Monte Carlo Methods in Statistical Physics

M. E. J. NEWMAN

Santa Fe Institute

and

G. T. BARKEMA

Institute for Theoretical Physics
Utrecht University

CLARENDON PRESS • OXFORD

OXFORD
UNIVERSITY PRESS
Great Clarendon Street, Oxford OX2 6DP
Oxford University Press is a department of the University of Oxford.
It furthers the University's objective of excellence in research, scholarship,
and education by publishing worldwide in
Oxford New York
Athens Auckland Bangkok Bogotá Buenos Aires Cape Town
Chennai Dar es Salaam Delhi Florence Hong Kong Istanbul Karachi
Kolkata Kuala Lumpur Madrid Melbourne Mexico City Mumbai Nairobi
Paris São Paulo Shanghai Singapore Taipei Tokyo Toronto Warsaw
with associated companies in Berlin Ibadan
Oxford is a registered trade mark of Oxford University Press
in the UK and in certain other countries
Published in the United States
by Oxford University Press Inc., New York

First published 1999
Reprinted (with corrections) 2001

A catalogue record for this book is available from the British Library
Library of Congress Cataloging in Publication Data
(Data available)
ISBN 0 19 851796 3 (Hbk)
ISBN 0 19 851797 1 (Pbk)
Printed in India by Thomson Press

Preface

This book is intended for those who are interested in the use of Monte Carlo simulations in classical statistical mechanics. It would be suitable for use in a course on simulation methods or on statistical physics. It would also be a good choice for those who wish to teach themselves about Monte Carlo methods, or for experienced researchers who want to learn more about some of the sophisticated new simulation techniques which have appeared in the last decade or so.

The primary goal of the book is to explain how to perform Monte Carlo simulations efficiently. For many people, Monte Carlo simulation just means applying the Metropolis algorithm to the problem in hand. Although this famous algorithm is very easy to program, it is rarely the most efficient way to perform a simulation. The Metropolis algorithm is certainly important and we do discuss it in some detail (Chapter 3 is devoted to it), but we also show that for most problems a little work with a pencil and paper can usually turn up a better algorithm, in some cases thousands or millions of times faster. In recent years there has been quite a flurry of interesting new Monte Carlo algorithms described in the literature, many of which are specifically designed to accelerate the simulation of particular classes of problems in statistical physics. Amongst others, we describe cluster algorithms, multigrid methods, non-local algorithms for conserved-order-parameter models, entropic sampling, simulated tempering and continuous time Monte Carlo. The book is divided into parts covering equilibrium and non-equilibrium simulations, and throughout we give pointers to how the algorithms can be most efficiently implemented. At the end of the book we include a number of chapters on general implementation issues for Monte Carlo simulations. We also cover data analysis methods in some detail, including generic methods for estimating observable quantities, equilibration and correlation times, correlation functions, and standard errors, as well as a number of techniques which are specific to Monte Carlo simulation, such as the single and multiple histogram methods, finite-size scaling and the Monte Carlo renormalization group.

The *modus operandi* of this book is teaching by example. We have tried

to include as many as possible of the important Monte Carlo algorithms in use today, and each one we introduce in the context of a particular model or models. For example, we illustrate the Metropolis algorithm by applying it to the simulation of the Ising model. We have not assumed however that the reader is familiar with the models studied, and give a brief outline of the physics behind each one at the start of the corresponding chapter. All we assume is that the reader has a working knowledge of basic statistical mechanics and thermodynamics at the level typical of a student beginning graduate study in physics. Reasonable fluency in a computer programming language such as FORTRAN or C will certainly be necessary if you actually want to write a Monte Carlo program yourself, but the book can be understood without it. We have avoided giving examples of actual computer code in the body of the book. There are a couple of reasons for this. Doing so would require the reader to know a particular language, or to learn that language if he or she was not already familiar with it. Furthermore, we do not believe that inclusion of the code helps much with the understanding of an algorithm. It is better to get a clear idea of the principles behind an algorithm by working through the physics and mathematics involved than to try to learn by reading someone else's program. With a clear understanding of the principles, you should be able to write your own program with little difficulty. However, inspecting other people's programs can be useful in one respect: it is a good way to learn programming tricks and techniques for writing efficient code. For this reason we have included programs for some of the more common Monte Carlo algorithms in an appendix at the end of the book. The programs are written in C, which is fast replacing FORTRAN as the most commonly used language for scientific programming.

We have also included a number of problems at the end of each chapter for the reader to work through if he or she wishes. Some of these are purely analytic and can be done on paper. Others ask the reader to write a short computer program. Answers to the analytic problems are given at the end of the book. The ones which require you to write a computer program have many equally good solutions, so we have by and large resorted to giving hints rather than answers for these problems.

There are many people and organizations who have assisted us in the writing of this book. We would like to thank our editors Sönke Adlung, Donald Degenhardt and Julia Tompson at Oxford University Press for their help and patience. We are also grateful to a number of institutions who have offered us hospitality during the four-year process of preparing the manuscript, including Cornell University, Oxford University, the Institute for Advanced Study in Princeton, Forschungszentrum Jülich, the Santa Fe Institute and Utrecht University. Finally, we would like to thank the many colleagues and friends who have offered suggestions and encouragement, including James Binney, Geoffrey Chester, Eytan Domany, Peter Grass-

berger, Harvey Gould, Daniel Kandel, Yongyut Laosiritaworn, Jim Louck, Jon Machta, Nick Metropolis, Cris Moore, Richard Palmer, Gunter Schütz, Jim Sethna, Kan Shen, Alan Sokal and Ben Widom. Responsibility for any mistakes which may lurk in the text rests of course with the authors. We would be very grateful to learn from our keen-eyed readers of any such problems.

June 1998

Mark Newman
Santa Fe, New Mexico, USA
Gerard Barkema
Utrecht, The Netherlands

This page intentionally left blank

Contents

II Out-of-equilibrium simulations

III Implementation

Part I

Equilibrium Monte Carlo simulations

This page intentionally left blank

1

Introduction

This book is about the use of computers to solve problems in statistical physics. In particular, it is about **Monte Carlo methods**, which form the largest and most important class of numerical methods used for solving statistical physics problems. In this opening chapter of the book we look first at what we mean by statistical physics, giving a brief overview of the discipline we call **statistical mechanics**. Whole books have been written on statistical mechanics, and our synopsis takes only a few pages, so we must necessarily deal only with the very basics of the subject. We are assuming that these basics are actually already familiar to you, but writing them down here will give us a chance to bring back to mind some of the ideas that are most relevant to the study of Monte Carlo methods. In this chapter we also look at some of the difficulties associated with solving problems in statistical physics using a computer, and outline what Monte Carlo techniques are, and why they are useful. In the last section of the chapter, purely for fun, we give a brief synopsis of the history of computational physics and Monte Carlo methods.

1.1 Statistical mechanics

Statistical mechanics is primarily concerned with the calculation of properties of condensed matter systems. The crucial difficulty associated with these systems is that they are composed of very many parts, typically atoms or molecules. These parts are usually all the same or of a small number of different types and they often obey quite simple equations of motion so that the behaviour of the entire system can be expressed mathematically in a straightforward manner. But the sheer number of equations—just the magnitude of the problem—makes it impossible to solve the mathematics exactly. A standard example is that of a volume of gas in a container. One

litre of, say, oxygen at standard temperature and pressure consists of about 3×10^{22} oxygen molecules, all moving around and colliding with one another and the walls of the container. One litre of air under the same conditions contains the same number of molecules, but they are now a mixture of oxygen, nitrogen, carbon dioxide and a few other things. The atmosphere of the Earth contains 4×10^{21} litres of air, or about 1×10^{44} molecules, all moving around and colliding with each other and the ground and trees and houses and people. These are large systems. It is not feasible to solve Hamilton's equations for these systems because there are simply too many equations, and yet when we look at the macroscopic properties of the gas, they are very well-behaved and predictable. Clearly, there is something special about the behaviour of the solutions of these many equations that "averages out" to give us a predictable behaviour for the entire system. For example, the pressure and temperature of the gas obey quite simple laws although both are measures of rather gross average properties of the gas. Statistical mechanics attempts to side-step the problem of solving the equations of motion and cut straight to the business of calculating these gross properties of large systems by treating them in a probabilistic fashion. Instead of looking for exact solutions, we deal with the probabilities of the system being in one state or another, having this value of the pressure or that—hence the name *statistical* mechanics. Such probabilistic statements turn out to be extremely useful, because we usually find that for large systems the range of behaviours of the system that are anything more than phenomenally unlikely is very small; all the reasonably probable behaviours fall into a narrow range, allowing us to state with extremely high confidence that the real system will display behaviour within that range. Let us look at how statistical mechanics treats these systems and demonstrates these conclusions.

The typical paradigm for the systems we will be studying in this book is one of a system governed by a Hamiltonian function H which gives us the total energy of the system in any particular state. Most of the examples we will be looking at have discrete sets of states each with its own energy, ranging from the lowest, or ground state energy E_0 upwards, E_1, E_2, $E_3 \ldots$, possibly without limit. Statistical mechanics, and the Monte Carlo methods we will be introducing, are also applicable to systems with continuous energy spectra, and we will be giving some examples of such applications.

If our Hamiltonian system were all we had, life would be dull. Being a Hamiltonian system, energy would be conserved, which means that the system would stay in the same energy state all the time (or if there were a number of degenerate states with the same energy, maybe it would make transitions between those, but that's as far as it would get).[1] However,

[1] For a classical system which has a continuum of energy states there can be a continuous set of degenerate states through which the system passes, and an average over those states can sometimes give a good answer for certain properties of the system. Such sets of

there's another component to our paradigm, and that is the **thermal reservoir**. This is an external system which acts as a source and sink of heat, constantly exchanging energy with our Hamiltonian system in such a way as always to push the temperature of the system—defined as in classical thermodynamics—towards the temperature of the reservoir. In effect the reservoir is a weak perturbation on the Hamiltonian, which we ignore in our calculation of the energy levels of our system, but which pushes the system frequently from one energy level to another. We can incorporate the effects of the reservoir in our calculations by giving the system a **dynamics**, a rule whereby the system changes periodically from one state to another. The exact nature of the dynamics is dictated by the form of the perturbation that the reservoir produces in the Hamiltonian. We will discuss many different possible types of dynamics in the later chapters of this book. However, there are a number of general conclusions that we can reach without specifying the exact form of the dynamics, and we will examine these first.

Suppose our system is in a state μ. Let us define $R(\mu \to \nu)\,\mathrm{d}t$ to be the probability that it is in state ν a time $\mathrm{d}t$ later. $R(\mu \to \nu)$ is the **transition rate** for the transition from μ to ν. The transition rate is normally assumed to be time-independent and we will make that assumption here. We can define a transition rate like this for every possible state ν that the system can reach. These transition rates are usually all we know about the dynamics, which means that even if we know the state μ that the system starts off in, we need only wait a short interval of time and it could be in any one of a very large number of other possible states. This is where our probabilistic treatment of the problem comes in. We define a set of weights $w_\mu(t)$ which represent the probability that the system will be in state μ at time t. Statistical mechanics deals with these weights, and they represent our entire knowledge about the state of the system. We can write a **master equation** for the evolution of $w_\mu(t)$ in terms of the rates $R(\mu \to \nu)$ thus:[2]

$$\frac{\mathrm{d}w_\mu}{\mathrm{d}t} = \sum_\nu \bigl[w_\nu(t)R(\nu \to \mu) - w_\mu(t)R(\mu \to \nu)\bigr]. \tag{1.1}$$

The first term on the right-hand side of this equation represents the rate at which the system is undergoing transitions into state μ; the second term is the rate at which it is undergoing transitions out of μ into other states. The probabilities $w_\mu(t)$ must also obey the sum rule

$$\sum_\mu w_\mu(t) = 1 \tag{1.2}$$

---

degenerate states are said to form a **microcanonical ensemble**. The more general case we consider here, in which there is a thermal reservoir causing the energy of the system to fluctuate, is known as a **canonical ensemble**.

[2]The master equation is really a set of equations, one for each state μ, although people always call it *the* master equation, as if there were only one equation here.

for all t, since the system must always be in *some* state. The solution of Equation (1.1), subject to the constraint (1.2), tells us how the weights w_μ vary over time.

And how are the weights w_μ related to the macroscopic properties of the system which we want to know about? Well, if we are interested in some quantity Q, which takes the value Q_μ in state μ, then we can define the **expectation** of Q at time t for our system as

$$\langle Q \rangle = \sum_\mu Q_\mu w_\mu(t). \tag{1.3}$$

Clearly this quantity contains important information about the real value of Q that we might expect to measure in an experiment. For example, if our system is definitely in one state τ then $\langle Q \rangle$ will take the corresponding value Q_τ. And if the system is equally likely to be in any of perhaps three states, and has zero probability of being in any other state, then $\langle Q \rangle$ is equal to the mean of the values of Q in those three states, and so forth. However, the precise relation of $\langle Q \rangle$ to the observed value of Q is perhaps not very clear. There are really two ways to look at it. The first, and more rigorous, is to imagine having a large number of copies of our system all interacting with their own thermal reservoirs and whizzing between one state and another all the time. $\langle Q \rangle$ is then a good estimate of the number we would get if we were to measure the instantaneous value of the quantity Q in each of these systems and then take the mean of all of them. People who worry about the conceptual foundations of statistical mechanics like to take this "many systems" approach to defining the expectation of a quantity.[3] The trouble with it however is that it's not very much like what happens in a real experiment. In a real experiment we normally only have one system and we make all our measurements of Q on that system, though we probably don't just make a single instantaneous measurement, but rather integrate our results over some period of time. There is another way of looking at the expectation value which is similar to this experimental picture, though it is less rigorous than the many systems approach. This is to envisage the expectation as a *time average* of the quantity Q. Imagine recording the value of Q every second for a thousand seconds and taking the average of those one thousand values. This will correspond roughly to the quantity calculated in Equation (1.3) as long as the system passes through a representative selection of the states in the probability distribution w_μ in those thousand seconds. And if we make ten thousand measurements of Q instead of one thousand,

[3] In fact the word *ensemble*, as in the "canonical ensemble" which was mentioned in a previous footnote, was originally introduced by Gibbs to describe an ensemble of *systems* like this, and not an ensemble of, say, molecules, or any other kind of ensemble. These days however, use of this word no longer implies that the writer is necessarily thinking of a many systems formulation of statistical mechanics.

or a million or more, we will get an increasingly accurate fit between our experimental average and the expectation $\langle Q \rangle$.

Why is this a less rigorous approach? The main problem is the question of what we mean by a "representative selection of the states". There is no guarantee that the system will pass through anything like a representative sample of the states of the system in our one thousand seconds. It could easily be that the system only hops from one state to another every ten thousand seconds, and so turns out to be in the same state for all of our one thousand measurements. Or maybe it changes state very rapidly, but because of the nature of the dynamics spends long periods of time in small portions of the state space. This can happen for example if the transition rates $R(\mu \to \nu)$ are only large for states of the system that differ in very small ways, so that the only way to make a large change in the state of the system is to go through very many small steps. This is a very common problem in a lot of the systems we will be looking at in this book. Another potential problem with the time average interpretation of (1.3) is that the weights $w_\mu(t)$, which are functions of time, may change considerably over the course of our measurements, making the expression invalid. This can be a genuine problem in both experiments and simulations of non-equilibrium systems, which are the topic of the second part of this book. For equilibrium systems, as discussed below, the weights are by definition not time-varying, so this problem does not arise.

Despite these problems however, this time-average interpretation of the expectation value of a quantity is the most widely used and most experimentally relevant interpretation, and it is the one that we will adopt in this book. The calculation of expectation values is one of the fundamental goals of statistical mechanics, and of Monte Carlo simulation in statistical physics, and much of our time will be concerned with it.

1.2 Equilibrium

Consider the master equation (1.1) again. If our system ever reaches a state in which the two terms on the right-hand side exactly cancel one another for all μ, then the rates of change $\mathrm{d}w_\mu/\mathrm{d}t$ will all vanish and the weights will all take constant values for the rest of time. This is an **equilibrium** state. Since the master equation is first order with real parameters, and since the variables w_μ are constrained to lie between zero and one (which effectively prohibits exponentially growing solutions to the equations) we can see that all systems governed by these equations must come to equilibrium in the end. A large part of this book will be concerned with Monte Carlo techniques for simulating equilibrium systems and in this section we develop some of the important statistical mechanical concepts that apply to these systems.

The transition rates $R(\mu \to \nu)$ appearing in the master equation (1.1)

do not just take *any* values. They take particular values which arise out of the thermal nature of the interaction between the system and the thermal reservoir. In the later chapters of this book we will have to choose values for these rates when we simulate thermal systems in our Monte Carlo calculations, and it is crucial that we choose them so that they mimic the interactions with the thermal reservoir correctly. The important point is that we know *a priori* what the equilibrium values of the weights w_μ are for our system. We call these equilibrium values the **equilibrium occupation probabilities** and denote them by

$$p_\mu = \lim_{t\to\infty} w_\mu(t). \tag{1.4}$$

It was Gibbs (1902) who showed that for a system in thermal equilibrium with a reservoir at temperature T, the equilibrium occupation probabilities are

$$p_\mu = \frac{1}{Z}\,\mathrm{e}^{-E_\mu/kT}. \tag{1.5}$$

Here E_μ is the energy of state μ and k is Boltzmann's constant, whose value is $1.38 \times 10^{-23}\ \mathrm{J\,K^{-1}}$. It is conventional to denote the quantity $(kT)^{-1}$ by the symbol β, and we will follow that convention in this book. Z is a normalizing constant, whose value is given by

$$Z = \sum_\mu \mathrm{e}^{-E_\mu/kT} = \sum_\mu \mathrm{e}^{-\beta E_\mu}. \tag{1.6}$$

Z is also known as the **partition function**, and it figures a lot more heavily in the mathematical development of statistical mechanics than a mere normalizing constant might be expected to. It turns out in fact that a knowledge of the variation of Z with temperature and any other parameters affecting the system (like the volume of the box enclosing a sample of gas, or the magnetic field applied to a magnet) can tell us virtually everything we might want to know about the macroscopic behaviour of the system. The probability distribution (1.5) is known as the **Boltzmann distribution**, after Ludwig Boltzmann, one of the pioneers of statistical mechanics. For a discussion of the origins of the Boltzmann distribution and the arguments that lead to it, the reader is referred to the exposition by Walter Grandy in his excellent book *Foundations of Statistical Mechanics* (1987). In our treatment we will take Equation (1.5) as our starting point for further developments.

From Equations (1.3), (1.4) and (1.5) the expectation of a quantity Q for a system in equilibrium is

$$\langle Q\rangle = \sum_\mu Q_\mu p_\mu = \frac{1}{Z}\sum_\mu Q_\mu\,\mathrm{e}^{-\beta E_\mu}. \tag{1.7}$$

For example, the expectation value of the energy $\langle E \rangle$, which is also the quantity we know from thermodynamics as the internal energy U, is given by

$$U = \frac{1}{Z} \sum_{\mu} E_{\mu} \, \mathrm{e}^{-\beta E_{\mu}}. \tag{1.8}$$

From Equation (1.6) we can see that this can also be written in terms of a derivative of the partition function:

$$U = -\frac{1}{Z} \frac{\partial Z}{\partial \beta} = -\frac{\partial \log Z}{\partial \beta}. \tag{1.9}$$

The specific heat is given by the derivative of the internal energy:

$$C = \frac{\partial U}{\partial T} = -k\beta^2 \frac{\partial U}{\partial \beta} = k\beta^2 \frac{\partial^2 \log Z}{\partial \beta^2}. \tag{1.10}$$

However, from thermodynamics we know that the specific heat is also related to the entropy:

$$C = T \frac{\partial S}{\partial T} = -\beta \frac{\partial S}{\partial \beta}, \tag{1.11}$$

and, equating these two expressions for C and integrating with respect to β, we find the following expression for the entropy:

$$S = -k\beta \frac{\partial \log Z}{\partial \beta} + k \log Z. \tag{1.12}$$

(There is in theory an integration constant in this equation, but it is set to zero under the convention known as the third law of thermodynamics, which fixes the arbitrary origin of entropy by saying that the entropy of a system should tend to zero as the temperature does.) We can also write an expression for the (Helmholtz) free energy F of the system, using Equations (1.9) and (1.12):

$$F = U - TS = -kT \log Z. \tag{1.13}$$

We have thus shown how U, F, C and S can all be calculated directly from the partition function Z. The last equation also tells us how we can deal with other parameters affecting the system. In classical thermodynamics, parameters and constraints and fields interacting with the system each have conjugate variables which represent the response of the system to the perturbation of the corresponding parameter. For example, the response of a gas system in a box to a change in the confining volume is a change in the pressure of the gas. The pressure p is the conjugate variable to the parameter V. Similarly, the magnetization M of a magnet changes in response

to the applied magnetic field B; M and B are conjugate variables. Thermodynamics tells us that we can calculate the values of conjugate variables from derivatives of the free energy:

$$p = -\frac{\partial F}{\partial V}, \tag{1.14}$$

$$M = \frac{\partial F}{\partial B}. \tag{1.15}$$

Thus, if we can calculate the free energy using Equation (1.13), then we can calculate the effects of parameter variations too.

In performing Monte Carlo calculations of the properties of equilibrium systems, it is sometimes appropriate to calculate the partition function and then evaluate other quantities from it. More often it is better to calculate the quantities of interest directly, but many times in considering the theory behind our simulations we will return to the idea of the partition function, because in principle the entire range of thermodynamic properties of a system can be deduced from this function, and any numerical method that can make a good estimate of the partition function is at heart a sound method.

1.2.1 Fluctuations, correlations and responses

Statistical mechanics can tell us about other properties of a system apart from the macroscopic ones that classical equilibrium thermodynamics deals with such as entropy and pressure. One of the most physically interesting classes of properties is **fluctuations** in observable quantities. We described in the first part of Section 1.1 how the calculation of an expectation could be regarded as a time average over many measurements of the same property of a single system. In addition to calculating the mean value of these many measurements, it is often useful also to calculate their standard deviation, which gives us a measure of the variation over time of the quantity we are looking at, and so tells us quantitatively how much of an approximation we are making by giving just the one mean value for the expectation. To take an example, let us consider the internal energy again. The mean square deviation of individual, instantaneous measurements of the energy away from the mean value $U = \langle E \rangle$ is

$$\langle (E - \langle E \rangle)^2 \rangle = \langle E^2 \rangle - \langle E \rangle^2. \tag{1.16}$$

We can calculate $\langle E^2 \rangle$ from derivatives of the partition function in a way similar to our calculation of $\langle E \rangle$:

$$\langle E^2 \rangle = \frac{1}{Z} \sum_{\mu} E_{\mu}^2 \mathrm{e}^{-\beta E_{\mu}} = \frac{1}{Z} \frac{\partial^2 Z}{\partial \beta^2}. \tag{1.17}$$

So

$$\langle E^2 \rangle - \langle E \rangle^2 = \frac{1}{Z}\frac{\partial^2 Z}{\partial \beta^2} - \left[\frac{1}{Z}\frac{\partial Z}{\partial \beta}\right]^2 = \frac{\partial^2 \log Z}{\partial \beta^2}. \tag{1.18}$$

Using Equation (1.10) to eliminate the second derivative, we can also write this as

$$\langle E^2 \rangle - \langle E \rangle^2 = \frac{C}{k\beta^2}. \tag{1.19}$$

And the standard deviation of E, the RMS fluctuation in the internal energy, is just the square root of this expression.

This result is interesting for a number of reasons. First, it gives us the magnitude of the fluctuations in terms of the specific heat C or alternatively in terms of $\log Z = -\beta F$. In other words we can calculate the fluctuations entirely from quantities that are available within classical thermodynamics. However, this result could never have been derived within the framework of thermodynamics, since it depends on microscopic details that thermodynamics has no access to. Second, let us look at what sort of numbers we get out for the size of the energy fluctuations of a typical system. Let us go back to our litre of gas in a box. A typical specific heat for such a system is $1\ \mathrm{J\,K^{-1}}$ at room temperature and atmospheric pressure, giving RMS energy fluctuations of about 10^{-18} J. The internal energy itself on the other hand will be around 10^2 J, so the fluctuations are only about one part in 10^{20}. This lends some credence to our earlier contention that statistical treatments can often give a very accurate estimate of the expected behaviour of a system. We see that in the case of the internal energy at least, the variation of the actual value of U around the expectation value $\langle E \rangle$ is tiny by comparison with the kind of energies we are considering for the whole system, and probably not within the resolution of our measuring equipment. So quoting the expectation value gives a very good guide to what we should expect to see in an experiment. Furthermore, note that, since the specific heat C is an extensive quantity, the RMS energy fluctuations, which are the square root of Equation (1.19), scale like $\sqrt{V}$ with the volume V of the system. The internal energy itself on the other hand scales like V, so that the relative size of the fluctuations compared to the internal energy decreases as $1/\sqrt{V}$ as the system becomes large. In the limit of a very large system, therefore, we can ignore the fluctuations altogether. For this reason, the limit of a large system is called the **thermodynamic limit**. Most of the questions we would like to answer about condensed matter systems are questions about behaviour in the thermodynamic limit. Unfortunately, in Monte Carlo simulations it is often not feasible to simulate a system large enough that its behaviour is a good approximation to a large system. Much of the effort we put into designing algorithms will be aimed at making them efficient enough that we can simulate the largest systems possible in the available computer time, in

the hope of getting results which are at least a reasonable approximation to the thermodynamic limit.

What about fluctuations in other thermodynamic variables? As we discussed in Section 1.2, each parameter of the system that we fix, such as a volume or an external field, has a conjugate variable, such as a pressure or a magnetization, which is given as a derivative of the free energy by an equation such as (1.14) or (1.15). Derivatives of this general form are produced by terms in the Hamiltonian of the form $-XY$, where Y is a "field" whose value we fix, and X is the conjugate variable to which it couples. For example, the effect of a magnetic field on a magnet can be accounted for by a magnetic energy term in the Hamiltonian of the form $-MB$, where M is the magnetization of the system, and B is the applied magnetic field. We can write the expectation value of X in the form of Equations (1.14) and (1.15) thus:

$$\langle X\rangle = \frac{1}{\beta Z}\sum_\mu X_\mu \mathrm{e}^{-\beta E_\mu} = \frac{1}{\beta Z}\frac{\partial}{\partial Y}\sum_\mu \mathrm{e}^{-\beta E_\mu}, \tag{1.20}$$

since E_μ now contains the term $-X_\mu Y$ which the derivative acts on. Here X_μ is the value of the quantity X in the state μ. We can then write this in terms of the free energy thus:

$$\langle X\rangle = \frac{1}{\beta}\frac{\partial \log Z}{\partial Y} = -\frac{\partial F}{\partial Y}. \tag{1.21}$$

This is a useful technique for calculating the thermal average of a quantity, even if no appropriate field coupling to that quantity appears in the Hamiltonian. We can simply make up a fictitious field which couples to our quantity in the appropriate way—just add a term to the Hamiltonian anyway to allow us to calculate the expectation of the quantity we are interested in—and then set the field to zero after performing the derivative, making the fictitious term vanish from the Hamiltonian again. This is a very common trick in statistical mechanics.

Another derivative of $\log Z$ with respect to Y produces another factor of X_μ in the sum over states, and we find

$$-\frac{1}{\beta}\frac{\partial^2 F}{\partial Y^2} = \frac{1}{\beta}\frac{\partial\langle X\rangle}{\partial Y} = \langle X^2\rangle - \langle X\rangle^2, \tag{1.22}$$

which we recognize as the mean square fluctuation in the variable X. Thus we can find the fluctuations in all sorts of quantities from second derivatives of the free energy with respect to the appropriate fields, just as we can find the energy fluctuations from the second derivative with respect to β. The derivative $\partial\langle X\rangle/\partial Y$, which measures the strength of the response of X to changes in Y is called the **susceptibility** of X to Y, and is usually denoted by χ:

$$\chi \equiv \frac{\partial\langle X\rangle}{\partial Y}. \tag{1.23}$$

Thus the fluctuations in a variable are proportional to the susceptibility of that variable to its conjugate field. This fact is known as the **linear response theorem** and it gives us a way to calculate susceptibilities within Monte Carlo calculations by measuring the size of the fluctuations of a variable.

Extending the idea of the susceptibility, and at the same time moving a step further from the realm of classical thermodynamics, we can also consider what happens when we change the value of a parameter or field at one particular position in our system and ask what effect that has on the conjugate variable at other positions. To study this question we will consider for the moment a system on a lattice. Similar developments are possible for continuous systems like gases, but most of the examples considered in this book are systems which fall on lattices, so it will be of more use to us to go through this for a lattice system here. The interested reader might like to develop the corresponding theory for a continuous system as an exercise.

Let us then suppose that we now have a field which is spatially varying and takes the value Y_i on the i^{th} site of the lattice. The conjugate variables to this field[4] are denoted x_i, and the two are linked via a term in the Hamiltonian $-\sum_i x_i Y_i$. Clearly if we set $Y_i = Y$ and $x_i = X/N$ for all sites i, where N is the total number of sites on the lattice, then this becomes equal once more to the homogeneous situation we considered above. Now in a direct parallel with Equation (1.20) we can write the average value of x_i as

$$\langle x_i \rangle = \frac{1}{Z} \sum_\mu x_i^\mu \mathrm{e}^{-\beta E_\mu} = \frac{1}{\beta} \frac{\partial \log Z}{\partial Y_i}, \tag{1.24}$$

where x_i^μ is the value of x_i in state μ. Then we can define a generalized susceptibility χ_{ij} which is a measure of the response of $\langle x_i \rangle$ to a variation of the field Y_j at a different lattice site:

$$\chi_{ij} = \frac{\partial \langle x_i \rangle}{\partial Y_j} = \frac{1}{\beta} \frac{\partial^2 \log Z}{\partial Y_i \partial Y_j}. \tag{1.25}$$

Again the susceptibility is a second derivative of the free energy. If we make the substitution $Z = \sum_\mu \mathrm{e}^{-\beta E_\mu}$ again (Equation (1.6)), we see that this is also equal to

$$\begin{aligned} \chi_{ij} &= \frac{\beta}{Z} \sum_\mu x_i^\mu x_j^\mu \mathrm{e}^{-\beta E_\mu} - \beta \left[\frac{1}{Z} \sum_\mu x_i^\mu \mathrm{e}^{-\beta E_\mu} \right] \left[\frac{1}{Z} \sum_\nu x_j^\nu \mathrm{e}^{-\beta E_\nu} \right] \\ &= \beta (\langle x_i x_j \rangle - \langle x_i \rangle \langle x_j \rangle) = \beta G_c^{(2)}(i,j). \end{aligned} \tag{1.26}$$

[4] We use lower-case x_i to denote an intensive variable. X by contrast was extensive, i.e., its value scales with the size of the system. We will use this convention to distinguish intensive and extensive variables throughout much of this book.

The quantity $G_c^{(2)}(i,j)$ is called the **two-point connected correlation function** of x between sites i and j, or just the connected correlation, for short. The superscript (2) is to distinguish this function from higher order correlation functions, which are discussed below. As its name suggests, this function is a measure of the correlation between the values of the variable x on the two sites; it takes a positive value if the values of x on those two sites fluctuate in the same direction together, and a negative one if they fluctuate in opposite directions. If their fluctuations are completely unrelated, then its value will be zero. To see why it behaves this way consider first the simpler **disconnected correlation function** $G^{(2)}(i,j)$ which is defined to be

$$G^{(2)}(i,j) \equiv \langle x_i x_j \rangle. \tag{1.27}$$

If the variables x_i and x_j are fluctuating roughly together, around zero, both becoming positive at once and then both becoming negative, at least most of the time, then all or most of the values of the product $x_i x_j$ that we average will be positive, and this function will take a positive value. Conversely, if they fluctuate in opposite directions, then it will take a negative value. If they sometimes fluctuate in the same direction as one another and sometimes in the opposite direction, then the values of $x_i x_j$ will take a mixture of positive and negative values, and the correlation function will average out close to zero. This function therefore has pretty much the properties we desire of a correlation function, and it can tell us a lot of useful things about the behaviour of our system. However, it is not perfect, because we must also consider what happens if we apply our field Y to the system. This can have the effect that the mean value of x at a site $\langle x_i \rangle$ can be non-zero. The same thing can happen even in the absence of an external field if our system undergoes a phase transition to a **spontaneously symmetry broken state** where a variable such as x spontaneously develops a non-zero expectation value. (The Ising model of Section 1.2.2, for instance, does this.) In cases like these, the disconnected correlation function above can have a large positive value simply because the values of the variables x_i and x_j are always either both positive or both negative, even though this has nothing to do with them being correlated to one another. The *fluctuations* of x_i and x_j can be completely unrelated and still the disconnected correlation function takes a non-zero value. To obviate this problem we define the connected correlation function as above:

$$\begin{aligned} G_c^{(2)}(i,j) &\equiv \langle x_i x_j \rangle - \langle x_i \rangle \langle x_j \rangle \\ &= \big\langle (x_i - \langle x_i \rangle) \times (x_j - \langle x_j \rangle) \big\rangle. \end{aligned} \tag{1.28}$$

When the expectations $\langle x_i \rangle$ and $\langle x_j \rangle$ are zero and x_i and x_j are just fluctuating around zero, this function is exactly equal to the disconnected correlation function. But when the expectations are non-zero, the connected correlation

function correctly averages only the fluctuations about those expectations—the term we subtract exactly takes care of any trivial contribution arising because of external fields or spontaneous symmetry breaking. If such trivial contributions are the only reason why $G^{(2)}$ is non-zero then $G_c^{(2)}$ will be zero, which is what we would like. If it is not zero, then we have a genuine correlation between the fluctuations of x_i and x_j.

Although they are not often used in the sorts of systems we will be studying in this book and we will not have call to calculate their values in any of the calculations we will describe here, it is worth mentioning, in case you ever need to use them, that there are also higher-order connected correlation functions, defined by generalizing Equation (1.25) like this:

$$G_c^{(3)}(i,j,k) = \frac{1}{\beta^3}\frac{\partial^3 \log Z}{\partial Y_i \partial Y_j \partial Y_k},$$
$$G_c^{(4)}(i,j,k,l) = \frac{1}{\beta^4}\frac{\partial^4 \log Z}{\partial Y_i \partial Y_j \partial Y_k \partial Y_l}, \tag{1.29}$$

and so on. These are measures of the correlation between simultaneous fluctuations on three and four sites respectively. For a more detailed discussion of these correlation functions and other related ones, see for example Binney *et al.* (1992).

1.2.2 An example: the Ising model

To try to make all of this a bit more concrete, we now introduce a particular model which we can try these concepts out on. That model is the Ising model, which is certainly the most thoroughly researched model in the whole of statistical physics. Without doubt more person-hours have been spent investigating the properties of this model than any other, and although an exact solution of its properties in three dimensions still eludes us, despite many valiant and increasingly sophisticated attempts, a great deal about it is known from computer simulations, and also from approximate methods such as series expansions and ϵ-expansions. We will spend three whole chapters of this book (Chapters 3, 4 and 10) discussing Monte Carlo techniques for studying the model's equilibrium and non-equilibrium properties. Here we will just introduce it briefly and avoid getting too deeply into the discussion of its properties.

The Ising model is a model of a magnet. The essential premise behind it, and behind many magnetic models, is that the magnetism of a bulk material is made up of the combined magnetic dipole moments of many atomic spins within the material. The model postulates a lattice (which can be of any geometry we choose—the simple cubic lattice in three dimensions is a common choice) with a magnetic dipole or spin on each site. In the Ising model

these spins assume the simplest form possible, which is not particularly realistic, of scalar variables s_i which can take only two values ± 1, representing up-pointing or down-pointing dipoles of unit magnitude. In a real magnetic material the spins interact, for example through exchange interactions or RKKY interactions (see, for instance, Ashcroft and Mermin 1976), and the Ising model mimics this by including terms in the Hamiltonian proportional to products $s_i s_j$ of the spins. In the simplest case, the interactions are all of the same strength, denoted by J which has the dimensions of an energy, and are only between spins on sites which are nearest neighbours on the lattice. We can also introduce an external magnetic field B coupling to the spins. The Hamiltonian then takes the form

$$H = -J \sum_{\langle ij \rangle} s_i s_j - B \sum_i s_i, \tag{1.30}$$

where the notation $\langle ij \rangle$ indicates that the sites i and j appearing in the sum are nearest neighbours.[5] The minus signs here are conventional. They merely dictate the choice of sign for the interaction parameter J and the external field B. With the signs as they are here, a positive value of J makes the spins want to line up with one another—a ferromagnetic model as opposed to an anti-ferromagnetic one which is what we get if J is negative—and the spins also want to line up in the same direction as the external field—they want to be positive if $B > 0$ and negative if $B < 0$.

The states of the Ising system are the different sets of values that the spins can take. Since each spin can take two values, there are a total of 2^N states for a lattice with N spins on it. The partition function of the model is the sum

$$Z = \sum_{s_1=\pm 1} \sum_{s_2=\pm 1} \cdots \sum_{s_N=\pm 1} \exp\Bigl[\beta J \sum_{\langle ij \rangle} s_i s_j + \beta B \sum_i s_i\Bigr]. \tag{1.31}$$

To save the eyes, we'll write this in the shorter notation

$$Z = \sum_{\{s_i\}} \mathrm{e}^{-\beta H}. \tag{1.32}$$

If we can perform this sum, either analytically or using a computer, then we can apply all the results of the previous sections to find the internal energy, the entropy, the free energy, the specific heat, and so forth. We can also calculate the mean magnetization $\langle M \rangle$ of the model from the partition

[5] This notation is confusingly similar to the notation for a thermal average, but unfortunately both are sufficiently standard that we feel compelled to use them here. In context it is almost always possible to tell them apart because one involves site labels and the other involves physical variables appearing in the model.

function using Equation (1.15), although as we will see it is usually simpler to evaluate $\langle M \rangle$ directly from an average over states:

$$\langle M \rangle = \Big\langle \sum_i s_i \Big\rangle. \tag{1.33}$$

Often, in fact, we are more interested in the mean magnetization per spin $\langle m \rangle$, which is just

$$\langle m \rangle = \frac{1}{N} \Big\langle \sum_i s_i \Big\rangle. \tag{1.34}$$

(In the later chapters of this book, we frequently use the letter m alone to denote the average magnetization per spin, and omit the brackets $\langle \ldots \rangle$ around it indicating the average. This is also the common practice of many other authors. In almost all cases it is clear from the context when an average over states is to be understood.)

We can calculate fluctuations in the magnetization or the internal energy by calculating derivatives of the partition function. Or, as we mentioned in Section 1.2.1, if we have some way of calculating the size of the fluctuations in the magnetization, we can use those to evaluate the **magnetic susceptibility**

$$\frac{\partial \langle M \rangle}{\partial B} = \beta(\langle M^2 \rangle - \langle M \rangle^2). \tag{1.35}$$

(See Equation (1.22).) Again, it is actually more common to calculate the magnetic susceptibility per spin:

$$\chi = \frac{\beta}{N}(\langle M^2 \rangle - \langle M \rangle^2) = \beta N(\langle m^2 \rangle - \langle m \rangle^2). \tag{1.36}$$

(Note the leading factor of N here, which is easily overlooked when calculating χ from Monte Carlo data.) Similarly we can calculate the specific heat per spin c from the energy fluctuations thus:

$$c = \frac{k\beta^2}{N}(\langle E^2 \rangle - \langle E \rangle^2). \tag{1.37}$$

(See Equation (1.19).)

We can also introduce a spatially varying magnetic field into the Hamiltonian thus:

$$H = -J \sum_{\langle ij \rangle} s_i s_j - \sum_i B_i s_i. \tag{1.38}$$

This gives us a different mean magnetization on each site:

$$\langle m_i \rangle = \langle s_i \rangle = \frac{1}{\beta} \frac{\partial \log Z}{\partial B_i}, \tag{1.39}$$

and allows us to calculate the connected correlation function

$$G_c^{(2)}(i,j) = \frac{1}{\beta^2}\frac{\partial^2 \log Z}{\partial B_i \partial B_j}. \tag{1.40}$$

When we look at the equilibrium simulation of the Ising model in Chapters 3 and 4, all of these will be quantities of interest, and relations like these between them give us useful ways of extracting good results from our numerical data.

1.3 Numerical methods

While the formal developments of statistical mechanics are in many ways very elegant, the actual process of calculating the properties of a particular model is almost always messy and taxing. If we consider calculating the partition function Z, from which, as we have shown, a large number of interesting properties of a system can be deduced, we see that we are going to have to perform a sum over a potentially very large number of states. Indeed, if we are interested in the thermodynamic limit, the sum is over an infinite number of states, and performing such sums is a notoriously difficult exercise. It has been accomplished exactly for a number of simple models with discrete energy states, most famously the Ising model in two dimensions (Onsager 1944). This and other exact solutions are discussed at some length by Baxter (1982). However, for the majority of models of interest today, it has not yet proved possible to find an exact analytic expression for the partition function, or for any other equivalent thermodynamic quantity. In the absence of such exact solutions a number of approximate techniques have been developed including series expansions, field theoretical methods and computational methods. The focus of this book is on the last of these, the computational methods.

The most straightforward computational method for solving problems in statistical physics is to take the model we are interested in and put it on a lattice of finite size, so that the partition function becomes a sum with a finite number of terms. (Or in the case of a model with a continuous energy spectrum it becomes an integral of finite dimension.) Then we can employ our computer to evaluate that sum (or integral) numerically, by simply evaluating each term in turn and adding them up. Let's see what happens when we apply this technique to the Ising model of Section 1.2.2.

If we were really interested in tackling an unsolved problem, we might look at the Ising model in three dimensions, whose exact properties have not yet been found by any method. However, rather than jump in at the deep end, let's first look at the two-dimensional case. For a system of a given linear dimension, this model will have fewer energy states than the three-dimensional one, making the sum over states simpler and quicker to perform,

and the model has the added pedagogical advantage that its behaviour has been solved exactly, so we can compare our numerical calculations with the exact solution. Let's take a smallish system to start with, of 25 spins on a square lattice in a 5×5 arrangement. By convention we apply periodic boundary conditions, so that there are interactions between spins on the border of the array and the opposing spins on the other side. We will also set the external magnetic field B to zero, to make things simpler still.

With each spin taking two possible states, represented by ± 1, our 25 spin system has a total of $2^{25} = 33\,554\,432$ possible states. However, we can save ourselves from summing over half of these, because the system has up/down symmetry, which means that for every state there is another one in which every spin is simply flipped upside down, which has exactly the same energy in zero magnetic field. So we can simplify the calculation of the partition function by just taking one out of every pair of such states, for a total of 16 777 216 states, and summing up the corresponding terms in the partition function, Equation (1.6), and then doubling the sum.[6]

In Figure 1.1 we show the mean magnetization per spin and the specific heat per spin for this 5×5 system, calculated from Equations (1.10) and (1.34). On the same axes we show the exact solutions for these quantities on an infinite lattice, as calculated by Onsager. The differences between the two are clear, and this is precisely the difference between our small finite-sized system and the infinite thermodynamic-limit system which we discussed in Section 1.2.1. Notice in particular that the exact solution has a non-analytic point at about $kT = 2.3J$ which is not reproduced even moderately accurately by our small numerical calculation. This point is the so-called "critical temperature" at which the length-scale ξ of the fluctuations in the magnetization, also called the "correlation length", diverges. (This point is discussed in more detail in Section 3.7.1.) Because of this divergence of the length-scale, it is never possible to get good results for the behaviour of the system at the critical temperature out of any calculation performed on a finite lattice—the lattice is never large enough to include all of the important physics of the critical point. Does this mean that calculations on finite lattices are useless? No, it certainly does not. To start with, at temperatures well away from the critical point the problems are much less severe, and the numerical calculation and the exact solution agree better,

[6] If we were really serious about this, we could save ourselves further time by making use of other symmetries too. For example the square system we are investigating here also has a reflection symmetry and a four-fold rotational symmetry (the symmetry group is C_4), meaning that the states actually group into sets of 16 states (including the up–down symmetry pairs), all of which have the same energy. This would reduce the number of terms we have to evaluate to 2 105 872. (The reader may like to ponder why this number is not exactly $2^{25}/16$, as one might expect.) However, such efforts are not really worthwhile, since, as we will see very shortly, this direct evaluation of the partition function is not a promising method for solving models.

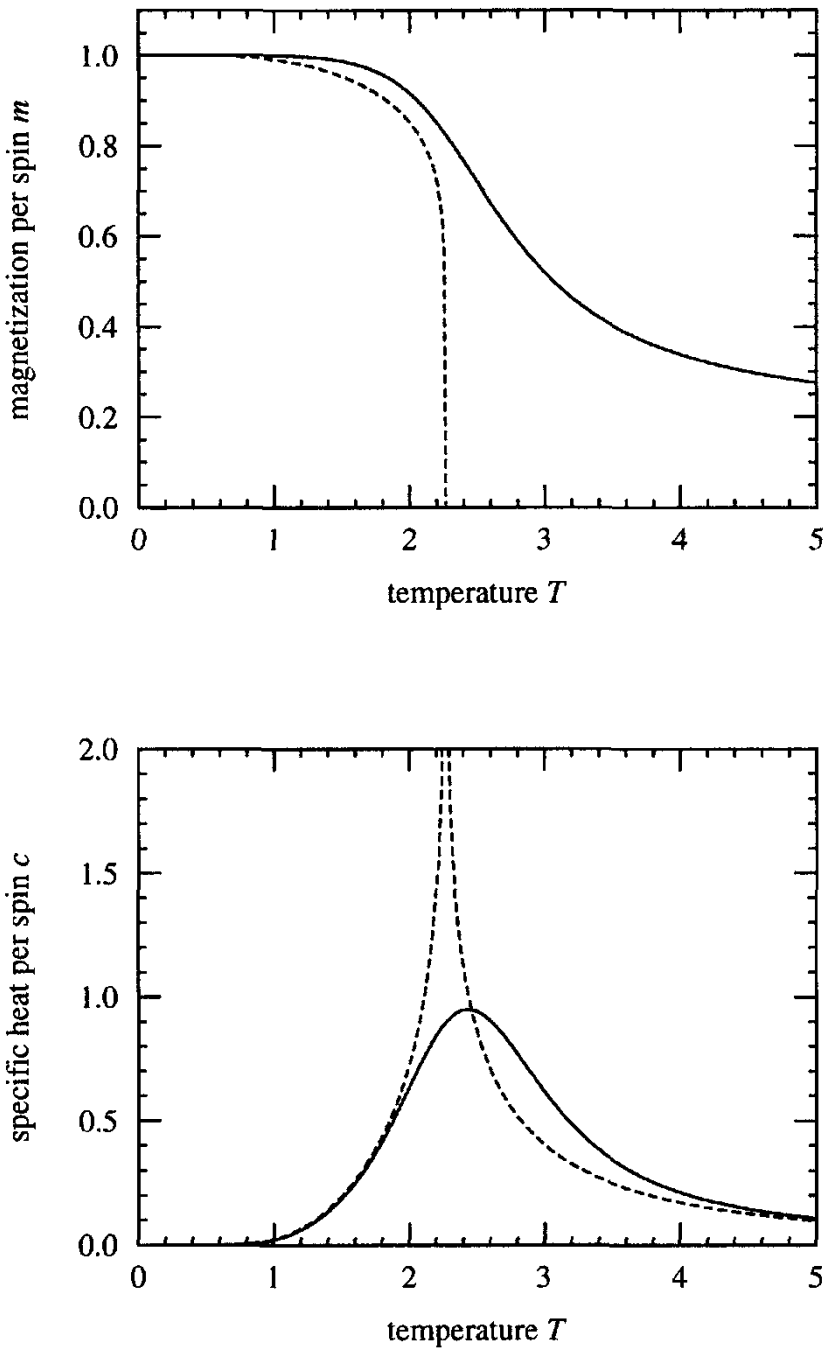

FIGURE 1.1 Top: the mean magnetization per spin m of a 5×5 Ising model on a square lattice in two dimensions (solid line) and the same quantity on an infinitely big square lattice (dashed line). Bottom: the specific heat per spin c for the same two cases.

as we can see in the figure. If we are interested in physics in this regime, then a calculation on a small lattice may well suffice. Second, the technique of "finite size scaling", which is discussed in Section 8.3, allows us to extrapolate results for finite lattices to the limit of infinite system size, and extract good results for the behaviour in the thermodynamic limit. Another technique, that of "Monte Carlo renormalization", discussed in Section 8.4, provides us with a cunning indirect way of calculating some of the features of the critical regime from just the short length-scale phenomena that we get out of a calculation on a small lattice, even though the direct cause of the features that we are interested in is the large length-scale fluctuations that we mentioned.

However, although these techniques can give answers for the critical properties of the system, the accuracy of the answers they give still depends on the size of the system we perform the calculation on, with the answers improving steadily as the system size grows. Therefore it is in our interest to study the largest system we can. However, the calculation which appears as the solid lines in Figure 1.1 took eight hours on a moderately powerful computer. The bulk of this time is spent running through the terms in the sum (1.6). For a system of N spins there are 2^N terms, of which, as we mentioned, we only need actually calculate a half, or 2^{N-1}. This number increases exponentially with the size of the lattice, so we can expect the time taken by the program to increase very rapidly with lattice size. The next size of square lattice up from the present one would be 6×6 or $N = 36$, which should take about $2^{36-1}/2^{25-1} = 2048$ times as long as the previous calculation, or about two years. Clearly this is an unacceptably long time to wait for the answer to this problem. If we are interested in results for any system larger than 5×5, we are going to have to find other ways of getting them.

1.3.1 Monte Carlo simulation

There is essentially only one known numerical method for calculating the partition function of a model such as the Ising model on a large lattice, and that method is Monte Carlo simulation, which is the subject of this book. The basic idea behind Monte Carlo simulation is to simulate the random thermal fluctuation of the system from state to state over the course of an experiment. In Section 1.1 we pointed out that for our purposes it is most convenient to regard the calculation of an expectation value as a time average over the states that a system passes through. In a Monte Carlo calculation we directly simulate this process, creating a model system on our computer and making it pass through a variety of states in such a way that the probability of it being in any particular state μ at a given time t is equal to the weight $w_\mu(t)$ which that state would have in a real system.

In order to achieve this we have to choose a dynamics for our simulation—a rule for changing from one state to another during the simulation—which results in each state appearing with exactly the probability appropriate to it. In the next chapter we will discuss at length a number of strategies for doing this, but the essential idea is that we try to simulate the physical processes that give rise to the master equation, Equation (1.1). We choose a set of rates $R(\mu \to \nu)$ for transitions from one state to another, and we choose them in such a way that the equilibrium solution to the corresponding master equation is precisely the Boltzmann distribution (1.5). Then we use these rates to choose the states which our simulated system passes through during the course of a simulation, and from these states we make estimates of whatever observable quantities we are interested in.

The advantage of this technique is that we need only sample quite a small fraction of the states of the system in order to get accurate estimates of physical quantities. For example, we do not need to include every state of the system in order to get a decent value for the partition function, as we would if we were to evaluate it directly from Equation (1.6). The principal disadvantage of the technique is that there are statistical errors in the calculation due to this same fact that we don't include every state in our calculation, but only some small fraction of the states. In particular this means that there will be statistical noise in the partition function. Taking the derivative of a noisy function is always problematic, so that calculating expectation values from derivatives of the partition function as discussed in Section 1.2 is usually not a good way to proceed. Instead it is normally better in Monte Carlo simulations to calculate as many expectations as we can directly, using equations such as (1.34). We can also make use of relations such as (1.36) to calculate quantities like susceptibilities without having to evaluate a derivative.

In the next chapter we will consider the theory of Monte Carlo simulation in equilibrium thermal systems, and the rest of the first part of the book will deal with the design of algorithms to investigate these systems. In the second part of the book we look at algorithms for non-equilibrium systems.

1.4 A brief history of the Monte Carlo method

In this section we outline the important historical developments in the evolution of the Monte Carlo method. This section is just for fun; feel free to skip over it to the next chapter if you're not interested.

The idea of Monte Carlo calculation is a lot older than the computer. The name "Monte Carlo" is relatively recent—it was coined by Nicolas Metropolis in 1949—but under the older name of "statistical sampling" the method has a history stretching back well into the last century, when numerical calculations were performed by hand using pencil and paper and perhaps

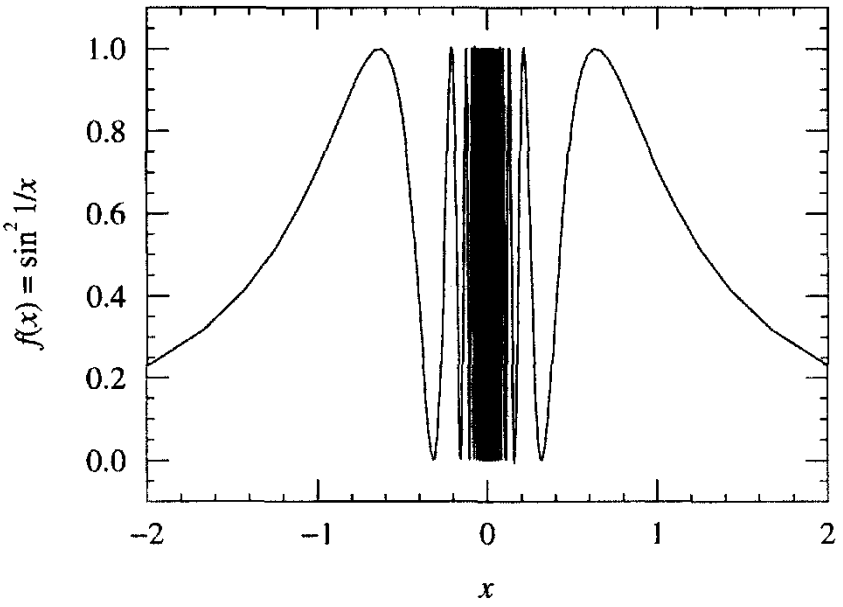

FIGURE 1.2 The pathological function $f(x) \equiv \sin^2 \frac{1}{x}$, whose integral with respect to x, though hard to evaluate analytically, can be evaluated in a straightforward manner using the Monte Carlo integration technique described in the text.

a slide-rule. As first envisaged, Monte Carlo was not a method for solving problems in physics, but a method for estimating integrals which could not be performed by other means. Integrals over poorly-behaved functions and integrals in high-dimensional spaces are two areas in which the method has traditionally proved profitable, and indeed it is still an important technique for problems of these types. To give an example, consider the function

$$f(x) \equiv \sin^2 \frac{1}{x} \tag{1.41}$$

which is pictured in Figure 1.2. The values of this function lie entirely between zero and one, but it is increasingly rapidly varying in the neighbourhood of $x = 0$. Clearly the integral

$$I(x) \equiv \int_0^x f(x')\,\mathrm{d}x' \tag{1.42}$$

which is the area under this curve between 0 and x, takes a finite value somewhere in the range $0 < I(x) < x$, but it is not simple to calculate this value exactly because of the pathologies of the function near the origin. However, we can make an estimate of it by the following method. If we choose a random real number h, uniformly distributed between zero and x, and another v between zero and one and plot on Figure 1.2 the point for which these are the horizontal and vertical coordinates, the probability that

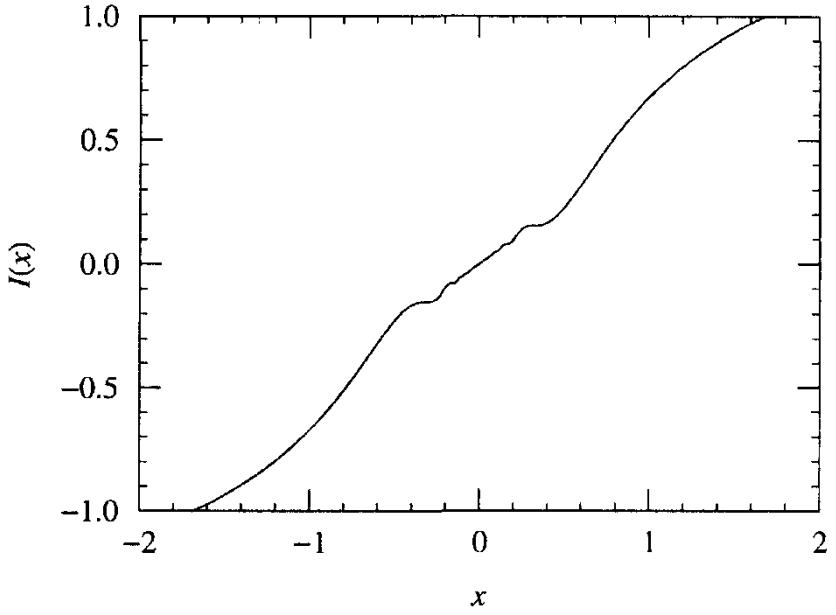

FIGURE 1.3 The function $I(x)$, calculated by Monte Carlo integration as described in the text.

this point will be below the line of $f(x)$ is just $I(x)/x$. It is easy to determine whether the point is in fact below the line: it is below it if $h < f(v)$. Thus if we simply pick a large number N of these random points and count up the number M which fall below the line, we can estimate $I(x)$ from

$$I(x) = \lim_{N\to\infty} \frac{Mx}{N}. \tag{1.43}$$

You can get an answer accurate to one figure by taking a thousand points, which would be about the limit of what one could have reasonably done in the days before computers. Nowadays, even a cheap desktop computer can comfortably run through a million points in a few seconds, giving an answer accurate to about three figures. In Figure 1.3 we have plotted the results of such a calculation for a range of values of x. The errors in this calculation are smaller than the width of the line in the figure.[7]

A famous early example of this type of calculation is the experiment known as "Buffon's needle" (Dörrie 1965), in which the mathematical constant π is determined by repeatedly dropping a needle onto a sheet of paper ruled with evenly spaced lines. The experiment is named after Georges-Louis Leclerc, Comte de Buffon who in 1777 was the first to show that if we throw a needle of length l completely at random onto a sheet of paper ruled with lines a distance d apart, then the chances that the needle will fall so as to

[7]In fact there exist a number of more sophisticated Monte Carlo integration techniques which give more accurate answers than the simple "hit or miss" method we have described here. A discussion can be found in the book by Kalos and Whitlock (1986).

intersect one of the lines is $2l/\pi d$, provided that $d \geq l$. It was Laplace in 1820 who then pointed out that if the needle is thrown down N times and is observed to land on a line M of those times, we can make an estimate of π from

$$\pi = \lim_{N\to\infty} \frac{2Nl}{Md}. \tag{1.44}$$

(Perhaps the connection between this and the Monte Carlo evaluation of integrals is not immediately apparent, but it will certainly become clear if you try to derive Equation (1.44) for yourself, or if you follow Dörrie's derivation.) A number of investigators made use of this method over the years to calculate approximate values for π. The most famous of these is Mario Lazzarini, who in 1901 announced that he had calculated a value of 3.1415929 for π from an experiment in which a $2\frac{1}{2}$ cm needle was dropped 3408 times onto a sheet of paper ruled with lines 3 cm apart. This value, accurate to better than three parts in ten million, would be an impressive example of the power of the statistical sampling method were it not for the fact that it is almost certainly faked. Badger (1994) has demonstrated extremely convincingly that, even supposing Lazzarini had the technology at his disposal to measure the length of his needle and the spaces between his lines to a few parts in 10^7 (a step necessary to ensure the accuracy of Equation (1.44)), still the chances of his finding the results he did were poorer than three in a million; Lazzarini was imprudent enough to publish details of the progress of the experiment through the 3408 castings of the needle, and it turns out that the statistical "fluctuations" in the numbers of intersections of the needle with the ruled lines are much smaller than one would expect in a real experiment. All indications are that Lazzarini forged his results. However, other, less well known attempts at the experiment were certainly genuine, and yielded reasonable figures for π: 3.1596 (Wolf 1850), 3.1553 (Smith 1855). Apparently, performing the Buffon's needle experiment was for a while quite a sophisticated pastime amongst Europe's intellectual gentry.

With the advent of mechanical calculating machines at the end of the nineteenth century, numerical methods took a large step forward. These machines increased enormously the number and reliability of the arithmetic operations that could be performed in a numerical "experiment", and made the application of statistical sampling techniques to research problems in physics a realistic possibility for the first time. An early example of what was effectively a Monte Carlo calculation of the motion and collision of the molecules in a gas was described by William Thomson (later Lord Kelvin) in 1901. Thomson's calculations were aimed at demonstrating the truth of the equipartition theorem for the internal energy of a classical system. However, after the fashion of the time, he did not perform the laborious analysis himself, and a lot of the credit for the results must go to Thomson's

secretary, William Anderson, who apparently solved the kinetic equations for more than five thousand molecular collisions using nothing more than a pencil and a mechanical adding machine.

Aided by mechanical calculators, numerical methods, particularly the method of finite differences, became an important tool during the First World War. The authors recently heard the intriguing story of the Herculean efforts of French mathematician Henri Soudée, who in 1916 calculated firing tables for the new 400 mm cannons being set up at Verdun, directly from his knowledge of the hydrodynamic properties of gases. The tables were used when the cannons were brought to bear on the German-occupied Fort de Douaumont, and as a result the fort was taken by the allies. Soudée was later honoured by the French. By the time of the Second World War the mechanical calculation of firing angles for large guns was an important element of military technology. The physicist Richard Feynman tells the story of his employment in Philadelphia during the summer of 1940 working for the army on a mechanical device for predicting the trajectories of planes as they flew past (Feynman 1985). The device was to be used to guide anti-aircraft guns in attacking the planes. Despite some success with the machine, Feynman left the army's employ after only a few months, joking that the subject of mechanical computation was too difficult for him. He was shrewd enough to realize he was working on a dinosaur, and that the revolution of electronic computing was just around the corner. It was some years however before that particular dream would become reality, and before it did Feynman had plenty more chance to spar with the mechanical calculators. As a group leader during the Manhattan Project at Los Alamos he created what was effectively a highly pipelined human CPU, by employing a large number of people armed with Marchant mechanical adding machines in an arithmetic assembly line in which little cards with numbers on were passed from one worker to the next for processing on the machines. A number of numerical calculations crucial to the design of the atomic bomb were performed in this way.

The first real applications of the statistical sampling method to research problems in physics seem to have been those of Enrico Fermi, who was working on neutron diffusion in Rome in the early 1930s. Fermi never published his numerical methods—apparently he considered only the results to be of interest, not the methods used to obtain them—but according to his influential student and collaborator Emilio Segrè those methods were, in everything but name, precisely the Monte Carlo methods later employed by Ulam and Metropolis and their collaborators in the construction of the hydrogen bomb (Segrè 1980).

So it was that when the Monte Carlo method finally caught the attention of the physics community, it was again as the result of armed conflict. The important developments took place at the Los Alamos National Laboratory

in New Mexico, where Nick Metropolis, Stanislaw Ulam and John von Neumann gathered in the last months of the Second World War shortly after the epochal bomb test at Alamagordo, to collaborate on numerical calculations to be performed on the new ENIAC electronic computer, a mammoth, room-filing machine containing some 18 000 triode valves, whose construction was nearing completion at the University of Pennsylvania. Metropolis (1980) has remarked that the technology that went into the ENIAC existed well before 1941, but that it took the pressure of America's entry into the war to spur the construction of the machine.

It seems to have been Stan Ulam who was responsible for reinventing Fermi's statistical sampling methods. He tells of how the idea of calculating the average effect of a frequently repeated physical process by simply simulating the process over and over again on a digital computer came to him whilst huddled over a pack of cards, playing patience[8] one day. The game he was playing was "Canfield" patience, which is one of those forms of patience where the goal is simply to turn up every card in the pack, and he wondered how often on average one could actually expect to win the game. After abandoning the hopelessly complex combinatorics involved in answering this question analytically, it occurred to him that you could get an approximate answer simply by playing a very large number of games and seeing how often you win. With his mind never far from the exciting new prospect of the ENIAC computer, the thought immediately crossed his mind that he might be able to get the machine to play these games for him far faster than he ever could himself, and it was only a short conceptual leap to applying the same idea to some of the problems of the physics of the hydrogen bomb that were filling his work hours at Los Alamos. He later described his idea to John von Neumann who was very enthusiastic about it, and the two of them began making plans to perform actual calculations. Though Ulam's idea may appear simple and obvious to us today, there are actually many subtle questions involved in this idea that a physical problem with an exact answer can be approximately solved by studying a suitably chosen random process. It is a tribute to the ingenuity of the early Los Alamos workers that, rather than plunging headlong into the computer calculations, they considered most of these subtleties right from the start.

The war ended before the first Monte Carlo calculations were performed on the ENIAC. There was some uncertainty about whether the Los Alamos laboratory would continue to exist in peacetime, and Edward Teller, who was leading the project to develop the hydrogen bomb, was keen to apply the power of the computer to the problems of building the new bomb, in order to show that significant work was still going on at Los Alamos. Von Neumann developed a detailed plan of how the Monte Carlo method could be

[8] Also called "solitaire" in the USA.

implemented on the ENIAC to solve a number of problems concerned with neutron transport in the bomb, and throughout 1947 worked with Metropolis on preparations for the calculations. They had to wait to try their ideas out however, because the ENIAC was to be moved from Philadelphia where it was built to the army's Ballistics Research Laboratory in Maryland. For a modern computer this would not be a problem, but for the gigantic ENIAC, with its thousands of fragile components, it was a difficult task, and there were many who did not believe the computer would survive the journey. It did, however, and by the end of the year it was working once again in its new home. Before von Neumann and the others put it to work on the calculations for the hydrogen bomb, Richard Clippinger of the Ballistics Lab suggested a modification to the machine which allowed it to store programs in its electronic memory. Previously a program had to be set up by plugging and unplugging cables at the front of the machine, an arduous task which made the machine inflexible and inconvenient to use. Von Neumann was in favour of changing to the new "stored program" model, and Nick Metropolis and von Neumann's wife, Klari, made the necessary modifications to the computer themselves. It was the end of 1947 before the machine was at last ready, and Metropolis and von Neumann set to work on the planned Monte Carlo calculations.

The early neutron diffusion calculations were an impressive success, but Metropolis and von Neumann were not able to publish their results, because they were classified as secret. Over the following two years however, they and others, including Stan Ulam and Stanley Frankel, applied the new statistical sampling method to a variety of more mundane problems in physics, such as the calculation of the properties of hard-sphere gases in two and three dimensions, and published a number of papers which drew the world's attention to this emerging technique. The 1949 paper by Metropolis and Ulam on statistical techniques for studying integro-differential equations is of interest because it contained in its title the first use of the term "Monte Carlo" to describe this type of calculation. Also in 1949 the first conference on Monte Carlo methods was held in Los Alamos, attracting more than a hundred participants. It was quickly followed by another similar meeting in Gainesville, Florida.

The calculations received a further boost in 1948 with the arrival at Los Alamos of a new computer, humorously called the MANIAC. (Apparently the name was suggested by Enrico Fermi, who was tiring of computers with contrived acronyms for names—he claimed that it stood for "Metropolis and Neumann Invent Awful Contraption". Nowadays, with all our computers called things like XFK–23/z we would no doubt appreciate a few pronounceable names.) Apart from the advantage of being in New Mexico rather than Maryland, the MANIAC was a significant technical improvement over the ENIAC which Presper Eckert (1980), its principal architect,

refers to as a "hastily built first try". It was faster and contained a larger memory (40 kilobits, or 5 kilobytes in modern terms). It was built under the direction of Metropolis, who had been lured back to Los Alamos after a brief stint on the faculty at Chicago by the prospect of the new machine. The design was based on ideas put forward by John von Neumann and incorporated a number of technical refinements proposed by Jim Richardson, an engineer working on the project. A still more sophisticated computer, the MANIAC 2, was built at Los Alamos two years later, and both machines remained in service until the late fifties, producing a stream of results, many of which have proved to be seminal contributions to the field of Monte Carlo simulation. Of particular note to us is the publication in 1953 of the paper by Nick Metropolis, Marshall and Arianna Rosenbluth, and Edward and Mici Teller, in which they describe for the first time the Monte Carlo technique that has come to be known as the Metropolis algorithm. This algorithm was the first example of a thermal "importance sampling" method, and it is to this day easily the most widely used such method. We will be discussing it in some detail in Chapter 3. Also of interest are the Monte Carlo studies of nuclear cascades performed by Antony Turkevich and Nick Metropolis, and Edward Teller's work on phase changes in interacting hard-sphere gases using the Metropolis algorithm.

The exponential growth in computer power since those early days is by now a familiar story to us all, and with this increase in computational resources Monte Carlo techniques have looked deeper and deeper into the subject of statistical physics. Monte Carlo simulations have also become more accurate as a result of the invention of new algorithms. Particularly in the last twenty years, many new ideas have been put forward, of which we describe a good number in the rest of this book.

Problems

1.1 "If a system is in equilibrium with a thermal reservoir at temperature T, the probability of its having a total energy E varies with E in proportion to $e^{-\beta E}$." True or false?

1.2 A certain simple system has only two energy states, with energies E_0 and E_1, and transitions between the two states take place at rates $R(0 \to 1) = R_0 \exp[-\beta(E_1 - E_0)]$ and $R(1 \to 0) = R_0$. Solve the master equation (1.1) for the probabilities w_0 and w_1 of occupation of the two states as a function of time with the initial conditions $w_0 = 0$, $w_1 = 1$. Show that as $t \to \infty$ these solutions tend to the Boltzmann probabilities, Equation (1.5).

1.3 A slightly more complex system contains N distinguishable particles, each of which can be in one of two boxes. The particles in the first box have energy $E_0 = 0$ and the particles in the second have energy E_1, and particles

are allowed to move back and forward between the boxes under the influence of thermal excitations from a reservoir at temperature T. Find the partition function for this system and then use this result to calculate the internal energy.

1.4 Solve the Ising model, whose Hamiltonian is given in Equation (1.30), in one dimension for the case where $B = 0$ as follows. Define a new set of variables σ_i which take values 0 and 1 according to $\sigma_i = \frac{1}{2}(1 - s_i s_{i+1})$ and rewrite the Hamiltonian in terms of these variables for a system of N spins with periodic boundary conditions. Show that the resulting system is equivalent to the one studied in Problem 1.3 in the limit of large N and hence calculate the internal energy as a function of temperature.

2

The principles of equilibrium thermal Monte Carlo simulation

In Section 1.3.1 we looked briefly at the general ideas behind equilibrium thermal Monte Carlo simulations. In this chapter we discuss these ideas in more detail in preparation for the discussion in the following chapters of a variety of specific algorithms for use with specific problems. The three crucial ideas that we introduce in this chapter are "importance sampling", "detailed balance" and "acceptance ratios". If you know what these phrases mean, you can understand most of the thermal Monte Carlo simulations that have been performed in the last thirty years.

2.1 The estimator

The usual goal in the Monte Carlo simulation of a thermal system is the calculation of the expectation value $\langle Q \rangle$ of some observable quantity Q, such as the internal energy in a model of a gas, or the magnetization in a magnetic model. As we showed in Section 1.3, the ideal route to calculating such an expectation, that of averaging the quantity of interest over all states μ of the system, weighting each with its own Boltzmann probability

$$\langle Q \rangle = \frac{\sum_\mu Q_\mu \, \mathrm{e}^{-\beta E_\mu}}{\sum_\mu \mathrm{e}^{-\beta E_\mu}} \tag{2.1}$$

is only tractable in the very smallest of systems. In larger systems, the best we can do is average over some subset of the states, though this necessarily introduces some inaccuracy into the calculation. Monte Carlo techniques work by choosing a subset of states at random from some probability distribution p_μ which we specify. Suppose we choose M such states $\{\mu_1 \ldots \mu_M\}$.

Our best estimate of the quantity Q will then be given by

$$Q_M = \frac{\sum_{i=1}^{M} Q_{\mu_i} p_{\mu_i}^{-1} \mathrm{e}^{-\beta E_{\mu_i}}}{\sum_{j=1}^{M} p_{\mu_j}^{-1} \mathrm{e}^{-\beta E_{\mu_j}}}. \tag{2.2}$$

Q_M is called the **estimator** of Q. It has the property that, as the number M of states sampled increases, it becomes a more and more accurate estimate of $\langle Q \rangle$, and when $M \to \infty$ we have $Q_M = \langle Q \rangle$.

The question we would like to answer now is how should we choose our M states in order that Q_M be an accurate estimate of $\langle Q \rangle$? In other words, how should we choose the probability distribution p_μ? The simplest choice is to pick all states with equal probability; in other words make all p_μ equal. Substituting this choice into Equation (2.2), we get

$$Q_M = \frac{\sum_{i=1}^{M} Q_{\mu_i} \mathrm{e}^{-\beta E_{\mu_i}}}{\sum_{j=1}^{M} \mathrm{e}^{-\beta E_{\mu_j}}}. \tag{2.3}$$

It turns out however, that this is usually a rather poor choice to make. In most numerical calculations it is only possible to sample a very small fraction of the total number of states. Consider, for example, the Ising model of Section 1.2.2 again. A small three-dimensional cubic system of $10 \times 10 \times 10$ Ising spins would have $2^{1000} \simeq 10^{300}$ states, and a typical numerical calculation could only hope to sample up to about 10^8 of those in a few hours on a good computer, which would mean we were only sampling one in every 10^{292} states of the system, a very small fraction indeed. The estimator given above is normally a poor guide to the value of $\langle Q \rangle$ under these circumstances. The reason is that one or both of the sums appearing in Equation (2.1) may be dominated by a small number of states, with all the other states, the vast majority, contributing a negligible amount even when we add them all together. This effect is often especially obvious at low temperatures, where these sums may be dominated by a hundred states, or ten states, or even one state, because at low temperatures there is not enough thermal energy to lift the system into the higher excited states, and so it spends almost all of its time sitting in the ground state, or one of the lowest of the excited states. In the example described above, the chances of one of the 10^8 random states we sample in our simulation being the ground state are one in 10^{292}, which means there is essentially no chance of our picking it, which makes Q_M a very inaccurate estimate of $\langle Q \rangle$ if the sums are dominated by the contribution from this state.

On the other hand, if we had some way of knowing which states made the important contributions to the sums in Equation (2.1) and if we could pick our sample of M states from just those states and ignore all the others, we could get a very good estimate of $\langle Q \rangle$ with only a small number of terms. This is the essence of the idea behind thermal Monte Carlo methods. The

technique for picking out the important states from amongst the very large number of possibilities is called **importance sampling**.

2.2 Importance sampling

As discussed in Section 1.1, we can regard an expectation value as a time average over the states that a system passes through during the course of a measurement. We do not assume that the system passes through every state during the measurement, even though every state appears in the sums of Equation (2.1). When you count how many states a typical system has you realize that this would never be possible. For instance, consider again the example we took in the last chapter of a litre container of gas at room temperature and atmospheric pressure. Such a system contains on the order of 10^{22} molecules. Typical speeds for these molecules are in the region of 100 $\mathrm{m\,s^{-1}}$, giving them a de Broglie wavelength of around 10^{-10} m. Each molecule will then have about 10^{27} different quantum states within the one litre box, and the complete gas will have around $(10^{27})^{10^{22}}$ states, which is a spectacularly large number.[1] The molecules will change from one state to another when they undergo collisions with one another or with the walls of the container, which they do at a rate of about 10^9 collisions per second, or 10^{31} changes of state per second for the whole gas. At this rate, it will take about $10^{10^{23}}$ times the lifetime of the universe for our litre of gas to move through every possible state. Clearly then, our laboratory systems are only sampling the tiniest portion of their state spaces during the time that we conduct our experiments on them. In effect, real systems are carrying out a sort of Monte Carlo calculation of their own properties; they are "analogue computers" which evaluate expectations by taking a small but representative sample of their own states and averaging over that sample.[2] So it should not come as a great surprise to learn that we can also perform a reasonable calculation of the properties of a system using a simulation which only samples a small fraction of its states.

In fact, our calculations are often significantly better than this simple argument suggests. In Section 1.2.1 we showed that the range of energies of the states sampled by a typical system is very small compared with the total

[1] Actually, this is probably an overestimate, since it counts states which are classically distinguishable but quantum mechanically identical. For the purpose of the present rough estimation however, it will do fine.

[2] There are some systems which, because they have certain conservation laws, will not in fact sample their state spaces representatively, and this can lead to discrepancies between theory and experiment. Special Monte Carlo techniques exist for simulating these "conservative" systems, and we will touch on one or two of them in the coming chapters. For the moment, however, we will make the assumption that our system takes a representative sample of its own states.

energy of the system—the ratio was about 10^{-20} in the case of our litre of gas, for instance. Similar arguments can be used to show that systems sample very narrow ranges of other quantities as well. The reason for this, as we saw, is that the system is not sampling all states with equal probability, but instead sampling them according to the Boltzmann probability distribution, Equation (1.5). If we can mimic this effect in our simulations, we can exploit these narrow ranges of energy and other quantities to make our estimates of such quantities very accurate. For this reason, we normally try to take a sample of the states of the system in which the likelihood of any particular one appearing is proportional to its Boltzmann weight. This is the most common form of importance sampling, and most of the algorithms in this book make use of this idea in one form or another.

Our strategy then is this: instead of picking our M states in such a way that every state of the system is as likely to get chosen as every other, we pick them so that the probability that a particular state μ gets chosen is $p_\mu = Z^{-1}\mathrm{e}^{-\beta E_\mu}$. Then our estimator for $\langle Q \rangle$, Equation (2.2), becomes just

$$Q_M = \frac{1}{M}\sum_{i=1}^{M} Q_{\mu_i}. \tag{2.4}$$

Notice that the Boltzmann factors have now cancelled out of the estimator, top and bottom, leaving a particularly simple expression. This definition of Q_M works much better than (2.3), especially when the system is spending the majority of its time in a small number of states (such as, for example, the lowest-lying ones when we are at low temperatures), since these will be precisely the states that we pick most often, and the relative frequency with which we pick them will exactly correspond to the amount of time the real system would spend in those states.

The only remaining question is *how* exactly we pick our states so that each one appears with its correct Boltzmann probability. This is by no means a simple task. In the remainder of this chapter we describe the standard solution to the problem, which makes use of a "Markov process".

2.2.1 Markov processes

The tricky part of performing a Monte Carlo simulation is the generation of an appropriate random set of states according to the Boltzmann probability distribution. For a start, one cannot simply choose states at random and accept or reject them with a probability proportional to $\mathrm{e}^{-\beta E_\mu}$. That would be no better than our original scheme of sampling states at random; we would end up rejecting virtually all states, since the probabilities for their acceptance would be exponentially small. Instead, almost all Monte Carlo schemes rely on **Markov processes** as the generating engine for the set of states used.

For our purposes, a Markov process is a mechanism which, given a system in one state μ, generates a new state of that system ν. It does so in a random fashion; it will not generate the same new state every time it is given the initial state μ. The probability of generating the state ν given μ is called the **transition probability** $P(\mu \to \nu)$ for the transition from μ to ν, and for a true Markov process all the transition probabilities should satisfy two conditions: (1) they should not vary over time, and (2) they should depend only on the properties of the current states μ and ν, and not on any other states the system has passed through. These conditions mean that the probability of the Markov process generating the state ν on being fed the state μ is the same every time it is fed the state μ, irrespective of anything else that has happened. The transition probabilities $P(\mu \to \nu)$ must also satisfy the constraint

$$\sum_{\nu} P(\mu \to \nu) = 1, \tag{2.5}$$

since the Markov process must generate *some* state ν when handed a system in the state μ. Note however, that the transition probability $P(\mu \to \mu)$, which is the probability that the new state generated will be the same as the old one, need not be zero. This amounts to saying there may be a finite probability that the Markov process will just stay in state μ.

In a Monte Carlo simulation we use a Markov process repeatedly to generate a **Markov chain** of states. Starting with a state μ, we use the process to generate a new one ν, and then we feed *that* state into the process to generate another λ, and so on. The Markov process is chosen specially so that when it is run for long enough starting from any state of the system it will eventually produce a succession of states which appear with probabilities given by the Boltzmann distribution. (We call the process of reaching the Boltzmann distribution "coming to equilibrium", since it is exactly the process that a real system goes through with its "analogue computer" as it reaches equilibrium at the ambient temperature.) In order to achieve this, we place two further conditions on our Markov process, in addition to the ones specified above, the conditions of "ergodicity" and "detailed balance"

2.2.2 Ergodicity

The **condition of ergodicity** is the requirement that it should be possible for our Markov process to reach any state of the system from any other state, if we run it for long enough. This is necessary to achieve our stated goal of generating states with their correct Boltzmann probabilities. Every state ν appears with some non-zero probability p_ν in the Boltzmann distribution, and if that state were inaccessible from another state μ no matter how long we continue our process for, then our goal is thwarted if we start in state μ: the probability of finding ν in our Markov chain of states will be zero, and

not p_ν as we require it to be.

The condition of ergodicity tells us that we are allowed to make some of the transition probabilities of our Markov process zero, but that there must be at least one path of non-zero transition probabilities between any two states that we pick. In practice, most Monte Carlo algorithms set almost all of the transition probabilities to zero, and we must be careful that in so doing we do not create an algorithm which violates ergodicity. For most of the algorithms we describe in this book we will explicitly prove that ergodicity is satisfied before making use of the algorithm.

2.2.3 Detailed balance

The other condition we place on our Markov process is the **condition of detailed balance**. This condition is the one which ensures that it is the Boltzmann probability distribution which we generate after our system has come to equilibrium, rather than any other distribution. Its derivation is quite subtle. Consider first what it means to say that the system is in equilibrium. The crucial defining condition is that the rate at which the system makes transitions into and out of any state μ must be equal. Mathematically we can express this as[3]

$$\sum_\nu p_\mu P(\mu \to \nu) = \sum_\nu p_\nu P(\nu \to \mu). \tag{2.6}$$

Making use of the sum rule, Equation (2.5), we can simplify this to

$$p_\mu = \sum_\nu p_\nu P(\nu \to \mu). \tag{2.7}$$

For any set of transition probabilities satisfying this equation, the probability distribution p_μ will be an equilibrium of the dynamics of the Markov process. Unfortunately, however, simply satisfying this equation is not sufficient to guarantee that the probability distribution will tend to p_μ from any state of the system if we run the process for long enough. We can demonstrate this as follows.

The transition probabilities $P(\mu \to \nu)$ can be thought of as the elements of a matrix $\mathbf{P}$. This matrix is called the **Markov matrix** or the **stochastic matrix** for the Markov process. Let us return to the notation of Section 1.1, in which we denoted by $w_\mu(t)$, the probability that our system is in a state μ at time t. If we measure time in steps along our Markov chain, then the

[3]This equation is essentially just a discrete-time version of the one we would get if we were to set the derivative in the master equation, Equation (1.1), to zero.

probability $w_\nu(t+1)$ of being in state ν at time $t+1$ is given by[4]

$$w_\nu(t+1) = \sum_\mu P(\mu \to \nu)\, w_\mu(t). \tag{2.8}$$

In matrix notation, this becomes

$$\mathbf{w}(t+1) = \mathbf{P} \cdot \mathbf{w}(t), \tag{2.9}$$

where $\mathbf{w}(t)$ is the vector whose elements are the weights $w_\mu(t)$. If the Markov process reaches a simple equilibrium state $\mathbf{w}(\infty)$ as $t \to \infty$, then that state satisfies

$$\mathbf{w}(\infty) = \mathbf{P} \cdot \mathbf{w}(\infty). \tag{2.10}$$

However, it is also possible for the process to reach a **dynamic equilibrium** in which the probability distribution $\mathbf{w}$ rotates around a number of different values. Such a rotation is called a **limit cycle**. In this case $\mathbf{w}(\infty)$ would satisfy

$$\mathbf{w}(\infty) = \mathbf{P}^n \cdot \mathbf{w}(\infty), \tag{2.11}$$

where n is the length of the limit cycle. If we choose our transition probabilities (or equivalently our Markov matrix) to satisfy Equation (2.7) we guarantee that the Markov chain will have a simple equilibrium probability distribution p_μ, but it may also have any number of limit cycles of the form (2.11). This means that there is no guarantee that the actual states generated will have anything like the desired probability distribution.

We get around this problem by applying an additional condition to our transition probabilities thus:

$$p_\mu P(\mu \to \nu) = p_\nu P(\nu \to \mu). \tag{2.12}$$

This is the condition of detailed balance. It is clear that any set of transition probabilities which satisfy this condition also satisfy Equation (2.6). (To prove it, simply sum both sides of Equation (2.12) over ν.) We can also show that this condition eliminates limit cycles. To see this, look first at the left-hand side of the equation, which is the probability of being in a state μ multiplied by the probability of making a transition from that state to another state ν. In other words, it is the overall rate at which transitions from μ to ν happen in our system. The right-hand side is the overall rate for the reverse transition. The condition of detailed balance tells us that on average the system should go from μ to ν just as often as it goes from ν to μ. In a limit cycle, in which the probability of occupation of some or all of the states changes in a cyclic fashion, there must be states for which this

[4]This equation is also closely related to Equation (1.1). The reader may like to work out how the one can be transformed into the other.

condition is violated on any particular step of the Markov chain; in order for the probability of occupation of a particular state to increase, for instance, there must be more transitions into that state than out of it, on average. The condition of detailed balance forbids dynamics of this kind and hence forbids limit cycles.

Once we remove the limit cycles in this way, it is straightforward to show that the system will always tend to the probability distribution p_μ as $t \to \infty$. As $t \to \infty$, $\mathbf{w}(t)$ will tend exponentially towards the eigenvector corresponding to the largest eigenvalue of $\mathbf{P}$. This may be obvious to you if you are familiar with stochastic matrices. If not, we prove it in Section 3.3.2. For the moment, let us take it as given. Looking at Equation (2.10) we see that the largest eigenvalue of the Markov matrix must in fact be one.[5] If limit cycles of the form (2.11) were present, then we could also have eigenvalues which are complex roots of one, but the condition of detailed balance prevents this from happening. Now look back at Equation (2.7) again. We can express this equation in matrix notation as

$$\mathbf{p} = \mathbf{P} \cdot \mathbf{p}. \tag{2.13}$$

In other words, if Equation (2.7) (or equivalently the condition of detailed balance) holds for our Markov process, then the vector $\mathbf{p}$ whose elements are the probabilities p_μ is precisely the one correctly normalized eigenvector of the Markov matrix which has eigenvalue one. Putting this together with Equation (2.10) we see that the equilibrium probability distribution over states $\mathbf{w}(\infty)$ is none other than $\mathbf{p}$, and hence $\mathbf{w}(t)$ must tend exponentially to $\mathbf{p}$ as $t \to \infty$.

There is another reason why detailed balance makes sense for Monte Carlo simulations: the "analogue computers" which constitute the real physical systems we are trying to mimic almost always obey the condition of detailed balance. The reason is that they are based on standard quantum or classical mechanics, which is time-reversal symmetric. If they did not obey detailed balance, then in equilibrium they could have one or more limit cycles around which the system passes in one particular direction. If we take such a system and reverse it in time, the motion around this cycle is also reversed, and it becomes clear that the dynamics of the system in equilibrium is not the same forward as it is in reverse. Such a violation of time-reversal symmetry is forbidden for most systems, implying that they must satisfy detailed balance. Although this does not mean that we are necessarily obliged

[5] All Markov matrices have at least one eigenvector with corresponding eigenvalue one, a fact which is easily proven since Equation (2.5) implies that the vector $(1, 1, 1, \ldots)$ is a left eigenvector of $\mathbf{P}$ with eigenvalue one. It is possible to have more than one eigenvector with eigenvalue one if the states of the system divide into two or more mutually inaccessible subsets. However, if the condition of ergodicity is satisfied then such subsets are forbidden and hence there is only one such eigenvector.

to enforce detailed balance in our simulations as well, it is helpful if we do, because it makes the behaviour of our model system more similar to that of the real one we are trying to understand.

So, we have shown that we can arrange for the probability distribution of states generated by our Markov process to tend to any distribution p_μ we please by choosing a set of transition probabilities which satisfy Equation (2.12). Given that we wish the equilibrium distribution to be the Boltzmann distribution,[6] clearly we want to choose the values of p_μ to be the Boltzmann probabilities, Equation (1.5). The detailed balance equation then tells us that the transition probabilities should satisfy

$$\frac{P(\mu \to \nu)}{P(\nu \to \mu)} = \frac{p_\nu}{p_\mu} = \mathrm{e}^{-\beta(E_\nu - E_\mu)}. \tag{2.14}$$

This equation and Equation (2.5) are the constraints on our choice of transition probabilities $P(\mu \to \nu)$. If we satisfy these, as well as the condition of ergodicity, then the equilibrium distribution of states in our Markov process will be the Boltzmann distribution. Given a suitable set of transition probabilities, our plan is then to write a computer program which implements the Markov process corresponding to these transition probabilities so as to generate a chain of states. After waiting a suitable length of time[7] to allow the probability distribution of states $w_\mu(t)$ to get sufficiently close to the Boltzmann distribution, we average the observable Q that we are interested in over M states and we have calculated the estimator Q_M defined in Equation (2.4). A number of refinements on this outline are possible and we will discuss some of those in the remainder of this chapter and in later chapters of the book, but this is the basic principle on which virtually all modern equilibrium Monte Carlo calculations are based.

Our constraints still leave us a good deal of freedom over how we choose the transition probabilities. There are many ways in which to satisfy them. One simple choice for example is

$$P(\mu \to \nu) \propto \mathrm{e}^{-\frac{1}{2}\beta(E_\mu - E_\nu)}, \tag{2.15}$$

although as we will show in Section 3.1 this choice is not a very good one. There are some other choices which are known to work well in many cases, such as the "Metropolis algorithm" proposed by Metropolis and co-workers in 1953, and we will discuss the most important of these in the coming chapters. However, it must be stressed—and this is one of the most important

[6] Occasionally, in fact, we want to generate equilibrium distributions other than the Boltzmann distribution. An example is the entropic sampling algorithm of Section 6.3. In this case the arguments here still apply. We simply feed our required distribution into the condition of detailed balance.

[7] Exactly how long we have to wait can be a difficult thing to decide. A number of possible criteria are discussed in Section 3.2.

things this book has to say—that the standard algorithms are *very rarely* the best ones for solving new problems with. In most cases they will work, and in some cases they will even give quite good answers, but you can almost always do a better job by giving a little extra thought to choosing the best set of transition probabilities to construct an algorithm that will answer the particular questions that you are interested in. A purpose-built algorithm can often give a much faster simulation than an equivalent standard algorithm, and the improvement in efficiency can easily make the difference between finding an answer to a problem and not finding one.

2.3 Acceptance ratios

Our little summary above makes rather light work of the problems of constructing a Monte Carlo algorithm. Given a desired set of transition probabilities $P(\mu \to \nu)$ satisfying the conditions (2.5) and (2.14), we say, we simply concoct some Markov process that generates states with exactly those transition probabilities, and *presto!* we produce a string of states of our system with exactly their correct Boltzmann probabilities. However, it is often very far from obvious what the appropriate Markov process is that has the required transition probabilities, and finding one can be a haphazard, trial-and-error process. For some problems we can use known algorithms such as the Metropolis method (see Section 3.1), but for many problems the standard methods are far from ideal, and we will do much better if we can tailor a new algorithm to our specific needs. But though we may be able to suggest many candidate Markov processes—different ways of creating a new state ν from an old one μ—still we may not find one which gives exactly the right set of transition probabilities. The good news however is that we don't have to. In fact it turns out that we can choose any algorithm we like for generating the new states, and still have our desired set of transition probabilities, by introducing something called an **acceptance ratio**. The idea behind the trick is this.

We mentioned in Section 2.2.1 that we are allowed to make the "stay-at-home" transition probability $P(\mu \to \mu)$ non-zero if we want. If we set $\nu = \mu$ in Equation (2.14), we get the simple tautology $1 = 1$, which means that the condition of detailed balance is always satisfied for $P(\mu \to \mu)$, no matter what value we choose for it. This gives us some flexibility about how we choose the other transition probabilities with $\mu \neq \nu$. For a start, it means that we can adjust the value of any $P(\mu \to \nu)$ and keep the sum rule (2.5) satisfied, by simply compensating for that adjustment with an equal but opposite adjustment of $P(\mu \to \mu)$. The only thing we need to watch is that $P(\mu \to \mu)$ never passes out of its allowed range between zero and one. If we make an adjustment like this in $P(\mu \to \nu)$, we can also arrange for Equation (2.14) to remain satisfied, by simultaneously making a change in

$P(\nu \to \mu)$, so that the ratio of the two is preserved.

It turns out that these considerations actually give us enough freedom that we can make the transition probabilities take any set of values we like by tweaking the values of the probabilities $P(\mu \to \mu)$. To see this, we break the transition probability down into two parts:

$$P(\mu \to \nu) = g(\mu \to \nu)\, A(\mu \to \nu). \tag{2.16}$$

The quantity $g(\mu \to \nu)$ is the **selection probability**, which is the probability, given an initial state μ, that our algorithm will generate a new target state ν, and $A(\mu \to \nu)$ is the acceptance ratio (sometimes also called the "acceptance probability"). The acceptance ratio says that if we start off in a state μ and our algorithm generates a new state ν from it, we should accept that state and change our system to the new state ν a fraction of the time $A(\mu \to \nu)$. The rest of the time we should just stay in the state μ. We are free to choose the acceptance ratio to be any number we like between zero and one; choosing it to be zero for all transitions is equivalent to choosing $P(\mu \to \mu) = 1$, which is the largest value it can take, and means that we will never leave the state μ. (Not a very desirable situation. We would never choose an acceptance ratio of zero for an actual calculation.)

This gives us complete freedom about how we choose the selection probabilities $g(\mu \to \nu)$, since the constraint (2.14) only fixes the ratio

$$\frac{P(\mu \to \nu)}{P(\nu \to \mu)} = \frac{g(\mu \to \nu)A(\mu \to \nu)}{g(\nu \to \mu)A(\nu \to \mu)}. \tag{2.17}$$

The ratio $A(\mu \to \nu)/A(\nu \to \mu)$ can take any value we choose between zero and infinity, which means that both $g(\mu \to \nu)$ and $g(\nu \to \mu)$ can take any values we like.

Our other constraint, the sum rule of Equation (2.5), is still satisfied, since the system must end up in *some* state after each step in the Markov chain, even if that state is just the state we started in.

So, in order to create our Monte Carlo algorithm what we actually do is think up an algorithm which generates random new states ν given old ones μ, with some set of probabilities $g(\mu \to \nu)$, and then we accept or reject those states with acceptance ratios $A(\mu \to \nu)$ which we choose to satisfy Equation (2.17). This will then satisfy all the requirements for the transition probabilities, and so produce a string of states which, when the algorithm reaches equilibrium, will each appear with their correct Boltzmann probability.

This all seems delightful, but there is a catch which we must always bear in mind, and which is one of the most important considerations in the design of Monte Carlo algorithms. If the acceptance ratios for our moves are low, then the algorithm will on most time steps simply stay in the state

that it is in, and not go anywhere. The step on which it actually accepts a change to a new state will be rare, and this is wasteful of time. We want an algorithm that moves nimbly about state space and samples a wide selection of different states. We don't want to take a million time steps and find that our algorithm has only sampled a dozen states. The solution to this problem is to make the acceptance ratio as close to unity as possible. One way to do this is to note that Equation (2.17) fixes only the ratio $A(\mu \to \nu)/A(\nu \to \mu)$ of the acceptance ratios for the transitions in either direction between any two states. Thus we are free to multiply both $A(\mu \to \nu)$ and $A(\nu \to \mu)$ by the same factor, and the equation will still be obeyed. The only constraint is that both acceptance ratios should remain between zero and one. In practice then, what we do is to set the larger of the two acceptance ratios to one, and have the other one take whatever value is necessary for the ratio of the two to satisfy (2.17). This ensures that the acceptance ratios will be as large as they can be while still satisfying the relevant conditions, and indeed that the ratio in one direction will be unity, which means that in that direction at least, moves will always be accepted.

However, the best thing we can do to keep the acceptance ratios large is to try to embody in the selection probabilities $g(\mu \to \nu)$ as much as we can of the dependence of $P(\mu \to \nu)$ on the characteristics of the states μ and ν, and put as little as we can in the acceptance ratio. The ideal algorithm is one in which the new states are selected with exactly the correct transition probabilities all the time, and the acceptance ratio is always one. A good algorithm is one in which the acceptance probability is usually close to one. Much of the effort invested in the algorithms described in this book is directed at making the acceptance ratios large.

2.4 Continuous time Monte Carlo

There is a another twist we can add to our Markov process to allow ourselves further freedom about the way in which we choose states, without letting the acceptance ratios get too low. It is called **continuous time Monte Carlo**, or sometimes the **BKL algorithm**, after Bortz, Kalos and Lebowitz (1975), who invented it. Continuous time Monte Carlo is not nearly as widely used as it ought to be; it is an important and powerful technique and many calculations can be helped enormously by making use of it.

Consider a system at low temperature. Such systems are always a problem where Monte Carlo methods are concerned—cool systems move from state to state very slowly in real life and the problem is no less apparent in simulations. A low-temperature system is a good example of the sort of problem system that was described in the last section. Once it reaches equilibrium at its low temperature it will spend a lot of its time in the ground state. Maybe it will spend a hundred consecutive time-steps of the simu-

lation in the ground state, then move up to the first excited state for one time-step and then relax back to the ground state again. Such behaviour is not unreasonable for a cold system but we waste a lot of computer time simulating it. Time-step after time-step our algorithm selects a possible move to some excited state, but the acceptance ratio is very low and virtually all of these possible moves are rejected, and the system just ends up spending most of its time in the ground state.

Well, what if we were to accept that this is the case, and take a look at the acceptance ratio for a move from the ground state to the first excited state, and say to ourselves, "Judging by this acceptance ratio, this system is going to spend a hundred time-steps in the ground state before it accepts a move to the first excited state". Then we could jump the gun by *assuming* that the system will do this, miss out the calculations involved in the intervening useless one hundred time-steps, and progress straight to the one time-step in which something interesting happens. This is the essence of the idea behind the continuous time method. In this technique, we have a time-step which corresponds to a varying length of time, depending on how long we expect the system to remain in its present state before moving to a new one. Then when we come to take the average of our observable Q over many states, we weight the states in which the system spends longest the most heavily—the calculation of the estimator of Q is no more than a time average, so each value Q_μ for Q in state μ should be weighted by how long the system spends in that state.

How can we adapt our previous ideas concerning the transition probabilities for our Markov process to take this new idea into account? Well, assuming that the system is in some state μ, we can calculate how long a time Δt (measured in steps of the simulation) it will stay there for before a move to another state is accepted by considering the "stay-at-home" probability $P(\mu \to \mu)$. The probability that it is still in this same state μ after t time-steps is just

$$[P(\mu \to \mu)]^t = \mathrm{e}^{t \log P(\mu \to \mu)}, \tag{2.18}$$

and so the time-scale Δt is

$$\begin{aligned} \Delta t &= -\frac{1}{\log P(\mu \to \mu)} = -\frac{1}{\log[1 - \sum_{\nu \neq \mu} P(\mu \to \nu)]} \\ &\simeq \frac{1}{\sum_{\nu \neq \mu} P(\mu \to \nu)}. \end{aligned} \tag{2.19}$$

So, if we can calculate this quantity Δt, then rather than wait this many time-steps for a Monte Carlo move to get accepted, we can simply pretend that we have done the waiting and go right ahead and change the state of the system to a new state $\nu \neq \mu$. Which state should we choose for ν? We should choose one at random, but in proportion to $P(\mu \to \nu)$. Thus our continuous time Monte Carlo algorithm consists of the following steps:

1. We calculate the probabilities $P(\mu \to \nu)$ for transitions to all states which can be reached in one Monte Carlo step from the current state μ. We choose a new state ν with probability proportional to $P(\mu \to \nu)$ and change the state of the system to ν.

2. Using our values for the $P(\mu \to \nu)$ we also calculate the time interval Δt. Notice that we have to recalculate Δt at each step, since in general it will change from one step to the next.

3. We increment the time t by Δt, to mimic the effect of waiting Δt Monte Carlo steps. The variable t keeps a record of how long the simulation has gone on for in "equivalent Monte Carlo steps".

While this technique is in many respects a very elegant solution to the problem of simulating a system at low temperatures (or any other system which has a low acceptance ratio), it does suffer from one obvious drawback, which is that step (1) above involves calculating $P(\mu \to \nu)$ for every possible state ν which is accessible from μ. There may be very many such states (for some systems the number goes up exponentially with the size of the system), and so this step may take a very long time. However, in some cases, it turns out that the set of transition probabilities is very similar from one step of the algorithm to the next, only a few of them changing at each step, and hence it is possible to keep a table of probabilities and update only a few entries at each step to keep the table current. In cases such as these the continuous time method becomes very efficient and can save us a great deal of CPU time, despite being more complex than the accept/reject method discussed in the previous section. One example of a continuous time Monte Carlo algorithm is presented in Section 5.2.1 for the conserved-order-parameter Ising model.

In the next few chapters, we will examine a number of common models used for calculating the equilibrium properties of condensed matter systems, and show how the general ideas presented in this chapter can be used to find efficient numerical solutions to these physical problems.

Problems

2.1 Derive Equation (2.8) from Equation (1.1).

2.2 Consider a system which has just three energy states, with energies $E_0 < E_1 < E_2$. Suppose that the only allowed transitions are ones of the form $\mu \to \nu$, where $\nu = (\mu + 1) \bmod 3$. Such a system cannot satisfy detailed balance. Show nonetheless that it is possible to choose the transition probabilities $P(\mu \to \nu)$ so that the Boltzmann distribution is an equilibrium of the dynamics.

3

The Ising model and the Metropolis algorithm

In Section 1.2.2 we introduced the Ising model, which is one of the simplest and best studied of statistical mechanical models. In this chapter and the next we look in detail at the Monte Carlo methods that have been used to investigate the properties of this model. As well as demonstrating the application of the basic principles described in the last chapter, the study of the Ising model provides an excellent introduction to the most important Monte Carlo algorithms in use today. Along the way we will also look at some of the tricks used for implementing Monte Carlo algorithms in computer programs and at some of the standard techniques used to analyse the data those programs generate.

To recap briefly, the Ising model is a simple model of a magnet, in which dipoles or "spins" s_i are placed on the sites i of a lattice. Each spin can take either of two values: $+1$ and -1. If there are N sites on the lattice, then the system can be in 2^N states, and the energy of any particular state is given by the Ising Hamiltonian:

$$H = -J \sum_{\langle ij \rangle} s_i s_j - B \sum_i s_i, \tag{3.1}$$

where J is an interaction energy between nearest-neighbour spins $\langle ij \rangle$, and B is an external magnetic field. We are interested in simulating an Ising system of finite size using Monte Carlo methods, so that we can estimate the values of quantities such as the magnetization m (Equation (1.34)) or the specific heat c (Equation (1.37)) at any given temperature. Most of the interesting questions concerning the Ising model can be answered by performing simulations in zero magnetic field $B = 0$, so for the moment at least we will concentrate on this case.

3.1 The Metropolis algorithm

The very first Monte Carlo algorithm we introduce in this book is the most famous and widely used algorithm of them all, the **Metropolis algorithm**, which was introduced by Nicolas Metropolis and his co-workers in a 1953 paper on simulations of hard-sphere gases (Metropolis *et al.* 1953). We will use this algorithm to illustrate many of the general concepts involved in a real Monte Carlo calculation, including equilibration, measurement of expectation values, and the calculation of errors. First however, let us see how the algorithm is arrived at, and how one might go about implementing it on a computer.

The derivation of the Metropolis algorithm follows exactly the plan we outlined in Section 2.3. We choose a set of selection probabilities $g(\mu \to \nu)$, one for each possible transition from one state to another, $\mu \to \nu$, and then we choose a set of acceptance probabilities $A(\mu \to \nu)$ such that Equation (2.17) satisfies the condition of detailed balance, Equation (2.14). The algorithm works by repeatedly choosing a new state ν, and then accepting or rejecting it at random with our chosen acceptance probability. If the state is accepted, the computer changes the system to the new state ν. If not, it just leaves it as it is. And then the process is repeated again and again.

The selection probabilities $g(\mu \to \nu)$ should be chosen so that the condition of ergodicity—the requirement that every state be accessible from every other in a finite number of steps—is fulfilled (see Section 2.2.2). This still leaves us a good deal of latitude about how they are chosen; given an initial state μ we can generate any number of candidate states ν simply by flipping different subsets of the spins on the lattice. However, as we demonstrated in Section 1.2.1, the energies of systems in thermal equilibrium stay within a very narrow range—the energy fluctuations are small by comparison with the energy of the entire system. In other words, the real system spends most of its time in a subset of states with a narrow range of energies and rarely makes transitions that change the energy of the system dramatically. This tells us that we probably don't want to spend much time in our simulation considering transitions to states whose energy is very different from the energy of the present state. The simplest way of achieving this in the Ising model is to consider only those states which differ from the present one by the flip of a single spin. An algorithm which does this is said to have **single-spin-flip dynamics**. The algorithm we describe in this chapter has single-spin-flip dynamics, although this is not what makes it the Metropolis algorithm. (As discussed below, it is the particular choice of acceptance ratio that characterizes the Metropolis algorithm. Our algorithm would still be a Metropolis algorithm even if it flipped many spins at once.)

Using single-spin-flip dynamics guarantees that the new state ν will have an energy E_ν differing from the current energy E_μ by at most $2J$ for each

bond between the spin we flip and its neighbours. For example, on a square lattice in two dimensions each spin has four neighbours, so the maximum difference in energy would be $8J$. The general expression is $2zJ$, where z is the **lattice coordination number**, i.e., the number of neighbours that each site on the lattice has.[1] Using single-spin-flip dynamics also ensures that our algorithm obeys ergodicity, since it is clear that we can get from any state to any other on a finite lattice by flipping one by one each of the spins by which the two states differ.

In the Metropolis algorithm the selection probabilities $g(\mu \to \nu)$ for each of the possible states ν are all chosen to be equal. The selection probabilities of all other states are set to zero. Suppose there are N spins in the system we are simulating. With single-spin-flip dynamics there are then N different spins that we could flip, and hence N possible states ν which we can reach from a given state μ. Thus there are N selection probabilities $g(\mu \to \nu)$ which are non-zero, and each of them takes the value

$$g(\mu \to \nu) = \frac{1}{N}. \tag{3.2}$$

With these selection probabilities, the condition of detailed balance, Equation (2.14), takes the form

$$\frac{P(\mu \to \nu)}{P(\nu \to \mu)} = \frac{g(\mu \to \nu)A(\mu \to \nu)}{g(\nu \to \mu)A(\nu \to \mu)} = \frac{A(\mu \to \nu)}{A(\nu \to \mu)} = \mathrm{e}^{-\beta(E_\nu - E_\mu)}. \tag{3.3}$$

Now we have to choose the acceptance ratios $A(\mu \to \nu)$ to satisfy this equation. As we pointed out in Section 2.2.3, one possibility is to choose

$$A(\mu \to \nu) = A_0 \mathrm{e}^{-\frac{1}{2}\beta(E_\nu - E_\mu)}. \tag{3.4}$$

The constant of proportionality A_0 cancels out in Equation (3.3), so we can choose any value for it that we like, except that $A(\mu \to \nu)$, being a probability, should never be allowed to become greater than one. As we mentioned above, the largest difference in energy $E_\nu - E_\mu$ that we can have between our two states is $2zJ$, where z is the lattice coordination number. That means that the largest value of $\mathrm{e}^{-\frac{1}{2}\beta(E_\nu - E_\mu)}$ is $\mathrm{e}^{\beta z J}$. Thus, in order to make sure $A(\mu \to \nu) \leq 1$ we want to choose

$$A_0 \leq \mathrm{e}^{-\beta z J}. \tag{3.5}$$

To make the algorithm as efficient as possible, we want the acceptance probabilities to be as large as possible, so we make A_0 as large as it is allowed to

[1] This is not the same thing as the "spin coordination number" which we introduce in Chapter 5. The spin coordination number is the number of spins j neighbouring i which have the same value as spin i: $s_j = s_i$.

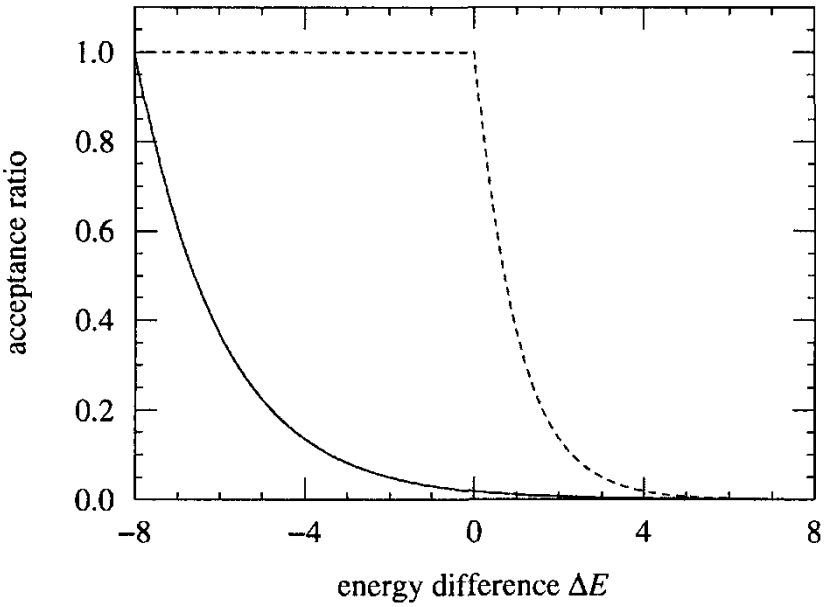

FIGURE 3.1 Plot of the acceptance ratio given in Equation (3.6) (solid line). This acceptance ratio gives rise to an algorithm which samples the Boltzmann distribution correctly, but is very inefficient, since it rejects the vast majority of the moves it selects for consideration. The Metropolis acceptance ratio (dashed line) is much more efficient.

be, which gives us

$$A(\mu \to \nu) = \mathrm{e}^{-\frac{1}{2}\beta(E_\nu - E_\mu + 2zJ)}. \tag{3.6}$$

This is not the Metropolis algorithm (we are coming to that), but using this acceptance probability we can perform a Monte Carlo simulation of the Ising model, and it will correctly sample the Boltzmann distribution. However, the simulation will be very inefficient, because the acceptance ratio, Equation (3.6), is very small for almost all moves. Figure 3.1 shows the acceptance ratio (solid line) as a function of the energy difference $\Delta E = E_\nu - E_\mu$ over the allowed range of values for a simulation with $\beta = J = 1$ and a lattice coordination number $z = 4$, as on a square lattice for example. As we can see, although $A(\mu \to \nu)$ starts off at 1 for $\Delta E = -8$, it quickly falls to only about 0.13 at $\Delta E = -4$, and to only 0.02 when $\Delta E = 0$. The chances of making any move for which $\Delta E > 0$ are pitifully small, and in practice this means that an algorithm making use of this acceptance ratio would be tremendously slow, spending most of its time rejecting moves and not flipping any spins at all. The solution to this problem is as follows.

In Equation (3.4) we have assumed a particular functional form for the acceptance ratio, but the condition of detailed balance, Equation (3.3), doesn't actually require that it take this form. Equation (3.3) only specifies the ratio of pairs of acceptance probabilities, which still leaves us quite a lot of room to manœuvre. In fact, as we pointed out in Section 2.3, when given a con-

straint like (3.3) the way to maximize the acceptance ratios (and therefore produce the most efficient algorithm) is always to give the larger of the two ratios the largest value possible—namely 1—and then adjust the other to satisfy the constraint. To see how that works out in this case, suppose that of the two states μ and ν we are considering here, μ has the lower energy and ν the higher: $E_\mu < E_\nu$. Then the larger of the two acceptance ratios is $A(\nu \to \mu)$, so we set that equal to one. In order to satisfy Equation (3.3), $A(\mu \to \nu)$ must then take the value $e^{-\beta(E_\nu - E_\mu)}$. Thus the optimal algorithm is one in which

$$A(\mu \to \nu) = \begin{cases} e^{-\beta(E_\nu - E_\mu)} & \text{if } E_\nu - E_\mu > 0 \\ 1 & \text{otherwise.} \end{cases} \tag{3.7}$$

In other words, if we select a new state which has an energy lower than or equal to the present one, we should always accept the transition to that state. If it has a higher energy then we maybe accept it, with the probability given above. This is the Metropolis algorithm for the Ising model with single-spin-flip dynamics. It is Equation (3.7) which makes it the Metropolis algorithm. This is the part that was pioneered by Metropolis and co-workers in their paper on hard-sphere gases, and any algorithm, applied to any model, which chooses selection probabilities according to a rule like (3.7) can be said to be a Metropolis algorithm. At first, this rule may seem a little strange, especially the part about how we *always* accept a move that will lower the energy of the system. The first algorithm we suggested, Equation (3.6), seems much more natural in this respect, since it sometimes rejects moves to lower energy. However, as we have shown, the Metropolis algorithm satisfies detailed balance, and is by far the more efficient algorithm, so, natural or not, it has become the algorithm of choice in the overwhelming majority of Monte Carlo studies of simple statistical mechanical models in the last forty years. We have also plotted Equation (3.7) in Figure 3.1 (dashed line) for comparison between the two algorithms.

3.1.1 Implementing the Metropolis algorithm

Let us look now at how we would actually go about writing a computer program to perform a simulation of the Ising model using the Metropolis algorithm. For simplicity we will continue to focus on the case of zero magnetic field $B = 0$, although the generalization to the case $B \neq 0$ is not hard (see Problem 3.1). In fact almost all the past studies of the Ising model, including Onsager's exact solution in two dimensions, have looked only at the zero-field case.

First, we need an actual lattice of spins to work with, so we would define a set of N variables—an array—which can take the values ± 1. Probably we would use integer variables, so it would be an integer array. Normally, we

apply **periodic boundary conditions** to the array. That is, we specify that the spins on one edge of the lattice are neighbours of the corresponding spins on the other edge. This ensures that all spins have the same number of neighbours and local geometry, and that there are no special edge spins which have different properties from the others; all the spins are equivalent and the system is completely translationally invariant. In practice this considerably improves the quality of the results from our simulation.

A variation on the idea of periodic boundary conditions is to use "helical boundary conditions" which are only very slightly different from periodic ones and possess all the same benefits but are usually considerably simpler to implement and can make our simulation significantly faster. The various types of boundary conditions and their implementation are described in detail in Section 13.1, along with methods for representing most common lattice geometries using arrays.

Next we need to decide at what temperature, or alternatively at what value of β we want to perform our simulation, and we need to choose some starting value for each of the spins—the initial state of the system. In a lot of cases, the initial state we choose is not particularly important, though sometimes a judicious choice can reduce the time taken to come to equilibrium (see Section 3.2). The two most commonly used initial states are the zero-temperature state and the infinite temperature state. At $T = 0$ the Ising model will be in its ground state. When the interaction energy J is greater than zero and the external field B is zero (as is the case in the simulations we will present in this chapter) there are actually two ground states. These are the states in which the spins are all up or all down. It is easy to see that these must be ground states, since in these states each pair of spins in the first term of Equation (3.1) contributes the lowest possible energy $-J$ to the Hamiltonian. In any other state there will be pairs of spins which contribute $+J$ to the Hamiltonian, so that its overall value will be higher. (If $B \neq 0$ then there will only be one ground state—the field ensures that one of the two is favoured over the other.) The other commonly used initial state is the $T = \infty$ state. When $T = \infty$ the thermal energy kT available to flip the spins is infinitely larger than the energy due to the spin–spin interaction J, so the spins are just oriented randomly up or down in an uncorrelated fashion.

These two choices of initial state are popular because they each correspond to a known, well defined temperature, and they are easy to generate. There is, however, one other initial state which can sometimes be very useful, which we should mention. Often we don't just perform one simulation at a single temperature, but rather a set of simulations one after another at a range of different values of T, to probe the behaviour of the model with varying temperature. In this case it is often advantageous to us to choose as the initial state of our system the *final* state of the system for a simulation

at a nearby temperature. For example, suppose we are interested in probing a range of temperatures between $T = 1.0$ and $T = 2.0$ in steps of 0.1. (Here and throughout much of the rest of this book, we measure temperature in energy units, so that $k = 1$. Thus when we say $T = 2.0$ we mean that $\beta^{-1} = 2.0$.) Then we might start off by performing a simulation at $T = 1.0$ using the zero-temperature state with all spins aligned as our initial state. At the end of the simulation, the system will be in equilibrium at $T = 1.0$, and we can use the final state of that simulation as the initial state for the simulation at $T = 1.1$, and so on. The justification for doing this is clear: we hope that the equilibrium state at $T = 1.0$ will be more similar to that at $T = 1.1$ than will the zero-temperature state. In most cases this is a correct assumption and our system will come to equilibrium quicker with this initial state than with either a $T = 0$ or a $T = \infty$ one.

Now we start our simulation. The first step is to generate a new state—the one we called ν in the discussion above. The new state should differ from the present one by the flip of just one spin, and every such state should be exactly as likely as every other to be generated. This is an easy task to perform. We just pick a single spin k at random from the lattice to be flipped. Next, we need to calculate the difference in energy $E_\nu - E_\mu$ between the new state and the old one, in order to apply Equation (3.7). The most straightforward way to do this would be to calculate E_μ directly by substituting the values s_i^μ of the spins in state μ into the Hamiltonian (3.1), then flip spin k and calculate E_ν, and take the difference. This, however, is not a very efficient way to do it. Even in zero magnetic field $B = 0$ we still have to perform the sum in the first term of (3.1), which has as many terms as there are bonds on the lattice, which is $\frac{1}{2}Nz$. But most of these terms don't change when we flip our single spin. The only ones that change are those that involve the flipped spin. The others stay the same and so cancel out when we take the difference $E_\nu - E_\mu$. The *change* in energy between the two states is thus

$$\begin{aligned} E_\nu - E_\mu &= -J \sum_{\langle ij \rangle} s_i^\nu s_j^\nu + J \sum_{\langle ij \rangle} s_i^\mu s_j^\mu \\ &= -J \sum_{i \text{ n.n. to } k} s_i^\mu (s_k^\nu - s_k^\mu). \end{aligned} \tag{3.8}$$

In the second line the sum is over only those spins i which are nearest neighbours of the flipped spin k and we have made use of the fact that all of these spins do not themselves flip, so that $s_i^\nu = s_i^\mu$. Now if $s_k^\mu = +1$, then after spin k has been flipped we must have $s_k^\nu = -1$, so that $s_k^\nu - s_k^\mu = -2$. On the other hand, if $s_k^\mu = -1$ then $s_k^\nu - s_k^\mu = +2$. Thus we can write

$$s_k^\nu - s_k^\mu = -2s_k^\mu, \tag{3.9}$$

and so

$$E_\nu - E_\mu = 2J \sum_{i \text{ n.n. to } k} s_i^\mu s_k^\mu$$
$$= 2J s_k^\mu \sum_{i \text{ n.n. to } k} s_i^\mu. \tag{3.10}$$

This expression only involves summing over z terms, rather than $\frac{1}{2}Nz$, and it doesn't require us to perform any multiplications for the terms in the sum, so it is much more efficient than evaluating the change in energy directly. What's more, it involves only the values of the spins in state μ, so we can evaluate it *before* we actually flip the spin k.

The algorithm thus involves calculating $E_\nu - E_\mu$ from Equation (3.10) and then following the rule given in Equation (3.7): if $E_\nu - E_\mu \leq 0$ we definitely accept the move and flip the spin $s_k \to -s_k$. If $E_\nu - E_\mu > 0$ we still may want to flip the spin. The Metropolis algorithm tells us to flip it with probability $A(\mu \to \nu) = \mathrm{e}^{-\beta(E_\nu - E_\mu)}$. We can do this as follows. We evaluate the acceptance ratio $A(\mu \to \nu)$ using our value of $E_\nu - E_\mu$ from Equation (3.10), and then we choose a random number r between zero and one. (Strictly the number can be equal to zero, but it must be less than one: $0 \leq r < 1$.) If that number is *less* than our acceptance ratio, $r < A(\mu \to \nu)$, then we flip the spin. If it isn't, we leave the spin alone.

And that is our complete algorithm. Now we just keep on repeating the same calculations over and over again, choosing a spin, calculating the energy change we would get if we flipped it, and then deciding whether to flip it according to Equation (3.7). Actually, there is one other trick that we can pull that makes our algorithm a bit faster still. (In fact, on most computers it will make it a lot faster.) One of the slowest parts of the algorithm as we have described it is the calculation of the exponential, which we have to perform if the energy of the new state we choose is greater than that of the current state. Calculating exponentials on a computer is usually done using a polynomial approximation which involves performing a number of floating-point multiplications and additions, and can take a considerable amount of time. We can save ourselves this effort, and thereby speed up our simulation, if we notice that the quantity, Equation (3.10), which we are calculating the exponential of, can only take a rather small number of values. Each of the terms in the sum can only take the values $+1$ and -1. So the entire sum, which has z terms, can only take the values $-z, -z+2, -z+4 \ldots$ and so on up to $+z$—a total of $z+1$ possible values. And we only actually need to calculate the exponential when the sum is negative (see Equation (3.7) again), so in fact there are only $\frac{1}{2}z$ values of $E_\mu - E_\nu$ for which we ever need to calculate exponentials. Thus, it makes good sense to calculate the values of these $\frac{1}{2}z$ exponentials before we start the calculation proper, and store them in the computer's memory (usually in an array), where we can

simply look them up when we need them during the simulation. We pay the one-time cost of evaluating them at the beginning, and save a great deal more by never having to evaluate any exponentials again during the rest of the simulation. Not only does this save us the effort of evaluating all those exponentials, it also means that we hardly have to perform any floating-point arithmetic during the simulation. The only floating-point calculations will be in the generation of the random number r. (We discuss techniques for doing this in Chapter 16.) All the other calculations involve only integers, which on most computers are much quicker to deal with than real numbers.

3.2 Equilibration

So what do we do with our Monte Carlo program for the Ising model, once we have written it? Well, we probably want to know the answer to some questions like "What is the magnetization at such-and-such a temperature?", or "How does the internal energy behave with temperature over such-and-such a range?" To answer these questions we have to do two things. First we have to run our simulation for a suitably long period of time until it has come to equilibrium at the temperature we are interested in—this period is called the **equilibration time** τ_{eq}—and then we have to measure the quantity we are interested in over another suitably long period of time and average it, to evaluate the estimator of that quantity (see Equation (2.4)). This leads us to several other questions. What exactly do we mean by "allowing the system to come to equilibrium"? And how long is a "suitably long" time for it to happen? How do we go about measuring our quantity of interest, and how long do we have to average over to get a result of a desired degree of accuracy? These are very general questions which we need to consider every time we do a Monte Carlo calculation. Although we will be discussing them here using our Ising model simulation as an example, the conclusions we will draw in this and the following sections are applicable to all equilibrium Monte Carlo calculations. These sections are some of the most important in this book.

As we discussed in Section 1.2, "equilibrium" means that the average probability of finding our system in any particular state μ is proportional to the Boltzmann weight $e^{-\beta E_\mu}$ of that state. If we start our system off in a state such as the $T = 0$ or $T = \infty$ states described in the last section and we want to perform a simulation at some finite non-zero temperature, it will take a little time before we reach equilibrium. To see this, recall that, as we demonstrated in Section 1.2.1, a system at equilibrium spends the overwhelming majority of its time in a small subset of states in which its internal energy and other properties take a narrow range of values. In order to get a good estimate of the equilibrium value of any property of the system therefore, we need to wait until it has found its way to one of the states that

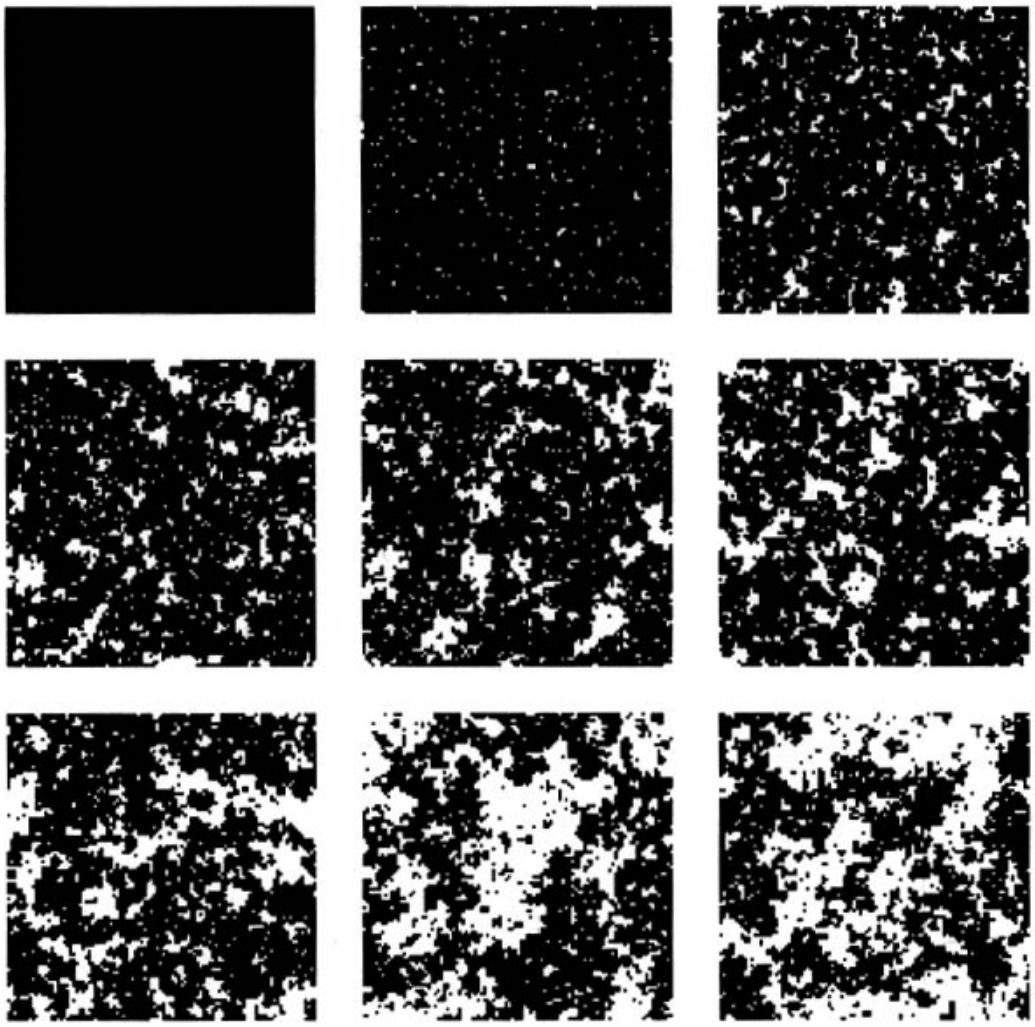

FIGURE 3.2 Nine snapshots of a 100×100 Ising model on a square lattice with $J = 1$ coming to equilibrium at a temperature $T = 2.4$ using the Metropolis algorithm. In these pictures the up-spins ($s_i = +1$) are represented by black squares and the down-spins ($s_i = -1$) by white ones. The starting configuration is one in which all the spins are pointing up. The progression of the figures is horizontally across the top row, then the middle row, then the bottom one. They show the lattice after 0, 1, 2, 4, 6, 10, 20, 40 and 100 times 100 000 steps of the simulation. In the last frame the system has reached equilibrium according to the criteria given in this section.

fall in this narrow range. Then, we assume, the Monte Carlo algorithm we have designed will ensure that it stays roughly within that range for the rest of the simulation—it should do since we designed the algorithm specifically to simulate the behaviour of the system at equilibrium. But it may take some time to find a state that lies within the correct range. In the version of the Metropolis algorithm which we have described here, we can only flip one spin at a time, and since we are choosing the spins we flip at random, it could take quite a while before we hit on the correct sequence of spins to

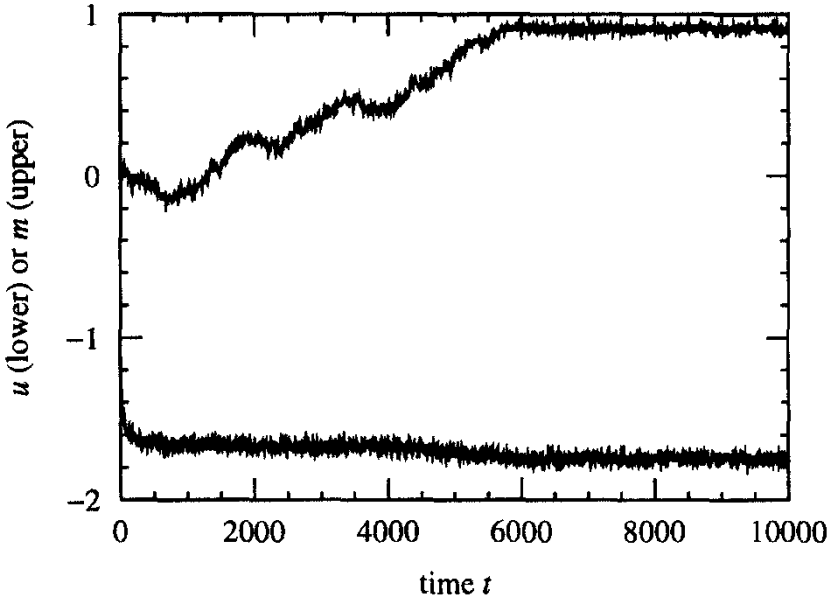

FIGURE 3.3 The magnetization (upper curve) and internal energy (lower curve) per site of a two-dimensional Ising model on a square lattice of 100×100 sites with $J = 1$ simulated using the Metropolis algorithm of Section 3.1. The simulation was started at $T = \infty$ (i.e., the initial states of the spins were chosen completely at random) and "cooled" to equilibrium at $T = 2.0$. Time is measured in Monte Carlo steps per lattice site, and equilibrium is reached after about 6000 steps per site (in other words, 6×10^7 steps altogether).

flip in order to get us to one of the states we want to be in. At the very least we can expect it to take about N Monte Carlo steps to reach a state in the appropriate energy range, where N is the number of spins on the lattice, since we need to allow every spin the chance to flip at least once. In Figure 3.2 we show a succession of states of a two-dimension Ising model on a square lattice of 100×100 spins with $J = 1$, as it is "warmed" up to a temperature $T = 2.4$ from an initial $T = 0$ state in which all the spins are aligned. In these pictures the $+1$ and -1 spins are depicted as black and white squares. By the time we reach the last frame out of nine, the system has equilibrated. The whole process takes on the order of 10^7 steps in this case.

However looking at pictures of the lattice is not a reliable way of gauging when the system has come to equilibrium. A better way, which takes very little extra effort, is to plot a graph of some quantity of interest, like the magnetization per spin m of the system or the energy of the system E, as a function of time from the start of the simulation. We have done this in Figure 3.3. (We will discuss the best ways of measuring these quantities in the next section, but for the moment let's just assume that we calculate them directly. For example, the energy of a given state can be calculated by

feeding all the values of the spins s_i into the Hamiltonian, Equation (3.1).) It is not hard to guess simply by looking at this graph that the system came to equilibrium at around time $t = 6000$. Up until this point the energy and the magnetization are changing, but after this point they just fluctuate around a steady average value.

The horizontal axis in Figure 3.3 measures time in Monte Carlo steps *per lattice site*, which is the normal practice for simulations of this kind. The reason is that if time is measured in this way, then the average frequency with which any particular spin is selected for flipping is independent of the total number of spins N on the lattice. This average frequency is called the "attempt frequency" for the spin. In the simulation we are considering here the attempt frequency has the value 1. It is natural that we should arrange for the attempt frequency to be independent of the lattice size; in an experimental system, the rate at which spins or atoms or molecules change from one state to another does not depend on how many there are in the whole system. An atom in a tiny sample will change state as often as one in a sample the size of a house. Attempt frequencies are discussed further in Section 11.1.1.

When we perform N Monte Carlo steps—one for each spin in the system, on average—we say we have completed one **sweep** of the lattice. We could therefore also say that the time axis of Figure 3.3 was calibrated in sweeps.

Judging the equilibration of a system by eye from a plot such as Figure 3.3 is a reasonable method, provided we know that the system will come to equilibrium in a smooth and predictable fashion as it does in this case. The trouble is that we usually know no such thing. In many cases it is possible for the system to get stuck in some metastable region of its state space for a while, giving roughly constant values for all the quantities we are observing and so appearing to have reached equilibrium. In statistical mechanical terms, there can be a **local energy minimum** in which the system can remain temporarily, and we may mistake this for the **global energy minimum**, which is the region of state space that the equilibrium system is most likely to inhabit. (These ideas are discussed in more detail in the first few sections of Chapter 6.) To avoid this potential pitfall, we commonly adopt a different strategy for determining the equilibration time, in which we perform two different simulations of the same system, starting them in different initial states. In the case of the Ising model we might, for example, start one in the $T = 0$ state with all spins aligned, and one in the $T = \infty$ state with random spins. Or we could choose two different $T = \infty$ random-spin states. We should also run the two simulations with different "seeds" for the random number generator (see Section 16.1.2), to ensure that they take different paths to equilibrium. Then we watch the value of the magnetization or energy or other quantity in the two systems and when we see them reach the same approximately constant value, we deduce that

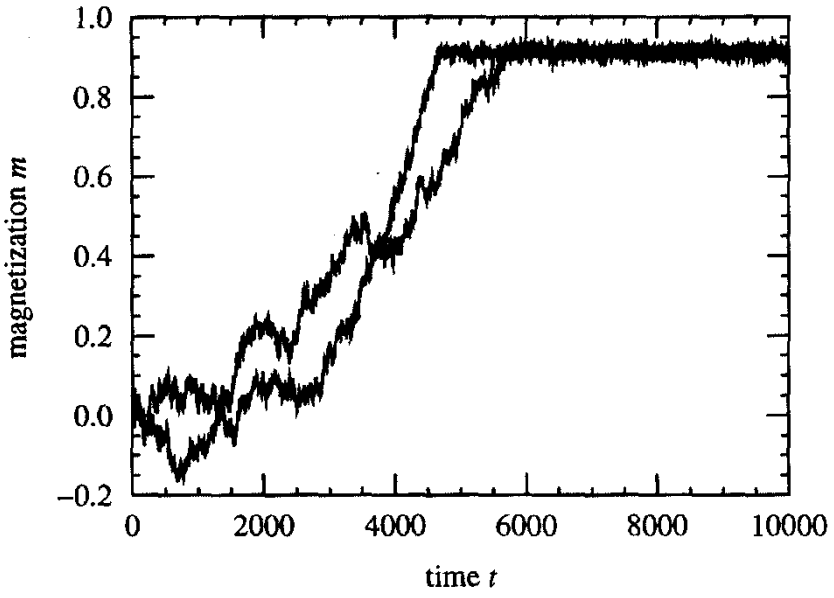

FIGURE 3.4 The magnetization of our 100 × 100 Ising model as a function of time (measured in Monte Carlo steps per lattice site) for two different simulations using the Metropolis algorithm. The two simulations were started off in two different $T = \infty$ (random-spin) states. By about time $t = 6000$ the two simulations have converged to the same value of the mean magnetization, within the statistical errors due to fluctuations, and so we conclude that both have equilibrated.

both systems have reached equilibrium. We have done this for two 100×100 Ising systems in Figure 3.4. Again, we clearly see that it takes about 6000 Monte Carlo steps for the two systems to reach a consensus about the value of the magnetization. This technique avoids the problem mentioned above, since if one of the systems finds itself in some metastable region, and the other reaches equilibrium or gets stuck in another metastable region, this will be apparent from the graph, because the magnetization (or other quantity) will take different values for the two systems. Only in the unlikely event that the two systems coincidentally become trapped in the same metastable region (for example, if we choose two initial states that are too similar to one another) will we be misled into thinking they have reached equilibrium when they haven't. If we are worried about this possibility, we can run *three* different simulations from different starting points, or four, or five. Usually, however, two is sufficient.

3.3 Measurement

Once we are sure the system has reached equilibrium, we need to measure whatever quantity it is that we are interested in. The most likely candidates

for the Ising model are the energy and the magnetization of the system. As we pointed out above, the energy E_μ of the current state μ of the system can be evaluated directly from the Hamiltonian by substituting in the values of the spins s_i from our array of integers. However, this is not an especially efficient way of doing it, and there is a much better way. As part of our implementation of the Metropolis algorithm, you will recall we calculated the energy difference $\Delta E = E_\nu - E_\mu$ in going from state μ to state ν (see Equation (3.10)). So, if we know the energy of the current state μ, we can calculate the new energy when we flip a spin, using only a single addition:

$$E_\nu = E_\mu + \Delta E. \tag{3.11}$$

So the clever thing to do is to calculate the energy of the system from the Hamiltonian at the very start of the simulation, and then every time we flip a spin calculate the new energy from Equation (3.11) using the value of ΔE, which we have to calculate anyway.

Calculating the magnetization is even easier. The total magnetization M_μ of the whole system in state μ (as opposed to the magnetization per *spin*—we'll calculate that in a moment), is given by the sum

$$M_\mu = \sum_i s_i^\mu. \tag{3.12}$$

As with the energy, it is not a shrewd idea to evaluate the magnetization directly from this sum every time we want to know it. It is much better to notice that only one spin k flips at a time in the Metropolis algorithm, so the change of magnetization from state μ to state ν is

$$\Delta M = M_\nu - M_\mu = \sum_i s_i^\nu - \sum_i s_i^\mu = s_k^\nu - s_k^\mu = 2s_k^\nu, \tag{3.13}$$

where the last equality follows from Equation (3.9). Thus, the clever way to evaluate the magnetization is to calculate its value at the beginning of the simulation, and then make use of the formula

$$M_\nu = M_\mu + \Delta M = M_\mu + 2s_k^\nu \tag{3.14}$$

every time we flip a spin.[2]

[2]However, to be absolutely fair, we should point out that doing this involves performing at least one addition operation every time we flip a spin, or one addition every $\bar{A}^{-1}$ steps, where $\bar{A}$ is the mean acceptance ratio. Direct evaluation of Equation (3.12) on the other hand involves N additions every time we want to know the magnetization. Thus, if we want to make measurements less often than once every $N/\bar{A}$ steps, it may pay to use the direct method rather than employing Equation (3.14). Similar considerations apply to the measurement of the energy also.

Given the energy and the magnetization of our Ising system at a selection of times during the simulation, we can average them to find the estimators of the internal energy and average magnetization. Then dividing these figures by the number of sites N gives us the internal energy and average magnetization per site.

We can also average the squares of the energy and magnetization to find quantities like the specific heat and the magnetic susceptibility:

$$c = \frac{\beta^2}{N}(\langle E^2\rangle - \langle E\rangle^2), \tag{3.15}$$

$$\chi = \beta N(\langle m^2\rangle - \langle m\rangle^2). \tag{3.16}$$

(See Equations (1.36) and (1.37). Note that we have set $k = 1$ again.)

In order to average quantities like E and M, we need to know how long a run we have to average them over to get a good estimate of their expectation values. One simple solution would again be just to look at a graph like Figure 3.3 and guess how long we need to wait. However (as you might imagine) this is not a very satisfactory solution. What we really need is a measure of the **correlation time** τ of the simulation. The correlation time is a measure of how long it takes the system to get from one state to another one which is significantly different from the first, i.e., a state in which the number of spins which are the same as in the initial state is no more than what you would expect to find just by chance. (We will give a more rigorous definition in a moment.) There are a number of ways to estimate the correlation time. One that is sometimes used is just to assume that it is equal to the equilibration time. This is usually a fairly safe assumption:[3] usually the equilibration time is considerably longer than the correlation time, $\tau_{\text{eq}} > \tau$, because two states close to equilibrium are qualitatively more similar than a state far from equilibrium (like the $T = 0$ or $T = \infty$ states we suggested for starting this simulation with) and one close to equilibrium. However, this is again a rather unrigorous supposition, and there are more watertight ways to estimate τ. The most direct of these is to calculate the "time-displaced autocorrelation function" of some property of the model.

3.3.1 Autocorrelation functions

Let us take the example of the magnetization m of our Ising model. The **time-displaced autocorrelation** $\chi(t)$ of the magnetization is given by

$$\begin{aligned}\chi(t) &= \int \mathrm{d}t' \, [m(t') - \langle m\rangle][m(t'+t) - \langle m\rangle] \\ &= \int \mathrm{d}t' \, \big[m(t')m(t'+t) - \langle m\rangle^2\big].\end{aligned} \tag{3.17}$$

[3] In particular, it works fine for the Metropolis simulation of the Ising model which we are considering here.

where $m(t)$ is the instantaneous value of the magnetization at time t and $\langle m \rangle$ is the average value. This is rather similar to the connected correlation function which we defined in Equation (1.26). That measured the correlation between the values of a quantity (such as the magnetization) on two different sites, i and j, on the lattice. The autocorrelation gives us a similar measure of the correlation at two different times, one an interval t later than the other. If we measure the difference between the magnetization $m(t')$ at time t' and its mean value, and then we do the same thing at time $t' + t$, and we multiply them together, we will get a positive value if they were fluctuating in the same direction at those two times, and a negative one if they were fluctuating in opposite directions. If we then integrate over time as in Equation (3.17), then $\chi(t)$ will take a non-zero value if on average the fluctuations are correlated, or it will be zero if they are not. For our Metropolis simulation of the Ising model it is clear that if we measure the magnetization at two times just a single Monte Carlo step apart, the values we get will be very similar, so we will have a large positive autocorrelation. On the other hand, for two times a long way apart the magnetizations will probably be totally unrelated, and their autocorrelation will be close to zero. Ideally, we should calculate $\chi(t)$ by integrating over an infinite time, but this is obviously impractical in a simulation, so we do the best we can and just sum over all the measurements of m that we have, from beginning to end of our run. Figure 3.5 shows the magnetization autocorrelation of our 100×100 Ising model at temperature $T = 2.4$ and interaction energy $J = 1$, calculated in exactly this manner using results from our Metropolis Monte Carlo simulation. As we can see, the autocorrelation does indeed drop from a significant non-zero value at short times t towards zero at very long times. In this case, we have divided $\chi(t)$ by its value $\chi(0)$ at $t = 0$, so that its maximum value is one. The typical time-scale (if there is one) on which it falls off is a measure of the correlation time τ of the simulation. In fact, this is the *definition* of the correlation time. It is the typical time-scale on which the autocorrelation drops off; the autocorrelation is expected to fall off exponentially at long times thus:

$$\chi(t) \sim \mathrm{e}^{-t/\tau}. \tag{3.18}$$

(In the next section we show why it should take this form.) With this definition, we see that in fact there is still a significant correlation between two samples taken a correlation time apart: at time $t = \tau$ the autocorrelation function, which is a measure of the similarity of the two states, is only a factor of $1/\mathrm{e}$ down from its maximum value at $t = 0$. If we want truly independent samples then, we may want to draw them at intervals of greater than one correlation time. In fact, the most natural definition of statistical independence turns out to be samples drawn at intervals of 2τ. We discuss this point further in Section 3.4.1.

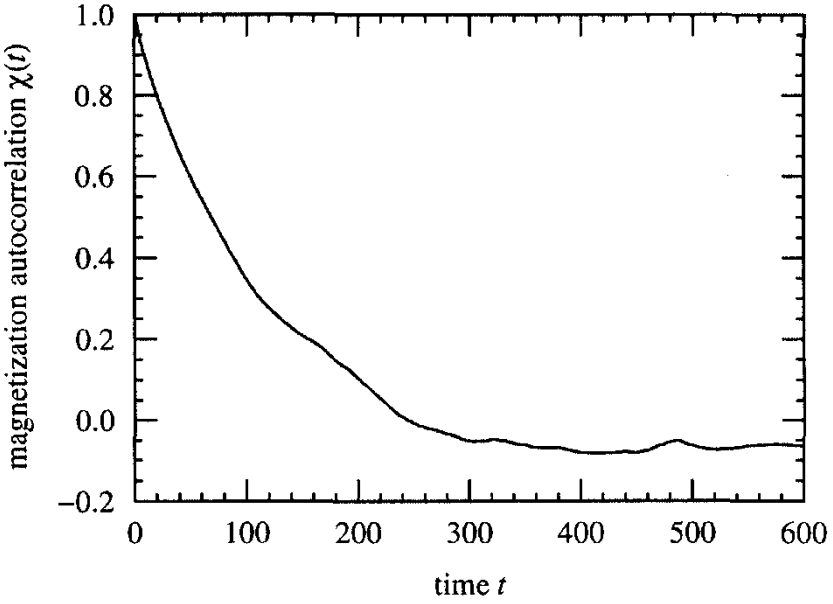

FIGURE 3.5 The magnetization autocorrelation function $\chi(t)$ for a two-dimensional Ising model at temperature $T = 2.4$ on a square lattice of 100×100 sites with $J = 1$ simulated using the Metropolis algorithm of Section 3.1. Time is measured in Monte Carlo steps per site.

We can make a reasonable estimate of τ by eye from Figure 3.5. At a guess we'd probably say τ was about 100 in this case. This is an accurate enough figure for estimating how long a Monte Carlo run we need to do in order to get decent statistics. It tells us that we expect to get a new independent spin configuration about once every $2\tau = 200$ sweeps of the lattice in our simulation. So if we want, say, 10 independent measurements of the magnetization, we need to run our Metropolis algorithm for about 2000 sweeps after equilibration, or 2×10^7 Monte Carlo steps. If we want 100 measurements we need to do 2×10^8 steps. In general, if a run lasts a time t_{max}, then the number of independent measurements we can get out of the run, after waiting a time τ_{eq} for equilibration, is on the order of

$$n = \frac{t_{\text{max}}}{2\tau}. \tag{3.19}$$

It is normal practice in a Monte Carlo simulation to make measurements at intervals of less than the correlation time. For example, in the case of the Metropolis algorithm we might make one measurement every sweep of the lattice. Thus the total number of measurements we make of magnetization or energy (or whatever) during the run is usually greater than the number of independent measurements. There are a number of reasons why we do it this way. First of all, we usually don't know what the correlation time is until after the simulation has finished, or at least until after it has run

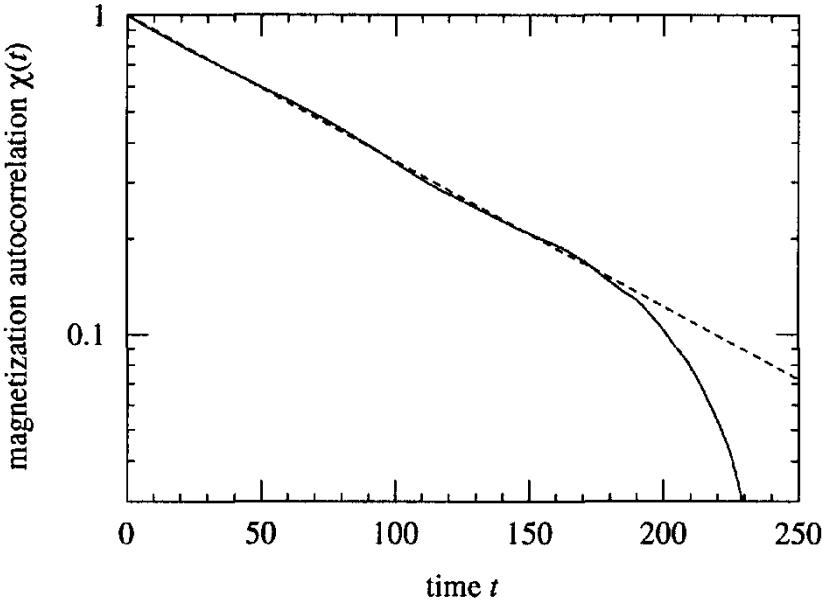

FIGURE 3.6 The autocorrelation function of Figure 3.5 replotted on semi-logarithmic axes. The dashed line is a straight line fit which yields a figure of $\tau = 95 \pm 5$ in this case. Note that the horizontal scale is not the same as in Figure 3.5.

for a certain amount of time, and we want to be sure of having at least one measurement every two correlation times. Another reason is that we want to be able to calculate the autocorrelation function for times less than a correlation time, so that we can use it to make an accurate estimate of the correlation time. If we only had one measurement every 2τ, we wouldn't be able to calculate τ with any accuracy at all.

If we want a more reliable figure for τ, we can replot our autocorrelation function on semi-logarithmic axes as we have done in Figure 3.6, so that the slope of the line gives us the correlation time. Then we can estimate τ by fitting the straight-line portion of the plot using a least-squares method. The dotted line in the figure is just such a fit and its slope gives us a figure of $\tau = 95 \pm 5$ for the correlation time in this case.

An alternative is to calculate the **integrated correlation time.**[4] If we assume that Equation (3.18) is accurate for all times t then

$$\int_0^\infty \frac{\chi(t)}{\chi(0)}\,\mathrm{d}t = \int_0^\infty \mathrm{e}^{-t/\tau}\,\mathrm{d}t = \tau. \tag{3.20}$$

This form has a number of advantages. First, it is often easier to apply Equation (3.20) than it is to perform the exponential fit to the autocorrelation

[4]This is a rather poor name for this quantity, since it is not the correlation time that is integrated but the autocorrelation function. However, it is the name in common use so we use it here too.

function. Second, the method for estimating τ illustrated in Figure 3.6 is rather sensitive to the range over which we perform the fit. In particular, the very long time behaviour of the autocorrelation function is often noisy, and it is important to exclude this portion from the fitted data. However, the exact point at which we truncate the fit can have quite a large effect on the resulting value for τ. The integrated correlation time is much less sensitive to the way in which the data are truncated, although it is by no means perfect either, since, as we will demonstrate in Section 3.3.2, Equation (3.18) is only strictly correct for long times t, and we introduce an uncontrolled error by assuming it to be true for all times. On the other hand, the direct fitting method also suffers from this problem, unless we only perform our fit over the exact range of times for which true exponential behaviour is present. Normally, we don't know what this range is, so the fitting method is no more accurate than calculating the integrated correlation time. Usually then, Equation (3.20) is the method of choice for calculating τ. Applying it to the data from Figure 3.6 gives a figure of $\tau = 86 \pm 5$, which is in moderately good agreement with our previous figure.

The autocorrelation function in Figure 3.5 was calculated directly from a discrete form of Equation (3.17). If we have a set of samples of the magnetization $m(t)$ measured at evenly-spaced times up to some maximum time t_{max}, then the correct formula for the autocorrelation function is[5]

$$\chi(t) = \frac{1}{t_{\mathrm{max}} - t} \sum_{t'=0}^{t_{\mathrm{max}}-t} m(t')\, m(t'+t) - \frac{1}{t_{\mathrm{max}} - t} \sum_{t'=0}^{t_{\mathrm{max}}-t} m(t') \times \frac{1}{t_{\mathrm{max}} - t} \sum_{t'=0}^{t_{\mathrm{max}}-t} m(t'+t). \qquad (3.21)$$

Notice how we have evaluated the mean magnetization m in the second term using the same subsets of the data that we used in the first term. This is not strictly speaking necessary, but it makes $\chi(t)$ a little better behaved. In Figure 3.5 we have also normalized $\chi(t)$ by dividing throughout by $\chi(0)$, but this is optional. We've just done it for neatness.

Note that one should be careful about using Equation (3.21) to evaluate $\chi(t)$ at long times. When t gets close to t_{max}, the upper limit of the sums becomes small and we end up integrating over a rather small time interval to get our answer. This means that the statistical errors in $\chi(t)$ due to the random nature of the fluctuations in $m(t)$ may become large. A really satisfactory simulation would always run for many correlation times, in which case we will probably not be interested in the very tails of $\chi(t)$, since the correlations will have died away by then, by definition. However, it is not

[5] In fact, this formula differs from (3.17) by a multiplicative constant, but this makes no difference as far as the calculation of the correlation time is concerned.

always possible, because we only have limited computer resources, to perform simulations as long as we would like, and one should always be aware that errors of this type can crop up with shorter data sets.

Calculating the autocorrelation function from Equation (3.21) takes time of order n^2, where n is the number of samples. For most applications, this is not a problem. Even for simulations where several thousand samples are taken, the time needed to evaluate the autocorrelation is only a few seconds on a modern computer, and the simplicity of the formulae makes their programming very straightforward, which is a big advantage—people-time is usually more expensive than computer-time. However, sometimes it is desirable to calculate the autocorrelation function more quickly. This is particularly the case when one needs to do it very often during the course of a calculation, for some reason. If you need to calculate an autocorrelation a thousand times, and each time takes a few seconds on the computer, then the seconds start to add up. In this case, at the expense of rather greater programming effort, we can often speed up the process by the following trick. Instead of calculating $\chi(t)$ directly, we calculate the Fourier transform $\widetilde{\chi}(\omega)$ of the autocorrelation and then we invert the Fourier transform to get $\chi(t)$. The Fourier transform is related to the magnetization as follows:

$$\begin{aligned}\widetilde{\chi}(\omega) &= \int \mathrm{d}t\, \mathrm{e}^{\mathrm{i}\omega t} \int \mathrm{d}t'\, [m(t') - \langle m\rangle][m(t'+t) - \langle m\rangle] \\ &= \int \mathrm{d}t \int \mathrm{d}t'\, \mathrm{e}^{-\mathrm{i}\omega t'}[m(t') - \langle m\rangle]\, \mathrm{e}^{\mathrm{i}\omega(t'+t)}[m(t'+t) - \langle m\rangle] \\ &= \widetilde{m}'(\omega)\,\widetilde{m}'(-\omega) = \left|\widetilde{m}'(\omega)\right|^2, \end{aligned} \tag{3.22}$$

where $\widetilde{m}'(\omega)$ is the Fourier transform of $m'(t) \equiv m(t) - \langle m\rangle$.[6]

So all we need to do is calculate the Fourier transform of $m'(t)$ and feed it into this formula to get $\widetilde{\chi}(\omega)$. The advantage in doing this is that the Fourier transform can be evaluated using the so-called **fast Fourier transform** or FFT algorithm, which was given by Cooley and Tukey in 1965. This is a standard algorithm which can evaluate the Fourier transform in a time which goes like $n \log n$, where n is the number of measurements of the magnetization.[7] Furthermore, there exist a large number of ready-made computer software packages that perform FFTs. These packages have

[6] Since $m(t)$ and $m'(t)$ differ only by a constant, $\widetilde{m}(\omega)$ and $\widetilde{m}'(\omega)$ differ only in their $\omega = 0$ component (which is zero in the latter case, but may be non-zero in the former). For this reason, it is often simplest to calculate $\widetilde{m}'(\omega)$ by first calculating $\widetilde{m}(\omega)$ and then just setting the $\omega = 0$ component to zero.

[7] On a technical note, if we simply apply the FFT algorithm directly to our magnetization data, the result produced is the Fourier transform of an infinite periodic repetition of the data set, which is not quite what we want in Equation (3.22). A simple way of getting around this problem is to add n zeros to the end of the data set before we perform the transform. It can be shown that this then gives a good estimate of the autocorrelation function (Futrelle and McGinty 1971).

been written very carefully by people who understand the exact workings of the computer and are designed to be as fast as possible at performing this particular calculation, so it usually saves us much time and effort to make use of the programs in these packages. Having calculated $\widetilde{\chi}(\omega)$, we can then invert the Fourier transform, again using a streamlined inverse FFT routine which also runs in time proportional to $n \log n$, and so recover the function $\chi(t)$.

We might imagine that, if we wanted to calculate the integrated correlation time, Equation (3.20), we could avoid inverting the Fourier transform, since

$$\widetilde{\chi}(0) = \int_0^\infty \chi(t)\, dt, \tag{3.23}$$

and thus

$$\tau = \frac{\widetilde{\chi}(0)}{\chi(0)}, \tag{3.24}$$

where $\chi(0)$ is simply the magnetization fluctuation

$$\chi(0) = \langle m^2 \rangle - \langle m \rangle^2. \tag{3.25}$$

However, you should avoid calculating τ this way because, as footnote 6 on the previous page makes clear, $\widetilde{\chi}(0)$ is zero when calculated directly for finite datasets. Equation (3.24) is only applicable if $\langle m \rangle$ is calculated over a much longer run than $\widetilde{\chi}(\omega)$.

3.3.2 Correlation times and Markov matrices

The techniques outlined in the previous section are in most cases quite sufficient for estimating correlation times in Monte Carlo simulations. However, we have simplified the discussion somewhat by supposing there to be only one correlation time in the system. In real life there are as many correlation times as there are states of the system, and the interplay between these different times can sometimes cause the methods of the last section to give inaccurate results. In this section we look at these issues in more detail. The reader who is interested only in how to calculate a rough measure of τ could reasonably skip this section.

In Section 2.2.3 we showed that the probabilities $w_\mu(t)$ and $w_\mu(t+1)$ of being in a particular state μ at consecutive Monte Carlo steps are related by the Markov matrix $\mathbf{P}$ for the algorithm. In matrix notation we wrote

$$\mathbf{w}(t+1) = \mathbf{P} \cdot \mathbf{w}(t), \tag{3.26}$$

where $\mathbf{w}$ is the vector whose elements are the probabilities w_μ (see Equation (2.9)). By iterating this equation from time $t = 0$ we can then show that

$$\mathbf{w}(t) = \mathbf{P}^t \cdot \mathbf{w}(0). \tag{3.27}$$

Now $\mathbf{w}(0)$ can be expressed as a linear combination of the right eigenvectors $\mathbf{v}_i$ of $\mathbf{P}$ thus:[8]

$$\mathbf{w}(0) = \sum_i a_i \mathbf{v}_i, \tag{3.28}$$

where the quantities a_i are coefficients whose values depend on the configuration of the system at $t = 0$. Then

$$\mathbf{w}(t) = \mathbf{P}^t \cdot \sum_i a_i \mathbf{v}_i = \sum_i a_i \lambda_i^t \mathbf{v}_i, \tag{3.29}$$

where λ_i is the eigenvalue of $\mathbf{P}$ corresponding to the eigenvector $\mathbf{v}_i$. As $t \to \infty$, the right-hand side of this equation will be dominated by the term involving the largest eigenvalue λ_0 of the Markov matrix. This means that in the limit of long times, the probability distribution $\mathbf{w}(t)$ becomes proportional to $\mathbf{v}_0$, the eigenvector corresponding to the largest eigenvalue. We made use of this result in Section 2.2.3 to demonstrate that $\mathbf{w}(t)$ tends to the Boltzmann distribution at long times.

Now suppose that we are interested in knowing the value of some observable quantity Q, such as the magnetization. The expectation value of this quantity at time t can be calculated from the formula

$$Q(t) = \sum_\mu w_\mu(t)\, Q_\mu \tag{3.30}$$

or

$$Q(t) = \mathbf{q} \cdot \mathbf{w}(t), \tag{3.31}$$

where $\mathbf{q}$ is the vector whose elements are the values Q_μ of the quantity in the various states of the system. Substituting Equation (3.29) into (3.31) we then get

$$Q(t) = \sum_i a_i \lambda_i^t\, \mathbf{q} \cdot \mathbf{v}_i = \sum_i a_i \lambda_i^t\, q_i. \tag{3.32}$$

Here $q_i \equiv \mathbf{q} \cdot \mathbf{v}_i$ is the expectation value of Q in the i^{th} eigenstate. The long time limit $Q(\infty)$ of this expression is also dominated by the largest eigenvalue λ_0 and is proportional to q_0. If we now define a set of quantities τ_i thus:

$$\tau_i = -\frac{1}{\log \lambda_i} \tag{3.33}$$

for all $i \neq 0$, then Equation (3.32) can be written

$$Q(t) = Q(\infty) + \sum_{i \neq 0} a_i q_i \mathrm{e}^{-t/\tau_i}. \tag{3.34}$$

[8] $\mathbf{P}$ is in general not symmetric, so its right and left eigenvectors are not the same.

The quantities τ_i are the correlation times for the system, as we can demonstrate by showing that they also govern the decay of the autocorrelation function. Noting that the long-time limit $Q(\infty)$ of the expectation of Q is none other than the equilibrium expectation $\langle Q \rangle$, we can write the autocorrelation of Q as the correlation between the expectations at zero time and some later time t thus:

$$\chi(t) = [Q(0) - Q(\infty)][Q(t) - Q(\infty)] = \sum_{i \neq 0} b_i e^{-t/\tau_i}, \tag{3.35}$$

with

$$b_i = \sum_{j \neq 0} a_i a_j q_i q_j. \tag{3.36}$$

Equation (3.35) is the appropriate generalization of Equation (3.18) to all times t (not just long ones).

As we said, there are as many correlation times as there are states of the system since that is the rank of the matrix $\mathbf{P}$. (Well, strictly there are as many of them as the rank of the matrix *less one*, since there is no τ_0 corresponding to the highest eigenvalue. There are $2^N - 1$ correlation times in the case of the Ising model, for example. However, the rank of the matrix is usually very large, so let's not quibble over one correlation time.) The longest of these correlation times is τ_1, the one which corresponds to the second largest eigenvalue of the matrix. This is the correlation time we called τ in the last section. Clearly, for large enough times t, this will be the only correlation time we need to worry about, since all the other terms in Equation (3.35) will have decayed away to insignificance. (This is how Equation (3.18) is derived.) However, depending on how close together the higher eigenvalues of the Markov matrix are, and how long our simulation runs for, we may or may not be able to extract reliable results for τ_1 by simply ignoring all the other terms. In general the most accurate results are obtained by fitting our autocorrelation function to a sum of a small number of decaying exponentials, of the form of Equation (3.35), choosing values for the quantities b_i by a least-squares or similar method. In work on the three-dimensional Ising model, for example, Wansleben and Landau (1991) showed that including three terms was sufficient to get a good fit to the magnetization autocorrelation function, and thus get an accurate measure of the longest correlation time $\tau \equiv \tau_1$. In studies of the dynamical properties of statistical mechanical models, this is the most correct way to measure the correlation time. Strictly speaking it gives only a lower bound on τ since it is always possible that correlations exist beyond the longest times that one can measure. However, in practice it usually gives good results.

3.4 Calculation of errors

Normally, as well as measuring expectation values, we also want to calculate the errors on those values, so that we have an idea of how accurate they are. As with experiments, the errors on Monte Carlo results divide into two classes: **statistical errors** and **systematic errors**.[9] Statistical errors are errors which arise as a result of random changes in the simulated system from measurement to measurement—thermal fluctuations, for example—and they can be estimated simply by taking many measurements of the quantity we are interested in and calculating the spread of the values. Systematic errors on the other hand, are errors due to the procedure we have used to make the measurements, and they affect the whole simulation. An example is the error introduced by waiting only a finite amount of time for our system to equilibrate. (Ideally, we should allow an infinite amount of time for this, in order to be sure the system has completely equilibrated. However, this, of course, is not practical.)

3.4.1 Estimation of statistical errors

In a Monte Carlo calculation the principal source of statistical error in the measured value of a quantity is usually the fluctuation of that quantity from one time step to the next. This error is inherent in the Monte Carlo method. As the name "Monte Carlo" itself makes clear, there is an innate randomness and statistical nature to Monte Carlo calculations. (In Chapter 6 on glassy spin models, we will see another source of statistical error: "sample-to-sample" fluctuations in the actual system being simulated. However, for the simple Ising model we have been considering, thermal fluctuations are the only source of statistical error. All other errors fall into the category of systematic errors.) It is often straightforward to estimate the statistical error in a measured quantity, since the assumption that the error is statistical—i.e., that it arises through random deviations in the measured value of the quantity—implies that we can estimate the true value by taking the mean of several different measurements, and that the error on that estimate is simply the error on the mean. Thus, if we are performing the Ising model simulation described in the last section, and we make n measurements m_i of the magnetization of the system during a particular run, then our best estimate of the true thermal average of the magnetization is the mean $\overline{m}$ of those n measurements (which is just the estimator of m, as defined in Section 2.1),

[9] Monte Carlo simulations are in many ways rather similar to experiments. It often helps to regard them as "computer experiments", and analyse the results in the same way as we would analyse the results of a laboratory experiment.

and our best estimate of the standard deviation on the mean is given by[10]

$$\sigma = \sqrt{\frac{\frac{1}{n}\sum_{i=0}^{n}(m_i - \overline{m})^2}{n-1}} = \sqrt{\frac{1}{n-1}(\overline{m^2} - \overline{m}^2)}. \tag{3.37}$$

This expression assumes that our samples m_i are statistically independent, which in general they won't be. As we pointed out in Section 3.3.1, it is normal to sample at intervals of less than a correlation time, which means that successive samples will in general be correlated. A simple and usually perfectly adequate solution to this problem is to use the value of n given by Equation (3.19), rather than the actual number of samples taken. In fact, it can be shown (Müller-Krumbhaar and Binder 1973) that the correct expression for σ in terms of the actual number of samples is

$$\sigma = \sqrt{\frac{1 + 2\tau/\Delta t}{n-1}(\overline{m^2} - \overline{m}^2)}, \tag{3.38}$$

where τ is the correlation time and Δt is the time interval at which the samples were taken. Clearly this becomes equal to Equation (3.37) when $\Delta t \gg \tau$, but more often we have $\Delta t \ll \tau$. In this case, we can ignore the 1 in the numerator of Equation (3.38). Noting that for a run of length $t_{\max}$ (after equilibration) the interval Δt is related to the total number of samples by

$$n = \frac{t_{\max}}{\Delta t}, \tag{3.39}$$

we then find that for large n

$$\sigma = \sqrt{\frac{2\tau}{t_{\max}}(\overline{m^2} - \overline{m}^2)}, \tag{3.40}$$

which is the same result as we would get by simply using Equation (3.19) for n in Equation (3.37). This in fact was the basis for our assertion in Section 3.3.1 that the appropriate sampling interval for getting independent samples was twice the correlation time. Note that the value of σ in Equation (3.40) is independent of the value of Δt, which means we are free to choose Δt in whatever way is most convenient.

3.4.2 The blocking method

There are some cases where it is either not possible or not straightforward to estimate the error in a quantity using the direct method described in the last

[10]The origin of the $n - 1$ in this and following expressions for error estimates is a little obscure. The curious reader is referred to any good book on data analysis for an explanation, such as Bevington and Robinson (1992).

section. This happens when the result we want is not merely the average of some measurement repeated many times over the course of the simulation, as the magnetization is, but is instead derived in some more complex way from measurements we make during the run. An example is the specific heat c, Equation (3.15), which is inherently an average macroscopic quantity. Unlike the magnetization, the specific heat is not defined at a single time step in the simulation. It is only defined in terms of averages of many measurements of E and E^2 over a longer period of time. We might imagine that we could calculate the error on $\langle E \rangle$ and the error on $\langle E^2 \rangle$ using the techniques we employed for the magnetization, and then combine them in some fashion to give an estimate of the error in c. But this is not as straightforward as it seems at first, since the errors in these two quantities are correlated—when $\langle E \rangle$ goes up, so does $\langle E^2 \rangle$. It is possible to do the analysis necessary to calculate the error on c in this fashion. However, it is not particularly simple, and there are other more general methods of error estimation which lend themselves to this problem. As the quantities we want to measure become more complex, these methods—"blocking", the "bootstrap" and the "jackknife"—will save us a great deal of effort in estimating errors. We illustrate these methods here for the case of the specific heat, though it should be clear that they are applicable to almost any quantity that can be measured in a Monte Carlo simulation.

The simplest of our general-purpose error estimation methods is the **blocking method**. Applied to the specific heat, the idea is that we take the measurements of E that we made during the simulation and divide them into several groups, or **blocks**. We then calculate c separately for each block, and the spread of values from one block to another gives us an estimate of the error. To see how this works, suppose we make 200 measurements of the energy during our Ising model simulation, and then split those into 10 groups of 20 measurements. We can evaluate the specific heat from Equation (3.15) for each group and then find the mean of those 10 results exactly as we did for the magnetization above. The error on the mean is given again by Equation (3.37), except that n is now replaced by the number n_b of blocks, which would be 10 in our example. This method is intuitive, and will give a reasonable estimate of the order of magnitude of the error in a quantity such as c. However, the estimates it gives vary depending on the number of different blocks you divide your data up into, with the smallest being associated with large numbers of blocks, and the largest with small numbers of blocks, so it is clearly not a very rigorous method. A related but more reliable method, which can be used for error estimation in a wide variety of different circumstances, is the **bootstrap method**, which we now describe.

3.4.3 The bootstrap method

The bootstrap method is a **resampling method**. Applied to our problem of calculating the specific heat for the Ising model, it would work like this. We take our list of measurements of the energy of the model and from the n numbers in this list we pick out n at random. (Strictly, n should be the number of *independent* measurements. In practice the measurements made are usually not all independent, but luckily it transpires that the bootstrap method is not much affected by this difference, a point which is discussed further below.) We specifically allow ourselves to pick the same number twice from the list, so that we end up with n numbers each of which appears on the original list, and some of which may be duplicates of one another. (In fact, if you do your sums, you can show that about a fraction $1-1/\mathrm{e} \simeq 63\%$ of the numbers will be duplicates.) We calculate the specific heat from these n numbers just as we would normally, and then we repeat the process, picking (or **resampling**) another n numbers at random from the original measurements. It can be shown (Efron 1979) that after we have repeated this calculation several times, the standard deviation of the distribution in the results for c is a measure of the error in the value of c. In other words if we make several of these "bootstrap" calculations of the specific heat, our estimate of the error σ is given by

$$\sigma = \sqrt{\overline{c^2} - \bar{c}^2}. \tag{3.41}$$

Notice that there is no extra factor of $1/(n-1)$ here as there was in Equation (3.37). (It is clear that the latter would not give a correct result, since it would imply that our estimate of the error could be reduced by simply resampling our data more times.)

As we mentioned, it is not necessary for the working of the bootstrap method that all the measurements made be independent in the sense of Section 3.3.1 (i.e., one every two correlation times or more). As we pointed out earlier, it is more common in a Monte Carlo simulation to make measurements at comfortably short intervals throughout the simulation so as to be sure of making at least one every correlation time or so and then calculate the number of independent measurements made using Equation (3.19). Thus the number of samples taken usually exceeds the number which actually constitute independent measurements. One of the nice things about the bootstrap method is that it is not necessary to compensate for this difference in applying the method. You get fundamentally the same estimate of the error if you simply resample your n measurements from the entire set of measurements that were made. In this case, still about 63% of the samples will be duplicates of one another, but many others will effectively be duplicates as well because they will be measurements taken at times less than a correlation time apart. Nonetheless, the resulting estimate of σ is the

same.

The bootstrap method is a good general method for estimating errors in quantities measured by Monte Carlo simulation. Although the method initially met with some opposition from mathematicians who were not convinced of its statistical validity, it is now widely accepted as giving good estimates of errors (Efron 1979).

3.4.4 The jackknife method

A slightly different but related method of error estimation is the **jackknife**. For this method, unlike the bootstrap method, we really do need to choose n independent samples out of those that were made during the run, taking one approximately every two correlation times or more. Applying the jackknife method to the case of the specific heat, we would first use these samples to calculate a value c for the specific heat. Now however, we also calculate n other estimates c_i as follows. We take our set of n measurements, and we remove the first one, leaving $n-1$, and we calculate the specific heat c_1 from that subset. Then we put the first one back, but remove the second and calculate c_2 from that subset, and so forth. Each c_i is the specific heat calculated with the i^{th} measurement of the energy removed from the set, leaving $n-1$ measurements. It can then be shown that an estimate of the error in our value of c is

$$\sigma = \sqrt{\sum_{i=1}^{n}(c_i - c)^2}, \tag{3.42}$$

where c is our estimate of the specific heat using all the data.[11]

Both the jackknife and the bootstrap give good estimates of errors for large data sets, and as the size of the data set becomes infinite they give exact estimates. Which one we choose in a particular case usually depends on how much work is involved applying them. In order to get a decent error estimate from the bootstrap method we usually need to take at least 100 resampled sets of data, and 1000 would not be excessive. (100 would give the error to a bit better than 10% accuracy.) With the jackknife we have to recalculate the quantity we are interested in exactly n times to get the error estimate. So, if n is much larger than 100 or so, the bootstrap is probably the more efficient. Otherwise, we should use the jackknife.[12]

---

[11] In fact, it *is* possible to use the jackknife method with samples taken at intervals Δt less than 2τ. In this case we just reduce the sum inside the square root by a factor of $\Delta t/2\tau$ to get an estimate of σ which is independent of the sampling interval.

[12] For a more detailed discussion of these two methods, we refer the interested reader to the review article by Efron (1979).

3.4.5 Systematic errors

Just as in experiments, systematic errors are much harder to gauge than statistical ones; they don't show up in the fluctuations of the individual measurements of a quantity. (That's why they are systematic errors.) The main source of systematic error in the Ising model simulation described in this chapter is the fact that we wait only a finite amount of time for the system to equilibrate. There is no good general method for estimating systematic errors; each source of error has to be considered separately and a strategy for estimating it evolved. This is essentially what we were doing when we discussed ways of estimating the equilibration time for the Metropolis algorithm. Another possible source of systematic error would be not running the simulation for a long enough time after equilibration to make good independent measurements of the quantities of interest. When we discussed methods for estimating the correlation time τ, we were dealing with this problem. In the later sections of this chapter, and indeed throughout this book, we will discuss methods for estimating and controlling systematic errors as they crop up in various situations.

3.5 Measuring the entropy

We have described how we go about measuring the internal energy, specific heat, magnetization and magnetic susceptibility from our Metropolis Monte Carlo simulation of the Ising model. However, there are three quantities of interest which were mentioned in Chapter 1 which we have not yet discussed: the free energy F, the entropy S and the correlation function $G_c^{(2)}(i,j)$. The correlation function we will consider in the next section. The other two we consider now.

The free energy and the entropy are related by

$$F = U - TS \tag{3.43}$$

so that if we can calculate one, we can easily find the other using the known value of the total internal energy U. Normally, we calculate the entropy, which we do by integrating the specific heat over temperature as follows.

We can calculate the specific heat of our system from the fluctuations in the internal energy as described in Section 3.3. Moreover, we know that the specific heat C is equal to

$$C = T\frac{\mathrm{d}S}{\mathrm{d}T}. \tag{3.44}$$

Thus the entropy $S(T)$ at temperature T is

$$S(T) = S(T_0) + \int_{T_0}^{T} \frac{C}{T}\,\mathrm{d}T. \tag{3.45}$$

If we are only interested in how S varies with T, then it is not necessary to know the value of the integration constant $S(T_0)$ and we can give it any value we like. If we want to know the absolute value of S, then we have to fix $S(T_0)$ by choosing T_0 to be some temperature at which we know the value of the entropy. The conventional choice, known as the third law of thermodynamics, is to make the entropy zero[13] when $T_0 = 0$. In other words

$$S(T) = \int_0^T \frac{C}{T}\,\mathrm{d}T. \tag{3.46}$$

As with the other quantities we have discussed, we often prefer to calculate the entropy per spin $s(T)$ of our system, which is given in terms of the specific heat per spin c by

$$s(T) = \int_0^T \frac{c}{T}\,\mathrm{d}T. \tag{3.47}$$

Of course, evaluating either of these expressions involves calculating the specific heat over a range of temperatures up to the temperature we are interested in, at sufficiently small intervals that its variation with T is well approximated. Then we have to perform a numerical integral using, for example, the trapezium rule or any of a variety of more sophisticated integration techniques (see, for instance, Press *et al.* 1988). Calculating the entropy (or equivalently the free energy) of a system is therefore a more complex task than calculating the internal energy, and may use up considerably more computer time. On the other hand, if we are interested in probing the behaviour of our system over a range of temperatures anyway (as we often are), we may as well make use of the data to calculate the entropy; the integration is a small extra computational effort to make by comparison with the simulation itself.

The integration involved in the entropy calculation can give problems. In particular, if there is any very sharp behaviour in the curve of specific heat as a function of temperature, we may miss it in our simulation, which would give the integral the wrong value. This is a particular problem near "phase transitions", where the specific heat often diverges (see Section 3.7.1).

3.6 Measuring correlation functions

One other quantity which we frequently want to measure is the two-point connected correlation function $G_c^{(2)}(i,j)$. Let us see how we would go about

[13] Systems which have more than one ground state may violate the third law. This point is discussed in more detail in Section 7.1.2.

calculating this correlation function for the Ising model. The most straightforward way is to evaluate it directly from the definition, Equation (1.26), using the values of the spins s_i from our simulation:

$$G_c^{(2)}(i,j) = \langle s_i s_j \rangle - \langle s_i \rangle \langle s_j \rangle = \langle s_i s_j \rangle - m^2. \tag{3.48}$$

(Here m denotes the expectation value of the magnetization.) In fact, since our Ising system is translationally invariant, $G_c^{(2)}(i,j)$ is dependent only on the displacement $\mathbf{r}$ between the sites i and j, and not on exactly where they are. In other words, if $\mathbf{r}_i$ is the position vector of the i^{th} spin, then we should have

$$G_c^{(2)}(\mathbf{r}_i, \mathbf{r}_i + \mathbf{r}) = G_c^{(2)}(\mathbf{r}) \tag{3.49}$$

independent of the value of $\mathbf{r}_i$. This means that we can improve our estimate of the correlation function $G_c^{(2)}(\mathbf{r})$ by averaging its value over the whole lattice for all pairs of spins separated by a displacement $\mathbf{r}$:

$$G_c^{(2)}(\mathbf{r}) = \frac{1}{N} \sum_{\substack{i,j \text{ with} \\ \mathbf{r}_j - \mathbf{r}_i = \mathbf{r}}} \left[\langle s_i s_j \rangle - m^2 \right]. \tag{3.50}$$

If, as with our Ising model simulation, the system we are simulating has periodic boundary conditions (see Section 3.1), then $G_c^{(2)}(\mathbf{r})$ will not die away for very large values of $\mathbf{r}$. Instead, it will be periodic, dying away for values of $\mathbf{r}$ up to half the width of the lattice, and then building up again to another maximum when we have gone all the way across the lattice and got back to the spin we started with.

In order to evaluate $G_c^{(2)}(\mathbf{r})$ using Equation (3.50) we have to record the value of every single spin on the lattice at intervals during the simulation. This is not usually a big problem given the generous amounts of storage space provided by modern computers. However, if we want to calculate $G_c^{(2)}(\mathbf{r})$ for every value of $\mathbf{r}$ on the lattice, this kind of direct calculation does take an amount of time which scales with the number N of spins on the lattice as N^2. As with the calculation of the autocorrelation function in Section 3.3.1, it actually turns out to be quicker to calculate the Fourier transform of the correlation function instead.

The spatial Fourier transform $\widetilde{G}_c^{(2)}(\mathbf{k})$ is defined by

$$\begin{aligned}
\widetilde{G}_c^{(2)}(\mathbf{k}) &= \sum_{\mathbf{r}} \mathrm{e}^{\mathrm{i}\mathbf{k}\cdot\mathbf{r}} G_c^{(2)}(\mathbf{r}) \\
&= \frac{1}{N} \sum_{\mathbf{r}} \sum_{\substack{i,j \text{ with} \\ \mathbf{r}_j - \mathbf{r}_i = \mathbf{r}}} \mathrm{e}^{\mathrm{i}\mathbf{k}\cdot(\mathbf{r}_j - \mathbf{r}_i)} \left[\langle s_i s_j \rangle - m^2 \right] \\
&= \frac{1}{N} \left\langle \sum_{\mathbf{r}_i} \mathrm{e}^{-\mathrm{i}\mathbf{k}\cdot\mathbf{r}_i} (s_i - m) \sum_{\mathbf{r}_j} \mathrm{e}^{\mathrm{i}\mathbf{k}\cdot\mathbf{r}_j} (s_j - m) \right\rangle
\end{aligned}$$

$$= \frac{1}{N}\langle|\tilde{s}'(\mathbf{k})|^2\rangle, \tag{3.51}$$

where $\tilde{s}'(\mathbf{k})$ is the Fourier transform of $s'_i \equiv s_i - m$. In other words, in order to calculate $\widetilde{G}_c^{(2)}(\mathbf{k})$, we just need to perform a Fourier transform of the spins at a succession of different times throughout the simulation and then feed the results into Equation (3.51). As in the case of the autocorrelation function of Section 3.2, this can be done using a standard FFT algorithm. To get the correlation function in real space we then have to use the algorithm again to Fourier transform back, but the whole process still only takes a time which scales as $N \log N$, and so for the large lattices of today's Monte Carlo studies it is usually faster than direct calculation from Equation (3.50).[14]

Occasionally we also need to calculate the disconnected correlation function defined in Section 1.2.1. The equivalent of (3.51) in this case is simply

$$\widetilde{G}^{(2)}(\mathbf{k}) = \frac{1}{N}\langle|\tilde{s}(\mathbf{k})|^2\rangle. \tag{3.52}$$

Note that s_i and s'_i differ only by the average magnetization m, which is a constant. As a result, $\widetilde{G}^{(2)}(\mathbf{k})$ and $\widetilde{G}_c^{(2)}(\mathbf{k})$ are in fact identical, except at $\mathbf{k} = 0$. For this reason, it is often simpler to calculate $\widetilde{G}_c^{(2)}(\mathbf{k})$ by first calculating $\widetilde{G}^{(2)}(\mathbf{k})$ and then just setting the $\mathbf{k} = 0$ component to zero.

3.7 An actual calculation

In this section we go through the details of an actual Monte Carlo simulation and demonstrate how the calculation proceeds. The example that we take is that of the simulation of the two-dimensional Ising model on a square lattice using the Metropolis algorithm. This system has the advantage that its properties in the thermodynamic limit are known exactly, following the analytic solution given by Onsager (1944). Comparing the results from our simulation with the exact solution will give us a feel for the sort of accuracy one can expect to achieve using the Monte Carlo method. Some of the results have already been presented (see Figures 3.3 and 3.5 for example). Here we

[14] Again we should point out that this does not necessarily mean that one should always calculate the correlation function this way. As with the calculation of the autocorrelation function, using the Fourier transform is a more complicated method than direct calculation of the correlation function, and if your goal is to get an estimate quickly, and your lattice is not very large, you may be better advised to go the direct route. However, the Fourier transform method is more often of use in the present case of the two-point correlation function, since in order to perform the thermal average appearing in Equations (3.50) and (3.51) we need to repeat the calculation about once every two correlation times throughout the entire simulation, which might mean doing it a hundred or a thousand times in one run. Under these circumstances the FFT method may well be advantageous.

describe in detail how these and other results are arrived at, and discuss what conclusions we can draw from them.

The first step in performing any Monte Carlo calculation, once we have decided on the algorithm we are going to use, is to write the computer program to implement that algorithm. The code used for our Metropolis simulation of the Ising model is given in Appendix B. It is written in the computer language C.

As a test of whether the program is correct, we have first used it to simulate a small 5×5 Ising model for a variety of temperatures between $T = 0$ and $T = 5.0$ with J set equal to 1. For such a small system our program runs very fast, and the entire simulation only took about a second at each temperature. In Section 1.3 we performed an exact calculation of the magnetization and specific heat for the 5×5 Ising system by directly evaluating the partition function from a sum over all the states of the system. This gives us something to compare our Monte Carlo results with, so that we can tell if our program is doing the right thing. At this stage, we are not interested in doing a very rigorous calculation, only in performing a quick check of the program, so we have not made much effort to ensure the equilibration of the system or to measure the correlation time. Instead, we simply ran our program for 20 000 Monte Carlo steps per site (i.e., $20\,000 \times 25 = 500\,000$ steps in all), and averaged over the last 18 000 of these to measure the magnetization and the energy. Then we calculated m from Equation (3.12) and c from Equation (3.15). If the results do not agree with our exact calculation then it could mean either that there is a problem with the program, or that we have not waited long enough in either the equilibration or the measurement sections of the simulation. However, as shown in Figure 3.7, the numerical results agree rather well with the exact ones. Even though we have not calculated the statistical errors on our data in order to determine the degree of agreement, these results still give us enough confidence in our program to proceed with a more thorough calculation on a larger system.

For our large-scale simulation, we have chosen to examine a system of 100×100 spins on a square lattice. We started the program with randomly chosen values of all the spins—the $T = \infty$ state of Section 3.1.1—and ran the simulations at a variety of temperatures from $T = 0.2$ to $T = 5.0$ in steps of 0.2, for a total of 25 simulations in all.[15] Again we ran our simulations

[15]Note that it is not possible to perform a simulation at $T = 0$ because the acceptance ratio, Equation (3.7), for spin flips which increase the energy of the system becomes zero in this limit. This means that it is not possible to guarantee that the system will come to equilibrium, because the requirement of ergodicity is violated; there are some states which it is not possible to get to in a finite number of moves. It is true in general of thermal Monte Carlo methods that they break down at $T = 0$, and often they become very slow close to $T = 0$. The continuous time Monte Carlo method of Section 2.4 can sometimes by used to overcome this problem in cases where we are particularly interested

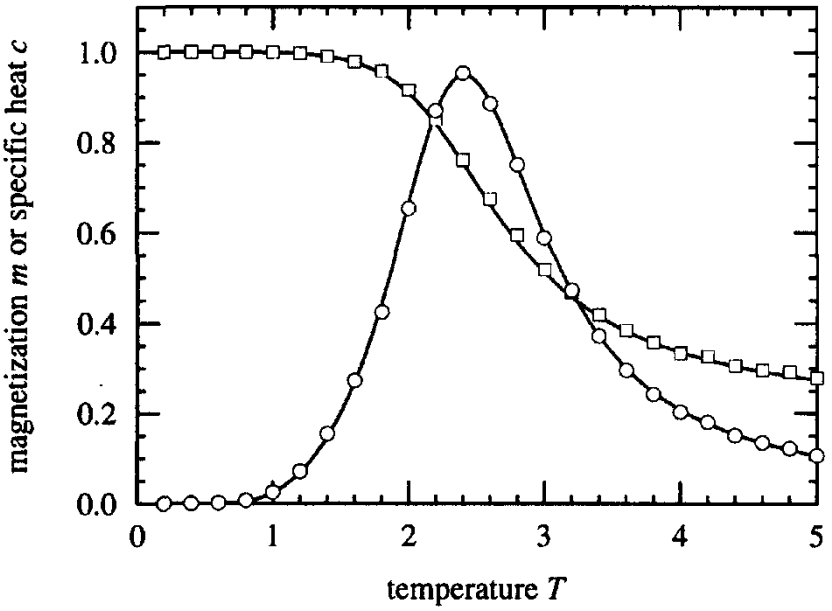

FIGURE 3.7 The magnetization (squares) and specific heat (circles) per spin of an Ising model in two dimensions on a 5×5 square lattice. The points are the results of the Metropolis Monte Carlo calculation described in the text. The lines are the exact calculations performed in Section 1.3, in which we evaluated the partition function by summing over all the states of the system.

for 20 000 Monte Carlo steps per lattice site. This is a fairly generous first run, and is only possible because we are looking at quite a small system still. In the case of larger or more complex models, one might well first perform a shorter run to get a rough measure of the equilibration and correlation times for the system, before deciding how long a simulation to perform. A still more sophisticated approach is to perform a short run and then store the configuration of the spins on the lattice before the program ends. Then, after deciding how long the entire calculation should last on the basis of the measurements during that short run, we can pick up exactly where we left off using the stored configuration, thus saving ourselves the effort of equilibrating the system twice.

Taking the data from our 25 simulations at different temperatures, we first estimate the equilibration times τ_{eq} at each temperature using the methods described in Section 3.2. In this case we found that all the equilibration times were less than about 1000 Monte Carlo steps per site, except for the simulations performed at $T = 2.0$ and $T = 2.2$, which both had equilibration times on the order of 6000 steps per site. (The reason for this anomaly is explained in the next section.) Allowing ourselves a margin for error in these estimates, we therefore took the data from time 2000 onwards as our equi-

in the behaviour of a model close to $T = 0$.

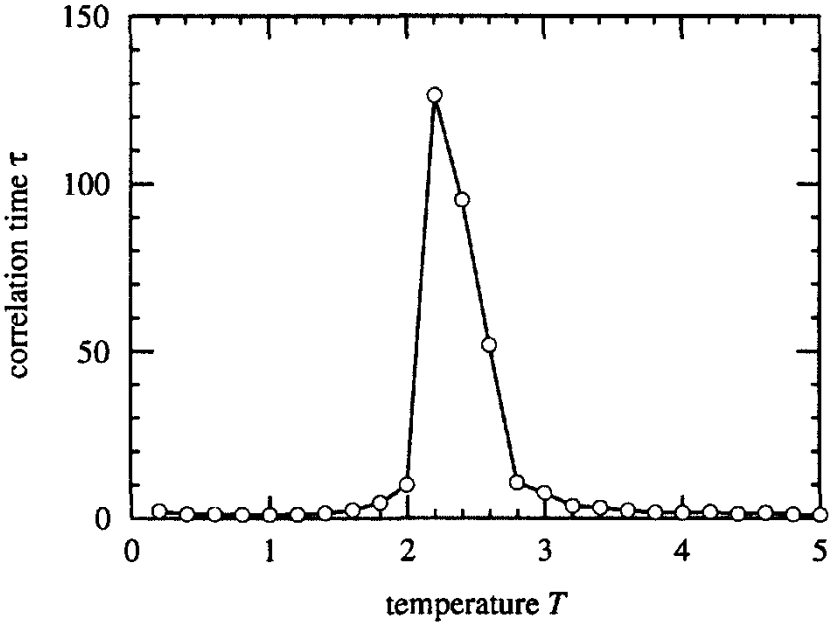

FIGURE 3.8 The correlation time for the 100×100 Ising model simulated using the Metropolis algorithm. The correlation time is measured in Monte Carlo steps per lattice site (i.e., in multiples of 10 000 Monte Carlo steps in this case). The straight lines joining the points are just to guide the eye.

librium measurements for all the temperatures except the two slower ones, for which we took the data for times 10 000 onwards.

Next we need to estimate how many independent measurements these data constitute, which means estimating the correlation time. To do this, we calculate the magnetization autocorrelation function at each temperature from Equation (3.21), for times t up to 1000. (We must be careful only to use our equilibrium data for this calculation since the autocorrelation function is an equilibrium quantity. That is, we should not use the data from the early part of the simulation during which the system was coming to equilibrium.) Performing a fit to these functions as in Figure 3.6, we make an estimate of the correlation time τ at each temperature. The results are shown in Figure 3.8. Note the peak in the correlation time around $T = 2.2$. This effect is called "critical slowing down", and we will discuss it in more detail in Section 3.7.2. Given the length of the simulation t_{max} and our estimates of τ_{eq} and τ for each temperature, we can calculate the number n of independent measurements to which our simulations correspond using Equation (3.19).

Using these figures we can now calculate the equilibrium properties of the 100×100 Ising model in two dimensions. As an example, we have calculated the magnetization and the specific heat again. Our estimate of the magnetization is calculated by averaging over the magnetization measurements from the simulation, again excluding the data from the early portion

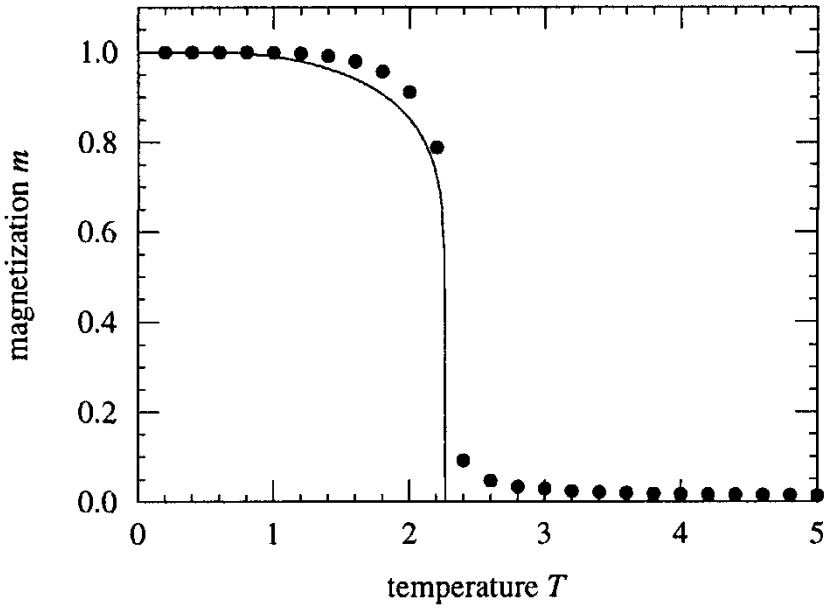

FIGURE 3.9 The magnetization per spin of the two-dimensional Ising model. The points are the results from our Monte Carlo simulation using the Metropolis algorithm. The errors are actually smaller than the points in this figure because the calculation is so accurate. The solid line is the known exact solution for the Ising model on an infinite two-dimensional square lattice.

of the run where the system was not equilibrated. The results are shown in Figure 3.9, along with the known exact solution for the infinite system. Calculating the errors on the magnetization from Equation (3.37), we find that the errors are so small that the error bars would be completely covered by the points themselves, so we have not bothered to put them in the figure. The agreement between the numerical calculation and the exact one for the infinite lattice is much better than it was for the smaller system in Figure 1.1, although there is still some discrepancy between the two. This discrepancy arises because the quantity plotted in the figure is in fact the average $\langle |m| \rangle$ of the *magnitude* of the magnetization, and not the average magnetization itself; we discuss our reasons for doing this in Section 3.7.1 when we examine the spontaneous magnetization of the Ising model.

The figure clearly shows the benefits of the Monte Carlo method. The calculation on the 5×5 system which we performed in Section 1.3 was exact, whereas the Monte Carlo calculation on the 100×100 system is not. However, the Monte Carlo calculation still gives a better estimate of the magnetization of the infinite system. The errors due to statistical fluctuations in our measurements of the magnetization are much smaller than the inaccuracy of working with a tiny 5×5 system.

Using the energy measurements from our simulation, we have also calculated the specific heat for our Ising model from Equation (3.15). To calculate

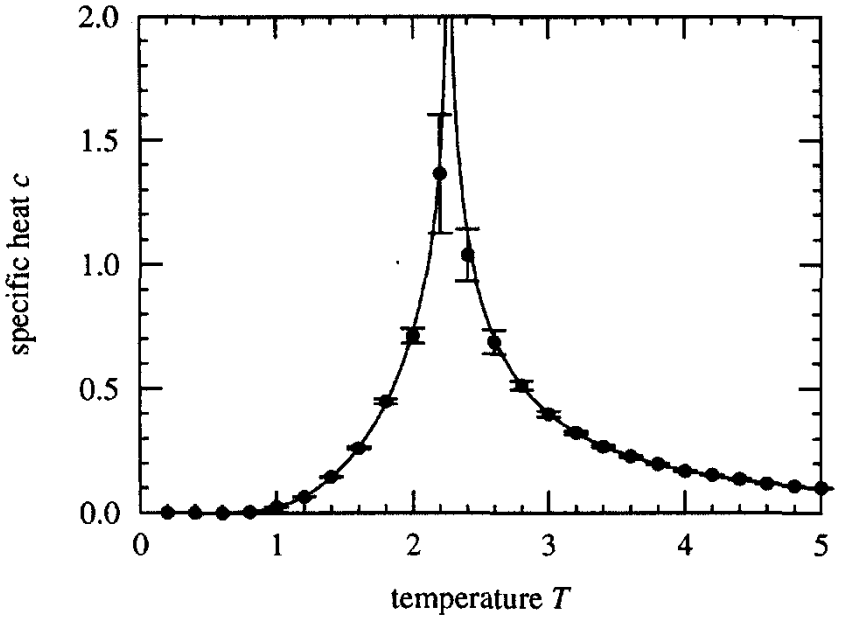

FIGURE 3.10 The specific heat per spin of the two-dimensional Ising model calculated by Monte Carlo simulation (points with error bars) and the exact solution for the same quantity (solid line). Note how the error bars get bigger close to the peak in the specific heat. This phenomenon is discussed in detail in the next section.

the errors on the resulting numbers we could use the blocking method of Section 3.4.2 to get a rough estimate. Here we are interested in doing a more accurate calculation, and for that we should use either the bootstrap or the jackknife method (see Sections 3.4.3 and 3.4.4). The number of independent samples n is for most temperatures considerably greater than 100, so by the criterion given in Section 3.4.4, the bootstrap method is the more efficient one to use. In Figure 3.10 we show our results for the specific heat with error bars calculated from 200 bootstrap resamplings of the data (giving errors accurate to about 5%). The agreement here with the known exact result for the specific heat is excellent—better than for the magnetization—though the errors are larger, especially in the region close to the peak. If we were particularly interested to know the value of c in this region it would make sense for us to go back and do a longer run of our program in this region to get more accurate data. For example, if we wanted to calculate the entropy difference from one side of the peak to the other using Equation (3.45), then large error bars in this region would make the integral inaccurate and we might well benefit from expending a little more effort in this region to get more accurate results.

3.7.1 The phase transition

It is not really the intention of this book to discuss the properties of the Ising model in detail, or the properties of any other model. However, there are a few things about the Ising model which we need to look into in a little more detail in order to understand the strengths and weaknesses of the Metropolis algorithm, and to explain why other algorithms may be better for certain calculations.

If we look at Figures 3.9 and 3.10, the most obvious feature that strikes us is the behaviour of the model around temperature $T = 2.2$ or so. The results shown in Figure 3.9 indicate that above this temperature the mean magnetization per site m is quite small, whereas below it the magnetization is definitely non-zero and for the most part quite close to its maximum possible value of 1. This seems like a sensible way for the model to behave, since we know (see Section 3.1.1) that the $T = \infty$ state is one in which all the spins are randomly oriented up or down so that the net magnetization will be zero on average, and we know that the $T = 0$ state is one in which all the spins line up with one another, either all up or all down, so that the magnetization per site is either $+1$ or -1. If we only had results for a small Ising system to go by, like the ones depicted in Figure 3.7, we might imagine that the true behaviour of the system was simply that the magnetization rose smoothly from zero in the $T \to \infty$ limit to 1 as the temperature tends to zero. However, our results for the larger 100×100 system indicate that the transition from small m to large m becomes sharper as we go to larger systems, and in fact we know in the case of this particular model, because we have an exact solution, that the change from one regime to the other is actually infinitely sharp in the thermodynamic limit. This kind of change is called a **phase transition**. The Ising model is known to have a phase transition in two or more dimensions. The two regimes we observe are called the **phases** of the model. The phase transition between them takes place at a temperature T_c, which we call the **critical temperature**, whose value in the particular case of the two-dimensional Ising model is known to be

$$T_c = \frac{2J}{\log(1+\sqrt{2})} \simeq 2.269J. \tag{3.53}$$

Above this temperature the system is in the **paramagnetic phase**, in which the average magnetization is zero. Below it, the system is in the **ferromagnetic phase** and develops a **spontaneous magnetization** (i.e., most of the spins align one way or the other and the magnetization becomes non-zero all of its own accord without the application of a magnetic field to the model). This spontaneous magnetization rises from zero at the phase transition to unity at absolute zero. The magnetization is referred to as the **order parameter** of the Ising model because of this behaviour. In general, an order parameter is any quantity which is zero on one side of a phase transition and

non-zero on the other. A phase transition in which the order parameter is continuous at T_c, as it is here, is called a **continuous phase transition**.

In fact, to be strictly correct, the *mean* magnetization of the Ising model below the critical temperature is still zero, since the system is equally happy to have most of its spins pointing either down or up. Thus if we average over a long period of time we will find that the magnetization is close to $+1$ half the time and close to -1 for the other half, with occasional transitions between the two, so that the average is still close to zero. However, the average of the magnitude $|m|$ of the magnetization will be close to $+1$, whereas it will be close to zero above the phase transition. In Figure 3.9 we therefore actually plotted the average of $|m|$, and not m. This explains why the magnetization above the transition temperature is still slightly greater than zero. The average magnetization in this phase is definitely zero (give or take the statistical error) but the average of the *magnitude* of m is always greater than zero, since we are taking the average of a number which is never negative. Still, as we go to the thermodynamic limit we expect this quantity to tend to zero, so that the numerical result and the exact solution should agree.[16]

We can look in detail at what happens to the spins in our Ising system as we pass through the phase transition from high to low temperatures by examining pictures such as those in Figure 3.2. At high temperatures the spins are random and uncorrelated, but as the temperature is lowered the interactions between them encourage nearby spins to point in the same direction, giving rise to correlations in the system. Groups of adjacent spins which are correlated in this fashion and tend to point in the same direction are called **clusters**.[17] As we approach T_c, the typical size ξ of these clusters—also called the **correlation length**—diverges, so that when we are precisely at the transition, we may encounter arbitrarily large areas in which the spins are pointing mostly up or mostly down. Then, as we pass below the transition temperature, the system spontaneously chooses to have the majority of its spins in either the up or the down direction, and develops a non-zero magnetization in that direction. Which direction it chooses depends solely

[16] This also provides an explanation of why the agreement between the analytic solution and the Monte Carlo calculation was better for the specific heat, Figure 3.10, than it was for the magnetization. The process of taking the mean of the magnitude $|m|$ of the magnetization means that we consistently overestimate the magnetization above the critical temperature, and in fact this problem extends to temperatures a little below T_c as well (see Figure 3.9). No such adjustments are necessary when calculating the specific heat, and as a result our simulation agrees much better with the known values for c, even though the error bars are larger in this case.

[17] A number of different mathematical definitions of a cluster are possible. Some of them require that all spins in the cluster point in the same direction whilst others are less strict. We discuss these definitions in detail in the next chapter, particularly in Sections 4.2 and 4.4.2.

on the random details of the thermal fluctuations it was going through as we passed the critical temperature, and so is itself completely random. As the temperature drops further towards $T = 0$, more and more of the spins line up in the same direction, and eventually as $T \to 0$ we get $|m| = 1$.

The study of phase transitions is an entire subject in itself and we refer the interested reader to other sources for more details of this interesting field. For our purposes the brief summary given above will be enough.

3.7.2 Critical fluctuations and critical slowing down

We are interested in the behaviour of the Ising model in the region close to T_c. This region is called the **critical region**, and the processes typical of the critical region are called **critical phenomena**. As we mentioned, the system tends to form into large clusters of predominantly up- or down-pointing spins as we approach the critical temperature from above. These clusters contribute significantly to both the magnetization and the energy of the system, so that, as they flip from one orientation to another, they produce large fluctuations in m and E, often called **critical fluctuations**. As the typical size ξ of the clusters diverges as $T \to T_c$, the size of the fluctuations does too. And since fluctuations in m and E are related to the magnetic susceptibility and the specific heat through Equations (3.15) and (3.16), we expect to get divergences in these quantities at T_c also. This is what we see in Figure 3.10. These divergences are some of the most interesting of critical phenomena, and a lot of effort, particularly using Monte Carlo methods, has been devoted to investigating their exact nature. Many Monte Carlo studies of many different models have focused exclusively on the critical region to the exclusion of all else. Unfortunately, it is in precisely this region that our Metropolis algorithm is least accurate.

There are two reasons for this. The first has to do with the critical fluctuations. The statistical errors in the measured values of quantities like the magnetization and the internal energy are proportional to the size of these critical fluctuations (see Section 3.4) and so grow as we approach T_c. In a finite-sized system like the ones we study in our simulations, the size of the fluctuations never actually diverges—that can only happen in the thermodynamic limit—but they can become very large, and this makes for large statistical errors in the measured quantities.

What can we do about this? Well, recall that the error on, for example, the magnetization m indeed increases with the size of the magnetization fluctuations, but it also decreases with the number of independent measurements of m that we make during our simulation (see Equation (3.37)). Thus, in order to reduce the error bars on measurements close to T_c, we need to run our program for longer, so that we get a larger number of measurements. This however, is where the other problem with the Metropolis algorithm

comes in. As we saw in Figure 3.8, the correlation time τ of the simulation is also large in the region around T_c. In fact, like the susceptibility and the specific heat, the correlation time actually diverges at T_c in the thermodynamic limit. For the finite-sized systems of our Monte Carlo simulations τ does not diverge, but it can still become very large in the critical region, and a large correlation time means that the number of independent measurements we can extract from a simulation of a certain length is small (see Equation (3.19)). This effect on its own would increase the size of the errors on measurements from our simulation, even without the large critical fluctuations. The combination of both effects is particularly unfortunate, because it means that in order to increase the number of independent measurements we make during our simulation, we have to perform a much longer run of the program; the computer time necessary to reduce the error bars to a size comparable with those away from T_c increases very rapidly as we approach the phase transition.

The critical fluctuations which increase the size of our error bars are an innate physical feature of the Ising model. Any Monte Carlo algorithm which correctly samples the Boltzmann distribution will also give critical fluctuations. There is nothing we can do to change our algorithm which will reduce this source of error. However, the same is not true of the increase in correlation time. This effect, known as **critical slowing down**, is a property of the Monte Carlo algorithm we have used to perform the simulation, but not of the Ising model in general. Different algorithms can have different values of the correlation time at any given temperature, and the degree to which the correlation time grows as we approach T_c, if it grows at all, depends on the precise details of the algorithm. Therefore, if we are particularly interested in the behaviour of a model in the critical region, it may be possible to construct an algorithm which suffers less from critical slowing down than does the Metropolis algorithm, or even eliminates it completely, allowing us to achieve much greater accuracy for our measurements. In the next chapter we look at a number of other algorithms which do just this and which allow us to study the critical region of the Ising model more accurately.

Problems

3.1 Derive the appropriate generalization of Equation (3.10) for a simulation of an Ising model with non-zero external magnetic field B.

3.2 Suppose we have a set of n measurements $x_1 \ldots x_n$ of a real quantity x. Find an approximate expression for the error on our best estimate of the mean of their squares. Take the numbers below and estimate this error.

20.27	19.61	20.06	20.73
20.09	20.68	19.37	20.40
19.95	20.55	19.64	19.94

Now estimate the same quantity for the same set of numbers using the jackknife method of Section 3.4.4.

3.3 In this chapter we described methods for calculating a variety of quantities from Monte Carlo data, including internal energies, specific heats and entropies. Suggest a way in which we might measure the partition function using data from a Monte Carlo simulation.

3.4 Write a computer program to carry out a Metropolis Monte Carlo simulation of the one-dimensional Ising model in zero field. Use it to calculate the internal energy of the model for a variety of temperatures and check the results against the analytic solution from Problem 1.4.