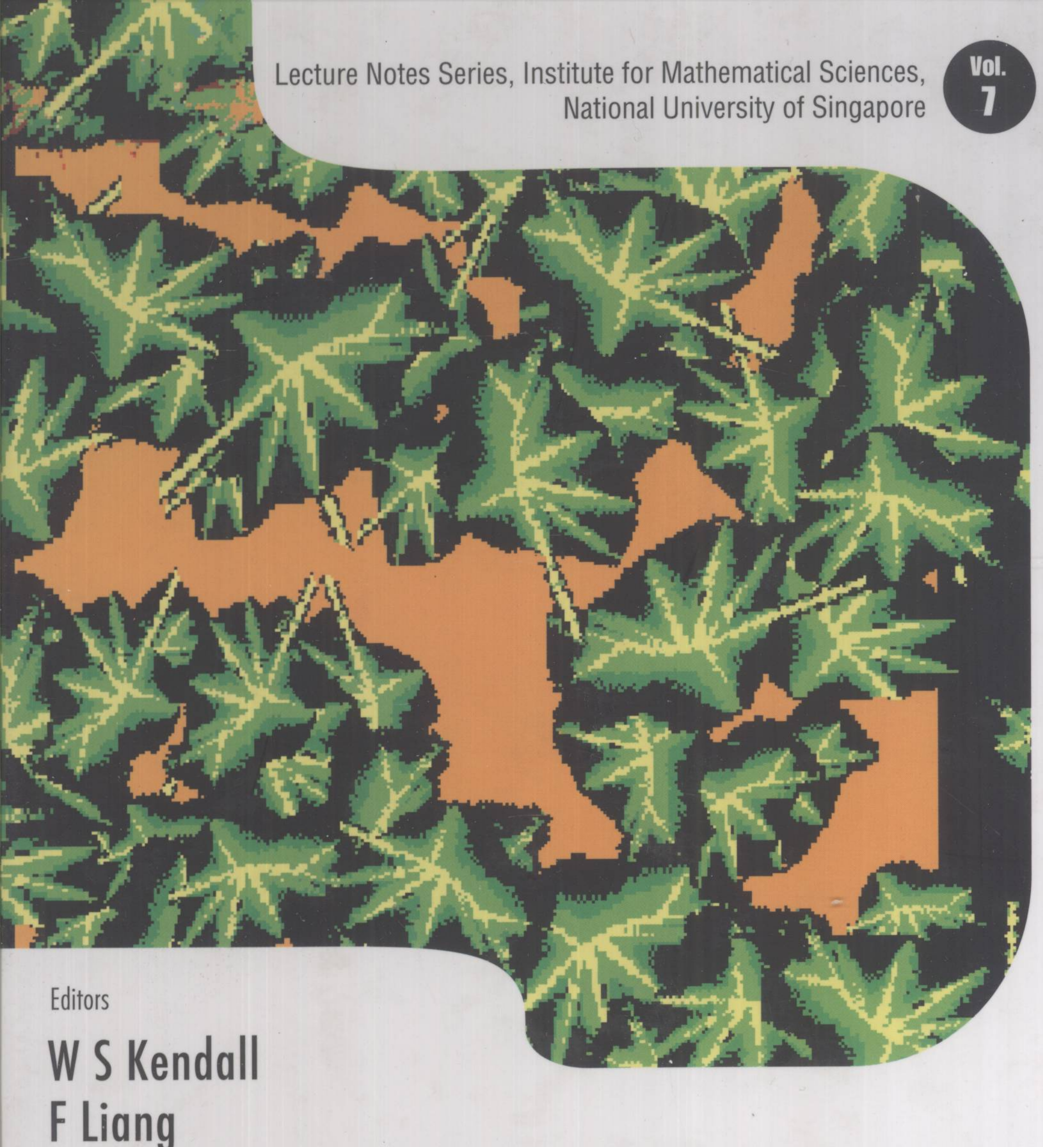
Lecture Notes Series, Institute for Mathematical Sciences,
National University of Singapore

Vol. 7



Editors

W S Kendall
F Liang
J-S Wang

MARKOV CHAIN MONTE CARLO

Innovations and Applications

Lecture Notes Series, Institute for Mathematical Sciences,
National University of Singapore Vol. 7

MARKOV CHAIN MONTE CARLO

Innovations and Applications

Editors

W S Kendall
University of Warwick, UK

F Liang
Texas A & M University, USA

J-S Wang
National University of Singapore, Singapore

World Scientific

NEW JERSEY • LONDON • SINGAPORE • BEIJING • SHANGHAI • HONG KONG • TAIPEI • CHENNAI

Published by

World Scientific Publishing Co. Pte. Ltd.
5 Toh Tuck Link, Singapore 596224
USA office: 27 Warren Street, Suite 401-402, Hackensack, NJ 07601
UK office: 57 Shelton Street, Covent Garden, London WC2H 9HE

British Library Cataloguing-in-Publication Data
A catalogue record for this book is available from the British Library.

MARKOV CHAIN MONTE CARLO
Innovations and Applications

ISBN 981-256-427-6

Printed in Singapore by World Scientific Printers (S) Pte Ltd

MARKOV CHAIN MONTE CARLO

Innovations and Applications

LECTURE NOTES SERIES
Institute for Mathematical Sciences, National University of Singapore

Series Editors: Louis H. Y. Chen and Denny Leung
Institute for Mathematical Sciences
National University of Singapore

Published

Vol. 1 Coding Theory and Cryptology
edited by Harald Niederreiter

Vol. 2 Representations of Real and *p*-Adic Groups
edited by Eng-Chye Tan & Chen-Bo Zhu

Vol. 3 Selected Topics in Post-Genome Knowledge Discovery
edited by Limsoon Wong & Louxin Zhang

Vol. 4 An Introduction to Stein's Method
edited by A. D. Barbour & Louis H. Y. Chen

Vol. 5 Stein's Method and Applications
edited by A. D. Barbour & Louis H. Y. Chen

Vol. 6 Computational Methods in Large Scale Simulation
edited by K.-Y. Lam & H.-P. Lee

Vol. 7 Markov Chain Monte Carlo: Innovations and Applications
edited by W. S. Kendall, F. Liang & J.-S. Wang

FOREWORD

The Institute for Mathematical Sciences at the National University of Singapore was established on 1 July 2000 with funding from the Ministry of Education and the University. Its mission is to provide an international center of excellence in mathematical research and, in particular, to promote within Singapore and the region active research in the mathematical sciences and their applications. It seeks to serve as a focal point for scientists of diverse backgrounds to interact and collaborate in research through tutorials, workshops, seminars and informal discussions.

The Institute organizes thematic programs of duration ranging from one to six months. The theme or themes of each program will be in accordance with the developing trends of the mathematical sciences and the needs and interests of the local scientific community. Generally, for each program there will be tutorial lectures on background material followed by workshops at the research level.

As the tutorial lectures form a core component of a program, the lecture notes are usually made available to the participants for their immediate benefit during the period of the tutorial. The main objective of the Institute's Lecture Notes Series is to bring these lectures to a wider audience. Occasionally, the Series may also include the proceedings of workshops and expository lectures organized by the Institute. The World Scientific Publishing Company and the Singapore University Press have kindly agreed to publish jointly the Lecture Notes Series. This volume, "Markov Chain Monte Carlo: Innovations and Applications" is the seventh of this Series. We hope that through regular publication of lecture notes the Institute will achieve, in part, its objective of promoting research in the mathematical sciences and their applications.

July 2005

Louis H. Y. Chen
Denny Leung
Series Editors

PREFACE

The technique of Markov chain Monte Carlo (MCMC) first arose in statistical physics, marked by the celebrated 1953 paper of Metropolis, Rosenbluth, Rosenbluth, Teller and Teller. The underlying principle is simple: if one wishes to sample randomly from a specific probability distribution then design a Markov chain whose long-time equilibrium is that distribution, write a computer program to simulate the Markov chain, and run the programmed chain for a time long enough to be confident that approximate equilibrium has been attained; finally record the state of the Markov chain as an approximate draw from equilibrium. The Metropolis *et al.* paper used a symmetric Markov chain; later developments included adaptation to the case of non-symmetric Markov chains.

The technique has developed strongly in the statistical physics community but also in separate ways and with rather different emphases in the computer science community concerned with the study of random algorithms (where the emphasis is on whether the resulting algorithm scales well with increasing size of the problem), in the spatial statistics community (where one is interested in understanding what kinds of patterns arise from complex stochastic models), and also in the applied statistics community (where it is applied largely in Bayesian contexts, enabling researchers to formulate statistical models which would otherwise be intransigent to effective statistical analyses).

Within the statistical physics community, the MCMC technique lies at the heart of the tradition of "simulation physics": understanding phase transition and other physical behaviour by constructing careful simulation experiments on the computer. A particular line of development for the past 10 years in the statistical physics community is that of extended ensemble methods, beginning with Berg's work on the multicanonical method, followed by simulated tempering, parallel tempering, broad histogram Monte Carlo, transition matrix Monte Carlo, etc. These methods substantially extend the ability to simulate complicated systems that are very difficult to deal with directly, such as spin-glasses, or protein models.

Within the statistics community, landmark papers include the famous Geman–Geman 1984 paper on image restoration, work by Gelfand and Smith in 1990 showing that MCMC can be applied effectively to Bayesian problems, and Green's (1995) work on dimension-varying problems. The resulting impact on applied statistics has been truly revolutionary.

A recent theoretical development is that of *perfect simulation*, addressing the following central question: how long should one run the Markov chain so as to ensure that it is close to equilibrium? This rather startling development is as follows: in favourable cases one can adjust the Markov chain Monte Carlo algorithm so as to generate exact draws from the target distribution. It was given practical effect in two different ways in the seminal papers of Propp and Wilson (1996) and of Fill (1998). Subsequent work has filled out the mathematical picture by clarifying how recent developments relate to previous work (both the Propp–Wilson and Fill algorithms relate in interesting ways to each other and to prior theoretical concepts, while adding their own attractively empirical flavour).

The development of theory also benefits applications: simulation techniques have been applied to develop practical statistical inferences for almost all problems in (bio)statistics, for example, the problems in longitudinal data analysis, image analysis, genetics, contagious disease epidemics, random spatial pattern, and financial statistical models such as GARCH and stochastic volatility. The techniques also constitute a major part of today's bioinformatics toolbox.

The expositions which make up this book arose from a desire to bring together people who work on innovative developments and applications across the range of statistics, physics, and bioinformatics, to encourage cross-fertilization and to challenge each other with varied problems. The Institute of Mathematical Sciences of the National University of Singapore kindly and generously agreed to fund a month-long programme of activity in March 2004, which allowed us to invite a number of distinguished lecturers from the different fields, to present courses at graduate level; this resulted in a memorable and most productive month, greatly facilitated by the kindness and efficiency of the IMS Director, Prof Louis Chen, and his talented and able staff. The chapters of this book correspond to several of these courses: in Chapter 1 Bernd Berg introduces Markov chain Monte Carlo from the perspective of statistical physics, starting from simple ideas of probability, and developing MCMC ideas right up to multicanonical ensembles, illustrated using FORTRAN computer code available on the web. Chapter 2 presents a complementary view from David Landau, with particular emphasis on issues of finite-size effects, the peculiarities of random number generators, and a spectrum of ingenious techniques to assess phase transition effects. In a change of pace, Wilfrid Kendall uses Chapter 3 to describe the various ideas involved in the transformations of algorithms known collectively as perfect simulation, which in favourable cases deliver exact draws in random rather than deterministic runtimes. A very different theme is treated by Rong Chen in Chapter 4: simulation algorithms that process information sequentially (known as Sequential Monte Carlo), either because this is natural to the algorithm itself, or because it is useful to decompose the problem in such a manner. Finally, in Chapter 5, Elizabeth Thompson presents a careful study of how MCMC is put into practice in the analysis of pedigrees in genetics.

It is our hope as editors that this ensemble of expositions, and the diversity of ideas contained therein, will form an attractive invitation to readers, to introduce them to the fascinating and various worlds of Markov chain Monte Carlo in mathematical science. We trust you will enjoy this book as much as we enjoyed the task of its compilation!

W. S. Kendall
F. Liang
J.-S. Wang

GLOSSARY

Contributors to this volume come from several different fields, each with their own preferred terminology, which can often overlap. To aid the reader, we have therefore assembled the following glossary of terms and brief definitions, which we have organized under the titles of Probability, Statistical Physics, and Mathematical Genetics.

1. Probability

- *Bernoulli distribution:* A random variable X has a Bernoulli distribution if it has probability p of being equal to 1, probability $1-p$ of being equal to -1.
- *coalescence:* A family of random processes X, Y, Z, $\ldots$ are said to coalesce if there is some random time T (the *coalescence time*) at which they are all equal: $X(T) = Y(T) = Z(T) = \ldots$. Sometimes called *grand coupling*, since two processes X, Y are said to *couple* if $X(T) = Y(T)$ for some random time T.
- *conditional probability:* The conditional probability $\mathbb{P}[A|B]$ of A given B is the ratio $\mathbb{P}[A \text{ and } B]/\mathbb{P}[B]$ of the probability of both A and B to the probability of B.
- *coupling technique:* The technique of constructing two random processes X and Y such that (a) individually both X and Y have specified statistical behaviour, but (b) the joint behaviour of X and Y meets some useful requirement (perhaps X always lies below Y, or perhaps X and Y *couple* with $X(T) = Y(T)$ at some random time T, or $\ldots$). *Coupling from the Past* (CFTP) uses coupling techniques to convert favourable MCMC algorithms into exact simulation algorithms.
- *Gibbs sampler:* A specific form of MCMC in which values X_n at successive sites n are updated using the full conditional distribution of X_n given the values X_m at all other sites $m \neq n$. Also

known as the *heat bath sampler*. Successive sites n may be chosen *systematically* or *randomly*.

- *Ising model:* A random field, giving a random value or *spin* $X_{i,j} = \pm 1$ to each site (i, j). The probability distribution of the value $X_{i,j} = \pm 1$ depends on the pattern of its neighbouring values.
- *occlusion:* A term used in image analysis when one item partially covers or *occludes* another item.
- *Poisson process:* A random point pattern such that the number $X(A)$ of points falling in a region A has a Poisson distribution of mean proportional to the size of A; numbers of points falling in non-overlapping regions are statistically independent.
- *posterior distribution:* The conditional probability distribution of an unknown parameter θ after data is observed, and hence conditional on that data. Thus if we observe the result $Y = y$ then the posterior distribution of θ lying in the region A is given by the conditional probability $\mathbb{P}[\theta \in A | Y = y]$.
- *prior distribution:* The probability distribution of an unknown parameter θ before data is observed (in the Bayesian paradigm of statistics, the prior distribution expresses one's beliefs about what value might be taken by θ).
- *probability distribution:* The probability measure obtained by considering the probabilities $\mathbb{P}[X \in A]$ of a random object X taking on values in various regions A. Often abbreviated to the *distribution* of X, used as shorthand to refer to the statistical behaviour of X considered on its own.
- *probability measure:* The mathematical entity capturing the notion of probability: informally, a probability measure $\mathbb{P}$ assigns a probability $\mathbb{P}[A]$ to each of a family of possible events A. Probability measures must obey additive and countably-additive laws, and their values must lie between 0 (expressing almost impossibility) and 1 (expressing almost certainty).
- *random walk:* A random process which moves by independent identically distributed jumps.
- *resampling:* Given a set of values obtained by drawing from a probability distribution (for example, the sample obtained at step n of a sequential Monte Carlo scheme); *resampling* is the procedure of drawing a new sample from these values, typically according to appropriate *resampling weights*.

2. Statistical Physics

- *autocorrelation time:* A typical time scale for the dynamical correlation (time-displacement) function, $\langle Q(t)Q(0)\rangle - \langle Q(0)\rangle^2$, for some observable Q.
- *Boltzmann weight:* When system is in thermal equilibrium, the probability of a state is assumed to be proportional to $e^{-E/(k_B T)}$, where E is the energy of the state, T is temperature, and k_B Boltzmann constant. Such a distribution is also called the canonical distribution.
- *canonical (Gibbs) ensemble:* see Boltzmann weight.
- *coexisting phases:* A particular set of model parameters or physical conditions, such that two or more phases exist, such as the coexistence of water and ice.
- *coupling constant:* the constant J in Ising model where the energy is given by $-J\sum_{\langle i,j\rangle}\sigma_i\sigma_j$.
- *dynamic universality class:* A class of models with the same static and dynamic (time-dependent) critical exponents in a second-order phase transition.
- *energy function:* The total energy of a system, also known as Hamiltonian.
- *entropy:* One of the most important thermal dynamic functions related to the degree of disorder. It is given by Boltzmann's famous formula $S = k_B \ln \Omega$ where Ω is the number of microstates.
- *equilibration:* The Monte Carlo steps used to let the system reach equilibrium or limiting distribution.
- *external magnetic field:* extra term of energy in the form e.g., $-h\sigma_i$, in an Ising model; h is called the magnetic field.
- *free energy:* thermodynamical functions, defined, e.g. for Helmholtz free energy, $F = -k_B T \ln Z$,where Z is partition function.
- *Glauber dynamics:* A Markov chain dynamics in continuous time, with the transition rate $\sigma_i \to -\sigma_i$ (for the case of Ising model),

$$\frac{1}{2}\left[1 - \sigma_i \tanh((k_B T)^{-1}\sum_j J_{ij}\sigma_j)\right].$$

- *Hamiltonian:* (a) The energy function viewed as variables of coordinates and momenta. This function $H(p,q)$ gives the Hamilton's equation of motion, $\dot{q} = \partial H/\partial p$, $\dot{p} = -\partial H/\partial q$. (b) the operator in Schrödinger's equation $i\hbar\partial\Psi/\partial t = \hat{H}\Psi$. (c) Sometimes the terminology is used more loosely; the "Hamiltonian" can be used to refer to the energy of a system.

- *heat-bath sampler:* See *Gibbs sampler.*
- *heat-bath update:* A single step of a *heat bath* or *Gibbs sampler* (see elsewhere).
- *Helmholtz free energy:* see free energy.
- *hysteresis:* a metastable process where increasing a parameter of a model slowly (say the magnetic field) from h_0 to h_1 traces out a function $f_+(h)$ which does not agree with a reverse process of h_1 to h_0 in same observable $f_-(h)$.
- *metastable:* A state of a system which looks like in equilibrium for finite period of times, but is in fact not in equilibrium in the limit of time going to infinity.
- *Metropolis update:* A popular choice of a transition rate in Monte Carlo dynamics, with a form $\min\big[1, \exp\big(-(E'-E)/(k_B T)\big)\big]$, where E' is new energy and E is old energy.
- *microcanonical temperature:* defined as $1/T = \partial S/\partial E$ where S is entropy, and E is (internal) energy.
- *microstate, configuration:* A state described by a set of dynamical variables, also known as a configuration.
- *multicanonical simulation:* A Monte Carlo simulation in an artificial ensemble with probability distribution of energy being a constant.
- *$O(3)$ σ-model:* the $O(n)$ model with $n = 3$.
- *$O(n)$ model:* a model with Hamiltonian $-\sum_{ij} J_{ij} \boldsymbol{\sigma}_i \cdot \boldsymbol{\sigma}_j$, where $\boldsymbol{\sigma}_i$ is an n-dimensional unit vector.
- *observable:* Average of any variables with respect to a distribution.
- *partition function:* The sum of the Boltzmann weights over all states, or the normalization constant of the Boltzmann distribution function, commonly denoted by Z.
- *Potts model, Potts spin:* A generalization of the Ising model with Hamiltonian $-\sum_{i,j} J_{ij} \delta_{\sigma_i, \sigma_j}$, where $\sigma_i = 1, 2, ..., q$ is known as Potts spins.
- *specific heat, heat capacity:* Defined as $d\langle E\rangle/dT$, where $\langle E\rangle$ is the ensemble average of energy, and T is temperature.
- *supercritical slowing down:* is a process with correlation times that depend on system dimensions exponentially, such as in a first-order phase transition.

3. Mathematical Genetics

- *allele:* One of two or more alternative forms of a gene, only one of which can be present in a chromosome.
- *Baum-Welch algorithm:* An algorithm to estimate hidden Markov model parameters with the maximum likelihood of generating the given symbol sequence in the observation vector.
- *centromere:* A constricted region of a chromosome that joins the two sister chromatids to each other during cell divisions. See also *chromatid.*
- *chromatid:* A duplicated chromosome that is held together in the middle. On each of the sides of the chromatid is an exact copy of the original chromosome.
- *crossover:* The process of exchange of genetic material between pairs of homologous chromosomes during meiosis. See also *homologous* and *meiosis.*
- *DNA:* An abbreviation for **d**eoxyribo**n**ucleic **a**cid. The material inside the nucleus of cells that carries genetic information.
- *genetic interference:* The effect that the presence of one crossover reduces the chance of another occurring in its vicinity.
- *genetic marker:* Sequence of DNA that can be easily identified and which therefore can be used as a reference point for mapping other genes.
- *genome:* The total genetic material of an organism, comprising the genes carried on its chromosomes.
- *genotype:* The genetic information carried by a pair of alleles. See also *allele.*
- *hidden Markov Model:* (in bioinformatics) A probabilistic model used to align and analyze DNA or protein sequence datasets by generalization from a sequence profile.
- *homologous:* (in genetics) Describing a pair of chromosomes having identical gene loci. One member of the pair is derived from the mother, the other from the father. See also *locus.*
- *inheritance:* The transmission of genetic characteristics from parents to offspring.
- *linkage:* The tendency for certain genes to be inherited together due to their physical proximity on the chromosome.
- *locus:* (in genetics) A position on a chromosome occupied by a gene.
- *lod score:* The likelihood (value) that two genes are linked.

- *MCMC:* Abbreviation for **M**arkov **c**hain **M**onte **C**arlo.
- *meiosis:* A type of nuclear division such that each child nuclei contains half the number of chromosomes of the parent.
- *Metropolis-Hasting algorithm:* A Markov chain Monte Carlo algorithm that is used to simulate from a complex distribution. See also *Metropolis update.*
- *missing data:* Missing data occur when some or all of the values for a sampled unit are absent in the dataset.
- *mitosis:* A type of cell division in which a single cell produces two genetically identical cells.
- *pedigree:* (in genetics) A digram showing the descent relationship of a group of related individuals.
- *peeling algorithm:* (in bioinformatics) An algorithm for computing the likelihood of an evolutionary tree.
- *phenotype:* Visible biochemical characteristics of an organism that are produced by the interaction of the genes and the environment.
- *quantitative trait:* A phenotypic character associated with particular genes. The phenotype can be described with a continuous (rather than a discrete) distribution. See also *phenotype.*
- *Rao-Blackwellization:* In statistics, it refers to the "Rao-Blackwell theorem" by C. R. Rao (1945 Bull. Calcutta Math. Soc. 37, 81-91) and David Blackwell (1947 Ann. Math. Stat., 18, 105-110).
- *recombination:* The process of exchange of DNA between homologous chromosomes in sexually reproducing organisms. See also *homologous.*
- *sequential imputation:* A sequential Monte Carlo algorithm which is used to impute the missing data.
- *SNP:* An abbreviation for **S**ingle **N**ucleotide **P**olymorphism, a single basepair change in a sequence of DNA.

CONTENTS

Foreword vii

Preface ix

Glossary xiii

Introduction to Markov Chain Monte Carlo Simulations and Their Statistical Analysis
B. A. Berg 1

An Introduction to Monte Carlo Methods in Statistical Physics
D. P. Landau 53

Notes on Perfect Simulation
W. S. Kendall 93

Sequential Monte Carlo Methods and Their Applications
R. Chen 147

MCMC in the Analysis of Genetic Data on Pedigrees
E. A. Thompson 183

Index 217

INTRODUCTION TO MARKOV CHAIN MONTE CARLO SIMULATIONS AND THEIR STATISTICAL ANALYSIS

Bernd A. Berg

Department of Physics
Florida State University
Tallahassee, Florida 32306-4350, USA
and
School of Computational Science
Florida State University
Tallahassee, Florida 32306-4120, USA
E-mail: berg@csit.fsu.edu

This article is a tutorial on Markov chain Monte Carlo simulations and their statistical analysis. The theoretical concepts are illustrated through many numerical assignments from the author's book [7] on the subject. Computer code (in Fortran) is available for all subjects covered and can be downloaded from the web.

Contents

1 Introduction 2
2 Probability Distributions and Sampling 3
3 Random Numbers and Fortran Code 4
3.1 How to Get and Run the Fortran Code 5
4 Confidence Intervals and Heapsort 6
5 The Central Limit Theorem and Binning 8
6 Gaussian Error Analysis for Large and Small Samples 11
6.1 χ^2 Distribution, Error of the Error Bar, F-Test 15
6.2 The Jackknife Approach 16
7 Statistical Physics and Potts Models 17
8 Sampling and Re-weighting 19
9 Importance Sampling and Markov Chain Monte Carlo 21
9.1 Metropolis and Heat Bath Algorithm for Potts Models 23
9.2 The $O(3)$ σ Model and the Heat Bath Algorithm 25