

Phase transitions

(Entry for Rouse Ball essay prize; title approved by Prof. Leader)

Tim D. Austin

June 26, 2008

Contents

1	Introduction and motivation: physical phase transitions	2
2	Mathematical abstraction: macroscopic order from microscopic disorder	4
3	The Thermodynamic Formalism and lattice gases	7
3.1	The configuration space	7
3.2	Gibbs states	8
3.3	An alternative approach to Gibbs states	10
3.4	Homogeneous states: translation invariance and the free energy	11
4	Phase transitions for homogeneous lattice gases	13
4.1	High inverse temperatures and ground states	13
4.2	Splitting the degeneracy	15
4.3	Examples	17
4.3.1	The Ising ferromagnet	17
4.3.2	The Potts model	19
4.3.3	The Heisenberg ferromagnet	20
4.3.4	The hard core repulsive gas on a general graph	21
4.3.5	One dimensional lattice gases	21
5	Bond percolation	22
6	The Ising ferromagnet and bond percolation in two dimensions, and Peierls' argument	23
7	Other important ideas	28
7.1	Renormalization	29
7.2	Universality	30
8	Epilogue: phase transitions in combinatorics	30

Prerequisites

This essay covers a number of different mathematical topics, none of them in very great depth. I will, however, have to assume basic notions of graph theory and some more sophisticated probability, measure theory and functional analysis. In particular, I will assume a basic working knowledge of topological measure theory (including the Riesz Representation Theorem, Radon measures and the vague topology on the space of Radon probabilities on a Polish space) and also ergodic theory

and Choquet's theory of barycentric decompositions. Much of the necessary technology is developed in the more comprehensive standard texts on the mathematical foundations of statistical mechanics (Chapter I of Simon [23] is a particularly useful introduction); alternatively, the more general background ideas can be found in any advanced probability text such as Kallenberg [15], or in comprehensive detail and generality in Volume 4 of Fremlin [9].

1 Introduction and motivation: physical phase transitions

Among the laws of physics, certain hold a distinguished position denoted by the term 'fundamental'. These are the laws from which we feel all others that we have yet identified should – in principal – follow; colloquially, these are the laws that actually *govern* the universe, as opposed to merely describing a few things that we see in it. As we have developed ever more intricate models of physical reality, those laws that we identify as fundamental have changed: in the seventeenth century Newton's laws of motion were regarded as such; now the standard model in quantum field theory holds that place, albeit with the physicists' full expectation that it will, in time, be superseded.

Why, then, do we need other laws of physics? Simply because we are not clever enough to deduce everything that we need from the fundamental laws. Newton's laws of motion describe the behaviour of a single particle under the action of a force; but what of the behaviour of all the particles in a glass of water, or a river, or a star? The natural world is complicated beyond our capacity to imagine. Although its behaviour is governed by the fundamental laws (or, more correctly, the fundamental laws should provide an adequate explanation for its behaviour within the model of reality in question), those laws give rise to a vast range of intricate behaviour when applied to systems of such size and complexity. In order to understand this behaviour, we have at our disposal a large collection of supplementary laws that apply specifically to phenomena on this scale. These laws are not fundamental in the presence of Newton's laws of motion or the laws of quantum physics: had we sufficient computing power to apply those other laws to all systems of real-world complexity we would not need the supplementary laws. However, that is a hopeless proposition, hence our need for empirical laws that are specific to behaviour at this scale. These include the laws of thermodynamics, of fluid mechanics, and of some areas of chemistry (and, arguably, almost all of biology and the Earth sciences).

One particularly compelling class of phenomena that emerge only when dealing with large, complicated systems is that of phase transitions: a physical system whose state can be controlled through the slow, continuous variation of a small number of external parameters undergoes a sudden qualitative change in behaviour at certain critical values of those parameters.

We meet examples of phase transitions every day of our lives:

- melting ice in a soft drink;
- boiling water for a cup of coffee;
- the formation of permanent magnets (which occurs in certain metals only below a certain critical temperature, the Curie point);
- the transformation upon cooling of liquid silicates into glass – a solid without a crystalline structure;
- the transition to superconducting behaviour of certain conductors below a critical temperature;
- the formation of dew on grass below the dew point;
- the changing of the liquid crystal display on your digital watch.

Most of these transitions are understood only phenomenologically. They result from a wide range of different mechanisms, and it is not obvious to what extent they can be described in a unified way, although some features are common to many. Often they can be associated with a discontinuous change in the value of some thermodynamic variable, such as temperature, pressure, magnetic or electric field. Often we have a qualitative description of the changes that take place in a substance involve first changes on the microscopic scale that then grow to the macroscopic; this is called ‘nucleation’ (for example, small drops of liquid appear in the gas phase before it fully condenses into liquid form).

Another feature common to many phase transitions is ‘spontaneous symmetry breaking’. Suppose we are considering a physical system that is, we think, described accurately by some mathematical model with the property that all of the ingredients in the model are invariant under the action of some group X . It may nevertheless occur that in practice we observe a loss of X -symmetry in certain phases (and so a phase transition from one phase to another will be marked by the spontaneous breaking of the X -symmetry as we cross a critical value of some control variable). The classic physical example of this behaviour is spontaneous magnetization: in some cases a block of metal, symmetric at least under certain reflections and with all ambient fields set as close as possible to zero, will, when cooled below a certain critical temperature, gain a magnetization in a certain particular direction and so lose its symmetry.

To explain this apparently paradoxical behaviour is an interesting challenge; we will see one detailed mathematical model, the Heisenberg model, that mimics it in Subsection 4.3. The qualitative account runs as follows. A magnet can be thought of as comprising a large lattice of small components, each with a magnetic dipole (‘spin’). If the directions of these spins are jumbled and ‘random’, then the net field will be nearly zero; if, on the other hand, they are largely aligned in the same direction, then they will yield a net field in that direction.

Crucially, more energy is needed for two close-by spins to be pointing in very different directions than for them to be aligned. At high temperatures, above the critical point, the system will have enough energy for much non-alignment, and so, since overall there are many more ways in which the system can be jumbled than in which it can be largely aligned, it is much more likely to be jumbled. However, at low temperatures, the magnet does not have enough energy for the spins not to be mostly aligned.

Of course, we were not able to set the ambient magnetic field strictly to zero; so as the block of metal is cooled below a certain critical temperature (the Curie point) the slight ambient field is enough to force it into a more stable, lower energy state with all spins aligned in similar directions. This gives rise to a large spontaneous magnetization of the block, even if the original ambient field was tiny.

Although such behaviour as symmetry breaking occurs in many phase transitions, it seems that no detailed (and, in particular, quantitative) description of how phase transitions arise applies to many different examples at once. The central problem in the study of phase transitions, then, is to build as representative as possible a mathematical model of a particular material or other physical system, and then analyze its behaviour in order to understand how the macroscopic properties of that system result from its microscopic structure, and why these properties then respond to external changes as they do.

It is in the analysis of such models that we begin to encounter hard mathematics, in addition to the physics we had to do to build the model. This essay is mostly about that mathematics. In fact, very little is known about models of most physical systems; we will spend most of our time on one particular family of models known as ‘lattice gases’, for which substantial results are available. These models describe a few physical situations, but leave most untouched.

It should be made clear that in real-world systems, the system does not actually change discontinuously as the control parameters are varied. Only in the ideal limit of a system comprising infinitely many units is there any chance of a strict phase transition; large but finite systems can only approximate to the behaviour of this ‘thermodynamic limit’. However, many real systems are large enough (a gas in a closed box might contain $\sim 10^{27}$ particles) that the change in the system with a continuous variation of the control parameters can be so rapid as to appear instantaneous.

Thus, a mathematical explanation of phase transitions works with a strictly infinite mathematical system, on the understanding that those real systems under study are sufficiently large that their behaviour may be identified with that of this infinite limit.

This essay makes no attempt to be comprehensive, self-contained or to describe the historical development of the subject. Many important results have been left out, as have many proofs (although appropriate references are given). It is hoped, however, to give the flavour of the mathematical side of the subject, and to convince the reader that here, at least, a fruitful interaction between pure mathematicians and physicists is very much alive.

2 Mathematical abstraction: macroscopic order from microscopic disorder

The work of this essay starts properly in the next section. Here we daydream a little: below is presented a very abstract discussion of the kind of mathematical phenomena that we will be considering, at least some of the time. This section is not strictly necessary later in the essay, and so can be skipped by those who find very soft measure theory and functional analysis distasteful. However, it may be that starting with a brief discussion at this level of abstraction both motivates and helps in understanding what follows. References will occasionally be made back to this section for ideas or notation.

It should perhaps be noted that there is no generally accepted definition of a ‘phase transition’ in mathematics; as with the word ‘fractal’, it seems to be better left a little vague, partly because no one definition seems adequate and partly because we do not know what other examples will arise in the future that we will want to refer to as phase transitions.

And so this essay does not seek to give a unified definition of ‘phase transition’. However, as we progress and meet more examples we will formulate three different notions of phase transition and pull them out of the main body of the text (one below and another in Section 4). These three are not exhaustive (we shall hint at a fourth, very different, notion of phase transition in the Epilogue), but it is hoped that presenting them thus will help the reader to see the similarities between them, and also to see why no one definition would be adequate.

What might be the ingredients of a description of a phase transition in pure mathematics? Keeping in mind the physical motivation of the last section, we consider the following:

1. We start with a measurable space (Ω, Σ) . This will have a familiar interpretation from probability: it is the space of configurations that some system can be in, together with those properties of the system that we can check (or, more precisely, those subsets of the configuration space that are precisely defined by such a property). Very often, as elsewhere in probability theory, Ω will also have some topological structure; typically it will be at least Polish, and often compact metric. Our Ω will be compact metric for the rest of this essay; the technical difficulties surrounding generalizations to other types of topological space are of no interest to us now. In addition Σ will always be its Borel σ -algebra.
2. In addition, we suppose that we have a distinguished σ -subalgebra $T \subseteq \Sigma$. The idea here is that our configuration space Ω is very complicated, and so has both a ‘large’ σ -algebra of subsets Σ , sufficient for a very detailed (‘microscopic’) description of the configuration of the system, and also a smaller σ -algebra T corresponding to a much coarser and less complete (‘macroscopic’) description of the configuration: the description that we are able to make in practice with only crude equipment to measure macroscopic properties of the configuration. We are now free to interpret T -measurable \mathbb{R} -valued functions as ‘observables’: macroscopic quantities depending on the configuration that we are actually able to measure.
3. What is a state of our system? As usual, in the absence of precise knowledge of the configuration of the system, its ‘state’ will be a probability measure μ on (Ω, Σ) (important: not just on (Ω, T)). Note that in examples arising from statistical physics this probability

results, not from our simple ignorance of the configuration of the system, but because the configuration may be changing continually with time rather than static. In this case μ should be an invariant ergodic probability for the associated time-indexed dynamical system, and so $\mu(A)$ has the usual interpretation of the ‘long-run proportion of time spent in A ’. In fact in the practice of statistical physics we obtain μ in a different way (to be discussed in Section 3), and this ergodicity for μ has been proved in almost no cases, and certainly none of physical interest. The question of why certain probabilities arise in statistical mechanics remains one of the great open problems of the area. It is, however, only peripherally related to phase transitions once we have our μ , and we shall rarely mention it again.

We will assume that our measures μ are Radon on the compact metric space Ω .

4. Typically, our system will not have only one possible state. Just as we have certain macroscopic properties of the system that we can measure, corresponding to T -measurable functions, we will suppose that we also have some macroscopic control over the system, in the form of a ‘control parameter’ on which the measure μ depends; that is, we actually have a measure-valued function $y \mapsto \mu_y$. In addition to Ω being compact metric we will suppose that the parameter y lies in some suitable connected Polish space Y , that all the μ_y are Radon, and that the measure-valued function $y \mapsto \mu_y$ is continuous with respect to the vague topology on $\text{Pr}(\Omega)$. Very often Y will be an interval of \mathbb{R} or a connected open subset of Euclidean space.
5. Now we come to a crucial point: we want our system, as described so far, to appear deterministic on the macroscopic scale. What does this mean? Unless the measures μ_y are actually concentrated on single points, then most measurable – even continuous – functions f on Ω will not be almost surely constant: their laws as random variables will not be concentrated on single points in \mathbb{R} . This is, after all, what it means for our system to display ‘randomness’. However, we could ask that all functions that are actually measurable with respect to T be almost surely constant for every μ_y : that is, for any $f \in L^\infty(T)$ and for any given $y \in Y$ we ask that f be μ_y -almost everywhere equal to a fixed value (which must then equal its μ_y expectation $\langle f, \mu_y \rangle = \int_\Omega f \, d\mu_y$).

Note that we do not expect these fixed values for a given f to be the same for different y . However, if they differ for, say, y_1 and y_2 , then μ_{y_1} and μ_{y_2} must be mutually singular, as they are concentrated on the disjoint sets

$$\{\omega \in \Omega : f(\omega) = \langle f, \mu_{y_1} \rangle\} \quad \text{and} \quad \{\omega \in \Omega : f(\omega) = \langle f, \mu_{y_2} \rangle\}$$

respectively.

Let us summarize our supposed position so far. We have a parameter y with which we can specify a probability μ_y on the configuration space (Ω, Σ) . In general this will not be a point measure, but if we consider only the ‘macroscopic observables’ – members of $L^\infty(T)$ – then they are all almost surely determined for each μ_y . Thus, for any $f \in L^\infty(T)$, we have a corresponding function $y \mapsto \langle f, \mu_y \rangle$ which sends y to the expectation of f , which must equal the value f takes μ_y -almost everywhere. (It is routine to show that this map $Y \rightarrow \mathbb{R}$ is Borel measurable, since the function $y \mapsto \langle f, \mu_y \rangle$ is measurable for *any* Σ -measurable f .)

The warning should be given now that we will *not* always find ourselves in this position in this essay. In particular, when we come to study homogeneous Gibbs states for a lattice gas, we will find that we cannot specify our desired measures μ_y uniquely, and that almost-sure determination of some macroscopic observables can fail precisely where we have more than one candidate μ_y ; these failures themselves are associated with certain phases of the model in question, and so a phase transition occurs where this behaviour begins. We will end up approaching phase transitions in lattice gases from a rather different direction from that described in this section, but not so different that we cannot relate the two in many important cases; see the discussion at the end of Subsection 4.2.

6. We are now ready to formulate our first notions of what a phase transition is. So far we have described a mathematical situation that corresponds to our idea of a system that is random at the microscopic scale but from which order emerges – in the form of almost determined random variables – at the macroscopic scale. In order to think about phase transitions we suppose that in addition we have some special collection \mathcal{A} of $L^\infty(\mathbb{T})$ of functions that correspond to observables which we feel should vary continuously as the state of the system is controlled continuously (such as the pressure or temperature of a gas, or the magnetization of a bar of metal in an ambient magnetic field). However, it may happen that at certain critical parameter values y_c , the function $y \mapsto \langle f, \mu_y \rangle$ for some $f \in \mathcal{A}$ is *not* continuous at y_c .

In fact, we very often face the following situation: the parameter space Y actually decomposes into the disjoint union of a collection of connected open subsets $(U_\alpha)_{\alpha \in S}$ together with some nowhere dense set F such that $y \mapsto \langle f, \mu_y \rangle$ is continuous on each U_α for all $f \in \mathcal{A}$ but jumps at points of F . This gives us:

Phase transition: notion 1A In the above situations we refer to the different open sets U_α as **phases** and say that a **phase transition** occurs across F , where our maps $\langle f, \mu_y \rangle$ can be discontinuous. The decomposition of Y into the U_α is called a **phase diagram** for our system.

An alternative interpretation of phase transitions might be that we have a special collection $\mathcal{F} \subseteq \mathbb{T}$ of events that are not only macroscopic but that we consider actually delimit some major structural change to the configuration of our system. In this case our parameter space may decompose into connected open sets U_α and a meagre set F as above, but this time such that for any point $y_0 \in F$ there is some $A \in \mathcal{F}$ for which $\mu_y(A)$ jumps discontinuously between 0 and 1 in the neighbourhood of y_0 .

Phase transition: notion 1B In the above situation we refer to the different open sets U_α as **phases** and say that a **phase transition** occurs across F , where the structure of our configuration as described by the sets $A \in \mathcal{F}$ changes in that the values $\mu_y(A)$ jump between 0 and 1. The decomposition of Y into the U_α is again called a **phase diagram** for our system.

We will make explicit two more notions of phase transitions when we look at lattice gases later in the essay, but the above mental image gets us started and may be helpful to give a feel for the phenomenon.

The above description is far too vague to take us much further: we have assumed that \mathbb{T} , $y \mapsto \mu_y$ and \mathcal{A} or \mathcal{F} have all come from somewhere, but there are no a priori reasons for such objects to play the roles described above. The real mathematics will begin when we examine particular cases in which the above behaviour – or some other form of critical phenomenon – does arise and try to understand why.

By far the most important – and heavily studied – mathematical phase transitions occur in the precise mathematical models used in statistical physics, and in particular in the so-called lattice gases, about which the most is known. These models for physics are also the historical motivations for studying phase transitions mathematically, arising as they do from the rich and important physics outlined in the Introduction above. We will study lattice gases in Sections 3 and 4, and they will reappear in Section 6.

However, this essay is not about statistical physics, nor is it about only those mathematical phase transitions that arise straight from it. We will also describe another form of phase transition, bond percolation, in Section 5, and Section 6 will then be devoted to one of the links between it and lattice gases.

Section 7 mentions some important ideas that relate to the earlier topics of the essay but do not actually appear among them, and in the Epilogue we glimpse another, rather different, notion of phase transition in combinatorics.

3 The Thermodynamic Formalism and lattice gases

3.1 The configuration space

In order to go beyond the purely phenomenological account of large physical systems given by thermodynamics, we need to build suitable mathematical models of the systems in question and develop a rigorous mathematical formalism for studying them. Naturally this is very difficult, and to date such models exist only for a few of the simplest phase transitions found in nature – in particular, those described by the ‘lattice gas’ models. However, these themselves have given rise to much beautiful mathematics. As Simon puts it in the first sentence of his book [23]:

Lattice models are caricatures invented to illuminate various aspects of elementary statistical mechanics, especially the phenomena of phase transitions and spontaneously broken symmetry.

This subsection is taken heavily from Simon’s book, which is a very comprehensive and clear exposition.

The original formalism we will use appeared first in Ruelle’s work in the 1960’s; the canonical reference for the resulting *Thermodynamic Formalism* is still his book [21]. Our mental picture is of the most obvious mathematical interpretation of a system with ‘infinitely many degrees of freedom’: we consider a countably infinite set L , and a non-empty set K of ‘values’, and our state space is then $\Omega = K^L$ (or, more generally, a suitable subspace of this).

Typically, each K will be a compact metric space and our states will be Radon probabilities on Ω with its product topology.

It is worth giving a warning right away about levels of generality. It is possible to develop the thermodynamic formalism with a different set K_x for each $x \in L$, and in [21] Ruelle does just that; but there seems to be no additional interesting mathematics that results from this, and more modern treatments start with a fixed K . K is always taken to be compact metric; in fact, almost all detailed results involve either a finite set or a sphere S^{d-1} or other well-known compact manifold. In most cases Ω will be the whole product K^L (certainly in all of the more classical work models), but recently a few special cases in which Ω is a natural subspace of K^L are considered; in fact, this distinction is not vital, for these latter cases can be reformulated in terms of measures on the whole of K^L that automatically (as a result of the way the model is built) are concentrated on the desired subspace. Finally, in almost all of the classical models, L is \mathbb{Z}^n , or at least some lattice in Euclidean space, but more recently results have appeared for L the vertex set of an arbitrary infinite graph (although often these results still trivialize unless the graph is particularly ‘nice’, such as the natural nearest-neighbour lattice graph on \mathbb{Z}^n). A model to be studied in this more general case, the hard core repulsive model, will be introduced in Subsection 4.3.

The earlier literature on the thermodynamic formalism (particularly [21] and [24]) sometimes seems very schizophrenic about appropriate levels of generality. The impression given is that the theory was developed in whatever additional generality was easy, and otherwise was left alone. In more modern treatments such as Simon [23], the extra easy-but-not-interesting generality seems to have been dropped; this essay largely tries to follow that trend.

Traditionally, two special cases with $L = \mathbb{Z}^n$ have been subjected to the most study: ‘lattice gases’ (in the old sense of the term – see below), in which $K = \{0, 1\}$, and ‘spin systems’, in which $K = \{-1, 1\}$ or, more generally, S^{d-1} . These two cases reflect the curious fact that the basic thermodynamic formalism can yield models of physical systems in more than one way:

- In the traditional lattice gas, each $x \in L$ is thought of as a point in space at which there either is a particle (corresponding to 1) or there is not (0). The state of the system (which has a natural interpretation as a member of $\mathcal{P}L$) then describes the locations of a (finite or infinite) collection of particles in space, constrained to occupy positions on the lattice. It is worth noting that these particles are **indistinguishable**: we care only what lattice sites they collectively occupy, not who they are as individuals.

- In a spin system, each $x \in L$ is thought of, not only as a point in space, but as the location of a particular particle in a crystal lattice in space. These particles are supposed to have a magnetic moment or ‘spin’, and the values in K tell us where the spin is pointing: in the simplest case $K = \{-1, 1\}$ the spins are either ‘up’ (+1) or ‘down’ (-1), and more generally for $K = S^{d-1}$ they point in some direction (specified by a unit vector) in d -dimensional Euclidean space.

Recently, the term ‘lattice gas’ has been broadened to include all models with configuration space K^L and with the Gibbs states to be described below; henceforth we will keep with this modern usage and refer to our models as lattice gases.

3.2 Gibbs states

Having decided to work with the configuration space K^L , we need to decide how to describe the state of our system; this requires us to select suitable Radon probabilities on K^L .

The probabilities we will consider are the Gibbs states. These arise in statistical physics as the natural probabilities describing the state of a statistical mechanical system that is in equilibrium with its surroundings; they are specified by a recipe depending on the energy of the different states that the system can be in and an additional parameter from thermodynamics: the temperature.

We will first describe the construction of Gibbs states in a simpler case than the lattice gas. Suppose that the energy of a configuration of a system is given by a measurable function $H : \Omega \rightarrow \mathbb{R}$ (the **Hamiltonian**), and that the temperature is $T > 0$ (in fact we will work instead with the inverse temperature, $\beta = 1/T$, and will also allow the limiting cases $T = \infty, \beta = 0$). Suppose also that we are given some fixed finite Radon measure ν on Ω with respect to which H is integrable – often ν will have arisen naturally in our choice of configuration space, for example as a Haar measure. The **Gibbs state relative to** ν , say μ , is given by normalizing the indefinite-integral measure $\nu \llcorner e^{-\beta H}$ to be a probability:

$$\mu(A) = \frac{1}{Z_H(\beta)} \int_A e^{-\beta H(\omega)} \nu(d\omega),$$

where the normalizing constant $Z_H(\beta) = \int_{\Omega} e^{-\beta H} d\nu$ is called the **partition function**. Henceforth we refer to these Gibbs states as ‘simple’ Gibbs states, to distinguish them from their lattice gas counterparts.

Why should these simple Gibbs states describe the ‘average’ behaviour of the system (which will actually evolve following the deterministic dynamics given by the laws of motion)? This is the difficult question mentioned in Step 3 of Section 2, to which no truly satisfactory answer is known. Nevertheless, the Gibbs distributions have been hugely successful in describing statistical mechanical systems, and there is no doubt that they are the appropriate choice in many cases.

Reassured, we will now define a Gibbs state on a lattice gas. This is harder; typically, in the thermodynamic limit of a system with infinitely many degrees of freedom, there will not be a suitable finite energy function. Instead we need the notion of **interaction**.

Definition 3.1 An *interaction* is a family of real-valued functions $\Phi = (\Phi_{\Lambda})_{\Lambda \in [L]^{<\omega}}$, with Φ_{Λ} a continuous function on the finite product K^{Λ} for each finite $\Lambda \subset L$, and such that for each $x \in L$

$$\sum_{\Lambda \in [L]^{<\omega}, \Lambda \ni x} \|\Phi_{\Lambda}\| < \infty,$$

where $\|\Phi_{\Lambda}\|$ is the sup-norm of Φ_{Λ} as a function on K^{Λ} .

The idea is that the interaction ‘breaks down’ the total energy of a configuration $\omega = (\omega_x)_{x \in L}$ into a sum of finite contributions from the different finite subsystems of the overall system: the contribution from $\omega|_{\Lambda} = (\omega_x)_{x \in \Lambda}$ for finite $\Lambda \subset L$ is precisely $\Phi_{\Lambda}(\omega|_{\Lambda})$. In fact the total energy

may be infinite or otherwise undefined, but we can suppose that these finite contributions still make sense in the form of the interaction.

Sometimes interactions are defined by the weaker condition

$$\sum_{\Lambda \in [L]^{<\omega}, \Lambda \ni x} \frac{1}{|\Lambda|} \|\Phi_\Lambda\| < \infty,$$

such as in [21]. In this case some of the subsequent results change or require different proofs; here we will restrict ourselves to the stricter class.

Given Φ , we define for each finite $\Lambda \subset L$ the **internal energy function** $U_\Lambda : K^\Lambda \rightarrow \mathbb{R}$ by

$$U_\Lambda(\eta) = \sum_{S \subseteq \Lambda} \Phi_S(\eta|_S)$$

and the **interaction energy function** $W_\Lambda : K^\Lambda \times K^{L \setminus \Lambda} \rightarrow \mathbb{R}$ by

$$W_\Lambda(\eta, \xi) = \sum_{S \in [L]^{<\omega}, S \cap \Lambda \neq \emptyset, S \setminus \Lambda \neq \emptyset} \Phi_S(\eta|_{S \cap \Lambda} \hat{\cap} \xi|_{S \setminus \Lambda}),$$

where we write $\eta|_{S \cap \Lambda} \hat{\cap} \xi|_{S \setminus \Lambda}$ for the member of K^S with terms equal to those of ω for coordinates in $S \cap \Lambda$ and those of ξ for coordinates in $S \setminus \Lambda$. The interpretation here is that $U_\Lambda(\eta)$ is the internal energy of only those lattice sites lying in Λ with values described by η , and $W_\Lambda(\eta, \xi)$ is the energy of interaction between them and the ‘boundary condition’ ξ : the configuration of the rest of the system, described by a member of $K^{L \setminus \Lambda}$ (hence the sum over sets S hitting both Λ and its complement). Now the **total energy of $\eta \in K^\Lambda$ given boundary condition $\xi \in K^{L \setminus \Lambda}$** is $U_\Lambda(\eta) + W_\Lambda(\eta, \xi)$; this is finite by our definition of interaction.

Suppose we are also given a value for the inverse temperature β , and a fixed finite Radon measure ν on K . We now specify the Gibbs states on our system by the following consideration of their marginals. For any subset $S \subset L$ write $\pi_S : K^L \rightarrow K^S$ for the projection map. Now given $\Lambda \subseteq L$ finite and a measure $\sigma \in \Pr(\Omega)$, we can disintegrate $\sigma \circ (\pi_\Lambda)^{-1}$ over $\pi_{L \setminus \Lambda}$: for any $A \in \mathcal{B}(K^L)$ depending only on coordinates in Λ (hence an A that corresponds to a Borel subset of K^Λ) we have

$$\sigma(A) = \int_{K^{L \setminus \Lambda}} \mu_\xi(A) \tau(d\xi)$$

for the probability $\tau = \sigma \circ (\pi_{L \setminus \Lambda})^{-1} \in \Pr(K^{L \setminus \Lambda})$ and a disintegration (sometimes called a ‘regular conditional probability’) $\mu_\xi \in \Pr(K^\Lambda)$ indexed by $\xi \in K^{L \setminus \Lambda}$.

Definition 3.2 *The probability σ is a **Gibbs state relative to ν for Φ at inverse temperature β** if for every finite $\Lambda \subset L$ it disintegrates with*

$$\mu_\xi(A) = \frac{1}{Z_{\Lambda, \Phi, \xi}(\beta)} \int_A e^{-\beta(U_\Lambda(\eta) + W_\Lambda(\eta, \xi))} \nu^{\otimes \Lambda}(d\eta)$$

for $A \in \mathcal{B}(K^\Lambda)$, $\xi \in K^{L \setminus \Lambda}$, where the **partition function $Z_{\Lambda, \Phi, \xi}(\beta)$ with boundary condition ξ** is given by

$$\int_{K^\Lambda} e^{-\beta(U_\Lambda(\eta) + W_\Lambda(\eta, \xi))} \nu^{\otimes \Lambda}(d\eta)$$

Notice that in the above definition, working with Φ and β is equivalent to working with $\beta\Phi = (\beta\Phi_\Lambda)_{\Lambda \in [L]^{<\omega}}$ and 1.

The intuitive picture corresponding to the above definition is a natural extension of the earlier simple definition of Gibbs distribution for finite energy. Here a Gibbs state relative to ν is a probability on the space K^L of possible configurations such that for any finite subset Λ of degrees of freedom, the probability of a Borel subset A of K^Λ occurring is dependent on the boundary

conditions: the values taken by the degrees of freedom outside Λ . For any given boundary condition $\xi \in K^{L \setminus \Lambda}$ the probability distribution on K^Λ is just the simple Gibbs state on K^Λ relative to $\nu^{\otimes \Lambda}$ in the earlier sense with energy function $H = U_\Lambda + W_\Lambda(\cdot, \xi)$, the sum of the internal and interaction energies. Now a general Gibbs state on K^L is any state such that whenever we consider such a Λ , the resulting measure on K^Λ is a mixture of these simple Gibbs states with different boundary conditions.

The following observation is crucial: Gibbs states may not be unique! Given ν and $\beta\Phi$ there may be several Radon probabilities σ satisfying the above definition. Writing $K_{\beta\Phi}$ for the set of such probabilities (suppressing the dependence on ν), we actually have the following:

Theorem 3.3 *The space $K_{\beta\Phi}$ is a non-empty Choquet simplex in $\text{Pr}(\Omega)$. A measure σ is an extreme point of $K_{\beta\Phi}$ if and only if for any $A \in \Sigma$ depending only on coordinates in a finite $\Lambda \subset L$ and any $\varepsilon > 0$ there is a finite $M \supset \Lambda$ such that whenever $B \in \Sigma$ depends only on coordinates in $L \setminus M$ we have*

$$|\sigma(A \cap B) - \sigma(A)\sigma(B)| < \varepsilon.$$

These extreme points are mutually singular.

Proof Chapter 1 of Ruelle [21]. □

Extreme points of $K_{\beta\Phi}$ are referred to as **pure Gibbs states**.

3.3 An alternative approach to Gibbs states

The above definition of Gibbs states is well established as the best way in to the theory. However, we might have been tempted to try a more naive approach to Gibbs states for the lattice gas corresponding to a simpler interpretation of ‘taking the thermodynamic limit’.

In this approach we first consider the simple Gibbs states on the finite-dimensional subsystems K^Λ of Ω : for each finite $\Lambda \subset L$ let σ_Λ be the simple Gibbs state on K^Λ relative to $\nu^{\otimes \Lambda}$ arising from a suitably-chosen Hamiltonian H_Λ . There are two possible methods for choosing these Hamiltonians:

1. we can take $H_\Lambda = U_\Lambda$, the internal energy function alone, or;
2. we can add a **boundary term** by choosing one particular $\psi \in \Omega$ and then for each Λ letting $H_\Lambda = U_\Lambda + B_\Lambda$, where for $\eta \in K^\Lambda$ the additional term is

$$B_\Lambda(\eta) = W_\Lambda(\eta, \psi|_{L \setminus \Lambda}),$$

the interaction energy with a boundary condition agreeing with ψ outside Λ .

In either of the above cases, we obtain a family of simple Gibbs states σ_Λ on the finite-dimensional subsystems K^Λ . We now extend each of these arbitrarily to probabilities $\tilde{\sigma}_\Lambda$ on Ω (that is, $\tilde{\sigma}_\Lambda \circ (\pi_\Lambda)^{-1} = \sigma_\Lambda$) to obtain a net $(\tilde{\sigma}_\Lambda)_{\Lambda \in [L]^{<\omega}}$. Our general Gibbs state σ could now be taken to be a limit point of this net in the vague topology on $\text{Pr}(\Omega)$; these certainly exist since $\text{Pr}(\Omega)$ is compact. We will call such a σ a **thermodynamic limit** of the net. Note that such limits may not be unique. Note also that it does not matter how we extend each σ_Λ to $\tilde{\sigma}_\Lambda$: for any given finite $M \subset L$ the restriction of $\tilde{\sigma}_\Lambda$ to sets depending only on coordinates in M is uniquely specified by Λ for all finite $\Lambda \supset M$, and so by the definition of the vague topology, whether and how the net $(\tilde{\sigma}_\Lambda)_{\Lambda \in [L]^{<\omega}}$ converges to a given set of limits depends on only the measures σ_Λ . Henceforth we will be sloppy and refer to the convergence of these measures themselves.

We will call a limit point of the internal-energy-only net a **simple limit Gibbs state**, and a limit point of the net with boundary terms arising from $\psi \in \Omega$ a **limit Gibbs state with boundary condition ψ** .

In fact, in many cases of importance the situation is that the nets constructed above actually converge to one particular measure; we will have more to say about this at the end of Subsection 4.2.

How do such limits relate to the Gibbs states we defined previously? Happily they give us only what we already know:

Theorem 3.4 *The thermodynamic limits σ of the nets constructed above (with or without boundary terms) are Gibbs states. If we allow the use of boundary terms corresponding to different $\psi \in \Omega$ then the set K_Φ is the closed convex hull of the set of all thermodynamic limits thus obtained.*

Proof Theorem 1.9 in Ruelle [21], or Section III.2 in Simon [23]. □

3.4 Homogeneous states: translation invariance and the free energy

An important special case occurs when $L = \mathbb{Z}^n$ and we restrict ourselves to translation-invariant interactions (so $\Phi_S = \Phi_{S+x}$ for any finite $S \subset L$ and $x \in L$) and resulting translation-invariant Gibbs states. These are shown in Ruelle [21] to form a non-empty sub-simplex of $K_{\beta\Phi}$ (Chapter 3). These states are referred to as **homogeneous**. We write I_Φ for the sub-simplex of K_Φ containing the translation-invariant states; and note also the fact that the extreme points of I_Φ are precisely those that are ergodic with respect to translations (see again Corollary III.3.10 in Simon [23]).

We will concentrate only on the homogeneous case *for the rest of the essay*, and may henceforth sometimes suppress the word ‘homogeneous’. There are interesting things that can be said about the non-homogeneous case, but they do not relate to the phase transitions that we wish to study.

It turns out that in the homogeneous case, in a certain sense, ‘most’ interactions Φ have a unique Gibbs state: I_Φ is a singleton. To make this precise, write \mathcal{B}_1 for the space of all translation-invariant interactions with the norm

$$\|\Phi\| = \sum_{\Lambda \in [L]^{<\omega}, \Lambda \ni 0} \|\Phi_\Lambda\|;$$

we can check easily that \mathcal{B}_1 is a separable Banach space. In Simon’s notation ([23]) it is the ‘smaller Banach space of interactions’ (the ‘big’ space corresponds to the weaker definition of interaction mentioned in Subsection 3.2).

We now have the following:

Theorem 3.5 (Gibbs phase rule) *The homogeneous Gibbs state for Φ is unique (that is, the simplex I_Φ is a singleton) for residually many $\Phi \in \mathcal{B}_1$.*

Proof Theorem 3.7(c) of Ruelle [21] for the case of finite K , or Theorem III.3.5 in Simon [23]. □

The above result is, in fact, not proved directly for Gibbs states at all, but for ‘equilibrium states’, which are then shown to be equivalent to Gibbs states. In the remainder of this subsection we will give a whistlestop account of this alternative theory and its linchpin: the specific free energy of an interaction. In the next section we will return to this object briefly as an alternative way to approach phase transitions.

Recall the definition of the simple Gibbs state $\sigma_{H,\beta}$ arising from an energy function $H : \Omega \rightarrow \mathbb{R}$ and inverse temperature β relative to a given finite Radon measure ν :

$$\sigma_{H,\beta}(A) = \frac{1}{Z_H(\beta)} \int_A e^{-\beta H(\omega)} \nu(d\omega),$$

where $Z_H(\beta) = \int_\Omega e^{-\beta H(\omega)} \nu(d\omega)$. We define the **free energy of H at inverse temperature β** to be

$$f_H(\beta) = -\log Z_H(\beta),$$

and observe the relation, familiar in classical thermodynamics,

$$\langle H, \sigma_{H,\beta} \rangle = \frac{1}{Z_H(\beta)} \int_A H(\omega) e^{-\beta H(\omega)} \nu(d\omega) = \frac{df_H}{d\beta}(\beta)$$

(assuming H is not so wild that we cannot differentiate under the integral sign). Thus we can find the average energy at a given value of β if we know the form of the function $\beta \mapsto f_H(\beta)$ locally at β .

We will find this situation is repeated in the case of the lattice: many of the macroscopic quantities of interest to us will be expressible in terms of a suitably defined function of β , the specific free energy.

Clearly if we do not have a sensible finite energy function we cannot use the above simple definition. However, given a translation-invariant interaction Φ , for each finite $\Lambda \subset L$ we can define the **simple partition function on Λ**

$$Z_{\Phi, \Lambda} = \int_{K^\Lambda} e^{-U_\Lambda(\eta)} \nu^{\otimes \Lambda}(d\eta),$$

and now define $f_\Lambda(\Phi) = \frac{1}{|\Lambda|} \log Z_{\Phi, \Lambda}$. Writing $B_s(0)$ for the cube in \mathbb{Z}^n of side length $2s$ and centred at 0, we have the following result:

Theorem 3.6 *The ‘thermodynamic’ limit $f(\Phi) = \lim_{s \rightarrow \infty} f_{B_s(0)}(\Phi)$ exists.*

Proof Theorem 3.4 in Ruelle [21]. □

We call $f(\Phi)$ the **specific free energy of Φ** . Remarkably, it is possible to characterize the homogeneous Gibbs states in terms of it.

Consider f as a function on the Banach space \mathcal{B}_1 . Given $\Phi \in \mathcal{B}_1$, we can define the **specific energy function $g_\Phi : \Omega \rightarrow \mathbb{R}$** by

$$g_\Phi(\omega) = \sum_{\Lambda \in [L]^{<\omega}, \Lambda \ni 0} \Phi_\Lambda(\omega|_\Lambda);$$

that this sum converges to define a continuous function follows from the definition of \mathcal{B}_1 .

Now, given a Radon probability μ on ω , we define the linear functional α_μ on \mathcal{B}_1 by $\langle \Phi, \alpha_\mu \rangle = \langle g_\Phi, \mu \rangle$.

We are ready for the alternative characterization.

Theorem 3.7 *The function f is convex on \mathcal{B}_1 .*

Proof Theorem II.2.1 in Simon [23]. □

We will say that a linear functional α on \mathcal{B}_1 **supports f at Φ** if for all $\Psi \in \mathcal{B}_1$ we have

$$f(\Psi) - f(\Phi) \geq \langle \Psi, \alpha \rangle - \langle \Phi, \alpha \rangle.$$

Theorem 3.8 *For a given $\Phi \in \mathcal{B}_1$ the homogeneous Gibbs states for Φ are precisely those Radon probabilities μ on Ω such that α_μ supports f at Φ .*

Proof Theorem III.3.2 in Simon [23]. □

Ruelle [21] also proves versions of the above results, but formulated in a rather different way that relies on a definition of specific free energy (or, in his terminology, pressure) for any continuous function on Ω .

The quick sketch above does not give a true idea of how extensive are the results on specific free energy in lattice gases; Simon [23] spends Chapter II developing their theory before moving on to states at all. However, from the above we at least begin to see the picture: the Gibbs states of Φ are (or rather, are those that give rise to) the supporting functionals to the specific free energy at Φ . If the specific free energy is differentiable at Φ , then such a measure is unique and is called the **tangent to f at Φ** , so I_Φ is a singleton; in fact this is the usual approach to proving the Gibbs phase rule, Theorem 3.5. The Gibbs states fail to be unique only where f is not differentiable.

Now we come to an important note. In practice phase transitions are often studied by restricting f to a family of interactions differentially parameterized by a finite number of real variables and by then considering its differentiability as a function of those variables. The idea is that many of the phase transitions we are interested in can be picked up in this way, without considering all homogeneous interactions Φ . This should be understood as corresponding to a slightly different notion of phase transition from those we have already seen; perhaps we should make this explicit:

Phase transition: notion 2 In specific cases we often have a continuous parameterization of a family of interactions $y \mapsto \Phi_y$ for y in some connected open subset U of a Euclidean space, and then find that U decomposes into the disjoint union of a family $(U_\alpha)_{\alpha \in S}$ of connected open subsets and a nowhere dense set F (often actually a union of hypersurfaces bounding the U_α) such that $f(\Phi_y)$ is smooth as a function of y within each U_α , but at each point of F some derivative of $y \mapsto f(\Phi_y)$ is discontinuous. In this case we refer to the U_α as **phases**, and say that a **phase transition occurs across** F . The **order** of the transition at a point $y \in F$ is the order of the lowest derivative that is not continuous there.

This notion is justified by our previous observation that many of the macroscopic variables in which we are interested in a given problem can be simply expressed in terms of f and its derivatives; for example the specific energy, A_Φ , is just

$$\frac{\partial f(\beta\Phi)}{\partial \beta}(1),$$

if this derivative exists.

In fact, even more is true: in many cases f will be an analytic function of the control parameters except at a few points, which will then be identified as critical points for phase transitions; many works interpret phase transitions solely in this way and develop a theory accordingly. This theory has grown out of work by Yang and Lee in the 1950's; see [19] for their original proposition. Here we have chosen largely to sideline this approach, feeling that the theory to be developed in the next three sections gives a much richer and more beautiful mathematical picture. Furthermore, there has been relatively little progress on the mathematical study of phase transitions in this way since the 1950's, although recent years have seen a few interesting exceptions to this (including, very recently, the work of Scott and Sokal [22] on hard core repulsive lattice gases and their relations to combinatorics, to be described briefly in Subsection 4.3).

4 Phase transitions for homogeneous lattice gases

4.1 High inverse temperatures and ground states

We could take the space \mathcal{B}_1 of all interactions as our parameter space for studying phase transitions in lattice gases, but arguably this is unphysical (we would certainly never have so much control over the interactions arising in a real-world problem) and leads us further into the very intricate theory of the specific free energy function on \mathcal{B}_1 (see Chapters 3 and 4 of Ruelle [21] and Chapter III of Simon [23]). A much more natural problem is to study phase of the parameter $\beta \in [0, \infty)$ for a fixed interaction Φ , with the possibility of introducing further parameters as necessary.

We note at once that in spite of Theorem 3.5, in general for a given Φ the simplex $I_{\beta\Phi}$ will be non-trivial for many β ; typically for all sufficiently large β . Although the set of potentials Φ for which I_Φ is not a singleton is meagre in the whole of \mathcal{B}_1 , it can and does happen that our one-parameter family $(\beta\Phi)_{\beta \geq 0}$ spends a lot of time there.

On the other hand, under fairly general conditions it is true that I_Φ – in fact, the whole of K_Φ , not just the homogeneous states – is a singleton for all sufficiently small β . An exact formulation of some results of this type can be found in the papers by Dobrushin [5], [6], [7].

Let us describe heuristically what is going on as β varies.

At $\beta = 0$, a quick check of the definition of Gibbs state shows us that the interaction is irrelevant: we obtain simply a product of independent copies of the normalized background measure $\nu/\nu(K)$. The above results of Dobrushin give conditions under which this state is stable: for β small, we still have only one Gibbs state, and it is close to an independent product: the correlations between the values taken at different coordinates are very slight.

However, as β increases, we feel that any Gibbs measure for $\beta\Phi$ is favouring those configurations with low energy (consider where the integrand is small and where it is large in Definition 3.2).

As $\beta \rightarrow \infty$ we want to say that the Gibbs states become concentrated on those configurations of minimal energy, although we cannot yet make this precise, recalling that in general there is no sensible finite energy function on the whole of Ω . We admit the possibility that we may have to identify several such configurations, and so there may be several different Gibbs states for large β . If this happens there will be a natural critical value β_c , $0 < \beta_c < \infty$: the infimum of those β with several Gibbs states.

A detailed examination of various results for $\beta \rightarrow \infty$ is given by Sinai in [24]; the discussion below is based heavily on his. Here we will restrict ourselves to the special case of a finite value space K ; this case is treated in Chapter II of [24]. In Chapter III, Sinai goes on to study more general K that come with the action of a continuous group of symmetries G ; we will not treat this theory here, although will catch a glimpse of it when we look at the example of the Heisenberg model in Subsection 4.3.

Sinai's idea is that under some circumstances pure homogeneous Gibbs states in I_Φ can be described as being concentrated on the set of small local distortions of a fixed periodic configuration $\psi \in \Omega$, in the following sense.

Suppose that for each β (or at least for all $\beta \in [0, \infty)$ sufficiently large) we have a particular Gibbs state σ_β in $I_{\beta\Phi}$, and suppose $\psi \in \Omega$. Often in practice σ_β varies continuously with β , but we need not assume this. We say that σ_β **describes small local distortions of ψ** as $\beta \rightarrow \infty$ if for sufficiently large β the random set $\{x \in \mathbb{Z}^n : \omega_x \neq \psi_x\}$ decomposes almost surely into a countable union of finite connected subsets of (the nearest-neighbour graph on) \mathbb{Z}^n , and furthermore for any $\varepsilon > 0$ and natural number s there is some β_ε such that for $\beta > \beta_\varepsilon$ we have

$$\sup_{x \in \mathbb{Z}^n} \sigma_\beta \{ \omega : \omega|_{B_s(x)} \neq \psi|_{B_s(x)} \} < \varepsilon,$$

where $B_s(x)$ denotes the cube in \mathbb{Z}^n of side length $2s$ and centred at x . The second part of this condition is saying that as $\beta \rightarrow \infty$, the probability of ω differing *at all* from ψ on a cube of given side-length tends to zero uniformly over all cubes of that side length.

It turns out that it is possible to identify certain distinguished periodic $\psi \in \Omega$ to which with can associate Gibbs states σ_β describing small local distortions. Heuristically, these are those ψ that have minimal energy among those configurations that differ from them in only finitely many coordinates. Given $\omega, \psi \in \Omega$ differing in only finitely many coordinates, write

$$H(\omega, \psi) = \sum_{\Lambda \in [L]^{<\omega}} (\Phi(\omega|_\Lambda) - \Phi(\psi|_\Lambda));$$

note that this sum converges, since $\omega|_\Lambda$ and $\psi|_\Lambda$ differ only if Λ hits the finite set $\{x \in L : \omega_x \neq \psi_x\}$, and so we can appeal to the finiteness of $\|\Phi\|$.

Definition 4.1 A configuration ψ is **ground state** if $H(\omega, \psi) \geq 0$ for any ω differing from it in only finitely many coordinates. If $H(\omega, \psi) > 0$ for all such $\omega \neq \psi$ then ψ is said to be **isolated**.

This gives us a suitable precise sense in which ψ can have (locally) minimal energy, and so can be favoured by Gibbs states for large β , as mentioned in the informal description above. However, in the next subsection we will need to know the following alternative definition of periodic ground state:

Proposition 4.2 A periodic configuration ψ is a ground state for Φ if and only if it minimizes the **specific energy**

$$h_\Phi(\psi) = \lim_{s \rightarrow \infty} \frac{1}{|B_s(0)|} \sum_{x \in B_s(0)} \left(\sum_{\Lambda \in [L]^{<\omega}, \Lambda \ni x} \frac{1}{|\Lambda|} \Phi_\Lambda(\psi|_\Lambda) \right)$$

(note that this limit exists, since ψ is periodic).

Proof Lemma 2.1 in Sinai [24]. □

We can think of this as providing an alternative precise sense in which ψ is ‘energy minimizing’, which happily turns out equivalent to our first attempt.

We note that a given Φ can have more than one ground state, in which case we say the ground states are **degenerate**. The term ‘ground state’ is unfortunate, as we should like to reserve the word ‘state’ for probabilities on Ω ; but this usage is entrenched in the literature, and there is some sense in identifying these ground states with the point measures concentrated on them so perhaps this does not matter.

Sinai goes on to consider in detail the following situation, which is known to hold in a number of the more heavily-studied lattice gas models:

- First we suppose that we have an interaction $\Phi \in \mathcal{B}_1$ with the property that the set $g(\Psi)$ of *periodic* ground states is finite, say $\{\psi_1, \dots, \psi_r\}$, and also that each of these is isolated. Suppose furthermore that Ψ satisfies **Peierls’ stability condition**: for some constant $\rho > 0$ and some natural number $s \geq 1$, the ‘energy difference’ $H(\omega, \psi)$ for ω differing from $\psi \in g(\Phi)$ in only finitely many places satisfies

$$H(\omega, \psi) \geq \rho |\{x \in \mathbb{Z}^n : \text{dist}(x, \{x_1 \in \mathbb{Z}^n : \omega_{x_1} \neq \psi_{x_1}\}) \leq s\}|,$$

where dist refers to the distance on the nearest-neighbour graph on \mathbb{Z}^n . Thus Peierls’ condition says that the energy difference $H(\omega, \psi)$ grows at least linearly with the number of coordinates within a certain distance of the points at which ω and ψ differ.

- We suppose also that for each $\psi \in g(\Phi)$ the limit Gibbs states $\sigma_{\beta, \psi}$ with boundary condition ψ , defined as at the end of Subsection 3.3, exist and are such that the $\sigma_{\beta, \psi}$ describe small local distortions of ψ as $\beta \rightarrow \infty$.

If our system satisfies the above conditions then we shall say it exhibits **Sinai behaviour** (this is not standard terminology).

It is important to note the following: the Gibbs states $\sigma_{\beta, \psi}$ need not span the whole of $I_{\beta\Phi}$: there may be other pure Gibbs states not accessible to us via the above theory. Sinai’s study was of the behaviour of only those Gibbs states describing local distortions of ground states.

I do not know under what general conditions the above situation is known to obtain; we will prove it for the two-dimensional Ising model in Section 6. If Sinai behaviour is exhibited then it is possible to develop a detailed theory of the ground states for large β , and in particular how they can be distinguished by perturbing the interaction Φ by adding linear multiples of $r-1 = |g(\Phi)| - 1$ suitably-chosen other interactions. We will look at this in the next subsection.

4.2 Splitting the degeneracy

Suppose our lattice gas exhibits Sinai behaviour with a non-singleton set of ground states $g(\Phi) = \{\psi_1, \dots, \psi_r\}$. Our next idea is to see whether we can remove this degeneracy by perturbing the interaction Φ slightly.

Suppose that in addition to our basic interaction Φ we have some other interactions $\Psi_1, \dots, \Psi_{r-1}$, one fewer than there are ground states. For $\lambda = (\lambda_1, \lambda_2, \dots, \lambda_{r-1})$ we write $\Phi_\lambda = \Phi + \sum_{i \leq r-1} \lambda_i \Psi_i$, a **perturbation** of Φ .

Definition 4.3 *The perturbation Φ_λ completely splits the degeneracy of the ground states of Φ if for any nonempty subset $S \subseteq \{1, 2, \dots, r-1\}$ and any vector $b = (b_1, b_2, \dots, b_r)$ with $b_i > 0$ for $i \notin S$ and $b_i = 0$ for $i \in S$ we can find a value λ of the parameter such that the specific energy $h_{\Phi_\lambda}(\psi_i)$ (in the sense of Proposition 4.2) attains its minimum over all $1 \leq i \leq r$ precisely for those $i \in S$, and for $i \notin S$ the excess $h_{\Phi_\lambda}(\psi_i) - \min_{1 \leq j \leq r} h_{\Phi_\lambda}(\psi_j)$ is exactly b_i .*

In words: For the right choice of perturbing interactions $\Psi_1, \dots, \Psi_{r-1}$, we can distort Φ to Φ_λ so that (a) for any nonempty subfamily of the ground states of Φ , the ground states of Φ_λ are precisely the members of that subfamily, and (b) the specific energies of the other ground states of Φ now exceed the specific energy of those that are still ground states of Φ_λ by any positive amounts we care to specify.

We will write V_r for the boundary of the positive quadrant in \mathbb{R}^r : $V_r = \{(b_1, \dots, b_r) : b_i \geq 0 \forall i, \min_{1 \leq i \leq r} b_i = 0\}$.

The crucial theorem is now the following: suppose we have the above situation (Sinai behaviour and complete splitting of the degeneracy by the perturbing interactions). Then

Theorem 4.4 *There exist $\beta_c > 0$ and a neighbourhood U of the origin in \mathbb{R}^{r-1} such that for any $\beta \geq \beta_c$ we have a continuous map $\phi_\beta : U \rightarrow V_r$ with the following properties:*

1. *the image $\phi_\beta(U)$ contains a neighbourhood of 0 in V_r ;*
2. *if $\lambda \in U$ has $\phi_\beta(\lambda) = (b_1, b_2, \dots, b_r)$, then for those i with $b_i = 0$ the limit Gibbs states σ_{β, ψ_i} for the perturbed interaction Φ_λ exist and are distinct pure states.*

Proof Main Theorem B in Section II.5 of Sinai [24]. □

In words: For β greater than some critical value β_c , for any nonempty subset S of $\{1, 2, \dots, r\}$ we can always perturb our basic interaction Φ using the Ψ_i to Φ_λ for a suitable λ so that all the limit Gibbs distributions σ_{β, ψ_i} for $i \in S$ with the interaction Φ_λ exist and are distinct and pure. In particular, we can choose λ so that $\phi_\beta(\lambda) = 0$, in which case *all* the limit Gibbs states σ_{β, ψ_i} for Φ_λ are distinct: the splitting of the degeneracy in the ground states that we assumed for the perturbing interactions Ψ_i results in a ‘splitting of the degeneracy’ for the corresponding limit Gibbs states.

To summarize from the last two subsections: For β small, there is only one limit Gibbs state, and in it the values at different lattice sites are nearly independent. The physical picture is that at high temperature (recall that $\beta = 1/T$) there is a lot of energy washing around, and so the different degrees of freedom can do largely what they like. However, above a certain critical value β_c (that is, *below* a certain critical temperature) the behaviour changes: the system, having less energy, prefers certain low energy configurations (the ground states) and with high probability it stays close to these. If there is only one ground state, well and good; if, however, there are several, then our limit Gibbs states describing the (homogeneous) states of our system are not unique, depending on the boundary conditions. Mathematically, the system may not just be in a state that is always close to one ground state, but may be in a ‘mixture’ of such states, so that it is always close to one or other ground state, but there may be more than one such involved. In this case we can choose suitable perturbing potentials (to be thought of in the physical situation as new control parameters for the ambient conditions of our physical system) so that, if we adjust these new parameters suitably, we can force the system to prefer one (or a few) of its ground states over the others, so that it will enter a pure homogeneous state that is close to that ground state.

It may be time to note our next notion of phase transition.

Phase transition: notion 3 In the above situation, we often find that the parameter space $[0, \infty]$ containing β splits into finitely many intervals such that the simplex $I_{\beta\Phi}$ has the same structure within each, but changes in structure (for example, changing from being a singleton to not being a singleton) across their end points. In this case we would refer to the intervals as the **phases** of the system, and would say that **phase transitions** occur at the end points of the intervals.

In the next subsection we will meet the Ising model of ferromagnetism, which will illustrate degeneracy splitting in action. Before leaving this section, we take a moment to see how the above theory can sometimes allow us to re-establish contact with our notions 1A and 1B of a phase transition from Section 2.

In that Section we thought of phase transition in a system as arising from discontinuities in the value of $\langle f, \mu_y \rangle$ as y varies continuously, for a suitable macroscopic observable f and a state of the system μ_y parameterized by y . However, in the thermodynamic formalism our picture of a phase transition does not obviously fit this pattern: instead, we look for critical values $\beta = \beta_c$ at which the structure of $I_{\beta\Phi}$ changes (typically from being a singleton at low β to a larger simplex at high β); in the case of Sinai behaviour we develop a detailed description of this structure in the more complicated high- β phases in terms of ground states and perturbations of the interaction.

Very often we can diagnose this phase transition by studying the expected values of suitably-chosen macroscopic observables for different Gibbs states; and very often these observables are almost surely determined for the *pure* Gibbs states, but not for mixtures of them. However, when the Gibbs state for $\beta\Phi$ is not unique, it is hard to see how we can interpret this in terms of a parameterization of measures $y \mapsto \mu_y$.

In *some* cases, however, we can overcome this. Recall the simple limit Gibbs states of Subsection 3.3: limits of the net of Gibbs states on finite products K^Λ given by the internal energy on these spaces alone. It can happen that these nets actually converge to one particular Gibbs state, even when $I_{\beta\Phi}$ is not a singleton (and this Gibbs state will then be invariant under any symmetries inherent in the original set-up, even when this group acts non-trivially on the set of extreme points of $I_{\beta\Phi}$). In this case we can select β for our parameter y and choose for $y = \beta > \beta_c$ the unique *simple* limit Gibbs state for our measure μ_y (choosing from among the states in $I_{\beta\Phi}$).

Now the Sinai theory of degeneracy splitting may allow us to observe a discontinuity in a parameterized measure – thus characterizing the phase transition – when that degeneracy splitting takes the following form. It is possible we can introduce a suitable new interaction Ψ (in fact, we may need several) and a new control parameter λ such that for λ in some punctured neighbourhood of 0 the Gibbs states for $\beta\Phi + \lambda\Psi$ are unique (call them $\mu_{\beta,\lambda}$) and vary continuously in both β and λ ; *but* there is a discontinuity around $\lambda = 0$. Our mental picture is that the simple limit Gibbs state $\mu_\beta = \mu_{\beta,0}$ is somehow ‘in the middle’ of the simplex $I_{\beta\Phi}$, but as $\lambda \rightarrow 0$, $\lambda \neq 0$, the uniquely determined Gibbs state $\mu_{\beta,\lambda}$ does not approach the ‘middle’ but rather tends to an extreme point of $I_{\beta\Phi}$, so that the resulting function $(\beta, \lambda) \mapsto \mu_{\beta,\lambda}$ is discontinuous around $\lambda = 0$. This behaviour will first manifest itself only above a certain critical value β_c – the infimum of those β for which $I_{\beta\Phi}$ is not just a singleton – and so we observe the phase transition as occurring at the infimal β where the map $(\beta, \lambda) \mapsto \mu_{\beta,\lambda}$ is not continuous around $\lambda = 0$.

The above description seems replete with ‘if’s: it is. In the general case we are not able to make contact with Section 2, and a substantial and compelling body of theory now exists that indicates that considering the structure of the simplex I_Φ really is the right way to formulate general results for lattice gases. However, in surprisingly many examples the above description does pertain: we will mention it briefly for the Ising model of ferromagnetism in the next subsection.

4.3 Examples

In this Subsection we describe three specific lattice gas models about which some precise results are known.

4.3.1 The Ising ferromagnet

The Ising ferromagnet is arguably the oldest and most venerable model in statistical mechanics; it has been studied extensively since Ising’s original work (which was on the one-dimensional model only) in the 1930’s. Let $K = \{-1, 1\}$: each point of the lattice has either ‘spin down’ or ‘spin up’. Let G be a connected graph on the countably infinite vertex set L with all degrees finite, and write E for its edge set. Our interaction Φ is nonzero only in the case of doubletons $\{x, y\}$ that are joined by an edge of G , in which case it is just $-\omega_x\omega_y$, so for finite $\Lambda \subset L$ we have

$$U_\Lambda(\eta) = - \sum_{\substack{x, y \in \Lambda \\ xy \in E}} \omega_x \omega_y,$$

and for $\xi \in \{-1, 1\}^{L \setminus \Lambda}$

$$W_\Lambda(\eta, \xi) = - \sum_{\substack{x \in \Lambda, y \in L \setminus \Lambda \\ xy \in E}} \omega_x \xi_y$$

(the first of these sums is clearly finite, and the second is also, by the property that all degrees of G are finite).

Results are known for the Ising ferromagnet on such a general graph G , but we will be primarily concerned with the case where G is the nearest-neighbour graph on \mathbb{Z}^n . In this case we can check easily that Φ has two periodic ground states, $\psi^+ \equiv 1$ and $\psi^- \equiv -1$, and also that the interaction satisfies Peierls' stability condition. In Section 6 we will show that for $n = 2$ the Ising ferromagnet actually exhibits Sinai behaviour: that is, that we can find limit Gibbs states with boundary conditions ψ^\pm that do describe small local distortions of ψ^\pm . In fact this is known to hold in all dimensions except one.

Most interest in the Ising model follows the introduction of a second control parameter: the ambient magnetic field h . We add to the basic interaction Φ a term depending on each ω_x separately, to be interpreted as the interaction energy of that spin with an external magnetic field of strength h :

$$U_\Lambda(\eta) = - \sum_{\substack{x, y \in \Lambda \\ xy \in E}} \omega_x \omega_y + h \sum_{x \in \Lambda} \omega_x.$$

Write $\Phi_{\beta, h}$ for this perturbed interaction at β and h .

We now specify also the chief macroscopic observable we are interested in: the magnetization of the ferromagnet, defined to be

$$m(\omega) = \begin{cases} \lim_{s \rightarrow \infty} \frac{1}{|B_s(0)|} \sum_{x \in B_s(0)} \omega_x & \text{if this limit exists} \\ 0 & \text{if it does not} \end{cases}$$

This corresponds to the physical idea of the specific magnetic field (that is, the average magnetic field per constituent lattice site) as the sum of the fields resulting from all the individual spins, divided by the number of spins.

Our physical intuition here is of a large block of metal with a magnetic spin at each lattice site in the presence of an ambient magnetic field; these spins align and so contribute an additional field to the ambient one, and the metal is magnetized. We are interested in the variation of this magnetization with temperature and ambient field strength.

The following behaviour is known. There is a critical value β_c such that for $\beta \leq \beta_c$ and any $h \in \mathbb{R}$ the Gibbs states are unique and vary (vaguely) continuously with β and h . This is true also for $\beta > \beta_c$ away from $h = 0$; in either case let us write $\sigma_{\beta, h}$ for this state.

Now, each homogeneous *pure* Gibbs state is an ergodic measure for the translations of \mathbb{Z}^n , and so for such states the magnetization m , being the limit of the ergodic averages of the random variable $\omega \mapsto \omega_0$, is defined and takes one particular value $M(\beta, h)$ $\sigma_{\beta, h}$ -almost everywhere, by the multi-parameter ergodic theorem. M varies continuously with (β, h) away from the ray $\beta > \beta_c$, $h = 0$.

However, for $\beta > \beta_c$ and $h = 0$ the homogeneous Gibbs states are not unique. In fact, M shows a jump on this line segment: for $\beta > \beta_c$ the right-hand limit $\lim_{h \downarrow 0} M(\beta, h)$ is strictly positive, and the corresponding left-hand limit, which equals minus the right-hand limit by symmetry, is thus strictly negative. We call this positive right-hand limit the **spontaneous magnetization**. The interaction with the ambient field, with control parameter h , splits the degeneracy in Sinai's sense.

The graph of M as a function of h for three different β is sketched in Figure 2.

Before leaving the Ising model for now, we remark that it connects to the Yang-Lee free energy approach to phase transitions; for the almost-sure value of the magnetization as a macroscopic observable in a Gibbs state is expressible in terms of a derivative of f . Indeed, with a little work

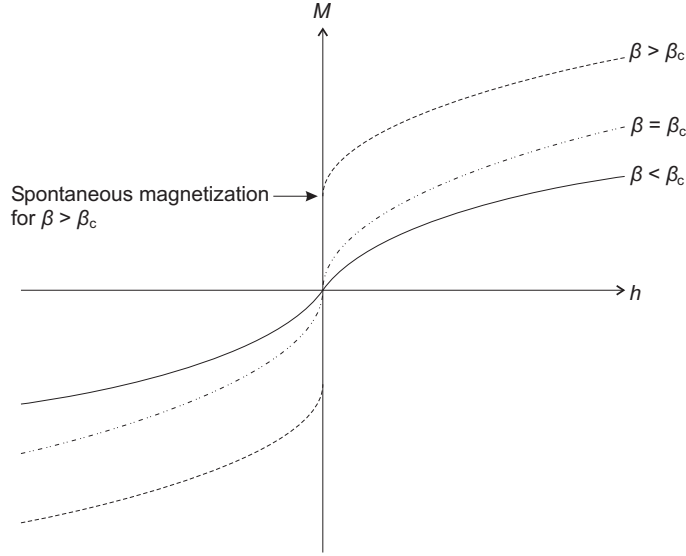


Figure 1: The spontaneous magnetization M of the Ising model as a function of h at three different values of β

it can be shown from the form of the interaction $\Phi_{\beta,h}$ that the magnetization M at (β, h) is given by

$$M = \frac{\partial f(\Phi_{\beta,h})}{\partial h};$$

thus, below β_c , we can identify the phase transition with a discontinuity in the first derivative of f with respect to h : the spontaneous magnetization is just $\lim_{h \downarrow 0} \frac{\partial f(\Phi_{\beta,h})}{\partial h}$.

The exact value of the critical inverse temperature β_c and spontaneous magnetization of the Ising model in two dimensions were obtained by Onsager in 1944; this remains one of the only exact calculations in the theory of lattice gases. Onsager's original proof was a tour de force relying on the use of the so-called 'transfer matrix'; we do not explain these here, but refer the reader to, for example, Sections II.5 and II.6 of Simon [23]. We content ourselves with stating the results:

$$\beta_c = \tanh^{-1}(\sqrt{2} - 1) = 0.44068679 \dots;$$

$$\lim_{h \downarrow 0} M(\beta, h) = \begin{cases} \left(1 - \frac{1}{(\sinh 2\beta)^4}\right)^{1/8} & \text{for } \beta > \beta_c \\ 0 & \text{for } \beta \leq \beta_c \end{cases}.$$

4.3.2 The Potts model

This is a natural generalization of the Ising model, in which K is the finite set $\{1, 2, \dots, q\}$ and the two-point nearest-neighbour part of the interaction is

$$\Phi_{\{x,y\}}(\omega) = \begin{cases} 1 & \text{if } \omega_x \neq \omega_y \\ -1 & \text{if } \omega_x = \omega_y \end{cases}.$$

values at neighbouring coordinates cost positive energy if they differ, but negative energy if they are the same.

Sinai's analysis (and, in particular, the 'Main Theorem B' of Section II.5 in [24]) applies to the Potts model also. Nevertheless, the detailed behaviour of this model is not so well understood as that of the Ising model; we mention it largely for completeness.

4.3.3 The Heisenberg ferromagnet

In many ways a more natural generalization of the Ising model than the Potts model, the Heisenberg ferromagnet is constructed on the lattice \mathbb{Z}^n with value space $K = S^{d-1}$, the sphere in d -dimensional Euclidean space, with the two-point nearest-neighbour interaction Φ given by $\Phi_{\{x,y\}}(\omega) = -\omega_x \bullet \omega_y$, where \bullet is the ordinary dot product in Euclidean space. Thus for finite $\Lambda \subset \mathbb{Z}^n$ and $\eta \in (S^{d-1})^\Lambda$

$$U_\Lambda(\eta) = - \sum_{x,y \in \Lambda, xy \in E} \omega_x \bullet \omega_y.$$

This clearly reduces to the Ising model for $d = 1$.

The physical interpretation here is that, whereas in the Ising ferromagnet the spin at each lattice site can only point either up or down, here it can point in any unit direction in a suitable Euclidean space (so, for ‘real world’ examples, d is typically 3).

The Heisenberg model for $d \geq 2$, like the Ising model, is known to exhibit a phase transition as β increases, but having a non-finite value space K it is not susceptible to the same description in terms of ground states as that due to Sinai above; in fact Chapter III of Sinai [24] is dedicated instead to models such as this where K is some compact Hausdorff space on which some compact Lie group G acts continuously. We will not describe that theory here, but will use the Heisenberg model to illustrate some other features of phase transitions that are of interest; in particular, that of spontaneous symmetry breaking.

Observe that our model is invariant under the obvious coordinatewise action of the group $O(d)$: the group acts by $(U, (\omega_x)_{x \in S}) \mapsto U\omega = (U\omega_x)_{x \in S}$ for $U \in O(d)$ and any $S \subseteq \mathbb{Z}^n$, and we have $\Phi_\Lambda(U\eta) = \Phi_\Lambda(\eta)$ for any finite subset $\Lambda \subset \mathbb{Z}^n$.

For β small the Gibbs states are unique; by the above symmetry, they too are invariant under the action of $O(d)$. However, for higher β the situation is different.

Theorem 4.5 (Fröhlich, Simon & Spencer) *There exists a critical value $\beta_c > 0$ (depending on n and d such that*

$$M = \liminf_{\Lambda \uparrow \mathbb{Z}^n} \left\langle \left\| \frac{1}{|\Lambda|} \sum_{x \in \Lambda} \omega_x \right\|^2, \sigma_{\Phi, \beta, \Lambda} \right\rangle > 0$$

whenever $\beta > \beta_0$.

Proof Theorem 3.2 in Sinai [24], or see the original paper by Fröhlich, Simon and Spencer [10]. \square

Now, the average $\frac{1}{|\Lambda|} \sum_{x \in \Lambda} \omega_x$ is the average spin over Λ (the ‘specific magnetization’ in this region). Upon working out the details of taking a homogeneous limit Gibbs state $\sigma_{\Phi, \beta}$ (which we omit here, although they are not immediately transparent; see the above references), we see that the theorem tells us that the thermodynamic limit of the specific magnetization of the model (that is, the limit of the above specific magnetization for $\Lambda \uparrow \mathbb{Z}^n$) is not almost surely zero, for the square of its length has positive expectation value. If there were only one homogeneous limit Gibbs state this would be impossible; for this state would then be ergodic with respect to translations, so by the multi-parameter ergodic theorem the thermodynamic limit specific magnetization would exist and equal a fixed constant almost everywhere, but this constant would then have to be zero by the $O(d)$ -symmetry.

In fact our situation is the following. Suppose σ is some extremal Gibbs state. Then it is ergodic with respect to translations, and so by the ergodic theorem the spontaneous magnetization $m(\omega) = \lim_{s \rightarrow \infty} \frac{1}{|B_s(0)|} \sum_{x \in B_s(0)} \omega_x$ exists and equals a fixed constant almost everywhere; call this $\phi \in \mathbb{R}^d$, with norm the above limit $M > 0$. Supposing that ϕ uniquely specifies our extremal Gibbs measure (this is in fact true), we may denote it σ_ϕ . Now our general Gibbs state σ is a mixture of these extremal Gibbs states: $m(\omega)$ exists and has norm M almost surely, but has a distribution around the sphere MS^{d-1} of radius M in \mathbb{R}^d corresponding to the measure ν occurring in the ergodic decomposition of σ :

$$\sigma = \int_{MS^{d-1}} \sigma_\phi \nu(d\phi).$$

In this case we say that the $O(d)$ symmetry has been spontaneously broken as β increases above β_c : below β_c there is only one homogeneous pure Gibbs state and it is $O(d)$ -invariant, but above β_c there are many homogeneous pure Gibbs states, none of them $O(d)$ invariant, and the general Gibbs state will be a mixture of them. Of course, now the continuous action of $O(d)$ on Ω gives rise to a non-trivial action on the set of all homogeneous pure Gibbs states.

In a real-world magnet, inevitable tiny ambient magnetic fields force the magnet modeled as above to settle in one of the homogeneous pure states, and this will be the state corresponding to the ϕ parallel to the direction of the ambient magnetic field (even though ϕ may be much longer than the strength of this ambient field). Thus the original symmetry really will appear to break: the magnet will spontaneously gain the non-zero magnetization ϕ in spite of the apparent near-exact symmetry of the original ingredients.

4.3.4 The hard core repulsive gas on a general graph

We describe briefly this lattice gas model as it has undergone some interesting study in the context of the Yang-Lee theory of phase transitions as failures of analyticity in the specific free energy f (see Subsection 3.3).

In the study of this model it is customary to replace $\beta \in [0, \infty]$ with another control parameter, the **fugacity** $w = e^{-\beta} \in [0, 1]$.

Suppose first that G is a finite graph on the vertex set L . We take $K = \{0, 1\}$, so a configuration has a natural interpretation as a subset of L , and our simple Gibbs state will be such that for a subset $X \subset L$

$$\mu_w\{X\} = \begin{cases} \frac{1}{Z_G(w)} w^{|X|} & \text{if } X \text{ is an independent subset of } L \\ 0 & \text{otherwise} \end{cases},$$

where

$$Z_G(w) = \sum_{X \subseteq L \text{ independent}} w^{|X|}$$

is the partition function. This corresponds to the interaction Φ concentrated on the two-point sets with $\Phi_{\{x,y\}}(\omega) = \infty$ if xy is an edge of G and $\omega_x = \omega_y = 1$, $\Phi_{\{x,y\}}(\omega) = w$ if xy is not an edge of G and $\omega_x = \omega_y = 1$, and $\Phi_{\{x,y\}}(\omega) = 0$ otherwise. It turns out that this is closely related to the independent set polynomial of the graph (see Bollobás [2]).

Given the above partition function, we define the specific free energy by $f_G(w) = \frac{1}{|G|} \log Z_G(w)$. Clearly for a finite graph this is an analytic function away from zeros of the polynomial Z_G . However, suppose now that G is a connected graph on a countably infinite vertex set L in which all vertices have finite degree, and that G_k is an increasing sequence of finite subgraphs converging to it. We define $f_G = \lim_{k \rightarrow \infty} f_{G_k}$, the specific free energy of G , if this limit exists. We could now study points of non-analyticity of f_G , following Yang and Lee.

A number of results are now known concerning this specific free energy function, including a surprising recent result of Scott and Sokal [22] relating conditions for Z_G not to vanish to the bounds on various conditional probabilities occurring in the Lovász Local Lemma (see, for example, Bollobás [3]) for a family of events with dependency graph G .

We will not describe these results here, our intention having been merely to show that there is worthwhile mathematics associated with our second notion of phase transition from Subsection 3.3, but with a very different feel from the results built on Ruelle's thermodynamic formalism described above, and to hint at the connections that exist between such models as this in statistical mechanics and graph theory.

There is a slightly more general model, the soft core repulsive lattice gas, to which the above analysis can be extended with a little work; again see Scott and Sokal [22].

4.3.5 One dimensional lattice gases

This section would not be complete without a mention of phase transitions in one dimensional lattice gases (that is, homogeneous lattice gases on the nearest-neighbour graph of \mathbb{Z}): there are

none. It is a surprising and substantial result that under fairly general conditions on the interaction Φ the simplex I_Φ is always a singleton; see, for example, Chapter 5 of Ruelle [21].

5 Bond percolation

In this Section we introduce another important mathematical model that exhibits a phase transition. Mercifully, bond percolation is much quicker to set up than the thermodynamic formalism; this section will be quite short, but we will have more to say about bond percolation in the next. There are no proofs in this section; the canonical reference throughout is the comprehensive treatment by Grimmett [13].

We start with a connected graph G on a countably infinite set L and with edge-set E . Our configuration space Ω will now be $\{0, 1\}^E$; this is given the product space topology as usual. For $p \in [0, 1]$ we can consider the standard product probability μ_p on this space given by allowing the edges to have the value 1 with independently probability p . We will refer to an edge with value 1 as ‘open’, otherwise it is ‘closed’. Thus a configuration in this model is a subgraph Γ of G . Given such a graph and $x \in \Lambda$, we write $C_\Gamma(x)$ for the component of x in the graph defined by the open edges of Γ . We now wish to consider the macroscopic events $A = \{\Gamma \in \Omega : \Gamma \text{ has an infinite component}\}$ and $A_x = \{\Gamma \in \Omega : |C_\Gamma(x)| = \infty\}$ for $x \in \Lambda$. It is clear that $\mu_0(A) = \mu_0(A_x) = 0$ (for then Γ is the edgeless graph on L almost surely) and $\mu_1(A) = \mu_1(A_x) = 1$ (for then $\Gamma = G$ almost surely); what is the situation for $p \in (0, 1)$?

One thing that we certainly feel should be true is monotonicity: $\mu_p(A)$ and $\mu_p(A_x)$ should both be non-decreasing in p . This is in fact true; see Section 2.1 of Grimmett [13]. It follows that there is some critical probability $p_{c,x}$ such that $\mu_p(A_x) = 0$ for $p < p_{c,x}$ but $\mu_p(A_x) > 0$ for $p > p_{c,x}$; of course, we do not know that $p_{c,x}$ is not 0 or 1.

In order to obtain more precise results than this, we restrict ourselves to the most studied special case: the nearest-neighbour graph G on \mathbb{Z}^n for $n \geq 2$. By translational symmetry, we see in this case that the **percolation probability** $\theta(p) = \mu_p(A_x)$ is independent of x , and hence so is $p_c = p_{c,x}$. In this case we find the following surprising behaviour: the critical probability actually lies strictly between 0 and 1. This is Theorem 1.10 in Grimmett [13]. A routine application of Kolmogorov’s zero-one law shows that $\mu_p(A)$ must always be 0 or 1, and so is 1 if and only if $\theta(p) > 0$. Finally, it can also be shown that for $p > p_c$ the infinite open cluster not only exists, but is unique (Theorem 8.1 in Grimmett [13]). This change in behaviour at p_c – the appearance of a single infinite component where before there were only finite components – is our next example of a phase transition.

What we are really interested in when studying bond percolation is the connectivity of the resulting graph; our rough idea of ‘macroscopic structure’ here corresponds to the size and shape of the connected components. In addition to θ , we are particularly interested in three other functions that describe this structure:

1. the **mean open cluster size at the origin**

$$\chi(p) = \int_{\{0,1\}^E} |C_\Gamma(0)| \mu_p(d\Gamma)$$

(this is clearly ∞ for $\theta(p) > 0$, hence for $p > p_c$; less obviously it turns out to be finite for $p < p_c$ and increases to ∞ as $p \uparrow p_c$; see Chapter 6 of Grimmett [13]);

2. the **mean size of the finite open cluster at the origin**, that is, the expectation of $|C_\Gamma(0)|$ conditional on the event $\{\Gamma : |C_\Gamma(0)| < \infty\}$ (when this event has positive probability)

$$\chi^f(p) = \frac{1}{\mu_p\{\Gamma : |C_\Gamma(0)| < \infty\}} \int_{\{\Gamma : |C_\Gamma(0)| < \infty\}} |C_\Gamma(0)| \mu_p(d\Gamma);$$

3. the number of open clusters per vertex

$$\kappa(p) = \int_{\{0,1\}^E} \frac{1}{|C_\Gamma(0)|} \mu_p(d\Gamma).$$

Various results are known describing the behaviour of these functions in the subcritical phase ($p < p_c$), the supercritical phase ($p > p_c$) and at the critical probability p_c . Importantly, these other functions also exhibit a phase transition at the same critical temperature p_c as θ , and are smooth (even analytic) elsewhere: we can say that the bond percolation process on \mathbb{Z}^n has only one phase transition.

Note that this phase transition essentially fits notion 1B of Section 2: here the parameter space is $[0, 1] \ni p$ and the crucial event is A . The system has the two phases (subcritical and supercritical) described above, corresponding to different values for $\mu_p(A)$, and a phase transition occurs at the boundary point p_c . As suggested above, we can find other ‘special’ functions describing the macroscopic structure, such as χ and κ , whose behaviour is discontinuous (even divergent) at p_c , again making contact with notions 1A and 1B.

We will not say more much about bond percolation now, except to mention one of the most celebrated results of the theory: the exact determination that for bond percolation on \mathbb{Z}^2 , $p_c = \frac{1}{2}$. This was proved by Kesten in 1982 (see Kesten [16]) by a beautiful argument using the self-duality of the nearest-neighbour graph on \mathbb{Z}^2 . We will meet this duality in the next section when we consider the Peierls’ argument, which will show for bond percolation on \mathbb{Z}^2 that $p_c < 1$ (Theorem 6.2); it turns out that this is crucial step in proving the upper half of the statement $0 < p_c < 1$ for bond percolation on any \mathbb{Z}^n .

More recently critical probabilities have been evaluated for a number of other planar graphs.

6 The Ising ferromagnet and bond percolation in two dimensions, and Peierls’ argument

In this Section we prove some results about two of the best understood among the models we have already met: the Ising ferromagnet and bond percolation in two dimensions. These models are both made more tractable by special properties of the nearest-neighbour graph on \mathbb{Z}^2 ; we have extracted the analyses that follow into a separate section to show how essentially one beautiful argument (due to Peierls in [20]) tells us something about both. We write G for the nearest-neighbour graph on \mathbb{Z}^2 .

The two theorems we are going to prove are the following:

Theorem 6.1 *For the Ising ferromagnet in two dimensions there exists $\beta_c > 0$ such that for $\beta > \beta_c$ there exist at least two homogeneous pure Gibbs states. As $\beta \rightarrow \infty$ these Gibbs states describe small local distortions of the ground states ψ^+ and ψ^- (see the first part of Subsection 4.3).*

Theorem 6.2 *For bond percolation in two dimensions there exists $p_c \in (0, 1)$ such that $\theta(p) > 0$ for $p > p_c$; in fact $\theta(p) \uparrow 1$ as $p \uparrow 1$.*

The point is that the idea behind both of these proofs is the same: Peierls’ argument ([20]), originally developed for the Ising model, was one of the earliest major contributions to the subject of phase transitions. Strangely, it seems to be rarely explained outside original papers; Sinai [24] gives a rather unclear treatment (Section II.4), or see Subsection 2.1.4 of Grosse [14] and the references listed there.

Before broaching the substance of the proofs, we need a run up, particularly to Theorem 6.1. In the case of the Ising model, for each finite $\Lambda \subset L$ we know how to construct the simple limit Gibbs state $\sigma_{\Lambda, \beta, \psi^\pm}$ with boundary condition ψ^\pm on $\{-1, 1\}^\Lambda$. For each β we select a single limit Gibbs state σ_{β, ψ^\pm} that is a thermodynamic limit of these states.

It is these limit Gibbs states $(\sigma_{\beta, \psi^\pm})_{\beta \geq 0}$ that we will show describe small local distortions of ψ^\pm . Observe that there is potentially some arbitrariness in our choice of σ_{β, ψ^\pm} : for a given β we do not know that our net has only one limit, and even if we were to select an increasing subsequence $\Lambda_k \uparrow L$ such that the sequence $\sigma_{\Lambda_k, \beta, \psi^\pm}$ converges for one β , we would not know that it converges for other values of β . This apparent imprecision will not matter for the proof that follows below, but for peace of mind we note that it is spurious: in fact for any interaction in any number of dimensions with finitely many ground states and satisfying Peierls' stability condition it is possible to prove that the nets do all have a unique limit, and also that they describe small local distortions of those ground states. However, this requires a very detailed argument (in fact a far-reaching extension of Peierls' argument as we will present it below) called the 'method of contours', for which we have not the space here; it is explained in Sections II.6 through II.10 of Sinai [24].

We let μ_p denote the measure on $\{0, 1\}^E$ corresponding to the bond percolation model with parameter p .

Next, in order to understand the proof of Theorems 6.1 and 6.2, we must consider contours in the dual graph G_d to G . This has vertices at the points of $\mathbb{Z}^2 + (\frac{1}{2}, \frac{1}{2})$ and edges unit line segments that perpendicularly bisect the edges of G . It is clear that G_d is just G translated by $(\frac{1}{2}, \frac{1}{2})$ (G is said to be **self-dual**). A **contour** γ in G_d is a subgraph of G_d that is a self-avoiding closed loop. Intuitively it is clear that such a contour separates finitely many vertices of \mathbb{Z}^2 'from infinity'. Without stopping to make this rigorous (it is an exercise in tedious detail; see p.386 of Kesten [17] for a careful treatment in the bond percolation context) we will write this subset as $\Lambda(\gamma)$.

Contours in G_d arise naturally in both of our models, in different ways:

1. Suppose $\omega \in \{-1, 1\}^{\mathbb{Z}^2}$ is a configuration in the Ising model. In any edge xy of G , ω_x and ω_y are either the same or they are different; we define the **boundary** of ω , $\partial\omega$, to be the set of such edges with $\omega_x \neq \omega_y$, and the **dual boundary**, $\partial_d\omega$, to be the set of all edges of G_d that cross edges of $\partial\omega$. See Figure 2.
2. Suppose now that $\Gamma \in \{0, 1\}^E$ is a graph resulting from the bond percolation model. We define its **G -dual**, $D\Gamma$, to be the graph formed from those edges of G_d that do *not* cross edges of Γ . See Figure 3.

Peierls' idea was that we can understand the random image of \pm 's arising in the Ising model by studying what contours γ can appear in the dual boundary $\partial_d\omega$ of a configuration. It turns out that we can use contours similarly for the study of bond percolation.

Henceforth we will work with the case of ψ^+ for the Ising model; the other case is clearly similar by symmetry. We want to prove that for sufficiently large β the set $\{x \in \mathbb{Z}^2 : \omega_x \neq 1\}$ is broken up into countably many finite 'islands' for σ_{β, ψ^+} -almost every ω , and that these disperse and decrease in size as $\beta \rightarrow \infty$ in the sense made precise by the notion of small local distortions. This will follow from:

Proposition 6.3 (Peierls' inequality) *Let γ denote a fixed contour and for finite $\Lambda \subset L$ identify $\sigma_{\Lambda, \beta, \psi^+}$ with its extension to all of $\{-1, 1\}^{\mathbb{Z}^2}$ that almost surely gives the value 1 on any coordinate outside Λ . Then for any Λ*

$$\sigma_{\Lambda, \beta, \psi^+} \{\omega : \gamma \subseteq \partial_d\omega\} \leq e^{-2\beta|\gamma|},$$

writing $|\gamma|$ for the length of γ .

Proof If any edge of γ is not adjacent to some point of Λ then the probability is clearly 0, so we suppose they all are. We compute for $\eta \in \{-1, 1\}^\Lambda$

$$H_\Lambda(\eta) = U_\Lambda(\eta) + W_\Lambda(\eta, 1) = - \sum_{\substack{x, y \in \Lambda \\ xy \in E}} \eta_x \eta_y - \sum_{\substack{x \in \Lambda, y \in \mathbb{Z}^2 \setminus \Lambda \\ xy \in E}} \eta_x$$

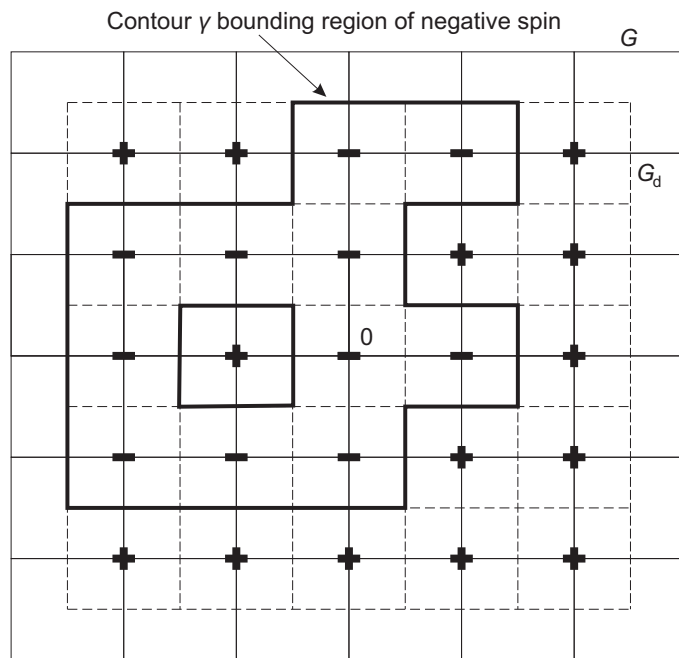


Figure 2: A contour γ within $\partial_d \omega$ in the Ising model

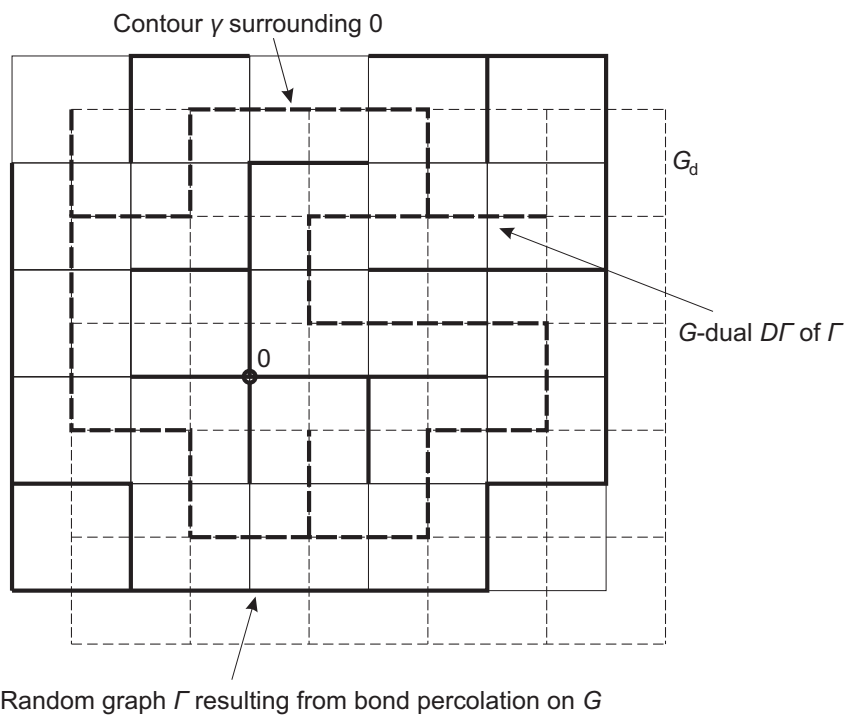


Figure 3: A contour γ of $D\Gamma$ surrounding 0

$$\begin{aligned}
&= -|\{xy \in E : (\eta \frown 1)_x = (\eta \frown 1)_y\}| \\
&\quad + |\{xy \in E : (\eta \frown 1)_x \neq (\eta \frown 1)_y\}| \\
&= -|\Lambda| + 2|\partial_d(\eta \frown 1)|
\end{aligned}$$

(recalling that $\eta \frown 1$ denotes the configuration $\omega \in \{-1, 1\}^{\mathbb{Z}^2}$ equal to η inside Λ and 1 outside), by some easy re-counting. Hence, from our definition of simple Gibbs state,

$$\begin{aligned}
\sigma_{\Lambda, \beta, \psi+} \{\omega : \gamma \subseteq \partial_d \omega\} &= \frac{\sum_{\eta \in \{-1, 1\}^\Lambda, \gamma \subseteq \partial_d(\eta \frown 1)} e^{-\beta H_\Lambda(\eta)}}{\sum_{\eta \in \{-1, 1\}^\Lambda} e^{-\beta H_\Lambda(\eta)}} \\
&= \frac{\sum_{\eta \in \{-1, 1\}^\Lambda, \gamma \subseteq \partial_d(\eta \frown 1)} e^{-2\beta |\partial_d(\eta \frown 1)|}}{\sum_{\eta \in \{-1, 1\}^\Lambda} e^{-2\beta |\partial_d(\eta \frown 1)|}},
\end{aligned}$$

taking out a common factor of $e^{-\beta |\Lambda|}$.

Now write A_γ for the set of all configuration ω with $\gamma \subseteq \partial_d \omega$, and B_γ for the set of ω with $\gamma \cap \partial_d \omega = \emptyset$. Our crucial observation is that we can define a bijection $\tau_\gamma : A_\gamma \rightarrow B_\gamma$ by the condition that $\tau_\gamma(\omega)_x = -\omega_x$ if $x \in \Lambda(\gamma)$ and otherwise $\tau_\gamma(\omega)_x = \omega_x$. It is clear that if γ is a part of the dual boundary of $\omega \in A_\gamma$, then is removed from the dual boundary of $\tau_\gamma(\omega) \in B_\gamma$; that is, $\partial_d \omega = \partial_d(\tau_\gamma(\omega)) \cup \gamma$ and $|\partial_d \omega| = |\partial_d(\tau_\gamma(\omega))| + |\gamma|$. Therefore

$$\begin{aligned}
\sigma_{\Lambda, \beta, \psi+} \{\omega : \gamma \subseteq \partial_d \omega\} &= \frac{\sum_{\eta \in \{-1, 1\}^\Lambda, (\eta \frown 1) \in A_\gamma} e^{-2\beta |\partial_d(\eta \frown 1)|}}{\sum_{\eta \in \{-1, 1\}^\Lambda} e^{-2\beta |\partial_d(\eta \frown 1)|}} \\
&= e^{-2\beta |\gamma|} \frac{\sum_{\eta \in \{-1, 1\}^\Lambda, (\eta \frown 1) \in B_\gamma} e^{-2\beta |\partial_d(\eta \frown 1)|}}{\sum_{\eta \in \{-1, 1\}^\Lambda} e^{-2\beta |\partial_d(\eta \frown 1)|}} \leq e^{-2\beta |\gamma|},
\end{aligned}$$

as required. \square

In the same spirit as the above, and to continue the analogy between the Ising model and bond percolation, we observe that given bond percolation with parameter $p \in [0, 1]$ yielding a random subgraph Γ of G , the resulting random subgraph $D\Gamma$ of G_d obeys the law of bond percolation with parameter $1 - p$; for an edge of $D\Gamma$ is present if and only if the edge of Γ that it could cross is absent, and for different edges these are independent events all with probability $1 - p$. We now see at once the truth of the following:

Proposition 6.4 *The following holds for all contours γ and $p \in [0, 1]$:*

$$\mu_p \{\Gamma : \gamma \subseteq D\Gamma\} = (1 - p)^{|\gamma|}.$$

\square

Thus we have two propositions yielding exponential decay with $|\gamma|$ for the probabilities of a given contour γ arising in natural constructions from our models ($\partial_d \omega$ and $D\Gamma$), with the decay rate tending to ∞ as either $\beta \rightarrow \infty$ or $p \uparrow 1$.

Let us consider now what it is we want to prove for our two models:

1. For the Ising model, we wish to show firstly that for sufficiently large β there is almost surely *no* infinite connected ‘island’ of -1 ’s, and secondly that the probability of a configuration differing from 1 anywhere on a square block $B_s(x)$ of side-length $2s$ decays to 0 as $\beta \rightarrow \infty$ uniformly over all blocks of this size. The handle we need here is that:

- (a) if ω contains an infinite connected island of -1 ’s, then as the finite set Λ increases to \mathbb{Z}^2 the configurations $\omega|_\Lambda \frown 1$ obtained by replacing all terms of ω with 1 outside Λ will be such that $\partial_d \omega$ contains arbitrarily long contours, and;

(b) if we know the -1 's appear only in finite connected islands then any given vertex, say 0 , can lie in a finite connected island of -1 's only if it lies in $\Lambda(\gamma)$ for some contour in $\partial_d \omega$ (although this is not an equivalent condition : look at the isolated '+' in Figure 2, and so the probability of the former is bounded by the probability of the latter.

2. For bond percolation, we wish to show that for sufficiently large p there is almost surely an infinite connected component in Γ , and that as $p \uparrow 1$ the probability of a given vertex lying in this component increases to 1. Similarly to the above case, contours give us a handle for attacking this problem: any given vertex, say 0 , lies in an infinite connected component of Γ if and only if it does not lie in $\Lambda(\gamma)$ for any contour $\gamma \subset D\Gamma$.

Note that we have had to break the result for the Ising model into two parts, whereas the bond percolation case has only one; this is merely a quirk arising from the precise problems at hand, and not something to worry about.

The proofs of both theorems will hinge on our exponential decay results for contour containment. So let us prove them.

Proof of Theorem 6.1 As indicated above, there are two stages.

Consider first the set $A = \{\omega : 0 \text{ lies in an infinite island of } -1\text{'s}\}$. If $\omega \in A$ then as $s \rightarrow \infty$ we see that the configurations $\omega^{(s)} = \omega|_{B_s(0)} \widehat{1}$ are such that $\partial_d \omega^{(s)}$ contain increasingly large contours that surround 0 (as $B_s(0)$ hits more and more of $\partial_d \omega$). For $m \geq 1$ and $s \geq 1$ let $A_{s,m}$ be the set $\{\omega : \partial_d \omega^{(s)}$ contains a contour of length at least m surrounding $0\}$; these $A_{s,m}$ are increasing in s and decreasing in m . Then

$$\sigma_{\beta, \psi^+}(A) \leq \sigma_{\beta, \psi^+} \left(\bigcap_{m \geq 1} \bigcup_{s \geq 1} A_{s,m} \right) = \lim_{m \rightarrow \infty} \sigma_{\beta, \psi^+} \left(\bigcup_{s \geq 1} A_{s,m} \right) \leq \lim_{m \rightarrow \infty} \sup_{s \geq 1} \sigma_{\beta, \psi^+}(A_{s,m}).$$

Now, each event $A_{s,m}$ depends on only finitely many coordinates (those in $B_s(0)$), and so for a suitable increasing sequence of finite subsets Λ_k of \mathbb{Z}^2 we have

$$\sigma_{\beta, \psi^+}(A_{s,m}) = \lim_{k \rightarrow \infty} \sigma_{\Lambda_k, \beta, \psi^+}(A_{s,m})$$

(by the definition of the vague topology). For k sufficiently large we have $\Lambda_k \supseteq B_{s+2}(0)$; thus for such k we see that there is a contour in $\partial_d(\omega|_{B_s(0)} \widehat{1})$ of length at least m and surrounding 0 only if either 0 is in an infinite connected island of -1 's in the configuration ω (which has probability zero for any $\sigma_{\Lambda_k, \beta, \psi^+}$, since for this measure every value outside Λ_k is almost surely equal to 1) or there is a contour in the original $\partial_d \omega$ of length at least m and surrounding 0 . Therefore, by Proposition 6.3,

$$\begin{aligned} \sigma_{\Lambda_k, \beta, \psi^+}(A_{s,m}) &\leq \sum_{r \geq m} \sum_{\substack{\text{contour } \gamma \text{ surrounds } 0 \\ |\gamma| \geq m}} \sigma_{\Lambda_k, \beta, \psi^+} \{\omega : \gamma \subseteq \partial_d \omega\} \\ &\leq \sum_{r \geq m} C_r e^{-2\beta r} \end{aligned}$$

where C_r is the number of possible contours of G_d of length r and containing 0 . A crude count of C_r can be performed using the facts that any such γ must cross the positive axis within distance r of 0 and that there are at most 3^r contours of length r passing through a given vertex (for if we traverse the contour, at any point of its length it can be continued in at most 3 ways); we find $C_r \leq r3^r$, and hence

$$\sigma_{\Lambda_k, \beta, \psi^+}(A_{s,m}) \leq \sum_{r \geq m} r3^r e^{-2\beta r}.$$

Now we need only choose β so large that $3e^{-2\beta} < 1$ to see that the first part of the theorem follows from the above inequalities, and the fact that the set $\{\omega : \text{there is an infinite connected island of } -1\text{'s}\}$ is the countable union of the sets $\{\omega : x \text{ lies in an infinite connected island of } -1\text{'s}\}$ for $x \in \mathbb{Z}^2$.

The second part follows in short order: consider the probability $\sigma_{\beta,\psi^+}\{\omega : \omega_0 = -1\}$. We know from the above that there are no infinite connected islands of -1 's; therefore this probability equals $\sigma_{\beta,\psi^+}\{\omega : \text{some contour } \gamma \subseteq \partial_d\omega \text{ surrounds } 0\}$, and the same inequalities as above yield for this the upper bound

$$\sum_{\substack{\text{contour } \gamma \text{ surrounds } 0 \\ |\gamma| \geq m}} \sigma_{\beta,\psi^+}\{\omega : \gamma \subseteq \partial_d\omega\} \leq \sum_{r \geq 1} r 3^r e^{-2\beta r},$$

which tends to 0 as $\beta \rightarrow \infty$. Therefore for any finite set Λ the probability

$$\sigma_{\beta,\psi^+}\{\omega : \omega|_{\Lambda} \not\equiv 1\} \leq \sum_{x \in \Lambda} \sum_{\substack{\text{contour } \gamma \text{ surrounds } x \\ |\gamma| \geq m}} \sigma_{\beta,\psi^+}\{\omega : \gamma \subseteq \partial_d\omega\}$$

also tends to 0 uniformly in $|\Lambda|$, so the result follows. \square

Proof of Theorem 6.2 The similarity of this proof to that above should quickly become apparent. We want to show that

$$\mu_p\{\Gamma : 0 \text{ lies in an infinite component of } \Gamma\} \rightarrow 1$$

as $p \uparrow 1$, or, equivalently, that

$$\mu_p\{\Gamma : 0 \text{ is surrounded by some contour } \gamma \text{ of } D\Gamma\} \rightarrow 0.$$

By Proposition 6.4 this latter probability is bounded by

$$\sum_{\text{contour } \gamma \text{ surrounds } 0} \mu_p\{\Gamma : \gamma \subseteq D\Gamma\} \leq \sum_{m \geq 1} C_m (1-p)^m \leq \sum_{m \geq 1} m(3(1-p))^m,$$

which we see tends to 0 as $p \uparrow 1$. This completes the proof. \square

7 Other important ideas

An essay as short as this necessarily leaves out more than it can include. There are many mathematical models, both in the mathematical theory of statistical mechanics and elsewhere, that have been left out.

Within statistical mechanics, there has been growing interest in the models called spin glasses during the last ten years; these are like lattice gas models, but have an additional randomness in the interaction itself. An introduction to spin glasses written by Michel Talagrand appears in the proceedings of the 2000 Saint Flour Summer School in Probability, and is also available to download from his website at

www.math.ohio-state.edu/~talagran/expository/index.html.

In addition, classical statistical mechanics is known to have some surprising connections with quantum field theory. Certain gauge theories involve models built using a version of the thermodynamic formalism, and may involve their own phase transitions.

Further afield, there are several models loosely related to bond percolation that may exhibit phase transitions of some kind. In particular we mention site percolation (in one sense a generalization of bond percolation); fractal percolation for subsets of the plane; and the random-cluster model, a very general model built on a graph that reduces to bond percolation, the Ising model and the Potts model at different values of the one of the input parameters, thereby strengthening our understanding of the relationships among the three. A good introduction to these models and several further references can be found in the Chapters 12 and 13 of Grimmett [13].

In addition to other models, there are also techniques for studying phase transitions that we have yet not mentioned in this essay. As things stand, many of these techniques are very heuristic, and are much more accessible in the physics literature than the mathematical. The oldest technique is ‘mean field theory’, dating to the 1930’s, which essentially amounts to approximating a complicated interaction with a suitable ‘linearized’ version. Mean field theory contributed greatly to the the early qualitative understanding of various models, but give few quantitative results and poor bounds. More recently known results have been considerably improved by the introduction of ‘renormalization theory’, which we take the space to describe briefly below.

We will also give a quick account of the idea of ‘universality’, which, if true, would give us a much more detailed understanding of phase transitions in different models.

A good physics book that touches on both mean field theory and renormalization, as well as other recent developments, is Goldenfield [12].

7.1 Renormalization

Renormalization theory is relatively new, having first arisen in particle physics during the 1960’s and then been imported into statistical physics by K. Wilson (see Wilson [25] for a subsequent account). Frustratingly, in spite of great effort, renormalization continues to resist attempts to make it rigorous. Sinai [24] gives some tentative first steps for a theory of renormalization of lattice gas models with value space \mathbb{R} and a Gaussian distribution, but it is not very clear where he is going; and a few specific results can be formulated and proved using the basic ideas of ‘scaling method’ and ‘block transformation’ that underlies renormalization (see Chapters 9 and 10 of Grimmett [13], for example); but the method remains largely a heuristic. A friendly recent treatment from the physicist’s point of view is Cardy’s book [4].

We will try to illustrate the basic ideas of scaling and renormalization for the lattice gas, although these methods have been used to attack other models also.

Consider some homogeneous Gibbs state σ_β at inverse temperature β for a lattice gas on $\Omega = K^L$. Suppose further that we have found some self-map $R : \Omega \rightarrow \Omega$ with the following two properties:

1. The image measure $\sigma_\beta \circ R^{-1}$ is a homogeneous Gibbs state at a *different inverse temperature*, say $R_1(\beta)$;
2. The macroscopic structure of a configuration $\omega \in \Omega$ (that is, the collection of values taken by those macroscopic observables that we care about) is the same for $R(\omega)$.

In this case, we infer that β and $R_1(\beta)$ lie in the same temperature phase: for by (1) the state of the lattice gas having been acted upon by R is also a Gibbs state and so we can still talk about its macroscopic properties, and then by (2) these must be the same as for σ_β . Thus the map R_1 gives us a self map of the parameter space $[0, \infty]$ that leaves different phases invariant. We then hope that we can determine the critical values β_c by studying this self-map; for example, if our model is known to have only two phases, one for low β and one for high β , then we know that the critical value β_c must be a fixed point for R_1 . Were it to happen that R_1 has only one fixed point between 0 and ∞ , we would know that it was the critical value.

The problem is that of the two conditions on R above, 1 is almost impossible to meet and 2 is usually very hard to make precise and prove. Nevertheless, numerical results have been obtained by assumed that the image measure $\sigma_\beta \circ R^{-1}$ is ‘close enough’ to another Gibbs state and then letting $R_1(\beta)$ be a suitably chosen inverse temperature such that the associated Gibbs state $\sigma_{R_1(\beta)}$ gives an adequate such approximation.

For example, in the two dimensional Ising model we can use the **scaling transformation** defined as follows. Pick b a natural number, at least 2, and break \mathbb{Z}^n into blocks of side length b :

$$\Lambda_{x,b} = \{z \in \mathbb{Z}^n : bx_i \leq z_i < bx_{i+1} \ \forall i \leq n\}$$

for $x \in \mathbb{Z}^n$, and define the self-map $R_b : \Omega \rightarrow \Omega$ by

$$(R_b(\omega))_x = \begin{cases} 1 & \text{if } \sum_{z \in \Lambda_{x,b}} \omega_z \geq 0 \\ 0 & \text{if } \sum_{z \in \Lambda_{x,b}} \omega_z < 0 \end{cases}$$

(note that if we take b odd then we can never have $\sum_{z \in \Lambda_{x,b}} \omega_z = 0$).

It turns out that the new state that results from applying this transformation to a suitable σ_β is ‘close to’ $\sigma_{R_{b,1}(\beta)}$ with

$$R_{b,1}(\beta) = 2\beta \left(\frac{e^{3\beta} + e^{-\beta}}{e^{3\beta} + 3e^{-\beta}} \right)^2;$$

this map has only three fixed points: $0, \infty$, and a third at approximately 0.34. This last is taken as our approximate value of the critical inverse temperature; it is not very close to the value of about 0.44 obtained by Onsager (see our discussion of the Ising model in Subsection 4.3), but it is also not hopelessly far away.

It is hoped that renormalization methods can be constructed with more care that will yield much better approximations to various critical values through their fixed points, and also that the validity or otherwise of renormalization methods involving scaling transformations such as the above will shed some light on the phenomenon of universality. Much work remains to be done.

7.2 Universality

So far we have not mentioned at all one of the most mysterious properties of phase transitions: in many systems, the behaviour of the interesting macroscopic observables as the control parameters approach a critical value seems to obey a (usually non-integral) power law and, furthermore, the exponents that arise in this way seem to be the same for large families of models, many of them not obviously closely related.

For example, consider bond percolation on \mathbb{Z}^n , $n \geq 2$. We know that $\theta(p) = 0$ for $p < p_c$ and $\theta(p) > 0$ for $p > p_c$, and similarly that $\chi(p) < \infty$ for $p < p_c$ and $\chi(p) \rightarrow \infty$ as $p \uparrow p_c$, but much more is conjectured: that as $p \downarrow p_c$ (resp. $p \uparrow p_c$), $\theta(p)$ (resp. $\chi(p)$) behaves as a power in $|p - p_c|$; that is, the limits

$$\lim_{p \downarrow p_c} \frac{\log \theta(p)}{\log(p - p_c)}, \quad - \lim_{p \uparrow p_c} \frac{\log \chi(p)}{\log |p - p_c|}$$

exist (and similarly for various other functions of p naturally arising in the description of the component structure of the random graph).

Furthermore, it is even conjectured that these so-called **critical exponents** are the same for all other bond percolation processes on lattice graphs in \mathbb{Z}^n : they depend only on the dimension $n \geq 2$, and not on the exact form of the lattice. This is the origin of the name ‘universality’.

Similar conjectures grouping various related models into the same ‘universality class’ have been proposed elsewhere; for example, various power law relationships with critical exponents depending only on the dimension n have been proposed for large classes of interaction for the lattice gas in \mathbb{Z}^n .

At present, however, very little has been proved. There is some hope that the scaling methods mentioned in the previous subsection will provide a way of proving the existence of, and even relations among, these critical exponents, as in many models scaling transformations involve their own specific exponents in some kind of normalizing constant.

For a summary of what is currently known for bond percolation see Chapter 9 of Grimmett [13], and for non-rigorous results for models in statistical mechanics see Goldenfield [12].

8 Epilogue: phase transitions in combinatorics

It seems inadequate to end this essay without mentioning another setting in which the phrase ‘phase transition’ occurs in pure mathematics: probabilistic combinatorics.

Since the groundbreaking work of Erdős and Rényi in the 1950's and 60's, the 'probabilistic method' has become one of the most important and powerful ideas in combinatorics. It has yielded stunning solutions to a number of hard problems in traditional (deterministic) combinatorics, in addition to spawning the study of random combinatorial structures themselves, which now have a mathematical life all their own.

One major area of study within probabilistic combinatorics has been the limiting behaviour of certain random graphs as their size tends to ∞ . The classic such random graph is $\mathcal{G}(n, p)$, the random graph on n vertices constructed by including each possible edge independently with probability p . We allow p to depend on n , and see what structural results we can prove for $\mathcal{G}(n, p)$ for suitable sequences $p(n)$ as $n \rightarrow \infty$. I have based the brief description that follows on the very nice treatment in Chapter 10 of the book by Alon, Spencer and Erdős [1]; a more detailed exposition can also be found in the standard reference for random graphs, Bollobás [3].

In their original 1960 paper ([8]), Erdős and Rényi discovered that the structure of the connected components of $\mathcal{G}(n, p)$ changes sharply at $p = 1/n$, in the following way. If $p = c/n$ with $c < 1$, then (with probability increasing to 1) as $n \rightarrow \infty$ $\mathcal{G}(n, p)$ will decompose into small components, the largest of which has size of order $\log n$.

On the other hand for any $c > 1$, with only a slightly higher probability of including any given edge, many of these small components will 'have been joined': as $n \rightarrow \infty$ there will now be (with probability increasing to 1) one giant component with size of order n , together with a few remaining little components with sizes of order at most $\log n$.

This phenomenon is referred to as a phase transition at $p = 1/n$. It bears an apparent resemblance to both the Ising model and bond percolation:

- In the Ising model, we know that in the unique homogeneous Gibbs states for $\beta < \beta_c$ the $+$'s and $-$'s occur almost independently: neither of the sets $\{x : \omega_x = 1\}$ and $\{x : \omega_x = -1\}$ shows any particular structure, and certainly neither of them dominates the whole of \mathbb{Z}^n . On the other hand, for $\beta > \beta_c$ we have two homogeneous pure Gibbs states describing small local distortions of the configuration containing only $+$'s and that containing only $-$'s; for the former, our configuration will appear as a giant connected 'sea' of $+$'s dotted with small 'islands' of $-$'s. This situation for the latter is clearly the reverse. Order, in the form of a giant connected component of either spin 'up' or spin 'down', has emerged.
- In bond percolation on \mathbb{Z}^n , the resemblance is even more striking: for p below the critical probability p_c , the random graph generated breaks up into many finite connected components almost surely, but above p_c one infinite ('giant') connected component emerges, with the rest of the lattice taken up by finite components.

However, notwithstanding this similarity, I am not aware of any major features that are common to the analysis of either of the above models and the analysis of $\mathcal{G}(n, p)$. There does not seem to be even a sensible notion of 'thermodynamic limit' as $n \rightarrow \infty$ for the random graphs $\mathcal{G}(n, p)$, and so it is hard even to put their study into the same language of parameterized families of measures as the Ising model (σ_β and friends) and bond percolation (μ_p). The curious distance between phase transitions in combinatorics and in the models considered earlier in this essay may be worthy of further study, but it is not at all clear what a unification would look like.

References

- [1] Alon N., Erdős P. & Spencer J.H., *The Probabilistic Method*, Wiley, New York, 1992;
- [2] Bollobás B., *Modern Graph Theory*, Springer, Berlin, 1998;
- [3] Bollobás B., *Random Graphs*, 2nd Ed., Cambridge University Press, Cambridge, 2001;
- [4] Cardy J., *Scaling and Renormalization in Statistical Physics*, Cambridge University Press, Cambridge, 1996;

- [5] Dobrushin R.L., “Problem of uniqueness of a Gibbs random field and phase transitions”, *Funkts. Anal. Prilozh.* 2 (1968), 44 – 57;
- [6] Dobrushin R.L., “The description of the random field by its conditional distributions and its regularity conditions”, *Teor. Veroyatn. Primen.* 2 (1968), 201 – 229;
- [7] Dobrushin R.L., “Gibbs field: the general case”, *Funkts. Anal. Prilozh.* 3 (1969), 27 – 35;
- [8] Erdős P. & Rényi A., “On the evolution of random graphs”, *Magyar Tud. Acad. Mat. Kut. Int. Közl* 5 (1960), 17 – 61;
- [9] Fremlin D.H., *Measure Theory*, Volume 4, Torres Fremlin, Colchester, 2003;
- [10] Fröhlich J., Simon B. & Spencer T., “Infrared bounds, phase transitions and continuous symmetry breaking”, *Comm. Math. Phys.* 50 (1976), 79 – 85;
- [11] Gallavotti G., *Statistical Mechanics: A Short Treatise*, Springer, Berlin, 1999;
- [12] Goldenfield N., *Lectures on Phase Transitions and the Renormalization Group*, Addison-Wesley, Reading, Massachusetts, 1992;
- [13] Grimmett G., *Percolation*, 2nd Ed., Springer, Berlin, 1999;
- [14] Grosse H., *Models in Statistical Physics and Quantum Field Theory*, Springer, Berlin, 1988;
- [15] Kallenberg O., *Foundations of Modern Probability*, Springer, Berlin, 2002;
- [16] Kesten H., “The critical probability of bond percolation on the square lattice equals $\frac{1}{2}$ ”, *Comm. Math. Phys.* 74 (1980), 41 – 59;
- [17] Kesten H., *Percolation Theory for Mathematicians*, Birkhäuser, Boston, 1982;
- [18] Leblond J., Meijer P.H.E., & Papon P., *The Physics of Phase Transitions*, Springer, Berlin, 2002;
- [19] Lee T.D. & Yang C.N., “Statistical theory of equations of state and phase transitions, I: Theory of condensation”, *Phys. Rev.* 87 (1952), no.3, 404 – 409;
- [20] Peierls R., “On Ising’s model of ferromagnetism”, *Proc. Cambridge Philos. Soc.* 36 (1936), 477 – 481;
- [21] Ruelle D., *Thermodynamic Formalism*, 2nd Ed., Cambridge University Press, Cambridge, 2004;
- [22] Scott A.D. & Sokal A.D., “The repulsive lattice gas, the independent set polynomial, and the Lovász Local Lemma”, *J. Statist. Phys.* 118 (2005), 1151 – 1261, con-mat/0309352 at arXiv.org;
- [23] Simon B., *The Statistical Mechanics of Lattice Gases*, Princeton University Press, Princeton, 1993;
- [24] Sinai Ya.G., *Theory of Phase Transitions: Rigorous Results*, Akadémiai Kiadó (co-edition with Pergamon Press Ltd.), Budapest, 1982;
- [25] Wilson K.G., “The renormalization group and critical phenomena”, *Review of Modern Physics* 55 (1983), 583 – 600.