Algorithms and Monte Carlo Methods in Computational Biology

A Dissertation

Presented in Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

with a

Major in Bioinformatics and Computational Biology

in the

College of Graduate Studies

University of Idaho

by

Yongtao Guan

May 2006

Major Professor: Stephen M. Krone, Ph.D.

Authorization to Submit Dissertation

This thesis of Yongtao Guan, submitted for the degree of Philosophy of Science with a major in Bioinformatics and Computational Biology and titled "Algorithms and Monte Carlo Methods in Computational Biology," has been reviewed in final form. Permission, as indicated by the signatures and dates given below, is now granted to submit final copies to the College of Graduate Studies for approval.

Major Professor _____ Date _____
Stephen M. Krone

Committee Members _____ Date _____
Fuchang Gao

_____ Date _____
Paul Joyce

_____ Date _____
Holly A. Wichman

Department Head _____ Date _____
Christopher J. Williams

College of Science Dean _____ Date _____
Judith Totman Parrish

Final Approval and Acceptance by the College of Graduate Studies

_____ Date _____
Margrit von Braun

Abstract

We study three different aspects of computational biology: interacting particle systems, phylogenetic distance methods, and Markov Chain Monte Carlo.

Our study of interacting particle systems has two components: First, a Windows-based simulator has been developed. It serves as a platform for many spatial biology studies; for example, plasmid transfer in biofilms. Second, a specific ecological model of multi-species cross-feeding has been studied using the simulator, where we have observed spatial self-organization and pattern formation (particularly, spiral waves). A mathematical derivation, which is based on analysis of partial differential equations, explains the phase-transitions observed in simulations.

Distance methods have been widely used to infer phylogenies (evolution trees among different species) because they are fast and reasonably accurate. In this study, we have proved that a popular distance method named neighbor-joining is in fact a special case of least squares methods. The proof relates neighbor-joining to least squares statistics. Hence it provides an explanation for the robustness and accuracy of neighbor-joining.

In an empirical study of MCMC, we investigate, through extensive simulations, how a mixture of local and heavy-tailed proposals (a.k.a., "small world proposals") can dramatically increase the convergence rates of Markov chains that admit stationary measures that are multi-modal. In the theoretical part of the study, we have proved, using techniques of state decomposition and isoperimetric inequalities for log-concave distributions, that the "small world proposals" turn a "slowly mixing" chain into a "rapidly mixing" chain.

To mom and dad, for everything...

Acknowledgments

I would like to thank my advisor and mentor, Professor Stephen M. Krone, for his skillful supervision, giving me an enormous amount of freedom to try many different things, allowing me to make my own mistakes, while steering me at the right moments to keep me on the right course.

Thanks also to my committee members, Professor Fuchang Gao, Professor Paul Joyce, and Professor Holly A. Wichman. I have constantly benefited from their insights in many areas in which I have been working. Many parts of this dissertation work contain their intellectual contributions, directly or indirectly.

I would like to thank members from research groups of Professor Krone, Professor Joyce, Professor Wichman, Professor Top, and Professor Forney, for many interactions, enlightening discussions, and companionship along the way. Thanks also to IBEST, the group that provided me a stimulating environment and much intellectual support during my Ph.D. study. My best wishes to this great group – IBEST BEGETS EARNEST STUDENTS.

Last but not least, I would like to thank my family in China, for their love and support as always.

Contents

List of Tables

List of Figures

Preface

Mathematics is biology's next microscope, only better;

biology is mathematics' next physics, only better.

– J. E. Cohen (2004)

This dissertation contains three parts. Each part is self-contained and independent of the others, although together they fit into a broad interdisciplinary framework: computational biology, an interface where biology meets mathematics.

The first part is about interacting particle systems and consists of two chapters. The first chapter is an announcement and description of a software package developed to simulate interacting particle systems. In the second chapter we use the simulator to study a multi-species crossing-feeding model of spatially distributed bacterial communities.

The second part is about phylogenetic distance methods. Particularly, in chapter three, we prove that the popular cluster method named neighbor-joining is a special case of least squares methods for inferring phylogenies using pair-wise distance matrices.

The third part is devoted to Markov chain Monte Carlo methods and it contains two chapters. In chapter four, we show, through extensive simulations, how a simple modification of a local proposal scheme can dramatically increase the performance of Metropolis-Hastings sampler in a multi-modal space. Such a modification is motivated by the notion of "small world networks" in that, by adding occasional heavy-tailed proposals onto the local ones, a regular state space transforms into a "small world network." Following the same spirit, but in a more mathematical vein, we show in chapter five that the small-world proposals turn a "slowly mixing chain" into a "rapidly mixing chain," therefore putting the simulation work in chapter four on a solid theoretical foundation.

Part I

Interacting Particle Systems

Chapter 1

WinSSS: Stochastic Spatial Simulator

1.1 Introduction

WinSSS (http://www.webpages.uidaho.edu/~krone) is a Windows-based program for simulating stochastic spatial models that are individual based, have discrete spatial structure and continuous time. This class of models is commonly referred to as interacting particle systems or asynchronously updated probabilistic cellular automata. It is ideally suited for developing insight and making predictions in spatial ecology. Ecological examples can be found in Dieckmann et al. (2000) and Durrett and Levin (1994).

WinSSS features an elaborate graphical interface Figure 1.1 that allows one to chose from various models, specify parameters such as birth/death rates and interaction strengths, initialize with various starting configurations, view spatial dynamics as well as time series and phase diagrams corresponding to spatial windows of various sizes. One can download freely at the above URL the ready-to-run simulator, which includes pre-programmed models and an HTML tutorial and help window. For those who would like to code their own models, the C++ code can also be obtained.

The models in WinSSS include mechanisms for invasion of new territory and competition for resources, head-to-head competition, pathogen spread, and various types of successional dynamics. The HTML tutorial gives a brief introduction to these spatially extended individual-based models and provides some references for further reading.

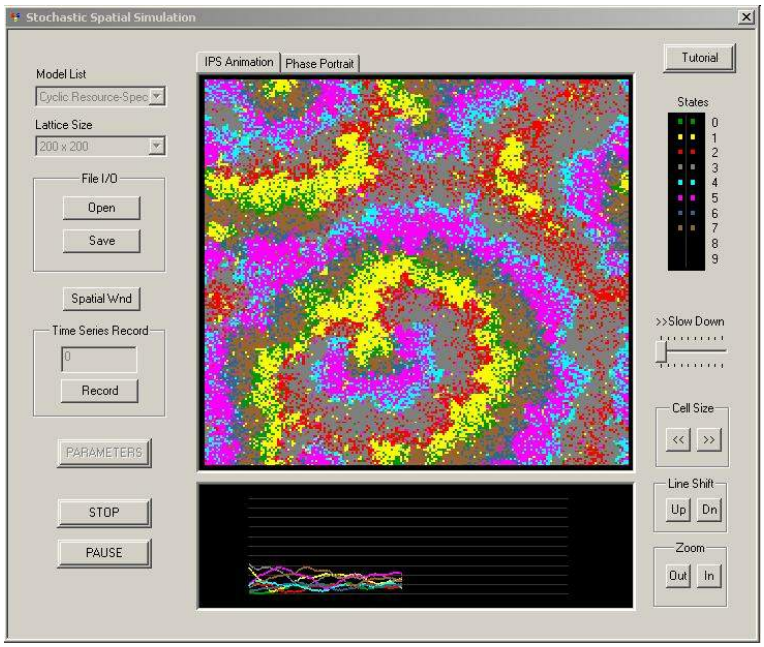

Figure 1.1: The graphical interface of WinSSS

1.2 Model specification and parameters

To describe the models and simulations, we begin by noting that all the action takes place on a 2-dimensional rectangular lattice (or grid) of sites, with a number of options for the lattice size. Each site in the lattice can be in a number of different states (represented as colors), depending on the specific model. One can think of a site as an individual or a group of individuals, say in a habitat patch. The state of a given site can change to other states at rates that depend in general on the configuration of states at "neighboring" sites. These changes occur in continuous time and very quickly, so when watching the simulation one typically observes sites changing all over the lattice. However, the changes are asynchronous due to the continuous time nature of this (Markov) process. The way to think of this is that every site has associated with it an (exponential) alarm clock with rate that depends on the state at that site and the states at neighboring sites. The site whose alarm rings first makes the appropriate change and all neighboring sites recalculate

their rates. All the alarm clocks then start over and we wait for the next one to ring. (We remark that the behavior of synchronously updated cellular automata can be similar in some respects but very different in others. For example, updating all the sites at once can lead to very rigid behavior that produces patterns not typically seen in biological populations.)

Rates and interaction neighborhoods: There are two basic types of rates that allow one to build most models of interest. These are "contact" and "spontaneous" rates. Contact rates are for events that depend on the types at neighboring sites. For example, a vacant site might become occupied by an offspring from a given species at a rate that is proportional to the number of that species currently within some distance of the vacant site. Contact rates can depend linearly or nonlinearly (e.g., a threshold event) on the states at neighboring sites. There are several options for neighborhood size in the simulator. Spontaneous rates are for events that occur independently of nearby sites. For example, an individual might die or change its life stage after some random time through no effect from other individuals.

Window size and time series: The overall lattice size can be selected from a number of options ranging from 100×100 up to 500×500. The densities of the different species appear, color-coded, in a separate window below the main simulation. These densities are averages over a spatial window that the user chooses. They can be recorded in an accessible file and used to obtain information about spatial length scales, as in Rand and Wilson (1995). The user can also choose to watch the phase plane trajectories corresponding to any two species. All of these observations of densities under various window sizes yield perspective on the effects of randomness, correlations between sites at various distances, and comparisons with the corresponding mass-action ordinary differential equations.

1.3 Implementation

The models in WinSSS were developed using Visual C++. The graphical interface employs OpenGL, the premier environment for developing portable, interactive $2D$ and $3D$ graphics applications.

To run WinSSS at reasonable speeds with lattice size 250×250 and above, a configuration that is better than Pentium III 866 with 256M RAM is recommended. WinSSS has been tested on Windows 2000 and Windows XP. Other operating systems in the Windows family (e.g., Windows 98 and Window NT) should also work, but we have not tested them.

Chapter 2

Spatial Self-Organization in a Cyclic Resource-Species Model

2.1 Introduction

Mathematical models and observations from macroscopic ecology tell us that spatial structure can have an enormous impact on the stability and diversity of interacting species. Bacteria typically live in structured environments such as biofilms and soils and, as in the case of biofilms, this structure is often largely the work of the bacteria themselves. In addition to this spatial self-organization, microbial communities often play an active role in helping to structure their geochemical environments.

Through their phenomenal diversity and adaptability, populations in microbial communities find ways to cooperate to efficiently consume nutrients. Pelz et al. (1999) and Bradshaw et al. (1994) give examples of such communities that, as biological units, are robust and able to extract the maximal energetic benefit from available nutrients. This cooperativity can work in concert with the afore-mentioned spatial organization.

We consider, as a representative of the class of spatially explicit cyclic resource-species dynamics to be treated here, a microbial scenario in which a number of different bacterial species coexist by cooperatively degrading a single nutrient in a succession of steps and providing catalytic support for the primary degrader. This "catalytic support," which can be anything that facilitates the growth of the primary degrader, introduces feedback into the system and provides a mechanism for spatial self-organization that can give rise

to coexistence of the species. Ours is a stochastic spatial model that is individual based with discrete spatial structure and continuous time. Such models fall under the heading of interacting particle systems or (asynchronously updated) probabilistic cellular automata.

Several factors make it a challenge to observe and understand the complex dynamics in microbial communities, and this is where mathematical models can lead to significant advances. Pelz et al. (1999) point out that, while metabolic pathways in individual microorganisms have been well studied since 1950, very little is known about multi-species metabolic networks in bacterial communities, especially those found in nature. In addition, the spatio-temporal dynamics of microbial communities are notoriously difficult to observe in any detail. This is partly due to the fact that most sampling protocols are invasive and destructive. In the process of trying to extract crude estimates of the diversity of a sample, the microscopic spatial structure is typically broken down (by homogenizing the sample) and one obtains instead only a spatially averaged view, often at a single time point. Theoretical advances can play an important role in determining the types of species interactions that lead to stable coexistence in a dynamic spatial setting, and they can reveal signatures of such interactions that make them more easily recognizable. For example, microscopic mechanisms sometimes lead to macroscopic patterns. While the latter do not uniquely determine the former, they often suggest hypotheses that can be tested in the lab.

The complex spatial nonlinear dynamics described by our model lead to interesting properties. We are particularly concerned with conditions leading to persistence of the species and the emergence of macroscopic spatio-temporal patterns from the microscopic dynamics. Our simulations imply, for example, that for certain parameter settings the species cannot coexist in the absence of spatial patterning. In these instances, coexistence is enhanced by the self-organization of species (and their resources). The fact that the resources can become patterned in such a process suggests potential mechanisms for detecting spatial structure in microbial communities.

2.2 The model

To motivate the model, let us imagine a collection of bacterial species that take part in the successive degradation of a single nutrient. Assume that these bacterial players live on a surface, say at a rock-water interface. The compound to be degraded adsorbs onto the surface from the bulk fluid. One species (the primary degrader) is able to break down this initial compound, taking in nutrients in the process. The product left over is then further broken down by a second species, leaving a second product, and so on. This sequence of steps is known as cross-feeding (e.g., Rosenzweig et al. 1994; Doebeli 2002). The last species in the sequence we are modeling, after taking up its share of the nutrient, then provides some form of catalytic support that enhances the growth of the first species. This last step in the dynamics causes feedback and thus results in a cyclic progression of states at each individual site. Such feedback mechanisms can arise in many ways. The last species, for example, could remove a toxin produced by any one of the previous species that would otherwise kill or inhibit the growth of the first species. Pelz et al. (1999) discuss such a situation; they study a microbial consortium with four primary species that are involved in carbon sharing and with one of the species scavenging toxic metabolites that would kill the primary degrader if they were allowed to accumulate. While their system does not constitute an exact match to our model, it gives a nice example of the type of feedback we are modeling. As an alternative mechanism producing feedback, the last species might simply create conditions under which the primary nutrient can adsorb from the bulk fluid.

The above scenario is meant to provide some intuition for the model which follows. There are many plausible systems that will provide the same dynamics. Our goal is to understand how such spatially explicit cyclic systems behave and to give some indication of the stability properties that make them likely to arise in the first place.

The stochastic spatial model introduced here is constructed on a 2-dimensional grid (the square lattice $\mathbb{Z}^2$ or a finite sublattice), with periodic boundary conditions imposed

when the space is finite. Each site of the lattice can be in any one of the following states:

$$R_1, S_1, R_2, S_2, \ldots, R_L, S_L,$$

where L denotes the number of species, S_i is the ith species, and R_i is the ith resource (or stage in the decomposition of the initial compound). If a site is occupied by an individual from species S_i, this individual will reproduce at rate β_i and deposit its (single) offspring onto a randomly chosen neighboring site; only if the offspring lands on a site in state R_i will it be viable. In such an event, the neighboring site will harbor the ith species on the ith resource; this, however, will be denoted by S_i instead of the bulkier $R_i + S_i$ (and we will simply say that the site is occupied by the ith species, greatly simplifying our explanations later). Thus, species S_i consumes or degrades its "required resource" R_i to produce its "product resource" R_{i+1}, the subscript being understood cyclically (i.e., $R_{L+1} = R_1$). To keep the model from having too many states, we assume that a species will disappear from a site once it has converted its required resource to its product resource. (This could be due to the removal of the species as a result of starvation or being out-competed by the next species.) Thus, if a site is in state S_i, it will change to state R_{i+1} at constant rate δ_i.

In summary, a given site proceeds through a cyclic progression of states

$$R_1 \to S_1 \to R_2 \to S_2 \to \cdots \to R_L \to S_L \to R_1 \to \cdots .$$

The transitions at such a site x come in two flavors:

$$R_i \to S_i \text{ at rate } \beta_i f_i(x), \tag{2.1}$$

and

$$S_i \to R_{i+1} \text{ at rate } \delta_i, \tag{2.2}$$

where $f_i(x)$ denotes the fraction of species S_i in some neighborhood about the site x. We

will refer to β_i and δ_i as the growth and death rates for species S_i.

The above transitions and the cyclic nature of the progression of states at a given site are illustrated in Figure 2.1. The sizes of the neighborhoods are important and will be discussed below. We say the transition in (2.1) is *"by contact"* since it depends on the states at nearby sites, requiring the presence of at least one S_i near site x. The transition in (2.2) is "*spontaneous*" in that it happens automatically, independently of the states at the surrounding sites. We remark that since ours is a stochastic model, the word "rate" always refers to exponential waiting times; i.e., some event happens at rate λ if the amount of time it takes is an exponential random variable with rate λ (mean $1/\lambda$). Figure 2.1 illustrates the transition rates *at a given site*. This is an example of what we call cyclic local dynamics, our model being additionally characterized by the alternating of contact and spontaneous transitions. Note that the model reduces to the basic contact process (cf., Durrett 2002) in the case of one species and one resource ($L = 1$).

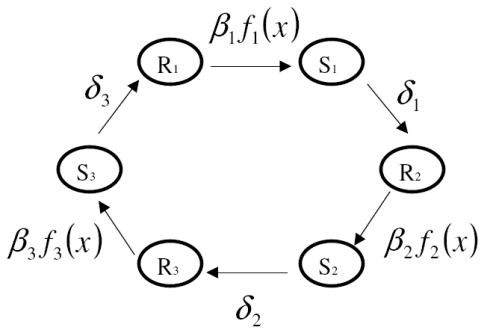

Figure 2.1: The cyclic local dynamics at site x for the model with 3 species and 3 resources.

To complete the description of the contact transition rates above, we must specify the neighborhoods used to calculate the fraction of neighboring sites in a given state. In our simulations, we treated each of the following three neighborhood types for a site x:

- *4 nearest neighbors*: the four adjacent sites north, south, east, and west of the site x;

- *8 nearest neighbors*: the sites in the 3×3 box centered at x (except for x) made up of the four sites above plus the four diagonally adjacent sites;
- *24 neighbors*: the sites in the 5×5 box centered at x (except for x).

Otherwise put, the 4 nearest neighbors are the sites within L^1 distance one of x (i.e., the von Neumann neighborhood of x), the 8 nearest neighbors are the sites within L^∞ distance one of x (i.e., the Moore neighborhood of x), and 24 neighbors correspond to the sites within L^∞ distance two of x.

Now that the local dynamics are specified, the goal is to understand how these translate into global behavior by keeping track of the entire configuration of states on the spatial grid. Such stochastic spatial models are sometimes called *interacting particle systems* (IPS) or (asynchronously-updated) probabilistic *cellular automata* (CA). While the latter term is more familiar to biologists, some confusion can occur if one fails to distinguish between the different types of CA model. Hereafter, we will use the term IPS model.

2.3 Particle system behavior

In this section, we first characterize the different patterns and behaviors of our model. Then we investigate the corresponding parameter regions that yield these different behaviors through extensive simulations of the particle system.

The model exhibits configurations that can be classified by four major patterns: frozen, synchronous waves, asynchronous waves, and mixing. These patterns are shown in Figure 2.2 and are understood to refer to patterns in a simulation that persist for a long time. A *frozen* pattern results when all the species die out, leaving only resources; once all species are absent, the configuration can no longer change. The other three types of patterns involve survival of all species. *Synchronous waves* include spiral waves and traveling waves that are spatially large (compared to the grid size) and temporally

synchronized. *Asynchronous waves* consist of spiral waves and traveling waves that are not large (compared to the grid size) and temporally unsynchronized. *Mixing* refers to spatially homogeneous configurations and corresponds to species survival in the absence of spatial patterning. We remark that the classification of a given simulation run as one of the four patterns was done "by eye." Also, the difference between synchronous and asynchronous waves is one of degree, with asynchronous waves amounting to synchronous waves that exist only on small length scales.

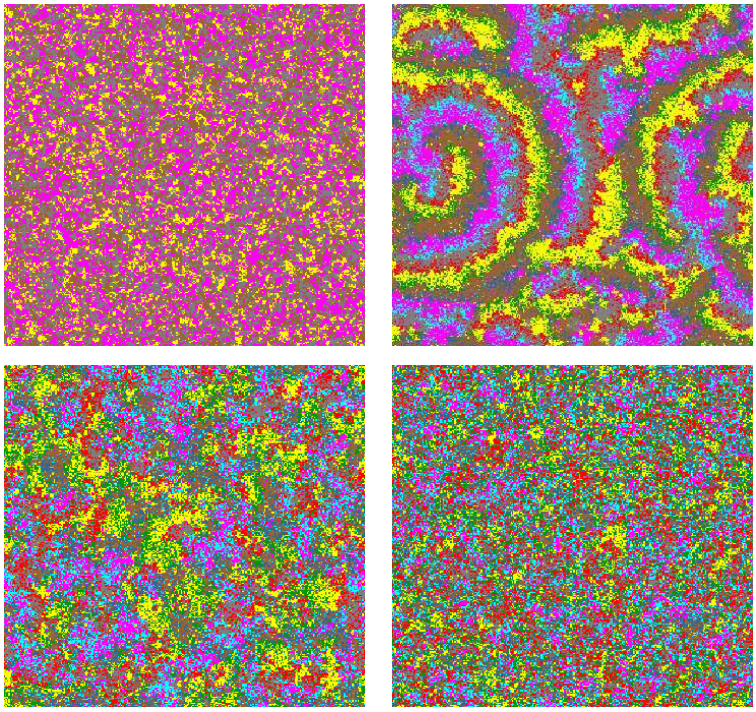

Figure 2.2: Four different patterns for an 8-state version of the model. Top left: frozen; Top right: synchronous waves; Bottom left: asynchronous waves; Bottom right: mixing. The colors designate species (green, red, light blue, blue) and resources (yellow, gray, purple, brown). For example, yellow represents the product resource of the green species and the required resource of the red species.

All simulations were run on a lattice of size 300×300. Each run starts with a random initial configuration and data are collected after 1000 units of time, where one "unit of time" corresponds to an average of $90,000$ (i.e., $300 \cdot 300$) possible random events: on

average, each site updates once during a unit of time. To fully investigate the parameter space for this model is difficult, if not impossible, because of the large number of parameters ($2L$ different parameters when there are L species). However, we can obtain important information by investigating *two special cases*. One is the *symmetric case* in which all species growth rates are equal and all death rates equal 1. The other is an *asymmetric case* in which we allow the growth rate of the first species to differ from the (majority) common growth rate of all other species. These cases are discussed in more detail below.

The afore-mentioned patterns should not be confused with short-lived transient patterns such as traveling waves of infection in a contact process or epidemic model. Once such a wave passes, the system relaxes to a steady state distribution that does not accommodate these waves. In our cyclic resource-species model, the patterns we refer to are recurring. Of course, any frozen state is absorbing (in the sense that it can never change) and as soon as one species dies the whole system is doomed to freeze. Since we are using a finite lattice, eventually the process will freeze, but this will take a very long time on our 300×300 lattice–much longer than the 1000 units of time on which our simulation results are based. The patterns we are referring to appear in simulations over the course of a reasonable length of time (from a few seconds to many hours). Thus, strictly speaking, they refer to quasi-stationary distributions for the particle system on a finite lattice. If one prefers to think of the particle system on an infinite lattice, these quasi-stationary states look like pieces of the stationary distribution for the infinite system.

2.3.1 Symmetric case

The results of the simulations in the symmetric case are summarized in Figure 2.3. We treated four different values for the number of states: $N = 2L = 4, 6, 8, 10$. For each of these, we considered three neighborhood sizes: $r = 4, 8, 24$. Once these two parameters were chosen, we ran simulations for a large range of values of the common growth rate β for 1000 units of time.

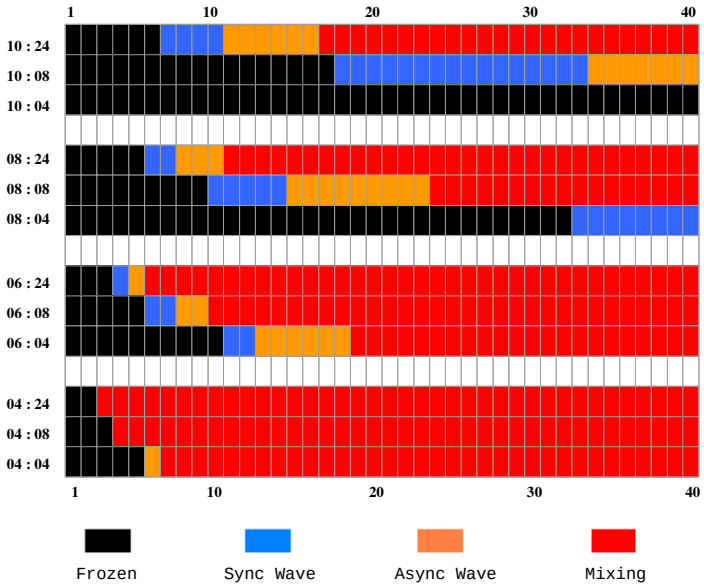

Figure 2.3: (Symmetric Case) Simulation results for IPS model with all species having the same growth rate β, and all death rates $\delta = 1$. The horizontal axis denotes the common growth rate β (sampled at integer values) and the vertical axis has labels of the form $N : r$, where $N = 2L$ denotes the total number of states and r denotes the neighborhood size. The outcome reported for a given values of β and $N : r$ corresponds to the majority outcome from 20 runs starting with different random configurations.

Growth rate. First consider the simulation results as a function of β, for fixed values of N and r. Small values of β lead to frozen patterns. As β increases, the patterns transform first to synchronous waves (in certain cases), then asynchronous waves, and, for large β, mixing.

Neighborhood size. To see the effects of neighborhood size, compare the three bands corresponding to each fixed value of N. First of all, larger neighborhood size leads to species survival (and hence coexistence) for a wider range of β values. This is a simple consequence of the fact that a larger range provides a given species a better chance to find its required resource. Also, larger values of r create more mixing of different types; this leads to less spatial structure in the sense that the wave clusters are smaller when r is larger. We will see below that this reduction in the amount of spatial structure can happen for other reasons, as well.

Number of states. The most striking observation about the effect of the number of states is that for $N = 4$ (i.e., 2 species) there are no synchronous waves and hence no spirals. In fact, the same is also true when $N = 2$ since, in that case, our model reduces to the single species contact process. Synchronous wave patterns are possible for all values of $N > 4$, provided β and r are suitably chosen. Similar observations have been made regarding catalytic hypercycle models where it is known that spiral patterns arise if and only if the number of states in the cycle is greater than or equal to 5; see (Dieckmann et al. 2000). Durrett and Griffeath (1993) also observed spiral waves for a suitably large number of states in two different cyclic spatial models. One is the Greenberg-Hastings model (one contact transition and all others spontaneous), and the other can be thought of as a generalized rock-scissors-paper model (all transitions are by contact); in each case, they use threshold updating for the contact transitions–a feature that encourages spiral formation. Bramson and Griffeath (1989) discovered another phase transition involving the number of states for a one-dimensional version of the generalized rock-scissors-paper model; they showed that the system freezes when the number of states exceeds four, and

otherwise keeps changing. Other interesting aspects relating to spiral formation in cyclic spatial systems can be found in the references in the above papers. Of course, for our model, the number of states is always even. In the next section, we will give a simple explanation for our $N > 4$ threshold based on a linearization argument for a related system of reaction-diffusion equations. As the number of states increases, one needs larger values of β to generate synchronous waves and to survive in general, especially when the neighborhood size is not large.

Initial configurations. For the above discussion, it was assumed that the initial configuration was random, with each site independently assigned one of the N states and probability $1/N$ per state. Thus any patterns that formed, such as spirals, were the result of self-organization. In some cases, parameters that led to species extinction, and hence a frozen pattern, could produce species persistence and spiral waves under appropriate initial configurations. This happens for β's in the frozen region that are not too far from the synchronous wave region. For such β, once enough spatial structure is present it can be maintained. There are several ways to get such "hard-to-start spirals." If we start with a random initial configuration, the species all die out, but before doing so the remaining resources will have become somewhat clustered; if we then "seed" the frozen configuration by randomly sprinkling in a small number of each species to the system, spiral waves will form. Such behavior has been observed in other cyclic systems (Comins et al. 1992). Often, only a very small amount of this pre-structuring is needed to get the spiral waves started. Another way to generate spirals from parameters that lead to frozen patterns under random initial configurations is to "train" the process, starting with a larger neighborhood size until synchronous waves start to form, and then decreasing the neighborhood size in the simulation. If synchronous waves have begun to form under the larger neighborhood size, then (with such a configuration as the initial state of the process) a reduction in neighborhood size causes to waves to become more focused and less fuzzy. It is easy to understand how, for certain parameters, only initial configurations with some

structure would lead to persistence and patterns. For example, in the model with 8 states and neighborhoods of size 4, a random configuration has low probability of presenting a given species with its required resource. This results in the various species using up the random bits of resource available to them and then dying out before they can reproduce again due to the lack of required nutrient in their neighborhoods.

The synchronous waves come in two forms: (segments of) circular traveling waves and spiral waves. Along a reactive wavefront (i.e., a species invading its required resource), we typically have a high density of the required resource. This allows a traveling wave that is locally similar to an infectious wave in an epidemic model. In the wake of this reactive wavefront, the invading species leaves behind its product resource, thus laying the foundation for the next species. The shape theorem of interacting particle systems (Durrett 2002) provides the essential ingredients for the dynamics of a traveling reactive wavefront started by a small localized cluster of a given species set in a uniform environment that has all other sites holding the required resource for that species. Such a wavefront is roughly circular in shape with the radius expanding at what is effectively a constant speed. Of course, our model is much more complicated than this. We have multiple species and resources, leaving an inhomogeneous mix of states in which a wavefront must spread, if at all. Instead of finding itself in a uniform field of the required resource, a given species will typically encounter sites that are holding other resources or other species as well as clusters of its required resource. The result is that the species is trying to spread in a fragmented environment. In the language of interacting particle systems, a species must spread on a (space-time) percolation cluster of its required resource. The more fragmented and sparse this percolation cluster is, the less likely it becomes that species members can reproduce before dying (i.e., before the site changes from the species state to the product resource). The only way the species can persist is to spread, and the only way it can spread in such a fragmented environment is to have a growth rate that is sufficiently large compared to its death rate. For example, a contact process in a random environment (that has some

sites inaccessible to the species) has a higher required growth rate than a contact process that has no such restrictions.

For our process, the cyclic nature of the local dynamics makes this effect of growth rate a double-edged sword. Large growth rates enable a species to quickly gobble up bits of its required resource almost as soon as they appear. This, however, leaves behind a fragmented landscape for the next species and leads to a breaking up the waves. This is why simulations with large β's lead to mixing patterns and not synchronous waves.

With this insight, we can now recap the simulation results that appear in Figure 2.3. If β is too small, the species members cannot spread fast enough and they die out, leaving behind a frozen state consisting only of resources. If we increase β far enough, the species will be able to persist and coexist by synchronizing their spread with other nearby members of the same species and leaving behind clustered regions of the resource required by the next species. In some cases, this structure is self-organized, even from a random initial configuration. For slightly smaller β, this synchronization may only be possible if the process starts with some structure. As we increase β more and more, the length scales on which the waves operate get smaller and smaller as wavefronts consistently encounter fragmented resources that cause them to split, thus removing the potential for spiral wave development. Asynchronous waves can be thought of as synchronous waves whose length scales are very small. As β increases still further, the species have no trouble persisting. Not only do they not require spatio-temporal synchronization to persist, but the high growth rates preclude such patterning and the result is a mixing configuration.

The dynamics of spiral waves have received much attention in the literature on reaction-diffusion equations and, to a lesser extent, in the literature on stochastic spatial models; cf. Grindrod (1996), Murray (1989), Pálsson and Cox (1996), Savill et al. (1997), and Dieckmann et al. (2000). We limit ourselves to a few comments on the spiral waves observed here. Each spiral has $N = 2L$ arms, appropriately ordered and forming a confluence at the eye of the spiral. A collision between two spirals occurs when two waves of the

same species approach each other from different directions, consuming the same resource. Such collisions often break off pieces of the spiral, but as long as the eye is intact the spiral can persist. The spiral waves in the IPS model are much more irregular than those that arise in reaction-diffusion equations and patch models; see next section.

2.3.2 Asymmetric case

Many of the above comments apply to the asymmetric case. The results of the asymmetric simulations are summarized in Figure 2.4. We see that the patterns are strongly influenced by the majority growth rate. However, an increase in the growth rate of species 1 can destroy or weaken the spatial structure due to fragmentation of the resource for the second species, much like what was explained about large β for the symmetric case. We remark that the diagonal region in this figure corresponds to $\beta_1 = \beta$ and hence to the $8 : 8$ symmetric case treated above. Notice that the results do not coincide exactly. This reflects two things. Firstly, the stochastic nature of the simulations (these are from separate runs) allows for different behaviors on individual runs, especially near the boundary regions. Secondly, as was mentioned earlier, the classification of a set of runs into a consensus state of frozen, synchronous waves, asynchronous waves, or mixed, was done "by eye" and states like asynchronous waves and mixed are hard to distinguish for certain parameter values. This can also be seen in the differential equation connection below in regions for which eigenvalues have positive real parts very close to zero.

2.4 Reaction-diffusion equations and linearization

Consider the following reaction-diffusion equation (RDE), where u_i denotes the density of resource R_i, v_i the density of species S_i, and Δ is the Laplacian.

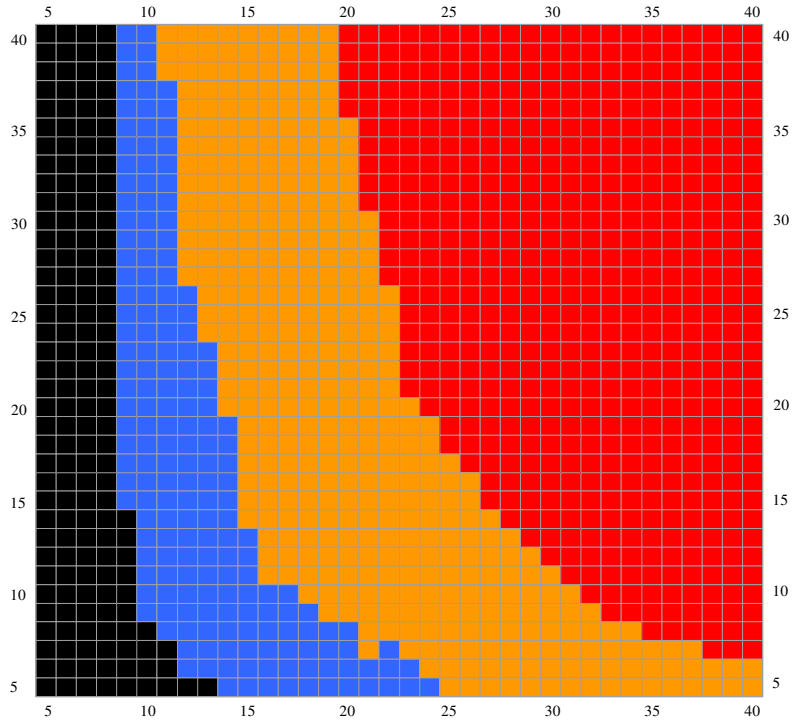

Figure 2.4: (Asymmetric Case) Simulation results for IPS model with all species having the same death rate $\delta = 1$, but the growth rate of species 1 is (possibly) different from the common growth rate of species $2, \ldots, L$. The horizontal axis denotes the majority growth rate β and the vertical axis denotes the growth rate β_1 of the first species. The number of states is $2L = 8$ and the neighborhood size is also 8. Each small square in the figure corresponds to the majority outcome from 10 runs starting with different random configurations and the growth rates assuming integer values. The color scheme is the same as in Figure 2.3.

$$\begin{cases} \dfrac{\partial u_1}{\partial t} = \Delta u_1 + \delta_L v_L - \beta_1 u_1 v_1, & \dfrac{\partial v_1}{\partial t} = \Delta v_1 - \delta_1 v_1 + \beta_1 u_1 v_1, \\ \dfrac{\partial u_2}{\partial t} = \Delta u_2 + \delta_1 v_1 - \beta_2 u_2 v_2, & \dfrac{\partial v_2}{\partial t} = \Delta v_2 - \delta_2 v_2 + \beta_2 u_2 v_2, \\ \quad \vdots & \quad \vdots \\ \dfrac{\partial u_L}{\partial t} = \Delta u_L + \delta_{L-1} v_{L-1} - \beta_L u_L v_L, & \dfrac{\partial v_L}{\partial t} = \Delta v_L - \delta_L v_L + \beta_L u_L v_L. \end{cases} \tag{2.3}$$

This is a partial differential equation that is related to the particle system in a natural way via a 'fast-stirring' limit (Durrett and Neuhauser 1994; Krone 2004). Roughly speaking, this says that if, in our original IPS model, between state changes at different sites we were to allow the 'particles' to jiggle around a bit by exchanging the states at some nearby sites, then the above reaction-diffusion equation (with the same rates as in the IPS) would serve as a good approximation to the large-scale spatial dynamics of the particle system. By this last part, we mean that the density at a spatial point for the solution to the RDE is approximately equal to the average density for a suitably chosen group of sites in some spatial window in the particle system. The above references can be consulted for a more thorough discussion on the relation between the IPS model and the RDE. For the purposes of the present chapter, it is enough to realize that the RDE can be used to provide information about the IPS.

In Figure 2.5 we see the behavior of the RDE (2.3) in the case of $N = 6$ states when all growth rates are the same, for two different values of the common growth rate. We solved the RDE numerically, starting with a random initial condition (in the numerical scheme) and periodic boundary. The initial condition introduces small inhomogeneities that are able to grow for certain parameters. When we set $\beta = 3.8$, spiraling and traveling waves emerge and persist. When $\beta = 5.8$, the patterns die out to a spatially homogeneous steady state. Figure 2.5 portrays the density of one of the resources; the bands of resource are wider than those of the species, so they are easier to see. Because of the cyclic nature

of the model and the symmetry in the parameters, it doesn't really matter which density is plotted.

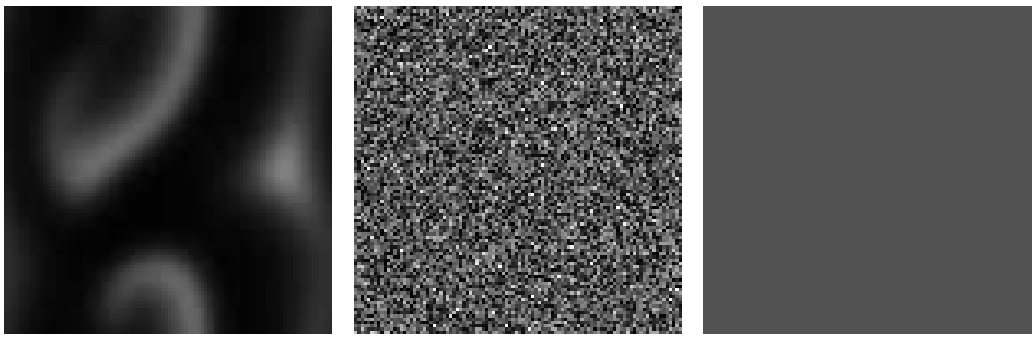

Figure 2.5: The density of resource R_1, with higher density corresponding to lighter gray level, in the numerical solution of the RDE (2.3). Here, we have $L = 3$ with all species having the same growth rate β; see symmetric case 2 below. The middle figure shows the random initial configuration. In the left figure, $\beta = 3.8$ and spiral waves develop. In the right figure, $\beta = 5.8$ and all spatial structure dies out to yield a homogeneous distribution.

Setting $X = (u_1, v_1, \ldots, u_L, v_L)$, we can write the RDE in vector form:

$$\frac{\partial X}{\partial t} = \Delta X + f(X) \tag{2.4}$$

The embedded ODE for this reaction-diffusion equation, sometimes referred to as the mean-field ODE for the IPS model, is

$$\frac{dX}{dt} = f(X). \tag{2.5}$$

We next attempt to use the RDE to predict results for the IPS simulations. We first use a simple linearization argument to show that eigenvalues of the linear system govern the behavior of the reaction-diffusion equation. We then apply this linearization argument to study the two special cases treated in the particle system simulations: symmetric and asymmetric. The study of these two mathematically tractable cases not only gives us some insight into the IPS model, but also enables us to compare the results of the differential equation and particle system models.

2.4.1 Linearization

Suppose the system (2.5) has a unique interior equilibrium $X^* = (u_1^*, v_1^*, \ldots, u_L^*, v_L^*)$; i.e., $f(X^*) = 0$ and $u_i^* > 0, v_i^* > 0,\ i = 1, \ldots, L$. We linearize the reaction term of (2.4) about the spatially homogeneous solution X^* of the RDE to get:

$$\frac{\partial Y}{\partial t} = \Delta Y + J\, Y \tag{2.6}$$

where Y is a small perturbation and $J = J_f(X^*)$ is the Jacobian of f at the equilibrium X^*. Equation (2.6) has a solution of form

$$Y(x,t) = \mathbf{c}_1 e^{i\sqrt{J}\, x} + \mathbf{c}_2 e^{J\, t}, \tag{2.7}$$

where $\mathbf{c}_1$ and $\mathbf{c}_2$ are vectors that depend on the initial condition. This suggests that the behavior of the reaction-diffusion equation (2.3) might be controlled, at least for small perturbations of the interior equilibrium, by the Jacobian for the embedded ODE (2.5). (For readers unfamiliar with taking functions of matrices, an exponential function of a matrix is obtained by writing the corresponding series expansion (Hirsch and Smale 1974); the square root of a matrix involves its spectral decomposition and is constructed from a diagonal matrix made up of square roots of the eigenvalues and two other matrices that involve left and right eigenvectors of the original matrix (Hoffman and Kunze 1971).) Rather than concerning ourselves with making the above equation rigorous, we simply take it as a heuristic for the method that we now propose. (But see Section 2.4.4 below.) We claim that the types of behavior in the IPS model seen in Figure 2.3 can be predicted by considering the eigenvalues of the mean-field ODE (2.5) at the interior equilibrium. In particular, we seek to explain why $N > 4$ is required to get spiral waves. Our approach will be to show that, up to a scaling of the parameters, the IPS simulations agree with the predictions based on the RDE. One should not expect a β-for-β agreement between these two models when the particle system has finite interaction range; the fast-stirring

limit RDE (which does have the same parameters as the original IPS model) only holds in a limiting sense. For example, in the one-dimensional contact process the critical growth rate is about 1.65, while the critical growth rate for the corresponding RDE is 1. What is important is that the two model types have the same qualitative behavior.

When the Jacobian matrix J has no eigenvalues with positive real part, the spatial inhomogeneities in the initial configuration do not grow. However, when the matrix J has eigenvalues with positive real part, certain wave solutions initiated by the spatial inhomogeneities can be amplified and propagate through space. Below, we investigate several representative cases that illustrate the connection between the linearization and the IPS results.

2.4.2 Symmetric case studies

As in the IPS model, the symmetric case here refers to parameters $\delta_1 = \cdots = \delta_L = 1$ and $\beta_1 = \cdots = \beta_L = \beta$. It is easy to see that the system (2.5) has only one interior equilibrium point X^* with components $u_1 = \cdots = u_L = \frac{1}{\beta}$ and $v_1 = \cdots = v_L = \frac{1}{L} - \frac{1}{\beta}$. To have all species surviving at equilibrium, and hence all densities u_i and v_i positive, we require $\beta > L$. To see how number of species affects the behavior of the model, we consider two different values of L.

Case 1 ($L = 2$): In this case, the total number of states in the system is 4 and

$$J_f(X^*) = \begin{pmatrix} 1 - \frac{\beta}{2} & -1 & 0 & 1 \\ -1 + \frac{\beta}{2} & 0 & 0 & 0 \\ 0 & 1 & 1 - \frac{\beta}{2} & -1 \\ 0 & 0 & -1 + \frac{\beta}{2} & 0 \end{pmatrix}. \tag{2.8}$$

Since $L = 2$, $\beta > 2$ is required for the existence of an interior equilibrium. The real parts of all the eigenvalues of the matrix (2.8) with respect to different β are plotted in Figure 2.6.

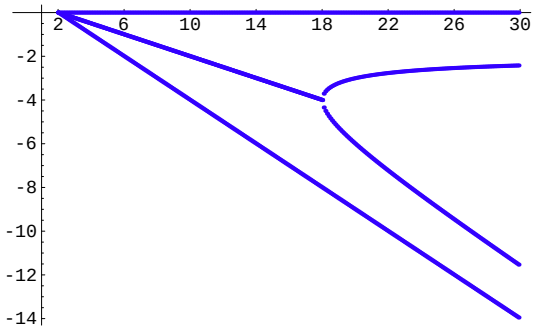

Figure 2.6: The real parts of all eigenvalues of the matrix (2.8) as a function of $\beta > 2$.

There is a zero eigenvalue for any β since the cyclic dynamics makes the matrix (2.8) singular. There is also a negative real eigenvalue and two complex conjugate eigenvalues with negative real parts when $\beta < 18$; for $\beta \geq 18$, all eigenvalues are real and the three that are nonzero are negative. The lack of positive real part in the eigenvalues implies that the system (2.5) is stable near the equilibrium X^* for any $\beta > 2$. The dissipative nature of the system explains why there are no synchronized waves in the corresponding particle system, as reported in Figure 2.3.

Case 2 ($L = 3$): In this case, the total number of states in the system is 6 and

$$J_f(X^*) = \begin{pmatrix} 1-\frac{\beta}{3} & -1 & 0 & 0 & 0 & 1 \\ -1+\frac{\beta}{3} & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1-\frac{\beta}{3} & -1 & 0 & 0 \\ 0 & 0 & -1+\frac{\beta}{3} & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1-\frac{\beta}{3} & -1 \\ 0 & 0 & 0 & 0 & -1+\frac{\beta}{3} & 0 \end{pmatrix}. \tag{2.9}$$

Notice that in this case, $\beta > 3$ is required to get positive equilibrium densities. The real parts of all the eigenvalues of the matrix (2.9), as functions of $\beta > 3$, are plotted in Figure 2.7.

Compared to the 4-state case, a major difference in the 6-state case is that for a

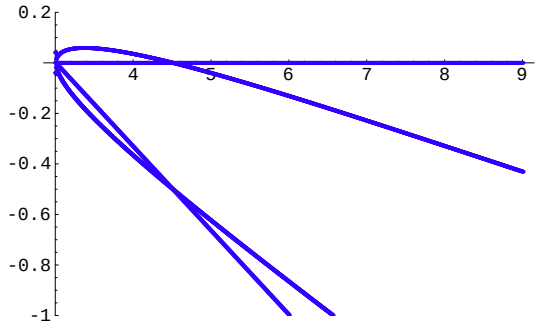

Figure 2.7: The real parts of all eigenvalues of the matrix (2.9) as a function of $\beta > 3$.

certain parameter range ($3 < \beta < 4.5$) there are eigenvalues with real part larger than 0. Therefore, within this parameter range, the system is no longer stable at the interior equilibrium and trajectories approach what is approximately a heteroclinic orbit, as shown in the Figure 2.8(a). When $\beta > 4.5$, the interior equilibrium is stable; Figure 2.8(b) shows a spiraling in trajectory when $\beta = 4.8$. Interestingly, $\beta = 4.5$ is a bifurcation point where we observe a limit cycle, as shown in Figure 2.8(c). This, again, agrees with the behavior of the IPS model (up to a scaling of the parameters) described in Figure 2.3. Notice, in particular, that the parameter region corresponding to coexistence via synchronous pattern formation is adjacent to the parameter values that do not permit coexistence. The trajectories shown in Figure 2.8 are representative, with the many other starting states we tried yielding similar behavior.

2.4.3 Asymmetric case study

In the asymmetric case, the first growth rate β_1 is different from other majority growth rates (β) that are equal. For comparison with the particle system results, we study the case $L = 4$. The interior equilibrium of the system (2.5) is given by $u_1 = \frac{1}{\beta_1}$ and $u_2 = u_3 = u_4 = \frac{1}{\beta}$, $v_1 = v_2 = v_3 = v_4 = \frac{1}{4} - \frac{1}{4\beta_1} - \frac{3}{4\beta}$, where positive species densities

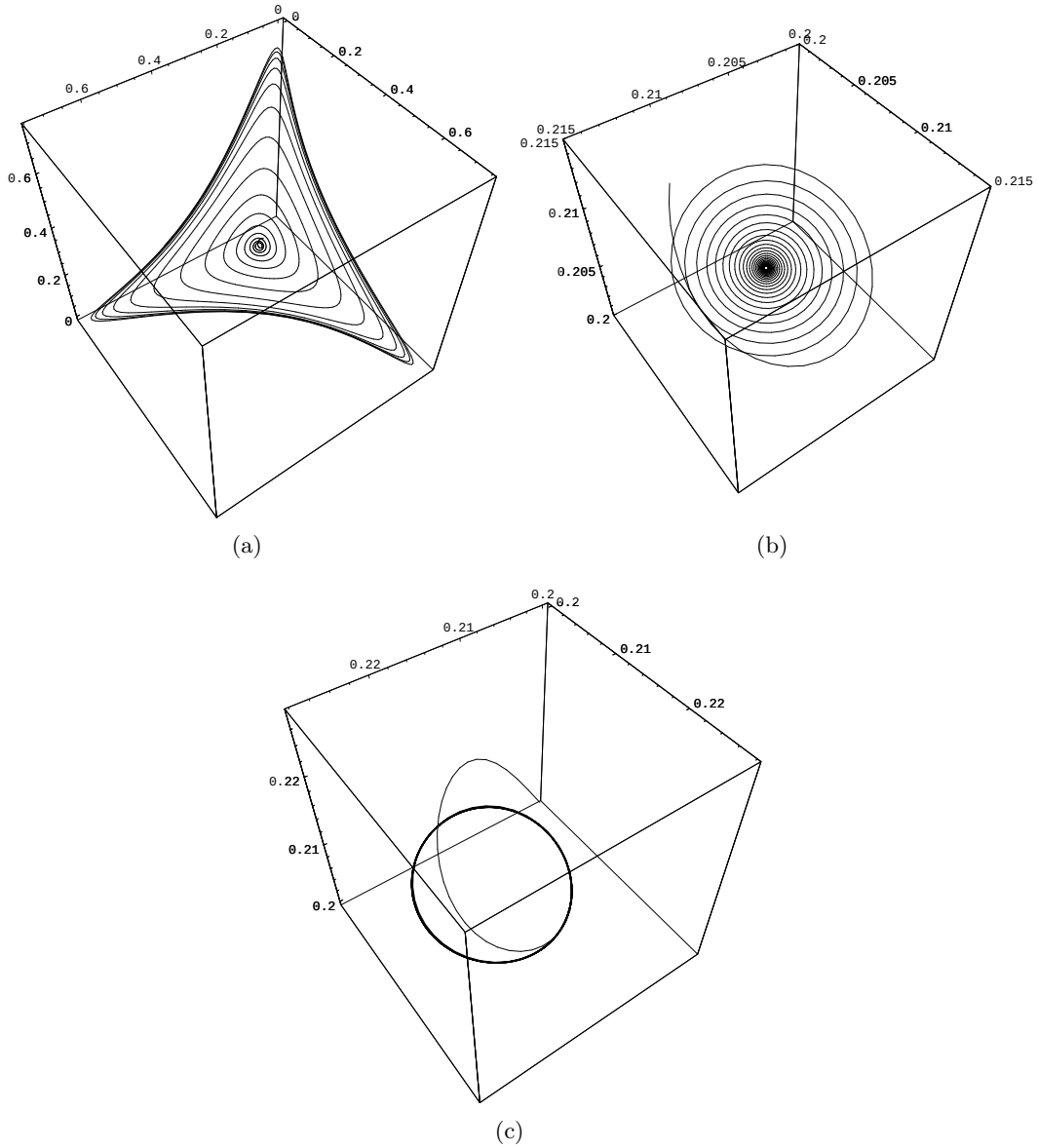

Figure 2.8: Representative trajectories plotting the densities of the three resources for different values of the common growth rate: (a) $\beta = 3.8$ giving spiraling-out (nearly) heteroclinic orbit; (b) $\beta = 4.8$ giving trajectory spiraling in to stable interior equilibrium; (c) $\beta = 4.5$ giving a limit cycle.

requires a constraint on β and β_1 in the form of positivity of the parameter

$$a \equiv \frac{1}{4} - \frac{1}{4\beta_1} - \frac{3}{4\beta}. \tag{2.10}$$

In terms of this parameter, the corresponding Jacobian matrix is

$$J_f(X^*) = \begin{pmatrix} -a\beta_1 & -1 & 0 & 0 & 0 & 0 & 0 & 1 \\ a\beta_1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & -a\beta & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & a\beta & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & -a\beta & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & a\beta & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & -a\beta & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 & a\beta & 0 \end{pmatrix}. \tag{2.11}$$

Figure 2.9 illustrates the linearization results for the asymmetric model. Notice that the hyperbolic curve separating the black and white regions comes from Equation (2.10). One should compare this figure to Figure 2.4 to appreciate how well the linearization and eigenvalue argument explains and predicts the particle system behavior in the asymmetric case.

2.4.4 A coupled map lattice interpretation

Another perspective on the connection between the behavior of the IPS model and the mean-field ODE can be gleaned from a kind of coupled map lattice interpretation. For specificity, we restrict this discussion to the symmetric case discussed in Sections 2.3.1 and 2.4.2.

The global (average) behavior of the IPS model is given approximately by the mean-field ODE if the neighborhood size (i.e., the interaction range) is so large that all sites

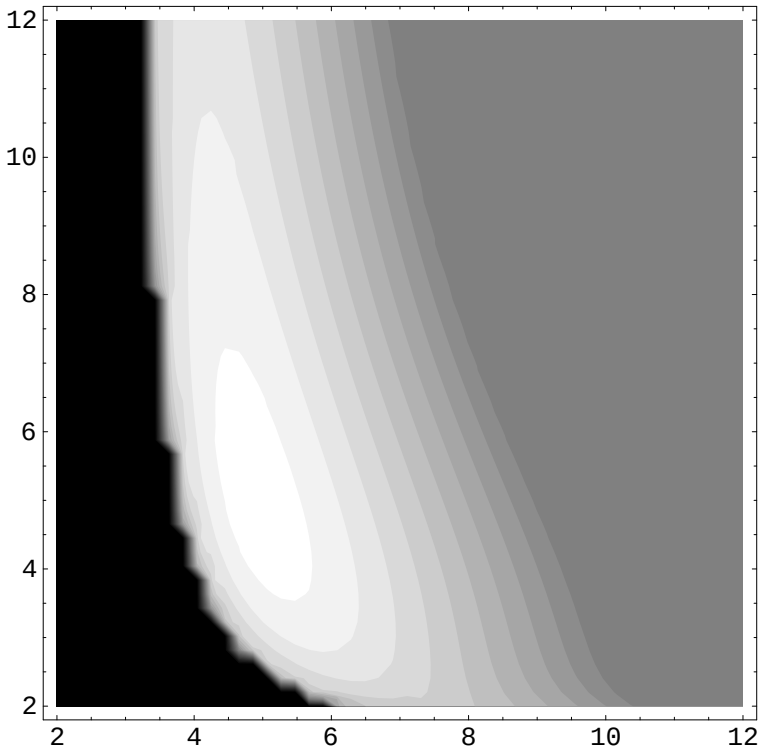

Figure 2.9: The largest real part for the 8 eigenvalues of the matrix (2.11) is plotted as a function of the majority growth rate β (horizontal axis) and the growth rate, β_1, of the first species (vertical axis). The dark region corresponds to parameter values that result in extinction of all species (i.e., no interior equilibrium and hence a frozen pattern); the gray region on the right is where none of the eigenvalues has positive real part (i.e., the largest real part is zero) and corresponds to mixing; in the remaining region, there are eigenvalues with positive real part, and the gray level is proportional to the size of the largest such real part, with lighter color corresponding to larger real part. This last region corresponds to synchronous waves, with lighter color implying larger length scales for the waves.

are neighbors (Durrett 1995). For smaller interaction ranges, the ODE still provides some information about global averages, although the parameters no longer match exactly and stochastic effects play a more prominent role. Now imagine that the lattice is partitioned into a collection of spatially distributed habitat patches and, with the above approximation in mind, the dynamics within a patch proceed according to the mean-field ODE and nearby patches interact to couple the whole system. Of course, the interaction ranges are not very large and the "patches" overlap considerably, but such simple approximations often lead to significant insight.

For large values of β, the mean-field ODE has a globally attracting interior equilibrium (cf. Figure 2.8(b)). This means that the patches will have type densities that are fairly stable; in particular, each species will be present at positive density with high probability. So even if a patch happens to be missing a species, it has the opportunity to quickly regain that species from surrounding patches and re-establish the approximate equilibrium. This suggests that large β leads to mixed configurations, in agreement with our simulations of the IPS model.

For intermediate values of β, one obtains eigenvalues with positive real part. The interior equilibrium of the mean-field ODE is then unstable with trajectories approaching something close to a heteroclinic orbit (cf. Figure 2.8(a)), with episodes of domination by a single resource (and low densities of all other resources and species) followed by rapid invasion of the preferred species (the one having the dominant resource as its required resource) and subsequent decay to domination by the next resource in the cycle. This means that the patches spend most of their time with type densities that are dominated by a single resource. If this domination is strong, in the sense that the "corners" of the trajectories in Figure 2.8(a) are close to the coordinate axes, then to keep the species from going extinct the patches must synchronize the surges of species invasion in a way that leads to waves of species spreading into regions dominated by a single resource. If this happens, then there will be contiguous regions dominated by the product resource, setting

the stage for the next species in the cycle, etc.

2.5 Discussion

Bacterial communities structure their environments to enhance persistence and coexistence. Biofilms provide a ubiquitous example of such structuring. Another common situation of structuring in natural communities is seen in bacterial mats. This is a striking example of niche partitioning in which different microbial species organize themselves in horizontal slabs along nutrient or electron acceptor gradients; here, exogenous heterogeneity in the environment seems to be the main force driving the spatial organization rather than the purely self-organized pattern formation in our system. Looking at a cross-section of a bacterial mat, one sees with the unaided eye sharp boundaries between different species. This niche partitioning is essentially constant in time. The spatio-temporal niche partitioning seen in the synchronous waves of our resource-species model can be thought of as space-time slabs of different species and resources; they are, however, significantly thinner than the static bands of a bacterial mat.

In the model presented here, the length scales of the macroscopic patterns are large (~ 1 mm) compared to the sizes of the bacterial cells (~ 1 μm). To see where these lengths come from in the IPS model, just note that lattice points correspond to the locations of individual cells, so accounting for a cell size of about one or two μm and intercellular distances also of the order of μm's, we can think of the real distance between lattice points as being roughly 1-10 μm. With hundreds (or thousands) of lattice points per side for a typical simulation grid, we should consider our viewing window to be several mm on a side, and hence this is also the length scale for patterns that extend across a large portion of this window.

This spatial structuring of the resources may lead to protocols for detecting such cyclic dynamics in the field or in laboratory bacterial communities. When spatial patterning is strongest, the band width of resources is much larger than that for species, so methods that

can distinguish between the different resources may be able to pick up the tell-tale patterns even without the species being present. The length scale estimates mentioned above should also help to direct this. In general, methods that allow one to use "macroscopic" observations, say of large-scale spatial patterns, to acquire knowledge about the underlying microscopic mechanisms can be quite useful.

To our knowledge, microbial systems with the specific dynamics described here have not yet been observed, although the types of feedback mechanisms we assume are known to exist and cyclic dynamics are quite common in nature. Furthermore, cross-feeding appears to arise quite naturally in microbial systems, even in the absence of spatial structure. For example, Rosenzweig et al. (1994) observed the evolution of such a polymorphism in a chemostat population of *E. coli* initiated with a single clone. As the population evolved to be dominated by individuals that could more quickly exploit the single limiting nutrient (glucose), the ability to assimilate secondary metabolites (acetate and glycerol) was rendered less efficient, thus opening the way for mutant strains that specialized on these secondary resources. Doebeli (2002) gives a nice mathematical account of the evolution of cross-feeding polymorphisms. Despite all this evidence, it is not surprising that microbial systems with the cyclic spatial dynamics presented here have not been observed. This is in large part due to the fact that microbial observations on such a fine spatial scale are not often undertaken, and in laboratory communities where it is easier to gain spatial resolution through confocal microscopy and fluorescent markers, one rarely deals with communities of more than a few species. We consider this a challenge to microbial ecologists.

This is an excellent area for the blending of theoretical and empirical studies in microbial ecology. A few other examples of theoretical and empirical work combined at the microbial level are Durrett and Levin (1997), Kerr et al. (2002), and Wei and Krone (2005). With individual-based lattice models, we can effectively "see" at an individual cell level the effects of certain mechanisms. This leads to biological hypotheses that can

be tested empirically. Our simulations show the types of patterns and length scales one should look for and also point out that it is enough to detect the nutrient patterning. This work also shows that spatial patterns in microbial (and other) systems can be a result of self-organization and not necessarily the result of inhomogeneities in the substrate. It is an example of microbial communities structuring both their abiotic and biotic environments.

In situations for which synchronous waves arise, another interesting feature holds. Even if the species were to be driven to extinction by external forces, say due to fluctuations in the environment (that are not modeled), the consumer-driven structuring of the resources leads to the potential for rapid re-establishment of the microbial community and its spatial structure provided a small influx of immigrants is able to re-seed the prepared environment before this structure erodes; cf. the comments about "seeding" and hard-to-start spirals in Section 2.3. This suggests an additional robustness to such cyclic systems.

All of these features help to illustrate why cyclic dynamics are likely to be prevalent in natural spatially structured systems and, in particular, why they could be a regular component of microbial communities. More empirical evidence is needed and the results of this paper illustrate the types of spatial patterns that one might look for. We intend to elaborate on the themes of this chapter elsewhere by including evolutionary dynamics in our model. [1]

[1]The author would like to thank Larry Forney for extensive comments on the manuscript, improving both presentation and biological precision. Thanks also to Zaid Abdo, and Holger Heuer for discussions.

Part II

Phylogenetic Distance Methods

Chapter 3

Least Squares and Neighbor Joining

3.1 Introduction

The neighbor-joining method of Saitou and Nei (1987) has been the most popular distance method for inferring phylogenies because it is fast and fairly accurate. Saitou and Nei proved that if the pairwise distance matrix were additive (cf. Saitou and Nei 1987 and Gascuel 1997a for precise definition), neighbor-joining always finds the true tree. In addition, Atteson (1999) has shown that the additivity condition can be relaxed a little for some distance methods, including neighbor-joining, to find the true tree. Felsenstein (2004, p.169) has made the following comment about the neighbor-joining method:

> "... although it computes their branch lengths by least squares, it does not use the sum of squares as the criterion for choosing which pair of species to join. Instead it uses the total length of the resulting tree, choosing that pair that minimizes this length. This is a form of the minimum evolution criterion. But we also cannot identify neighbor-joining as a star decomposition search for the minimum evolution tree, neighbor-joining allows negative branch lengths, while minimum evolution bans them."

Furthermore, he suggested that it would be of interest to see whether a star decomposition algorithm for least squares, or one for minimum evolution, could be developed that was comparable in speed to neighbor joining. This chapter shows that neighbor joining is indeed a special case of least squares methods.

Among distance methods, the least squares methods are of particular interest because they use a single objective function to solve for branch lengths and to choose among tree topologies and can thus be related to the least squares method of statistical estimation (Felsenstein 1997). In least squares methods, each tree topology τ is associated with a least squares score $Q(\tau)$ (up to a constant) and a set of branch lengths estimated under the least squares criterion given the tree topology τ. In general $Q(\tau)$ can be defined as

$$Q(\tau) = \sum_{i,j} w_{ij}(l_{ij} - d_{ij})^2, \tag{3.1}$$

where l_{ij} is the path length between taxa i and j on the topology τ, the w_{ij}'s are weights, and the d_{ij}'s are pairwise distances between taxa. Our goal is to find the tree with the smallest least squares score $Q(\tau)$. (In the case when the data are additive, Q would be zero.)

Although in theory we can calculate the least squares score for every possible tree topology, a brute force approach is not feasible in practice because the number of possible tree topologies increases exponentially with the number of taxa. Fortunately, we may use a star decomposition algorithm to search tree space for the (hopefully) best tree. As its name implies, a star decomposition algorithm starts with a totally unresolved star tree. At each step it coalesces two taxa as one (unity), and results in a one-taxa-less star tree. After the distances between the unity and the remaining taxa are estimated, we have a new distance matrix of one dimension less. Repeating this dimension reduction procedure, we obtain a totally resolved tree in the end. The choice of which two taxa to coalesce in each step and how to estimate the corresponding branch lengths are different in each method, of course. In neighbor-joining, the taxa to coalesce are chosen to minimize the total tree length (minimum evolution) while the branch lengths are estimated using the least squares criterion. Meanwhile, Weighbor (Bruno, Socci, and Halpern 2000) chooses taxa (to coalesce) to maximize the likelihood of additivity and positivity of the observed distances;

Bionj (Gascuel 1997a) picks taxa similar to neighbor-joining, but the new reduced distance matrix is estimated to minimize the variances of pairwise distances. In this least squares method, the choice of taxa and branch length estimation are done to minimize the least squares score of the resulting tree.

In what follows, we first develop the unweighted least squares algorithm, and then compare our algorithm with neighbor-joining to show that the two algorithms are in agreement with each other. A short discussion on generalizing to weighted least squares will conclude the chapter.

3.2 Unweighted least squares star decomposition

Suppose we have n taxa and pairwise distances matrix (d_{ij}). We begin with a star tree with the center node X and let l_{iX} be the i-th branch length. Setting $w_{ij} = 1$ in Equation (3.1), we have the unweighted least squares score of the star tree

$$Q = \sum_{i,j} (l_{iX} + l_{jX} - d_{ij})^2. \tag{3.2}$$

We often refer to unweighted least squares as ordinary least squares, or OLS for short. To compute the OLS estimate of each branch length of the star tree, we take derivatives of Q with respect to each l_{iX} and set them equal to zero to get a linear system:

$$(n-1)l_{iX} + \sum_{j:j\neq i} l_{jX} = \sum_{j:j\neq i} d_{ij} \quad (i = 1, 2, \cdots, n). \tag{3.3}$$

To write the above system in vector form, let $L = (l_{1X}, l_{2X}, \cdots, l_{nX})^T$ be the column vector of branch lengths, define $D_i = \sum_{j:j\neq i} d_{ij}$, column vector $D = (D_1, D_2, \cdots, D_n)^T$,

and

$$A = \begin{pmatrix} n-1 & 1 & \cdots & 1 \\ 1 & n-1 & \cdots & 1 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & \cdots & n-1 \end{pmatrix}. \tag{3.4}$$

We have $AL = D$. Since obviously A is a nonsingular matrix, we have

$$L = A^{-1}D. \tag{3.5}$$

Notice that the vector D is known. Since the matrix A has only one degree of freedom, it is not hard to find its inverse

$$A^{-1} = \begin{pmatrix} \frac{2n-3}{2(n-1)(n-2)} & \frac{-1}{2(n-1)(n-2)} & \cdots & \frac{-1}{2(n-1)(n-2)} \\ \frac{-1}{2(n-1)(n-2)} & \frac{2n-3}{2(n-1)(n-2)} & \cdots & \frac{-1}{2(n-1)(n-2)} \\ \vdots & \vdots & \vdots & \vdots \\ \frac{-1}{2(n-1)(n-2)} & \frac{-1}{2(n-1)(n-2)} & \cdots & \frac{2n-3}{2(n-1)(n-2)} \end{pmatrix}. \tag{3.6}$$

Proof of (3.6). Let $A = (n-2)I + B$, where I is the $n \times n$ identity matrix, B is the $n \times n$ matrix with each entry equal 1. Then (3.6) becomes

$$A^{-1} = \frac{1}{n-2}I + \frac{-1}{2(n-1)(n-2)}B.$$

So we have

$$\begin{aligned} A\ A^{-1} &= ((n-2)I + B)(\frac{1}{n-2}I + \frac{-1}{2(n-1)(n-2)}B) \\ &= I - \frac{1}{2(n-1)}B + \frac{1}{(n-2)}B - \frac{1}{2(n-1)(n-2)}B\ B \\ &= I - B(\frac{1}{2(n-1)} - \frac{1}{(n-2)} + \frac{n}{2(n-1)(n-2)}) \\ &= I. \end{aligned}$$

Notice the third equality hold because $B\ B = nB$. □

Substitute (3.6) back into Equation (3.5) and we get

$$l_{iX} = \frac{D_i}{n-2} - \frac{\sum_k D_k}{2(n-1)(n-2)} \quad (i = 1, 2, \cdots, n). \tag{3.7}$$

These are the OLS estimates of branch lengths of the star tree.

Our next step is to pick two taxa and coalesce them into a unity. The taxa are chosen such that Q decreases most. Define a new matrix (e_{ij}) with

$$e_{ij} = d_{ij} - l_{iX} - l_{jX} \tag{3.8}$$

and notice that

$$\sum_{i,j} e_{ij} = 0. \tag{3.9}$$

Proof of (3.9). From (3.8) and (3.7) we get,

$$\begin{aligned}
\sum_{i,j\neq i} e_{ij} &= \sum_{i,j\neq i} (d_{ij} - l_{iX} - l_{jX}) \\
&= \sum_i D_i - 2\sum_{i,j\neq i} l_{iX} \\
&= \sum_i D_i - 2(n-1)\sum_i l_{iX} \\
&= \sum_i D_i - 2(n-1)\sum_i (\frac{D_i}{n-2} - \frac{\sum_k D_k}{2(n-1)(n-2)}) \\
&= \sum_i D_i - \frac{2(n-1)}{n-2}\sum_i D_i + \frac{n}{n-2}\sum_k D_k \\
&= 0.
\end{aligned}$$

□

Equation (3.9) is also an immediate consequence of Property 5.1 of Gascuel (1997b). Obviously, Equations (3.2) and (3.8) imply $Q = \sum_{i,j} e_{ij}^2$. For each entry e_{ij}, we may have

three different cases.

1. $e_{ij} < 0$ implies the distance between taxa i and j (the sum of $l_{iX} + l_{jX}$) is *over*-estimated. If we coalesce taxa i and j and let them share an internal branch to node X (see Figure 3.1), we can decrease Q.

2. $e_{ij} = 0$ implies the distance between taxa i and j is exactly estimated.

3. $e_{ij} > 0$ implies the distance between taxa i and j is *under*-estimated, therefore, taxa i and j should be separated by internal branches in order to decrease Q.

Each negative entry in matrix (e_{ij}) provides a candidate taxa pair i and j to be coalesced. Equation (3.9) guarantees the supply of negative entries in (e_{ij}). (In the trivial case where all $e_{ij} = 0$, the star tree would be the best tree. Otherwise, we have at least one $e_{ij} < 0$.) We need to decide which pair to coalesce when we have multiple choices of negative entries.

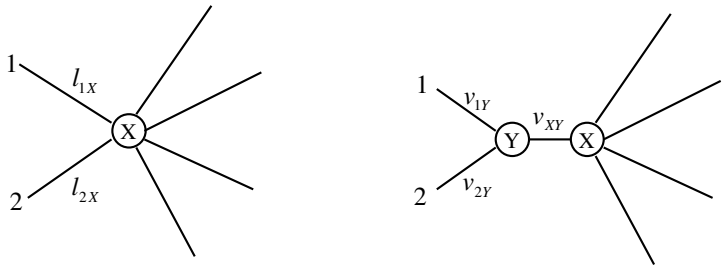

Figure 3.1: One step star decomposition. The left figure shows a star tree with internode X and the branch lengthes (l_{1X},l_{2X},$\cdots$) are estimated by least squares. The right figure shows coalescing of taxa 1 and 2 to form a unity node Y, and new branch lengths (v_{1Y}, v_{2Y}, v_{XY},$\cdots$) are estimated.

Without loss of generality, assume $e_{12} < 0$. Figure 3.1 shows the coalescence of taxa

1 and 2 and the corresponding branch lengths. Set

$$\begin{aligned} l_{1X} &= v_{1Y} + v_{XY} \\ l_{2X} &= v_{2Y} + v_{XY} \\ d_{12} &= v_{1Y} + v_{2Y} \end{aligned}$$

Note that in the above system, l_{1X}, l_{2X} and d_{12} are known, so we can easily solve for the three unknowns v_{1Y},v_{2Y} and v_{XY} and get

$$\begin{aligned} v_{XY} &= (l_{1X} + l_{2X} - d_{12})/2 \\ v_{1Y} &= (d_{12} + l_{1X} - l_{2X})/2 \\ v_{2Y} &= (d_{12} + l_{2X} - l_{1X})/2. \end{aligned} \tag{3.10}$$

After coalescing taxa 1 and 2 and computing the new branch length using Equation (3.10), we have the new error between taxa 1 and 2 as $e'_{12} = d_{12} - v_{1Y} - v_{2Y} = 0$. But e_{1k} and e_{2k} stay the same for $k = 3, 4, \cdots, n$. We have immediately that Q as defined in (3.2) decreases by e_{12}^2. This implies that in order to maximally decrease Q, we should pick the negative entry with the maximum absolute value in the matrix (e_{ij}) and coalesce taxa i and j.

Let Y be a new taxon representing the unity of taxa 1 and 2. We need to estimate the distance d_{kY} between taxon Y and k for all $k > 2$. We shall use the same formula as in (Saitou and Nei 1987) to estimate d_{kY}:

$$d_{kY} = (d_{1k} + d_{2k} - d_{12})/2. \tag{3.11}$$

In fact, Equation (3.11) is a special case of what Gascuel (1997a) called a "reduction formula." In the same paper, Gascuel has shown, among other things, that computing d_{kY} using Equation (3.11) will preserve the "additive property" of the data. The new

d_{kY}'s and the remaining d_{ij}'s, $i, j > 2$, give us a new distance matrix of one dimension less. Thus, we may repeat the above procedure until the tree is totally resolved.

We comment that, in general, if we calculate v_{XY}, v_{1Y} and v_{2Y} differently (see discussion), the decrease of Q might involve multiple terms of e_{ij}. However, once we use Equation (3.10) to calculate v_{XY}, v_{1Y} and v_{2Y}, e_{12} is the only modified error term as defined in Equation (3.8); others are all kept the same. In other words, Equation (3.10) in fact singles out one error term and screens out others, so that we may focus on one taxa-pair at each dimension reduction step. It is worthwhile to point out that the method presented here is to search (a tree space) for optimal trees (the ones with lowest least squares scores) using star decomposition. This approach belongs to a class of methods called "dynamic programming." Very loosely, the spirit of dynamic programming is to "worry about one thing at a time". Equation (3.10), by fixing the rest of the e_{ij}, enables us to "worry" about one of them (the negative one with maximum absolute value) at a time.

There is a resemblance (formula-wise) between the method presented here and the one in the appendix of Saitou and Nei (1987). We emphasize two major differences here. First, what Saitou and Nei have shown is that the total tree length as in Figure 3.1 (right) is the sum of the least squares estimates of branch lengths. This is exactly why they need the "minimum evolution" criterion — they use total tree lengths (instead of least squares scores) to pick two taxa to coalesce. On the other hand, we show here that one can use least squares scores to pick two taxa to coalesce, and the branch lengths are also estimated from the least squares criterion, so that it is a pure least squares method. Secondly, Saitou and Nei focus on trees as in Figure 3.1 (right), while we mainly work with star trees (Figure 3.1, left). Putting minimum evolution aside, we have an "advantage" in that the mathematics is simpler. For example, the inverse of matrix A as in Equation (3.4) is easy to guess.

In summary, the unweighted least squares star decomposition has following steps:

1. For each tip, compute a branch length using Equation (3.7).

2. Compute (e_{ij}) using Equation (3.8). Choose the i and j such that e_{ij} is the negative entry with maximum absolute value. If all $e_{ij} = 0$, stop and an optimal tree (may not be a totally resolved tree) is obtained.

3. Coalesce i and j as Y. Compute the branch length of (v_{iY}) and (v_{jY}) using Equation (3.10) with i replacing 1 and j replacing 2, respectively.

4. Compute the distance between the unity node Y and each of the remaining tips using Equation (3.11) with i replacing 1 and j replacing 2 respectively.

5. Delete tips i and j from the table and replace them by the unity Y, which is now treated as a tip.

6. If more than two nodes remain, go back to step 1. Otherwise connect the two remaining nodes (say h and m) by a branch of length d_{hm}.

3.3 Neighbor joining vs. least squares

The neighbor-joining algorithm (as modified by Studier and Keppler 1988; see also Felsenstein 2004, p.167) is:

1. For each tip i, compute $l_i = \sum_{j:j\neq i}^{n} d_{ij}/(n-2)$.

2. Choose the i and j for which $d_{ij} - l_i - l_j$ is smallest.

3. Join items i and j. Compute the branch length from i to the new node (v_i) and from j to the new node (v_j) as

$$v_i = (d_{ij} + l_i - l_j)/2 \qquad v_j = (d_{ij} - l_i + l_j)/2$$

4. Compute the distance between the new node (ij) and each of the remaining tips as

$$d_{(ij),k} = (d_{ik} + d_{jk} - d_{ij})/2$$

5. Delete tips i and j from the table and replace them by the new node, (ij), which is now treated as a tip.

6. If more than two nodes remain, go back to step 1. Otherwise connect the two remaining nodes (say h and m) by a branch of length d_{hm}.

If we compare least squares with neighbor-joining, we can see that the major difference between them is the computation of l_{iX} and l_i. In least squares,

$$l_{iX} = \frac{D_i}{n-2} - \frac{\sum_k D_k}{2(n-1)(n-2)},$$

while in neighbor joining,

$$l_i = \sum_{j:j\neq i}^{n} d_{ij}/(n-2).$$

However, $l_{iX} - l_i = \frac{\sum_k D_k}{2(n-1)(n-2)}$ is a constant, and in the subsequent computations of v_i, v_j and $d_{(ij),k}$, the constant has been cancelled. So, the least squares method and neighbor-joining result in the same topology and same branch lengths.

3.4 Discussion and conclusion

Weighted least squares star decomposition and its difficulty

It has been pointed out, for example, by Gascuel, Bryant, and Denis (2001), that unweighted least squares implicitly assumes each estimated pairwise distance is independent and has the same variance, which is not generally true. Weighted least squares accounts for the variable variance of the estimates and hence tends to be more reliable. The most

reliable method to estimate branch lengths is generalized least squares, which in addition takes into account the covariance between estimates. Such discussion is beyond the scope of our chapter and we recommend (Gascuel, Bryant, and Denis 2001) and (Felsenstein 2004, chap. 11) for more details.

For weighted least squares, we may take derivatives of Q in Equation (3.1) with respect to l_i and set them equal to 0. Borrowing the notation from the unweighted least squares, we can calculate the weighted least squares estimates of branch lengthes and we get:

$$L = B^{-1}D$$

where

$$B = \begin{pmatrix} \sum_{k\neq 1} w_{1k} & w_{12} & \cdots & w_{1n} \\ w_{21} & \sum_{k\neq 2} w_{2k} & \cdots & w_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ w_{n1} & w_{n2} & \cdots & \sum_{k\neq n} w_{nk} \end{pmatrix}. \tag{3.12}$$

However, by standard matrix manupulation, the complexity to find the inverse of the matrix B is $O(n^3)$ and this will slow down the speed of the algorithm to $O(n^5)$ comparing to $O(n^3)$ of the unweighted least squares case. The unweighted case is faster because we have a simple explicit formula to calculate the inverse. Although by assuming $w_{ij} = w_i w_j$ the explicit formula to calculate the inverse of matrix B has been found (result not shown), it does not ease the computation of the inverse. For people who are interested in fast numerical recipes, we recommend (Bryant and Waddell 1998) and references therein.

Branch length estimation

Felsenstein has pointed out (personal comm.) that there are other methods for estimating d_{kY} than the one given in Equation (3.11). For example, since Y represents two taxa it is reasonable to give more weight to the inner branch length v_{XY} (see Figure 3.1). Then the new distance matrix with one dimension less would be slightly different from

the current algorithm. Whether or not such a difference would affect the taxa choice in the following step, to what extent it will affect the final result, and how different weight would affect the robustness of the algorithm remain interesting open questions that might be answered by extensive simulations.

Bottom-up vs. top-down

Methods that use star decomposition, like neighbor-joining, adopt a bottom-up approach to search tree space in the sense that it starts from the tips of a tree and traces back to the root. It is natural to consider the possibility of the reversed "top-down" approach, i.e., one may start from the root of a tree and recursively partition the taxa into subgroups, reaching out to the tips. In a possible future work, we will report on such a "top-down" algorithm that is fast and is a by-product of the proof reported in this chapter.

To conclude, in this chapter we have not proposed a new algorithm that competes with neighbor-joining. Rather, we have shown that the neighbor-joining algorithm can be classified into the least squares framework, so that the two criteria of neighbor-joining (least squares and minimum evolution) are essentially one (least squares). We may thus regard the neighbor-joining method as an unweighted least squares method to search the optimal tree through star decomposition. Thus it explains why neighbor-joining is so robust. [1]

[1]The author would like to thank Joe Felsenstein for illuminating comments and discussions. Thanks also to Jack Sullivan and Roland Fleißner for their comments and discussions.

Part III

Markov Chain Monte Carlo

Chapter 4

Markov Chain Monte Carlo in Small Worlds

4.1 Introduction

Markov Chain Monte Carlo (Gamerman 1997) is a sampling scheme for surveying a space S with a prescribed probability measure π. It has particular importance in Bayesian analysis, where $x \in S$ represents a vector of parameters and $\pi(x)$ is the posterior distribution of the parameters conditional on the data. MCMC can as well be used to solve the so-called missing data problem in frequentist statistics. Here, $x \in S$ represents the value of a latent or unobserved random variable, and $\pi(x)$ is its distribution conditional on the data. In either case, MCMC serves as a tool for numerical computation of complex integrals and is often found to be the only workable approach for problems involving a large space with a complex structure where traditional numerical methods are not possible.

As its name implies, MCMC does not attempt to draw elements of S independently of each other, but instead relies on a Markov chain which moves through S. Probably the most widely used version of MCMC is the Metropolis-Hastings algorithm (Hastings 1970) which works in the following way: Suppose the chain is at a point $x \in S$, the algorithm then proposes a move to $y \in S$ following some proposal distribution $q(x, y)$. The move from x to y is either accepted or rejected and the acceptance probability is given by

$$a(x, y) = \min\left(1, \frac{\pi(y)q(y, x)}{\pi(x)q(x, y)}\right) \tag{4.1}$$

Therefore, the constructed Markov chain moves from state x to state y with prob-

ability $T(x,y) := q(x,y)a(x,y)$. It is easy to check that the detailed balance equation $\pi(x)T(x,y) = \pi(y)T(y,x)$ holds. Hence, setting up the acceptance probability in the above way results in an ergodic reversible Markov chain (Meyn and Tweedie 1996) with stationary distribution π provided that $q(x,y)$ is ergodic. This means that if we are able to run this chain long enough the frequency with which we observe $x \in S$ in our sample will converge to $\pi(x)$. However, if we do not pick the proposal distribution $q(x,y)$ wisely the Markov chain might reject most of the proposals and thus may not be active enough to reach the stationary distribution in a reasonable number of steps. Furthermore, if π has heavy tails or multiple modes it may become difficult for the chain to explore all the important regions of S.

To ensure activeness of the chain many implementations of the Metropolis-Hastings algorithm use local proposals. Assuming that S is a metric space with metric d, typically taken to be the Euclidean distance, we call a proposal distribution $q(x,y)$ a local proposal if $q(x,y)$ decreases rapidly with increasing distance between x and y, as is the case if $q(x,\cdot)$ is a normal distribution or if it has a compact support,i.e., $q(x,y) = 0$ if $d(x,y) > r$ for some finite r. If π is a smooth function then picking y in the neighbourhood of x will lead to a high acceptance rate $a(x,y)$. However, the resulting small step size means that it will take a large number of steps to move a substantial distance from the starting point. Therefore, it has been argued (Jarner and Roberts 2001) that, at least if π is heavy tailed, the proposal distributions should have heavy tails, too.

In the next two sections, we will demonstrate how very simple heavy-tailed proposal distributions – namely mixtures of local proposals and random draws – outperform pure local proposal schemes, especially if π is multi-modal. We will first work out some mathematics to illustrate why these proposal distributions should perform better than traditional MCMC (Section 4.2). Then we will present some simulations to support our argument (Section 4.3).

Throughout this chapter, we will assume that S, the state space of the Markov chain,

is a space with some metric $d(\cdot,\cdot)$ and that it is equipped with the Borel σ-field $\mathcal{B}(S)$ and two measures: a canonical measure μ and a probability measure π, the stationary distribution of the Markov chain on S. Let $|B| := \mu(B)$ for any $B \in \mathcal{B}(S)$. We assume $|S| < \infty$. In the continuous case, S typically is a compact subset of R^n typically with Lebesgue measure as its canonical measure, while in the discrete case, S is a finite set with μ being counting measure. Let $N_x = \{y : d(x,y) < r\}$ denote the local neighborhood of x. We further assume that the standard local proposal $q(x,y)$ is the uniform distribution over N_x, i.e., $q(x,y) = 1/|N_x|$ if $y \in N_x, 0$ otherwise. In order to keep the notation simple the rest of this chapter will only describe the discrete case. Yet, as any implementation of a continuous problem would necessitate the discretization of S, this does not affect the applicability of our method.

4.2 Metropolis-Hastings with small-world proposals

4.2.1 Small-world proposal distributions

Our choice for a modified proposal distribution is motivated by the so-called small-world networks (Watts and Strogatz 1998). These graphs are characterized by a much shorter average path length than regular lattices in spite of retaining considerable regularity. They can be constructed by randomly rewiring a small fraction of the edges of a regular network. It turns out that replacing a relatively small number of edges with long-range connections results in a drastic decrease in the average path length.

In the context of the Metropolis-Hastings algorithm we may impose such a "small-world effect" by altering the proposal distribution as follows. With a large probability $1 - p$, a move is proposed according to the local proposal distribution. However, with some small probability p, we propose, for example through a random draw from S, a move that is typically far away from the current state. These "wild" proposals play the role of the long-range connections in the small-world networks. Precisely, if $q(x,y)$ is our local

proposal distribution and $|S|$ represents the canonical measure of S then let

$$p(x, y) := (1 - p)q(x, y) + p/|S| \tag{4.2}$$

be the new proposal distribution. We call such a mixture of local proposals and random draws a *small-world proposal distribution*. In the following we refer to a Markov chain which uses small-world proposals as a small-world chain or SWC and we call a chain which only relies on local proposals a local-proposal chain or LPC. Below, we show how the probability of making a large jump in a SWC depends on π (Section 4.2.2), then calculate the cost of a SWC in terms of the chain's average acceptance rate (Section 4.2.3), and provide a rule for how to choose p (Section 4.2.4).

4.2.2 The probability of large jumps

Suppose A and B are two disjoint subsets of the state space S. Let us also assume that none of the points in B lies in the neighbourhood of a point in A. Then the probability that a SWC which is currently wandering in A jumps to a site in B in a single step is given by

$$\lambda = \sum_{x \in A} \sum_{y \in B} \frac{\pi(x)}{\pi(A)} \min\left(1, \frac{\pi(y)}{\pi(x)}\right) \frac{p}{|S|} \tag{4.3}$$

where $\pi(A) = \sum_{x \in A} \pi(x)$. To gain some insight into Equation (4.3), consider three special cases:

1. If A corresponds to a flat region and B to a hill, then $\pi(y) > \pi(x)$, for $y \in B, x \in A$. From (4.3), we get $\lambda = p\frac{|B|}{|S|}$, which is proportional to the relative size of B.

2. If A is a hill and B is a flat region, then $\pi(y) < \pi(x)$, for $y \in B, x \in A$. From (4.3), we get $\lambda = p\frac{1}{|S|}\frac{1}{\pi(A)/|A|}\pi(B)$. Notice that $\pi(A)/|A|$ is the average probability, or average height, of A. The higher A, the more difficult it is for a chain to jump out. λ is also proportional to the total measure of the flat region B.

3. If both, A and B, are hills. Using the fact that

$$\min(a,b) = \frac{1}{2}\left((a+b) - |a-b|\right), \tag{4.4}$$

we get

$$\lambda = \frac{1}{2}p\frac{|B|}{|S|}\left(1 + \frac{\pi(B)/|B|}{\pi(A)/|A|} - \frac{1}{\pi(A)|B|}\sum_A\sum_B |\pi(x) - \pi(y)|\right). \tag{4.5}$$

The last term in the parentheses is proportional to the average cross-variation between A and B. The second term in the parentheses is the ratio of the average height of hill B and the average height of hill A. $\frac{|B|}{|S|}$ is the relative size of set B. So if the cross-variation is small and if the ratio of the average heights is about 1, then we get $\lambda \approx p|B|/|S|$.

If we ignore all the paths that lead from A to B without a direct jump from the first subset to the latter, we may use λ as an indicator of how fast the SWC travels between distinct regions of S. In case (1) as well as in case (3), the speed is proportional to the relative size of the second hill. Note also that in case (3) the size of the valley between two hills does not influence the speed with which the chain jumps between the two hills. Notice also that the mean time to move from a flat spot to a hill depends only on the size of the space and not on its dimension. From case (2), we can see that the probability of jumping off a hill to a flat region is proportional to the total probability measure of the flat region. In the case that this total probability measure is small, most proposals of jumping off hills will be rejected. Hence, the Markov chain stays for a long time in important regions.

That the situation is quite different for local proposals can be seen in the following simple example of a multi-modal space where the mean time to move from the 'major hill' to the 'minor hill' can be calculated explicitly. The point of the exercise is to demonstrate that even for simple spaces the mean time to move from one hill to the next will be an

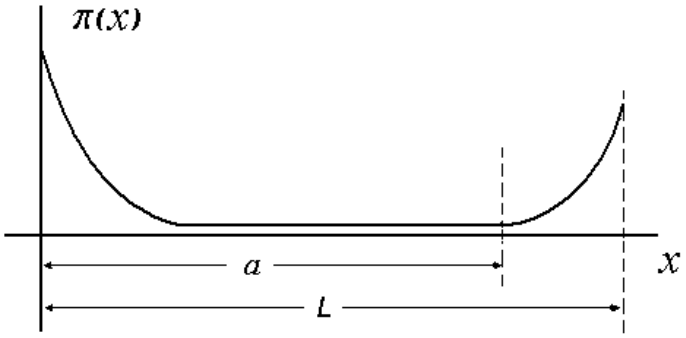

Figure 4.1: $\pi(x)$ as used in example 1.

exponential function of the distance between the hills for a local proposal distribution. Whereas, a SWC will move from one peak to the other in a time that is linear in the distance between the two peaks. Here our neighborhood size is one unit in length. A larger neighborhood will only change the scale factor in the exponential distribution.

Example 1 Let $x \in S = [0, L) \cap \mathbb{Z}$ and suppose that $\pi(x) \propto \alpha e^{-\alpha x}$ for $0 \leq x \leq a$ and $\pi(x) \propto e^{(x-L)}$ for $a < x < L$. Consider a very simple proposal distribution where a move one unit to the right or left is made with equal probability. That is $p(k, k+1) = p(k, k-1) = 0.5$. Note that $\frac{\pi(k+1)}{\pi(k)} = e^{-\alpha}$ for $k + 1 < a$. Now let T_k be the time it takes to move one step down the left hill from state k to state $k + 1$. To calculate $E(T_k)$ we develop a recursion equation by conditioning on all the possible one step moves that can be made while in state k. In order to move one step to the right down the hill, we must propose a move to the right and then accept that move. If instead we propose a move to the right, but reject that move then the process starts anew. We may also propose a move to the left and climb back up the hill. Conditioning on these three possibilities leads to the following recursion for $E(T_k)$:

$$E(T_k) = \frac{1}{2}e^{-\alpha} + \frac{1}{2}\left(1 - e^{-\alpha}\right)\left[E(T_k) + 1\right] + \frac{1}{2}\left[E(T_{k-1}) + E(T_k) + 1\right]. \tag{4.6}$$

This implies

$$E(T_k) = e^{\alpha} E(T_{k-1}) + 2e^{\alpha} - 1. \tag{4.7}$$

We are ultimately interested in the time it takes to move all the way down the hill to the reach the second mode. Denote this time by S_a where

$$S_a = \sum_{k=0}^{a-1} T_k.$$

Since $E(T_0) = e^{\alpha}$ it follows from (4.7) that

$$E(S_a) = e^{\alpha} E(S_{a-1}) + (a-1)(2e^{\alpha} - 1) + e^{\alpha}.$$

Since this is a simple linear recursion we can solve it explicitly and get

$$E(S_a) = e^{\alpha a} + [(a-1)(2e^{\alpha} - 1) + e^{\alpha}] \left[\frac{e^{\alpha a} - 1}{e^{\alpha} - 1} \right]. \tag{4.8}$$

The exponential function in this example descends much more slowly than the normal distribution which is often used to test MCMC algorithms. Still, Equation (4.8) shows that the mean time to move between the two hills is an exponential function of their distance. If we replaced the exponential with the normal distribution the situation would be even worse giving $E(S_a) \approx e^{\alpha a^2}$. A SWC, on the other hand, would move between this example's two hills in a time whose mean is a linear function of their distance, as can be seen by replacing $|B|$ in the discussion of Equation (4.3) with $L - a$ and $|S|$ with L.

4.2.3 Average acceptance rate

By (4.1) and (4.2), the average acceptance rate for a small-world chain at equilibrium can be expressed as:

$$a = \sum_{x \in S} \sum_{y \in S} \pi(x) \left((1-p) q(x,y) + \frac{p}{|S|} \right) \min \left(1, \frac{\pi(y)}{\pi(x)} \right). \tag{4.9}$$

Under the assumption that $q(x,y) = 1/|N_x|$ for $y \in N_x$, we get:

$$\sum_{x \in S} \sum_{y \in S} q(x,y) \min(\pi(x), \pi(y)) = 1 - \frac{1}{2|N_x|} \sum_{x \in S} \sum_{y \in N_x} | \pi(x) - \pi(y) | \tag{4.10}$$

and

$$\frac{p}{|S|} \sum_{x \in S} \sum_{y \in S} \min(\pi(x), \pi(y)) = p \left(1 - \frac{1}{2|S|} \right) \sum_{x \in S} \sum_{y \in S} | \pi(x) - \pi(y) |. \tag{4.11}$$

Denote

$$\begin{aligned} V_l &= \frac{1}{2|N_x|} \sum_{x \in S} \sum_{y \in N_x} |\pi(x) - \pi(y)| \\ V_g &= \frac{1}{2|S|} \sum_{x \in S} \sum_{y \in S} |\pi(x) - \pi(y)|. \end{aligned}$$

It is easy to see that V_l is the average local variation of π whereas V_g is its average global variation. Substituting them back into (4.9), we get the average acceptance rate as

$$a = 1 - V_l + p(V_l - V_g) \tag{4.12}$$

Notice that $1 - V_l$ is the average acceptance rate for the local proposal. In the case where the probability distribution is not too noisy, we have $V_l < V_g$. Therefore, the average acceptance rate of the SWC decreases compared to that of the LPC and this is what we have to pay when using small-world proposals. However, this is a small price, since the dominant term $1 - V_l$ is at least one order of magnitude larger than $p(V_l - V_g)$ and usually

the two terms in (4.12) differ by two to three orders of magnitude. In the case where the probability distribution is very noisy, one may get $V_l > V_g$. In that case the best choice would be to increase p all the way up to 1 and end up with random sampling.

Thus, allowing for wild proposals usually decreases the average acceptance rate. One should, however, keep in mind that a high acceptance rate does not mean fast convergence (Roberts, Gelman, and Gilks 1997). In all of the examples described below, the LPC had a reasonably good acceptance rate. This only means that the local proposal was doing a good job sampling one of the regions, but because the local proposal did not reach the other regions the LPC did not converge. In fact, we could improve the acceptance rate by making the neighborhood size smaller for the local proposal distribution, but this would only serve to sample one of the regions more thoroughly.

4.2.4 Simple strategies for choosing p

In this section we assume that the space S is partitioned into two disjoint subsets A and B. B will represent the "important" region, which can be thought of as a collection of hills. A will be the "unimportant" region which can be thought of as the flat region of the space S. We will assume that $\pi(y) > \pi(x)$ for all $y \in B$ and $x \in A$. If we again ignore any paths from A to B which do not involve large jumps, we can use the result from Section 4.2.2, that the probability of moving from a flat spot to a hill is $\lambda = \frac{|B|}{|S|}p$. Hence, the mean time it takes to move from the unimportant region A to the important region B is approximately

$$\frac{1}{\lambda} = \frac{|S|}{|B|p}.$$

However, the SWC also bears a cost due to its lower acceptance rate. Here, we describe how to choose p so as to minimize the effect of this trade off. We view the time spent in the flat region and the proposed moves from the hill back to the flat region as wasted steps in the chain; all other steps in the chain are used to explore the important part of the space. Suppose that a SWC is run for M steps and let $h(p)$ be the mean number of

wasted steps. Then

$$h(p) = \frac{|S|}{|B|p} + Mp\left(1 - \frac{|B|}{|S|}\right). \tag{4.13}$$

Denote by $r = |B|/|S|$ the fraction of the space that contains the hills. We now solve for p that minimizes $h(p)$. Indeed, since

$$h'(p) = -\frac{1}{rp^2} + M(1-r),$$

setting $h'(p) = 0$ and solving gives

$$p = \frac{1}{\sqrt{Mr(1-r)}} \approx \frac{1}{\sqrt{Mr}}.$$

For example, if you run the SWC for 1 million steps and the important region containing the hills represents one percent of the total then p should be set to 0.01. Normally, we will not know r in advance, but have to make a guess. However, if for example one's guess for r is two orders of magnitude too high, then, due to the square root, p will only be too low by one order of magnitude. Note also that in the simulations described in the next section we chose p in the range of 10^{-4} to 10^{-1} and no matter which value we picked the SWC always performed better than the LPC.

4.3 Simulations

4.3.1 A two dimensional distribution with four main modes

In the first simulation we use local-proposal and small-world chains to explore the probability distribution shown in Figure 4.2. The underlying space S is a 2-dimensional integer grid $\{1, \ldots, 5000\} \times \{1, \ldots, 5000\}$.

For both types of chains, the local proposal was simply proposing one of the 8 surrounding neighbours with equal probability. For the SWC we occasionaly proposed a random point of the grid. Figure 4.3 shows the results for three runs of the LPC and three

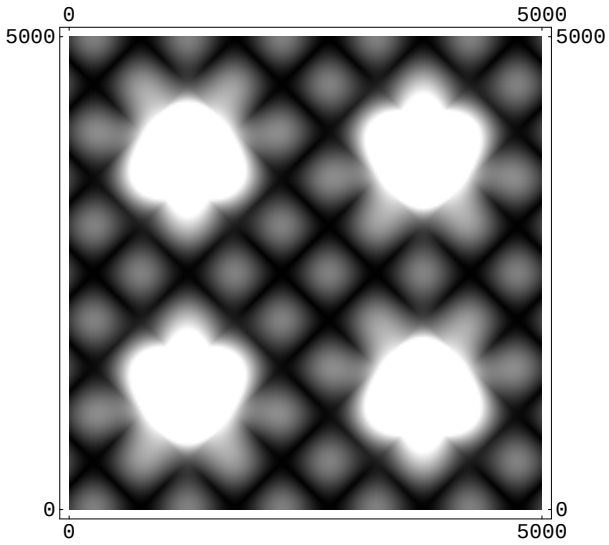

Figure 4.2: The probability distribution used in the first simulation; the lightness of a pixel is proportional to its point mass.

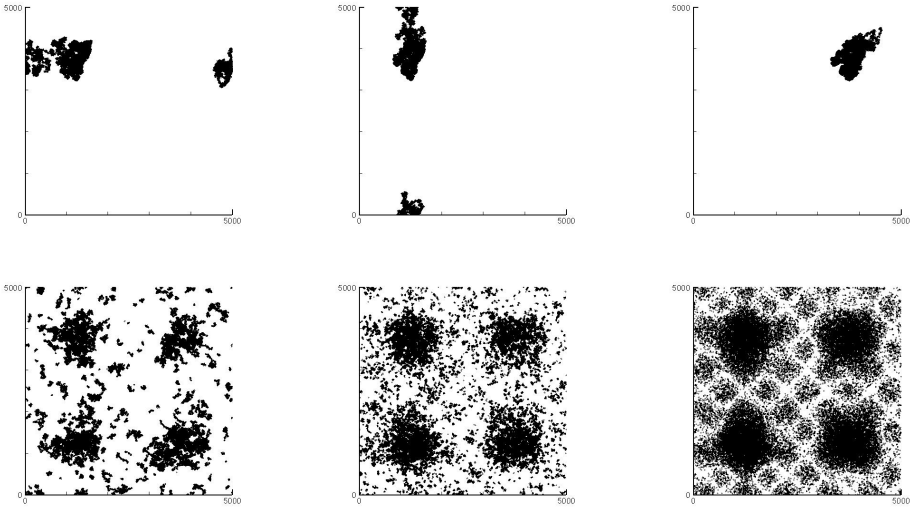

Figure 4.3: The results of the first simulation. The top row shows three runs of the LPC with 10^6 steps; the bottom row shows three runs of the SWC with 10^6 steps with different values of p (the leftmost being 0.0001, the middle one 0.001 and the rightmost 0.01).

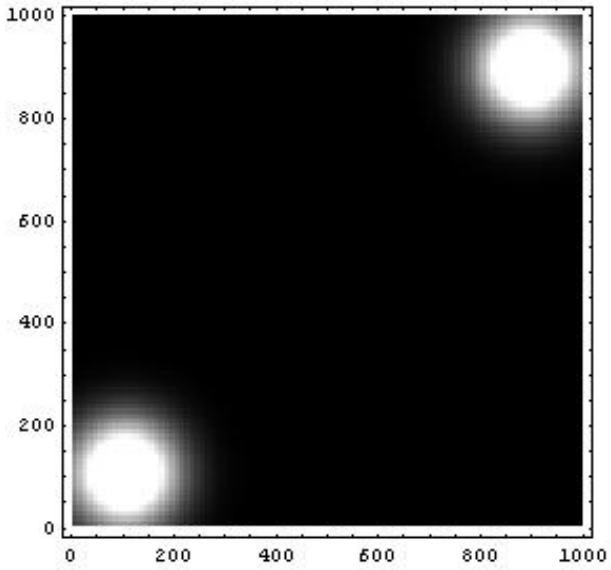

Figure 4.4: In the second simulation our probability distribution was the 10-dimensional equivalent of this figure.

runs of the SWC with varying p.

We can see that for this simulated distribution with four main modes, the LPC gets trapped within one peak, while the SWC surveys the whole distribution appropriately, the highest of the three values of p giving the best fit to the original distribution.

4.3.2 A high dimensional example

Here, we simulated a 2-modal mixed normal distribution over the grid $\{0, 1, ..., 999\}^{10}$. Thus, the state space had 10^{30} points, which should not be manageable for any numerical method. Figure 4.4 illustrates the situation with the corresponding 2-dimensional case.

Again, we sampled the space with both LPC and SWC. For the LPC, the proposals were randomly chosen from the $3^{10} - 1$ neighbours of the current state, while the SWC proposed a random state within the whole space with $p = 0.01$. To compare the results between LPC and SWC, we simply counted the sample points in the two peak regions: one being the set $A = (0, 200)^{10}$ and the other one being the set $B = (800, 1000)^{10}$. We

SWC		LPC	
visits to set A	visits to set B	visits to set A	visits to set B
2118493	2079152	0	4068403
2820473	1509184	0	4056310
2728981	1536989	0	4047301
3741123	689508	4497207	0
3374184	934000	0	4069861
3047729	1314131	4452966	0
2462776	1697934	0	4027776
2562090	1696232	0	4096586
2501613	1719777	4532164	0
3285900	1035378	0	4103339

Table 4.1: The results of the second simulation showing the number of times the SWC (left column) and the LPC (right column) visited region A and B of the state space. The rows correspond to ten independent runs.

ran 10 independent LPC and SWC. Each run had 10^8 steps, yet only every tenth step was recorded. Table 4.1 shows the results.

The left column is the result for the SWC. We can see that in each run the chain explores both peaks. The right column on the other hand is the result for the LPC. We can see that in each run the chain either gets trapped in set A or in set B.

4.3.3 SWC in infinite spaces

In order to demonstrate that the applicability of small-world proposals is not limited to the discrete case and finite spaces, we sampled a mixture of two normal distributions on $\mathbb{R}^4$, namely the density
$\frac{1}{2}f(x_1)f(x_2)f(x_3)f(x_4) + \frac{1}{2}g(x_1)g(x_2)g(x_3)g(x_4)$ where $f(x_i) \sim N(-10, 4)$ and $g(x_i) \sim N(10, 4)$. As in the previous simulation, we ran 10 independent LPC and SWC, each for 10^8 steps, and every tenth step was recorded. The local proposal incremented each component of the current point with a number drawn from $N(0, 0.5)$, while the wild proposals of the SWC used increments drawn from a Cauchy distribution whose full width at half maximum was 20. The frequency p of these wild proposals was set to 0.1. Table 4.2 summarizes the results of this simulation. Just like in the discrete case, each of the

SWC		LPC	
visits to A	visits to B	visits to A	visits to B
2484103	1937147	4419221	0
2239322	2179364	0	4430577
2240084	2197521	4417074	0
2179757	2231809	0	4432532
2147068	2268633	0	4426766
2272973	2151260	4417081	0
2228271	2199696	0	4414427
2296054	2123406	0	4421527
2191123	2222683	4430064	0
2297794	2107059	4425759	0

Table 4.2: The results of the third simulation showing the number of times the SWC (left column) and the LPC (right column) visited regions A and B of the state space, where A and B are hyper-balls with radius 3 centered at (-10,-10,-10,-10) and (10,10,10,10) respectively. The rows correspond to ten independent runs.

LPC got stuck at one of the peaks whereas the SWC always managed to explore both of them.

4.3.4 An example of a distribution with traps

The probability distribution used in our fourth simulation is shown in Figure 4.5. This distribution has one main hill in the center and four heaps at each corner surrounded by an almost null recurrent region. As the chance for an LPC to find the heaps is extremely small, we used only SWC in this simulation. We ran 10 independent chains for 10^7 steps, with $p = 0.1$ and recorded every tenth sample point.

The results of this simulation are collected in Table 4.3. All five hills were found in each run.

4.3.5 SWC in a heterogeneous space

To see how well an SWC performs in very heterogeneous spaces we took a picture which is often used in image processing, converted it to gray scale and increased the contrast by taking the fourth power of each gray level. Then we ran both LPC and SWC

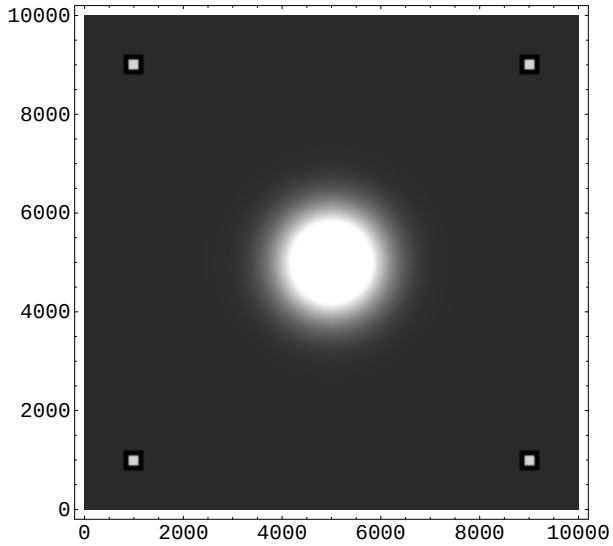

Figure 4.5: The two-dimensional distribution used in the fourth simulation. The whole space has size 10000×10000.

A	B	C	D	Center
62194	60062	59019	64989	504718
52897	55617	66427	58123	478288
57943	68062	57264	68815	508021
63832	55247	57274	62762	497264
53808	63851	60561	50230	467570
54432	61555	64119	64871	455125
58708	72168	63017	66713	491471
64462	60580	61251	60949	453775
70234	62663	67625	55654	489553
55683	60116	66087	53879	460234

Table 4.3: The results of the fourth simulation showing the number of times the SWC visited the four heaps (columns A-D) and the central hill ("Center") of the probability space shown in Figure 4.5. The rows correspond to ten independent runs.

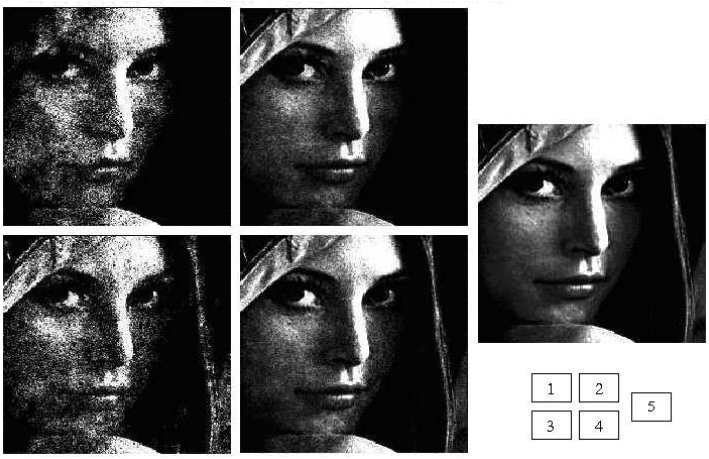

Figure 4.6: Results of sampling Lena. Top row: the results for the LPC. Bottom row: the results for the SWC. The original picture is displayed on the right-hand side. Figures 4.6.1 and 4.6.3 correspond to 50000 steps and Figures 4.6.2 and 4.6.4 to 10^7 steps, respectively.

to sample the picture. Again, p was set equal to 0.1 and every tenth sample point was recorded. The results as well as the original image are shown in Figure 4.6.

Although the difference between the LPC and SWC samples are subtle, one can see, that the image reconstructed with the LPC lacks some details like, for example, the highlight on the hair. One should also note that the SWC image is already very detailed after only 50000 steps.

4.3.6 SWC and importance sampling

Our final example illustrates how small-world proposals can be used in the context of importance sampling (Hastings 1970). Suppose we want to estimate the expectation $E_\pi(f(X))$ of a random variable $f(X)$ with respect to a probability measure π. Although Monte Carlo integration (Rice 1994), i.e., drawing a sample from π and averaging the obtained values of $f(X)$, is one way to achieve this goal, the resulting estimate might have a high variance since $f(x)$ might have maxima where $\pi(x)$ is small. Therefore, it is advisable (Press, Flannery, Teukolsky, and Vetterling 1992) to sample not from π but

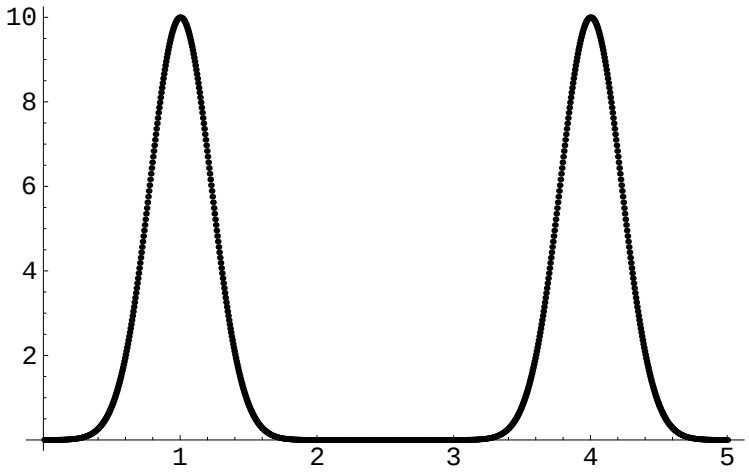

Figure 4.7: The function $f(x) = 10(e^{-10(x-1)^2} + e^{-10(x-4)^2})$ to be integrated in Section 4.3.6.

from a new distribution $\pi' \propto f\pi$ which gives an appropriate weight both to the function f and to the original distribution π. Assuming, without loss of generality, that $f > 0$, one can easily check that $E_\pi(f(X)) = 1/E_{\pi'}(1/f(X))$. Hence we can use the following estimator

$$\widehat{E_\pi(f(X))} = \frac{N}{\sum_{t=1}^{N} \frac{1}{f(X(t))}}, \tag{4.14}$$

where the $X(t)$ are taken from π'. Notice, however, that this importance sampling will in many cases require a chain to sample from a multi-modal space. Thus, its wide use has been restricted due to the limitations of local-proposal chains. Small-world chains, on the other hand, should be well suited for this problem.

In our final simulation we therefore compared the performance of local-proposal chains sampling directly from π and of small-world chains which sampled from $(f(x)+1)\pi(x)$ (see Figure 4.9) for the function $f(x)$ plotted in Figure 4.7 and the $\pi(x)$ shown in Figure 4.8. The reason for using $f+1$ instead of f is the high variance which might result from small values of f. This of course has to be taken into account when applying Equation (4.14). Each chain was run for 10^5 steps and every 100th step was recorded. The parameter p of the SWC was set to 0.1. We also ran simulations in which an LPC tried to sample from

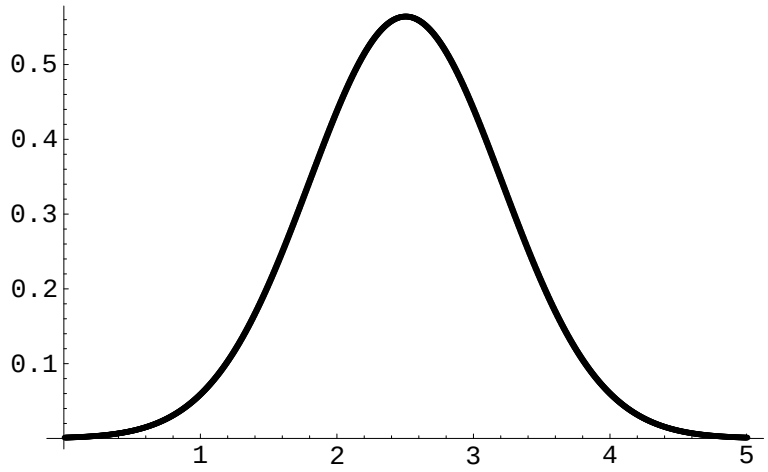

Figure 4.8: The density $\pi(x) \propto e^{-(x-2.5)^2}$ from which the LPC in Section 4.3.6 takes its sample.

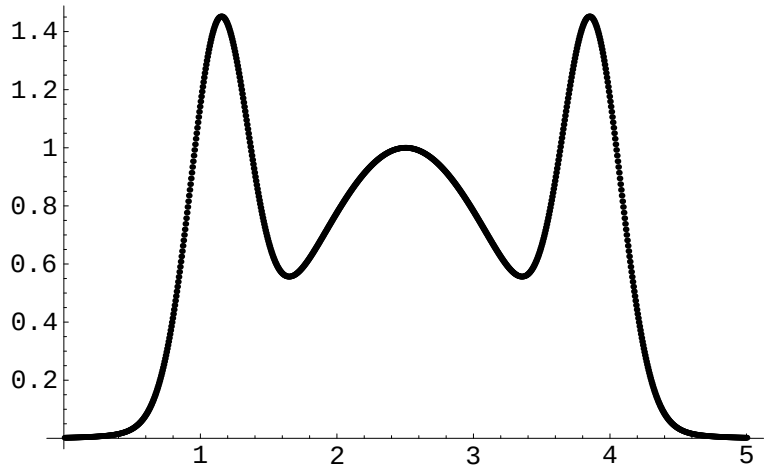

Figure 4.9: A plot of $(f(x)+1)\pi(x)$. The SWC in Section 4.3.6 samples from this function.

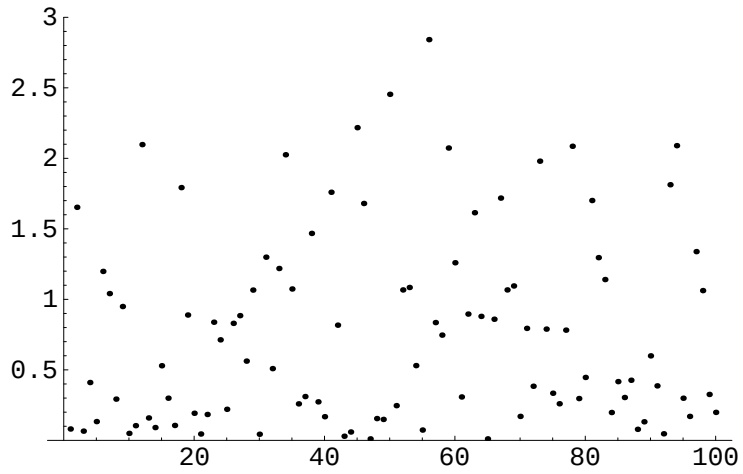

Figure 4.10: The estimates $\widehat{E_\pi(f(X))}$ in 100 independent runs of the LPC. The mean of the estimates is 0.770 with a standard deviation of 0.571.

$(f(x)+1)\pi(x)$, yet they always got stuck at one of the modes (data not shown).

Figures 4.10 and 4.11 show the estimated expectations for 100 independent runs of the LPC and the SWC respectively. While there is a considerable variation among the values obtained with the LPC, the estimates produced by the SWC are all very close to the true value which is approximately 0.780.

4.4 Discussion

The purpose of this chapter was to present a general yet simple idea for a proposal distribution that leads to a better convergence of MCMC. In all the examples in this chapter the small-world chains performed dramatically better than the chains which only relied on local proposals. While the local-proposal chains got stuck at one mode of the distribution (see Sections 4.3.1, 4.3.2 and 4.3.3) or missed important details (see Section 4.3.5), the small-world chains were even able to explore the extremely heterogeneous space of Section 4.3.4. The small-world chains' ability to sample multi-modal spaces also permitted the application of importance sampling in a case where local-proposal chains failed (see Section 4.3.6).

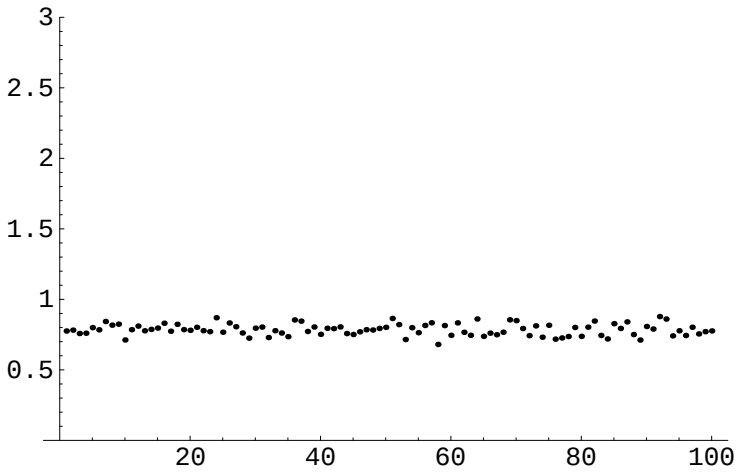

Figure 4.11: The estimates $\widehat{E_\pi(f(X))}$ in 100 independent runs of the SWC. The mean of the estimates is 0.785 with a standard deviation of 0.00897.

Through most of this chapter, we stuck to the mathematically tractable case of a Markov chain on a finite grid since this is also the case which is the most relevant one when implementing this algorithm. We also focused on the simplifying scenario that local proposals essentially behave like uniform random walks and that we can generate wild proposals by drawing a point from the space at random. Yet, as demonstrated in simulation 4.3.3, the small-world idea can also be applied when sampling from a continuous density on an unbounded space. There, one just has to use a sufficiently heavy-tailed distribution for the wild proposals like for example a wide Cauchy distribution. Taking a uniform random walk as exemplary local proposal instead of e.g. a normal random walk is also unproblematic since the central limit theorem guarantees that a uniform random walk and a normal random walk will be very similar local proposal distributions. Hence, replacing the uniform local proposals with the more typical normal proposal should not affect the principal results. We are not advocating exchanging every thinkable local proposal for a mixture of two uniform distributions, but rather relying on occasional wild proposals as a way to improve on any local proposal scheme. Therefore, we did not investigate the performance differences which still might exist between different types of local proposals.

Of course, there exists a great variety of other methods which try to ensure convergence of MCMC [cf. Tierney (1994), Chib and Greenberg (1995), Gilks, Richardson, and Spiegelhalter (1996)]. For example, we might know enough about the target distribution $\pi(x)$ so that we can tailor our proposal to this target (Chib, Greenberg, and Winkelmann 1998). Yet, in many cases that information will not be available.

Another widespread practice is running multiple chains which start from different points of the space. Although this has a superficial resemblance with our idea, exploring the space with multiple chains is in fact equivalent to an incorrect implementation of a single SWC. Suppose for example that an investigator decides to run 100 chains each of length 1 million, starting each chain at a random location. This is the same as running a single chain of length 100 million where every 1 million steps one proposes a wild move and accepts that move with probability 1. So, rather than proposing a wild move with a certain probability, one deterministically decides when to propose a wild move and rather than accepting the move with a certain probability, one always accepts it. These moves violate the assumptions of a homogeneous Markov chain and hence make it unlikely that running multiple chains produces samples from the stationary distribution unless stationarity is reached before the first wild move, in which case the problem is easy and both SWC and multiple chains are unnecessary. As Geyer points out (Geyer 1992; see also `http://www.stat.umn.edu/~charlie/mcmc/one.html`), if the problem is hard, then many short chains are likely to be sampling something closer to the initial distribution from which the starting points were chosen than the stationary distribution $\pi(x)$ of interest. Notice that the SWC accepts wild moves according to the details of the space being explored, as it should, mainly moving between hills or sampling the heavy tails and not returning to the flat regions. Whereas the multiple chain approach, does not make use of any information about the space when it effectively starts the chain over, which will typically be in a flat region. For this reason, a SWC of length N must do better at exploring the space than m multiple chains each of length n with $N = mn$. The only

advantage to multiple chains is that they can be run in parallel.

A method which uses multiple chains and indeed samples from the right distribution is the Metropolis-coupled MCMC (Geyer 1991). There, only one of the chains — the one whose states are recorded — explores π while the others run through flattened versions of that landscape. After every step one checks if it is worthwhile to recombine the chains. If we used this method with only two chains, one which samples from π and one which runs through a completely flattend landscape and if we only attempted to swap chains every $1/p$-th step, this algorithm should behave in the same way as SWC, be it at a higher computational burden.

Another method similar to SWC results from the mixing of transition kernels (Larget and Simon 1999). There, instead of having one expression for the Metropolis-Hastings ratio where the proposal probabilities can be written as a mixture, one has to decide before every step which proposal distribution to use and then only this one goes into the Metropolis-Hastings ratio. This mixing of the transition kernels will indeed give the same result as the mixing of the proposal distributions in the case of symmetric proposals. However, as already mentioned above, the vital point of our approach is not the usage of a mixture as proposal distribution, but the incidental wild jumps. A simple geometric view may give some insight into why the small-world proposals are so effective: If we think of the original space as the inside surface of a large sphere, then a small-world chain's wild proposals tie together distant regions of the sphere. Since these wild links are not static, we get an evolving geometry in which we occasionally stretch two distant points inside the sphere until they touch, then make our jump and let the surface snap back to its spherical shape. While a small-world chain moves through this evolving geometry, it is more likely to jump to another top of a hill than to jump into a valley. Therefore, the hills of the space will behave like neighbors. In that sense, small-world proposals rearrange the space that they are sampling.

It should be noted that this improvement was achieved in an almost automatic way.

The only parameter that has to be set in advance is p, the probability of wild proposals and as we have seen in Section 4.2.4 we do not need detailed knowledge of the space in order to produce a reasonable choice of p.

We do not allege that small-world chains always perform better than any other MCMC method, but since using small-world proposals is not more difficult than using local proposals, we are convinced that our method can be a simple and valuable add-on to any of the other methods. [1]

[1] The results presented in this chapter were joint work of the author and his advisor with Roland Fleißner and Paul Joyce when Roland was doing a Postdoc with Paul at U of I. The original publication (Guan et al. 2006) is available at `www.springerlink.com`

Chapter 5

Markov Chain Monte Carlo Convergence Rates

5.1 Introduction and main results

Many applications of Markov chain Monte Carlo (MCMC) involve very large and/or complex state spaces, and convergence rates are an important issue. A major problem in MCMC is thus to find sampling schemes whose mixing times do not grow too rapidly as the size or complexity of the space is increased. In Chapter 4, computer simulations were used to show that such problems can be handled simply and efficiently by using an idea from "small-world networks" (Watts and Strogatz 1998) to make a slight change in a given proposal scheme. This change amounts to augmenting a typical local proposal distribution with low probability long-distance jumps that effectively contract the space and lead to much faster convergence to multi-modal target distributions. In this chapter, we make rigorous comparisons of the convergence rates of these two types of chains.

Let π be a multi-modal probability measure on a convex set $\Omega \subseteq \mathbb{R}^n$. We wish to compare convergence rates to this measure by two different Metropolis–Hastings chains that are characterized by their proposal distributions: "local" and "small world." From now on, we refer to these two types of Markov chains as "local chains" and "small-world chains," respectively. Intuitively, a local proposal distribution is one that has thin tails, so that the mean distance of a proposed move away from the current state is small compared to the distances between modes; by a small-world proposal we mean a mixture of a local proposal and a heavy-tailed proposal, so that the mean distance of a proposed move away

from the current state is large.

In a multi-modal space, a local chain will equilibrate rapidly within a mode, but takes a long time to move from one mode to another. Hence the entire chain converges slowly to the target distribution. However, a small fraction of heavy-tailed proposals enables a small world chain to move from mode to mode much more quickly. While this reduces the efficiency of equilibrating within a mode, it is a small price to pay and easily outperforms purely local proposals. This is the spirit of our main results. We derive bounds on the spectral gaps for such local and small world chains and hence show how a small fraction of heavy-tailed proposals can turn a slowly mixing chain into a rapidly mixing chain.

Throughout this chapter, we assume the state space Ω is equipped with two measures: a reference measure, taken to be Lebesgue measure μ, and a Borel probability measure π which serves as the target distribution. Suppose π is absolutely continuous with respect to μ so that it admits a density $\pi(x)$:

$$\pi(B) = \int_B \pi(x)\mu(dx).$$

The most widely used Markov chain Monte Carlo method is the Metropolis–Hastings algorithm (Metropolis et al. 1953; Hastings 1970), which we now describe briefly.

5.1.1 Metropolis–Hastings algorithm

A transition probability kernel $P(x, dy)$ corresponds to a Metropolis–Hastings Markov chain on Ω if it is of the form

$$P(x, dy) = \alpha(x, y)\, k(x, y)\, \mu(dy) + r(x)\, \delta_x(dy), \tag{5.1}$$

where $k(x, y)$ is the *proposal distribution* and we say $k(x, y)$ *induces* $P(x, dy)$,

$$\alpha(x, y) = \min\left(\frac{\pi(y)\, k(y, x)}{\pi(x)\, k(x, y)}, 1\right)$$

is the *acceptance probability* of a proposed move, δ_x is the unit point mass at x, and

$$r(x) = \int_\Omega (1 - \alpha(x, y))k(x, y)\mu(dy)$$

is the probability that the proposed move from x is rejected. It is easy to check that the transition kernel $P(x, dy)$ satisfies the detailed balance equation $\pi(dx)\, P(x, dy) = \pi(dy)\, P(y, dx)$ as measures on $\Omega \times \Omega$, so that $P(x, dy)$ is reversible with respect to π and hence has π as its unique invariant measure. For simplicity, we consider only *symmetric* proposal distributions, $k(x, y) = k(|x - y|)$, in which case the acceptance probability simplifies to $\alpha(x, y) = \min\left(\frac{\pi(y)}{\pi(x)}, 1\right)$.

5.1.2 Geometric ergodicity and spectral gap

Let $L^2(\pi)$ denote the space of (Borel) measurable, complex functions on Ω satisfying

$$\int_\Omega |f(x)|^2 \pi(dx) < \infty.$$

This is a Hilbert space with inner product $\langle f, g\rangle = \int_\Omega f(x)\overline{g(x)}\pi(dx)$ and norm $\|f\| = \langle f, f\rangle^{\frac{1}{2}}$. The Metropolis–Hastings kernel $P(x, dy)$ induces a contraction operator P on $L^2(\pi)$ given by $Pf(x) = \int_\Omega f(y)P(x, dy)$. We say the operator P is *induced* by a proposal distribution $k(x, y)$ if the same is true of its transition kernel. $P(x, dy)$ being reversible with respect to π is equivalent to the operator P being self-adjoint, i.e.,

$$\langle Pf, g\rangle = \langle f, Pg\rangle, \qquad f, g \in L^2(\pi).$$

It is well known that the spectrum of P is a subset of $[-1, 1]$. (P being self-adjoint implies its spectrum is real, and $P(x, dy)$ being a transition probability kernel determines the range.)

A chain is $L^2(\pi)$-geometrically ergodic if there exists $\gamma < 1$ such that

$$\|\mu_0 P^n - \pi\| \leq \gamma^n \|\mu_0 - \pi\| \tag{5.2}$$

for any non-negative integer n and any probability measure $\mu_0 \in L^2(\pi)$ (i.e., $\mu_0 \ll \pi$ with $\int |\frac{d\mu}{d\pi}|^2 d\pi < \infty$). Roberts and Tweedie (2001) have shown that convergence in L^2 implies convergence in "total variation" norm

$$\|\mu_1 - \mu_2\|_{\text{tv}} = \sup_{A \subset \Omega} |\mu_1(A) - \mu_2(A)| = \frac{1}{2} \int_\Omega |f_1(x) - f_2(x)| \, dx,$$

where $f_i(x) = d\mu_i/dx$.

Let $L_0^2(\pi)$ denote the orthogonal complement of the constant function $\mathbf{1}$ in $L^2(\pi)$:

$$L_0^2(\pi) = \{f \in L^2(\pi) : \langle f, \mathbf{1} \rangle = \int_\Omega f(x)\pi(dx) = 0\}.$$

Clearly, as a subspace of $L^2(\pi)$, $L_0^2(\pi)$ is also a Hilbert space. Denote by P_0 the restriction of P to $L_0^2(\pi)$. Chan and Geyer (1994) proved that, for a geometrically ergodic chain, P_0 has no point spectrum (i.e., eigenvalues) of value ± 1. In addition, it has been shown (Roberts and Rosenthal 1997; Roberts and Tweedie 2001) that for reversible Markov chains, geometric ergodicity is equivalent to the condition

$$\|P_0\| \equiv \sup_{f \in L_0^2(\pi), \|f\| \leq 1} \|P_0 f\| < 1, \tag{5.3}$$

and any $\gamma \in [\|P_0\|, 1)$ satisfies Equation (5.2). The *spectral gap* of the chain P is defined by

$$\mathfrak{gap}\,(P) = 1 - \|P_0\|.$$

Thus the spectral gap provides a measure of the speed of convergence of a Markov chain to its stationary measure. Two of the main tools for studying spectral gaps in the

setting of MCMC are conductance and Cheeger's inequality, to which we now turn.

5.1.3 Conductance and Cheeger's Inequality

Let P be a Markov transition kernel that is reversible with respect to π. For $A \subseteq \Omega$ with $\pi(A) > 0$, define

$$\mathfrak{h}_P(A) = \frac{1}{\pi(A)} \int_A P(x, A^c)\pi(dx). \tag{5.4}$$

The quantity $\mathfrak{h}_P(A)$ can be thought of as the (probability) flow out of the set A in one step when the Markov chain is at stationary. Notice that $\pi(dx)/\pi(A)$ is the conditional stationary measure on the set A.

The *conductance* of the chain is defined by

$$\mathfrak{h}_P = \inf_{0<\pi(A)\leq 1/2} \mathfrak{h}_P(A). \tag{5.5}$$

Note that $0 \leq \mathfrak{h}_P \leq 1$. Intuitively, small $\mathfrak{h}_P$ implies that the chain can become stuck for a long time in some set whose measure is at most $1/2$, making it difficult for the chain to sample the rest of the distribution. As a result, such a chain converges slowly to the stationary measure. On the other hand, a large $\mathfrak{h}_P$ implies that the chain travels around swiftly and hence samples different parts of the distribution efficiently. As a result, such a chain converges rapidly. Lawler and Sokal (1988) have quantified this as a generalization of Cheeger's inequality.

Theorem 5.1.1 (Cheeger's Inequality)**.** *Let P be a reversible Markov transition kernel with invariant measure π. Then*

$$\frac{\mathfrak{h}_P^2}{2} \leq \mathfrak{gap}\,(P) \leq 2\,\mathfrak{h}_P. \tag{5.6}$$

Next, suppose that a proposal distribution $k(x, y)$ is a mixture of two proposal distributions $k_1(x, y)$ and $k_2(x, y)$. That is $k(x, y) = (1 - s)k_1(x, y) + s\,k_2(x, y)$, for some

$0 \le s \le 1$. Suppose operators P, P_1, and P_2 are induced by $k(x,y)$, $k_1(x,y)$, and $k_2(x,y)$, respectively. Clearly,

$$P = (1-s)P_1 + s\,P_2 \tag{5.7}$$

and, for any measurable set A, $\mathfrak{h}_P(A) = (1-s)\,\mathfrak{h}_{P_1}(A) + s\,\mathfrak{h}_{P_2}(A)$. As an immediate consequence we have the following lemma showing that conductance acts like a concave function on transition kernels and the spectral gap can be bounded from below by one of the components.

Lemma 5.1.2. *Suppose a reversible chain has a mixture kernel defined by* (5.7)*. Then the conductance of the chain satisfies* $\mathfrak{h}_P \ge (1-s)\mathfrak{h}_{P_1} + s\,\mathfrak{h}_{P_2}$*. In addition,*

$$\mathfrak{gap}\,(P) \ge \frac{1}{2}(1-s)^2\,\mathfrak{h}_{P_1}^2. \tag{5.8}$$

Proof. From (5.5),

$$\begin{aligned}
\mathfrak{h}_P &= \inf_{0<\pi(A)\le 1/2} ((1-s)\mathfrak{h}_{P_1}(A) + s\,\mathfrak{h}_{P_2}(A)) \\
&\ge (1-s) \inf_{0<\pi(A)\le 1/2} \mathfrak{h}_{P_1}(A) + s \inf_{0<\pi(B)\le 1/2} \mathfrak{h}_{P_2}(B) \\
&= (1-s)\mathfrak{h}_{P_1} + s\,\mathfrak{h}_{P_2} \ge (1-s)\mathfrak{h}_{P_1}.
\end{aligned}$$

Combine this with Cheeger's inequality (5.6) to get (5.8). □

5.1.4 Definitions and main results

Let $|\cdot|$ be a norm on $\Omega \subseteq \mathbb{R}^n$ and $B_r(x)$ the n-dimensional ball centered at x with radius r. Denote by $\partial B_r(x)$ the surface of the ball, and write $\pi^+(\partial A)$ for the surface measure (relative to π) of a set A in the sense that

$$\pi^+(\partial A) = \liminf_{\varepsilon \to 0} \frac{\pi(A^\varepsilon) - \pi(A)}{\varepsilon},$$

where $A^\varepsilon = \{x \in \Omega : \exists a \in A, |x - a| < \varepsilon\}$ is the ε-neighborhood of A, consisting of the union of A and its "ε-boundary" $A^\varepsilon \setminus A$.

The *barycenter* of a probability measure π is $\beta = \int_\Omega x\, \pi(dx)$. We say the measure π is *log-concave* if it has a density with respect to μ of the form $\pi(x) = \exp(-V(x))$, where $V : \Omega \to (-\infty, +\infty]$ can be an arbitrary convex function. Examples of log-concave distributions include uniform, exponential, normal, gamma, etc. For technical reasons, we restrict our attention to "smooth" log-concave functions (but see discussion at the end of Section (5.3)). We say a log-concave function $\exp(-V(x))$ is α-*smooth* if for any x, y, we have $|V(x) - V(y)| < \alpha\, |x - y|$. By Borell's theorem (Borell 1974), the tail of $\pi(x)$ is exponentially deceasing, i.e., there is a number $\nu_\pi > 0$, such that $\pi^+(\partial B_r(\beta)) \leq c\, \exp(-\nu_\pi\, r)$, for some constant c. We will refer to ν_π as a *decay exponent* for π. Define the first absolute centered moment of π as $M_\pi = \int_\Omega |x - \beta| \pi(dx)$.

Next, we characterize the multi-modal distributions that will serve as our target distributions. Let $\Omega = A_1 \cup \cdots \cup A_m$ be a partition of state space Ω into disjoint convex subsets. Suppose concentrated on each A_i we have a single α-smooth log-concave probability measure π_i with decay exponent ν_{π_i} and barycenter $\beta_i \in A_i$. Let $d_{ij} = |\beta_i - \beta_j|,\ i \neq j$, denote the pairwise distances between barycenters. The target distribution of interest is then defined as a mixture of these log-concave densities:

$$\pi(x) = \sum_{i=1}^{m} c\, \pi_i(x) 1_{A_i}(x), \tag{5.9}$$

where c is a normalization constant and 1_{A_i} is the indicator function of A_i. When the modes have different smoothness parameters, we take α to be the largest such.

We will refer to features of the above probability measure π that present barriers to mixing in the local Metropolis–Hastings chain as the "complexity of the target distribution." These include $\mu(\Omega)$ (if $\mu(\Omega) < \infty$), d_{ij}, and ν_{π_j}. In particular, we say a given chain is *slowly mixing in the complexity of* π if the spectral gap of the chain is an exponentially

decreasing function of at least one of these quantities. We say a chain is *rapidly mixing in the complexity of* π if the spectral gap is a polynomially decreasing function of all of these quantities.

To make our calculations concrete, we will always use for our symmetric local proposal distribution $k(x, y)$ a uniform distribution on an n-dimensional ball with radius δ. Such a proposal distribution captures the essence of "local proposals" and is easier to handle compare to other light-tailed proposals. We will sometimes refer to such a local proposal scheme as a "δ-ball walk."

Let $h(x, y)$ be a heavy-tailed distribution, i.e., one for which the tails decrease polynomially, instead of exponentially, on Ω. (We shall restrict ourselves to Cauchy distributions when Ω is unbounded, and uniform distributions when Ω is compact.) A *small-world proposal* distribution $g(x, y)$ is a mixture of a local and a heavy-tailed distributions:

$$g(x, y) = (1 - s)\, k(x, y) + s\, h(x, y), \tag{5.10}$$

for some $s \in (0, 1)$.

We are now ready to state our main result:

Theorem 5.1.3. *Let π be the multi-modal probability measure defined by* (5.9) *with α-smooth log-concave modes. Let $k(x, y)$ be the local proposal distribution and let $g(x, y)$ be defined by* (5.10)*, where $h(x, y)$ is a heavy-tailed proposal. Then the local Metropolis–Hastings chain is "slowly mixing" and the small-world chain is "rapidly mixing" in the complexity of π.*

The rest of the chapter is organized as follows. In the next section, we prove a new version of the state decomposition theorem of Madras and Randall (2002). This will play an important role in proving our main theorem. On each log-concave piece, an upper bound on conductance is easy to obtain. However, the lower bound requires some extra work. Thus we devote Section 5.3 to finding a lower bound through an isoperimetric

inequality for log-concave probability measures. The proof of the main theorem is given in Section 5.4. A discussion of possible applications of our result to convergence rates in Metropolis-coupled Markov chain Monte Carlo will conclude the chapter.

5.2 State decomposition theorem

In this section we state and prove a new version of the state decomposition theorem of Madras and Randall (2002). The setup of the new theorem is the same as that of their paper, but we repeat it here for convenience. Recall that $\{A_1, \ldots, A_m\}$ is a partition of Ω. We describe the "pieces" of a Metropolis–Hastings chain P by defining, for each $i = 1, \ldots, m$, a new Markov chain on A_i that rejects any transitions of P out of A_i. The transition kernel P_{A_i} of the new chain is given by

$$P_{A_i}(x, B) = P(x, B) + 1_B(x)P(x, A_i^c) \quad \text{for } x \in A_i, B \subset A_i. \tag{5.11}$$

It is easy to see that P_{A_i} is reversible on the state space A_i with respect to the measure π_i, which, by definition, is the restriction of π to the set A_i.

The movement of the original chain among the "pieces" can be modeled by a "component" Markov chain with state space $\{1, \ldots, m\}$ and transition probabilities:

$$P_H(i, j) = \frac{1}{2\,\pi(A_i)} \int_{A_i} P(x, A_j)\pi(dx), \quad \text{for } i \neq j, \tag{5.12}$$

and $P_H(i, i) = 1 - \sum_{j \neq i} P_H(i, j)$. This definition is quite similar to the definition of the quantity $\mathfrak{h}_P(A)$ except for the 2 in the denominator. The reason for this factor will become clear as we progress.

Our theorem is more or less a direct application of the following lemma, which is due to Caracciolo, Pelissetto, and Sokal, and was recorded, together with its proof, in Madras

and Randall (2002) as Theorem A.1.

Lemma 5.2.1 (Caracciolo, Pelissetto, and Sokal)**.** *In the setting stated at the beginning of this section, assume that $P(x, dy)$ and $Q(x, dy)$ are transition kernels that are reversible with respect to π. Assume further that Q is nonnegative definite and let $Q^{\frac{1}{2}}$ denote its nonnegative square root. Then*

$$\mathfrak{gap}\,(Q^{\frac{1}{2}}PQ^{\frac{1}{2}}) \geq \mathfrak{gap}\,(\overline{Q})(\min_{i=1,\ldots,m} \mathfrak{gap}\,(P_{A_i})), \tag{5.13}$$

where

$$\overline{Q}(i,j) = \frac{1}{\pi(A_i)} \int_{A_i} Q(x, A_j)\pi(dx), \quad \text{for } i \neq j,$$

and $\overline{Q}(i,i) = 1 - \sum_{j\neq i} \overline{Q}(i,j)$.

Theorem 5.2.2 (State Decomposition Theorem)**.** *In the preceding framework, as given by equations* (5.11) *and* (5.12)*, we have*

$$\mathfrak{gap}\,(P) \geq \frac{1}{2}\mathfrak{gap}\,(P_H)(\min_{i=1,\ldots,m} \mathfrak{gap}\,(P_{A_i})). \tag{5.14}$$

Remark 1. The theorem says the spectral gap for the whole Metropolis–Hastings chain can be bounded below by taking into account the mixing speed within each mode and the mixing speed between different modes.

Proof of Theorem 5.2.2. Let $Q = \frac{1}{2}(I+P)$, where I is the identity kernel. Reversibility of Q with respect to π follows from the same property for P. To see that Q is a non-negative definite (and hence can be used in Lemma 5.2.1), note first that since P is a self-adjoint probability operator, its spectrum is a subset of $[-1, 1]$ and hence $\|P\| \leq 1$. Thus,

$$\langle Qf, f\rangle = \langle \frac{1}{2}(I+P)f, f\rangle = \frac{1}{2}(\langle f, f\rangle + \langle Pf, f\rangle) \geq \frac{1}{2}(1 - \|P\|)\|f\|^2 \geq 0.$$

Since

$$Q^{\frac{1}{2}}PQ^{\frac{1}{2}} + Q = Q^{\frac{1}{2}}PQ^{\frac{1}{2}} + Q^{\frac{1}{2}}IQ^{\frac{1}{2}} = Q^{\frac{1}{2}}(P+I)Q^{\frac{1}{2}} = Q^{\frac{1}{2}}(2Q)Q^{\frac{1}{2}} = 2\,Q^2,$$

it follows that

$$Q^{\frac{1}{2}}PQ^{\frac{1}{2}} = 2Q^2 - Q = Q(2Q - I) = QP.$$

Furthermore, setting $\gamma = \|P_0\|$, we have $\mathfrak{gap}\,(P) = 1 - \gamma$ and, as a simple consequence of the spectral mapping theorem, $\mathfrak{gap}\,(QP) = 1-(1/2)\gamma(1+\gamma)$. Thus $2\,\mathfrak{gap}\,(P)-\mathfrak{gap}\,(QP) = 2(1-\gamma) - (1-(1/2)\gamma(1+\gamma)) = (1-\gamma)(1-\gamma/2) > 0$, and hence

$$\mathfrak{gap}\,(P) > \frac{1}{2}\mathfrak{gap}\,(QP) = \frac{1}{2}\,\mathfrak{gap}\,(Q^{\frac{1}{2}}PQ^{\frac{1}{2}}). \tag{5.15}$$

Following the definition in Lemma 5.2.1, we have

$$\overline{Q}(i,j) = \frac{\int_{A_i} Q(x,A_j)\pi(dx)}{\pi(A_i)} = \frac{\int_{A_i}(I(x,A_j)+P(x,A_j))\pi(dx)}{2\,\pi(A_i)} = \frac{\int_{A_i} P(x,A_j)\pi(dx)}{2\,\pi(A_i)}, \tag{5.16}$$

which is just $P_H(i,j)$.

Combine Equations (5.12), (5.13), and (5.15) to finish the proof. □

The same result has been obtained in Martin and Randall (2000). However, their proof was not applicable in the general situation for which P is not non-negative definite.

There is, of course, a resemblance between our state decomposition theorem and that of Madras and Randall (2002). We note that, firstly, our conclusion appears to be a bit stronger than theirs in that our result does not depend on the number of overlapping "pieces"; secondly and more importantly, in the original theorem the connection between different "pieces" of the state space is made via overlapping of the different "pieces". Jarner and Yuen (2004) have applied the original theorem to estimate the convergence rates of 1-dimensional local chains. Unfortunately, the original theorem is not readily

applicable to small-world chains because such chains can move from one region to another even when the two regions are not overlapping. On the other hand, in our theorem the connection between different "pieces" is made via the "probability flow" from one region to another. We emphasize that having a chain that jumps from one region to another without visiting the valleys in between is the key to sampling a multi-modal space efficiently. This is discussed in Chapter 4.

5.3 Lower bound for conductance

To apply the state decomposition theorem to a multi-modal probability measure defined by (5.9), we need a lower bound on the conductance (hence spectral gap) for each log-concave piece of the distribution. For this, we use an isoperimetric inequality.

The idea of using an isoperimetric inequality for log-concave probability measures to obtain a lower bound on the conductance of local chains is rather straightforward and has been used by many authors, including Applegate and Kannan (1990), Kannan and Li (1996), and Lovász and Vempala (2003a). Isoperimetric inequalities for log-concave probability measures have been studied by Bobkov (1999) and Kannan, Lovász, and Simonovits (1995). As noted by Bobkov, although the result presented in Kannan, Lovász, and Simonovits (1995) was for a uniform measure on a convex set, their method in fact extends naturally to general log-concave probability measures. The isoperimetric inequality in Kannan, Lovász, and Simonovits (1995) was studied using a "localization lemma" developed by Lovász and Simonovits (1993) which essentially reduces integral inequalities in an n-dimensional space to integral inequalities in a single variable. The original form of the result, applied to uniform measures, is the following, recorded as Theorem 5.2 in Kannan, Lovász, and Simonovits (1995).

Theorem 5.3.1 (Kannan, Lovasz, and Simonovits)**.** *Let K be a convex set and $K = K_1 \cup K_2 \cup K_3$ a partition of K into three measurable sets such that the distance between*

K_1 and K_2 is $d(K_1, K_2) > 0$. Let $b = \frac{1}{\mathrm{vol}(K)} \int_K x\,dx$ be the barycenter of K and $M_1(K) = \int_K |x - b|\,dx$. Then

$$\mathrm{vol}(K_3)\mathrm{vol}(K) \geq \frac{\ln 2}{M_1(K)} d(K_1, K_2)\mathrm{vol}(K_1)\mathrm{vol}(K_2).$$

The following is the log-concave version of the above isoperimetric inequality. See also Lovász and Vempala (2003b, Theorem 2.4).

Theorem 5.3.2. *Suppose π is a log-concave probability measure on a convex set K. Suppose further that π has barycenter 0 and set $M_\pi = \int_K |x|\,\pi(dx)$. Let $K = K_1 \cup K_2 \cup B$ be a partition of K into three measurable sets such that the distance between K_1, K_2 $d(K_1, K_2) > 0$. Then*

$$\pi(B) \geq \frac{\ln 2}{M_\pi}\, d(K_1, K_2)\, \pi(K_1)\, \pi(K_2).$$

As remarked above, the proof of Theorem 5.3.1 in Kannan, Lovász, and Simonovits (1995) extends to Theorem 5.3.2 via the "localization lemma" on log-concave probability measures (Kannan, Lovász, and Simonovits 1995, Theorem 2.7).

The next lemma makes the connection between Euclidean distance between two points and the total variation distance between the one-step Markov transition kernels starting from those two points. Both the idea and the proof are borrowed from Lovász and Vempala (2003b).

Lemma 5.3.3. *Let $K \subset \mathbb{R}^n$ be concave and suppose $u, v \in K$ satisfy $|u - v| < \frac{\delta}{8\sqrt{n}}$, for some $\delta > 0$. Suppose further that $P(x, dy)$ is a Metropolis–Hastings transition kernel induced by a δ-ball local proposal and having an α-smooth log-concave target distribution π on K. Then*

$$\|P(u, \cdot) - P(v, \cdot)\|_{\mathrm{tv}} \leq 1 - \frac{1}{2} e^{-\alpha\,\delta}.$$

Proof. Let $B_\delta(u)$ and $B_\delta(v)$ be the balls of radius δ around u and v, respectively. Write $\mathrm{vol}(B_\delta)$ for their Euclidean volume and set $C = B_\delta(u) \cap B_\delta(v)$. Since $|u - v| < \frac{\delta}{8\sqrt{n}}$,

we have $\mathrm{vol}(C) > \frac{1}{2}\mathrm{vol}(B_\delta)$. Since our target distribution is an α-smooth log-concave function, the Hastings ratio is of the form

$$\frac{\pi(y)}{\pi(x)} = e^{-|V(x)-V(y)|} \geq e^{-\alpha|x-y|}.$$

Thus, for any point $x \in C$, the probability density for an accepted δ-ball move from u to x is at least $\frac{1}{\mathrm{vol}(B_\delta)} e^{-\alpha\,\delta}$; similarly for an accepted move from v to x. Thus, computing the total variation distance as 1 minus the "overlapping area," we have

$$\|P(u,\cdot) - P(v,\cdot)\|_{\mathrm{tv}} \leq 1 - \frac{1}{\mathrm{vol}(B_\delta))} \int_C e^{-\alpha\,\delta} \mu(dx) = 1 - \frac{1}{2} e^{-\alpha\,\delta}.$$

□

Theorem 5.3.4. *Suppose π is an α-smooth log-concave probability measure on a convex set K. Suppose further that π has barycenter 0 and set $M_\pi = \int_K |x|\, \pi(dx)$. Then the conductance, $\mathfrak{h}_P$, of the Metropolis–Hastings chain with transition kernel $P(x, dy)$ induced by the uniform δ-ball proposal satisfies*

$$\mathfrak{h}_P \geq \frac{\delta\, e^{-\alpha\,\delta}}{512\,\sqrt{n}\,M_\pi},$$

provided δ is small compared to $1/M_\pi$.

Proof of Theorem 5.3.4. Let $K = S_1 \cup S_2$, where S_1 and S_2 are disjoint and measurable. We begin by proving that

$$\int_{S_1} P(x, S_2)\, \pi(dx) \geq \frac{\delta\, e^{-\alpha\,\delta}}{512\,\sqrt{n}\,M_\pi} \min\left(\pi(S_1), \pi(S_2)\right). \tag{5.17}$$

It is easy to see that

$$\int_{S_1} P(x, S_2)\pi(dx) = \int_{S_2} P(x, S_1)\pi(dx).$$

Now consider subsets that are "deep" inside S_1 and S_2, in the sense that the Metropolis–Hastings chain is unlikely to move out of them in one step:

$$S_1' = \left\{ x \in S_1 : P(x, S_2) < \frac{1}{4} e^{-\alpha\,\delta} \right\}$$

and

$$S_2' = \left\{ x \in S_2 : P(x, S_1) < \frac{1}{4} e^{-\alpha\,\delta} \right\}.$$

First consider the case $\pi(S_1') < \pi(S_1)/2$. Then

$$\int_{S_1} P(x, S_2)\pi(dx) \geq \frac{1}{4} e^{-\alpha\,\delta} \pi(S_1 \setminus S_1') > \frac{1}{8} e^{-\alpha\,\delta} \pi(S_1),$$

which proves (5.17) provided we choose δ small enough compared to $1/M_\pi$.

So we can assume that $\pi(S_1') \geq \pi(S_1)/2$ and, by the same reasoning, $\pi(S_2') \geq \pi(S_2)/2$. Then, for any $x \in S_1'$ and $y \in S_2'$,

$$\|P(x, \cdot) - P(y, \cdot)\|_{\text{tv}} \geq |P(x, S_1) - P(y, S_1)| \geq 1 - P(x, S_2) - P(y, S_1) > 1 - \frac{1}{2} e^{-\alpha\,\delta}.$$

Applying Lemma 5.3.3, we obtain for any $x \in S_1'$ and $y \in S_2'$ that

$$|x - y| \geq \frac{\delta}{8\sqrt{n}},$$

and hence $d(S_1', S_2') \geq \frac{\delta}{8\sqrt{n}}$. Set $B = K \setminus \{S_1' \cup S_2'\}$ and apply Theorem 5.3.2 to the partition $K = S_1' \cup S_2' \cup B$ to get

$$\pi(B) \geq \frac{\delta}{16\sqrt{n}\, M_\pi} \pi(S_1)\pi(S_2).$$

It follows that

$$\begin{aligned}\int_{S_1} P(x,S_2)\pi(dx) &= \frac{1}{2}\int_{S_1} P(x,S_2)\pi(dx) + \frac{1}{2}\int_{S_2} P(x,S_1)\pi(dx) \\ &\geq \frac{1}{16}\pi(B)e^{-\alpha\,\delta} \\ &\geq \frac{\delta\, e^{-\alpha\,\delta}}{256\,\sqrt{n}\,M_\pi}\pi(S_1)\pi(S_2),\end{aligned} \tag{5.18}$$

in agreement with (5.17) since $\pi(S_1)\pi(S_2) \geq \min\left(\pi(S_1), \pi(S_2)\right)/2$. (Note that the first inequality above holds because we may assume $\pi(B \cap S_1) \geq \pi(B)/2$ by symmetry, and on $B \cap S_1$ we have $P(x, S_2) \geq \frac{1}{4}e^{-\alpha\,\delta}$.)

Thus we have verified (5.17). To finish the proof of the theorem, just notice that (5.17) implies, for every set S_1 satisfying $\pi(S_1) \leq 1/2$ (and hence $\pi(S_2) \geq 1/2$), that

$$\frac{1}{\pi(S_1)}\int_{S_1} P(x,S_2)\pi(dx) \geq \frac{\delta\, e^{-\alpha\,\delta}}{512\,\sqrt{n}\,M_\pi},$$

and hence

$$\mathfrak{h}_P = \inf_{0<\pi(A)\leq 1/2} \mathfrak{h}_P(A) \geq \frac{\delta\, e^{-\alpha\,\delta}}{512\,\sqrt{n}\,M_\pi}.$$

□

Remark 2. We have freedom in choosing δ. The optimal δ (for the lower bound on conductance) is $\delta = 1/\alpha$. With this choice, we have

$$\mathfrak{h}_P \geq \frac{1}{512\,e\,\sqrt{n}\,M_\pi\,\alpha}.$$

This choice of δ makes sense. Imagine, for example, a chain starting at the apex of a 1-dimensional two-sided exponential density $e^{-\alpha|x|}$, with α large. A large value of δ causes proposed moves to be rejected most of the time, resulting in slower mixing. However, a chain with small δ has a reasonably large chance of moving away from the apex, and hence mixes faster.

In recent work, Lovász and Vempala (2003b) were able to demonstrate fast convergence when sampling a log-concave distribution without the "smoothness" assumption. The technique they used was, loosely, to "smooth out" the distribution by convolving the log-concave density with a uniform distribution of small variance. It is interesting to put their idea into a probability context. Suppose X and Y are two random variables such that X has a log-concave density, $f(x)$. Suppose the probability density of Y is smooth and log-concave, with $E[Y] = 0$ and $\mathrm{Var}(Y)$ small. Then the sum of these two random variables, $Z = X + Y$, has a density, $g(x)$, given by the convolution of two log-concave densities, and hence is also log-concave (Leindler 1972; Prékopa 1973). Intuitively, these two densities $f(x)$ and $g(x)$ should be close to each other if $\mathrm{Var}(Y)$ is sufficiently small, and $g(x)$ is smoother than $f(x)$ on the scale of the $\sqrt{\mathrm{Var}(Y)}$. Y can be interpreted as a small perturbation and this perturbation determines, in a way, how close a chain can get to the target distribution (if one leaves out the smoothness assumption on density of X).

The result of Lovász and Vempala (2003b), in a paraphrased form, says that

$$||\mu_0 P^n - \pi|| \leq M\epsilon + \gamma_\epsilon^n ||\mu_0 - \pi||, \tag{5.19}$$

where μ_0 is the starting measure, P is the Markov operator with target measure π, ϵ is a small term that determines the accuracy of the algorithm, M is a constant, and γ_ϵ is the convergence rate that is determined by ϵ. In fact, $\gamma_\epsilon = 1 - \Phi_\epsilon^2/2$, where Φ_ϵ is the ϵ-*conductance* defined by $\sup_{\epsilon < \pi(A) \leq 1/2} \frac{\int_A P(x, A^c)\pi(dx)}{\pi(A) - \epsilon}$. They were able to show that the ϵ-conductance can be bounded below by a quadratic function of ϵ.

In summary, if one ignores sets of small measure for a log-concave target density, a Metropolis–Hastings chain induced by a ball walk (even without the smoothness assumption on the target) is "geometrically ergodic." We would like to have directly applied this nice result, but we chose not to for two reasons. First, the state decomposition theorem applies in the context of spectral gap, while strictly speaking, Equation (5.19) does

not give geometric ergodicity, and hence it can not be applied directly in the state decomposition theorem. Secondly, if one chooses to cut off small sets, then all log-concave densities that decay faster than an exponential essentially have compact supports, and hence are "smooth." So the results in this section apply. We note here, however, that both Lemma 5.3.3 and Theorem 5.3.4 are borrowed from Lovász and Vempala (2003a) with some modifications to apply arguments on conductance instead of ϵ-conductance.

5.4 Proof of the main theorem

5.4.1 A 1-D example

To gain some insight into the role of the complexity of the target distribution and the idea behind the proof of Theorem 5.1.3, we begin with a simple 1-dimensional example in which Ω is a circle with perimeter $4L$ for some $L \gg 1$; i.e., the interval $[-2L, 2L]$ with the two ends connected. Consider a two-mode target distribution

$$\pi(x) = \begin{cases} c\,\nu\, e^{-\nu\,|x|} & \text{if } x \in [-L, L], \\ c\,\nu\, e^{-\nu\,(2L-|x|)} & \text{if } x \in [-2L, -L] \cup [L, 2L], \end{cases} \tag{5.20}$$

where c is the normalization constant. Here, we can think of L and ν as determining the complexity of the target distribution; increasing ν makes the modes more narrow, and increasing L increases the size of the space and places the modes further apart. We denote by π_1 the piece of π defined on $[-L, L]$ and by π_2 the other piece. We take for the local proposal the uniform distribution $k(x, y) = 2/\delta$ for $y \in [x-\delta, x+\delta]$ and 0 otherwise. Let $P_k(x, dy)$ be the transition kernel for the Metropolis–Hastings chain based on this local proposal and having target distribution π. Consider the partition $A = [-L, L]$, $A^c = [-2L, -L] \cup [L, 2L]$. Then

$$\mathfrak{h}_{P_k} \le \mathfrak{h}_{P_k}(A) < \frac{2}{\pi(A)} \int_{L-\delta}^{L} P_k(x, A^c)\pi(dx) < 2c\, e^{-\nu\,(L-\delta)}.$$

By Cheeger's inequality, we get

$$\mathfrak{gap}\,(P_k) \le 2\mathfrak{h}_{P_k} \le 4c\, e^{-\nu\,(L-\delta)}. \tag{5.21}$$

Thus, the spectral gap for the local Metropolis–Hastings chain decreases exponentially in L and ν, finishing the first part of our proof for this example.

Now consider a heavy-tailed proposal distribution $h(x,y) = 1/4L$, i.e., a uniform distribution on Ω, and the small world proposal $g(x,y) = (1-s)k(x,y) + s\,h(x,y)$. Let $P_{g,A}(x,dy)$ be the transition kernel for the small-world chain that is restricted to the set A. Then $P_{g,A}(x,dy) = (1-s)P_{k,A}(x,dy) + s\,P_{h,A}(x,dy)$, where $P_{k,A}$ and $P_{h,A}$ are the restrictions to A of the kernels induced by $k(x,y)$ and $h(x,y)$, respectively. By (5.8), we have $\mathfrak{h}_{P_{g,A}} \ge (1-s)\mathfrak{h}_{P_{k,A}}$. It is easy to check that, for the two-sided exponential distribution, $M_\pi = 1/\nu$. Then by Theorem 5.3.4,

$$\mathfrak{h}_{P_{k,A}} \ge \frac{\delta\,\nu\,e^{-\nu\,\delta}}{512}.$$

By Cheeger's inequality, we have

$$\mathfrak{gap}\,(P_{g,A}) \ge \frac{\mathfrak{h}^2_{P_{g,A}}}{2} \ge \frac{\delta^2\,\nu^2 e^{-2\,\nu\,\delta}}{2^{19}}\,(1-s)^2. \tag{5.22}$$

By symmetry, the small-world chain that is restricted to A^c has the same lower bound for its spectral gap.

Also, by symmetry, the matrix of transition probabilities for the component chain has the form $P_H = \left(\begin{smallmatrix} 1-a & a \\ a & 1-a \end{smallmatrix}\right)$. The spectral gap for this matrix is $\mathfrak{gap}\,(P_H) = 2a$. Now

we calculate $a = P_H(1,2)$. Set $I = \int_0^L \nu\, e^{-\nu\, x}\, dx$. Then $\pi(A) = 2\, c\, I$. By (5.12) we have

$$\begin{aligned} P_H(1,2) = \frac{\int_A P_g(x, A^c)\pi(dx)}{2\,\pi(A)} &> \frac{1}{4\,c\,I}\,\frac{s}{4L}\int_A\int_{A^c} \min(\pi(y),\pi(x))dydx \\ &= \frac{1}{4\,c\,I}\frac{c\,s\,\nu}{L}\int_0^L\int_0^L \min\left(e^{-\nu x}, e^{-\nu y}\right)\, dy\, dx \\ &= \frac{s\,\nu}{4\,I\,L}\int_0^L \left(\int_0^x + \int_x^L\right) \min\left(e^{-\nu x}, e^{-\nu y}\right)\, dy\, dx \\ &= \frac{s}{2\,I\,\nu\,L}(1 - e^{-\nu L} - \nu L e^{-\nu L}). \end{aligned} \tag{5.23}$$

When $\nu L \geq 2$, this yields $P_H(1,2) > s/(4\,\nu\,L)$. Note that instead of just using the fact that $2\pi(A) = 1$, we chose to do the calculation the "hard" way in order to show that the normalization constant c has no effect on the spectral gap. Using the state decomposition theorem to combine (5.22) and (5.23) we have

$$\mathfrak{gap}\,(P_g) > \frac{s(1-s)^2\,\delta^2\,\nu\,e^{-2\,\nu\,\delta}}{2^{21}L}, \quad \text{for } \nu\,L \geq 2. \tag{5.24}$$

Setting $\delta = 1/\nu$ in Equation (5.24) leads to

$$\mathfrak{gap}\,(P_g) > \frac{s(1-s)^2\,e^{-2}}{2^{21}\,\nu\,L}, \quad \text{for } \nu\,L \geq 2.$$

For a small world chain, the lower bound on the spectral gap decreases linearly with both L and ν. Moreover, the quantity $1/\nu$ determines the absolute "size" of a mode, and hence $1/(\nu\,L)$ reflects the relative size of each mode. Thus, we can see how the spectral gap is influenced by the relative size of each mode.

We have freedom in the choice of the value s. It is clear that $s = 0$ corresponds to a pure local chain and $s = 1$ corresponds to the rejection method. Either case will make the right side of (5.24) equal to 0, which either implies the lower bound is too rough, or the chain is slowly mixing. Note that, in the lower bound, the best value for s is $1/3$, which maximizes $s\,(1-s)^2$.

Using a uniform distribution for $h(x, y)$ does not make sense in an unbounded space. However, this is not a problem because we can always use, say, a Cauchy distribution $h(x) = \frac{1}{\pi} \frac{b}{x^2+b^2}$, where b is the half width at half maximum. Some prior knowledge about the target distribution will help in choosing b in a way that increases the lower bound on the spectral gap, and hence the convergence rate of the corresponding small-world chain. Even in a bounded space, the use of a Cauchy distribution, instead of a uniform, may increase the convergence rate in cases for which most of the mass is accumulated in a small portion of the state space.

5.4.2 The general case

Proof of Theorem 5.1.3. The proof of the general case is similar in spirit to the one-dimensional case. For the first part of the theorem we want to show that, under a local proposal, the spectral gap is exponentially small. It is sufficient to prove that the one-step probability flow going out of at least one mode is exponentially small. Among all m pieces of the partition, at least one piece has measure no bigger than $1/2$. Without loss of generality, suppose it is A_1. Consider any radius $L > 0$ such that $B = B_L(\beta_1) \subset A_1$ where β_1 is the barycenter of π_1. Let P_k be the operator induced by a local proposal $k(x, y)$ given by a δ-ball walk. Then

$$\begin{aligned}
\mathfrak{h}_{P_k} \leq \mathfrak{h}_{P_k}(B) = \frac{1}{\pi_1(B)} \int_B P_k(x, B^c)\, \pi(dx) \;&= \frac{1}{\pi_1(B)} \int_B \int_{B^c} \pi(x)\, k(x, y)\, \mu(dy)\, \mu(dx) \\
&\leq \frac{1}{\pi_1(B)} \int_{L-\delta}^{L} \pi_1^+(\partial B_u(\beta_1))\, du \\
&\leq \frac{1}{\pi_1(B)} \int_{L-\delta}^{L} e^{-\nu_1 u}\, du \\
&\leq \frac{1}{\pi_1(B)\, \nu_1}\, e^{-\nu_1(L-\delta)}
\end{aligned}$$

where the second inequality follows the fact $\int_{B^c} k(x, y)\, \mu(dy) \leq 1$, and we have written ν_1 for the decay exponent of π_1.

By Cheeger's inequality, we have

$$\mathfrak{gap}\,(P_k) \le 2\mathfrak{h}_{P_k} \le \frac{2}{\pi_1(B)\,\nu_1}\,e^{-\nu_1(L-\delta)},$$

and this finishes the first part of the proof.

To prove the second part of the theorem, let $P_g = (1-s)P_k + s\,P_h$ be the small world operator, where P_k and P_h are induced by the local proposal $k(x,y)$ and the heavy-tailed proposal $h(x,y)$, respectively. Let P_{g,A_j} be the restriction of the operator P_g on the set A_j, and P_{k,A_j}, P_{h,A_j} be the restrictions of P_k, P_h to A_j, respectively. We have $P_{g,A_j} = (1-s)P_{k,A_j} + s\,P_{h,A_j}$.

By Theorem 5.3.4 and $M_{\pi_j} \le c/\nu_j$, we have

$$\mathfrak{h}_{P_{g,A_j}} \ge \frac{\nu_j\,\delta\,e^{-\nu_j\,\delta}}{512\,c\,\sqrt{n}}\,(1-s)$$

and hence Cheeger's inequality implies

$$\mathfrak{gap}\,(P_{g,A_j}) \ge \frac{\nu_j^2\,\delta^2\,e^{-2\,\nu_j\,\delta}}{2^{19}\,c^2\,n}\,(1-s)^2. \tag{5.25}$$

Next we want to calculate $P_H(i,j)$. Let $b = \max_{i\neq j}|\beta_i - \beta_j|$ denote the maximum of the pairwise distances between barycenters. Let the heavy-tailed distribution be an n-dimensional Cauchy distribution with half width b:

$$h(x,y) = \frac{c_n\,b}{(|y-x|^2 + b^2)^{\frac{n+1}{2}}},$$

where $c_n = \pi^{(n+1)/2}/\Gamma(\frac{n+1}{2})$ is the normalization constant–a function of the dimension n.

On each partition piece A_i pick a ball $B_i = B_{R_i}(\beta_i) \subset A_i$ such that $\pi(B_i) = \frac{2}{3}\,\pi(A_i)$. Let $h_i = \inf_{x\in\partial B_i}\pi(x)$, the "height" of the density π_i along the boundary of B_i. Let $B_i^c = A_i \setminus B_i$ be the complement of B_i on the set A_i and set $c_{ij} = \min(h_i/h_j,\ h_j/h_i)$.

Then

$$
\begin{aligned}
I &\equiv \int_{A_i}\int_{A_j} h(x,y)\,\min(\pi(y),\pi(x))\mu(dx)\,\mu(dy) \\
&> \int_{B_i^c}\int_{B_j} h(x,y)\min(\pi(y),\pi(x))\mu(dx)\mu(dy) \;+ \int_{B_i}\int_{B_j^c} h(x,y)\min(\pi(y),\pi(x))\mu(dx)\mu(dy) \\
&> \int_{B_i^c}\int_{B_j} h(x,y)\pi(x)\min(\frac{h_i}{h_j},1)\mu(dx)\mu(dy) \;+ \int_{B_i}\int_{B_j^c} h(x,y)\pi(y)\min(\frac{h_j}{h_i},1)\mu(dx)\mu(dy) \\
&> c_{ij}\int_{B_i^c}\int_{B_j} \pi(x)\,h(x,y)\,\mu(dx)\,\mu(dy) \;+ c_{ij}\int_{B_j^c}\int_{B_i} \pi(y)\,h(x,y)\,\mu(dy)\,\mu(dx).
\end{aligned}
$$

Since $h(x,y) = h(|x-y|) = h(r)$ decreases polynomially, while both $\pi(x)$ and $\pi(y)$ decrease exponentially, there exists a ball $\hat{B}_w$ with radius wb such that $\pi_i(\hat{B}_w) > \frac{5}{6}\pi_i(A_i)$, $\pi_j(\hat{B}_w) > \frac{5}{6}\pi_j(A_j)$, and $\inf_{r\in\hat{B}_w} h(r) = \epsilon/c_n$, where $\epsilon = \epsilon(wb)$ is polynomially small in wb. Note that $\pi_i(B_i) = \frac{2}{3}\pi_i(A_i)$ and $\pi_i(B_j) = \frac{2}{3}\pi_j(A_j)$, so

$$
\begin{aligned}
I &> c_{ij}\int_{B_i^c\cap\hat{B}_w}\int_{B_j} \pi(x)\,\frac{\epsilon}{c_n}\,\mu(dx)\,\mu(dy) \;+ c_{ij}\int_{B_j^c\cap\hat{B}_w}\int_{B_i} \pi(y)\,\frac{\epsilon}{c_n}\,\mu(dy)\,\mu(dx) \\
&> \frac{c_{ij}\epsilon}{c_n}\left(\frac{1}{6}\pi(A_i)\,\mathrm{vol}(B_j) + \frac{1}{6}\pi(A_j)\,\mathrm{vol}(B_i)\right).
\end{aligned}
\tag{5.26}
$$

From (5.12) and (5.26) we get

$$
\begin{aligned}
P_H(i,j) = \frac{\int_{A_i} P_g(x,A_j)\pi(dx)}{2\,\pi(A_i)} &> \frac{s}{2\,\pi(A_i)}\,I \\
&> \frac{s}{2\,\pi(A_i)}\frac{c_{ij}\epsilon}{c_n}\left(\frac{1}{6}\pi(A_i)\,\mathrm{vol}(B_j) + \frac{1}{6}\pi(A_j)\,\mathrm{vol}(B_i)\right) \\
&> \frac{s\,c_{ij}\,\epsilon_{wb}}{12\,c_n}\,\mathrm{vol}(B_j).
\end{aligned}
\tag{5.27}
$$

For an $m\times m$ stochastic matrix $A = (a_{ij})$, the spectral gap can be bounded from below by Proposition 3.2 in (Peña 2005):

$$
\mathfrak{gap}\,(A) \geq m\min_{i\neq j} a_{ij}.
$$

Combining this with (5.27) results in

$$\mathfrak{gap}\,(P_H) \geq \frac{sm\,\epsilon_{wb}}{12\,c_n} \min_{i\neq j}(c_{ij}\mathrm{vol}(B_j)). \tag{5.28}$$

Using the state decomposition theorem to put (5.25) and (5.28) together, we get

$$\mathfrak{gap}\,(P_g) \geq s(1-s)^2\,\frac{m\,\epsilon_{wb}\,\delta^2}{2^{24}\,c^2\,n\,c_n} \min_j\,(\nu_j^2\,e^{-2\,\nu_j\,\delta}) \min_{i\neq j}(c_{ij}\mathrm{vol}(B_j)). \tag{5.29}$$

Setting $\delta = 1/\max_j\,(\nu_j)$ yields

$$\mathfrak{gap}\,(P_g) > s(1-s)^2\,\frac{m\,\epsilon_{wb}}{2^{24}\,c^2\,e^2\,n\,c_n} \min_{i\neq j}(c_{ij}\mathrm{vol}(B_j)).$$

Notice that $\mathrm{vol}(B_j)$ decreases polynomially with increase of ν_j. This concludes the proof.

□

Remark 3. In the proof, we essentially used a uniform distribution on a bounded set as a heavy-tailed distribution. Notice that, loosely, $\epsilon_{wb}\mathrm{vol}(B_j)/c_n$ determines the relative size of mode j. In our lower bound as shown in (5.29), we have the so-called "curse of dimensionality": c_n decreases exponentially with increase of dimension n. Interestingly, the best value for s is still 1/3.

5.5 Metropolis-coupled MCMC and simulated tempering

Metropolis-coupled MCMC (MCMCMC), proposed by Geyer (1991), shares the same spirit as "simulated tempering," which was independently proposed by Marinari and Parisi (1992). Both are based on an analogy with simulated annealing (Kirkpatrick, Gelatt, and Vecchi 1983), which is an optimization algorithm rather than a sampling scheme. It provides the useful metaphor of using some help from a "heated" version of the problem that makes valley crossing easier by flattening the state space, to obtain the result in

the original "cooled" version of the problem one is interested in. Simulated annealing uses a specific form of "heating" that is sometimes called "powering up." If $h_1(x)$ is the unnormalized density for the distribution of interest, $h_t(x) = h_1(x)^{1/t}$, for $t > 1$, are the heated unnormalized densities, including perhaps $t = \infty$ which gives $\pi(x) = 1$. However, as noted by Geyer and Thompson (1995), "powering up" is not an essential part of simulated tempering or of MCMCMC, and a different form of heating may work better in a specific real application.

Let $T = \{1, \ldots, t\}$. Both MCMCMC and simulated tempering simulate a sequence of t distributions specified by unnormalized densities $h_i(x)$ $(i \in T)$ on the same sample space Ω, where the index i is called the "temperature," $h_1(x)$ is the "cold" distribution, and $h_t(x)$ is the "hot" distribution. In fact, an MCMCMC chain lives in a product state space $\Omega \times T$ such that, for a given $i \in T$, the chain updates itself on Ω using a Metropolis–Hastings algorithm. For the move between different "temperatures," one keeps the $x \in \Omega$ and only updates the "temperature." Specifically, suppose $a(i)$ $(i = 1, \ldots, t)$ is the auxiliary probability distribution for the temperatures. Then one iteration of the "Metropolis-Hastings" version of the simulated tempering algorithm is as follow (Geyer and Thompson 1995):

1. Update x using a Metropolis-Hastings update for h_i.

2. Set $j = i \pm 1$ according to probabilities $q_{i,j}$, where $q_{1,2} = q_{m,m-1} = 1$ and $q_{i,i+1} = q_{i,i-1} = 1/2$ if $1 < i < m$ (i.e., reflecting random walk on different temperatures).

3. Calculate the Hastings ratio
$$r = \frac{h_j(x)a(j)q_{j,i}}{h_i(x)a(i)q_{i,j}}$$
and accept the transition (set i to j) or reject it according to the Metropolis rule: accept with probability $\min(r, 1)$.

An implicit assumption in the simulated tempering algorithm is that, at each temperature, the proposal distribution that is used to generate a new move $x \in \Omega$ is lo-

cal. For the sake of simplicity and clarity, let us assume that we have two temperatures, hot and cool, $a(1) = a(2) = 1/2$ and $q_{1,2} = q_{2,1} = 1$. Then r in step 3 becomes $h_j(x)/h_i(x)$, for $i, j \in \{1, 2\}$. Suppose now that the chain is at high temperature, $h_2(x)$. If x is in a mode, then $h_1(x)/h_2(x)$ is close to 1 (by powering up), so that the chain is likely to jump back to the cool state and collect samples. On the other hand, if x is in a valley, $h_1(x)/h_2(x)$ is small, so that the chain tends to stay at the hot temperature. When the hot chain has wandered far enough and proposes a move back to a cool temperature, it in fact proposes a move to the cool chain that is on average far away (as compared to the local proposal) from the state (in Ω) where the chain last visits the cool temperature. In summary, if one is only interested in the samples collected in cool state (i.e., the original distribution), then the only purpose of the hot state is to provide a far away proposal for the cool chain. This is the exact spirit of the occasional heavy-tailed proposals in the small-world chain.

We note, however, that although simulated tempering, or MCMCMC, is a way to generate heavy-tailed proposals to overcome bottlenecks in Ω, the computational cost is heavy–much heavier than for a small-world chain. Moreover, it has been shown by Bhatnagar and Randall (2004) that, in certain situations, the transition between different temperatures can have bottlenecks, which will slow down the frequency of "heavy-tailed" proposals, and hence slow down the overall convergence.

Nonetheless, if one can rule out the possible bottlenecks in transitions between the hot chain and the cool chian, our Theorem 5.1.3 for small-world chains readily applies to MCMCMC, or simulated tempering, to show that both of them are "rapidly mixing."

Note that the different temperatures in simulated tempering in fact correspond to different amounts of heaviness of the tail in a small-world chain. Particularly, when Ω is compact, $t = \infty$ corresponds to the heavy-tailed proposal being a uniform distribution. Therefore, we propose that a promising scheme for using Markov chain Monte Carlo methods to solve hard problems would be to run multiple small-world chains in parallel

with different chains having different heaviness of tails; for example, using different half-widths in Cauchy distributions, then coupling different chains via the Hastings ratio and Metropolis rule.

A Brief Outlook

In genome science, the three different parts of this dissertation may find a common playground. Furthermore, these interactions may happen on two different scales.

On a macroscopic scale, genomic projects (including human and other mammals) collect data from different geographical regions, so that a spatial population genetics model is required to draw meaningful conclusions from data. The inference of model parameters requires building phylogenetic (when data cross species) and/or coalescent (when data are from within a species) trees. Distance methods have a critical speed advantage over others when handling large data sets. The inference also requires one to sample the parameter space (including tree space), where Markov Chain Monte Carlo has played a dominant role.

On a microscopic scale, evolutionarily relevant parameters, for example, mutation and recombination rates, tend to have local dependence. Interacting particle systems could be a capable tool for modeling such dependence. On the other hand, the inference of recombination hotspots involves sampling ancestral recombination graphs, where a fast mixing MCMC algorithm is critical. Meanwhile, at each locus, an ancestral recombination graph breaks into trees. And distance methods could ease the computation, especially when data grows at a faster rate than Moore's Law.

Bibliography

Applegate, D. and R. Kannan (1990). Sampling and integration of near log-concave functions. In *Proc. 23rd ACM STOC.*, pp. 156–163.

Atteson, K. (1999). The performace of neighbor-joining methods of phylogenetic reconstruction. *Algorithmica 25*, 251–278.

Bhatnagar, N. and D. Randall (2004). Torpid mixing of simulated tempering on the Potts model. In *Proceedings of the 15th Annual ACM-SIAM Symposium on Discrete Algorithms (New Orleans, LA)*, pp. 478–487.

Bobkov, S. G. (1999). Isoperimetric and analytic inequalities for log-concave probability measures. *Ann. Probab. 27*, 1903–1921.

Borell, C. (1974). Convex measures on locally convex spaces. *Ark. Math. 12*, 239–252.

Bradshaw, D. J., K. A. Homer, P. D. Marsh, and D. Beighton (1994). Metabolic cooperation in oral microbial communities during growth on mucin. *Microbiology 140*, 3407–3412.

Bramson, M. and D. Griffeath (1989). Flux and fixation in cyclic particle systems. *Ann. Proba. 17*, 26–45.

Bruno, W. J., N. D. Socci, and A. L. Halpern (2000). Weighted neighbor joining: A likelihood-based approach to distance-based phylogeny reconstruction. *Molecular Biology and Evolution 17*, 189–197.

Bryant, D. and P. Waddell (1998). Rapid evaluation of least-squares and minimum-evolution criteria on phylogenetic trees. *Molecular Biology and Evolution 15*, 1346–1359.

Chan, K. S. and C. J. Geyer (1994). Discussion of the paper by Tierney. *Ann. Statist. 22*, 1747–1758.

Chib, S. and E. Greenberg (1995). Understanding the Metropolis-Hastings algorithm. *The American Statistician 49*, 327–335.

Chib, S., E. Greenberg, and R. Winkelmann (1998). Posterior simulation and Bayes factors in panel count data models. *Journal of Econometrics 86*, 33–54.

Cohen, J. E. (2004). Mathematics is biology's next microscope, only better; biology is mathematics' next physics, only better. *PLoS Biol. 2(12)*, e439.

Comins, H., M. Hassell, and R. M. May (1992). The spatial dynamics of host-parasitoid systems. *J. Anim. Ecol. 61*, 735–748.

Dieckmann, U., R. Law, and J. A. J. Metz (Eds.) (2000). *The Geometry of Ecological Interactions: Simplifying Spatial Complexity*. Cambridge University Press, Cambridge.

Doebeli, M. (2002). A model for the evolutionary dynamics of cross-feeding polymorphisms in microorganisms. *Popul. Ecol. 44*, 59–70.

Durrett, R. (1995). *Ten Lectures on Particle Systems*, Volume 1608 of *Lecture Notes in Mathematics*. Springer, Berlin.

Durrett, R. (2002). *Lecture Notes on Particle Systems and Percolation*. Wadsworth, Belmont, CA.

Durrett, R. and D. Griffeath (1993). Asymptotic behavior of excitable cellular automata. *Experiment. Math. 2*, 183–208.

Durrett, R. and S. Levin (1994). Stochastic spatial models: a user's guide to ecological applications. *Phil. Trans. R. Soc. Lond. B 343*, 329–350.

Durrett, R. and S. A. Levin (1997). Allelopathy in spatially distributed populations. *J. Theor. Biol. 185*, 165–172.

Durrett, R. and C. Neuhauser (1994). Particle systems and reaction diffusion equations. *Ann. Probab. 22*, 289–333.

Felsenstein, J. (1997). Fitting trees to distances by alternating least squares. *Syst. Biol. 46(1)*, 101–111.

Felsenstein, J. (2004). *Inferring Phylogenies*. Sinauer Associates, Sunderland, Massachusetts.

Gamerman, D. (1997). *Markov Chain Monte Carlo*. Chapman & Hall, London.

Gascuel, O. (1997a). BIONJ: An improved version of the NJ algorithm based on a simple model of sequence data. *Molecular Biology and Evolution 14*, 685–695.

Gascuel, O. (1997b). Concerning the NJ algorithm and its unweighted version, UNJ.

Gascuel, O., D. Bryant, and F. Denis (2001). Strengths and limitations of the minimum evolution principle. *Systematic Biology 50(5)*, 621–627.

Geyer, C. J. (1991). Markov chain Monte Carlo maximum likelihood. In E. M. Keramides (Ed.), *Computing Science and Statistics: Proceedings of the 23rd Symposium on the Interface*, pp. 156–163. Interface Foundation, Fairfax Station.

Geyer, C. J. (1992). Practical Markov chain Monte Carlo. *Statist. Sci. 7*, 473–483.

Geyer, C. J. and E. A. Thompson (1995). Annealing Markov chain Monte Carlo with applications to ancestral inference. *J. Amer. Statist. Assoc. 90*, 909–920.

Gilks, W. R., S. Richardson, and D. J. Spiegelhalter (1996). *Markov Chain Monte Carlo in Practice, 1st edition*. Chapman & Hall, London.

Grindrod, P. (1996). *The Theory and Applications of Reaction-Diffusion Equations: Patterns and Waves, 2nd ed.* Oxford Univ. Press.

Guan, Y., R. Fleißner, P. Joyce, and S. M. Krone (2006). Markov chain Monte Carlo in small worlds. *Stat. Comput. 16*, 193–202.

Hastings, W. K. (1970). Monte carlo sampling methods using markov chains and their applications. *Biometrika 57*, 97–109.

Hirsch, M. and S. Smale (1974). *Differential Equations, Dynamical Systems, and Linear Algebra*. Academic Press, San Diego.

Hoffman, K. and R. Kunze (1971). *Linear Algebra, 2nd ed.* Prentice Hall.

Jarner, S. F. and G. O. Roberts (2001). Convergence of heavy tailed MCMC algorithms. Technical report, Lancaster Univ. Available at `http://www.staslab.cam.ac.uk/mcmc`.

Jarner, S. F. and W. K. Yuen (2004). Conductance bounds on the L^2 convergence rate of Metropolis algorithms on unbounded state spaces. *Adv. in Appl. Probab. 36*, 243–266.

Kannan, R. and G. Li (1996). Sampling according to the multivariate normal density. In *37th Annual IEEE Symposium on Foundations of Computer Science (FOCS'96)*, pp. 204–212.

Kannan, R., L. Lovász, and M. Simonovits (1995). Isoperimetric problems for convex bodies and a localization lemma. *Discrete Comput. Geom. 13*, 541–559.

Kerr, B., M. A. Riley, M. W. Feldman, and B. J. M. Bohannan (2002). Local dispersal promotes biodiversity in a real-life game of rock-paper-scissors. *Nature 418*, 171–174.

Kirkpatrick, S., J. Gelatt, C. D., and M. P. Vecchi (1983). Optimization by simulated annealing. *Science 220*, 671–680.

Krone, S. M. (2004). Spatial models: stochastic and deterministic. *Math. Comp. Mod. 40*, 393–409.

Larget, B. and D. Simon (1999). Markov chain Monte Carlo algorithms for the Bayesian analysis of phylogenetic trees. *Mol. Biol. Evol. 16*, 750–759.

Lawler, G. F. and A. D. Sokal (1988). Bounds on the L^2 spectrum for Markov chains and Markov processes: a generalization of Cheeger's inequality. *Trans. Amer. Math. Soc. 309*, 557–580.

Leindler, L. (1972). On a certain converse of Hölder's Inequality II. *Acta Sci. Math. Szeged 33*, 217–223.

Lovász, L. and M. Simonovits (1993). Random walks in a convex body and an improved volume algorithm. *Random Structures Algorithms 4*, 359–412.

Lovász, L. and S. Vempala (2003a). The geometry of logconcave functions and an $O^*(n^3)$ sampling algorithm. Microsoft Research Tech. Rep. MSR-TR-2003–4. Available at : `http://www-math.mit.edu/~vempala/papers/logcon-ball.ps`.

Lovász, L. and S. Vempala (2003b). Logconcave functions: geometry and efficient sampling algorithm. In *44th Annual IEEE Symposium on Foundations of Computer Science (FOCS'03)*, pp. 1–10.

Madras, N. and D. Randall (2002). Markov chain decomposition for convergence rate analysis. *Ann. Appl. Probab. 12*, 581–606.

Marinari, E. and G. Parisi (1992). Simulated tempering: a new Monte Carlo scheme. *Europhysics Letters 19*, 451–458.

Martin, R. A. and D. Randall (2000). Sampling adsorbing staircase walks using a new Markov chain decomposition method. In *41st Annual IEEE Symposium on Foundations of Computer Science (FOCS'01)*, pp. 492–502.

Metropolis, N., A. E. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller (1953). Equation of state calculations by fast computing machines. *J. Chem. Phys. 21*, 1087–1091.

Meyn, S. P. and R. L. Tweedie (1996). *Markov Chains and Stochastic Stability*. Springer, New York.

Murray, J. D. (1989). *Mathematical Biology*. Springer.

Pálsson, E. and E. C. Cox (1996). Origin and evolution of circular waves and spirals in *Dictyostelium discoideum* territories. *Proc. Natl. Acad. Sci. USA 95*, 1151–1155.

Peña, J. M. (2005). Exclusion and inclusion intervals for the real eigenvalues of positive matrices. *SIAM Journal on Matrix Analysis and Applications 26(4)*, 908–917.

Pelz, O., M. Tesar, R. M. Wittich, E. R. B. Moore, K. N. Timmis, and W. R. Abraham (1999). Towards elucidation of microbial community metabolic pathways: unravelling the network of carbon sharing in a pollutant-degrading bacterial consortium by immunocapture and isotopic ratio mass spectrometry. *Environ. Microbiol. 1*, 167–174.

Prékopa, A. (1973). Logarithmic concave measures and functions. *Acta Sci. Math. Szeged 34*, 335–343.

Press, W. H., B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling (1992). *Numerical Recipes in C: The Art of Scientific Computing*. Cambridge University Press, Cambridge.

Rand, D. A. and H. B. Wilson (1995). Using spatio-temporal chaos and intermediate-scale determinism to quantify spatially extended ecosystems. *Proc. R. Soc. Lond. B 259*, 111–117.

Rice, J. A. (1994). *Mathematical Statistics and Data Analysis, 2nd edition*. Duxbury Press, Pacific Grove.

Roberts, G. O., A. Gelman, and W. R. Gilks (1997). Weak convergence and optimal scaling of random walk Metropolis algorithms. *Ann. Appl. Probab. 7*, 110–120.

Roberts, G. O. and J. S. Rosenthal (1997). Geometric ergodicity and hybrid Markov chains. *Electron. Comm. Probab. 2(2)*, 13–25.

Roberts, G. O. and R. L. Tweedie (2001). Geometric L^2 and L^1 convergence are equivalent for reversible Markov chains. *J. Appl. Probab. 38A*, 37–41.

Rosenzweig, R. F., R. R. Sharp, D. S. Treves, and J. Adams (1994). Microbial evolution in a simple unstructured environment: genetic differentiation in *Escherichia coli*. *Genetics 137*, 903–917.

Saitou, N. and M. Nei (1987). The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution 4*, 406–425.

Savill, N., P. Rohani, and P. Hogeweg (1997). Self-reinforcing spatial patterns enslave evolution in a host-parasitoid system. *J. Theor. Biol. 188*, 11–20.

Studier, J. A. and K. J. Keppler (1988). A note on the neighbor-joining algorithm of Saitou and Nei. *Molecular Biology and Evolution 4*, 729–731.

Tierney, L. (1994). Markov chains for exploring posterior distributions. *Ann. Stat. 22*, 1701–1762.

Watts, D. J. and S. H. Strogatz (1998). Collective dynamics of 'small-world' networks. *Nature 393*, 440–442.

Wei, W. and S. M. Krone (2005). Spatial invasion by a mutant pathogen. *J. Theor. Biol. 236*, 335–348.