Leftrightarrow \frac{\left|\mathbf{S}^{(d)}\left(\mathbf{U}^{(d)}\right)\right|}{\left(\overline{\alpha^2}(d)\right)^{1/2}} \xrightarrow{m.s.} 1$$

provided the eigenvalues of $\mathbf{S}^{(d)}$ *satisfy (3.73).*

Proof: *From Lemma 19*

$$\mathbb{E}\left[\left(\frac{\left|\mathbf{S}^{(d)}\left(\mathbf{U}^{(d)}\right)\right|}{\left(\overline{\alpha^2}(d)\right)^{1/2}} - 1\right)^2\right] - \mathbb{E}\left[\left(\left|\mathbf{U}^{(d)}\right| - 1\right)^2\right] = -2\left(\mathbb{E}\left[\frac{\left|\mathbf{S}^{(d)}\left(\mathbf{U}^{(d)}\right)\right|}{\left(\overline{\alpha^2}(d)\right)^{1/2}}\right] - \mathbb{E}\left[\left|\mathbf{U}^{(d)}\right|\right]\right)$$

So it is sufficient to prove that subject to (3.73)

$$\mathbb{E}\left[\left|\mathbf{U}^{(d)}\right|\right] \to 1 \Leftrightarrow \mathbb{E}\left[\frac{\left|\mathbf{S}^{(d)}\left(\mathbf{U}^{(d)}\right)\right|}{\left(\overline{\alpha^2}(d)\right)^{1/2}}\right] \to 1$$

We first require convergence of the second moment and convergence in probability. Lemma 19 in fact gives equivalence of second moments:

$$\frac{1}{\overline{\alpha^2}(d)}\mathbb{E}\left[\left|\mathbf{S}^{(d)}\left(\mathbf{U}^{(d)}\right)\right|^2\right] = \mathbb{E}\left[\left|\mathbf{U}^{(d)}\right|^2\right] = 1$$

Convergence in mean square implies convergence in probability so applying Lemma 18 and subject to (3.73)

$$\left|\mathbf{U}^{(d)}\right| \xrightarrow{m.s.} 1 \Rightarrow \frac{\left|\mathbf{S}^{(d)}\left(\mathbf{U}^{(d)}\right)\right|}{\left(\overline{\alpha^2}(d)\right)^{1/2}} \xrightarrow{p} 1 \quad \textit{and} \quad \frac{\left|\mathbf{S}^{(d)}\left(\mathbf{U}^{(d)}\right)\right|}{\left(\overline{\alpha^2}(d)\right)^{1/2}} \xrightarrow{m.s.} 1 \Rightarrow \left|\mathbf{U}^{(d)}\right| \xrightarrow{p} 1$$

We have proved equivalence of second moments and convergence in probability so the desired result follows by Scheffe's Lemma.

In Lemmas 18 and 21 the eigenvalue condition (3.73) applies to the map which transforms a spherically symmetric random variable ***to*** an elliptically symmetric random variable. It is crucial to the understanding of the application of Lemmas 18 and 21 to limiting forms for the random walk Metropolis that the reader keep the directionality of this map in mind.

3.3.2.2 Extension of limit results to unimodal elliptically symmetric targets

For Corollary 7 to be applicable in the transformed space we require there to exist $k_x^{*(d)}$ and $k_y^{*(d)}$ such that

$$\frac{\left|\mathbf{T}^{(d)}\left(\mathbf{X}^{(d)}\right)\right|}{k_x^{*(d)}} \xrightarrow{p} 1 \quad \text{and} \quad \frac{\left|\mathbf{T}^{(d)}\left(\mathbf{Y}^{(d)}\right)\right|}{k_y^{*(d)}} \xrightarrow{m.s.} 1 \tag{3.75}$$

The working statistician is free to chose a jump proposal such that $\left|\mathbf{Y}^{(d)}\right|/k_y^{(d)} \xrightarrow{m.s.} 1$. Lemma 21 makes explicit precisely when mean square convergence continues through to the transformed space if $\mathbf{Y}^{(d)}$ is spherically symmetric.

We also require convergence in probability of the transformed target $\mathbf{T}^{(d)}\left(\mathbf{X}^{(d)}\right)$. The most simply stated theorem would make no reference to the transformed space and would therefore be working with convergence in probability in the original

space and require from this the equivalent convergence in the transformed space. This would impose an additional constraint on the transformation (see Note 1 below). However the more natural convergence to request would be that of the sequence of transformed targets themselves since these are spherically symmetric and have unit scale parameter (this is analogous to the situation considered by Bedard (2006c)). We first consider spherically symmetric proposals.

Theorem 5 *Let* $\{\mathbf{X}^{(d)}\}$ *be a sequence of elliptically symmetric targets created by linear maps from spherically symmetric sequence* $\{\mathbf{X}_*^{(d)}\}$*, and let* $\{\mathbf{Y}^{(d)}\}$ *be a sequence of spherically symmetric proposals. Let there exist* $k_x^{*(d)}$ *and* $k_y^{(d)}$ *such that*

$$\frac{\mathbf{X}_*^{(d)}}{k_x^{*(d)}} \xrightarrow{p} 1 \quad \textit{and} \quad \frac{\mathbf{Y}^{(d)}}{k_y^{(d)}} \xrightarrow{m.s.} 1$$

Write $\{\mathbf{T}^{(d)}\}$ *for the sequence of linear maps such that* $\mathbf{T}^{(d)}\left(\mathbf{X}^{(d)}\right) = \mathbf{X}_*^{(d)}$ *is spherically symmetric with unit scale parameter and denote by* ν_i *the eigenvalues of* $\mathbf{T}^{(d)}$*. Also define*

$$k_y^{*(d)} = \left(\overline{\nu^2}\right)^{1/2} k_y^{(d)} \tag{3.76}$$

If

$$\frac{\nu_{max}(d)^2}{\sum \nu_i(d)^2} \to 0 \tag{3.77}$$

then for fixed

$$\mu := \frac{1}{2}\frac{d^{1/2}k_y^{*(d)}}{k_x^{*(d)}}\lambda \tag{3.78}$$

the expected acceptance rate and the ESJD satisfy

$$\overline{\alpha}_d(\mu) \rightarrow 2\Phi(-\mu) \tag{3.79}$$

$$\frac{d}{4k_x^{*(d)^2}} S_d^2(\mu) \rightarrow 2\mu^2\Phi(-\mu) \tag{3.80}$$

where $\Phi(x)$ is the cumulative distribution function of a standard Gaussian. Furthermore, the naive ESJD in the original space is

$$S^2_{d,\ naive} \sim \frac{1}{\overline{\nu^2}} S^2_d \tag{3.81}$$

Proof: *Apply Lemma 21 with $\mathbf{U}^{(d)} = \mathbf{Y}^{(d)}/k_y^{(d)}$ and $\mathbf{S}^{(d)} = \mathbf{T}^{(d)}$ to see that the second half of (3.75) holds with the new rescaling factor as in (3.76). We may therefore apply Corollary 7 in the transformed space with μ as defined in (3.78). Since acceptance is the same in the two spaces, this leads directly to (3.79) and (3.80).*

We now seek the naive ESJD in the original space. Here, the proposed squared jumping distance is $\left|\mathbf{Y}^{(d)}\right|^2$, and using (3.76)

$$\left(\overline{\nu^2}\right)^{1/2} \frac{\left|\mathbf{Y}^{(d)}\right|}{k_y^{*(d)}} \xrightarrow{m.s.} 1$$

Equation (3.32) gives

$$S^2_d = 2\lambda^2 \mathbb{E}\left[\left|\mathbf{Y}_*\right|^2 F_{1|d}\left(-\frac{1}{2}\lambda\left|\mathbf{Y}_*\right|\right)\right]$$

Considering the equivalent jumps in the original space

$$S^2_{d,\ naive} = 2\lambda^2 \mathbb{E}\left[\left|\mathbf{Y}\right|^2 F_{1|d}\left(-\frac{1}{2}\lambda\left|\mathbf{Y}_*\right|\right)\right]$$

Rearranging (3.78) and applying Lemma 15 we obtain

$$\begin{aligned} S^2_{d,\ naive}(\mu) &= \frac{8\mu^2\left(k_x^{*(d)}\right)^2}{d\left(k_y^{*(d)}\right)^2}\mathbb{E}\left[\left|\mathbf{Y}^{(d)}\right|^2 F_{1|d}\left(-\mu\left|\mathbf{Y}_*^{(d)}\right|\frac{k_x^{*(d)}}{d^{1/2}k_y^{*(d)}}\right)\right] \\ &= \frac{8\mu^2\left(k_x^{*(d)}\right)^2}{d}\left(\frac{1}{\overline{\overline{\nu^2}}}\right)\mathbb{E}\left[\overline{\nu^2}\left(\frac{\left|\mathbf{Y}^{(d)}\right|}{k_y^{*(d)}}\right)^2 F_{1|d}\left(-\mu\left|\mathbf{Y}_*^{(d)}\right|\frac{k_x^{*(d)}}{d^{1/2}k_y^{*(d)}}\right)\right] \\ &\sim \frac{4\mu^2 {k_x^{*(d)}}^2}{d\ \overline{\nu^2}}\times\overline{\alpha}_d(\mu) \\ &= \frac{1}{\overline{\overline{\nu^2}}}S_d^2 \end{aligned}$$

Notes:

1. If instead of convergence in probability of the (spherical) target in the transformed space, we are given convergence in probability of the elliptical target $\mathbf{X}^{(d)}$ we must additionally ensure that this leads to convergence in probability in the transformed space. We apply Lemma 18 with spherically symmetric $\mathbf{U}^{(d)} = \frac{1}{k_x^{(d)}}\left(\overline{\beta^2}\right)^{1/2}\mathbf{T}^{(d)}\left(\mathbf{X}^{(d)}\right)$ and $\mathbf{S}^{(d)} = \left(\mathbf{T}^{(d)}\right)^{-1}$ to obtain

$$\frac{\left|\mathbf{X}^{(d)}\right|}{k_x^{(d)}} = \frac{1}{k_x^{(d)}}\left(\overline{\beta^2}\right)^{1/2}\frac{\left|\left(\mathbf{T}^{(d)}\right)^{-1}\left(\mathbf{T}^{(d)}\left(\mathbf{X}^{(d)}\right)\right)\right|}{\left(\overline{\beta^2}\right)^{1/2}} \xrightarrow{p} 1$$

$$\Leftrightarrow \frac{1}{k_x^{(d)}}\left(\overline{\beta^2}\right)^{1/2}\left|\mathbf{T}^{(d)}\left(\mathbf{X}^{(d)}\right)\right| \xrightarrow{p} 1$$

provided

$$\frac{\beta_{max}(d)^2}{\sum\beta_i(d)^2}\to 0 \tag{3.82}$$

Here $\beta_i := 1/\nu_i$ are the eigenvalues of $\left(\mathbf{T}^{(d)}\right)^{-1}$ and the scale parameters of $\mathbf{X}^{(d)}$. This also allows us to relate the target rescalings in the two spaces

$$k_x^{(d)} = \left(\overline{\beta^2}\right)^{1/2}k_x^{*(d)} \tag{3.83}$$

2. Naturally (3.80) leads to the same optimal $\hat{\mu}_p$ as for a spherically symmetric target, so the optimal acceptance rate is still approximately 0.234 and the optimal scaling satisfies

$$\hat{\lambda} = 2\hat{\mu}_p \frac{k_x^{*(d)}}{d^{1/2} k_y^{(d)}} \times \frac{1}{\left(\overline{\nu^2}\right)^{1/2}}$$

If we knew nothing of the target shape then we would by default choose a spherically symmetric proposal. However given knowledge of the target shape we might choose a proposal with the same shape and orientation as the target; intuitively this is the optimum choice of proposal shape. Similar ideas for independent components are discussed by Roberts and Rosenthal (2001) and reviewed in Section 3.1.1.

Such a proposal is spherically symmetric in the transformed space. In a similar manner to Note 1 we apply Lemma 21 with spherically symmetric $\mathbf{U}^{(d)} = \frac{1}{k_y^{(d)}} \left(\overline{\beta^2}\right)^{1/2} \mathbf{T}^{(d)}\left(\mathbf{Y}^{(d)}\right)$ and $\mathbf{S}^{(d)} = \left(\mathbf{T}^{(d)}\right)^{-1}$ to obtain

$$\frac{\left|\mathbf{Y}^{(d)}\right|}{k_y^{(d)}} \xrightarrow{m.s.} 1 \Leftrightarrow \frac{1}{k_y^{(d)}} \left(\overline{\beta^2}\right)^{1/2} \left|\mathbf{T}^{(d)}\left(\mathbf{Y}^{(d)}\right)\right| \xrightarrow{m.s.} 1$$

provided (3.82) holds. This also provides a relation between the rescaling factors

$$k_y^{(d)} = \left(\overline{\beta^2}\right)^{1/2} k_y^{*(d)} \tag{3.84}$$

Therefore (3.79) and (3.80) of Theorem 5 apply in the transformed space, but now with the additional condition (3.82).

Also from (3.84) $\left|\mathbf{Y}^{(d)}\right| / \left(\left(\overline{\beta^2}\right)^{1/2} k_y^{*(d)}\right) \xrightarrow{m.s.} 1$ and we again apply Lemma 15 as in the proof of Theorem 5 to find

$$S^2_{d,\text{ naive}} = \overline{\beta^2} S_d^2 \tag{3.85}$$

We now wish to compare the efficiencies of the two proposal shapes subject to (3.77) and (3.82) i.e. a spherical proposal or an elliptical proposal with the same principle axes and eigenvalues (up to a constant of proportionality) as the target. One logical choice of the relative efficiency of the two proposal shapes is the ratio of the optimal ESJD when using a spherical jump proposal to the optimal ESJD when using the perfect elliptical shape.

If this definition is applied to our standard ESJD (3.15) then the relative efficiency is 1. To see this, note that since $k_x^{*(d)}$ is that for the spherical target with unit scale parameter, the optimal ESJD is independent of the proposal's elliptical shape (in the transformed space): i.e. there is always a proposal scale parameter that will produce the optimal ESJD.

However if we take the definition as applying to our naive ESJD in the original space (3.14) then, since the optimum S_d^2 does not depend on the shape of the target, we obtain (in the limit as $d \to \infty$)

$$\text{rel.eff} := \frac{\left(S^2_{d,\ \text{naive}}\right)_{sph}}{\left(S^2_{d,\ \text{naive}}\right)_{ell}} = \frac{1}{\overline{\beta^2\nu^2}} = \frac{\tilde{\beta^2}}{\overline{\beta^2}} = \frac{\tilde{\nu^2}}{\overline{\nu^2}} \tag{3.86}$$

where the tilde symbol denotes the harmonic mean. This is analagous to (3.105) which will be deduced in Section 3.4.2 from a theorem of Roberts and Rosenthal (2001).

We note that the harmonic mean is always less than or equal to the arithmetic mean, so that the optimal spherical proposal is never more efficient than the optimal elliptical proposal, and briefly consider two specific cases:

(i) If $\sigma^2 = Var[\beta_i^2] << E[\beta_i^2] = \mu$ then $\tilde{\beta^2}/\overline{\beta^2} \approx 1 - \sigma^2/\mu^2$.

(ii) If $\beta_i = d$ for odd i, and $\beta_i = 1$ for even i then $\overline{\beta^2} \approx d^2/2$ and $\tilde{\beta}^2 \approx 2$ so $\tilde{\beta}^2/\overline{\beta^2} \approx 4/d^2$.

Note that case (ii) satisfies both the criteria (3.77) and (3.82) on the scale parameters of the target distribution and is similar to Example 3 in Bedard (2006a).

It is only when both transformed target *and* transformed proposal are spherically symmetric that S_d^2/d is the expected square jump distance along any given component; otherwise the ESJD is simply an average over all components. In the latter case some components may be explored better than others.

We might ask if is possible to speed up exploration of the i^{th} component of $\mathbf{X}$ at the expense of the other components by altering the scale parameter λ to be closer to the optimum for that component. A moment's thought tells us this cannot be done while still updating all components in a single block. On spherically symmetric target $\mathbf{X}_*$, multiplying the optimal $\hat{\lambda}$ by a factor c is guaranteed to reduce the overall ESJD and hence (as we have not changed the shape of the proposal) the ESJD along all individual components in the transformed space and in the original space. Of course changing the *shape* of a proposal does allow us to speed up exploration of some components at the expense of others. An extreme case would be choosing a proposal which only ever updated the first component. This would of course never properly explore the target, but it does lead on to the idea partial blocking (see also Sections 1.3.1.2 and 3.1.1) where axes are partitioned into groups with each group updated separately.

3.3.3 Partial blocking: asymptotic results

We now consider the effect of updating components separately using several sub-blocks rather than a single block, and compare the limiting efficiency of such a scheme with that of a single block update.

As in Section 3.2.4 split the complete space into k components: $\Re^d = \mathcal{E}_1 \oplus \cdots \oplus \mathcal{E}_k$. Write $dim\left(\Re^d\right) = d$, $dim\left(\mathcal{E}_i\right) = d_i$ and require that $d_i \to \infty$ as $d \to \infty$ for all i, with $d_i/d \to f_i$, for some f_i. We also require that k remain fixed as $d \to \infty$, and update $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_k)$ to $\mathbf{x}' = (\mathbf{x}'_1, \ldots, \mathbf{x}'_k)$

1. Either via a single block update proposal

$$q\left(\mathbf{x}^*|\mathbf{x}\right) = q_1\left(\mathbf{x}_1^*|\mathbf{x}_1, \ldots, \mathbf{x}_k\right) \ldots q_k\left(\mathbf{x}_k^*|\mathbf{x}_1, \ldots, \mathbf{x}_k\right) \tag{3.87}$$

2. Or via k sub-blocks using k separate proposals with an accept/reject stage after each

$$q_1\left(\mathbf{x}_1^*|\mathbf{x}_1, \mathbf{x}_{-1}\right) \quad \text{to} \quad q_k\left(\mathbf{x}_k^*|\mathbf{x}_k, \mathbf{x}_{-k}\right) \tag{3.88}$$

where $\mathbf{x}_{-k}$ is as defined in (3.23).

3.3.3.1 Partial blocking on spherical targets

Restricting ourselves to the RWM on a unimodal spherically symmetric target we write $\mathbf{y} := \mathbf{x}^* - \mathbf{x}$ and $\mathbf{y}_i = \mathbf{x}_i^* - \mathbf{x}_i$. We will show that the limiting optimal ESJD using a single block and the limiting optimal ESJD's in each of the partial blocks are all equal.

Since $\mathbf{x}_i \perp \mathbf{x}_j \forall i \neq j$ we have $|\mathbf{x}|^2 = \sum_1^k |\mathbf{x}_i|^2$ so the acceptance regions are as follows

1. The single block acceptance region is

$$\{(\mathbf{x}, \mathbf{y}) : \pi(\mathbf{x}_1 + \mathbf{y}_1, \ldots, \mathbf{x}_k + \mathbf{y}_k) > \pi(\mathbf{x}_1, \ldots, \mathbf{x}_k)\}$$

   which (see Section 3.3.1.1) reduces to

$$\left\{(\mathbf{x},\mathbf{y}) : \mathbf{x}\cdot\hat{\mathbf{y}} \leq -\frac{1}{2}|\mathbf{y}|\right\}$$

2. With partial blocking, the acceptance region for block i is

$$\{y_1 : \pi(\mathbf{x}_i + \mathbf{y}_i, \mathbf{x}_{-i}) > \pi(\mathbf{x}_i, \mathbf{x}_{-i})\}$$

   which similarly reduces to

$$\left\{(\mathbf{x}_i, \mathbf{y}_i) : \mathbf{x}_i\cdot\hat{\mathbf{y}}_i \leq -\frac{1}{2}|\mathbf{y}_i|\right\}$$

For the single block update at equilibrium, the chain is reversible and we may apply Exchangeability Lemmas 1 and 2 and all the ensuing theory in order to maximise the ESJD. For the k block updates at equilibrium we may similarly apply Lemmas 4 and 5. Since each acceptance region has the same form within its subspace as the acceptance region for the full update within the full space, all the optimality theory so far derived applies to optimising the ESJD across each partial space as well.

Recall that we may decompose any spherically symmetric random variable as

$$\mathbf{X}^{(d)} = \frac{\mathbf{Z}^{(d)}}{|\mathbf{Z}^{(d)}|} \times R^{(d)}$$

where $\mathbf{Z}^{(d)} \sim N(\mathbf{0}, \mathbf{I}_d)$. Decompose both sides into the sub-spaces $\mathcal{E}_i$ to see that

$$\mathbf{X}_i^{(d_i)} = \frac{\mathbf{Z}_i^{(d_i)}}{|\mathbf{Z}^{(d)}|} \times R^{(d)}$$

Here $\mathbf{X}^{(d)} = \mathbf{X}_1^{(d_1)} \oplus \cdots \oplus \mathbf{X}_k^{(d_k)}$ and $\mathbf{Z}^{(d)} = \mathbf{Z}_1^{(d_1)} \oplus \cdots \oplus \mathbf{Z}_k^{(d_k)}$. If $R^{(d)}/k_x^{(d)} \xrightarrow{p} 1$ then by Lemma 8

$$\frac{\left|\mathbf{X}^{(d)}\right|}{k_x^{(d)}} \xrightarrow{p} 1 \quad \text{and} \quad \frac{d^{1/2}\left|\mathbf{X}_i^{(d_i)}\right|}{d_i^{1/2}k_x^{(d)}} \xrightarrow{p} 1$$

Hence the re-scaling factor for $\mathbf{X}_i^{(d_i)}$ is

$$k_{x_i}^{(d_i)} = \left(\frac{d_i}{d}\right)^{1/2} k_x^{(d)}$$

Consider now the ESJD's (or equivalently the naive ESJD's as the target is spherical). These have the same form for complete and partial blocks, up to a constant of proportionality, and are therefore maximised at the same scaling $\hat{\mu}_p$. The optimal ESJD for the single block update satisfies

$$\frac{d}{4\hat{\mu}_p^2 {k_x^{(d)}}^2} S_d^2(\hat{\mu}_p) \to 2\Phi\left(-\hat{\mu}_p\right)$$

The optimal ESJD for each partial update satisfies

$$\frac{d_i}{4\hat{\mu}_p^2 \left(k_{x_i}^{(d_i)}\right)^2} S_{d_i}^2(\hat{\mu}_p) \to 2\Phi\left(-\hat{\mu}_p\right)$$

But $d_i/\left(k_{x_i}^{(d_i)}\right)^2 = d/{k_x^{(d)}}^2$, so, in the limit as $d \to \infty$ the ratio of optimal ESJD's is

$$\frac{S_{d_i}^2(\hat{\mu}_p)}{S_d^2(\hat{\mu}_p)} = 1$$

Thus in the limit as $d \to \infty$, and provided there is a $k_x^{(d)}$ such that $\left|\mathbf{X}^{(d)}\right|/k_x^{(d)} \xrightarrow{p} 1$, *the optimal ESJD for a partial update is the same as that for a complete update.*

3.3.3.2 Partial blocking on elliptical targets

We now examine partial blocking on an elliptical target in the limit as $d \to \infty$, where each subspace $\mathcal{E}_i$ is the span of some subset of the principal axes. We as-

sume that the original target is explored using spherical proposals, that the spherical transformed target satisfies $\mathbf{X}_*^{(d)}/k_x^{*(d)} \xrightarrow{p} 1$, and that the eigenvalues of the orthogonal linear map transforming elliptical $\mathbf{X}^{(d)}$ to $\mathbf{X}_*^{(d)}$ satisfy (3.77). For an elliptical target explored using a spherical proposal and a single block update, the total naive ESJD is $\tilde{\beta}^2 S_d^2$ (see Theorem 5). Here S_d^2 is the ESJD and is equivalent to the naive ESJD in the transformed space, where the target is spherical; β_i are the scale parameters in the original space and $\tilde{\beta}^2 = 1/\overline{\nu^2}$ denotes the harmonic mean of their squares.

In the previous section we showed that the optimal ESJD for a partial block on a spherical target is equal to the optimal ESJD for a full block. However ESJD's are (by definition) invariant under transformation from a spherical to an elliptical target. Thus if a ratio of true ESJD's is taken as our measure of relative efficiency then on elliptical targets there is clearly no difference between updating via a single block or via partial blocking. However the naive total ESJD summed over all k partial blocks in the original space is

$$S_d^2 \sum_1^k \tilde{\beta}_i^2$$

Hence the relative efficiency with respect to naive ESJD's, of updating using k partial blockings compared to k updates using a single block is

$$\frac{\sum_1^k \tilde{\beta}_i^2}{k\ \tilde{\beta}^2}$$

But

$$\tilde{\beta}^2 = \left(\frac{1}{d} \sum_{i=1}^{k} \sum_{j=1}^{d_i} \frac{1}{\beta_{ij}^2} \right)^{-1} = \left(\frac{1}{d} \sum_{i=1}^{k} \frac{d_i}{\tilde{\beta}_i^2} \right)^{-1}$$

So partial blocking is more efficient if and only if

$$\frac{1}{k}\sum_{1}^{k}\tilde{\beta}_i^2 > \left(\sum_{i=1}^{k}\frac{d_i}{d}\frac{1}{\tilde{\beta}_i^2}\right)^{-1} \tag{3.89}$$

Equation (3.89) may be interpreted as a comparison between an arithmetic mean and a weighted harmonic mean. If the weights are all equal (i.e. the relative number of dimensions of each of the blocks is the same) then the inequality holds as the arithmetic mean is greater than the harmonic mean. Equality clearly arises when (for example) the harmonic mean squares of the scale parameters in each of the partial blocks is the same. However if there is a discrepancy between the blocks' scales and much greater weight is given to blocks with larger $\tilde{\beta}_i^2$ than to those with the smaller $\tilde{\beta}_i^2$ then the inequality does not hold. In other words if our blocking structure consists of large blocks with large scale parameters and much smaller blocks with small scale parameters then this can be less efficient than using a single block update. The limiting results for partial blocking are summarised in the following theorem.

Theorem 6 *Consider two stationary symmetric random walk Metropolis algorithms on an elliptically symmetric unimodal target distribution* $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_k)$ *with* $\mathbf{X}_i \in \mathcal{E}_i$, $\Re^d = \mathcal{E}_1 \oplus \cdots \oplus \mathcal{E}_k$ *and each* $\mathcal{E}_i$ *the span of a subset of the principal axes of the target. Let the first algorithm use a single proposal of the form given in (3.87) and the second use* k *sub-blocks as in (3.88). Provided that for all* i, $d_i := dim(\mathbf{X}_i) \to \infty$ *as* $d := dim(\mathbf{X}) \to \infty$ *with* $d_i/d \to f_i$, *and there is a* $k_x^{(d)}$ *such that* $|\mathbf{X}|/k_x^{(d)} \xrightarrow{p} 1$ *as* $d \to \infty$ *then, in the limit as* $d \to \infty$

(i) The optimal ESJD after any of the k *stages of the partial blocking algorithm is equal to the optimal ESJD for the single-block algorithm.*

(ii) For blocking such that $d_i/d \to 1/k \ \forall \ i$, the naive ESJD after all k stages of the second algorithm is at least as large as the naive ESJD after k repeats of the first. However if the dimensions of each partial block are unequal it is possible for the naive ESJD after all k partial updates to be less than the naive ESJD after k single block updates.

(iii) If the scaling parameters for each block have the same harmonic mean square then the naive ESJD for each partial-block is the same as that for the single block.

3.3.4 Optimal scaling of the random walk Metropolis for specific combinations of finite dimensional target and proposal

This section is concerned with the behaviour of the RWM on real (as opposed to limiting) finite dimensional spherically symmetric unimodal targets.

In Section 3.3.1.3 explicit results were derived for expected acceptance rate and ESJD on spherically symmetric unimodal targets in terms of simple double integrals involving the marginal radial density functions for the target and proposal. In Section 3.3.4.1 we simplify the formulae for specific combinations of target and proposal to a single integral for the general d-dimensional case and to simple analytical expressions for 1-dimensional targets. The former are then used to increase the efficiency of some of the numerical calculations in Section 3.3.4.2, and the latter are of interest purely for the simplicity of form.

In Section 3.3.4.2 the exact formulae are applied to finding the expected acceptance rate and ESJD for a variety of targets and proposals across a range of dimensions. Values are calculated using simple numerical integration routines written in R which produce in a few seconds graphs of optimal scalings and acceptance rates that would otherwise have required an extensive set of simulation studies. Behaviours are found to agree with a simulation study from the literature, and with the asymptotic results of Sections (3.3.1.7) and (3.3.1.8).

Section 3.3.4.3 details a simulation study involving real MCMC runs to illuminate the problem of defining a single true "optimal-scaling" on a target for which different portions vary on radically different scales.

3.3.4.1 Analytical results

In deriving analytical results we restrict attention to target densities of the form $\pi(\mathbf{x}) \propto \exp\left(-\frac{1}{\alpha}|\mathbf{x}|^{\alpha}\right)$ and proposal densities $q(\mathbf{y}) \propto |\mathbf{y}|^{\beta}\exp\left(-\frac{1}{\alpha}|\mathbf{y}|^{\alpha}\right)$ This includes as special cases

(i) Gaussian target with Gaussian proposal

(ii) (Spherical) exponential target with (spherical) exponential proposal

First note that the full expressions for the d-dimensional marginal radial densities are

$$\begin{aligned}\overline{f}_d(x) &= \frac{1}{\alpha^{\frac{d}{\alpha}-1}\Gamma\left(\frac{d}{\alpha}\right)}x^{d-1}\exp\left(-\frac{1}{\alpha}x^{\alpha}\right)\\ \overline{r}_d(y) &= \frac{1}{\alpha^{\frac{\beta+d}{\alpha}-1}\Gamma\left(\frac{\beta+d}{\alpha}\right)}y^{\beta+d-1}\exp\left(-\frac{1}{\alpha}y^{\alpha}\right)\end{aligned}$$

where the normalising constants are obtained through the changes of variable $z_1 = \frac{1}{\alpha}x^\alpha$ and $z_2 = \frac{1}{\alpha}y^\alpha$.

It is convenient to define the joint normalising constant

$$A_d(\alpha,\beta) := \left(\alpha^{\frac{\beta+2d}{\alpha}-2}\,\Gamma\left(\frac{d}{\alpha}\right)\Gamma\left(\frac{\beta+d}{\alpha}\right)\right)^{-1}$$

Substitution into (3.44) and (3.45) produces

$$\begin{aligned}
\overline{\alpha}_d(\lambda) &= A_d(\alpha,\beta)\int_0^\infty dy\ \int_{\frac{1}{2}\lambda y}^\infty dx\ x^{d-1}y^{\beta+d-1}\ \exp\left(-\frac{1}{\alpha}\left(x^\alpha+y^\alpha\right)\right)K_d\left(\frac{\lambda y}{2x}\right)\\
S_d^2(\lambda) &= A_d(\alpha,\beta)\lambda^2\int_0^\infty dy\ \int_{\frac{1}{2}\lambda y}^\infty dx\ x^{d-1}y^{\beta+d+1}\ \exp\left(-\frac{1}{\alpha}\left(x^\alpha+y^\alpha\right)\right)K_d\left(\frac{\lambda y}{2x}\right)
\end{aligned}$$

We then apply the following change of variable

$$u = \frac{y}{x} \quad \text{and} \quad v = \frac{1}{\alpha}\left(x^\alpha+y^\alpha\right)$$

from which

$$x = \left(\frac{\alpha v}{1+u^\alpha}\right)^{1/\alpha}, \qquad y = u\left(\frac{\alpha v}{1+u^\alpha}\right)^{1/\alpha} \quad \text{and} \quad \frac{\partial(x,y)}{\partial(u,v)} = \frac{1}{\alpha v}\left(\frac{\alpha v}{1+u^\alpha}\right)^{2/\alpha}$$

Hence

$$\overline{\alpha}_d(\lambda) = A_d(\alpha,\beta)\ \alpha^{\frac{\beta+2d}{\alpha}-1}\ \times \int_0^\infty dv\ \int_0^{2/\lambda} du\ v^{\frac{\beta+2d}{\alpha}-1}\frac{u^{\beta+d-1}}{(1+u^\alpha)^{\frac{\beta+2d}{\alpha}}}\ \exp\left(-v\right)K_d\left(\frac{\lambda u}{2}\right)$$

and

$$S_d^2(\lambda) = A_d(\alpha,\beta)\ \alpha^{\frac{\beta+2d+2}{\alpha}-1}\ \lambda^2\ \times \int_0^\infty dv\ \int_0^{2/\lambda} du\ v^{\frac{\beta+2d+2}{\alpha}-1}\frac{u^{\beta+d+1}}{(1+u^\alpha)^{\frac{\beta+2d+2}{\alpha}}}\ \exp\left(-v\right)K_d\left(\frac{\lambda u}{2}\right)$$

Integrating out v we obtain

$$\overline{\alpha}_d(\lambda) = \frac{\alpha}{B\left(\frac{d}{\alpha}, \frac{\beta+d}{\alpha}\right)} \int_0^{2/\lambda} du \; \frac{u^{\beta+d-1}}{(1+u^\alpha)^{\frac{\beta+2d}{\alpha}}} \; K_d\left(\frac{\lambda u}{2}\right) \tag{3.90}$$

$$S_d^2(\lambda) = \frac{\alpha^{1+2/\alpha}}{B\left(\frac{d}{\alpha}, \frac{\beta+d}{\alpha}\right)} \frac{\Gamma\left(\frac{\beta+2d+2}{\alpha}\right)}{\Gamma\left(\frac{\beta+2d}{\alpha}\right)} \lambda^2 \int_0^{2/\lambda} du \; \frac{u^{\beta+d+1}}{(1+u^\alpha)^{\frac{\beta+2d+2}{\alpha}}} K_d\left(\frac{\lambda u}{2}\right) \tag{3.91}$$

We now consider two special cases, obtaining exact results when $d = 1$.

Gaussian target and Gaussian proposal: substituting $\alpha = 2$ and $\beta = 0$ gives

$$\overline{\alpha}_d(\lambda) = \frac{2}{B\left(\frac{d}{2}, \frac{d}{2}\right)} \int_0^{2/\lambda} du \; \frac{u^{d-1}}{(1+u^2)^d} K_d\left(\frac{\lambda u}{2}\right) \tag{3.92}$$

$$S_d^2(\lambda) = \frac{4\lambda^2 d}{B\left(\frac{d}{2}, \frac{d}{2}\right)} \int_0^{2/\lambda} du \; \frac{u^{d+1}}{(1+u^2)^{d+1}} K_d\left(\frac{\lambda u}{2}\right) \tag{3.93}$$

For $d = 1$, since $K_1(v) = 1$ for $0 < v < 1$, this reduces to

$$\overline{\alpha}_1(\lambda) = \frac{2}{\pi} \int_0^{2/\lambda} du \; \frac{1}{1+u^2}$$

$$S_1^2(\lambda) = \frac{4\lambda^2}{\pi} \int_0^{2/\lambda} du \; \frac{u^2}{(1+u^2)^2}$$

But substituting $u = \tan\theta$ and defining $\gamma := 2/\lambda$

$$\int_0^{2/\lambda} du \frac{u^2}{(1+u^2)^2} = \int_0^{\tan^{-1}\gamma} d\theta \; \sin^2\theta = \frac{1}{2}\left(\tan^{-1}\gamma - \frac{\gamma}{1+\gamma^2}\right)$$

Therefore

$$\overline{\alpha}_1(\gamma) = \frac{2}{\pi}\tan^{-1}\gamma \tag{3.94}$$

$$S_1^2(\gamma) = \frac{8}{\pi\gamma^2}\left(\tan^{-1}\gamma - \frac{\gamma}{1+\gamma^2}\right) \tag{3.95}$$

Maximising (3.95) numerically gives an optimal-scaling of $\hat{\lambda} \approx 2.426$ (or $\hat{\gamma} \approx 0.8243$) which corresponds to an optimal acceptance rate of 0.4389 and an ESJD of 0.7442.

Exponential target and exponential proposal: substituting $\alpha = 1$ and $\beta = 0$ gives

$$\overline{\alpha}_d(\lambda) = \frac{1}{B(d,d)} \int_0^{2/\lambda} du \, \frac{u^{d-1}}{(1+u)^{2d}} K_d\left(\frac{\lambda u}{2}\right) \tag{3.96}$$

$$S_d^2(\lambda) = \frac{2d(2d+1)\lambda^2}{B(d,d)} \int_0^{2/\lambda} du \, \frac{u^{d+1}}{(1+u)^{2d+2}} K_d\left(\frac{\lambda u}{2}\right) \tag{3.97}$$

When $d = 1$, with $\gamma := 2/\lambda$ again, these reduce to

$$\overline{\alpha}_1(\lambda) = \int_0^{\gamma} du \, \frac{1}{(1+u)^2} = \frac{\gamma}{1+\gamma}$$

$$S_1^2(\lambda) = \frac{24}{\gamma^2} \int_0^{\gamma} du \, \frac{u^2}{(1+u)^4} = \frac{8\gamma}{(1+\gamma)^3}$$

S_1^2 and $\overline{\alpha}_1$ are thus related by the simple analytical expression

$$S_1^2 = 8\overline{\alpha}_1(1-\overline{\alpha}_1)^2 \tag{3.98}$$

Thus S_1^2 attains a maximum of $32/27$ at $\overline{\alpha}_1 = 1/3$; at this maximum $\gamma = 1/2$ and $\lambda = 4$.

Note that in both these special cases $S_1^2(\lambda) \to 0$ as $\lambda \to \infty$ and as $\lambda \to 0$, it is positive on $(0, \infty)$ and attains a (single) maximum, all of which is in agreement with Lemma 6 and Corollary 4. Further $\overline{\alpha}_d(\lambda)$ is a monotonically decreasing function of λ taking values throughout $(0, 1]$ as stated in Corollary 3.

3.3.4.2 Computational results

In this section we compare results for Gaussian and exponential targets using either Gaussian or exponential jump proposals. Initially we consider the effect of varying the scale parameter at fixed dimension, before proceeding to explore variation of optimal-scaling and acceptance rate with dimension.

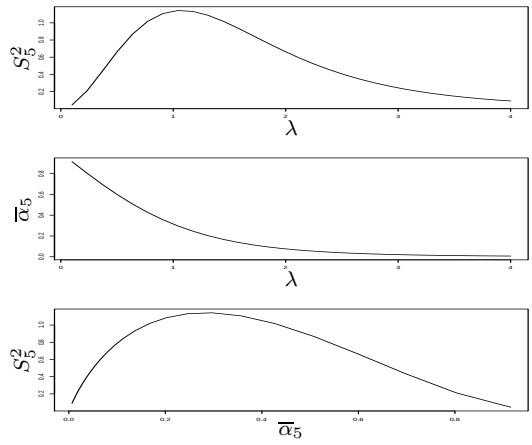

Figure 3.5: Plots for a Gaussian target with a Gaussian jump proposal at dimension $d = 5$: (i) ESJD against scaling, (ii) acceptance rate against scaling, and (iii) ESJD against acceptance rate.

Figure 3.5 shows the effect of changing the scale parameter when using a Gaussian jump proposal distribution to explore a Gaussian target at dimension $d = 5$. Increasing the scale parameter from 0 to ∞ decreases the acceptance rate from 1 to 0, as deduced in Corollary 3. Further, following Lemma, 6 the ESJD does indeed approach zero as the scale parameter approaches either zero or infinity, and as noted in Corollary 4 it achieves a global maximum somewhere between these extremes. The third graph shows ESJD plotted against acceptance rate. Since acceptance rate is a monotonic function of the scale parameter, this graph also shows a single maximum.

Figure 3.6 repeats the plot of ESJD against acceptance rate for the other three combinations of target and proposal. All three show the same general features as

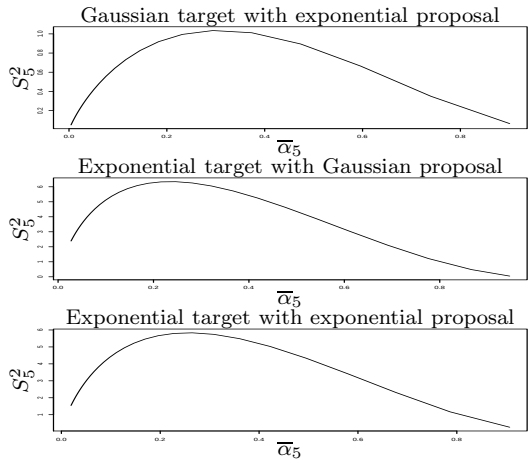

Figure 3.6: Plots of S_5^2 vs. $\overline{\alpha}_5$ for: (i) a Gaussian target with an exponential proposal, (ii) an exponential target with a Gaussian proposal, and (iii) an exponential target with an exponential proposal.

the plot for a Gaussian target with Gaussian proposal. The optimal ESJD's for both of the plots with a Gaussian target are very similar (1.145 and 1.035 respectively for Gaussian and exponential jumps) as are those for the exponential target (6.345 and 5.880 respectively), indicating that the choice of the type of (spherically symmetric) jump proposal makes little difference to the optimal efficiency.

For each combination of target and proposal simple numerical routines are employed to find the scaling $\hat{\lambda}$ that produces the largest ESJD. Substitution into (3.44) gives the corresponding optimal acceptance rate $\hat{\alpha}$. Figure 3.7 shows plots of optimal acceptance rate against dimension for the four combinations of Gaussian or exponential target and Gaussian or exponential proposal. Note that the first of these is entirely consistent with Figure 4 in Roberts and Rosenthal (2001),

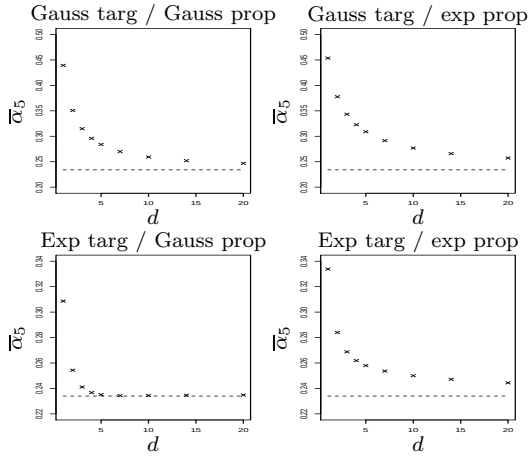

Figure 3.7: Plots of the optimal acceptance rate $\hat{\alpha}$ against dimension for the four combinations of a Gaussian or exponential target and either a Gaussian or exponential proposal. The asymptotic optimum acceptance rate of 0.234 is shown as a dotted line.

which shows optimal acceptance rates obtained through repeated runs of the RWM algorithm.

The rescaled modulus of sequences of either spherically symmetric exponential or Gaussian random variables (increasing in dimension) can be made to converge in probability (and mean square) to 1. The asymptotic theory of Section 3.3.1.8 indicates that in such cases the optimal acceptance rate should converge to 0.234 as $d \to \infty$ and this appears to be true in all four cases examined. We note also that in all four cases the limit is approached from above; it would be interesting to investigate criteria for this. The same theory indicates that asymptotically the optimal scale parameter should behave as $\hat{\lambda} \sim 2\hat{\mu}_p k_x^{(d)}/(d^{1/2} k_y^{(d)})$ with $\hat{\mu}_p \approx 1.19$. From Lemma 8 a standard Gaussian distribution has $k_d = d^{1/2}$; a standard ex-

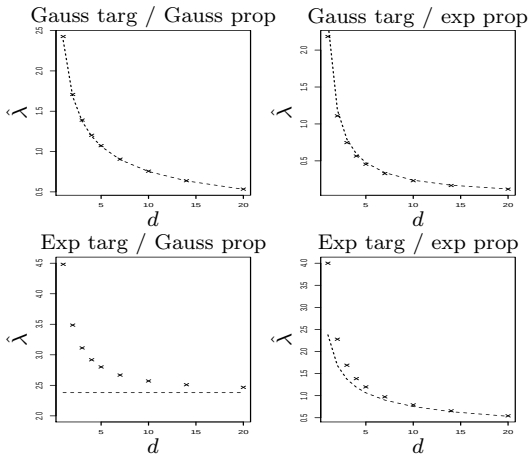

Figure 3.8: Plots of the optimal scale parameter $\hat{\lambda}$ against dimension for the four combinations of a Gaussian or exponential target and either a Gaussian or exponential proposal. Optimal values from the asymptotic theory appear as a dotted line.

ponential distribution has $k_d = d$ since the marginal radius is a Gamma random variate. The corresponding asymptotic behaviours for the different combinations are detailed in Table 3.1. Plots of $\hat{\lambda}$ against d are shown in Figure 3.8 with the expected asymptotic behaviour marked by a dotted line in each graph. All behave asymptotically as expected; for a Gaussian target very close agreement is attained even in one-dimension.

We now consider targets for which the rescaled modulus converges to some distribution other than a point mass at 1. Again using simple numerical integration of equations (3.44) and (3.45) we investigate the following three combinations of target and proposal

Target	Proposal	$k_x^{(d)}$	$k_y^{(d)}$	$\hat{\lambda}_{asymp}$
Gaussian	Gaussian	$d^{1/2}$	$d^{1/2}$	$2.38/d^{1/2}$
Gaussian	Exponential	$d^{1/2}$	d	$2.38/d$
Exponential	Gaussian	d	$d^{1/2}$	2.38
Exponential	Exponential	d	d	$2.38/d^{1/2}$

Table 3.1: Asymptotic optimal scaling behaviour for specific combinations of target and proposal.

1. A Gaussian proposal and a target with density

$$\pi_d(\mathbf{x}) \propto \frac{1}{x^{d-1}} e^{-\frac{1}{2}x^2}$$

2. An exponential proposal and a target with density

$$\pi_d(\mathbf{x}) \propto \frac{1}{x^{d-1}} e^{-x}$$

3. A Gaussian proposal and a target with density

$$\pi_d(\mathbf{x}) \propto \mathbf{1}_{\left\{x \leq e^{-(d-1)}\right\}} + e^{-\frac{1}{2}\left(\log\left(x/e^{-(d-1)}\right)\right)^2} \mathbf{1}_{\left\{x > e^{-(d-1)}\right\}}$$

All three targets are chosen so that a rescaling factor of $k_x^{(d)} = 1$ places all the radial mass in $(0, \infty)$ as $d \to \infty$. Clearly the marginal radial distributions of the first two targets are the unit Gaussian and unit exponential respectively, independent of dimension. The two segments of the third target simply ensure that it is unimodal; it has similar limiting properties to (3.47).

Figure 3.9 shows the variation of acceptance rate and optimal scaling with dimension for combinations (1) and (2). Scaling parameter plots are of $\log \hat{\lambda}$ against

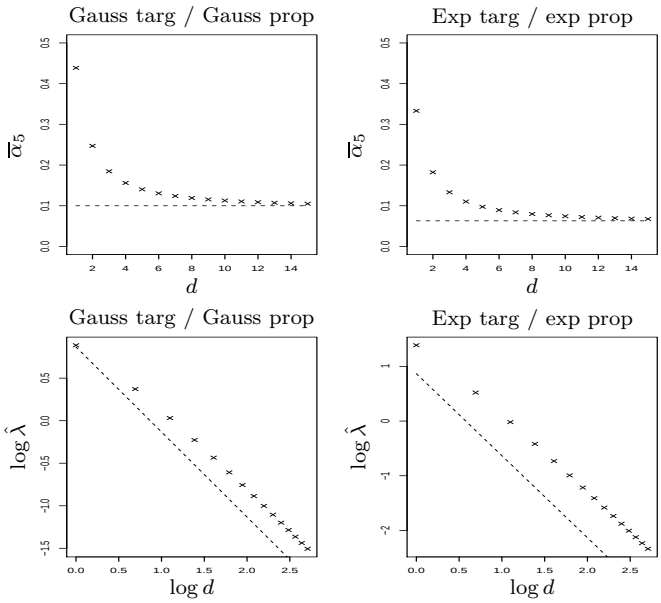

Figure 3.9: Plots for target and proposal combinations (1) and (2); acceptance rate is plotted against dimension with asymptotes of approximately 0.10 and 0.06 shown dotted; $\log \hat{\lambda}$ is plotted against $\log d$ with similar graphs for the asymptotically expected behaviour if the rescaled target modulus had converged in probability to 1.

$\log d$ so as to more easily compare with the asymptotically expected behaviour. As argued in Section 3.3.1.7 acceptance rates for each combination do approach asymptotically optimal values (approximately 0.10 and 0.063 respectively) both of which are less than 0.234 as was proved to be the case in Theorem 4. With $k_x^{(d)} = 1$ Corollary 8 implies that asymptotically $\hat{\lambda} \sim 2\hat{\mu}/d$ for combination (1), and $\hat{\lambda} \sim 2\hat{\mu}/d^{3/2}$ for combination (2). The dotted lines correspond to these formulae with $\hat{\mu} = \hat{\mu}_p \approx 1.19$. The optimal scale parameters clearly behave as expected, in both cases with $\hat{\mu} > \hat{\mu}_p$.

For combination (3) the optimal acceptance rate is expected to tend to zero as $d \to \infty$ (see Section 3.3.1.7). For $d = 1$ and $d = 2$ the optimal acceptance rates are approximately 0.111 and 0.010 respectively. Figure 3.10 shows plots of ESJD against scale parameter, acceptance rate against scale parameter and ESJD against acceptance rate for this combination at $d = 3$. The graphs are heuristically the same is those for the Gaussian target and proposal at $d = 5$ (Figure 3.5). However the optimal acceptance rate is approximately 5.7×10^{-4} and it certainly appears therefore to be approaching a limiting asymptotically optimal value of 0. This neatly brings together Lemma 6 and Corollary 4 for finite dimensional targets and Lemma 16 for the infinite dimensional limit. For each actual target in finite dimension d there is a finite optimal scaling and non-zero optimal acceptance rate. However as $d \to \infty$ the (rescaled) optimal scaling ($\hat{\mu}$) tends to infinity and the optimal acceptance rate tends to zero.

3.3.4.3 Simulation study on a target with a mixture of scales

In Section 3.3.1.8 we examined the limiting behaviour of the random walk Metropolis algorithm in terms of the limiting rescaled radial distribution. We considered convergence in distribution to 1 and convergence in distribution to some positive random variable with no mass at the origin or infinity. In Section 3.3.1.7 we considered limiting radii with masses at either or both of these extremes and argued that the optimality theory developed in this chapter could not apply. In this section we detail simulation studies on targets with a mixture of two very different scales of variation and discuss the implications for optimal scaling.

If each finite dimensional radial distribution contains a mixture of scales then intu-

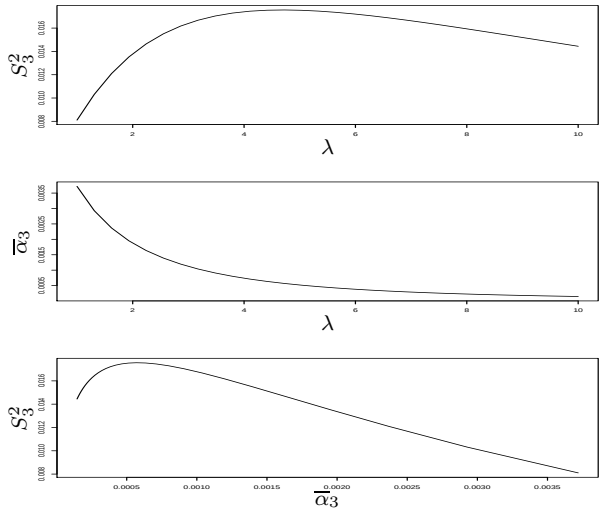

Figure 3.10: Plots for target and proposal combination (3) at $d = 3$; (i) ESJD vs scaling; (ii) expected acceptance rate vs scaling; and (ii) ESJD vs expected acceptance rate.

itively it would seem that a simple RWM with a single scale parameter could only efficiently explore one of these scales, and would be very inefficient on the other(s). If in the limit, the ratio of the mixed scales were to become infinite then rescaling according to any one of them would spread the radial mass on that particular scale over $(0, \infty)$ but add a point mass at zero or infinity or both.

To illustrate this we explored the target

$$\mathbf{X}^{(d)} \sim \left\{ \begin{array}{ll} N(\mathbf{0}, \mathbf{I}_d) & w.p.\ \ 0.5 \\ N(\mathbf{0}, 10^4 d\ \mathbf{I}_d) & w.p.\ \ 0.5 \end{array} \right\} \tag{3.99}$$

for $d = 1$ and $d = 10$, using jump proposal $\mathbf{Y}^{(d)} \sim N(0, \lambda^2 \mathbf{I}_d)$. All simulations were run for 100000 iterations and were started at $\mathbf{0}$, in the main mass at the smaller scale, and then repeated, starting at $100 d^{1/2} \times \mathbf{1}$, in the main marginal radial mass

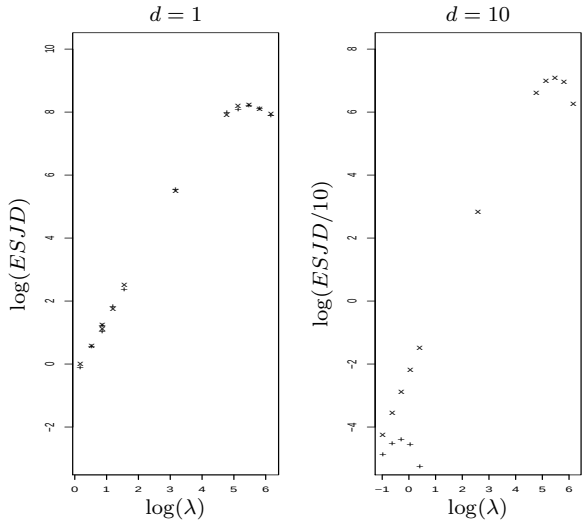

Figure 3.11: Plots (for $d = 1$ and $d = 10$) of log (ESJD/d) against log (scale parameter) for exploration of the Gaussian mixture target in (3.99). Runs started at the origin are plotted as '+', and runs started at $100d^{1/2}\ \mathbf{1}$ are plotted with 'x'.

at the larger scale. The scale parameters tested were clustered around the expected optima for each of the two distributions in the mixture: 2.38 and 238 at $d = 1$, and 0.753 and 238 at $d = 10$.

Figure 3.11 shows the variation of ESJD with scale parameter for $d = 1$ and $d = 10$, starting at the origin, or in the main radial mass of the outer mixture. In one-dimension the starting point appears immaterial and the mean square jump distance is maximised at $\lambda \approx 238$. By contrast, in ten dimensions, starting at the origin leads to a (small) maximum mean square jump distance at $\lambda \approx 0.753$ whereas starting in the outer shell gives an optimal scaling around $\lambda \approx 238$.

At $d = 10$ there is a complete lack of mixing between the two components of the target. Kernel density plots (not shown) for runs started from the origin with small scale parameters show only the Gaussian on the smaller of the two scales. Density decreases so quickly away from the origin that all 10^5 proposed jumps from the origin on any of the larger scales fail. Plots for runs started away from the origin show only the Gaussian with the larger of the two scales. The mass at the origin is concentrated over such a small volume V_s of $\Re^{10}$ compared to the larger scale, that when starting at a point on the larger scale, V_s is never found. This is because the chance of proposing a jump to V_s when using the larger scaling is almost zero, and when using the smaller scaling, the ramping up in density near to V_s is never seen so exploration never approaches V_s and again concentrates on the distribution with the larger scaling, albeit very inefficiently.

The same processes are at work when $d = 1$ but are not extreme enough to prevent mixing between the two components. Kernel density plots for explorations using scale parameters around 2.38 show a good Gaussian shape to the smaller component of the target and a very lopsided density curve for the larger component. By contrast, with $\lambda \approx 238$, the kernel density estimate on the scale of the larger component is much closer to the true shape than is that for the smaller component.

If the algorithms were allowed to run forever then clearly all mixture components would be explored thoroughly no matter what the (non-zero) scaling. However the whole search for an optimal scaling is motivated by a wish to achieve a good approximation to a large sample from the target in as few iterations as possible. We must therefore accept that if in practice the target varies on several very dif-

ferent scales, then there will be different optimal proposal scalings for exploring the different scales of the target. Further, that choosing a scaling which maximises the ESJD will allow efficient exploration of the component of the target with the largest scale parameter of all the components explored.

When $d = 1$ the theoretical acceptance rate at the optimal scaling is 0.22 (observed optimal acceptance rates were between 0.22 and 0.23), one half of the optimum for a single one-dimensional Gaussian. One half of the time the algorithm will be within the inner component, where proposals on this optimal scale are almost always into the outer component and are therefore almost always rejected. Likewise when $d = 10$ the theoretical optimal acceptance rate is $0.26/2 = 0.13$ but the absolute lack of mixing between the two components leads to only one of them ever being seen and so the observed optimal acceptance rates were around 0.26. Once more, if we had continued the algorithm for infintely many more iterations then eventually mixing would have occurred (many times) and the observed acceptance rate would have reduced to 0.13.

We summarise the three main points arising from the fact that simulations involve only a finite number of iterations.

- If the target varies on several very different scales, then there will be different optimal proposal scalings for exploring the different scales of the target.
- Choosing a scaling which maximises the ESJD will allow efficient exploration of the corresponding scale component of the target,
- Observed optimal acceptance rates may differ from their theoretical values as they only take account of scale-components of the target that are actually

explored.

3.4 Conclusion

In this chapter several strands of research have been developed and then linked together to produce new limiting results. We start with (Section 3.4.1) a relatively short guide to two of our key limiting results: Theorem 3 and Corollary 7. The motivation for this is two-fold: firstly it provides a summary of the contribution made by each of the strands and secondly it provides a useful geometric intuition into the processes at play in high dimension.

In Section 3.4.2 we draw comparisons between new results in this Chapter and the existing literature reviewed in Section 3.1.1. We then (Section 3.4.3) list some ideas for further work, and (Section 3.4.4) summarise and discuss this Chapter as a whole.

3.4.1 A selective tour of key results

This section provides some simple intuition behind two of our key results on the limiting behaviour of the RWM. The arguments here are not intended to be rigorous but to give the reader a geometrical feel for the limiting behaviour.

We have assumed a unimodal target density with either spherical or elliptical symmetry. The main assumptions in Theorem 3 are that the sequence of spherically symmetric targets $\mathbf{X}^{(d)}$ and proposals (with unit scale parameter) $\mathbf{Y}^{(d)}$ satisfy

$$\frac{\left|\mathbf{X}^{(d)}\right|}{k_x^{(d)}} \xrightarrow{D} R \quad \text{and} \quad \frac{\left|\mathbf{Y}^{(d)}\right|}{k_y^{(d)}} \xrightarrow{m.s.} 1 \tag{3.100}$$

for some R with no point mass at the origin, and some $k_x^{(d)}$ and $k_y^{(d)}$. We will provide the main intuition for the special case (3.101), which corresponds to Corollary 7, and then offer a simple generalisation.

$$\frac{\left|\mathbf{X}^{(d)}\right|}{k_x^{(d)}} \xrightarrow{p} 1 \quad \text{and} \quad \frac{\left|\mathbf{Y}^{(d)}\right|}{k_y^{(d)}} \xrightarrow{m.s.} 1 \tag{3.101}$$

Both assumptions (3.101) are satisfied by many common sequences of distributions, and the working statistician is in any case free to choose his or her proposals to satisfy the second. It has been shown (Section 3.3.1.8) that subject to these conditions the optimal scaling is approximately

$$\hat{\lambda} \sim 2\hat{\mu}\frac{k_x^{(d)}}{d^{1/2}k_y^{(d)}}$$

with $\hat{\mu} = \hat{\mu}_p \approx 1.19$, and that the acceptance rate at this scaling is 0.234. With the more general conditions (3.100) the form of the optimal scaling is unchanged (but with $\hat{\mu}$ potentially different from $\hat{\mu}_p$) and the optimal acceptance rate less than 0.234. In Section 3.3.4 several specific combinations of target and proposal were examined (over a variety of dimensions) and found to be completely consistent with the theory.

We have shown by a simple and intuitive argument (see Theorem 1 and Corollary 5) that the 1-dimensional marginal $X^{(1|d)}$ of any spherically symmetric random variable which satisfies the weaker of the assumptions (3.101) has

$$\frac{d^{1/2}}{k_x^{(d)}}X^{(1|d)} \xrightarrow{D} N(0,1)$$

We have also proved (see Section 3.2.2) two Exchangeability Lemmas. As a direct consequence of the chain's reversibility at equilibrium and of an "exchangeability" between two main regions of integration, for (almost) any Metropolis-Hastings algorithm the expected acceptance rate is twice the probability of proposing a jump

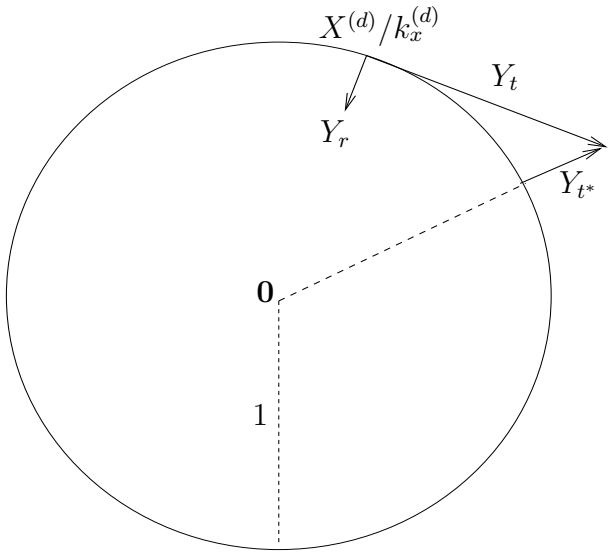

Figure 3.12: Rescaled current point $X^{(d)}/k_x^{(d)}$ on the unit hypershpere, together with tangential (Y_t) and radial (Y_r) components of the proposed move, and the radial motion due to the tangential movement (Y_{t^*}).

that has acceptance probability 1. Further, the naive ESJD is twice the integral of the squared jumping distance over the region where the acceptance probability is 1.

With assumptions (3.101) the rescaled radial mass is (in some sense) converging to a point. In other words a chain at stationarity will spend nearly all its time at a distance approximately $k_x^{(d)}$ from the origin and make jumps of magnitude approximately $k_y^{(d)}$. Figure 3.12 shows a point in the rescaled target distribution $\mathbf{X}^{(d)}/k_x^{(d)}$ for which all the mass is on the surface of a hypersphere of radius 1.

With proposal scale parameter λ, we split the next proposed jump $\lambda\mathbf{Y}^{(d)}$ into radial and tangential components with magnitudes Y_r and Y_t; we keep the superscript

(d) implicit for simplicity of notation. The next proposed point will be on the surface of a hypersphere centred at the current point and with radius approximately $\lambda k_y^{(d)}/k_x^{(d)}$ in the rescaled figure.

The distribution of Y_r is simply the marginal 1-dimensional distribution of $\lambda \mathbf{Y}^{(d)}/k_x^{(d)}$ along the radius and so approximately satisfies

$$\frac{1}{\sigma} Y_r \;\sim\; N(0,1) \quad \text{where} \quad \sigma := \frac{\lambda k_y^{(d)}}{d^{1/2} k_x^{(d)}}$$

All but one of the d components will be tangential so that, provided $\left(k_x^{(d)} Y_t\right) / \left(\lambda k_y^{(d)}\right)$ becomes deterministic in the same way as $\left|\mathbf{Y}^{(d)}\right| / k_y^{(d)}$ we have

$$\frac{\left(k_x^{(d)}\right)^2}{\left(k_y^{(d)}\right)^2} \frac{Y_t^2}{\lambda^2} \xrightarrow{m.s.} \frac{d-1}{d} \to 1$$

In practice it is found that near the optimal scaling, the total tangential movement $Y_{t^*} \to 0$ as $d \to \infty$ (or note that $\lambda k_y^{(d)}/k_x^{(d)} \sim 2\mu/d^{1/2}$). This tangential movement effects a radial movement Y_{t^*} as shown in the figure, where (by Taylor expansion)

$$Y_{t^*} = \left(1 + {Y_t}^2\right)^{1/2} - 1 \approx \frac{1}{2} {Y_t}^2 \approx \frac{1}{2} \lambda^2 \left(\frac{k_y^{(d)}}{k_x^{(d)}}\right)^2 =: \eta$$

Thus the effective proposed radial movement Y_e approximately satisfies

$$\frac{1}{\sigma} \left(Y_e - \eta\right) \;\sim\; N\left(0,1\right)$$

The target is unimodal and hence the acceptance probability is 1 exactly when $Y_e \leq 0$; so from the first Exchangeability Lemma the expected square jumping

distance is simply

$$S^2(\lambda) = 2\int_{-\infty}^{0} dy_e \; \frac{1}{\sigma}\phi\left(\frac{y_e-\eta}{\sigma}\right)\left(\frac{\lambda k_y^{(d)}}{k_x^{(d)}}\right)^2$$
$$= \frac{8}{d}\left(\frac{\eta}{\sigma}\right)^2 \int_{-\infty}^{0} dv \; \phi\left(v-\frac{\eta}{\sigma}\right) = \frac{8}{d}\left(\frac{\eta}{\sigma}\right)^2 \Phi\left(-\frac{\eta}{\sigma}\right)$$

This depends only on $\mu := \eta/\sigma = \frac{1}{2}\,\lambda\, d^{1/2}\, k_y^{(d)}/k_x^{(d)}$. and is maximised at $\hat{\mu}_p \approx 1.19$. Thus the scaling which maximises the ESJD is

$$\hat{\lambda} = 2\hat{\mu}_p \frac{k_x^{(d)}}{d^{1/2} k_y^{(d)}}$$

Finally the first Exchangeability Lemma gives an expected acceptance rate of

$$\overline{\alpha}_d(\lambda) = 2\int_{-\infty}^{0} dy_e \; \frac{1}{\sigma}\phi\left(\frac{y_e-\eta}{\sigma}\right) = 2\int_{-\infty}^{0} dv \; \phi\left(v-\frac{\eta}{\sigma}\right) = 2\Phi(-\mu)$$

So that the optimal scaling corresponds to one particular expected acceptance rate $2\Phi(-\hat{\mu}_p) \approx 0.234$ irrespective of the target and proposal.

Let us now return to the more general assumptions (3.100). Repeating the above arguments on a (rescaled) circle of radius R gives Y_r as before, but

$$Y_{t^*} = \left(R^2 + {Y_t}^2\right)^{1/2} - R \approx \frac{1}{2R}{Y_t}^2 \approx \frac{1}{2R}\lambda^2\left(\frac{k_y^{(d)}}{k_x^{(d)}}\right)^2 = \frac{\eta}{R}$$

so that

$$S_d^2(\lambda) = \frac{8}{d}\mu^2 \mathbb{E}\left[\Phi\left(-\frac{\mu}{R}\right)\right] \tag{3.102}$$
$$\overline{\alpha}_d(\lambda) = 2\mathbb{E}\left[\Phi\left(-\frac{\mu}{R}\right)\right] \tag{3.103}$$

This is Theorem 3 which may therefore be seen to arise from the predictability (in the limit) of the tangential motion and a need to (on average) balance the outward radial motion effected by this finite tangential motion against the single component

of the stochastic radial motion.

The above suggests a possible improvement to the RWM algorithm in high dimension, wherein the proposed jump is centred at $\mathbf{x} - \eta\hat{\mathbf{x}}$ rather than at $\mathbf{x}$, and compensation for the centrepetal motion is therefore included in the proposal. For a Gaussian proposal this turns out to be equivalent (in high dimension only) to the MALA algorithm. The idea of a correction to pull motion back towards the centre of the target is also reminiscent of the polar slice sampler suggested by Roberts and Rosenthal (2002).

3.4.2 Comparison with existing literature

We now compare new results derived in this chapter with the existing literature reviewed in Section 3.1.1. Specifically we examine

- The similarity between our asymptotic form for the ESJD (3.63) and the form for the speed of the limiting diffusion derived in Roberts et al. (1997).
- The link between our formulae for asymptotic relative efficiencies for elliptical targets explored by spherical and then elliptical proposals (3.86) and the speed of the limiting diffusion for targets with independent components, identical up to a scaling, as described by Roberts and Rosenthal (2001).
- The similarities and differences between our results for partial blocking algorithms and those of Neal and Roberts (2006).
- The relationship between the condition (3.77) for the 0.234 limiting optimal acceptance rate to apply to elliptical targets under a spherical proposal, and

the condition of Bedard (2006c), for the 0.234 acceptance rate to apply to targets with independent components up to a scaling, explored with a spherical Gaussian proposal.

We start with a comparison between the basic formulae for limiting efficiency. The asymptotic form (3.63) for the ESJD when there is convergence in probability to 1 of the modulus of the rescaled target, as derived in this chapter, is remarkably similar to the forms for the speeds of the limiting diffusions that appear in the current literature (e.g. (3.4)). The main differences appear to be the presence of the rescalings $k_x^{(d)}$ and $k_y^{(d)}$ in the new theory, and the roughness constant I in the diffusion based results. We now examine the two forms more closely and demonstrate that the apparent differences simply arise from the particular forms of proposal and target used. For simplicity we consider spherically symmetric unimodal targets explored using spherically symmetric proposals and contrast this with the exploration of targets with i.i.d components using a Gaussian proposal as performed in Roberts et al. (1997). In such cases the ESJD along any single component is simply $1/d$ times the overall ESJD. In Section 1.4.1 it was shown that for a diffusion evolving over a small time interval Δt, the ESJD along any single component is approximately the speed of the diffusion, h, multiplied by the time increment, Δt. Consider the transformed chain

$$\mathbf{Z}_{(t)}^{(d)} = \mathbf{X}_{[at]}^{(d)} \quad \text{where} \quad a = \left(\frac{d}{k_x^{(d)}}\right)^2$$

If as $d \to \infty$, $\mathbf{Z}_t^{(d)}$ were to approach a diffusion $\mathbf{Z}_t$ then from (3.63) the ESJD for a single component of $\mathbf{Z}_t$ in time Δt would be

$$\left(\frac{d}{k_x^{(d)}}\right)^2 \Delta t \times \frac{1}{d} S_d^2 \sim \left(\frac{d}{k_x^{(d)}}\right)^2 \Delta t \times \frac{8\left(k_x^{(d)}\right)^2}{d} \mu^2 \Phi\left(-\mu\right) = 8\mu^2 \Phi\left(-\mu\right) \Delta t$$

With $l = 2\mu$ this is the same as would arise from (3.4) except for the roughness constant I as defined in (3.5). However it was shown in Theorem 1 that as $d \to \infty$ each component of $\mathbf{X}^{(d)}$ approaches a multiple, $k_x^{(d)}/d^{1/2}$, of a standard Gaussian. Since (in this case) $k_x^{(d)} = d^{1/2}$ the limiting component is exactly a standard Gaussian, for which $I = 1$. Hence the newly derived form (3.63) for the limiting ESJD is indeed very closely related to the existing form (3.4) for the speed of the limiting diffusion.

The optimal scale parameter is given for example in Roberts et al. (1997) as $2.38/d^{1/2}$, and not $2.38k_x^{(d)}/(d^{1/2}k_y^{(d)})$ as presented in Corollary 8. However the targets considered in the earlier work had independent components, for which $k_x^{(d)} = d^{1/2}$, and similarly $k_y^{(d)} = d^{1/2}$ for the Gaussian proposal employed. The two forms are again equivalent.

We now consider exploration of a target with independent components as described by Roberts and Rosenthal (2001). We compare relative efficiencies for different shaped proposals for this scenario with those for exploration of a unimodal elliptical target as described in Section 3.3.2.2. In Section 3.1.1 we review Theorem 5 of Roberts and Rosenthal (2001) in which the speed of the limiting diffusion for the exploration of the first component of a target of the form (3.7) with a spherical Gaussian proposal is $(C_1^2/b) \times h(lb^{1/2})$ where $b = \mathbb{E}[C_i^2]$. It was noted that this is the speed of the limiting "scaleless" diffusion and that the equivalent limiting diffusion on the correct scale has speed $(1/b) \times h(lb^{1/2})$. This was compared with the speed of the limiting diffusion when exploring such a target with a similarly shaped Gaussian proposal. Since this is equivalent to exploring a target with i.i.d. components with a spherical Gaussian proposal, the speed of the limiting "scale-

less" diffusion was simply $h(l)$ as given in (3.4).

Now sum over all components, first in the "scaleless" space where the i^{th} component has been multiplied by C_i so that it has scale factor 1. If we were exploring a target of d dimensions then the ratio of total speeds of travel around the scaleless target would be approximately

$$\text{rel.eff}_{scaleless} = \frac{1}{\mathbb{E}\left[C_i^2\right]} \frac{\sum_1^d C_i^2}{\sum_1^d 1} \approx \frac{\mathbb{E}\left[C_i^2\right]}{\mathbb{E}\left[C_i^2\right]} = 1 \tag{3.104}$$

Speeds in the original space are those in the transformed space multiplied by $1/C_i^2$ and so the ratio of efficiencies along the first (or any other) component remain the same, but ratio of total speeds of travel around the original target would be approximately

$$\text{rel.eff}_{orig} = \frac{1}{\mathbb{E}\left[C_i^2\right]} \frac{\sum_1^d 1}{\sum_1^d \frac{1}{C_i^2}} \approx \frac{1}{\mathbb{E}\left[C_i^2\right] \mathbb{E}\left[C_i^{-2}\right]} \tag{3.105}$$

Both formulae become more exact as $d \to \infty$. The latter is strongly reminiscent of (3.86), its equivalent for the exploration of a unimodal elliptical target with spherical and then with elliptical proposals. As is also noted in Section 3.3.2.2, the efficiency ratio of such an exploration in the transformed (scaleless) space is 1, which concurs with (3.104).

Next consider partial blocking algorithms and compare the new results of Section 3.3.3 with those of Neal and Roberts (2006). We examined partial-blocking algorithms on elliptically symmetric targets for which (after an orthogonal linear map, if necessary) the rescaled modulus converges in probability to 1. For spherically symmetric targets, partial blocking was shown to make no difference to the efficiency of the algorithm, as measured by ESJD (or naive ESJD). These results hold for any jump proposal density including the spherically symmetric, for which the

ESJD along any component is simply the total ESJD divided by the dimension d. The results are therefore directly comparable with, and in agreement with those of Neal and Roberts (2006) who investigated the exploration of i.i.d. product densities using spherical Gaussian proposals and found the speed of the limiting diffusion along a single component to be unaffected by partial blocking. We then found (Section 3.3.3.2) that, with ESJD as a measure of efficiency, this continued to hold for elliptically symmetric unimodal targets. However if efficiency was measured through naive ESJD, partial blocking was found to increase the efficiency of the algorithm for elliptically symmetric targets unless the blocking scheme consisted of relatively large blocks with large scale parameters and relatively small blocks with small parameters, in which case the efficiency could actually decrease. At first glance this appears contrary to the theoretical and simulation results of Neal and Roberts (2006) for non-i.i.d. targets, which found no change in efficiency; however Neal and Roberts (2006) consider the random scan algorithm and Section 3.3.3 considers sequential updates. Applying the weak law of large numbers, in the limit as $d \to \infty$ the average scale parameter of a partial block selected by random scan is the overall mean scale parameter and does not vary between partial blocks. Compared to a single block update, our sequential scan would also produce no change in efficiency if the partial blocks all had the same mean (or in fact harmonic mean square) scale parameter. Further, efficiency in the simulation study of Neal and Roberts (2006) is measured in terms of expected square jumping distance along the first target component only. For non-spherical targets and proposals this is not directly comparable with either ESJD or naive ESJD.

Finally consider the condition (3.12) of Bedard (2006c) for the existence of a limiting Langevin diffusion with asymptotically optimal acceptance rate 0.234. Note that if the largest power $\gamma_{max} := \max_i \gamma_i$ of d occurs infinitely often then (3.12) certainly holds with $\lambda = \gamma_{max}$ and therefore it also holds with λ the largest power that occurs finitely often. The quoted result is thus equivalent to setting $\lambda = \gamma_{max}$. This is identical to the condition we require on the inverse scale parameters of elliptically symmetric targets in Section 3.3.2.2. It ensures continued convergence in mean square of the rescaled (initially spherical) proposal after the orthogonal linear transformation that also turns the elliptical target into a spherical target.

3.4.3 Further work

It has not been possible to persue exhaustively all of the concepts and ideas arising from the work presented in this chapter. This section presents some possible avenues for future exploration.

The Exchangeability Lemmas of Section 3.2 apply to all but the very unusual Metropolis-Hastings algorithms. In this chapter we have investigated some of their implications for the random walk Metropolis but the implications for other algorithms are yet to be explored in detail. Two algorithms have been briefly considered, though the work is not included in this thesis: the independence sampler, and the MALA algorithm. Difficulties arise in both cases through the more complex forms of the acceptance regions and in the case of the MALA algorithm also through the lack of separability of the proposal. However progress appears possible with both algorithms for certain combinations of target and proposal.

In Section 3.3.2 we examined exploration of an elliptically symmetric target with a spherically symmetric proposal. We derived a condition on the eigenvalues of the ellipse under which the limiting acceptance rate remained 0.234. It would be interesting to investigate limiting behaviour if this condition were to fail.

Limit theorems might also be possible for more general forms of target. Ideas from Sections 3.3.1.8 and 3.4.1 about the form of the optimal scaling and on when the optimal acceptance rate is equal to or strictly less than 0.234 might well carry over to more general distributions. Intuitively the important point is the curvature of contours in the d dimensional space. A conjecture would be that if the curvature is of the same order of magnitude thoughout the main target mass and if this mass is predominately spread over a relatively thin hyper-shell, then a single optimal scaling should exist and the optimal acceptance rate would be 0.234. If the hypershell were not "thin" then the acceptance rate would be less than 0.234 and if the curvature varied enormously across the main target mass then any single scaling would not suffice to properly explore the target. Such work would entail a different general approach to that in this thesis and might be better tackled by considering components of limiting diffusions both tangential to and perpendicular to the current contour.

The intuition in Section 3.4.1 of a stochastic radial component counteracting a deterministic tangential motion also suggests an adaptation of the RWM through the addition of a deterministic inward radial component. If a simple RWM were tuned to give an acceptance rate of 0.234 then the optimal scale parameter $\hat{\lambda}$ would provide an estimate of the ratio $k_y^{(d)}/k_x^{(d)}$, which is required in calculating the de-

terministic offset. A more thorough investigation of the relationship between this and the MALA algorithm could also be undertaken.

Finally it is worth noting that since the Exchangeability Lemmas (Section 3.2.2) only apply to stationary chains, they provide a means for assessing stationarity, taking into account all components of a chain. For a stationary chain the number of proposals in the acceptance region (all of which are accepted) should be roughly equal to the sum of the acceptance rates of proposals in the rejection region. Any one of a number of other measures (such as the naive ESJD) to which the Exchangeability Lemmas apply could be used instead of acceptance rate.

3.4.4 Discussion

We have investigated optimal-scaling of the random walk Metropolis algorithm on unimodal elliptically symmetric targets using the expected square jumping distance (ESJD) as a measure of efficiency. In this section we summarise the new results.

We obtained exact analytical expressions for the expected acceptance rate and the ESJD in finite dimension d. This became feasible through two "exchangeability lemmas" (and their extensions) which are valid for most sensible Metropolis-Hastings algorithms. Initial results were presented for spherically symmetric unimodal targets with any proposal. From these exact forms it was straightforward to show that expected acceptance rate does indeed vary monotonically from 1 to 0 as the proposal scaling parameter increases from 0 to ∞. This bijective mapping justifies to an extent the use of acceptance rate as a proxy for the scale parameter. It was also shown that all RWM's on finite dimensional targets with finite second

moments posses scaling(s) that maximise the ESJD and that this (or these) are finite. Explicit forms for expected acceptance rate and ESJD in terms of marginal radial densities were also used to explore specific combinations of target and proposal in finite dimensions. Numerical and analytical results agreed with our theory and with a simulation study in Roberts and Rosenthal (2001). All theoretical results were also shown to extend to elliptically symmetric targets through the use of a simple orthogonal linear map.

An asymptotic theory was developed for the behaviour of the algorithm as dimension $d \to \infty$. The asymptotically optimal acceptance rate of 0.234 was shown to extend to the class of spherically symmetric unimodal targets that can be rescaled so that their absolute value converges in probability to 1. An asymptotic form for the optimal scale parameter was also derived. The class for which the results are valid was then extended to include all elliptically symmetric targets which satisfy the same condition once they have been transformed to spherical symmetry by an orthogonal linear map. If the original target is being explored by a spherically symmetric proposal then an additional constraint applies to the eigenvalues of the linear map which forbids the scale parameter of the smallest principle component from being "too much smaller" than all the other scale parameters. This condition is equivalent to that of of Bedard (2006c), derived for targets with independent components identical up to a scaling. An expression was derived for the ratio of the naive ESJD's for the exploration of an elliptically symmetric target using either a spherically symmetric proposal or an elliptical proposal of the same shape as the target. An equivalent formula was derived in Section 3.4.2 for a target with similar independent components using a theorem of Roberts and Rosenthal (2001).

We also considered partial blocking, again on targets for which (after a linear map, if necessary) the rescaled modulus converges in probability to 1. For spherically symmetric targets, where results were comparable, they agreed with those of Neal and Roberts (2006): partial blocking does not affect efficiency. However we found that partial blocking does affect naive ESJD on elliptically symmetric targets, whereas partial blocking did not affect efficiency of exploration of the non-i.i.d. targets investigated by Neal and Roberts (2006). Although the efficiency measure used to guage exploration of non i.i.d targets is not directly comparable with naive ESJD (or ESJD) the different choice of update scheme (random scan in Neal and Roberts (2006) and sequential in Section 3.3.3) is more likely to have lead to the heuristically different results.

Given that the above limiting results are all dependent on convergence to 1 of the rescaled modulus of the target, it was of interest to explore the types of target for which this condition holds, and the limiting behaviour when it fails. A sufficient condition was derived for convergence in probability to 1 of the rescaled modulus. The condition specifies the limiting behaviour of the target's tails and is satisfied by many common distributions; however several counter-examples were demonstrated. We therefore considered spherically symmetric targets for which the rescaled modulus converges to some fixed distribution other than a unit mass at 1. In such cases the asymptotically optimal acceptance rate was shown to be strictly less than 0.234. However provided the asymptotically optimal acceptance rate was strictly greater than zero, the optimal scale parameter was shown to exhibit very similar behaviour to that when the optimal acceptance rate was 0.234.

This theory was supported by numerical results obtained for specific combinations of target and proposal in finite dimensions. The optimality results do not cover targets for which the limiting radius has a point mass at the origin or infinity; this corresponds to at least two very different scales of variation. A simulation study on spherically symmetric unimodal targets with two very different scales of variation illustrated this problem and called into question the very concept of a single optimal scaling in such cases.

The theory for optimal scaling of the random walk Metropolis presented in this Chapter provides a substantially different approach to the problem; it both agrees with and extends the existing literature on the subject, as well as providing interesting possibilities for further work.

Appendix A

Cubic expansion of the MMPP log-likelihood

A Taylor expansion of the log-likelihood of a two-dimensional MMPP with $\lambda_1 \approx \lambda_2$ was stated in Section 2.5.4.2. The expansion is derived in detail in this appendix, starting with a form for a general d-dimensional MMPP and then simplifying the expression for the two-dimensional case. Further details of the $(\overline{\lambda}, q, \alpha, \beta)$ reparameterisation are also provided.

A.1 General d-dimensional MMPP

For a general MMPP, first reparameterise to $(\overline{\lambda}, \mathbf{\Lambda}_*, q, \mathbf{Q}_*)$ with

$$\overline{\lambda} = \boldsymbol{\nu}^t \boldsymbol{\lambda} \ , \quad \mathbf{\Lambda} = \overline{\lambda}(\mathbf{I} + \mathbf{\Lambda}_*) \ , \quad \mathbf{Q} = -q\mathbf{Q}_*$$

for some (at present) arbitrary q.

With this reparameterisation

$$e^{(\mathbf{Q}-\mathbf{\Lambda})t_i} = e^{-\overline{\lambda}t_i}e^{-(\mathbf{Q}_*qt_i+\mathbf{\Lambda}_*\overline{\lambda}t_i)}$$

and therefore

$$L(\mathbf{Q},\mathbf{\Lambda},\mathbf{t}) = \overline{\lambda}^n e^{-\overline{\lambda}t_{obs}}\boldsymbol{\nu}^t e^{-(\mathbf{Q}_*qt_1+\mathbf{\Lambda}_*\overline{\lambda}t_1)}(\mathbf{I}+\mathbf{\Lambda}_*)\ldots$$
$$\ldots e^{-(\mathbf{Q}_*qt_n+\mathbf{\Lambda}_*\overline{\lambda}t_n)}(\mathbf{I}+\mathbf{\Lambda}_*)e^{-(\mathbf{Q}_*qt_{n+1}+\mathbf{\Lambda}_*\overline{\lambda}t_{n+1})}\mathbf{1}$$

But

$$e^{-(\mathbf{Q}_*qt_i+\mathbf{\Lambda}_*\overline{\lambda}t_i)} = \mathbf{I} - (\mathbf{Q}_*qt_i+\mathbf{\Lambda}_*\overline{\lambda}t_i) + \frac{1}{2}(\mathbf{Q}_*qt_i+\mathbf{\Lambda}_*\overline{\lambda}t_i)^2 + \ldots$$

Expand the likelihood in terms of $\mathbf{\Lambda}_*$ and for notational simplicity, temporarily ignore the factor $\overline{\lambda}^n e^{-\overline{\lambda}t_{obs}}$ and products of powers of $\overline{\lambda}t_i$ and qt_i. Terms in $\mathbf{\Lambda}_*$ are then multiples of

$$\boldsymbol{\nu}^t{\mathbf{Q}_*}^a\mathbf{\Lambda}_*{\mathbf{Q}_*}^b\mathbf{1} \text{ with } a\geq 0, b\geq 0$$

Terms in $\mathbf{\Lambda}_*^2$ are multiples of

$$\boldsymbol{\nu}^t{\mathbf{Q}_*}^a\mathbf{\Lambda}_*{\mathbf{Q}_*}^b\mathbf{\Lambda}_*{\mathbf{Q}_*}^c\mathbf{1} \text{ with } a\geq 0, b\geq 0, c\geq 0$$

and terms in $\mathbf{\Lambda}_*^3$ are multiples of

$$\boldsymbol{\nu}^t{\mathbf{Q}_*}^a\mathbf{\Lambda}_*{\mathbf{Q}_*}^b\mathbf{\Lambda}_*{\mathbf{Q}_*}^c\mathbf{\Lambda}_*{\mathbf{Q}_*}^d\mathbf{1} \text{ with } a\geq 0, b\geq 0, c\geq 0, d\geq 0$$

From their definitions

$$\boldsymbol{\nu}^t\mathbf{Q} = \mathbf{Q}\mathbf{1} = \boldsymbol{\nu}^t\mathbf{\Lambda}_*\mathbf{1} = 0$$

Therefore terms in $\mathbf{\Lambda}_*$ vanish and remaining square and cubic terms are of respective forms

$$\boldsymbol{\nu}^t\mathbf{\Lambda}_*{\mathbf{Q}_*}^b\mathbf{\Lambda}_*\mathbf{1} \text{ with } b\geq 0$$

and

$$\boldsymbol{\nu}^t \boldsymbol{\Lambda}_* \mathbf{Q}_*{}^b \boldsymbol{\Lambda}_* \mathbf{Q}_*{}^c \boldsymbol{\Lambda}_* \mathbf{1} \text{ with } b \geq 0, c \geq 0$$

Note that if quadratic and higher terms in $\boldsymbol{\Lambda}_*$ are ignored, the log-likelihood is that of a simple Poisson process

$$l(\overline{\lambda}, \boldsymbol{\Lambda}_*, q, \mathbf{Q}_*) \approx n \log \overline{\lambda} - \overline{\lambda} t_{obs} \tag{A.1}$$

A.2 Two-dimensional MMPP

We now focus on the two-dimensional MMPP, setting $q := q_{12} + q_{21}$ and $\delta := (\lambda_2 - \lambda_1)/\overline{\lambda}$. In this case

$$\boldsymbol{\Lambda}_* = \delta \begin{bmatrix} -\nu_2 & 0 \\ 0 & \nu_1 \end{bmatrix} \quad \text{and} \quad \mathbf{Q}_* = \begin{bmatrix} \nu_2 & -\nu_2 \\ -\nu_1 & \nu_1 \end{bmatrix}$$

Moreover

$$\mathbf{Q}_*{}^n = \mathbf{Q}_*$$

Quadratic terms in $\boldsymbol{\Lambda}_*$ are therefore multiples of

$$\boldsymbol{\nu}^t \boldsymbol{\Lambda}_*^2 \mathbf{1} \text{ or } \boldsymbol{\nu}^t \boldsymbol{\Lambda}_* \mathbf{Q}_* \boldsymbol{\Lambda}_* \mathbf{1}$$

and cubic terms are multiples of

$$\boldsymbol{\nu}^t \boldsymbol{\Lambda}_*^3 \mathbf{1} \text{ or } \boldsymbol{\nu}^t \boldsymbol{\Lambda}_* \mathbf{Q}_* \boldsymbol{\Lambda}_*^2 \mathbf{1} \text{ or } \boldsymbol{\nu}^t \boldsymbol{\Lambda}_*^2 \mathbf{Q}_* \boldsymbol{\Lambda}_* \mathbf{1} \text{ or } \boldsymbol{\nu}^t \boldsymbol{\Lambda}_* \mathbf{Q}_* \boldsymbol{\Lambda}_* \mathbf{Q}_* \boldsymbol{\Lambda}_* \mathbf{1}$$

But $\boldsymbol{\Lambda}_* \mathbf{1} = \delta[-\nu_2, \nu_1]^t$ is a right eigenvector of $\mathbf{Q}^*$ and $\boldsymbol{\nu}^t \boldsymbol{\Lambda}_* = \delta[\nu_1, \nu_2]$ is a left eigenvector of $\mathbf{Q}^*$, both with eigenvalues 1. Hence in the above products $\mathbf{Q}_*$ has no effect; both quadratic terms evaluate to $\delta^2 \nu_1 \nu_2$, and all cubic terms evaluate to $\delta^3 \nu_1 \nu_2 (\nu_2 - \nu_1)$.

To cubic terms in δ, the likelihood is therefore

$$L(\overline{\lambda}, q, \delta, \nu_1) \approx \overline{\lambda}^n e^{-\overline{\lambda} t_{obs}} \left(1 + 2\delta^2 \nu_1 \nu_2 f(\overline{\lambda}\mathbf{t}, q\mathbf{t}) + \delta^3 \nu_1 \nu_2 (\nu_2 - \nu_1) g(\overline{\lambda}\mathbf{t}, q\mathbf{t})\right)$$

where $f(\cdot,\cdot)$ and $g(\cdot,\cdot)$ are the sums of the many product terms in the expansion of the likelihood involving respectively two and three occurences of $\mathbf{\Lambda}_*$. From this, by a further Taylor expansion, the log-likelihood is

$$l(\overline{\lambda}, q, \delta, \nu_1) = n \log \overline{\lambda} - \overline{\lambda} t_{obs} + 2\delta^2 \nu_1 \nu_2 f(\overline{\lambda}\mathbf{t}, q\mathbf{t}) + \delta^3 \nu_1 \nu_2 (\nu_2 - \nu_1) g(\overline{\lambda}\mathbf{t}, q\mathbf{t}) + O(\delta^4)$$

This is Equation (2.23).

A.3 The $(\overline{\lambda}, q, \alpha, \beta)$ reparameterisation

The likelihood expansion (2.23) suggests a further reparameterisation (as described in Section 2.5.4.2) to $\overline{\lambda}$ and q as defined above and

$$\alpha := 2\delta(\nu_1\nu_2)^{1/2} \quad \text{and} \quad \beta := \delta(\nu_2 - \nu_1)$$

Parameters $\overline{\lambda}, \alpha$ and β (in this order) capture decreasing amounts of variation in the log-likelihood and so, conversely, it might be anticipated that there be corresponding decreasing amounts of information about the parameters contained in the likelihood, and so very different scalings required for each. This section provides further details of the reparameterisation.

Viewed in terms of the original parameters, we have

$$\begin{aligned}
\overline{\lambda} &:= \frac{q_{21}\lambda_1 + q_{12}\lambda_2}{q_{12} + q_{21}} \\
q &:= q_{12} + q_{21} \\
\alpha &:= 2\frac{(\lambda_2 - \lambda_1)(q_{12}q_{21})^{1/2}}{q_{21}\lambda_1 + q_{12}\lambda_2} \\
\beta &:= \frac{(\lambda_2 - \lambda_1)(q_{12} - q_{21})}{q_{21}\lambda_1 + q_{12}\lambda_2}
\end{aligned}$$

The inverse transformation has intermediate steps

$$\begin{aligned}
\delta &= \text{sign}(\alpha)(\alpha^2 + \beta^2)^{1/2} \\
\nu_1 &= \frac{1}{2}(1 - \frac{\beta}{\delta}) \\
\nu_2 &= \frac{1}{2}(1 + \frac{\beta}{\delta})
\end{aligned}$$

and is

$$\begin{aligned}
\lambda_1 &= \overline{\lambda}(1 - \nu_2\delta) \\
\lambda_2 &= \overline{\lambda}(1 + \nu_1\delta) \\
q_{12} &= \nu_2 q \\
q_{21} &= \nu_1 q
\end{aligned}$$

The Jacobian of the transformation is

$$\frac{\partial(\overline{\lambda}, q, \alpha, \beta)}{\partial(\lambda_1, \lambda_2, q_{12}, q_{21})} = \frac{|\lambda_2 - \lambda_1|(q_{12} + q_{21})^2}{(q_{21}\lambda_1 + q_{12}\lambda_2)^2(q_{12}q_{21})^{1/2}}$$

Appendix B

Additional simulated MMPP data sets and comparisons

Table B.1 lists additional simulated data sets used for comparison between our Gibbs sampler and the Metropolis-Hastings random walk algorithms (M1-M5). As with the core runs, in all runs on these additional data sets the priors are invariant to label-switching. Prior distributions for all q and λ parameters are exponential with the mean for q components set to 1 (even for HL and LH runs) and the λ component means calculated as described in Section 2.6.

Tables B.2, B.3, and B.4 show the relative integrated autocorrelation times for the additional runs. Estimates of the information matrices for replicate 1 of S2 and S3 are given in Table 2.5. CPU timings for 1000 iterations of each algorithm on two of the data sets appear in Table B.6.

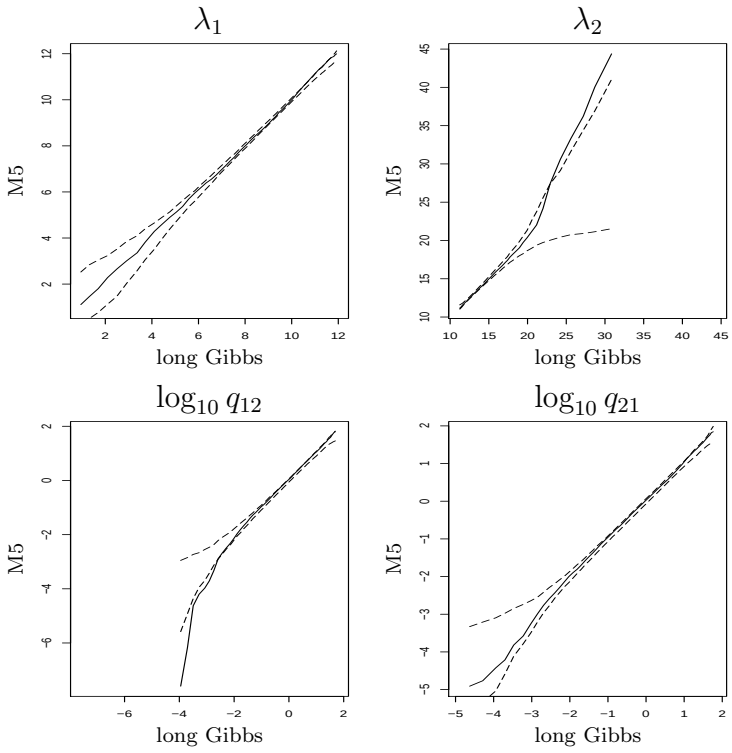

Figure B.1: qq plots for replicate 1 of S4, comparing the first 10 000 iterations of algorithm M5 against iterations 11 000 - 100 000 of the Gibbs sampler. Dashed lines are approximate 95% confidence limits obtained by repeated sampling from iterations 11 000 to 100 000 of the Long Gibbs data; sample sizes were 10 000/ACT, which is the effective sample size of the data being compared to the Long Gibbs run.

Dataset	λ_1	λ_2	q_{12}	q_{21}	t	replicates
S3*	10	17	1	1	400	2
HL1	10	90	2	0.5	100	1
HL4	10	13	2	0.5	100	1
LH1	10	90	0.5	2	100	1
LH4	10	13	0.5	2	100	1

Table B.1: Parameter values for additional simulated data sets.

		replicate 2				replicate 3			
Data	Alg	λ_1	λ_2	q_{12}	q_{21}	λ_1	λ_2	q_{12}	q_{21}
S1	Gibbs	1.2	1.2	1.5	1.6	1.4	1.1	1.4	1.3
	M1	13.9	13.9	14.7	15.0	14.8	14.3	14.0	13.7
	M2	13.7	13.9	15.0	14.5				
	M3	37.1	6.3	10.2	10.7				
	M4	7.5	65.6	12.9	14.1				
S2	Gibbs	4.8	4.7	6.4	7.4	5.2	5.1	7.5	8.3
	M1	21.7	24.8	29.6	30.8	19.9	22.7	29.4	32.4
	M2	22.0	21.7	27.0	30.5				
	M3	11.7	16.6	16.1	20.2				
	M4	15.6	26.4	28.4	31.7				
S3	Gibbs	10.6	9.1	14.4	17.8	44.4	20.2	70.4	38.5
	M1	43.4	39.2	43.3	51.4	173.6	75.3	221.3	91.5
	M2	38.9	38.6	46.1	58.7				
	M3	24.8	25.5	18.0	20.0				
	M4	115.9	46.6	257.8	59.8				
S4	Gibbs	8.1	4.5	9.2	16.0	11.4	3.7	10.1	25.7
	M1	128.4	64.1	47.9	52.7	136.4	77.3	48.0	64.8
	M2	79.7	63.6	37.7	75.7				
	M3	127.1	51.7	62.4	66.2				
	M4	32.2	99.7	36.7	40.7				

Table B.2: Estimated ACT_{rel} for replicates 2 and 3 of simulated data sets S1-S4. The poor mixing of M3 for λ_1 on replicate 2 of S1 is simply due to bad tuning (the random walk standard deviation for λ_1 was too small) and serves to emphasise the difficulty of optimal tuning for block random walks.

		replicate 1				replicate 2			
Data	Alg	λ_1	λ_2	q_{12}	q_{21}	λ_1	λ_2	q_{12}	q_{21}
S3*	Gibbs	46.3	20.5	73.3	40.3	41.9	23.1	52.1	32.5
	M1	127.7	74.1	173.6	102.6	147.7	89.3	162.5	91.5
	M2	95.2	69.5	134.4	97.0	107.4	80.6	132.9	78.6
	M3	49.1	39.1	61.5	41.5	50.3	41.7	52.6	38.5
	M4	142.6	79.5	187.9	93.1	110.9	91.1	115.0	78.6
	M5	24.5	16.6	34.7	24.8				

Table B.3: Estimated ACT_{rel} for replicates 1-2 of S3*.

Data	Alg	λ_1	λ_2	q_{12}	q_{21}	Data	λ_1	λ_2	q_{12}	q_{21}
HL1	Gibbs	2.0	1.2	1.7	1.9	LH1	1.1	1.6	1.6	1.6
	M1	12.6	13.7	16.5	17.1		14.9	14.4	14.7	14.9
	M2	13.6	13.7	14.4	16.2		13.7	13.5	13.7	16.6
	M5	15.0	13.1	213.3	163.4		16.5	26.2	47.0	177.1
HL4	Gibbs	10.3	4.6	9.8	20.4	LH4	39.9	35.4	47.5	31.3
	M1	160.9	141.4	50.5	61.4		250.7	188.5	195.8	144.7
	M2	117.1	895.8	127.0	84.7		145.2	217.2	198.1	136.6
	M5	42.2	45.9	24.1	24.3		22.7	41.7	25.4	22.1

Table B.4: Estimated ACT_{rel} for HL1, LH1, HL4, and LH4

Data	Information Matrix
S2	$\begin{bmatrix} 2.74 & -0.37 & 2.84 & -0.65 \\ -0.37 & 1.13 & -.41 & -1.75 \\ 2.84 & 0.41 & 28.94 & -15.57 \\ -0.65 & -1.75 & -15.57 & 28.12 \end{bmatrix}$
S3	$\begin{bmatrix} 2.07 & -0.13 & 5.25 & -3.84 \\ -0.13 & 1.72 & 3.42 & -5.18 \\ 5.25 & 3.42 & 34.96 & -29.91 \\ -3.84 & -5.18 & -29.91 & 42.90 \end{bmatrix}$

Table B.5: Information matrices for replicate 1 of S2 and S3 at the MLE, estimated by numerical differentiation.

Data	Gibbs	M1	M2	M3	M4	M5
S1	287.7	915.5	909.2	228.4	905.4	1002.8
S4	57.9	187.9	177.8	46.8	169.4	205.5

Table B.6: CPU timings (secs) for 1000 iterations of each algorithm on replicate 1 of S1 and S4 with an AMD Athlon 1458MHz CPU.

Appendix C

Proofs of limit theorems in Chapter 3

Proof of Lemma 8:

We require Chebyshev's inequality (see for example Grimmett and Stirzaker (2001), Chapter 7).

$$P(|X-\mu| \geq a) \leq \frac{\mathbb{E}\left[|X-\mu|^2\right]}{a^2}$$

Since $\left|\mathbf{Z}^{(n)}\right|^2 \sim \chi^2_n$, $\mathbb{E}\left[\left|\mathbf{Z}^{(n)}\right|^2\right] = n$ and $\text{Var}\left[\left|\mathbf{Z}^{(n)}\right|^2\right] = 2n$. So by Chebyshev's inequality

$$P\left(\left|\frac{\left|\mathbf{Z}^{(n)}\right|^2}{n} - 1\right| > \epsilon\right) = P\left(\left|\frac{\left|\mathbf{Z}^{(n)}\right|^2 - n}{n}\right| > \epsilon\right) \leq \frac{2n}{n^2\epsilon^2} = \frac{2}{n\epsilon^2}$$

Thus $\frac{\left|\mathbf{Z}^{(n)}\right|^2}{n} \xrightarrow{p} 1$ and hence the result.

Note that although the method of proof is identical, this is **not** just the weak law of large numbers for χ^2_1 random variables since we are considering a sequence of

sequences of χ^2_1 random variables.

Proof of Lemma 9:

Observe that the limit of a series of (identically) bounded monotonic functions is also bounded (with the same bounds) and monotonic. Without loss of generality we assume $G(x)$ is increasing (otherwise simply consider $G(-x)$), with upper limit 1 and lower limit 0 (otherwise simply rescale).

First deal with the ends of the domain. Since $G(x)$ is monotonic with lower limit 0 and upper limit 1 there are x_1 and x_2 such that $G(x) < \epsilon/2$ for all $x < x_1$ and $1 - \epsilon/2 < G(x) \leq 1$ for all $x > x_2$. Also there is a d_1 such that for all $d > d_1$, $|G(x_1) - G_d(x_1)| < \epsilon/2$ and $|G(x_2) - G_d(x_2)| < \epsilon/2$. Since the $G_d(x)$ are also increasing we see that for all $d > d_1$ and all $x \in (-\infty, x_1) \cup (x_2, \infty)$

$$|G(x) - G_d(x)| < \epsilon$$

Consider $[x_1, x_2]$ the (compact) remainder of the domain, and recall that $G(x)$ is continuous throughout this interval. Since $G(x)$ is continuous, it is uniformly continuous on $[x_1, x_2]$. So there is a δ such that $|G(x) - G(y)| < \epsilon/2$ for all $|x - y| < \delta$. Let P be the net of points

$$\left\{ x_1, x_1 + \delta/2, \ldots, x_1 + k\delta/2, \ldots, x_1 + \left[\frac{x_2 - x_1}{\delta/2} \right] \delta/2, x_2 \right\}$$

This is a finite net, so there is also a d_2 such that for all $d > d_2$, and all $x \in P$, $|G(x) - G_d(x)| < \epsilon/2$.

For any $x \in [x_1, x_2] \setminus P$ we may pick any adjacent pair $x_p, x_q \in P$ with $x_p < x < x_q$. Since $x_q - x_p < \delta$ we have $G(x_q) - G(x_p) < \epsilon/2$; also for all $d > d_2$ we have that

$|G(x_p) - G_d(x_p)| < \epsilon/2$ and $|G(x_q) - G_d(x_q)| < \epsilon/2$. But both $G(x)$ and $G_d(x)$ are increasing so that for any $x \in (x_p, x_q)$ we have that $G(x_p) \leq G(x) \leq G(x_q)$ and $G_d(x_p) \leq G_d(x) \leq G_d(x_q)$ therefore $|G(x) - G_d(x)| < \epsilon$. Combined with (uniform) convergence on P, this gives uniform convergence on $[x_1, x_2]$.

Proof of Theorem 2:

In proving Theorem 2 we use the following:

Lemma 22 *Let* $U_d \sim Beta\left(\frac{1}{2}, \frac{d-1}{2}\right)$ *then* $d^{1/2}U_d^{1/2} \xrightarrow{D} |Z|$ *, where* $Z \sim N(0,1)$.

Proof: *We will show that* $d\ U_d \xrightarrow{D} \chi_1^2$ *from which the required result follows since* $U_d^{1/2}$ *is positive.*

We may represent any Beta random variate in terms of two independent Gamma variates (see for example Hogg and Craig, 1995). Here

$$U_d = \frac{X_d}{X_d + Y_d}$$

where $X_d \sim Gam\left(\frac{1}{2}, 1\right)$ *and* $Y_d \sim Gam\left(\frac{d-1}{2}, 1\right)$. *Since* $X_d + Y_d \sim \chi_d^2$ *we apply Lemma 8 to obtain*

$$dU_d = \frac{X_d}{(X_d + Y_d)/d} \xrightarrow{D} \chi_1^2$$

Now, return to the main theorem and swap the order of integration

$$\begin{aligned}
\mathbb{E}_R\left[\Phi\left(\frac{x_1}{R}\right)\right] &= \int_0^\infty d\overline{\Theta}\,(r)\int_{-\infty}^{x_1/r} d\Phi(z) \\
&= \frac{1}{2} + sign(x_1)\int_0^\infty d\overline{\Theta}\,(r)\int_0^{|x_1|/r} d\Phi(z) \\
&= \frac{1}{2} + sign(x_1)\int_0^\infty d\Phi(z)\int_0^{|x_1|/z} d\overline{\Theta}\,(r) \\
&= \frac{1}{2}\left(1 + sign(x_1)\mathbb{E}_Z\left[\overline{\Theta}\left(\frac{|x_1|}{|Z|}\right)\right]\right) \qquad \text{(C.1)}
\end{aligned}$$

where $Z \sim N(0,1)$. But from Lemma 7 we have

$$F_{1|d}\left(\frac{k_d}{d^{1/2}}x_1\right) = \frac{1}{2}\left(1 + sign(x_1)\mathbb{E}\left[\overline{F}_d\left(\frac{|k_d\ x_1|}{d^{1/2}\ U_d^{1/2}}\right)\right]\right)$$

where $U_d \sim Beta\left(\frac{1}{2}, \frac{d-1}{2}\right)$. So it is sufficient to show that

$$\mathbb{E}\left[\overline{F}_d\left(\frac{k_d x_1}{d^{1/2}U_d^{1/2}}\right)\right] \to \mathbb{E}_Z\left[\overline{\Theta}\left(\frac{x_1}{|Z|}\right)\right] \quad \text{for } X_1 \geq 0$$

Writing the density of U_d as $g_d(\cdot)$ and noting that $lim_{d\to\infty}(d-3)/d = 1$, we must equivalently show that

$$\lim_{d\to\infty}\int_0^1 du\ \overline{F}_d\left(\frac{k_d\ x_1}{(d-3)^{1/2}\ u^{1/2}}\right)\ g_d(u) = \int_0^\infty dv\ \overline{\Theta}\left(\frac{x_1}{v}\right)\ 2\phi(v)$$

We first change variables, setting $V_d := (d-3)^{1/2}U_d^{1/2}$. The density function of V_d is therefore

$$\begin{aligned}
g_d^*(v) &= \frac{2}{(d-3)^{1/2}B\left(\frac{1}{2}, \frac{d-1}{2}\right)}\left(1 - \frac{v^2}{d-3}\right)^{(d-3)/2} \qquad (0 \leq v \leq d-3) \\
&\quad 0 \quad (v < 0 \text{ or } v > d-3)
\end{aligned}$$

and we will denote the corresponding distribution function as $G_d^*(\cdot)$. So

$$\int_0^1 du\ \overline{F}_d\left(\frac{k_d\ x_1}{(d-3)^{1/2}\ u^{1/2}}\right) g_d(u) = \int_0^\infty dv\ \overline{F}_d\left(\frac{k_d\ x_1}{v}\right) g_d^*(v)$$

We must therefore show that

$$\left| \int_0^\infty dv \ \overline{\Theta}\left(\frac{x_1}{v}\right) \ 2\phi(v) - \overline{F}_d\left(\frac{k_d \ x_1}{v}\right) \ g_d^*(v) \right| \to 0$$

Now

$$\left| \int_0^\infty dv \ \overline{\Theta}\left(\frac{x_1}{v}\right) \ 2\phi(v) - \overline{F}_d\left(\frac{k_d \ x_1}{v}\right) \ g_d^*(v) \right| \leq$$
$$\left| \int_0^\infty dv \ \overline{\Theta}\left(\frac{x_1}{v}\right) (2\phi(v) - g_d^*(v)) \right| + \left| \int_0^\infty dv \ g_d^*(v) \left(\overline{\Theta}\left(\frac{x_1}{v}\right) - \overline{F}_d\left(\frac{k_d \ x_1}{v}\right) \right) \right|$$

But

$$\left| \int_0^\infty dv \ g_d^*(v) \left(\overline{\Theta}\left(\frac{x_1}{v}\right) - \overline{F}_d\left(\frac{k_d \ x_1}{v}\right) \right) \right| = \left| \mathbb{E}\left[\overline{\Theta}\left(\frac{x_1}{V_d}\right) - \overline{F}_d\left(\frac{k_d \ x_1}{V_d}\right) \right] \right| \to 0$$

since $\overline{\Theta}(\cdot)$ is continuous and

$$\overline{\Theta}(a) \to \overline{F}_d(k_d \ a)$$

and this convergence is uniform by Lemma 9. Therefore it remains to show that

$$\left| \int_0^\infty dv \ \overline{\Theta}\left(\frac{x_1}{v}\right) (2\phi(v) - g_d^*(v)) \right| \to 0$$

Since $0 \leq \overline{\Theta}\left(\frac{x_1}{v}\right) \leq 1$

$$\left| \int_0^\infty dv \ \overline{\Theta}\left(\frac{x_1}{v}\right) (2\phi(v) - g_d^*(v)) \right| \leq \int_0^\infty dv \ \ |2\phi(v) - g_d^*(v)|$$

Split the integration region $[0, \infty)$ into $[0, v^*] \cup [v^*, \infty)$ with v^* chosen such that $1 - \Phi(v^*) < \epsilon/8$. From Lemma 22 we know that for $v \geq 0$, $G_d^*(v) \to 2\Phi(v) - 1$, and by Lemma 9 this convergence is uniform. So there is a d_1 such that for all $d > d_1$,

$$|1 - G_d^*(v) + 2\Phi(v) - 2| = |G_d^*(v) - 2\Phi(v) + 1| < \epsilon/4$$

Thus for all $d > d_1$

$$1 - G_d^*(v*) < |1 - G_d^*(v*) + 2\Phi(v*) - 2| + 2 - 2\Phi(v*) < \frac{\epsilon}{2}$$

Hence

$$\int_{v^*}^{\infty} dv \ |2\phi(v) - g_d^*(v)| \leq \int_{v^*}^{\infty} dv \ 2\phi(v) + \int_{v^*}^{\infty} dv \ g_d^*(v)$$
$$= 2(1 - \Phi(v^*)) + (1 - G_d^*(v^*)) < \frac{3\epsilon}{4} \quad \text{(C.2)}$$

In Section 1.5.1 it was shown that $(d-3)^{1/2} B\left(\frac{1}{2}, \frac{d-1}{2}\right) \to (2\pi)^{1/2}$. Therefore $g_d^*(v) \to 2\phi(v)$ pointwise. We also note that $g_d^*(v)$ is bounded and monotone decreasing in v, as is $2\phi(v)$. Therefore by a trivial extension of Lemma 9 (since $\sup_v g_d^*(v) = g_d^*(0) \to 2\phi(0)$) the convergence is uniform. Hence there is a d_2 such that for all $d > d_2$ and all v, $|2\phi(v) - g_d^*(v)| < \epsilon/4v^*$. So for all $d > d_2$

$$\int_0^{v^*} dv \ |2\phi(v) - g_d^*(v)| < \frac{\epsilon}{4} \quad \text{(C.3)}$$

Combining (C.2) and (C.3) we see that for all $d > \max(d_1, d_2)$

$$\int_0^{\infty} dv \ |2\phi(v) - g_d^*(v)| < \epsilon$$

Proof of Lemma 13

Our proof requires the following simple result.

Lemma 23 *For any given finite $r_0 > 0$*

$$\int_0^{r_0} dr \ f_d^*(r) \to 0 \quad as \quad d \to \infty$$

Proof: *First define*

$$I(s) := \int_0^s dr \ r^{d-1} \exp(-g(r))$$

Then

$$I(r_0) < r_0^{d-1} \int_0^{r_0} dr \ \exp\left(-g(r)\right)$$

and

$$I(\infty) > \int_{2r_0}^{\infty} dr \ r^{d-1} \exp\left(-g(r)\right) > \left(2r_0\right)^{d-1} \int_{2r_0}^{\infty} dr \ \exp\left(-g(r)\right)$$

Therefore

$$\int_0^{r_0} dr \ f_d^*(r) = \frac{I(r_0)}{I(\infty)} < \frac{1}{2^{d-1}} \frac{\int_0^{r_0} dr \ \exp\left(-g(r)\right)}{\int_{2r_0}^{\infty} dr \ \exp\left(-g(r)\right)} \to 0 \quad as \quad d \to \infty$$

Note that both the integrals in this expression are strictly positive as $g(\cdot)$ is bounded above over any finite interval.

We now prove Lemma 13 itself. Note that we will find (Equation C.10) that the density of the rescaled modulus $f_d^{**}(u)$ is approximately that of a $Gam(kb_d, kb_d)$ random variable with $b_d \to \infty$ as $d \to \infty$, from which the final result is intuitively clear but still takes some considerable algebra to prove.

Given ϵ with $0 < \epsilon < 1$ choose r_0 such that for $r > r_0$, (3.55) gives

$$k(1-\epsilon) < \frac{h(r)}{r^a} < k(1+\epsilon)$$

To simplify notation later on we now define $k_l := k(1-\epsilon)$ and $k_u := k(1+\epsilon)$.

Multiply (3.53) by r and differentiate to find

$$\left(r \left(\log f_d^*(r)\right)'\right)' = -\left(rg'(r)\right)' = -\frac{h(r)}{r}$$

Thus for $r > r_0$ we have

$$-k_u \ r^{a-1} < \left(r \left(\log f_d^*(r)\right)'\right)' < -k_l \ r^{a-1} \tag{C.4}$$

From (3.55) there is an $r^* < r_0$ such that $h(r) > 0$ for all $r > r^*$ and so by Lemma 12 we may choose d_1 such that $r_d > r_0$ for all $d > d_1$. Consider first the right hand inequality for $d > d_1$ and integrate between r_d and r to obtain

$$r \left(\log f_d^* (r)\right)' < -\frac{k_l}{a} \left(r^a - r_d^a\right)$$

Dividing by r and integrating once more produces

$$\log f_d^* (r) - \log f_d^* (r_d) < -\frac{k_l}{a} \left(\frac{r^a - r_d^a}{a} - r_d^a \log \left(\frac{r}{r_d} \right) \right)$$

from which

$$f_d^* (r) < f_d^* (r_d) \ \exp \left(\frac{k_l}{a^2} r_d^a \right) \left(\frac{r}{r_d} \right)^{\frac{k_l}{a} r_d^a} \exp \left(-\frac{k_l}{a^2} r^a \right)$$

Substitute $u = (r/r_d)^a$ so that $\frac{dr}{du} = \frac{r_d}{a} \ u^{1/a-1}$ and for $u > \frac{r_0}{r_d}$ the density for $U_d := \left(\left| \mathbf{X}^{(d)} \right| / r_d \right)^a$ satisfies

$$f_d^{**} (u) < f_d^* (r_d) \ \frac{r_d}{a} \exp \left(\frac{k_l}{a^2} r_d^a \right) \ u^{1/a-1} u^{\frac{k_l}{a^2} r_d^a} \exp \left(-\frac{k_l}{a^2} r_d^a u \right)$$

To simplify notation define $b_d := r_d^a / a^2$. Then for $u > r_0 / r_d$

$$\begin{aligned} f_d^{**} (u) &< f_d^* (r_d) \frac{r_d}{a} \ \exp (k_l b_d) \, u^{k_l b_d + 1/a - 1} \exp (-k_l b_d u) \\ &= f_d^* (r_d) \frac{r_d}{a} \ \frac{\Gamma (k_l b_d + 1/a)}{(k_l b_d)^{(k_l b_d + 1/a)} \exp (-k_l b_d)} Gam (u; k_l b_d + 1/a, k_l b_d) \end{aligned}$$

where $Gam(u; a, b)$ is a Gamma density function with shape parameter a and rate parameter b.

Note that $b_d \to \infty$ as $d \to \infty$ but by Stirling's approximation, as $b_d \to \infty$

$$\frac{\Gamma (k_l b_d + 1/a)}{(k_l b_d + 1/a - 1)^{k_l b_d + 1/a - 1/2} \exp (-(k_l b_d + 1/a - 1))} \to (2\pi)^{1/2}$$

So that for $u > r_0/r_d$ (after some algebra)

$$(k_l b_d)^{1/2} \frac{\Gamma(k_l b_d + 1/a)}{(k_l b_d)^{(k_l b_d + 1/a)} \exp(-k_l b_d)} \to (2\pi)^{1/2}$$

Therefore given any $\epsilon > 0$ there is a $d_2 > d_1$ such that for all $d > d_2$

$$f_d^{**}(u) < (1+\epsilon)(2\pi)^{1/2} f_d^*(r_d) \ \frac{r_d}{a} (k_l b_d)^{-1/2} Gam(u; k_l b_d + 1/a, k_l b_d) \tag{C.5}$$

Similarly, considering the left hand inequality in (C.4) we may by ensuring d_2 is large enough force

$$f_d^{**}(u) > (1-\epsilon)(2\pi)^{1/2} f_d^*(r_d) \ \frac{r_d}{a} (k_u b_d)^{-1/2} Gam(u; k_u b_d + 1/a, k_u b_d) \tag{C.6}$$

for all $d > d_2$. Now both (C.5) and (C.6) are valid for $u > r_0/r_d$, so defining

$$c_d := (2\pi)^{1/2} f_d^*(r_d) \ \frac{r_d}{a} (k b_d)^{-1/2}$$

we obtain

$$\begin{aligned} \frac{1-\epsilon}{(1+\epsilon)^{1/2}} \ c_d \ Gam(u; k_u b_d + 1/a, k_u b_d) &< f_d^{**}(u) \\ &< \frac{1+\epsilon}{(1-\epsilon)^{1/2}} \ c_d \ Gam(u; k_l b_d + 1/a, k_l b_d) \end{aligned} \tag{C.7}$$

for $d > d_2$ and $u > r_0/r_d$. But as $d \to \infty$ the mean of a $Gam(k(1 \pm \epsilon) b_d + 1/a, k(1 \pm \epsilon) b_d)$ random variable approaches 1 and the variance approaches 0. So by Chebyshev's inequality the area outside of some region $(1-\delta, 1+\delta)$ tends to 0. Thus, since $r_0/r_d < 1$ we can choose a $d_3 > d_2$ such that for all $d > d_3$

$$1 - \epsilon < \int_{r_0/r_d}^{\infty} du \ Gam(u; k(1+\epsilon) b_d + 1/a, k(1+\epsilon) b_d) < 1 \tag{C.8}$$

But by Lemma 23, we may take d_3 large enough to ensure that for all $d > d_3$

$$1 - \epsilon < \int_{r_0}^{\infty} dr \ f_d^*(r) = \int_{r_0/r_d}^{\infty} du \ f_d^{**}(u) < 1 \tag{C.9}$$

Integrating (C.7) over its range of validity $(r_0/r_d, \infty)$, applying (C.8) and (C.9) and rearranging, we obtain for $d > d_3$

$$\frac{(1-\epsilon)^{3/2}}{1+\epsilon} < c_d < \frac{(1+\epsilon)^{1/2}}{(1-\epsilon)^2}$$

Substitute back into (C.7) to obtain for $d > d_3$ and $u > r_0/r_d$,

$$\frac{(1-\epsilon)^{5/2}}{(1+\epsilon)^{3/2}} Gam\left(u; k_u b_d + 1/a, k_u b_d\right) < f_d^{**}(u) < \frac{(1+\epsilon)^{3/2}}{(1-\epsilon)^{5/2}} Gam\left(u; k_l b_d + 1/a, k_l b_d\right) \tag{C.10}$$

The Gamma random variables in (C.10) both converge in probability to 1 as $d \to \infty$ so that for large enough d, the integral between 0 and r_0/r_d can be made smaller than ϵ. This taken together with Lemma 23 allows us to bound the moment generating function for U_d.

$$\begin{aligned}
\tilde{f}_d^{(u)}(t) &> \frac{(1-\epsilon)^{5/2}}{(1+\epsilon)^{3/2}} \left(\left(1 - \frac{t}{k(1+\epsilon)b_d}\right)^{-k(1+\epsilon)b_d+1/a} - \epsilon \exp\left(\frac{r_0 t}{r_d}\right) \right) \\
\tilde{f}_d^{(u)}(t) &< \frac{(1+\epsilon)^{3/2}}{(1-\epsilon)^{5/2}} \left(1 - \frac{t}{k(1-\epsilon)b_d}\right)^{-k(1-\epsilon)b_d+1/a} + \epsilon \exp\left(\frac{r_0 t}{r_d}\right)
\end{aligned}$$

So the moment generating function of $V_d := (U_d - 1)(kb_d)^{1/2}$ is bounded

$$\begin{aligned}
\tilde{f}_d^{(v)}(t) > \exp\left(-(kb_d)^{1/2} t\right) & \frac{(1-\epsilon)^{5/2}}{(1+\epsilon)^{3/2}} \\
& \times \left(\left(1 - \frac{t}{(1+\epsilon)(kb_d)^{1/2}}\right)^{-k(1+\epsilon)b_d+1/a} - \epsilon \exp\left(\frac{r_0 t}{r_d}\right) \right)
\end{aligned}$$

$$\begin{aligned}
\tilde{f}_d^{(v)}(t) < \exp\left(-(kb_d)^{1/2} t\right) & \\
& \times \left(\frac{(1+\epsilon)^{3/2}}{(1-\epsilon)^{5/2}} \left(1 - \frac{t}{(1-\epsilon)(kb_d)^{1/2}}\right)^{-k(1-\epsilon)b_d+1/a} + \epsilon \exp\left(\frac{r_0 t}{r_d}\right) \right)
\end{aligned}$$

But as $b_d \to \infty$, for positive or negative ϵ,

$$\exp\left(-(kb_d)^{1/2} t\right) \left(1 - \frac{t}{(1+\epsilon)(kb_d)^{1/2}}\right)^{-k(1+\epsilon)b_d+1/a} \to \exp\left(\frac{t^2}{2(1+\epsilon)}\right)$$

So that as $d \to \infty$

$$\tilde{f}_d^{(v)}(t) \to \exp\left(\frac{1}{2}t^2\right)$$

This is the moment generating function of a $N(0,1)$ random variable.

Proof of Lemma 15

We first prove convergence to zero for $\mathbb{E}\left[(U_d-1)^2 G_d(U_d)\right]$ and $\mathbb{E}\left[(U_d-1) G_d(U_d)\right]$.

From the convergence in mean square of U_d and the bounds on $G_d(\cdot)$

$$\mathbb{E}\left[(U_d-1)^2 G_d(U_d)\right] \leq \mathbb{E}\left[(U_d-1)^2\right] \to 0$$

Also convergence in mean square implies convergence in expectation so

$$\left|\mathbb{E}\left[(U_d-1) G_d(U_d)\right]\right| \leq \mathbb{E}\left[\left|(U_d-1)\right| G_d(U_d)\right] \leq \mathbb{E}\left[\left|(U_d-1)\right|\right] \to 0$$

But

$$\mathbb{E}\left[(U_d-1)^2 G_d(U_d)\right] = \mathbb{E}\left[U_d^2 G_d(U_d)\right] - 2\mathbb{E}\left[(U_d-1) G_d(U_d)\right] - \mathbb{E}\left[G_d(U_d)\right]$$

So that

$$\mathbb{E}\left[U_d^2 G_d(U_d)\right] - \mathbb{E}\left[G_d(U_d)\right] \to 0$$

and the result follows.

Proof of Lemma 16

(i) If $h(\cdot)$ has a finite number of local maxima and is continuously differentiable then it has a finite number of local minima. Consider some neighbourhood

$(x^* - a, x^* + a)$ which contains no local minima: in this interval $h(\cdot)$ is decreasing away from x^*.

Pick a small $\delta < a$ and define $\epsilon := \frac{1}{2}\min(h(x^*) - h(x^* - \delta), h(x^*) - h(x^* + \delta))$ so that for all $x \in (-x^* - a, x^* - \delta] \cup [x^* + \delta, x^* + a)$ we have $h(x^*) - h(x) \geq 2\epsilon$.

The sequence of $h_d(x)$ converges uniformly to $h(x)$ on the compact interval $[x^* - a, x^* + a]$, so we may pick d_1 such that for all $d > d_1$ and $x \in (x^* - a, x^* + a)$, $|h(x) - h_d(x)| < \epsilon$. Thus for all $d > d_1$ and $x \in (x^* - a, x^* - \delta] \cup [x^* + \delta, x^* + a)$

$$h_d(x) < h(x) + \epsilon \leq h(x^*) - \epsilon < h_d(x^*)$$

Therefore over $(x^* - a, x^* + a)$, $h_d(x)$ achieves its maximum somewhere in $(x^* - \delta, x^* + \delta)$. But δ can be made arbitrarily small.

(ii) Pick a large $k > 0$ and positive $\epsilon \leq \frac{1}{2}(h(2k) - h(k))$. Convergence of the sequence $h_d(x)$ is uniform for $x \in [0, 2k]$ and so we may pick d_1 such that for all $d > d_1$, and $x \in [0, 2k]$, $|h_d(x) - h(x)| < \epsilon$. So that for $x \in [0, k]$

$$h_d(x) < h(x) + \epsilon \leq h(k) + \epsilon \leq h(2k) - \epsilon < h_d(2k)$$

so $x_d^* > k$, but k may be arbitrarily large.

Bibliography

Asmussen, S. (2000). Matrix-analytic models and their analysis. *Scandinavian Journal of Statistics*, 27:193–226.

Baum, L. E., Petrie, T., Soules, G., and Weiss, N. (1970). A maximisation technique occuring in the statistical analysis of probababilstic functions of Markov chains. *The Annals of Mathematical Statistics*, 41:164–171.

Bedard, M. (2006a). Efficient sampling using Metropolis algorithms: applications of optimal scaling results. *in preparation*.

Bedard, M. (2006b). Optimal acceptance rates for Metropolis algorithms: moving beyond 0.234. *submitted for publication*.

Bedard, M. (2006c). Weak convergence of Metropolis algorithms for non-iid target distributions. *submitted for publication*.

Bernardo, J. M. and Smith, A. F. M. (1995). *Bayesian Theory*. Wiley, Chichester, UK.

Blackwell, P. G. (2003). Bayesian inference for Markov processes with diffusion and discrete components. *Biometrika*, 90:613–627.

Bladt, M. and Sorensen, M. (2005). Statistical inference for discretely observed Markov jump processes. *Journal of the Royal Statistical Socienty, Series B*, 67:395–410.

Breyer, L. A. and Roberts, G. O. (2000). From Metropolis to diffusions: Gibbs states and optimal scaling. *Stochastic Processes and their Applications*, 90:181–206.

Burzykowski, T., Szubiakowski, J., and Ryden, T. (2003). Analysis of photon count data from single-molecule fluorescence experiments. *Chemical Physics*, 288:291–307.

Celeux, G., Hurn, M., and Robert, C. P. (2000). Computational and inferential difficulties with mixture posterior distributions. *Journal of the American Statistical Association*, 95:957–970.

Chib, S. (1995). Marginal likelihood from the Gibbs output. *Journal of the American Statistical Association*, 90:1313–1321.

Davison, A. C. and Ramesh, N. I. (1996). Some models for discretised series of events. *Journal of the American Statistical Association*, 91:601–609.

Dellaportas, P. and Roberts, G. O. (2003). An introduction to MCMC. In Moller, J., editor, *Spatial Statistics and Computational Methods*, number 173 in Lecture Notes in Statistics, pages 1–41. Springer, Berlin.

Eyrie-Walker, A. and Hurst, L. D. (2001). The evolution of isochores. *Nature Reviews*, 2:549–555.

Fang, K. T., Kotz, S., and Ng, K. W. (1990). *Symmetric multivariate and related*

distributions, volume 36 of *Monographs on Statistics and Applied Probability*. Chapman and Hall Ltd., London.

Fearnhead, P. and Meligkotsidou, L. (2004). Exact filtering for partially-observed continuous-time Markov models. *Journal of the Royal Statistical Society, Series B*, 66(3):771–789.

Fearnhead, P. and Sherlock, C. (2006). An exact Gibbs sampler for the Markov modulated Poisson processes. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 68(5):767–784.

Fischer, W. and Meier-Hellstern, K. (1992). The Markov-modulated Poisson process (MMPP) cookbook. *Performance evaluation*, 18:149–171.

Galtier, N., Piganeau, G., Mouchiroud, D., and Duret, L. (2001). GC-content evolution in mammalian genomes: The biased gene conversion hypothesis. *Genetics*, 159:907–911.

Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Trans. Pattn. Anal. Mach. Intel.*, 6:721–724.

Geyer, C. J. (1992). Practical Markov chain Monte Carlo. *Statistical Science*, 7:473–483.

Gilks, W. R., Richardson, S., and Spiegelhalter, D. J. (1996). *Markov Chain Monte Carlo in practice*. Chapman and Hall, London, UK.

Grimmett, G. R. and Stirzaker, D. R. (2001). *Probability and random processes*. Oxford University Press, New York, third edition.

Gruss, A. and Michel, B. (2001). The replication-recombination connection: insights from genomics. *Current Opinion in Microbiology*, 4:595–601.

Hogg, R. V. and Craig, A. T. (1995). *Introduction to Mathematical Statistics*. Prentice Hall, Upper Saddle River, New Jersey 07458, fifth edition.

Karoui, M. E., Biaudet, V., Schbath, S., and Gruss, A. (1999). Characteristics of chi distribution on different bacterial genomes. *Res. Microbiol.*, 150:579–587.

Kong, A., Gudbjartsson, D. F., Sainz, J., Jonsdottir, G. M., Gudjonsson, S. A., Richardsson, B., Sigurdardottir, S., Barnard, J., Hallbeck, B., Masson, G., Shlien, A., Palsson, S. T., Frigge, M. L., Thorgeirsson, T. E., Gulcher, J. R., and Stefansson, K. (2002). A high-resolution recombination map of the human genome. *Nature Genetics*, 31:241–247.

Kou, S. C., Xie, X. S., and Liu, J. S. (2005). Bayesian analysis of single-molecule experimental data. *Appl. Statist.*, 54:1–28.

Li, W., Bernaola-Galvan, P., Haghighi, F., and Grosse, I. (2002). Applications of recursive segmentation to the analysis of DNA sequences. *Computers and Chemistry*, 26:491–510.

Louis, T. A. (1982). Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 44:226–233.

Marais, G. (2003). Biased gene conversion: implications for genome and sex evolution. *TRENDS in Genetics*, 19:330–338.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equations of state calculations by fast computing machine. *J. Chem. Phys.*, 21:1087–1091.

Meyn, S. P. and Tweedie, R. L. (1993). *Markov chains and stochastic stability*. Communications and Control Engineering Series. Springer-Verlag London Ltd., London.

Neal, P. and Roberts, G. (2006). Optimal scaling for partially updating MCMC algorithm. *Ann. Appl. Probab.*, 16:475–515.

Øksendal, B. (1998). *Stochastic differential equations*. Universitext. Springer-Verlag, Berlin, fifth edition. An introduction with applications.

Papaspiliopoulos, O., Roberts, G. O., and Sköld, M. (2003). Non-centered parameterizations for hierarchical models and data augmentation. In *Bayesian statistics, 7 (Tenerife, 2002)*, pages 307–326. Oxford Univ. Press, New York. With a discussion by Alan E. Gelfand, Ole F. Christensen and Darren J. Wilkinson, and a reply by the authors.

Roberts, G. O. (1998). Optimal metropolis algorithms for product measures on the vertices of a hypercube. *Stochastics and Stochastic Reports*, 62:275–283.

Roberts, G. O., Gelman, A., and Gilks, W. R. (1997). Weak convergence and optimal scaling of random walk Metropolis algorithms. *The Annals of Applied Probability*, 7:110–120.

Roberts, G. O. and Rosenthal, J. S. (2001). Optimal scaling for various Metropolis-Hastings algorithms. *Statistical Science*, 16:351–367.

Roberts, G. O. and Rosenthal, J. S. (2002). The polar slice sampler. *Stoch. Models*, 18(2):257–280.

Ryden, T. (1996). An EM algorithm for estimation in Markov-modulated Poisson processes. *Computational Statistics*, 21:431–447.

Sahu, S. K. and Roberts, G. O. (1999). On convergence of the EM algorithm and the Gibbs sampler. *Statistics and Computing*, 9:55–64.

Scott, S. L. (1999). Bayesian analysis of a two-state Markov modulated Poisson process. *Journal of Computational and Graphical Statistics*, 8:662–670.

Scott, S. L. and Smyth, P. (2003). The Markov modulated Poisson process and Markov Poisson cascade with applications to web traffic modelling. *Bayesian Statistics*, 7:1–10.

Stephens, M. (2000). Dealing with label-switching in mixture models. *Journal of the Royal Statistical Society, Series B*, 62:795–809.

Williams, D. (1991). *Probability with Martingales*. Cambridge University Press, Cambridge, UK.